IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

ADAM-1: An AI Reasoning and Bioinformatics Model for Alzheimer's Disease Detection and Microbiome-Clinical Data Integration

Ziyuan Huang[145], Vishaldeep Kaur Sekhon[2], Roozbeh Sadeghian[3], Maria L. Vaida[3], Cynthia Jo[4], Beth A. McCormick[15], Doyle V. Ward[15], Vanni Bucci[15], & John P. Haran[145]

[1] Department of Microbiology, UMass Chan Medical School, Worcester, MA, 01655 USA
[2] Department of Geriatric Medicine and Gerontology, Johns Hopkins University, Baltimore, MD, 21224 USA
[3] Data Sciences, Harrisburg University of Science and Technology, Harrisburg, PA, 17101 USA
[4] Department of Emergency Medicine, UMass Chan Medical School, Worcester, MA, 01655 USA
[5] Program in Microbiome Dynamics, UMass Chan Medical School, Worcester, MA, 01655 USA

Corresponding author: Ziyuan Huang (e-mail: ziyuan.huang2@umassmed.edu).

This work was supported by the National Institutes of Health (NIH) under Grant R01AG067483-01. The project was conducted within the Haran Research Group at the University of Massachusetts Chan Medical School, under the direction of Principal Investigator John P. Haran.

ABSTRACT Alzheimer's Disease Analysis Model Generation 1 (ADAM-1) is a multi-agent reasoning large language model (LLM) framework designed to integrate and analyze multimodal data, including microbiome profiles, clinical datasets, and external knowledge bases, to enhance the understanding and classification of Alzheimer's disease (AD). By leveraging the agentic system with LLM, ADAM-1 produces insights from diverse data sources and contextualizes the findings with literature-driven evidence. A comparative evaluation with XGBoost revealed a significantly improved mean F1 score and significantly reduced variance for ADAM-1, highlighting its robustness and consistency, particularly when utilizing human biological data. Although currently tailored for binary classification tasks with two data modalities, future iterations will aim to incorporate additional data types, such as neuroimaging and peripheral biomarkers, and expand them to predict disease progression, thereby broadening ADAM-1's scalability and applicability in AD research and diagnostic applications.

INDEX TERMS alzheimer's disease, artificial intelligence, multi-agent systems, knowledge based systems.

I. INTRODUCTION

The integration of multimodal data sources in biomedical research has accelerated with advances in deep learning, particularly through the development of large language models (LLMs). The introduction of AlexNet in 2012 demonstrated the potential of deep neural networks for complex tasks such as image classification [1]. Subsequent developments, including transformer architecture in 2017 and the release of models like GPT-1 in 2018 and GPT-2 in 2019, marked significant milestones in natural language processing [2, 3]. Other transformer-based LLM families, such as LLaMA, Gemini, Claude, and DeepSeek, have also demonstrated powerful structured and unstructured biomedical data processing [4-10]. More recently, reasoning-focused LLMs, including GPT-4.5, OpenAI o1, Gemini 2.5, Claude Sonnet 3.7, and DeepSeek R1, have further extended the role of artificial intelligence (AI) in scientific domains by enabling more complex inference and task coordination [11-15].

In bioscience, LLMs have shown promise in clinical decision support, patient-trial matching, and biomedical question answering [16-18]. Concurrent developments like AlphaFold and ESM-2 have led to breakthroughs in protein structure prediction and variant interpretation [19, 20]. Frameworks like BioLunar, ASD-cancer, and CHIEF exemplify the value of integrating multimodal evidence,

including genomic, imaging, and text-based modalities, for enhanced diagnostics in cancer and other diseases [21-23].

While cancer research has rapidly integrated AI tools across diagnostics, treatment planning, and drug discovery, research into Alzheimer's disease (AD) has been slower to adopt these technologies on a comparable scale. AD is a multifactorial neurodegenerative disorder characterized by beta-amyloid plaques, tau protein tangles, immune system dysregulation, and changes in the gut microbiome [24, 25]. Traditional studies often depend on single-modality datasets, which can limit comprehensive insights into the disease's progression.

This study presents the Alzheimer's Disease Analysis Model Generation 1 (ADAM-1), a multi-agent reasoning framework leveraging large language models (LLMs) and retrieval-augmented generation (RAG) [26]. The current version, Generation 1 (ADAM-1, hereafter referred to as ADAM), integrates clinical and microbiome data, modular agents with Chain-of-Thought (CoT) [27] reasoning to enhance interpretability and contextual relevance in disease classification. Trained and tested on a laboratory dataset of 335 multimodal data samples from older adults, ADAM achieved a significantly higher mean F1 score and a much lower prediction variance than the XGBoost baseline model. These findings highlight the methodological benefits of agentic reasoning systems in maintaining performance under data-limited situations.

II. Multimodal Dataset Description and Visualization

This study utilizes the nursing home clinical and metagenomic dataset originally compiled by Haran et al. [28], which focused on the dysregulation of the anti-inflammatory P-glycoprotein pathway in AD. The original study received approval from the Institutional Review Board (IRB) of the University of Massachusetts Medical School (Docket H00010892), and informed consent was obtained from all participants. For the present analysis, only de-identified data were used. We repurposed this dataset to develop ADAM, an LLM-based classifier and reporting system. The dataset comprises clinical and microbiome data from five nursing home sites in central Massachusetts, focusing on individuals with and without AD. The analysis included 335 stool samples collected from 102 unique participants, with some individuals providing multiple samples over a period of up to five months. This dataset includes older adults with Alzheimer's disease and healthy controls, but does not include other types of dementia or individuals with mild cognitive impairment.

A. Clinical Data

Within the clinical dataset, each sample was annotated with detailed clinical metadata, including AD status, comorbidities, demographic information, and longitudinal sampling data. Of the total sample, 32.84% came from individuals diagnosed with AD, while the remaining 67.16% came from individuals without dementia. The cohort is predominantly female (85.7%), with a mean age of 84.5 years and a median of 86.0 years, spanning an age range from 52 to 102 years. This age distribution reflects the demographic characteristics typical of long-term care populations (FIGURE 1).

In addition to cross-sectional clinical features, the dataset included a longitudinal sampling component. Participants contributed between one and 12 stool samples, with a median of three samples per participant, enabling within-subject comparisons over time. Sampling timelines varied by individual, with consistent representation across the AD and non-AD groups. This temporal resolution supports dynamic analyses of microbiome profiles and disease status. The integrated design of the dataset, linking clinical diagnoses, demographic characteristics, comorbid conditions, and longitudinal biospecimens, provides a robust framework for exploring the complex interactions between neurodegenerative diseases, aging, and host-associated microbial communities. These clinical features were integrated into the reasoning and analytical processes of the ADAM framework.

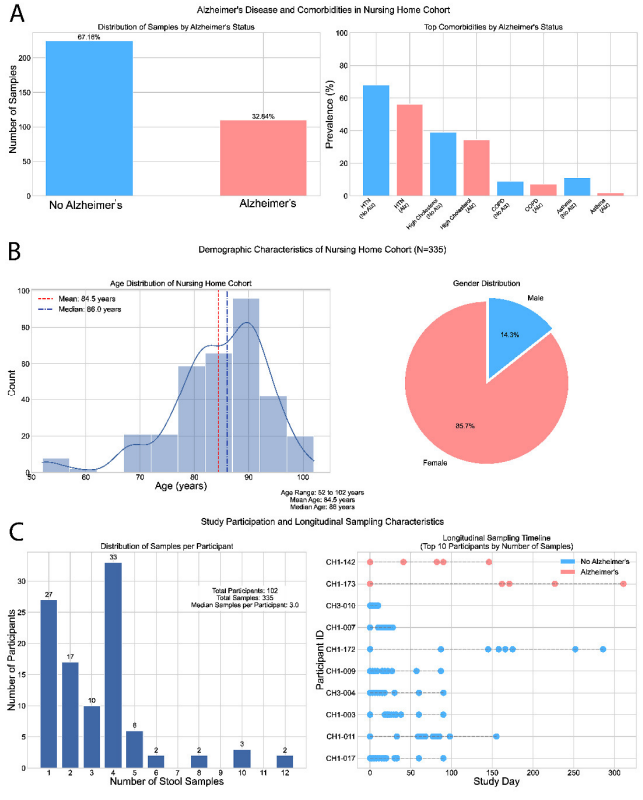

FIGURE 1. The Nursing Home Cohort Clinical Data Description (source: [28]). (A) Bar plots illustrate the distribution of participants based on Alzheimer's status and the prevalence of key comorbidities. 67.16% of participants do not have Alzheimer's disease, while 32.84% do. The most prevalent comorbidities include hypertension (HTN), and various cardiovascular diseases (CVD), with prevalence categorized by Alzheimer's status. (B) The demographic characteristics of the cohort (N=335) encompass age distribution, with a mean age of 84.5 years and a median age of 86.0 years, alongside a pie chart depicting gender distribution, which shows a predominance of females (85.7%) in the cohort. (C) Study participation metrics include the distribution of stool

samples per participant and a longitudinal sampling timeline for the top 10 participants based on the number of samples collected. The median number of samples per participant is 3. Alzheimer's status is color-coded across all visualizations

B. Gut Microbiome Data

The original microbiome dataset was generated through a standardized process, beginning with collecting stool samples from nursing home residents, followed by DNA extraction using a Qiagen DNeasy PowerSoil Pro Kit. The resulting DNA was used to create a pool containing 2 nM DNA, 12 µL RSB with Tween, and 4 µL of diluted PhiX prep, which was then pipetted into a P4 Illumina flow cell cartridge. The sequencing run was generated on the BaseSpace Illumina platform and validated using Illumina NextSeq 2000 prior to analysis. Post-sequencing, the dataset underwent prevalence and abundance-based filtering to reduce technical noise and enhance biological interpretability. Bacterial species were retained if they appeared in more than 5% of samples (i.e., prevalence >17 samples) and exhibited a relative abundance greater than 0.01% of the total community (i.e., proportion >1e-4). This procedure reduced the dataset from 940 to 247 species, focusing subsequent analyses on taxa consistently represented across the cohort.

