Beyond Classical Approaches: Fine-Tuning Clinical BERT Model on Structured Data for Alzheimer's Disease Diagnosis

Hager Saleh [1,2,3,*], Michael McCann [4], John G. Breslin [1,5], Shaker El-Sappagh [6,7], and for the Alzheimer's Disease Neuroimaging Initiative [**]

[1] Insight Research Ireland Centre for Data Analytics, University of Galway, Galway H91 TK33, Ireland
[2] Atlantic Technological University, Letterkenny, Ireland
[3] Faculty of Computers and Artificial Intelligence, Hurghada University, Hurghada, Egypt
[4] Department of Computing, Atlantic Technological University, Letterkenny, Ireland
[5] School of Engineering, University of Galway, Galway H91 TK33, Ireland
[6] Faculty of Computer Science and Engineering, Galala University, Suez, Egypt
[7] Faculty of Computers and Artificial Intelligence, Benha University, Banha, Egypt
Email: hager.saleh.fci@gmail.com (H.S.); Michael.McCann@atu.ie (M.M.); john.breslin@universityofgalway.ie (J.G.B.); sh.elsappagh@gmail.com (S.E.-S.)
[*]Corresponding author

[**] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).

***Abstract*—Alzheimer's Disease (AD) is a neurodegenerative disorder requiring early diagnosis for effective intervention. This study presents a comprehensive evaluation of classical machine learning models, fine-tuned BERT-based models, and Large Language Models (LLMs) for AD diagnosis using structured data. The comparative analysis demonstrates that fine-tuned BERT-based models, particularly BioBERT and ClinicalBERT, achieve superior performance, with BioBERT attaining an accuracy of 94.48%, precision of 94.64%, recall of 94.48%, and an F1-score of 94.52% based only on structured numerical data. These models outperform classical approaches such as random forest and AdaBoost, which achieved accuracies of 91.87% and 91.41%, respectively. In contrast, LLMs, including GPT-4 and LLama 3.2, exhibited suboptimal results, with GPT reaching a maximum accuracy of 53.988% in the zero-shot setting, highlighting their limitations in handling structured clinical data without extensive fine-tuning. The study highlights the importance of domain-specific fine-tuning, as general-purpose LLMs struggle with structured data due to their reliance on prompt engineering. While traditional ML models provide a solid baseline, fine-tuned BERT models offer enhanced diagnostic capabilities by capturing intricate data relationships.**

***Keywords*—Alzheimer's disease diagnosis, Large Language Model (LLM), language model, in-context learning**

I. Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that affects millions of individuals globally [1]. It is characterized by cognitive deterioration, memory impairment, and behavioral changes, underscoring the critical need for early and accurate diagnosis to facilitate effective treatment and management strategies. Conventional diagnostic methods, such as clinical evaluations and neuroimaging techniques, often encounter challenges related to subjectivity and resource intensiveness [2]. In recent years, Machine Learning (ML) models have demonstrated significant potential in enhancing clinical decision-making by leveraging structured patient data to identify patterns indicative of AD [3]. These models offer the promise of improving diagnostic precision, alleviating the workload on healthcare professionals, and enabling timely interventions [4]. For comprehensive survey of the current applications of ML models in AD domain, readers are guided to read these survey papers [5]–[9].

Many modalities have been used to train ML models for the diagnosis of AD. Numerous studies have investigated the application of machine learning techniques using only structured data [10]. Structured data, encompassing demographic information, cognitive test scores, biomarker levels, and neuroimaging-derived metrics, provide a comprehensive representation of an individual's health profile. Traditional ML algorithms, such as support vector machines (SVM), random forests, and deep neural networks, have been employed to analyze such data, yielding encouraging results in differentiating AD from Mild Cognitive Impairment (MCI) and Cognitively Normal (CN) individuals [11] [12].

Although structured data-based ML models have achieved substantial progress, several challenges persist [13]. The challenges in the current literature of ML models include data quality issues and a lack of generalizability and interpretability [3]. To build a generalizable model, huge datasets are needed to train and externally validate the models, but these data are not always available in the medical domain. In addition, many existing approaches rely on manually engineered features, which may fail to capture complex interactions within the data. Furthermore, traditional ML models can experience performance degra-

Manuscript received February 11, 2025; revised March 11, 2025; accepted March 27, 2025; published June 19, 2025.

dation when confronted with heterogeneous data sources or missing values. The fine-tuning of pre-trained language models, such as BERT and its specialized variants such as BioBERT, ClinicalBERT, and BioRoBERTa, presents a promising solution to these challenges [14]. Large Language Models (LLMs) have demonstrated their capability to capture intricate relationships within textual data, and recent advancements allow their adaptation to structured data by effectively encoding tabular information [15].

BERT-based models have shown considerable promise in utilizing structured data to develop classifiers for AD diagnosis [16]. By representing structured inputs as sequences of embeddings, BERT models can discern meaningful relationships among various features, thereby enhancing classification accuracy. Moreover, fine-tuning domain-specific models such as BioBERT and ClinicalBERT, which are pre-trained on biomedical and clinical corpora, can further enhance the comprehension of AD-related features, facilitating the creation of robust and interpretable decision support systems [17]. The hypothesis of this study is that fine-tuning of language models such as BERT or its variants can support the detection of AD. Disease-specific language representation model can better understand structured data that are serialized as text [18]. The primary contributions of this study are as follows:

- Implement an end-to-end training and evaluation architecture for building a clinical decision support system for AD diagnosis based on structured data and language models such as BERT, which can be tailored directly for a wide range of clinical predictive tasks.
- An exploration of the feasibility of fine-tuning BERT models and their variants, such as BioBERT and ClinicalBERT, to analyze structured data for Alzheimer's disease diagnosis.
- A comprehensive evaluation of these models in comparison to large language models such as GPT-4 and LLaMA 3.2, and classical machine learning models such as decision tree (DT), support vector machine (SVM), K-nearest neighbor (KNN), logistic regression (LR), AdaBoost, random forest (RF), and naive Bayes (NB), highlighting their advantages and limitations.
- The development of a novel approach to encoding structured data in a format compatible with BERT-based models, enhancing the classification performance.

The remainder of this paper is arranged as follows. The proposed framework introduced in Section III. The experimental results are discussed in Section V. Finally, conclusions are shown in Section VIII.

II. Related Work

Machine learning models have been used to detect and diagnose AD using structured data. For example, Alexander *et al*. [19] utilized Electronic Health Records (EHRs) from the Clinical Practice Research Datalink (CPRD) to identify and evaluate clinical subtypes of AD using unsupervised ML techniques. The dataset included 7,913 AD patients (mean age: 82, 66.2% female) and covered a broad range of clinical features, including 21 variables encompassing symptoms, comorbidities, and demographic factors. The key symptoms analyzed were memory loss, confusion, neuropsychological symptoms (e.g., depression, hallucinations), and motor symptoms (e.g., difficulty walking). The comorbidities assessed included anxiety, atrial fibrillation, cancer, depression, diabetes, hypertension, kidney disease, rheumatoid arthritis, and heart failure. Demographic factors such as age of onset, gender, smoking status, and alcohol consumption were also considered. To identify subtypes, the study applied four clustering methods: k-means, kernel k-means, affinity propagation, and Latent Class Analysis (LCA). Among these, k-means produced the most stable and clinically interpretable results, identifying five subgroups. A notable and recurring cluster across multiple methods consisted of younger-onset, predominantly female patients with high rates of depression and anxiety, who exhibited a faster rate of cognitive decline.

Li *et al.* [20] explored ML approaches for the early prediction of AD using structured EHR data from the OneFlorida+ Research Consortium, analyzing 23,835 ADRD cases and 1,038,643 controls. It compared knowledge-driven models and data-driven models using Logistic Regression (LR) and Gradient Boosting Trees (GBT). GBT models achieved the highest AUC scores of 0.939, 0.906, 0.884, and 0.854 for ADRD prediction at 0, 1, 3, and 5 years before diagnosis, respectively. The study identified key predictive features, including age, cerebrovascular disease, mood disorders, diabetes, fatigue, medication history, laboratory results, and behavioral factors (e.g., smoking, alcohol use). Results demonstrated that data-driven models outperformed knowledge-driven models, but challenges remain, including data variability, generalizability, and the need for external validation. Gao *et al.* [21] applied ML to predict AD risk by integrating polygenic risk scores (PRSs) and EHRs from the AD Genetics Consortium (ADGC) and UK Biobank (UKB). Using eXtreme Gradient Boosting (XGBoost) and SHapley Additive exPlanations (SHAP), the model identified PRSs as the strongest predictors in individuals aged 65+, surpassing age. The dataset included genetic markers, demographics, lifestyle factors, comorbidities, and ICD-10 codes, with key predictors such as urinary tract infections, syncope, and chest pain. The model achieved an AUC of 0.88 for AD risk in individuals aged 40+ and 0.78 in those 65+, demonstrating the value of combining genetic and clinical data for early detection.

Tang *et al.* [22] utilized EHRs from the University of California, San Francisco (UCSF) and knowledge networks (SPOKE) to predict AD onset and identify sex-specific biological insights. The dataset included 749 AD cases and 250,545 controls, with ML models trained to predict AD up to 7 years before diagnosis. Random forest models achieved an AUC of 0.72 (7 years prior) to 0.81 (1 day prior), improving to 0.86–0.90 with demographic and visit-related features. Key predictors included hyperlipidemia, hypertension, osteoporosis (in females), and cognitive symptoms. Knowledge networks identified shared genetic associations, such as APOE for hyperlipidemia-related AD risk and MS4A6A for female-specific osteoporosis-related AD risk. The study validated findings using external EHR datasets and genetic colocalization analysis, emphasizing the need for integrating multi-omics data for personalized risk stratification. Kumar *et al*. [23] examined the application of ML in AD dementia using clinical data, particularly from EHRs. The study highlighted the increasing use of ML over the

past five years, emphasizing its role in early AD detection, personalized disease modeling, and non-invasive diagnosis. The review identified decision trees, neural networks, and SVMs as the predominant ML models, while also noting the limited use of unsupervised and semi-supervised learning. The study highlighted the challenges of the current literature of ML models, such as data quality issues and lack of generalizability and interpretability [3].