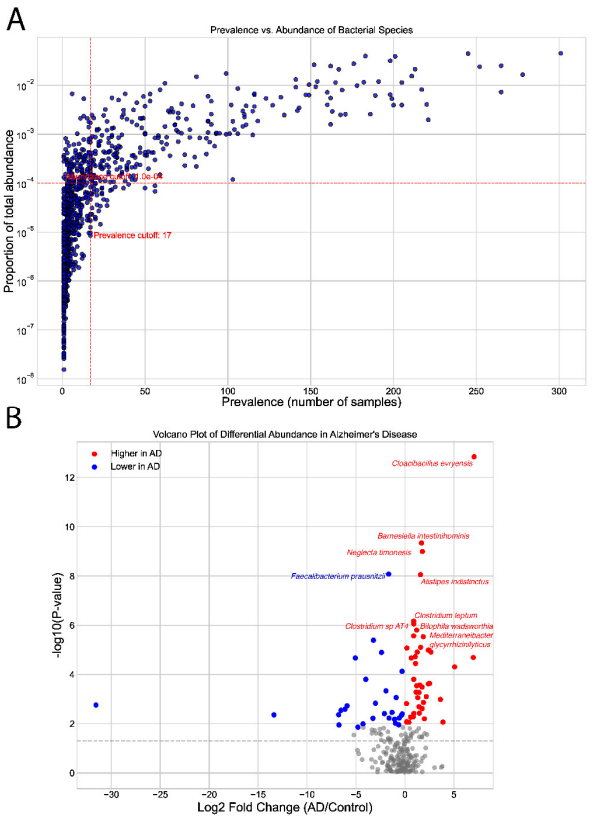

FIGURE 2. Nursing Home Cohort Bacteria Data Description (source: [28]). (A) Prevalence vs. Abundance of Bacterial Species: Scatter plot illustrating bacterial species by their prevalence (the number of samples in which each species appears) and proportional abundance across the dataset. Red dashed lines represent the filtering criteria applied during preprocessing: species present in fewer than 5% of samples (prevalence cutoff: 17) or with a total relative abundance below 1e-4 were removed to minimize noise and focus on biologically relevant taxa. (B) Volcano Plot of Differential Abundance: Log$_2$ fold change versus -log$_{10}$(p-value) for all bacterial species. Species with significantly higher abundance in Alzheimer's patients are shown in red; those lower in AD are shown in blue.

We characterized the dataset by analyzing species-level prevalence, abundance, and differential abundance metrics, as shown in FIGURE 2. The prevalence–abundance distribution (FIGURE 2(A)) exhibits a typical long-tailed pattern, where most bacterial species show low prevalence and low proportional abundance. These rare taxa often add to noise and variance, reducing statistical power. To address this issue, we implemented filtering thresholds, retaining only species present in at least 5% of samples (prevalence ≥17) and with a total abundance greater than 1e-4. This approach ensures that our analysis focuses on reproducible and biologically relevant species.

After preprocessing, we analyzed differential abundance to pinpoint taxa significantly linked to AD status. The volcano plot in FIGURE 2(B) displays log$_2$ fold changes versus –log$_{10}$(p-values), indicating species that are higher in AD (red) and those that are lower (blue). This thorough profiling approach improves the signal-to-noise ratio, ensuring that only reliable microbial features contribute to the ADAM framework. Other data types, like neuroimaging and peripheral biomarkers, were not included in this study.

III. Architecture of ADAM

The ADAM framework comprises an agentic system, a semantic search engine, two base LLMs, and a reporting module, along with new and existing laboratory data. FIGURE 3 illustrates the architecture of the ADAM framework. The ADAM framework begins with users inputting new data into the agentic system. The agentic system processes the new data in its computational agent and sends results to the summarization and classification agents, along with historical laboratory data. This information is then processed through a semantic research engine and the base LLMs, with cosine similarity controlling the return quality. Finally, a reporting system presents the output as a classification report.

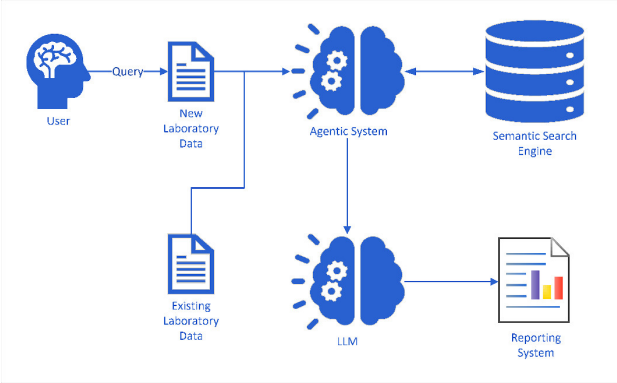

FIGURE 3. Workflow of the ADAM Framework. This diagram illustrates user interactions within the ADAM system. When a user submits a query along with new laboratory data, the framework employs the agentic system to refine the query and identify relevant literature evidence from

a semantic search engine, integrating it with existing laboratory data. The semantic search engine collaborates with the agentic system, alongside the LLM, to reason, analyze, and interpret data, ultimately delivering precise analytical reports.

A. Agentic System

The agentic system within the ADAM framework consists of three distinct yet interconnected agents: a computational agent, a summarization agent, and a classification agent. This multi-agent design produces both computational and peer-reviewed–informed semantic outputs, thereby improving analytical strength while also reducing hallucinations and the likelihood of misclassification inherent in the base LLMs. These agents transform clinical and metagenomic data into interpretable patient-specific insights. As illustrated in FIGURE 4, the computational agent processes the newly uploaded data to extract key features and analytical outputs. These outputs, along with historical laboratory data, are then used for summarization with a proposed context window size of approximately 100,000 tokens and classification with a proposed context window size of approximately 50,000 tokens, guided by CoT reasoning and supported by LLMs. The summarization agent generated context-aware narratives, whereas the classification agent determined the AD status.

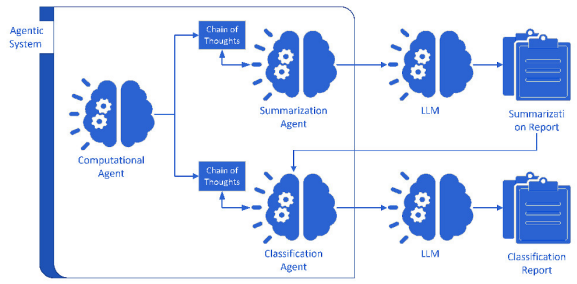

FIGURE 4. Agentic System Workflow. The Diagram illustrates the coordinated interaction among three AI agents—Computational, Summarization, and Classification—within the agentic system. Each agent leverages Chain-of-Thought reasoning and communicates with the LLM to generate structured outputs. The system collectively produces Alzheimer's-specific summarization and classification reports by integrating computational analysis with contextual interpretation.

Base LLM Selection Rationale: GPT-4o and its light variant GPT-4o-mini were selected as the base LLM for ADAM-1 due to their consistently superior performance across a wide range of biomedical and clinical benchmarks. In cancer genomics, GPT-4o achieved the highest accuracy (0.7318) when classifying clinically actionable variants from multiple knowledge bases, outperforming Qwen 2.5 and Llama 3.1, and showing increased concordance with expert annotations even in complex evidentiary contexts [29]. In medical education, GPT-4o demonstrated top-tier performance, achieving an accuracy of 89.2% on the Japanese National Medical Examination (JNME), significantly exceeding Claude 3 Opus, Gemini 1.5, and earlier GPT versions [30]. It was notably reliable for easy-to-moderate tasks, reflecting its capacity for high factual retention and instructional consistency. Additionally, GPT-4o achieved the highest accuracy across multiple medical licensing examinations, including the United States Medical Licensing Examination (USMLE), the Professional and Linguistic Assessments Board (PLAB), the Hong Kong Medical Licensing Examination (HKMLE), and the National Medical Licensing Examination (NMLE), demonstrating its robustness across diverse cultural and linguistic contexts in medicine [31]. These empirical advantages directly support its integration into ADAM-1, where high-stakes reasoning on clinical and microbiome data requires both precision and stability for the summarization and classification agents in biological studies.

Computational Agent: The computational agent comprises bioinformatics and machine learning. Alpha diversity (Shannon, Simpson, Berger-Parker dominance) and beta diversity (Bray-Curtis, Jaccard, Canberra) were calculated to provide a descriptive analysis of bacterial relationships. XGBoost [32] was chosen as the primary machine learning model, working in conjunction with SHAP [33] to generate feature importance information and SHAP values that describe feature interactions in relation to AD status from a global and an individual study subject perspective. The formula for the computational agent is as follows:

$$\mathrm{CA}_{\mathrm{comp}}(\mathrm{MD}, \mathrm{CF}) = \mathrm{SHAP}\big(\mathrm{XGB}(\mathrm{MD}, \mathrm{CF})\big), \mathrm{D}_{\alpha}(\mathrm{MD}), \mathrm{D}_{\beta}(\mathrm{MD}) \quad (1)$$

Where:

$\mathrm{CA}_{\mathrm{comp}}$ is the Computational Agent responsible for processing microbiome and clinical data to generate interpretable outputs.

MD is the Microbial Data input (e.g., microbial abundance profiles).

CF is a clinical feature input (e.g., patient demographics and health records).

SHAP(XGB(MD, CF)) represents the XGBoost model embedded within SHAP, which enables interpretable machine learning via feature attributes and interactions.

D_a(MD) denotes the Alpha diversity metrics (Shannon, Simpson, Berger-Parker dominance) computed from microbial data.

D_β(MD) denotes the Beta diversity metrics (Bray-Curtis, Jaccard, Canberra) representing microbial community dissimilarity.

Summarization Agent: The summarization agent integrates the computational agent's results, CoT reasoning, semantic search engine, and GPT-4o model to provide an overall context summary. The model demonstrates superior language understanding and coherence in long text **[34]**. Its ability to maintain contextual relevance and accuracy in summaries is well documented, particularly in academic and technical domains. This makes it an ideal choice for summarization tasks that require precision and reliability. The formula for the summarization agent is as follows:

$$\mathrm{SA}_{\mathrm{summary}}\left(\mathrm{CA}_{\mathrm{comp}},\mathrm{CoT}_{\mathrm{reasoning}},\mathrm{SS}_{\mathrm{search}},\mathrm{LLM}_{\mathrm{GPT\text{-}4o}}\right) = \mathrm{LLM}_{\mathrm{GPT\text{-}4o}}\left(\mathrm{Integrate}\left(\mathrm{CA}_{\mathrm{comp}},\mathrm{CoT}_{\mathrm{reasoning}},\mathrm{SS}_{\mathrm{search}}\right)\right) \quad (2)$$

Where:

- $\mathrm{SA}_{\mathrm{summary}}$ (*Summarization Agent*): Synthesizes context-aware insights
- $\mathrm{CA}_{\mathrm{comp}}$ (*Computational Agent*): Processes microbial and clinical data to generate structured outputs
- $\mathrm{CoT}_{\mathrm{reasoning}}$ (*Chain-of-Thought reasoning*): Provides structured logical interpretive paths
- $\mathrm{SS}_{\mathrm{search}}$ (*Semantic Search*): Retrieves relevant contextual information using RAG from a vector database.
- $\mathrm{LLM}_{\mathrm{GPT\text{-}4o}}$ (*Large Language Model*): GPT-4o model that generates accurate, coherent summaries
- Integrate(·): Combines multiple sources into a unified semantic context

We propose a logic for CoT reasoning for the summarization agent in the following eight steps to ensure a comprehensive understanding of the input:

1. Patient Overview: Starts with basic patient demographics and general health status
2. Key Clinical Markers: Identifies important biomarkers or health indicators
3. Gut Microbiome Profile: Analyzes the specific microorganisms present in the gut
4. Diversity Metrics Analysis: Evaluates biodiversity measures of the gut microbiome
5. Interactions and Mechanisms: Examines relationships between the microbiome and clinical markers
6. Descriptive Correlation: Identifies statistical relationships without implying causation
7. Machine Learning analysis and probabilistic assessment: ML models were used to evaluate the patterns and make probabilistic predictions
8. Final Comprehensive Descriptive Summary: Creates a holistic summary integrating all findings

Classification Agent: The classification agent integrates the output of the computational agent, the output of the summarization agent, CoT reasoning, our semantic search engine, and the GPT-4o-mini model for content classification. We used GPT-4o-mini for AD classification tasks because it efficiently processes extremely long texts, delivers high accuracy and reliability essential for medical diagnosis, and offers unmatched cost-effectiveness compared with similar models [35]. The formula for a classification agent is as follows:

$$\mathrm{CA}_{\mathrm{class}}\left(\mathrm{CA}_{\mathrm{comp}},\mathrm{SA}_{\mathrm{summary}},\mathrm{CoT}_{\mathrm{reasoning}},\mathrm{SS}_{\mathrm{search}},\mathrm{LLM}_{\mathrm{GPT\text{-}}}\right) = \mathrm{LLM}_{\mathrm{GPT\text{-}4o\text{-}mini}}\left(\mathrm{Integrate}\left(\mathrm{CA}_{\mathrm{comp}},\mathrm{SA},\mathrm{CoT}_{\mathrm{reasoning}},\mathrm{S}\right.\right. \quad (3)$$

Where:

- $\mathrm{CA}_{\mathrm{class}}$ (*Classification Agent*): Responsible for generating Alzheimer's disease classification outputs
- $\mathrm{CA}_{\mathrm{comp}}$ (*Computational Agent*) contains structured data features (e.g., predictions and feature importance)
- $\mathrm{SA}_{\mathrm{summary}}$ (*Summarization Agent*): Provides high-level contextual insight
- $\mathrm{CoT}_{\mathrm{reasoning}}$ (*Chain-of-Thought reasoning*): Guides interpretability through logical step-by-step inference.
- $\mathrm{SS}_{\mathrm{search}}$ (*Semantic Search*): retrieves conceptually relevant knowledge from a vector-based index
- $\mathrm{LLM}_{\mathrm{GPT\text{-}4o\text{-}mini}}$ (*Large Language Model*): Optimizes for fast and efficient classification tasks
- Integrate(·) is a function that merges multiple contextual and data-driven sources into a unified representation

We propose the CoT reasoning logic for the classification agent in the following eight steps to ensure robustness in the classification task.

1. Historical Data Insights: Analyzes past patient data to establish baselines and trends
2. Diversity Metrics & Classification Refinement: Uses microbiome diversity measures to improve classification accuracy
3. Adaptive Threshold Decisioning: Determines dynamic thresholds for classification rather than fixed values
4. Handling Edge Cases & Misclassifications: Identifies and addresses outliers or difficult-to-classify cases
5. Comprehensive Summary of this Visit: Provides a complete analysis of the current patient data
6. SHAP Feature Importance: Uses SHAP values and ML outputs to understand the features that influence the model's predictions most
7. Key Considerations for Prediction and Misclassification Adjustments: Identifying factors that might lead to misclassification and how to adjust for them

8. Prediction Decision Rules: Establishes clear rules for making final classification decisions

B. Base LLM Models

This study utilized two OpenAI LLM models: GPT-4o for summarization and GPT-4o-mini for improved speed and cost efficiency. Their selection aligns with the design concepts of the summarization and classification agents. Using an eight-step CoT reasoning logic, the summarization agent is designed to handle more contextual data, including single-visit and longitudinal data. Each step corresponds to a relevant RAG output, explaining the meaning of each step in summarizing an individual's clinical conditions and microbiome profiles. The summarization task was designed to process approximately 100,000 tokens of text data at a time, given its superior ability to handle clinical text summarization [36]. Therefore, as a downstream counterpart to GPT-4o, GPT-4o-mini was proposed for focused binary classification tasks utilizing a dedicated CoT reasoning framework for classification. It processes data that has been preprocessed by computational and summarization agents powered by GPT-4o and is optimized for high-throughput classification over approximately 50,000 tokens per instance.

C. Semantic Search Engine

The knowledge base comprised 76,751 full-text publications or abstracts programmatically retrieved in October 2024 from PubMed Central (PMC) using the NCBI Entrez system via E-utilities and Entrez Direct. Articles were identified based on user-defined query terms applied to specific fields, such as titles, abstracts, or text words, and retrieved in XML format using the efetch utility. The collected records were subsequently processed using an embedding model and indexed into two vector databases to support downstream retrieval tasks.

The publications were divided into 2,058,502 text chunks, each containing 2,000 characters, with a 20 percent overlap with the previous chunk to ensure continuity and preserve context. This approach minimizes the risk of losing critical information when analyzing the AD-related literature. The embedding model converts these chunks into high-dimensional vector representations, which are stored in two separate vector databases optimized for semantic search and retrieval (FIGURE 5). These databases enable rapid querying of relevant AD research, supporting tasks such as disease classification, summarization, and hypothesis generation within the ADAM framework.

Publications: This component serves as a source of existing literature and the reasoning logic of the ADAM, integrating 76,751 publications relevant to AD research at the time of this study. The keywords used for indexing included Alzheimer's disease, Gut-Brain Axis, Gut Microbiome, and Immunosenescence (TABLE I).

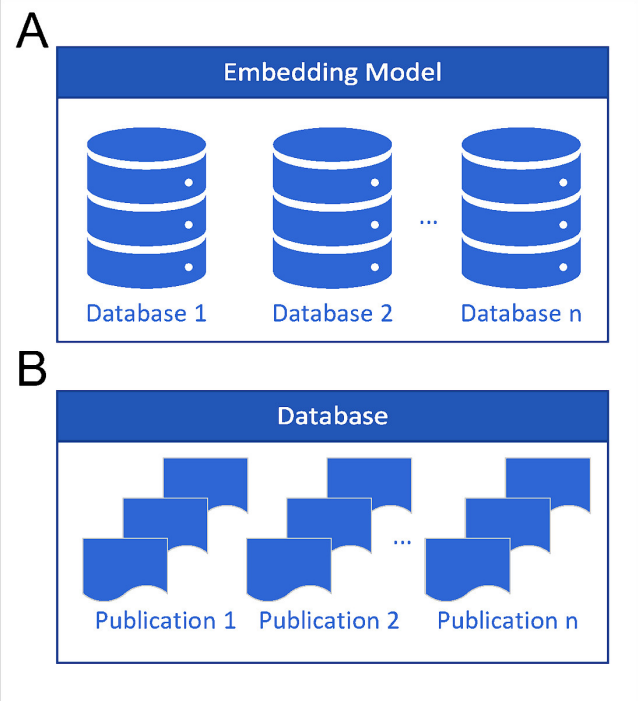

FIGURE 5. Semantic Search Engine Architecture. Semantic Search Engine. (A) The embedding model converts documents into and out of numerical vectors. (B) Vector databases work in parallel with the semantic search engine that returns the most relevant information measured by cosine similarity. Inside a Vector Database. Each publication is split into 2000-character vector chunks, with a 20 percent overlap between the prior chunk of text and the next.

TABLE I
PUBLICATIONS AND TEXT SEGMENT COUNTS BY TOPIC.

Keywords	Publications	Segments
Alzheimer's disease	62,478	1,591,441
Gut Microbiome	11,692	381,630
Immunosenescence	1,273	36,172
Gut-Brain Axis	1,308	49,259
Total	76,751	2,058,502

THE NUMBER OF PUBLICATIONS AND SEGMENTED TEXT UNITS ANALYZED IN THIS STUDY ACROSS KEY BIOMEDICAL TOPICS RELATED TO ALZHEIMER'S DISEASE AND MICROBIOME RESEARCH. THE DATASET INCLUDES OVER 76,000 PUBLICATIONS AND 2 MILLION SEGMENTS, MOST FOCUSED ON ALZHEIMER'S DISEASE AND THE GUT MICROBIOME.

Each publication P_i is indexed by a set of relevant keywords $K_{i,j}$, forming a keyword-embedding vector $K(P_i)$:

$$\boldsymbol{K}(P_i) = \sum_{j=1}^{m} w_{i,j} \cdot \text{EmbeddingModel}\left(K_{i,j}\right) \qquad (4)$$

Where:

P_i represents the i-th publication.

$K_{i,j}$ is the j-th keyword associated with P_i.

$w_{i,j}$ is the weight assigned to $K_{i,j}$ based on its relevance to P_i.

$\boldsymbol{K}(P_i)$ is the aggregated embedding vector representing the keywords of P_i, facilitates its retrieval from the knowledge base.

The vector $\boldsymbol{K}(P_i)$ enables the efficient retrieval of publications based on semantic relevance to query keywords in AD research.

Embedding Model: The embedding model transforms textual data into embeddings, thereby allowing the framework to process and effectively retrieve relevant information. We chose text-embedding-ada-002 because of its high-cost effectiveness and proven success in text processing since its release in December 2022. To determine the most suitable embedding models for this study, we compared three available embedding models from OpenAI: text-embedding-ada-002 (hereafter referred to as ada-002), text-embedding-3-small (hereafter referred to as 3-small), and text-embedding-3-large (hereafter referred to as 3-large) to determine the most suitable embedding models for this study [37]. These three models were evaluated based on the following criteria: semantic richness, computational efficiency, storage requirements, cost-effectiveness, versatility, and adoption (TABLE II).