Recent advancements in natural language processing, particularly the development of models such as BERT and other Large Language Models (LLMs), have significantly impacted the early diagnosis of AD and related conditions. These models excel in understanding and generating human-like text, enabling the analysis of subtle linguistic patterns associated with cognitive impairments. In a notable study, researchers developed AD-BERT, a model pretrained on clinical notes, to predict the progression from Mild Cognitive Impairment (MCI) to AD. This model demonstrated superior performance compared to traditional methods, achieving an Area Under the receiver operating characteristic Curve (AUC) of 0.8170 and an F1 score of 0.4178 on one dataset, and an AUC of 0.8830 and an F1 score of 0.6836 on another, highlighting its potential in clinical settings [17]. Further research explored the application of prompt learning techniques with pretrained language models for AD detection. By fine-tuning models like BERT and RoBERTa using prompt-based methods, researchers achieved a mean detection accuracy of 84.20% on manual transcripts and 82.64% on automatic speech recognition transcripts, underscoring the efficacy of prompt learning in enhancing diagnostic accuracy [24]. Additionally, a study employing the LLAMA-2 model combined with prompt engineering achieved an accuracy of 81.31% in AD detection, marking a 4.46% improvement over baseline models. This approach emphasizes the adaptability and efficiency of LLMs in processing speech data for diagnostic purposes [25]. Collectively, these studies demonstrate that leveraging BERT and other LLMs facilitates the identification of linguistic markers indicative of cognitive decline, offering a non-invasive, cost-effective, and scalable approach to early AD diagnosis. The integration of these models into clinical practice holds promise for improving patient outcomes through timely intervention [26].

Recently, Shakeri and Farmanbar [27] and Shankar *et al.* [28] comprehensively examined the applications of Natural Language Processing (NLP) in AD research. The study underscored the increasing role of NLP in AD diagnostics, patient monitoring, and the evaluation of social and cognitive burdens associated with the disease. Large Language Models (LLMs) have many applications which span from automated analysis of spontaneous speech to extracting insights from EHRs, social media discussions, and scientific literature. The integration of NLP models with structured EHR data has proven particularly useful for identifying AD biomarkers, predicting disease progression, and assessing risk factors. The study highlighted the significant role of BERT-based models in AD research, particularly in automated diagnosis and cognitive assessment. BERT and its variants, such as DistilBERT, BioBERT, and ClinicalBERT, have been widely utilized for feature extraction from speech transcripts, clinical notes, and biomedical literature. These models demonstrated strong performance in differentiating AD from healthy controls, predicting cognitive scores, and identifying AD-related risk factors in electronic health records. Their ability to leverage contextual embeddings has enhanced the accuracy of NLP-based classification tasks, outperforming traditional machine learning approaches. Furthermore, fine-tuned BERT models have been instrumental in multilingual AD detection, enabling improved generalizability across diverse linguistic datasets. Despite these advancements, challenges remain in optimizing BERT models for interpretability and ensuring their robustness in real-world clinical applications. However, the survey did not mention any studies that serialized the structured data and using it for AD classification.

BERT-based models demonstrated state-of-the-art performance in the detection of Alzheimer's Disease (AD) based on clinical notes text data. For example, Wang et al. [29] fine-tuned a ClinicalBERT model using Electronic Health Record (EHR) notes obtained from an academic health system. Their approach achieved an Area under the Curve (AUC) of 0.997 and an Area under the Precision-Recall Curve (AUPRC) of 0.929 on an independent test set. The model was capable of identifying early indicators of cognitive decline up to four years prior to the initial diagnosis of Mild Cognitive Impairment (MCI). Similarly, Du *et al.* [30] employed an ensemble approach that integrated BERT-based models with traditional machine learning techniques to detect MCI from EHR notes. Their method yielded a sensitivity of 94.2% and an F1 score of 92.2%. These transformer-based architectures effectively captured intricate semantic relationships within clinical text, thereby enhancing the accuracy of detecting subtle cognitive impairments. Using numerical data, Gu *et al.* [31] proposed a transformer-based multimodal model that converts structured clinical data into textual sequences and uses a BERT encoder alongside a vision transformer for MRI analysis. This model achieved 87.9% accuracy on ADNI, surpassing single-modality approaches. Liu *et al.* [32] developed a transformer network embedding tabular data (i.e., age, APOE genotype, cognitive scores) alongside MRI features, yielding state-of-the-art accuracy.

Incorporating numerical data into BERT-based models necessitates effective serialization techniques to ensure compatibility with BERT's architecture, which is primarily designed for textual inputs. Various serialization techniques have been explored including textual embedding, concatenation with embeddings, and specialized transformer architectures [33]. As a result, integrating BERT-based models with the structured data offers a promising approach to enhance AD diagnosis. However, no study int the literature has examined the role of multimodal numerical data such as cognitive assessments, biomarkers, and imaging metrics to uncover subtle patterns that may be overlooked by traditional ML methods.

III. Methodology

Machine learning model can be used as an engine to build a Clinical Decision Support System (CDSS). However, these models need huge datasets and the resulting models are specialized in specific tasks. Language models such as BERT and Large Language Models (LLMs) such as GPT-4o and LLaMA-3, on the other hand, can be used as CDSS

engines with little fine-tuning because they are pre-trained with huge dataset. The pre-trained models have general knowledge which are customizable in specific domains with little fine-tuning. The study proposes a CDSS architecture for AD diagnosis based on LM and LLM models. In this section, we discuss the used dataset for model fine-tuning and the proposed architecture.

A. Data Description

In this study, we use a dataset from the ADNI dataset [34]. The dataset contains 811 entries and 49 features. These features include the the most important features for AD diagnosis include demographic data (e.g., age, gender, education), biomarkers (e.g., FDG, AV45, and TAU), cognitive assessments (e.g., MMSE and ADAS-Cog scores), and imaging measurements (e.g., hippocampus volume and ventricles size). The dataset is selected from the baseline visits of the patients. No timeseries data are included in the study. In this study, the 811 participants (53.62% male and 47.75) aged between 71 and 75 years, fluent in English or Spanish, and with at least 15 years of education, are categorized based on 300 individuals (37.5%) who exhibit Mini-Mental State Examination (MMSE) scores of 29.03±1.11; (b) Mild Cognitive Impairment (MCI): 300 individuals (37.5%) consistently diagnosed with MCI, with MMSE scores of 27.66±1.84; and (d) Alzheimer's Disease (AD): 211 individuals (26.37%) diagnosed with AD throughout the study, with MMSE scores of 23.33±2.06. Statistical analyses revealed significant differences in MMSE scores among the groups, as well as variations in age and years of education between specific groups. Table I shows the statistics of some of the used numerical features.

B. Proposed Model

In the development of CDSS, LMs have demonstrated superior performance compared to traditional machine learning algorithms. This improvement can be attributed to the fundamental differences in their architecture, training paradigms, and ability to capture complex relationships within clinical text. Traditional machine learning models, including decision trees, support vector machines, and random forests, typically rely on handcrafted features and structured data inputs. These models are constrained by their dependence on feature engineering and often struggle to generalize across diverse and unstructured clinical data sources. In contrast, LMs and LLMs leverage deep neural networks, particularly transformer architectures, to model intricate relationships within unstructured clinical text, such as electronic health records, clinical notes, and medical literature. Pre-trained on vast amounts of biomedical and general-domain text, these models inherently possess a deep understanding of medical terminology, contextual relationships, and linguistic nuances, allowing them to capture subtle patterns that traditional models might overlook. For instance, BERT's bidirectional contextual embeddings enhance clinical named entity recognition and relation extraction tasks by considering both left and right context, leading to more accurate and contextually aware predictions.

Moreover, LLMs such as GPT-4o and LLaMA-3 go beyond contextual embeddings by incorporating generative capabilities, enabling them to provide sophisticated responses to clinical queries, summarize patient records, and support complex decision-making processes. Their extensive training on diverse datasets allows them to generalize across multiple clinical scenarios with minimal fine-tuning. Furthermore, their ability to perform zero-shot and few-shot learning further enhances their utility in clinical environments where labeled data may be limited. Another significant advantage of LLMs is their adaptability to multimodal inputs, combining textual data with structured clinical data such as lab results and imaging reports, thereby providing a more holistic view of patient health. This flexibility positions LLMs as robust tools for personalized medicine, risk stratification, and predictive analytics. The study proposed a CDSS based on LM for the diagnosis of AD. The problem is formulated as a three-class classification task. Fig. 1 shows the proposed architecture of the LM model for the diagnosis of AD. In the following, we discuss the steps of the proposed framework.

1) Data extraction: In this step, we collect the medically relevant collection of features for AD diagnosis from the ADNI dataset. These features have proved their relevance in the literature [11]. The used structured dataset is in tabular form and contains a collection of features from different modalities such as demographic, genetic, and neuropsychological features, which provide a complete picture about the AD patients. Before doing any data preprocessing steps, the raw dataset is split randomly into a training set (80%) and test set (20%) in a stratified way. This prevent any leakage between training and testing data [3].

2) Data preprocessing: The initial dataset has 2379 cases. Of the 2379, we selected 1522 cases with little missing values. All cases and features with more than 30% missing are removed from the dataset. The resulting dataset is randomly split into a training set (80%) and test set (20%) in a stratified way. Traning set is used to train and validate the classical ML models and to finetune the BERT-based models. The independent test set is used to evaluate the generalization performance of the resulting models. In this step, we improve the quality of the dataset. A set of two sequential steps are done on the raw data including (1) handling of missing values and (2) data balancing. We use the KNN technique to handle the missing values of the resulting dataset. The resulting data set is unbalanced. We perform the down-sampling technique to balance the dataset. This process results in a clean and balanced dataset of 811 instances. The resulting dataset is used to train and test the classical machine learning models such as KKN, DT, RF, SVM, AdaBoost, NB, and LR. The results of these models are used as a benchmark to compare them with BERT-based and LLM-based CDSS.

3) Data serialization: LM and LLM models can not work directly with tabular and numerical data. As a result, in this step, we serialize the numerical data of all cases as text data. The serialization process has been done programmingly by converting the vector of each example into a description of the case as a paragraph. The conversion is based on a prompt template that highlights the features in the dataset. Our methodology is inspired by [35]. Examples of cases from the three AD classes are as follows. Note that, for space restrictions, the cases are represented with subsets of the features

Example 1 (Class: CN): "Patient is a 67.5-year-old Male

TABLE I
DATASET DESCRIPTION

Feature	Total	CN	MCI	AD
AGE	71.90 ± 7.28	71.31 ± 6.54	71.69 ± 7.47	74.35 ± 8.23
PTEDUCAT	16.34 ± 2.53	16.71 ± 2.38	16.20 ± 2.58	15.72 ± 2.62
APOE4	0.53 ± 0.65	0.35 ± 0.54	0.58 ± 0.67	0.89 ± 0.73
FDG	1.24 ± 0.15	1.32 ± 0.11	1.26 ± 0.13	1.07 ± 0.14
AV45	1.21 ± 0.23	1.11 ± 0.17	1.22 ± 0.23	1.40 ± 0.22
TAU	279.79 ± 131.61	254.12 ± 112.45	272.36 ± 127.83	352.48 ± 141.25
MMSE	27.64 ± 2.50	29.03 ± 1.11	27.66 ± 1.84	23.33 ± 2.06
Hippocampus	7148.79 ± 1114.85	7485.32 ± 1023.67	7198.67 ± 1095.82	6283.12 ± 985.67
Ventricles	36871.38 ± 20673.16	31542.78 ± 18456.54	36124.65 ± 19845.23	47281.89 ± 21345.87
WholeBrain	1.05e6 ± 1.08e5	1.07e6 ± 1.02e5	1.05e6 ± 1.07e5	9.82e5 ± 1.12e5

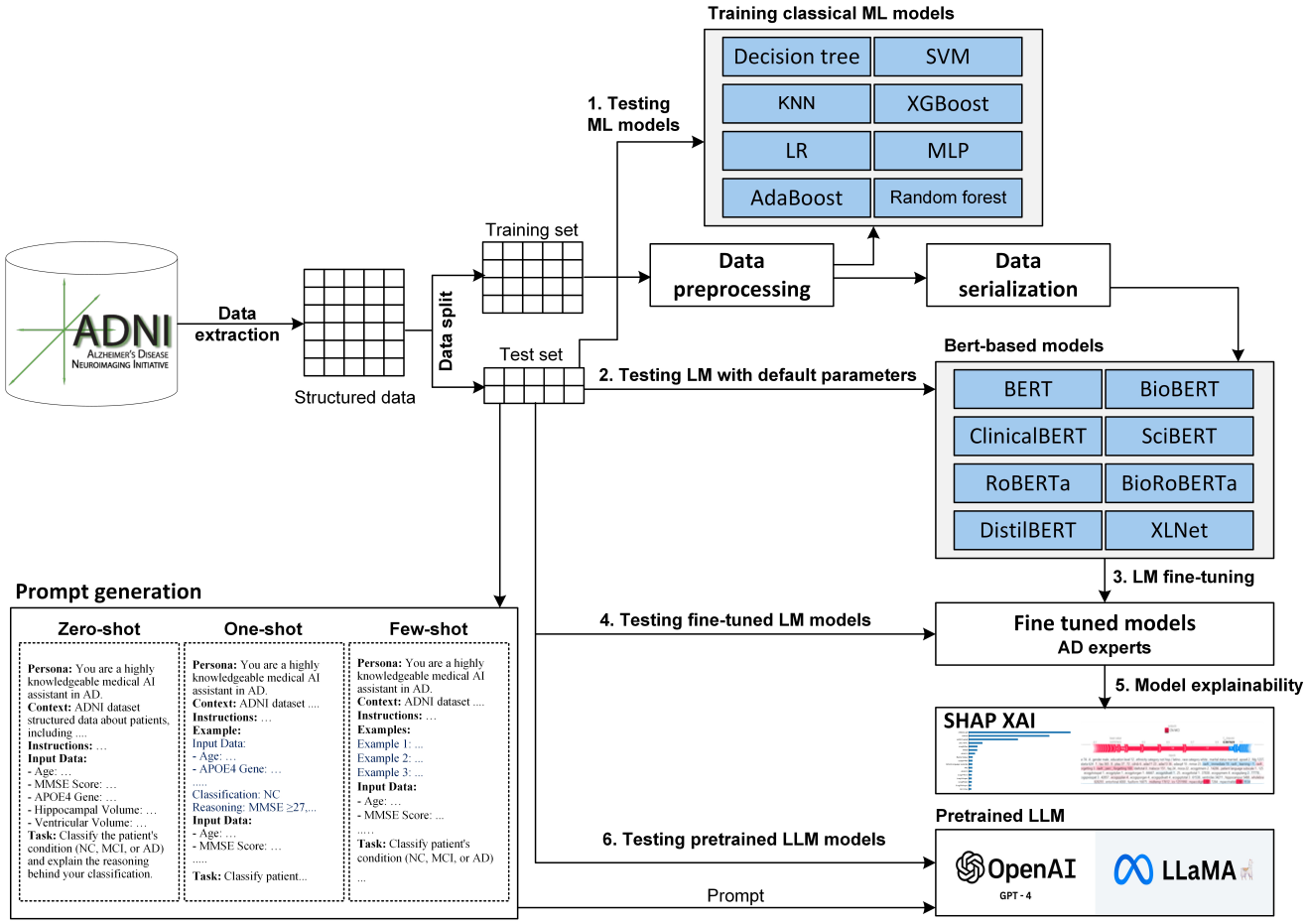

Fig. 1. The proposed architecture.

with 16 years of education, identified as Not Hisp/Latino and Black, and has never been married. APOE4 gene count: 0. FDG PET: 1.25096, AV45 PET: 0.9851, ABETA: 731.8, Hippocampus volume: 8297, Whole brain volume: 1165500. Neuropsychological scores: mPACCdigit: 3.09162, mPACCtrailsB: 2.411920, ... Diagnosis: CN."
Example 2 (Class: MCI): "Patient is a 64.8-year-old Male with 16 years of education, identified as Not Hisp/Latino and White, and is married. APOE4 gene count: 0. FDG PET: 1.20619, AV45 PET: 0.9692, ABETA: 1504.0, Hippocampus volume: 7960, Whole brain volume: 1135560. Neuropsychological scores: mPACCdigit: -1.57740, mPACCtrailsB: -0.738264, ... Diagnosis: MCI."
Example 3 (Class: AD): "Patient is a 62.7-year-old Female with 20 years of education, identified as Not Hisp/Latino and White, and is married. APOE4 gene count: 0. FDG PET: 1.35060, AV45 PET: 1.0180, Hippocampus volume: 7694, Whole brain volume: 1012760. Neuropsychological scores: mPACCdigit: -17.11820, mPACCtrailsB: -16.126200, ... Diagnosis: AD."

4) LM fine-tuning: BERT is an encode only transformer based model which can understand text and generate deep embedding. BERT is a pretrained large language model, which can be used directly for natural language understanding [36]. It has been used to build classification and decision support models in the medical domain [37], [38]. The [CLS] embedding vector contains the semantic and deep features of the whole case. A classification head based on Multilayer Perceprton (MLP) and SoftMax activation function at the output layer can be used to learn the [CLS] embedding predict the AD class, see Fig. 2. We explore the capabilities of different BERT variants including BioBERT, ClinicalBERT, SciBERT, RoBERTa, BioRoBERTa, TabNet, TabTransformer, and DistilBERT.

a) BERT: Bidirectional Encoder Representations from Transformers (BERT) [36] is an encoder only language representation model that pre-trains deep bidirectional representations by jointly conditioning on both left and right context across all Transformer layers, allowing for a more comprehensive understanding of text compared to unidirectional models, see Fig. 2. BERT is pre-trained using two self-supervised tasks: the Masked Language Model (MLM), which randomly masks 15% of tokens in the input sequence and predicts the original tokens based on bidirectional context, and the next sentence prediction (NSP), which determines whether a given sentence follows another, enhancing the model's ability to understand sentence relationships. The model employs a multi-layer Transformer encoder and is available in two configurations—$\text{BERT}_{\text{BASE}}$ with 12 layers, 768 hidden units, and 12 attention heads (110M parameters), and $\text{BERT}_{\text{LARGE}}$ with 24 layers, 1024 hidden units, and 16 attention heads (340M parameters). BERT-based LLM can be fine-tuned with minimal modifications by adding an output layer specific to the task and training the entire model end-to-end, making it suitable for a wide range of natural language processing (NLP) tasks, such as text classification, named entity recognition, and question answering. It utilizes a flexible input representation with special tokens, including `[CLS]` as a classification token and `[SEP]` as a separator for sentence pairs. BERT has demonstrated superior performance across numerous benchmarks, including the GLUE benchmark, SQuAD question answering, and SWAG sentence prediction, achieving notable improvements such as a +7.7% absolute gain on GLUE and +5.1% on SQuAD 2.0. For medical text classification, BERT can be fine-tuned by processing clinical text using its tokenization scheme and adding a classification head to the `[CLS]` token representation, followed by a fully connected layer with softmax or sigmoid activation for multi-class or binary classification tasks. Fine-tuning on domain-specific medical text corpora, coupled with optimized hyperparameters, can leverage BERT's deep bidirectional representations to enhance the handling of complex medical terminology and context, offering significant advantages over traditional methods by reducing the need for task-specific feature engineering and improving generalization through extensive pre-training on diverse corpora.

The different variations of BERT model are BioBERT, ClinicalBERT, SciBERT, RoBERTa, BioRoBERTa, TabNet, TabTransformer, and DistilBERT, as shown in Table II. We compare the performance of these variations to check the role of different pre-fine tuning processes of these models.

- *BioBERT [39]:* BioBERT is a domain-specific adaptation of BERT pre-trained on large-scale biomedical corpora, including PubMed abstracts and PMC full-text articles. It is designed to enhance biomedical natural language processing (BioNLP) tasks such as named entity recognition, relation extraction, and question answering. By leveraging biomedical text data, BioBERT achieves significant improvements over generic BERT models in understanding biomedical terminology and contextual relationships.
- *ClinicalBERT [40]:* ClinicalBERT is a specialized version of BERT fine-tuned on clinical notes from electronic health records (EHRs). It is optimized to capture the unique linguistic characteristics of clinical text, including medical abbreviations, jargon, and domain-specific contextual dependencies. ClinicalBERT is widely used for tasks such as phenotyping, diagnosis prediction, and clinical named entity recognition.
- *SciBERT [41]:* SciBERT is a variant of BERT trained on a corpus of scientific literature, including papers from diverse domains such as biomedicine and computer science. It introduces a domain-specific vocabulary to improve performance on tasks requiring scientific language understanding. SciBERT has shown superior performance in scientific text classification, information retrieval, and entity recognition tasks compared to general-purpose BERT models.
- *RoBERTa [42]:* RoBERTa (Robustly Optimized BERT Approach) is an enhancement of BERT that applies more extensive pre-training with larger datasets, dynamic masking, and optimized hyperparameters. It removes the next sentence prediction objective and focuses solely on the masked language modeling task. RoBERTa consistently outperforms BERT across multiple benchmark tasks due to its improved pre-training methodology.
- *BioRoBERTa [43]:* BioRoBERTa is an adaptation of RoBERTa to the biomedical domain, incorporating biomedical literature and domain-specific text to enhance performance on BioNLP tasks. It retains the advantages of RoBERTa, such as dynamic masking and larger training corpora, while achieving better results on biomedical text mining and classification tasks.