Furthermore, we applied these three embedding models to a small-scale test vector database derived from gut microbiome publications, where each text chunk contained 2,000 characters and had a 20 percent overlap with the previous chunk, utilizing a cosine similarity threshold of 0.7. Our empirical findings indicated that the ada-002 model yielded more consistent results within the LLM and vector database configurations during the RAG retrieval process.

TABLE II.
EMBEDDING MODEL COMPARISON. SUMMARY OF KEY PERFORMANCE AND USABILITY ATTRIBUTES OF THE ADA-002 EMBEDDING MODEL, HIGHLIGHTING ITS STRONG SEMANTIC REPRESENTATION, HIGH COMPUTATIONAL EFFICIENCY, MODERATE STORAGE DEMANDS, AND BROAD ADOPTION ACROSS APPLICATIONS.

Criterion	ada-002 (1,536 dims)	3-small (1,536 dims)	3-large (3,072 dims)
Semantic Richness	High	High	Very High
Computational Efficiency	High	Highest	Moderate to Low
Storage Requirements	Moderate	Moderate	High
Cost-effectiveness	High	Moderate	Lower
Versatility & Adoption	Proven, widely used	Newer, less proven	Newer, powerful, but resource-intensive

Given an input text $\boldsymbol{T}$, the embedding model generates an embedding vector $\vec{\boldsymbol{E}}$ as follows:

$$\vec{\boldsymbol{E}} = \text{EmbeddingModel}(\boldsymbol{T}) \quad (5)$$

Where:

$\boldsymbol{T}$ represents the i-th publication.

$\vec{\boldsymbol{E}}$ is the resulting embedding vector in high-dimensional space.

This embedding vector $\vec{\boldsymbol{E}}$ captures semantic and contextual information from $\boldsymbol{T}$, facilitating effective retrieval and matching within the framework.

Vector Databases: When embedding publications into a vector database, the publication text is divided into 2000-character segments, with a 20 percent overlap between segments, to maintain the information flow and provide a more continuous and coherent representation of the text.

Given:

Segment length $\boldsymbol{s} = \mathbf{2000}$ characters

Overlap $\boldsymbol{o} = \mathbf{0.2} \times \boldsymbol{s} = \mathbf{400}$ characters

Effective step size = $\boldsymbol{s} - \boldsymbol{o} = \mathbf{1600}$ characters

$$\boldsymbol{p_i} = \mathbf{1} + (\boldsymbol{i} - \mathbf{1}) \times (\boldsymbol{s} - \boldsymbol{o}) \quad (6)$$

This formula calculates the starting position of the $\boldsymbol{i}$-th segment:

Segment 1 starts at position 1

Segment 2 starts at position 1601

Segment 3 starts at position 3201, and so on...

Total number of segments $\boldsymbol{n}$ required to cover the entire text $\boldsymbol{L}$, each new segment after the first adds $(\boldsymbol{s} - \boldsymbol{o}) = \mathbf{1600}$ new characters. Subtraction of $\boldsymbol{o}$ from the numerator accounts for its overlap with the last segment.

$$\boldsymbol{n} = \left\lceil \frac{\boldsymbol{L} - \boldsymbol{o}}{\boldsymbol{s} - \boldsymbol{o}} \right\rceil \quad (7)$$

For example:

If $\boldsymbol{L} = \mathbf{5800}$ characters

First segment covers positions 1-2000

Second segment covers positions 1601-3600

Third segment covers positions 3201-5200

Fourth segment covers the position of 5211-7200

Using the original formula $\boldsymbol{n} = \lceil \frac{\mathbf{5800-400}}{\mathbf{2000-400}} \rceil = \lceil \frac{\mathbf{5400}}{\mathbf{1600}} \rceil = \lceil \mathbf{3.375} \rceil = \mathbf{4}$, as a result, 5800 characters need 4 segments.

This configuration incorporated a 20 percent overlap between consecutive segments to maintain continuity in the embeddings. Consequently, the vector databases in the semantic search engine comprised 2,058,502 discrete overlapping segments derived from 76,751 publications (TABLE I).

IV. Hardware, Software, and Computational Cost

The experiments were conducted on an Ubuntu 24.04.2 LTS workstation equipped with an Intel® Core™ i9-10900X (20 threads), 128 GB RAM, and four NVIDIA GeForce RTX™ 3090 GPUs, providing a robust computing environment for LLM and machine learning tasks. The software stack was built on Python 3.10.14, utilizing XGBoost 2.1.3, and Scikit-Learn 1.5.2 for model training, with Optuna 4.1.0 handling hyperparameter optimization.

For LLM processing, the system integrates OpenAI 1.55.1, PandasAI 2.4.2, LangChain 0.3.8, and LangChain-Chroma 0.1.4. Additionally, Scikit-Bio 0.6.2, SciPy 1.10.1, NumPy 1.26.4, and Pandas 1.5.3 facilitated data processing and analysis. For visualization and interpretability, the setup included Matplotlib 3.7.5, Seaborn 0.12.2, and SHAP 0.46.0.

This configuration ensures high computational efficiency and scalability for deep learning workflows.

Within the ADAM framework, the summarization agent typically requires 2 to 5 minutes per observation, while the classification agent takes approximately 5 to 10 minutes, depending on the availability of the OpenAI API. XGBoost, optimized using Optuna, consumes approximately 310 to 350 MB of VRAM on an NVIDIA 3090 GPU and completes training and testing in 2 to 3 minutes. For cost transparency, we also report token pricing: GPT-4o currently costs $2.50 per million input tokens and $10.00 per million output tokens, while GPT-4o-mini costs $0.15 per million input tokens and $0.60 per million output tokens [38].

VI. Data Split and Seeding Policy

A. Data Split Policy

We implemented two data-splitting policies: one for selecting the baseline model and the other for the ADAM framework. In the baseline model selection phase, the data were initially split at 75:25 by a unique Study ID and stratified according to the AD status. In the second phase, which involved the ADAM framework, 75% of the data was retained as the training or reference dataset. For testing, we randomly selected 15 positive and 15 negative cases from the 25% portion owing to hardware limitations. This approach allowed us to work efficiently with the LLM while maintaining statistical significance (n = 30).

B. Seeding and Measures

We implemented two seeding strategies to support the different phases of the study: one for baseline model selection and the other for creating the ADAM framework. The initial phase involved applying 10 random seeds to identify the best-performing classifier based on the accuracy, AUC, and F1 score. The selected model was then integrated into ADAM, a language-model-based system designed to enhance the binary classification of AD. A broader evaluation was conducted using 30 random seeds to examine whether ADAM improved the mean F1 score and reduced the performance variance compared with the baseline. This seeding policy design ensured that the performance metrics in both phases were derived from stable and reproducible evaluations, allowing for a fair assessment of ADAM's added value over the baseline model.

VII. Baseline Model Selection

We trained three classifiers on the AD dataset: XGBoost, random forest, and logistic regression. These models are commonly used in biological studies and are computationally efficient for laboratory computers. The best-performing model was selected as the baseline for this study. The model selection process involved two phases: feature selection and model training, both optimized using Optuna, which is a Bayesian-based hyperparameter-tuning framework [39]. To identify the most relevant features, we configured an XGBoost-based feature selector applied to all three machine learning models. The selected features are subsequently fed into the proposed models for further training and testing.

We implemented several performance metrics to evaluate the model's effectiveness, including accuracy, AUC, F1 score, and overall performance index. Each metric was calculated by averaging the results from 10 different random seeds, which helped capture the variability and enhance the robustness of the evaluation process. The results are shown in FIGURE 6 and TABLE III, which summarizes the comparative performance of each model across all metrics. By relying on multiple evaluation criteria, we aimed to assess not only how well each model classifies, but also how reliably and consistently it performs across different data splits, which is especially critical in clinical datasets where precision and balance are essential.

As shown in TABLE III, XGBoost consistently outperformed logistic regression and Random Forest across all key metrics. It achieved the highest average accuracy, AUC, and F1 score, demonstrating its superior discriminative ability and robustness. Furthermore, XGBoost exhibited the highest stability across different seeds, underscoring its reliability in managing complex and heterogeneous data related to AD. Owing to its strong performance and consistency, XGBoost was selected as the baseline model in this study. This served as the foundation against which the proposed ADAM framework was assessed, providing a high-performance and dependable benchmark for future model comparisons.

TABLE III.
BASE MODEL PERFORMANCE AVERAGED ACROSS 10 RANDOM SEEDS.

Model	Accuracy	AUC	F1	Mean
XGBoost	**0.769 ± 0.069**	**0.821 ± 0.061**	**0.651 ± 0.1**	**0.769 ± 0.069**
Random Forest	0.742 ± 0.066	0.804 ± 0.087	0.603 ± 0.087	0.742 ± 0.066
Logistic Regression	0.735 ± 0.067	0.772 ± 0.1	0.626 ± 0.095	0.735 ± 0.067

THE CLASSIFICATION PERFORMANCE OF XGBOOST, RANDOM FOREST, AND LOGISTIC REGRESSION MODELS WAS EVALUATED USING MEAN ACCURACY, AUC, AND F1 SCORES WITH STANDARD DEVIATION. RESULTS REFLECT MODEL STABILITY AND PREDICTIVE POWER ACROSS REPEATED RUNS, WITH XGBOOST ACHIEVING THE HIGHEST AVERAGE PERFORMANCE ACROSS ALL METRICS.

VIII. Results and Evaluation

A. Evaluation Strategy

The evaluation strategy aimed to determine whether ADAM outperformed the baseline XGBoost model in binary classification tasks using the F1 score. The F1 score is directly affected by false negatives (FN) because it is the harmonic mean of precision and recall, with recall particularly sensitive to false negatives [40]. This is especially critical in medical applications, where false negatives can lead to severe consequences [41, 42]. In the context of AD, false negatives can result in delayed diagnosis, leading to inadequate

monitoring and treatment, loss of social and financial benefits, increased emotional distress for patients and caregivers, and ultimately, worse patient outcomes [43-45]. The mean F1 score quantifies the overall predictive performance, whereas the variance of the F1 score assesses the consistency and stability of the ADAM performance compared to XGBoost.