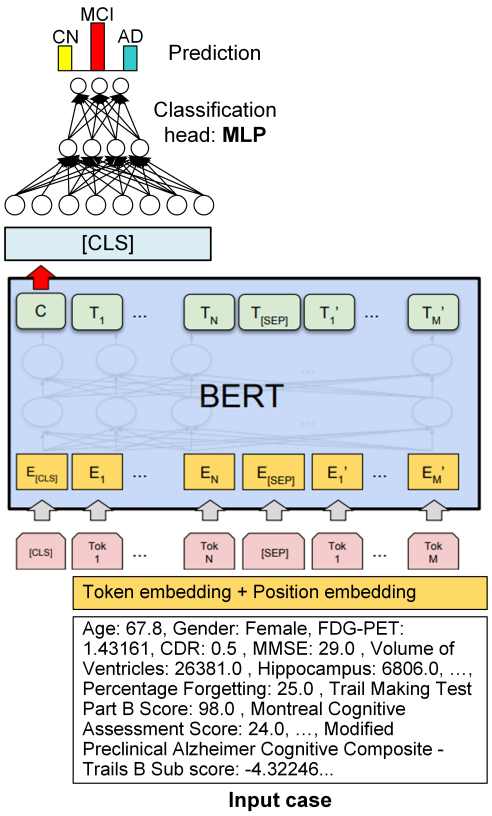

Fig. 2. BERT model. MLP: Multilayer perceprton, [CLS]: Classification token, [SEP]: Separator token, E_i: Embedding of token i.

- *TabNet [44]:* TabNet is a deep learning architecture designed for tabular data that combines decision tree interpretability with the power of deep learning. It uses sequential attention to select relevant features at each decision step, enabling it to outperform traditional machine learning models on structured data while providing interpretability through feature selection at each step.
- *TabTransformer [45]:* TabTransformer applies transformer-based deep learning to tabular data by encoding categorical features using embeddings and self-attention mechanisms. This model leverages the attention mechanism to learn complex relationships within structured data, improving performance in tasks such as fraud detection, customer segmentation, and healthcare analytics.
- *DistilBERT [46]:* DistilBERT is a lightweight version of BERT that reduces the model size while retaining most of its performance. It is trained using knowledge distillation, which transfers knowledge from a larger teacher model to a smaller student model. DistilBERT offers faster inference and lower memory requirements, making it suitable for deployment in resource-constrained environments.
- *XLNet [47]:* XLNet is a language pretraining model that improved upon BERT by using a permutation-based autoregressive approach to capture bidirectional context without relying on masked tokens. This eliminated BERT's pretraining-finetuning gap and better models dependencies between words. XLNet also incorporated Transformer-XL's architectural enhancements, enabling it to handle longer text sequences efficiently. It outperformed BERT on multiple NLP benchmarks, demonstrating superior performance in tasks like question answering and sentiment analysis.

5) LLM prompting: The performance of the BERT-based classifiers is compared with the pre-trained and fine-tuned generative LLM models,particularly GPT-4o [48] and LLaMA-3 [49]. For a fair comparison, these two LLMs are tested using the same test dataset that has been used to test BERT-based and classical machine learning models. Test cases are converted to prompts and the LLM are queried by each case. We test three prompt engineering techniques with LLMs including zero-shot, one-shot, and few-shot prompting. Zero-shot testing assess how well the model can classify without prior examples, one-shot testing evaluate model performance when given a single labeled example before the test case, and few-shot testing observe how the model generalizes with multiple labeled cases. In the following, we provide examples of prompts for the three prompt engineering methods. For space restrictions, the examples are given for the GPT-4o and LLaMA-3 prompts can be formulated in the same manner but using different separators.

GPT-4o zero-shot Prompt

Persona: You are a highly knowledgeable medical AI assistant specializing in Alzheimer's disease research and diagnostics.
Context: The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset contains structured data about patients, including clinical, genetic, and neuroimaging biomarkers. Your task is to classify patients into one of three categories: Normal Cognition (NC), Mild Cognitive Impairment (MCI), or Alzheimer's Disease (AD). The decision should be based on the provided structured data.
Instructions: Use the input data to make an informed classification. If the information is insufficient to classify accurately, state `Insufficient information.`
Input Data:

- Age: 50
- MMSE Score: 29
- APOE4 Gene: 1
- Hippocampal Volume: 7694
- Ventricular Volume: 14149

Task: Classify the patient's condition (NC, MCI, or AD).

GPT-4o one-shot Prompt

Persona: You are a highly knowledgeable medical AI assistant specializing in Alzheimer's disease research and diagnostics.
Context: The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset contains structured data about patients, including clinical, genetic, and neuroimaging biomarkers. Your task is to classify patients into one of three categories: Normal Cognition (NC), Mild Cognitive Impairment (MCI), or Alzheimer's Disease (AD). The decision should be based on the provided structured data.
Example:

```
Input Data:
- Age: 65
- MMSE Score: 29
- APOE4 Gene: 0
- Hippocampal Volume: 7694
- Ventricular Volume: 14149
Classification: NC
```

Instructions: Based on the provided structured data, classify the patient's condition and explain your reasoning.
Input Data:

- Age: 50
- MMSE Score: 29
- APOE4 Gene: 1
- Hippocampal Volume: 7694
- Ventricular Volume: 14149

Task: Classify the patient's condition (NC, MCI, or AD).

GPT-4o few-shot Prompt

Persona: You are a highly knowledgeable medical AI assistant specializing in Alzheimer's disease research and diagnostics.
Context: The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset contains structured data about patients, including clinical, genetic, and neuroimaging biomarkers. Your task is to classify patients into one of three categories: Normal Cognition (NC), Mild Cognitive Impairment (MCI), or Alzheimer's Disease (AD). The decision should be based on the provided structured data.
Examples:

```
1. Input Data:
   - Age: 70
   - MMSE Score: 23
   - APOE4 Gene: 1
   - Hippocampal Volume: 7694
   - Ventricular Volume: 14149
   Classification: MCI
2. Input Data:
   - Age: 80
   - MMSE Score: 15
   - APOE4 Gene: 2
   - Hippocampal Volume: 7694
   - Ventricular Volume: 7694
   Classification: AD
3. Input Data:
   - Age: 60
   - MMSE Score: 20
   - APOE4 Gene: 0
   - Hippocampal Volume: 8694
   - Ventricular Volume: 8694
   Classification: CN
```

Instructions: Based on the provided structured data, classify the patient's condition and explain your reasoning.
Input Data:

- Age: 50
- MMSE Score: 29
- APOE4 Gene: 1
- Hippocampal Volume: 7694
- Ventricular Volume: 14149

Task: Classify the patient's condition (NC, MCI, or AD).

6) Model explainability: Explainable artificial intelligence (XAI) refers to a suite of techniques designed to enhance model transparency and interpretability, model reliability, ethical considerations, and regulatory adherence. In addition, explainability helps identify erroneous predictions, leading to improved model refinement and robustness. In recent years, XAI has gained considerable attention, particularly in the medical domain, where decision making is critical for patient outcomes. By integrating explainability into AI-driven systems, clinicians can gain insight into model predictions, fostering trust, and facilitating regulatory

TABLE II
THE TESTED PRE-TRAINED AND FINE-TUNED LM AND LLM MODELS

Model	parameters	Pre-training data scale	Pre-training dataset	Primary tasks
BERT	110M-340M	3.3B tokens	Books + English Wikipedia	General NLP
RoBERTa	125M-355M	161GB	Extended web corpus	General NLP
BioBERT	110M	18B tokens	PubMed + PMC	Biomedical NLP
ClinicalBERT	110M	112k clinical notes	MIMIC-III	Clinical NLP
SciBERT	110M	3.17B tokens	Scientific Literature	Biomedical NLP
DistilBERT	66M	16GB	Wikipedia, BooksCorpus	Text classification, Text summarization
XLNet	340M	158GB	BooksCorpus, Wikipedia, Giga5, ClueWeb09	Language modeling, Text generation, Text classification
BioRoBERTa	125M	24GB	PubMed abstracts, PMC articles	Biomedical text classification, Named Entity Recognition
LLaMA-3.2-3B-Instruct	3B	Unspecified	Internet-scale datasets	Instruction-following tasks, Text generation, Language understanding
PT-4-turbo	Confidential	Confidential	Proprietary large-scale internet data	Text generation, Chat, Code generation, Multimodal

compliance. A diagnostic tool for AD detection requires rigorous validation to ensure that the model output aligns with clinical reasoning.

XAI techniques can be broadly classified into two categories: local and global explainability. Local explainability methods provide information about individual predictions rather than the overall behavior of the model. These techniques analyze the contribution of each input feature to a specific decision, making them particularly useful in scenarios where justifying a single prediction is necessary. Examples include SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). Global explainability aims to provide a comprehensive understanding of the overall decision-making process of a model by identifying the key patterns and contribution to the features that influence its predictions across the entire dataset. Global XAI techniques include feature importance ranking, decision trees, and rule-based explanations.

In this study, we extend the BERT model to provide XAI for each decision. We use a well-know XAI tool called SHAP to provide local and global XAI features. SHAP is a powerful method derived from cooperative game theory that provides a robust framework for feature attribution. It assigns each feature a contribution score based on its marginal impact on the model's output. Given a predictive model f and an input instance x, the SHAP value for a feature i is computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right], \tag{1}$$

where:

- N is the set of all features,
- S is a subset of features excluding feature i,
- $f(S)$ represents the model prediction using only the features in subset S,
- $f(S \cup \{i\}) - f(S)$ quantifies the marginal contribution of feature i.

This formulation ensures that SHAP values satisfy desirable properties, including efficiency, symmetry, and additivity.

IV. EXPERIMENT SETUP

The hardware configuration of the experimental platform is an Intel i7-6700 CPU, the graphics card is RTX 4090, the memory is 16G, the operating system is Windows 11, and the model is implemented based on the Python programming language and PyTorch frameworks. The data set was divided into 80% training set and 20% test set. Training set is used to train models, and testing set is used to evaluate models. The models are evaluated using accuracy (Acc), precision (Pre), recall (Rec), F1-score (F1).

V. RESULTS AND DISCUSSION

In this section, we discuss the results of AD diagnosis based on different ML algorithms and large language models. The results are reported with the standard deviations to measure model's stability. The standard deviation is calculated by training the models five times with random splits of data and testing it. We report the testing results based on the unseen test sets. Best results in each experiment are highlighted in the table by bold face. In Subsection V-A, we discuss the performance of different classical ML models. Note that these models are tuned and optimized to learn this dataset only, and they will not be able to work with data coming from different distribution. In Subsection V-B, we discuss the performance of the BERT-based models. In this experiment, we explore the capabilities of LLM models to work as classifiers and clinical decision support systems. General purpose LLM models can understand text and make general decisions such as summarization, sentiment analysis, and translations. however, for decisions that need deep knowledge about specific domains, a finetuning stage is needed to tune the model for this specific domain. The resulting model still has the capabilities to understand, but after finetuning it becomes powerful at making deep medical decisions. In Subsection V-C, we discuss the performance of the well-known generative models including LLama-3 and GPT-4o. This experiment explore the level of medial knowledge of the general generative LLM models. Although these models are great at doing different tasks, they lack deep knowledge in sensitive domains such as medicine.

A. Results of Classical Machine Learning

In this experiment, we evaluate the performance of different classical machine learning models such as DT, SVM, RF, LR, NB, and AdaBoost, see Table III. RF model achieved the highest Acc of 91.87%, outperforming all other models confidence interval of [91.61, 92.01]. AdaBoost follows closely with an Acc of 91.41%, demonstrating its ability to generalize well to the test data. DT achieved a slightly lower Acc of 90.80%, indicating its effectiveness in handling structured data. On the lower end, KNN exhibited the weakest accuracy performance at 39.26%, suggesting that it might struggle with high-dimensional structured data. SVM and LR displayed similar accuracies at 57.05% and 57.06%, respectively, while NB achieved a moderate accuracy of 69.94%.

The highest Pre was observed with the AdaBoost model at 91.41%, followed by RF at 91.07%. The Decision Tree

TABLE III
RESULTS OF THE CLASSICAL MACHINE LEARNING MODEL

Models	Accuracy	Precision	Recall	F1-score
DT	90.80±0.101	90.98±0.135	90.98±0.168	90.87±0.202
SVM	57.05±0.103	56.10±0.137	56.10±0.171	56.14±0.205
KNN	39.26±0.105	38.65±0.140	38.65±0.175	38.74±0.210
LR	57.06±0.150	56.91±0.200	56.91±0.250	56.66±0.300
AdaBoost	91.41±0.119	91.41±0.158	91.41±0.198	91.41±0.237
RF	**91.87±0.108**	**91.07±0.144**	**91.07±0.180**	**91.91±0.216**
NB	69.94±0.108	69.43±0.144	69.43±0.180	68.03±0.216

model also performed well with a Pre of 90.98%, reflecting its capability to make reliable positive predictions. The lowest Pre value was obtained by KNN at 38.65%, indicating a high rate of false positives. SVM and LR models demonstrated Pre values of 56.1% and 56.91%, respectively, showing their relatively moderate discriminative power. RF emerged as the best model in terms of Rec, achieving 91.87%, ensuring a higher rate of correct positive predictions. AdaBoost followed closely with a recall of 91.41%, indicating its strong sensitivity. The Decision Tree model had a Rec of 90.80%, which is consistent with its accuracy. KNN once again showed the weakest Rec at 39.26%, which indicates difficulties in identifying true positives. SVM and LR displayed recall values of 57.55% and 57.06%, respectively, with NB achieving a Rec of 69.94%.

RF achieved the highest F1 of 91.91%, demonstrating a well-balanced trade-off between Pre and Rec. AdaBoost followed with an F1 of 91.41%, further confirming its reliability. DT achieved an F1 of 90.87%, supporting its consistency across metrics. KNN had the lowest F1 of 38.74%, highlighting its inefficiency in handling the complexity of the data. NB obtained an F1 of 68. 03%, outperforming SVM (56. 144%) and LR (56. 66%). In conclusion, RF stands out as the best performing model in all metrics, i.e., Acc (91. 87%), Rec (91. 87%) and F1 (91. 91%), indicating its superior ability to effectively classify cases of AD from structured data. The findings from this study highlight that ensemble-based models, particularly Random Forest and AdaBoost, outperform other classical machine learning algorithms in diagnosing AD using structured data.

Despite the promising results achieved by classical machine learning models, they still face limitations in capturing complex, non-linear interactions between features. BERT-based models, which have demonstrated their effectiveness in processing textual and structured data, present an opportunity to further improve diagnostic accuracy. Using their deep contextual understanding and ability to represent structured data in a meaningful way, BERT-based models can potentially improve feature interactions, handle missing values effectively, and provide interpretable insights into AD progression. Future work should focus on fine-tuning domain-specific variants such as BioBERT and ClinicalBERT to explore their potential to outperform traditional ML models and improve clinical decision support systems.

B. Results of BERT-Based Language Models

The evaluation of BERT-based language models, both with (Table IV) and without (Table V) fine-tuning, provides valuable insights into their effectiveness compared to classical machine learning models for AD diagnosis. Without fine tuning, the BERT models struggled to match the performance of classical models such as RF (91. 87% Acc, 91. 07% Pre, 91. 87% Rec and 91. 91% F1) and AdaBoost (91.41% Acc, Pre, Rec, and F1). For instance, the general BERT model without fine-tuning achieved an Acc of only 37.423%, Pre of 40.014%, Rec of 37.423%, and an F1 of 21.096%, highlighting its initial inability to effectively process structured data. Similarly, BioBERT without fine-tuning achieved 37.423% Acc, 50.443% Pre, 37.423% Rec, and an F1 of 21.104%. The best performing model was XLNet with 71.77% Acc, 72.14% Pre, 72.14% Rec, and 71.33% F1. In addition, XLNet have the best confidence interval of [70.94, 71.52]. In contrast, fine-tuning significantly improved performance, with BioBERT achieving the highest Acc of 94.479%, Pre of 94.637%, Rec of 94.479%, and an F1 of 94.517%, surpassing all classical models.

The lower performance of non-fine-tuned BERT models can be attributed to their inability to process structured data directly, lack of domain-specific knowledge, and the risk of overfitting due to high model complexity. Unlike traditional models, BERT requires structured data to be serialized into textual form, which can introduce noise and reduce accuracy if not carefully managed. Additionally, without fine tuning on relevant biomedical data, general BERT models fail to capture meaningful feature relationships, leading to suboptimal performance. For example, SciBERT without fine-tuning achieved an Acc of 50.920%, Pre of 55.094%, Rec of 50.920%, and an F1 of 46.922%, whereas with fine-tuning, it reached 70.552%, 70.523%, 70.552%, and 68.185% in the same metrics.

Despite these challenges, fine-tuned BERT models offer significant advantages over classical ML methods. They provide a deeper contextual feature representation, leverage transfer learning capabilities, and offer scalability across different datasets. Fine-tuned versions such as BioBERT and ClinicalBERT demonstrate their potential by capturing intricate medical relationships that traditional models might overlook. ClinicalBERT achieved an Acc of 93.252%, Pre of 93.512%, Rec of 93.252%, and an F1 of 93.309%, making it a strong contender for structured data applications.

Among the fine-tuned BERT models, BioBERT stands out as the best-performing model, achieving an Acc of 94.479% with superior Pre (94.637%), Rec (94.479%), and F1 (94.517%), with confidence interval of [94.01, 94.83]. ClinicalBERT follows closely, further highlighting the effectiveness of domain-specific pre-training. DistilBERT, despite its lighter architecture, improved from a non-fine-tuned Acc of 36.81% to 44.785% after fine-tuning, showcasing the impact of domain adaptation. In conclusion, while classical machine learning models provide a strong baseline for AD, fine-tuned BERT-based models, particularly BioBERT, offer superior performance by effectively leveraging biomedical knowledge.

C. Results of Generative Models

The evaluation of LLMs, specifically LLama-3 and GPT-4o, across different prompt engineering techniques (zero-shot, one-shot, and few-shot) provided further insights. In the zero-shot setting, GPT-4o significantly outperformed LLama, achieving an accuracy of 53.988%, precision of 75.239%, recall of 53.988%, and an F1-score of 44.825%,

TABLE IV
RESULTS OF NOT FINE-TUNED BERT-BASED LANGUAGE MODELS

Models	Accuracy	Precision	Recall	F1-score
BERT	37.42±0.112	40.01±0.150	40.01±0.187	21.09±0.187
BioBERT	37.42±0.108	50.44±0.144	50.44±0.180	21.10±0.180
ClinicalBERT	39.87±0.105	40.45±0.140	40.45±0.175	25.94±0.175
SciBERT	50.92±0.101	55.09±0.135	55.09±0.168	46.92±0.202
RoBERTa	57.66±0.212	47.99±0.283	47.99±0.354	49.83±0.424
BioRoBERTa	36.81±0.108	13.55±0.144	13.55±0.180	19.80±0.216
DistilBERT	36.81±0.112	13.55±0.150	13.55±0.187	19.80±0.224
XLNet	**71.77±0.119**	**72.14±0.158**	**72.14±0.198**	**71.33±0.237**

TABLE V
RESULTS OF FINE-TUNED BERT-BASED LANGUAGE MODELS

Models	Accuracy	Precision	Recall	F1-score
BERT	36.81±0.150	13.55±0.200	13.55±0.250	19.80±0.300
BioBERT	**94.47±0.129**	**94.63±0.172**	**94.63±0.215**	**94.51±0.258**
ClinicalBERT	93.25±0.101	93.51±0.135	93.51±0.168	93.30±0.202
SciBERT	70.55±0.103	70.52±0.137	70.52±0.171	68.18±0.205
RoBERTa	36.81±0.105	13.55±0.140	13.55±0.175	19.80±0.210
BioRoBERTa	87.11±0.108	87.07±0.144	87.07±0.180	86.90±0.216
DistilBERT	44.78±0.112	27.57±0.150	27.57±0.187	32.27±0.224
XLNet	88.95±0.119	89.36±0.158	89.36±0.198	89.037±0.198

whereas LLama only managed 26.38% accuracy, 6.959% precision, 26.38% recall, and an F1-score of 11.013%. However, in the one-shot scenario, GPT's performance dropped to 31.902% accuracy, 25.61% precision, 31.902% recall, and 28.131% F1-score, while LLama maintained the same performance across different shot settings. Interestingly, in the few-shot setup, LLama showed improvement, achieving an accuracy of 28.834%, precision of 80.754%, recall of 28.834%, and an F1-score of 15.981%, while GPT's performance dropped further to 22.086% accuracy, 29.108% precision, 22.086% recall, and 21.310% F1-score.