Each model was fine-tuned using Optuna, with extensive cross-validation. For each of the 30 seeds, we conducted 200 Optuna trials using Bayesian optimization (TPE sampling), with each trial evaluated via stratified 3-fold cross-validation, resulting in 600 evaluations per seed. The best hyperparameters identified for each seed were used exclusively for that seed's final model, and the corresponding F1 scores were recorded. The statistical significance of the difference in mean F1 scores between ADAM and XGBoost was analyzed using the Mann-Whitney U test, a non-parametric test that does not require normality assumptions [46]. Additionally, an F-test was conducted to compare the variances in the F1 scores, providing insights into the performance stability of each model [47].

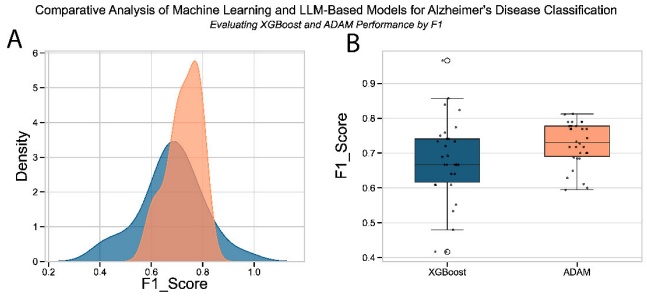

FIGURE 6. Comparative Analysis of XGBoost and ADAM. (A) Density plots and (B) boxplots of F1 scores comparing XGBoost and the ADAM framework across multiple runs. ADAM demonstrates a higher median F1 score with reduced variance, suggesting more consistent and reliable classification performance for AD prediction.

B. F1 Score and Variability Analysis

Following the evaluation strategy, we conducted a comparative analysis of ADAM and XGBoost based on 30 independent experimental runs. The focus was on two core aspects of model performance: accuracy, measured by the mean F1 score, and stability, indicated by the variability of F1 scores across runs. By assessing central tendency and dispersion, we determined which model performed better on average and which produced more consistent results.

Across these runs, ADAM achieved a higher mean F1 score (0.7263) than XGBoost (0.6774), as determined by the Mann-Whitney U test ($p = 0.0418$), indicating a statistically significant improvement at the 95% confidence level. In addition, ADAM demonstrated a lower standard deviation (0.0632) than XGBoost (0.1217), as supported by Levene's test ($p = 0.0300$), indicating a more stable and consistent performance. The medium effect size (Cohen's $d = 0.5038$) further underscores the practical relevance of ADAM's superior and reliable classification capability.

The box plots and density distributions (FIGURE 6) demonstrate that ADAM maintains a narrower F1 score distribution (FIGURE 6(B)) with a higher median and fewer extreme values than XGBoost. Although both models exhibit similar central performance around 0.7, the density plot (FIGURE 6(A)) reveals that ADAM's F1 scores cluster more tightly around 0.75, whereas XGBoost displays a broader spread, ranging from 0.4 to nearly 1.0, indicating greater performance variability across different runs. This higher consistency in the ADAM performance suggests that it may offer more reliable predictions for AD classification (FIGURE 7).

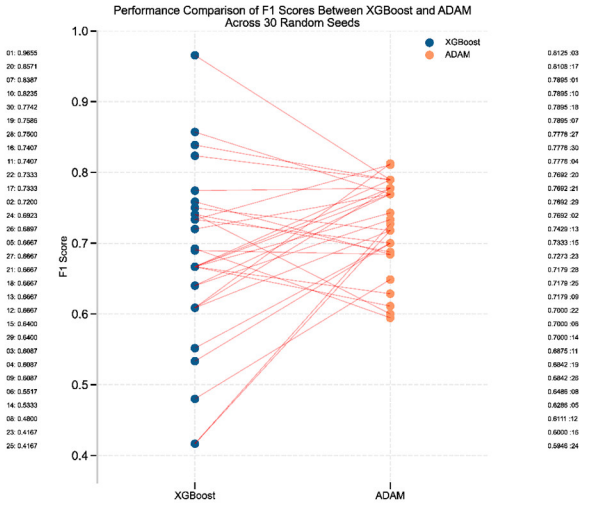

FIGURE 7. Performance Comparison of F1 Scores Between XGBoost and ADAM. Line plot comparing F1 scores of XGBoost and ADAM across 30 random seeds. Each line connects paired runs for a given seed. ADAM consistently shows improved or more stable F1 performance relative to XGBoost, highlighting its robustness and reduced variability across repeated evaluations.

These findings collectively demonstrate that ADAM significantly outperforms XGBoost, offering an improved average performance and greater consistency and stability across experimental runs.

IX. Reporting Module

ADAM is not only a classifier but also a reporting system. Its classification is based on the combined results of the three agentic systems presented in textual form. Its logical reasoning design classifies the AD status while generating analytical reports for each study subject. We listed two sample classification reports as follows:

A. Sample Reports

TABLE IV presents two sample reports produced by ADAM, including one positive and one negative case. Owing to the length of the entire report, only the conclusion section is listed in each of the following reports. The full set of ADAM-generated reports, totaling over 900 case-specific classifications, is available at https://github.com/melhzy/ADAM/reporting. Two

representative sample reports, positive and negative, were presented in a condensed format.

TABLE IV.
ADAM SAMPLE REPORTS: POSITIVE CLASSIFICATION AND NEGATIVE CLASSIFICATION. INTERPRETABILITY-FOCUSED SUMMARIES PRODUCED BY ADAM FOR ONE POSITIVELY AND ONE NEGATIVELY CLASSIFIED SAMPLE. EACH REPORT INCLUDES CLINICAL VARIABLES, MEDICATION USAGE, GUT MICROBIOME COMPOSITION, DIVERSITY METRICS, AND SHAP-BASED FEATURE ATTRIBUTIONS TO CLARIFY THE PREDICTED ALZHEIMER'S DISEASE STATUS.

Sample of a Positive Classification - Conclusion	Sample of a Negative Classification - Conclusion
Prediction: **Yes** - The prediction of Alzheimer's disease (AD) for Sample ID FB151 is strongly supported by multiple factors:	**Prediction**: **No** - The prediction for Sample ID FB128 is 'No' for Alzheimer's disease (AD) based on a probability of 24.20%. This low probability is influenced by several key factors:
1. **Clinical Indicators**: The patient is a 94-year-old female with a Clinical Frailty Scale score of 7, indicating severe frailty. This level of frailty is significantly associated with cognitive decline and AD progression. Additionally, the patient has a malnutrition score of 2, which places her at risk for nutritional deficiencies that can exacerbate cognitive decline.	1. **Demographics and Clinical Background**: - The patient is 81 years old, male, with a moderate to severe frailty score of 6. This level of frailty is concerning but does not alone indicate a high probability of AD. – The malnutrition score of 2 indicates a risk of malnutrition, which can exacerbate cognitive decline but is not definitive for AD.
2. **Polypharmacy**: The patient is on multiple medications (polypharmacy), which can alter gut microbiota composition and increase systemic inflammation, further contributing to cognitive health deterioration.	2. **Clinical Features**: - The presence of seizure medications and oral corticosteroids suggests potential neurological comorbidities, which could elevate AD risk. However, the absence of cardiovascular risk factors (e.g., hypertension, high cholesterol) is protective against vascular contributions to cognitive decline.
3. **Gut Microbiome Profile**: The microbiome analysis reveals an imbalance with elevated levels of pro-inflammatory species such as *Phocaeicola vulgatus* and *Clostridia bacterium*, while beneficial species like *Faecalibacterium prausnitzii* are absent. This dysbiosis is linked to increased inflammation and cognitive decline, which are critical factors in AD pathology.	3. **Gut Microbiome Profile**: - The microbiome analysis shows a moderate abundance of protective species like *Faecalibacterium prausnitzii* (1.28183), which has anti-inflammatory properties. This is counterbalanced by the presence of *Neglecta timonensis* (0.74327), which is positively associated with AD risk. – The overall alpha diversity metrics (Shannon Index: 3.50, Simpson Index: 0.93) indicate a relatively balanced microbiome, which is generally protective against systemic inflammation and cognitive decline.
4. **Diversity Metrics:** The Shannon Index of 2.98 indicates moderate diversity, but the presence of dominant pro-inflammatory species and high Bray-Curtis dissimilarity suggests significant deviations from a healthy microbiome. This imbalance can exacerbate systemic inflammation and cognitive decline.	4. **SHAP Feature Importance**: - The top SHAP features include *Neglecta timonensis* (SHAP: +0.7978), Seizure Medications (SHAP: +0.7696), and *Faecalibacterium prausnitzii* (SHAP: -0.6193). The positive contributions from Neglecta and seizure medications suggest increased risk, while Faecalibacterium's negative contribution indicates a protective effect.
5. **SHAP Feature Importance**: The SHAP analysis highlights that the malnutrition score, the presence of *Neglecta timonensis*, and the low levels of *Faecalibacterium prausnitzii* are the top contributors to the high probability of AD. These features align with known risk factors for AD.	5. **Diversity Metrics**: - The Bray-Curtis dissimilarity indicates high dissimilarity from healthy controls, suggesting a distinct microbial composition that may reflect gut dysbiosis linked to AD.

Analyses of samples FB151 and FB128 demonstrated that the gut microbiome composition significantly influenced AD classification. Sample FB151, from a 94-year-old female with severe frailty (CFS 7), exhibited clear microbiome dysbiosis, characterized by elevated pro-inflammatory species, absence of the beneficial *Faecalibacterium prausnitzii*, and lower diversity (Shannon index: 2.98), which led to a positive AD classification. In contrast, sample FB128, an 81-year-old male with moderate-to-severe frailty (score 6), maintained a more balanced microbiome with protective *Faecalibacterium prausnitzii* (1.28183) that counterbalanced the presence of AD-associated *Neglecta timonensis*, resulting in higher diversity metrics (Shannon, 3.50; Simpson, 0.93) and a negative classification despite a similar malnutrition risk (score 2). SHAP analyses for both patients highlighted the critical importance of *Faecalibacterium prausnitzii* as a protective factor and *Neglecta timonensis* as a risk-inducing factor, demonstrating how microbiome balance can mitigate clinical risk factors and potentially protect against AD development, even in advanced age and frailty.

Figures 8 and 9 show SHAP waterfall plots for the same two samples. These plots display how individual features contributed to the final model prediction and visually support the textual reports in Table IV. In Figure 8, the model gave a high probability of Alzheimer's disease for FB151, mainly driven by the malnutrition score, presence of *Neglecta timonensis*, and absence of *Faecalibacterium prausnitzii*. These factors increased the prediction toward AD, aligning with the patient's clinical frailty and gut dysbiosis. In Figure 9, the SHAP values for FB128 demonstrate how protective microbial features, such as *Faecalibacterium prausnitzii*, counteract the effects of seizure medications and *Neglecta timonensis*, reducing the predicted risk and supporting a non-AD classification. The inclusion of these visual explanations clarifies the alignment between computational attributions (SHAP+XGBoost) and the semantic narrative generated by the LLM agent. This combined approach strengthens interpretability and clinical relevance, satisfying both statistical and explanatory transparency.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3599857

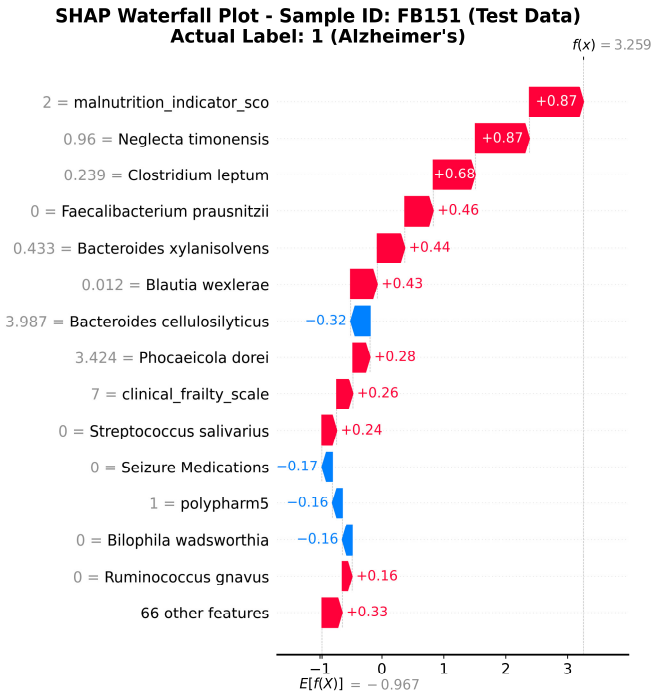

Figure 8. SHAP waterfall plot for Sample FB151 (AD-positive). Malnutrition and *Neglecta timonensis* contributed strongly to the prediction, while the absence of protective *Faecalibacterium prausnitzii* further increased AD risk.

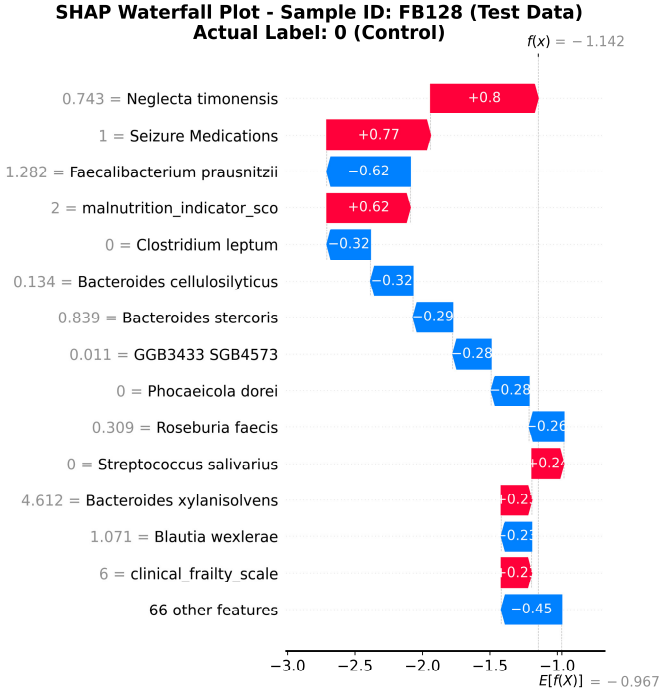

Figure 9. SHAP waterfall plot for Sample FB128 (AD-negative). The presence of *Faecalibacterium prausnitzii* lowered the prediction probability, counteracting other risk factors such as *Neglecta timonensis* and seizure medications.

X. Discussion and Limitations

The ADAM framework demonstrates the potential of leveraging LLMs via API calls by integrating bioinformatics, machine learning, explainable AI, and relevant literature with biological laboratory data analytics. RAG is facilitated through a web service hosting vector databases. ADAM is tailored to experimental settings common in Alzheimer's disease research, where around 100 to 300 data points are generated per experiment due to the high cost. We have also successfully run ADAM on low-resource lab computers, such as a Dell Inspiron (AMD Ryzen 5, 8 GB RAM) and an ASUS Vivobook S 14 (Intel Core Ultra 5, 16 GB RAM), where machine learning hardware requirements and GPT API availability bottleneck performance. Large-scale dataset experiments are beyond the scope of the current study. Biological datasets often exhibit substantial noise, which can hinder downstream analyses and affect the robustness of scientific conclusions. When dealing with human-derived data, the levels of noise and inter-individual variability are typically higher, introducing additional complexity to analysis and interpretation. A primary goal of ADAM is to reduce reliance on manual intervention by employing systematic reasoning, guided by literature, data, and computational logic, to identify and mitigate the influence of confounding factors through integrated computational and semantic strategies on a case-by-case basis.

Despite these challenges, ADAM demonstrates that when combined with RAG, CoT reasoning, and traditional machine learning models, large language models can effectively contextualize biological data, even in small-sample or high-noise scenarios such as the data used in this study, which was collected from five nursing home sites. In ADAM, CoT reasoning is guided by RAG to retrieve literature-grounded evidence for each inference step, while traditional models contribute statistical robustness. This combination yields both computational precision and semantically grounded outputs, thereby reducing hallucination and enhancing reliability. This capability is particularly crucial in microbiome clinical research, where important patterns are often subtly dispersed across different data types. Employing explainable AI methods, such as SHAP values, enhances interpretability, allowing researchers to connect a model's predictions to specific features or insights drawn from the literature.

However, this study has some limitations that should be acknowledged. First, although the evaluation was performed on a relatively small but high-dimensional dataset, the ADAM framework was designed to support various types of biological data by using a modular, agent-based architecture. This includes distinct agents for computation, summarization, and classification, each of which can be independently scaled and parallelized to handle increasing data complexity and volume. However, performance at larger scales still needs systematic validation. Second, while the ADAM framework demonstrated strong generalization in this study, additional training, fine-tuning, and reinforcement learning strategies are recommended to reduce hallucinations when applied to

unseen datasets or unfamiliar biomedical areas. Third, although the semantic search engine includes a broad literature corpus to reduce noise and variability in raw data and intermediate outputs, its effectiveness depends on the coverage, quality, and representativeness of the embedded knowledge. Gaps or biases in the literature may lead to missed connections or skewed interpretations, notably in underrepresented disease areas. Additionally, the system currently relies on two LLMs pretrained for general-purpose use, which may lack the domain-specific accuracy needed for tasks in biological and clinical settings where small differences can be critical. Finally, the reasoning mechanisms within the summarization and classification agents are manually tuned, which may limit adaptability across different datasets. Future improvements will involve automated tuning and domain-specific optimization to enhance portability and reduce reliance on manual adjustments.

Confounding remains a significant obstacle in modeling Alzheimer's disease because of clinical heterogeneity. We included all structured clinical and microbiome features in the model to enable data-driven adjustment of observed covariates, using XGBoost with regularization and embedded feature selection. Known confounders such as polypharmacy and proton pump inhibitor use appeared as top contributors. However, causal inference methods were not used, and residual confounding may still be present. SHAP values help interpretability but do not establish causality, so findings should be viewed with caution.

In addition, the computational overhead associated with running multiple AI agents and real-time retrieval over large document bases may not be practical in all laboratory settings, particularly for those with limited infrastructure. Furthermore, although ADAM enables comprehensive data analysis and biological reasoning, expert interpretation is necessary to validate the findings and ensure scientific rigor. It is also important to recognize that large language models have inherent limitations in consistency, as test–retest reliability remains imperfect even when the generation temperature is set to zero [48, 49]. As this study involves biological data and given the non-deterministic nature of large language models, we repeated the result generation 30 times, following the Central Limit Theorem, to support the statistical reliability of our conclusions. The findings are further supported and contextualized using the most relevant content retrieved from a corpus of 76,751 peer-reviewed publications on Alzheimer's disease, which are integrated into the system's semantic retrieval engine.

XI. Conclusion and Future Work

This study offers a thorough evaluation of the ADAM framework for classifying AD by comparing its performance with that of the widely used machine-learning baseline, XGBoost. By leveraging LLMs, explainable AI, and retrieval-augmented generation, ADAM achieved better predictive performance and showed greater consistency across repeated experimental runs.

Across 30 independent trials, ADAM consistently achieved a higher mean F1 score (0.7263) than XGBoost (0.6774), with a statistically significant difference confirmed by the Mann-Whitney U test ($p = 0.0418$). The lower standard deviation of the ADAM F1 scores (0.0632 vs. 0.1217) and its favorable distribution characteristics further underscore its stability and robustness. The F1 variance of XGBoost (0.0148) indicated that XGBoost shows 3.71 times more variability than ADAM (0.0040), as supported by Levene's test ($p = 0.0300$). These performance characteristics are particularly important in the medical domain, where model stability and reproducibility are essential for building trust in diagnostic tools and supporting consistent decision-making in patient care.

The ADAM framework offers a promising direction for integrating AI-driven inference with biological and clinical data, providing higher accuracy, greater interpretability, and operational stability. These results support its potential for adoption in biomedical research environments and clinical decision support systems, particularly in settings characterized by small, noisy, or imbalanced datasets. Future work will examine scalability and generalizability across more complex, multimodal biological inputs, including integrated multi-omics datasets, to broaden the framework's applicability and performance. It will also examine the integration of causal inference methods to enhance confounder adjustment and improve the interpretability of feature–outcome relationships in high-dimensional observational data.

Overall, ADAM addresses the crucial gap between experimental biology and agentic AI. Future enhancements, such as domain-specific LLMs and fine-tuning, expanded knowledge integration with RAG model optimization, adaptive reasoning logic using reinforcement learning, and large-scale, real-time validation mechanisms, are essential for broader adoption and translational impact. The reasoning logic will include three modes: manual, automatic, and hybrid, in a future release of ADAM, allowing flexible adaptation across various research and data environments. These developments may ultimately enable a foundational LLM trained and explicitly tuned for AD research, driving this line of research toward the realization of physical AI systems capable of interacting with and reasoning about complex biological data in real-time. It is important to note that ADAM is currently designed as a research tool only and should not be used for clinical application in its current form. This represents a promising future development for ADAM after further enhancement and validation outside of clinical settings.