The comparative analysis of fine-tuned BERT-based models and Large Language Models (LLMs) reveals notable performance differences, particularly in the context of Alzheimer's disease diagnosis using structured data. Fine-tuned BERT-based models, such as BioBERT and ClinicalBERT, demonstrated superior accuracy, precision, recall, and F1-scores, with BioBERT achieving an impressive accuracy of 94.479%. In contrast, LLMs such as GPT and LLama exhibited significantly lower performance, with the highest accuracy recorded at 53.988% for GPT in a zero-shot setting. The primary reasons for the suboptimal performance of LLMs can be attributed to their general-purpose nature and lack of domain-specific fine-tuning. Unlike BERT-based models that have been pre-trained and fine-tuned on biomedical corpora, LLMs rely on broader knowledge representations that may not effectively capture the intricate relationships present within structured clinical data. Furthermore, the reliance of LLMs on prompt engineering techniques, such as zero-shot and few-shot learning, often results in inconsistent performance due to their sensitivity to prompt phrasing and context variations. Classical machine learning models, on the other hand, leverage well-defined feature engineering approaches tailored specifically to structured data, leading to relatively stable performance compared to the generalist approach of LLMs. These findings highlight the necessity of domain-specific adaptations and structured data compatibility for achieving optimal diagnostic accuracy in Alzheimer's disease classification tasks.

TABLE VI
RESULTS OF LARGE LANGUAGE MODELS

Shot types	Models	Accuracy	Precision	Recall	F1-score
Zero-short prompting	LLama	26.38±0.108	06.95±0.144	06.95±0.18	11.01±0.216
	GPT	**53.98±0.150**	**75.23±0.200**	**75.23±0.250**	**44.82±0.300**
One-shot prompting	LLama	26.38±0.129	06.95±0.172	06.95±0.215	11.01±0.258
	GPT	31.90±0.112	25.61±0.150	25.61±0.187	28.13±0.224
Few-shot prompting	LLama	28.83±0.119	80.75±0.158	80.75±0.198	15.98±0.237
	GPT	22.08±0.103	29.10±0.137	29.10±0.171	21.31±0.205

D. Model Explainability

After training and validating a robust AD classifier, we extend the proposed architecture to provide explainability to its decisions. Explainable artificial intelligence is critical to make the model trustworthy and acceptable in a real medical environment [50]. Fig. 3 showed the global feature importance based on SHAP. The importance scores quantified the contribution of each feature to the model's decisions. The results showed that mPACCdigit, CDRSB, and mPACCtrailsB were the most influential features, exhibiting the highest SHAP values. Medically, these features played a crucial role in capturing cognitive impairment and disease progression. Other important features included LDELTOTAL, EcogPtPlan, and MMSE, which are commonly utilized in neuropsychological assessments for AD. Moreover, biomarkers such as ADAS11 and ADAS13 demonstrated moderate contributions, while demographic and neuroimaging variables, including marital status, ventricles, hippocampus volume, and entorhinal cortex thickness, contributed to a lesser extent. Features related to genetic markers, such as APOE4, and cerebrospinal fluid biomarkers, such as PTAU and TAU, had lower impact scores in this specific model. In summary, this global feature importance suggested that cognitive test scores and functional assessments were more influential in model decisions than demographic, genetic, and neuroimaging variables.

Next, we explore the local feature importance based on text data. We tested the explainability features of SHAP with Clinical BERT as a classifier by using the test set to measure XAI stability. For example, Fig. 4 showed an XAI for an AD case. The SHAP visualization demonstrated how specific features influenced the prediction, with red bars pushing the decision towards AD classification and blue bars opposing it. From the input features, variables such as mPACCdigit (-17.7) and mPACCtrailsB (-14.9) had the most significant negative SHAP values, strongly driving the prediction towards AD classification. Similarly, RAVLT forgetting and RAVLT percent forgetting showed significant positive contributions, reinforcing their importance in Alzheimer's disease progression. These cognitive assessment features played a dominant role in the decision-making process, aligning with the global SHAP analysis in the previous figure, where mPACCdigit, CDRSB, and mPACCtrailsB were among the most important global predictors. The features of MMSE, ADAS11, ADAS13, and LDELTOTAL, which were moderately ranked in the global SHAP feature importance analysis, also appeared in this local explanation, further confirming their relevance in distinguishing AD from MCI or CN. Neuroimaging biomarkers, such as hippocampus volume and ventricles, had a relatively lower impact on this specific prediction, which is also consistent with their lower ranking in the global feature importance chart. As it can be noticed, there is a consistency between global and local XAI methods.

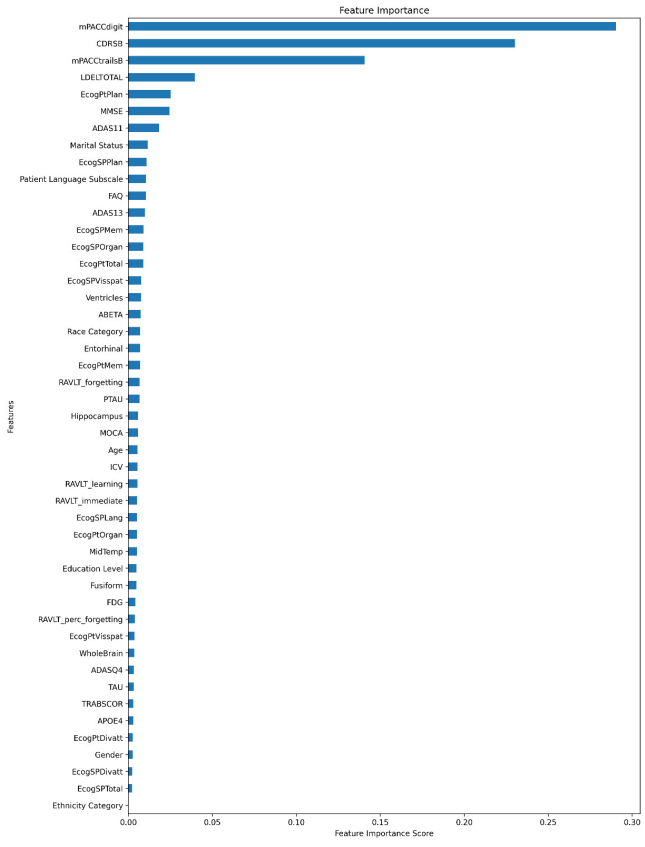

Fig. 3. Global feature importance.

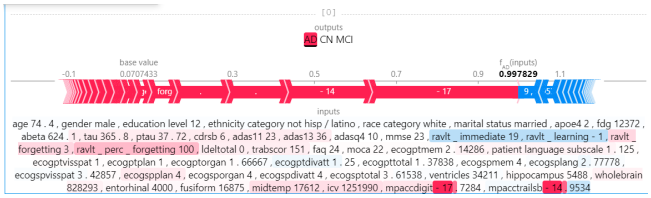

Fig. 4. Local XAI for an AD case.

Fig. 5 illustrated the local SHAP explanation for an MCI. A key observation in this case is the substantial positive contribution of RAVLT percent forgetting (100), which strongly supported the classification of MCI rather than AD. This aligns with clinical findings, as RAVLT (Rey Auditory Verbal Learning Test) scores reflect memory performance and are widely used in cognitive assessments for early AD. Conversely, features such as mPACCdigit (-6.85) and mPACCtrailsB (-6.9) contributed negatively, indicating that these digital cognitive assessments played a role in reducing the likelihood of an MCI classification, aligning with their high importance in global SHAP analysis. The CDRSB score (1.5) also positively influenced the prediction toward AD, reinforcing its role as a key feature in disease progression assessment. Other cognitive test scores, such as MMSE, ADAS11, ADAS13, and LDELTOTAL, appeared in the feature set, consistent with the global SHAP importance ranking. Moreover, neuroimaging biomarkers (e.g., hippocampus volume and ventricles) contributed less significantly, in line with the global feature importance results, where these features were of moderate influence. By comparing the local and global SHAP explanation, we notice that model's interpretability and consistency are

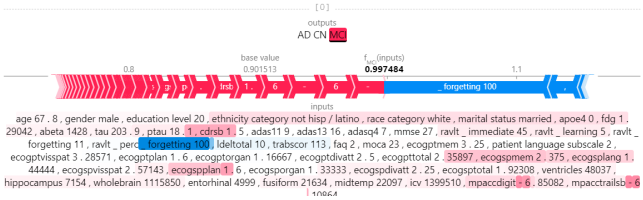

Fig. 5. Local XAI for an MCI case.

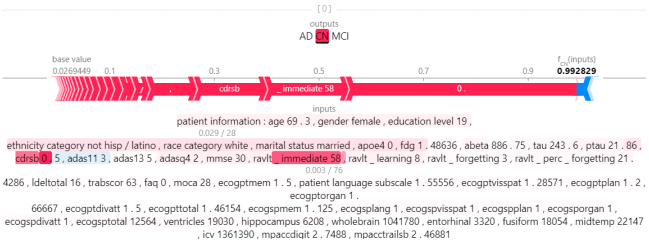

Fig. 6. Local XAI for an CN case.

preserved. In other words, the same high-importance features identified globally (mPACCdigit, CDRSB, and RAVLT forgetting scores) are also observed to be crucial in this specific prediction, increasing trust in the model's decision-making process.

Fig. 6 presented a local SHAP explanation for an CN case. The contribution of the features showed that CDRSB (0.5) and RAVLT immediate recall (58) emerged as key contributors. These cognitive assessment scores, widely used in neuropsychological evaluations, played a central role in differentiating normal cognitive aging from pathological decline. The presence of ADAS11, ADAS13, and MMSE scores further reinforced their importance in the diagnostic process, aligning with the global SHAP feature importance analysis, where these features consistently ranked among the most influential predictors. In addition, biological markers such as FDG metabolism, tau (243.6), and pTau (21.86) were observed in the feature set, though their impact on this specific classification was comparatively lower. Structural neuroimaging features, including hippocampus volume and ventricle size, contributed minimally, consistent with their moderate importance in the global ranking.