CODE AVAILABILITY

The complete code that supports all the findings and analyses presented in this study is available at https://github.com/melhzy/ADAM. This repository includes data preprocessing scripts, model training scripts, evaluation scripts, and documentation detailing the setup procedures and dependencies. The code is distributed under the MIT License, permitting reuse and modification with appropriate attribution.

Please use the repository issue tracker or contact the corresponding author, Dr. Ziyuan Huang, for any questions or concerns. For inquiries related to Alzheimer's and its microbiome, please contact Dr. John P. Haran. Dr. Bucci Vanni should be contacted for computational biology inquiries.

DATA AVAILABILITY

All data are available under BioProject accession number PRJNA529586 at NCBI and are further described in the Supplementary Data at doi.org/10.1128/mBio.00632-19 at mBio.

ETHICS APPROVAL

This prospective cohort study was approved by the Institutional Review Board (IRB) of the University of Massachusetts Medical School (Docket H00010892).

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the members of the Bucci Lab at UMass Chan Medical School for their valuable contributions to this work. Their insightful feedback, technical support, and collaborative spirit significantly enhanced the development and refinement of this manuscript. We are especially thankful for their expertise and dedication throughout the research process.

The ADAM framework automatically generated the sample reports provided in TABLE IV, utilizing GPT-4o and GPT-4o-mini as its backend LLMs, along with ADAM's computational, summarization, and classification agents, to process and analyze clinical and microbiome data for Alzheimer's disease summarization and classification.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this paper.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* vol. 25, 2012, doi: 10.1145/3065386.

[2] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI,* 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. [Accessed: Jul. 11, 2025].

[3] A. Radford *et al.*, "Language models are unsupervised multitask learners," *OpenAI,* vol. 1, no. 8, p. 9, 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. [Accessed: Jul. 11, 2025].

[4] Anthropic, "The Claude 3 Model Family: Opus, Sonnet, Haiku," *Anthropic,* 2024. [Online]. Available: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. [Accessed: Jul. 11, 2025].

[5] T. B. Brown *et al.*, "Language models are few-shot learners," presented at the Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 2020. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/3495724.3495883. [Accessed: Jul. 11, 2025].

[6] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *preprint arXiv:2203.02155,* 2022, doi: 10.48550/arXiv.2203.02155.

[7] Gemini Team Google *et al.*, "Gemini: A family of highly capable multimodal models," *preprint arXiv:2312.11805,* 2023, doi: 10.48550/arXiv.2312.11805.

[8] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *preprint arXiv:2302.13971,* 2023, doi: doi.org/10.48550/arXiv.2302.13971.

[9] A. Vaswani *et al.*, "Attention is all you need," presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295349. [Accessed: Jul. 11, 2025].

[10] X. Bi *et al.*, "Deepseek llm: Scaling open-source language models with longtermism," *preprint arXiv:2401.02954,* 2024, doi: 10.48550/arXiv.2401.02954.

[11] OpenAI, "OpenAI GPT-4.5 System Card," 2025. [Online]. Available: https://openai.com/index/gpt-4-5-system-card/. [Accessed: Jul. 11, 2025].

[12] OpenAI, "OpenAI o1 System Card," 2025. [Online]. Available: https://openai.com/index/openai-o1-system-card/. [Accessed: Jul. 11, 2025].

[13] K. Kavukcuoglu, "Gemini 2.5: Our most intelligent AI model." [Online]. Available: https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. [Accessed: Jul. 11, 2025].

[14] Anthropic, "Claude 3.7 Sonnet and Claude Code," 2025. [Online]. Available: https://www.anthropic.com/news/claude-3-7-sonnet. [Accessed: Jul. 11, 2025].

[15] DeepSeek-AI *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *preprint arXiv:2501.12948,* 2025, doi: 10.48550/arXiv.2501.12948.

[16] P. Zhang, J. Shi, and M. N. K. Boulos, "Generative AI in Medicine and Healthcare: Moving Beyond the 'Peak of Inflated Expectations'," *Future Internet,* vol. 16, no. 12, 2024, doi: 10.3390/fi16120462.

[17] A. Hanafi, M. Saad, N. Zahran, R. J. Hanafy, and M. E. Fouda, "A Comprehensive Evaluation of Large Language Models on Mental Illnesses," 2024, doi: 10.48550/arxiv.2409.15687.

[18] Q. Jin *et al.*, "Demystifying Large Language Models for Medicine: A Primer," 2024, doi: 10.48550/arxiv.2410.18856.

[19] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature,* vol. 596, no. 7873, pp. 583-589, 2021, doi: 10.1038/s41586-021-03819-2.

[20] Z. Lin *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science,* vol. 379, no. 6637, pp. 1123-1130, 2023, doi: 10.1126/science.ade2574.

[21] O. Wysocki *et al.*, "An LLM-based Knowledge Synthesis and Scientific Reasoning Framework for Biomedical Discovery," 2024, doi: 10.48550/arxiv.2406.18626.

[22] H. Zhang, X. Xiong, M. Cheng, L. Ji, and K. Ning, "Deep learning enabled integration of tumor microenvironment microbial profiles and host gene expressions for interpretable survival subtyping in diverse types of cancers," (in eng), *mSystems,* vol. 9, no. 12, p. e0139524, Dec 17 2024, doi: 10.1128/msystems.01395-24.

[23] X. Wang *et al.*, "A pathology foundation model for cancer diagnosis and prognosis prediction," *Nature,* pp. 1-9, 2024, doi: 10.1038/s41586-024-07894-z.

[24] E. H. Abdelaziz, R. Ismail, M. S. Mabrouk, and E. Amin, "Multi-omics data integration and analysis pipeline for precision medicine: Systematic review," (in eng), *Comput Biol Chem,* vol. 113, p. 108254, Dec 2024, doi: 10.1016/j.compbiolchem.2024.108254.

[25] J. Wu *et al.*, "Integrating transcriptomics, genomics, and imaging in Alzheimer's disease: A federated model," *Frontiers in radiology,* vol. 1, p. 777030, 2022, doi: 10.3389/fradi.2021.777030.

[26] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," presented at the Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 2020. [Online]. Available: https://dl.acm.org/doi/pdf/10.5555/3495724.3496517. [Accessed: Jul. 11, 2025].

[27] J. Wei *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," presented at the Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 2022. [Online]. Available: https://dl.acm.org/doi/10.5555/3600270.3602070. [Accessed: Jul. 11, 2025].

[28] J. P. Haran *et al.*, "Alzheimer's disease microbiome is associated with dysregulation of the anti-inflammatory P-glycoprotein pathway," *MBio,* vol. 10, no. 3, pp. 10-1128, 2019 2019, doi: 10.1128/mBio.00632-19.

[29] K.-H. Lin *et al.*, "Benchmarking large language models GPT-4o, llama 3.1, and qwen 2.5 for cancer genetic variant classification," *npj Precision Oncology,* vol. 9, no. 1, p. 141, 2025, doi: 10.1038/s41698-025-00935-4.

[30] M. Liu *et al.*, "Evaluating the Effectiveness of advanced large language models in medical Knowledge: A Comparative study using Japanese national medical examination," *International Journal of Medical Informatics,* vol. 193, p. 105673, 2025, doi: 10.1016/j.ijmedinf.2024.105673.

[31] Y. Chen *et al.*, "Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study," *BMC Medical Education,* vol. 24, no. 1, p. 1372, 2024, doi: 10.1186/s12909-024-06309-x.

[32] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016.

[33] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295230. [Accessed: Jul. 11, 2025].

[34] I. A. Murad, M. I. Khaleel, and M. Y. Shakor, "Unveiling GPT-4o: Enhanced Multimodal Capabilities and Comparative Insights with ChatGPT-4," *International Journal of Electronics and Communications Systems,* vol. 4, no. 2, 2024, doi: 10.24042/ijecs.v4i2.25079.

[35] OpenAI, "GPT-4o-mini: advancing cost-efficient intelligence," *OpenAI,* 2024. [Online]. Available: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. [Accessed: Jul. 11, 2025].

[36] D. Van Veen *et al.*, "Adapted large language models can outperform medical experts in clinical text summarization," *Nature Medicine,* vol. 30, no. 4, pp. 1134-1142, 2024, doi: 10.1038/s41591-024-02855-5.

[37] OpenAI, "OpenAI Platform Documentation: Embeddings," *OpenAI,* 2024. [Online]. Available: https://platform.openai.com/docs/guides/embeddings#embedding-models. [Accessed: Jul. 11, 2025].

[38] OpenAI, "Pricing," *OpenAI,* 2025. [Online]. Available: https://platform.openai.com/docs/pricing. [Accessed: Jul. 11, 2025].

[39] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019: Association for Computing Machinery, pp. 2623–2631, doi: 10.1145/3292500.3330701.

[40] D. J. Hand, P. Christen, and N. Kirielle, "F*: an interpretable transformation of the F-measure," *Machine Learning,* vol. 110, no. 3, pp. 451 - 456, 2021, doi: 10.1007/s10994-021-05964-1.

[41] W. Bokhari and A. Bansal, "Asymmetric Error Control for Binary Classification in Medical Disease Diagnosis," in *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Laguna Hills, CA, USA, 2020: IEEE, pp. 25-32, doi: 10.1109/AIKE48582.2020.00013.

[42] G. Charizanos, H. Demirhan, and D. İçen, "Binary classification with fuzzy logistic regression under class imbalance and complete separation in clinical studies," *BMC medical research methodology,* vol. 24, no. 1, p. 145, 2024, doi: 10.1186/s12874-024-02270-x.

[43] J. L. Andrew, "Getting it wrong: the clinical misdiagnosis of Alzheimer's disease," *International Journal of Clinical Practice,* vol. 58, no. 11, pp. 1092-1094, 2004, doi: 10.1111/J.1368-5031.2004.00314.X.

[44] C. E. Emily *et al.*, ""Missed" Mild Cognitive Impairment: High False-Negative Error Rate Based on Conventional Diagnostic Criteria," *Journal of Alzheimer's Disease,* vol. 52, no. 2, pp. 685-691, 2016, doi: 10.3233/JAD-150986.