The local and global SHAP importance analyses demonstrated that the model consistently prioritized medically relevant features in its decision-making process. The global feature importance rankings highlighted cognitive assessment scores, such as mPACCdigit, CDRSB, and RAVLT forgetting scores, as the most influential predictors, reflecting well-established clinical markers of cognitive decline. The local SHAP explanations further validated this finding by showing that these same features played a dominant role in individual predictions, whether distinguishing CN, MCI, or AD cases. The interpretation of neuropsychological assessments and clinical biomarkers indicated the model captured meaningful patterns associated with AD diagnosis. This XAI is critical to increase the trust of medical professionals because it provides transparent insights into how the model arrived at its predictions.

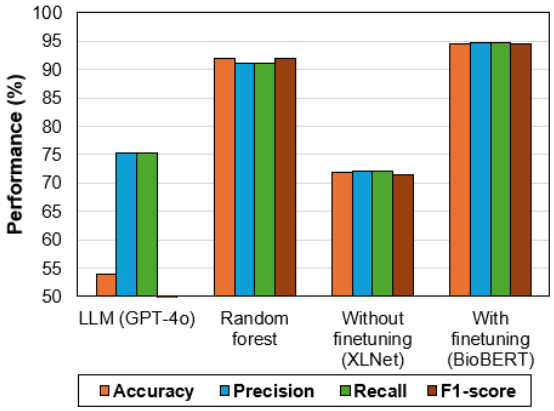

Fig. 7. Comparison of different AD classification models.

VI. Discussion

The comparative analysis of AD classification models in this study highlights the efficacy of different ML paradigms, including classical models, fine-tuned BERT-based LLM models, and generative LLMs models. The results suggest that fine-tuned domain-specific language models, such as BioBERT and ClinicalBERT, significantly outperform classical approaches and general-purpose LLMs in structured AD diagnosis, as shown in Fig. 7. This discussion provides deeper insight into these findings, explores their implications, and highlights their contributions to the field of computational AD diagnostics.

a) Performance of classical machine learning models: Classical ML models, particularly ensemble-based methods such as RF and AdaBoost achieved high classification performance, with accuracy values of 91.87% and 91.41%, respectively. These models demonstrated robustness in handling structured data, benefiting from explicit feature engineering and well-optimized hyperparameters. While RF and AdaBoost provided a strong baseline, their limited capacity to capture intricate feature dependencies constrained their performance. Unlike deep learning-based approaches, classical ML models rely heavily on predefined features and lack the ability to dynamically learn feature interactions from large datasets. This finding reinforces the necessity of more sophisticated models that integrate both structured and unstructured medical data for improved diagnostic precision.

b) Superiority of finetuned BERT-based models: Among all evaluated models, finetuned BERT-based models exhibited the highest classification performance, with BioBERT achieving 94.48% accuracy, 94.64% precision, 94.48% recall, and a 94.52% F1-score. ClinicalBERT followed closely, demonstrating similar performance metrics. These results indicate that transformer-based models fine-tuned on biomedical and clinical datasets can effectively capture domain-specific patterns in structured AD diagnosis. The primary advantage of these models lies in their ability to leverage transfer learning from extensive biomedical corpora. Finetuning allows these models to understand the deep relationships among structured features, improving their ability to differentiate between NC, MCI, and AD. The superior performance of BioBERT and ClinicalBERT underscores the importance of domain-specific pre-training in medical AI applications, particularly in complex diseases like AD.

c) Limitations of general-purpose large language models: In contrast to fine-tuned BERT models, general-purpose generative LLMs models, including GPT-4o and LLama-3, exhibited suboptimal performance, with accuracy values peaking at 53.99% in the zero-shot setting. The notable decline in classification accuracy in one-shot (31.90%) and few-shot (22.09%) prompting scenarios highlights the sensitivity of LLMs to prompt engineering. The primary limitation of these LLMs in AD diagnosis stems from their lack of structured data compatibility and lack of deep knowledge in AD domain. Unlike BERT-based models that are finetuned for biomedical tasks, LLMs rely heavily on natural language processing capabilities and struggle to interpret structured and medical numerical data effectively. The results reinforce the necessity of domain adaptation for LLMs, as general-purpose training alone is insufficient for complex clinical classification tasks. Additionally, prompt engineering approaches did not yield substantial improvements in LLM performance. While zero-shot classification showed relatively better accuracy, the inconsistency observed across one-shot and few-shot settings suggests that LLMs require more sophisticated adaptation techniques, such as fine-tuning or structured data serialization, to achieve clinically viable accuracy levels. The superior performance of fine-tuned BERT models demonstrates their applicability in CDSS, where accurate classification of AD stages is crucial for early intervention and treatment planning. Moreover, the study highlights the need for structured data adaptation techniques in LLMs. In summary, this study provided an empirical validation of fine-tuned BERT models which demonstrates that domain-specific finetuning significantly enhances structured data classification, outperforming both classical ML models and general-purpose generative LLMs.

VII. Study Limitations and Future Work

This study presents a significant advancement in AD diagnosis by demonstrating the superiority of fine-tuned BERT-based models, particularly BioBERT and ClinicalBERT, over classical machine learning approaches and general-purpose generative LLM models. The study makes notable contributions to the field by (1) developing a clinical decision support system (CDSS) tailored for structured AD data, (2) showcasing the efficacy of domain-specific language models in clinical diagnostics, (3) establishing an innovative serialization technique to adapt structured data for language models, and (4) providing an extensive comparative analysis between traditional, transformer-based, and generative AI models. The findings underscore the pivotal role of fine-tuning biomedical transformers to extract meaningful patterns from structured clinical data, thereby offering a robust and interpretable AI-driven diagnostic tool. Despite these contributions, some limitations must be acknowledged. First, while the fine-tuned BERT models exhibit high classification accuracy, concerns regarding the trustworthiness of AI systems remain. Potential biases in the dataset, adversarial vulnerabilities, and model uncertainty must be rigorously evaluated. Future research should incorporate fairness-aware training methods, adversarial testing, and uncertainty quantification to enhance the model's reliability in diverse clinical settings. Second, the study primarily relies on internal validation using the ADNI dataset, raising

concerns about model stability and external validity. To ensure generalizability, external validation on independent datasets, such as the National Alzheimer's Coordinating Center (NACC) and Open Access Series of Imaging Studies (OASIS), is essential. Such validation will assess the model's robustness across different demographic and clinical populations, mitigating overfitting risks associated with single-dataset training. Third, while the current approach leverages structured numerical data, multimodal learning could further enhance diagnostic accuracy. Alzheimer's disease diagnosis benefits from a combination of biomarkers, imaging modalities (e.g., MRI, PET scans), clinical text (e.g., physician notes), and structured tabular data. Future studies should explore the integration of multimodal deep learning frameworks, leveraging fusion architectures that combine textual, visual, and numerical features for a more comprehensive diagnosis.

Despite these limitations, the study demonstrates promising results, with fine-tuned BERT models achieving state-of-the-art performance in structured AD classification. The proposed approach holds great potential for real-world clinical application, facilitating early detection and personalized intervention. By integrating domain-specific NLP techniques with structured clinical data, this research paves the way for more precise, scalable, and interpretable AI-driven diagnostic tools in neurodegenerative disease management. In this study we proposed an end-to-end pipeline for finetuning an BERT-based models to predict AD. However, nothing special about AD. The model can be applied to predict other diseases such as Parkinson, diabetes, to name a few. In addition, our architecture can be extended to enhance clinical decision support systems, aiding healthcare professionals in diagnosing diseases, recommending treatments, predicting outcomes, and monitoring patient progress. In radiology, AI algorithms can assist in interpreting medical images, leading to more accurate and efficient diagnoses. In psychiatry, AI has the potential to predict treatment outcomes and support mental health interventions. Additionally, in primary care, AI can support decision-making, predictive modeling, and business analytics. By leveraging AI's capabilities, our model can contribute to improved patient outcomes and operational efficiencies across diverse healthcare settings.

VIII. Conclusion

This study emphasized the importance of model selection and fine-tuning in improving AD diagnosis using structured data. Fine-tuned BERT-based models, particularly BioBERT, achieved superior performance with an accuracy of 94.48%, surpassing classical models such as random forest, which obtained 91.87%. In contrast, LLMs such as GPT and LLama demonstrated lower accuracy, with GPT reaching a maximum of 53.99% in the zero-shot setting, highlighting their limitations in handling structured clinical data without specialized training. The enhanced performance of fine-tuned BERT models resulted from their ability to capture domain-specific nuances through targeted training, aligning with existing literature that supports the efficacy of specialized models in medical diagnostics. However, challenges persist in processing structured clinical data using general-purpose LLMs due to their lack of domain-specific adaptations. Future research will explore the generalizability of the proposed architecture by externally validating it using other datasets such as NACC and OASIS. We will explore also the Retrieval-Augmented Generation (RAG) to leverage external knowledge for improving the performance of LLMs, multimodal learning to integrate imaging and clinical data, and further fine-tuning of LLMs with specialized datasets such as ADNI and NACC to enhance their applicability. Additionally, developing explainable AI models will be essential for ensuring transparency and trustworthiness in clinical decision-making.

Ethics Statement

Data used in this study were obtained from the ADNI (http://adni.loni.usc.edu/). The Alzheimer's Disease Neuroimaging Initiative Data and Publications Committee (ADNI DPC) coordinates patient enrollment and ensures standard practice on the uses and distribution of the data as follows: The ADNI data were previously collected across 50 research sites. To participate in the study, each study subject gave written informed consent at the time of enrollment for imaging and genetic sample collection and completed questionnaires approved by each participating sites' Institutional Review Board (IRB). All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. A complete description of ADNI and up-to-date information is available at http://adni.loni.usc.edu/ and data access requests are to be sent to http://adni.loni.usc.edu/data-samples/access-data/. Detailed inclusion criteria for the diagnostic categories can be found at the ADNI website (http://adni.loni.usc.edu/methods).

Data Availability

The data that support the findings of this study are openly available at the ADNI web site (http://adni.loni.usc.edu/).

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization H.S and S.E.-S; Data curation, H.S and S.E.-S; Formal analysis, H.S and S.E.-S; Funding acquisition, M.M. and J.G.B, Investigation, H.S and S.E.-S; Methodology, H.S and S.E.-S.; Project administration, M.M. and J.G.B; Resources, H.S, S.E.-S, M.M. and J.G.B; Software, H.S.; Writing—original draft, H.S, S.E.-S, M.M. and J.G.B.; Writing—review and editing, H.S, S.E.-S, M.M. and J.G.B. All authors have read and agreed to the published version of the manuscript.

Acknowledgment

This publication has emanated from research conducted with the financial support of Taighde Éireann - Research Ireland under Grant Numbers 12/RC/2289_P2 (Insight) and 21/FFP-A/9174 (SustAIn). For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. *Data used in preparation of this article

were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this publication. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

REFERENCES

[1] S. Chen, Z. Cao, A. Nandi, N. Counts, L. Jiao, K. Prettner, M. Kuhn, B. Seligman, D. Tortorice, D. Vigo *et al.*, "The global macroeconomic burden of alzheimer's disease and other dementias: estimates and projections for 152 countries or territories," *The Lancet Global Health*, vol. 12, no. 9, pp. e1534–e1543, 2024.

[2] N. Bogdanovic, "The challenges of diagnosis in alzheimer's disease," *Eur Neurol Rev*, vol. 14, pp. 15–6, 2018.

[3] S. El-Sappagh, J. M. Alonso-Moral, T. Abuhmed, F. Ali, and A. Bugarín-Diz, "Trustworthy artificial intelligence in alzheimer's disease: state of the art, opportunities, and challenges," *Artificial Intelligence Review*, vol. 56, no. 10, pp. 11 149–11 296, 2023.

[4] S. El-Sappagh, T. Abuhmed, S. R. Islam, and K. S. Kwak, "Multimodal multitask deep learning model for alzheimer's disease progression detection based on time series data," *Neurocomputing*, vol. 412, pp. 197–215, 2020.

[5] S. Shah and M. Shah, "The effects of machine learning algorithms in magnetic resonance imaging (mri), and biomarkers on early detection of alzheimer's disease," *Advances in Biomarker Sciences and Technology*, 2024.

[6] Z. Rezaie and Y. Banad, "Machine learning applications in alzheimer's disease research: a comprehensive analysis of data sources, methodologies, and insights," *International Journal of Data Science and Analytics*, pp. 1–35, 2024.

[7] M. G. Alsubaie, S. Luo, and K. Shaukat, "Alzheimer's disease detection using deep learning on neuroimaging: a systematic review," *Machine Learning and Knowledge Extraction*, vol. 6, no. 1, pp. 464–505, 2024.

[8] S. G. Singh, D. Das, U. Barman, and M. J. Saikia, "Early alzheimer's disease detection: A review of machine learning techniques for forecasting transition from mild cognitive impairment," *Diagnostics*, vol. 14, no. 16, p. 1759, 2024.

[9] S. Song, T. Li, W. Lin, R. Liu, and Y. Zhang, "Application of artificial intelligence in alzheimer's disease: a bibliometric analysis," *Frontiers in Neuroscience*, vol. 19, p. 1511350, 2025.

[10] H. Saleh, N. El-Rashidy, T. Abuhmed, and S. El-Sappagh, "Lstm deep learning model for alzheimer's disease prediction based on cost-effective time series cognitive scores," in *2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*. IEEE, 2023, pp. 1–6.

[11] S. El-Sappagh, J. M. Alonso, S. R. Islam, A. M. Sultan, and K. S. Kwak, "A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease," *Scientific reports*, vol. 11, no. 1, p. 2660, 2021.

[12] A. AlMohimeed, R. M. Saad, S. Mostafa, N. M. El-Rashidy, S. Farrag, A. Gaballah, M. Abd Elaziz, S. El-Sappagh, and H. Saleh, "Explainable artificial intelligence of multi-level stacking ensemble for detection of alzheimer's disease based on particle swarm optimization and the sub-scores of cognitive biomarkers," *IEEE Access*, vol. 11, pp. 123 173–123 193, 2023.

[13] K. Patel, C. Mistry, D. Mehta, U. Thakker, S. Tanwar, R. Gupta, and N. Kumar, "A survey on artificial intelligence techniques for chronic diseases: open issues and challenges," *Artificial Intelligence Review*, pp. 1–54, 2022.

[14] J. Wang, J. X. Huang, X. Tu, J. Wang, A. J. Huang, M. T. R. Laskar, and A. Bhuiyan, "Utilizing bert for information retrieval: Survey, applications, resources, and challenges," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–33, 2024.

[15] K. Zhang, X. Meng, X. Yan, J. Ji, J. Liu, H. Xu, H. Zhang, D. Liu, J. Wang, X. Wang *et al.*, "Revolutionizing health care: The transformative impact of large language models in medicine," *Journal of Medical Internet Research*, vol. 27, p. e59069, 2025.

[16] T. Zheng, X. Xie, X. Peng, H. Chen, and F. Tian, "Alzheimer's disease detection based on large language model prompt engineering," *arXiv preprint arXiv:2501.00861*, 2025.

[17] C. Mao, J. Xu, L. Rasmussen, Y. Li, P. Adekkanattu, J. Pacheco, B. Bonakdarpour, R. Vassar, L. Shen, G. Jiang *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. 104442, 2023.

[18] C.-H. Lin, K.-C. Hsu, C.-K. Liang, T.-H. Lee, C.-W. Liou, J.-D. Lee, T.-I. Peng, C.-S. Shih, and Y. C. Fann, "A disease-specific language representation model for cerebrovascular disease research," *Computer methods and programs in biomedicine*, vol. 211, p. 106446, 2021.

[19] N. Alexander, D. C. Alexander, F. Barkhof, and S. Denaxas, "Identifying and evaluating clinical subtypes of alzheimer's disease in care electronic health records using unsupervised machine learning," *BMC Medical Informatics and Decision Making*, vol. 21, pp. 1–13, 2021.

[20] Q. Li, X. Yang, J. Xu, Y. Guo, X. He, H. Hu, T. Lyu, D. Marra, A. Miller, G. Smith *et al.*, "Early prediction of alzheimer's disease and related dementias using real-world electronic health records," *Alzheimer's & Dementia*, vol. 19, no. 8, pp. 3506–3518, 2023.

[21] X. R. Gao, M. Chiariglione, K. Qin, K. Nuytemans, D. W. Scharre, Y.-J. Li, and E. R. Martin, "Explainable machine learning aggregates polygenic risk scores and electronic health records for alzheimer's disease prediction," *Scientific reports*, vol. 13, no. 1, p. 450, 2023.

[22] A. S. Tang, K. P. Rankin, G. Cerono, S. Miramontes, H. Mills, J. Roger, B. Zeng, C. Nelson, K. Soman, S. Woldemariam *et al.*, "Leveraging electronic health records and knowledge networks for alzheimer's disease prediction and sex-specific biological insights," *Nature Aging*, vol. 4, no. 3, pp. 379–395, 2024.

[23] S. Kumar, I. Oh, S. Schindler, A. M. Lai, P. R. Payne, and A. Gupta, "Machine learning for modeling the progression of alzheimer disease dementia using clinical data: a systematic literature review," *JAMIA open*, vol. 4, no. 3, p. ooab052, 2021.

[24] Y. Wang, J. Deng, T. Wang, B. Zheng, S. Hu, X. Liu, and H. Meng, "Exploiting prompt learning with pre-trained language models for alzheimer's disease detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] T. Zheng, X. Xie, X. Peng, H. Chen, and F. Tian, "Alzheimer's disease detection based on large language model prompt engineering," in *International Conference on Social Robotics*. Springer, 2024, pp. 207–216.

[26] T. Mo, J. C. Lam, V. O. Li, and L. Y. Cheung, "Leveraging large language models for identifying interpretable linguistic markers and enhancing alzheimer's disease diagnostics," *medRxiv*, pp. 2024–08, 2024.

[27] A. Shakeri and M. Farmanbar, "Natural language processing in alzheimer's disease research: Systematic review of methods, data, and efficacy," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 17, no. 1, p. e70082, 2025.

[28] R. Shankar, A. Bundele, and A. Mukhopadhyay, "A systematic review of natural language processing techniques for early detection of cognitive impairment," *Mayo Clinic Proceedings: Digital Health*, p. 100205, 2025.

[29] L. Wang, J. Laurentiev, J. Yang, Y.-C. Lo, R. E. Amariglio, D. Blacker, R. A. Sperling, G. A. Marshall, and L. Zhou, "Development and validation of a deep learning model for earlier detection of cognitive decline from clinical notes in electronic health records," *JAMA network open*, vol. 4, no. 11, pp. e2 135 174–e2 135 174, 2021.

[30] X. Du, J. Novoa-Laurentiev, J. M. Plasek, Y.-W. Chuang, L. Wang, G. A. Marshall, S. K. Mueller, F. Chang, S. Datta, H. Paek *et al.*, "Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes," *EBioMedicine*, vol. 109, 2024.

[31] Q. Chen and Y. Hong, "Alifuse: Aligning and fusing multimodal medical data for computer-aided diagnosis," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 3082–3087.

[32] L. Liu, S. Liu, L. Zhang, X. V. To, F. Nasrallah, and S. S. Chandra, "Cascaded multi-modal mixing transformers for alzheimer's disease classification with incomplete data," *NeuroImage*, vol. 277, p. 120267, 2023.

[33] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," *Transactions of the association for computational linguistics*, vol. 8, pp. 842–866, 2021.

[34] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. Jack Jr, W. J. Jagust, L. M. Shaw, A. W. Toga *et al.*, "Alzheimer's disease neuroimaging initiative (adni) clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.

[35] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, "Tabllm: Few-shot classification of tabular data with large language models," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 5549–5581.

[36] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, no. 2. Minneapolis, Minnesota, 2019.

[37] R. Vavekanand, P. Karttunen, Y. Xu, S. Milani, and H. Li, "Large language models in healthcare decision support: a review," 2024.

[38] M. V. Koroteev, "Bert: a review of applications in natural language processing and understanding," *arXiv preprint arXiv:2103.11943*, 2021.

[39] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

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

[41] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[42] S. Chopra, P. Agarwal, J. Ahmed, S. S. Biswas, and A. J. Obaid, "Roberta and bert: Revolutionizing mental healthcare through natural language," *SN Computer Science*, vol. 5, no. 7, p. 889, 2024.

[43] A. Varela-Vega, A.-B. Posada-Reyes, and C.-F. Méndez-Cruz, "Fine-tuning bert models to extract transcriptional regulatory interactions of bacteria from biomedical literature," *bioRxiv*, pp. 2024–02, 2024.

[44] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 8, 2021, pp. 6679–6687.

[45] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "Tabtransformer: Tabular data modeling using contextual embeddings," *arXiv preprint arXiv:2012.06678*, 2020.

[46] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arxiv 2019," *arXiv preprint arXiv:1910.01108*, 2019.

[47] Z. Yang, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.

[48] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[49] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[50] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," *Information fusion*, vol. 99, p. 101805, 2023.