[45] D. M. Leslie, G. L. Timothy, and D. M. Keith, "Alzheimer's Disease: The Problem of Incorrect Clinical Diagnosis," *Journal of Geriatric Psychiatry and Neurology,* vol. 6, no. 4, pp. 230-234, 1993, doi: 10.1177/089198879300600409.

[46] N. Nachar, "The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution," *Tutorials in quantitative Methods for Psychology,* vol. 4, no. 1, pp. 13-20, 2008, doi: 10.20982/tqmp.04.1.p013.

[47] A. Gelman and J. Hill, "Analysis of variance," in *Data Analysis Using Regression and Multilevel/Hierarchical Models*: Cambridge University Press, 2006, ch. 22, pp. 487 - 502.

[48] E. Klishevich, Y. Denisov-Blanch, S. Obstbaum, I. Ciobanu, and M. Kosinski, "Measuring determinism in large language models for software code review," *preprint arXiv:2502.20747,* 2025, doi: 10.48550/arXiv.2502.20747.

[49] M. Lee, P. Liang, and Q. Yang, "Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities," in *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022, pp. 1-19, doi: 10.1145/3491102.3502030.

Ziyuan Huang (Member, IEEE) is a Postdoctoral Associate in the Department of Emergency Medicine and the Department of Biology at the University of Massachusetts Chan Medical School and part of the Microbiology & Microbiome Dynamics AI Hub in Worcester, MA, USA. He has expertise in large language models, deep learning, machine learning, and optimization. Dr. Huang earned his Ph.D. in data sciences from Harrisburg University of Science and Technology, Harrisburg, PA, USA, in 2023, with a focus on artificial neural networks and optimization. Additionally, he earned a master's degree in management information systems from Metropolitan State University, Minneapolis, USA, in 2014.

He is currently a Postdoctoral Associate at the University of Massachusetts Chan Medical School in Worcester, MA, USA. Previously, he worked as a Data Science Research Fellow at Harrisburg University of Science and Technology, where he supported the instruction of courses in data science and machine learning. During his academic tenure, he engaged in various research projects focused on the development and optimization of artificial neural network models, particularly in the areas of deep learning, retrieval-augmented generation, and large language model integration. He has contributed to publications, poster presentations, book chapters, and paper reviews that advance the understanding of machine learning optimization and the application of deep learning in biological and medical research.

Dr. Huang is a member of the Alzheimer's Association International Society to Advance Alzheimer's Research and Treatment (ISTAART) and the Gerontological Society of America (GSA). He is also an active member of the Microsoft Certified Professional program. Dr. Huang is dedicated to fostering interdisciplinary collaboration and advancing research at the intersection of artificial intelligence and biological science.

Vishaldeep Kaur Sekhon is a Senior Biostatistician in the Division of Geriatrics at the Johns Hopkins University School of Medicine. She is a member of the Gerontological Society of America (GSA). Her expertise includes analyzing large datasets, such as electronic medical records and claims data, focusing on oncology and Alzheimer's dementia. Her research encompasses studying cancer treatment patterns and understanding the epidemiological landscape of neurodegenerative diseases to inform evidence-based healthcare decisions.

Roozbeh Sadeghian (Member, IEEE), Dr. Roozbeh Sadeghian received his B.S. in electrical engineering from Isfahan University of Technology, Iran, in 2002, his M.S. in electrical engineering from Shiraz University, Iran, in 2005, and his Ph.D. in electrical engineering from the State University of New York at Binghamton, USA, in 2017. His major field of study is speech recognition and data analytics, focusing on the intersection of machine learning and healthcare applications.

He has extensive experience in academia and industry, currently serving as the Program Director of Data Analytics at Harrisburg University, Harrisburg, PA. His current role as Associate Professor includes projects in applying machine learning to diagnose Alzheimer's disease, using computer vision for wildlife-vehicle collision prevention, and designing AI systems for diagnosing and preventing several health disorders. He has also contributed significantly to mentoring Ph.D. and M.S. students, overseeing diverse research projects, and publishing numerous articles in prestigious journals.

Dr. Sadeghian is a member of IEEE and other professional societies, and he has been recognized with awards such as the PA Governor's Award for Environmental Excellence and the Best Paper Award at the IEEE CSCI conference.

Maria Vaida is an Assistant Professor of Data Science at Harrisburg University, specializing in advanced computational methods within the realms of genomics and metabolomics. With a focus on graph neural networks and ensemble modeling, Dr. Vaida's research aims to develop innovative machine learning and graph neural network frameworks that can uncover complex biological insights and enhance predictive accuracy in multi-omics data analysis.

Cynthia Jo is a Research Associate in Dr. John Haran's lab in the Department of Emergency Medicine at the University of Massachusetts Chan Medical School. She graduated with a B.S. in Biology from the University of Massachusetts Amherst in 2023 and is currently studying the relationships between immune system activity, variations in the gut microbiome, and the presence of Alzheimer's.

Beth A. McCormick received the B.A. degree in microbiology with a minor in history from the University of New Hampshire, Durham, NH, USA, and the Ph.D. degree in microbiology from the University of Rhode Island, Kingston, RI, USA. She completed postdoctoral training in gastrointestinal pathophysiology and infectious disease at Children's Hospital and Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

She is currently the Worcester Foundation for Biomedical Research Chair II and serves as Chair and Professor in the Department of Microbiology at the University of Massachusetts Chan Medical School, Worcester, MA, USA. She is also the Founding Director of both the Program in Microbiome Dynamics and the Center for Microbiome Research. Her research focuses on the interactions between gut microbes and gastrointestinal physiology, particularly the role of microbes and their metabolites in influencing host physiology through intestinal multidrug efflux transporters and epithelial cell nuclear factors. Her work has led to the development of novel therapeutic strategies for gastrointestinal disorders, including the co-founding of Adiso Therapeutics, which is advancing treatments for ulcerative colitis.

Among Dr. McCormock's honors and awards, she is an elected Fellow of the American Academy of Microbiology, an elected Fellow of the American Gastroenterological Association, a recipient of the American Gastroenterological Association Mentor of the Year, and was named a Distinguished Scientist by the American Physiological Society. She serves as Editor-in-Chief of the journal Gut Microbes and holds leadership positions in several professional organizations, including Chair of the Basic and Clinical Intestinal Disorders Council of the American Gastroenterological Association, and is an elected member of the Board of Directors for the American Physiological Society.

Doyle V. Ward received the PhD degree in Molecular Biology from Princeton University and completed postdoctoral training at the University of California, Berkeley. Dr. Ward is now Associate Professor in the Department of Microbiology and the Program in Microbiome Dynamics and is Director of Operations for the Center for Microbiome at UMass Chan Medical School in Worcester, MA. Dr. Ward's research focuses on microbiome and microbial genomics in human diseases.

Vanni Bucci is a Microbiology Professor and co-directs the Microbiome Dynamics Program at UMass Chan Medical School. Dr. Bucci earned his Bachelor of Science in Civil Engineering from the University of Florence in 2006. He then completed his Master of Science and Ph.D. in Civil Engineering at Northeastern University in 2008 and 2010, respectively. From 2010 to 2013, Dr. Bucci conducted postdoctoral research in computational biology at the Memorial Sloan Kettering Cancer Center. His research focuses on developing computational modeling and experimental methods to understand complex microbial/microbiome dynamics, particularly how environmental influences such as diet and medical interventions impact health and disease. With a background in engineering and computational biology, he possesses extensive experience in microbiology and synthetic biology. Dr. Bucci's laboratory pioneered frequentist and Bayesian regression techniques to infer host-microbiome dynamics from time-series data. These methods have been applied to predict the dynamics of intestinal commensal and enteropathogenic bacteria and optimize bacterial consortia that induce microbiome-dependent immune responses.

Furthermore, Dr. Bucci's lab developed novel interpretable machine and deep learning methods to link microbiome dynamics to clinical outcomes. This work has uncovered associations between the microbiome, innate immunity induction, and Alzheimer's Disease risk, as well as microbiome links to *C. difficile* infection. These innovative computational and machine learning techniques have opened new research avenues into the relationship between the microbiome and host immunity in response to lung pathogens like Myc tuberculosis and SARS-CoV-2, as well as during interventions to address multiple inflammatory conditions.

On the experimental front, Dr. Bucci's lab develops microbiome-based therapeutics to treat infectious, inflammatory, and neurological disorders. His team created one of the first synthetically engineered probiotics that inhibits pathogenic Salmonella by sensing Salmonella-induced intestinal inflammation. Leveraging this, his group explores targeted microbiome intervention strategies to eradicate drug-resistant Enterobacteriaceae, such as Carbapenem-Resistant *Klebsiella* and ESBL-producing *E. coli*, from the gastrointestinal tract. Recently, Dr. Bucci discovered how succinate-producing bacteria promote the expansion of colonic Tuft Cells, demonstrating that supplementation with these microbes protects against *C. difficile*-associated disease in a Pou2f3-dependent manner.

John P. Haran received his B.S. in biochemistry from Rensselaer Polytechnic Institute in 1999 and his M.D. from the University of Massachusetts Chan Medical School in 2007. He finished his residency training in Emergency Medicine at the Alpert Medical School of Brown University in 2011 and then his Ph.D. degree in biomedical sciences in 2018 at the University of Massachusetts Chan Medical School in Worcester, MA, USA.

He is currently Professor of Emergency Medicine and Microbiology at UMass Chan Medical School in Worcester Massachusetts. He is Research Director for the Department of Emergency Medicine and is also the Clinical Director for the Program in Microbiome Dynamics at UMass Chan. Dr. Haran's research background focuses on investigations into older adult health and associations with microbiome composition and health outcomes in both nursing home and community-dwelling patients. His earlier work had focused on the microbiome and how dysbiosis can lead to increased risk of bacterial colonization and infection. The centerpiece of Dr. Haran's current lab's focus is on the microbiome-gut-brain axis and cognitive function in older adults, especially those suffering from Alzheimer's disease. Working with longitudinal cohorts of older adults, his lab has begun to unravel the microbiome's role in inflammation, immune dysfunction, and neurodegeneration. Importantly, they have published findings on the dysbiotic microbiome in older adults with Alzheimer's disease and have replicated these results in both in vitro systems and transgenic Alzheimer's disease mouse models. His lab includes a collaborative team that has demonstrated the power of machine learning and AI in interpreting complex findings and discovering possible causal mechanisms in Alzheimer's disease.