Enhancing Alzheimer's Detection: Leveraging ADNI Data and Large Language Models for High-Accuracy Diagnosis

Hassan Almalki[1], Alaa O. Khadidos[2], Nawaf Alhebaishi[3]
Department of Information Technology, College of Technology for Communications and Information, Jeddah, Saudi Arabia[1]
Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia[2, 3]
Center of Research Excellence in Artificial Intelligence and Data Science, King Abdulaziz University, Jeddah, Saudi Arabia[2]

***Abstract*—Alzheimer's disease (AD), the most common type of dementia, is expected to affect 152 million people by 2050, emphasizing the importance of early diagnosis. This study uses the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, combining cognitive tests, biomarkers, demographic details, and genetic data to build predictive models. Using large language models (LLMs), specifically ChatGPT 3.5, we achieved high classification accuracy, with ROC AUC values of 0.98 for cognitively normal (CN) individuals, 0.99 for dementia, and 0.98 for mild cognitive impairment (MCI). These findings show that LLMs can handle complex data quickly and accurately. By focusing on numerical and text-based data instead of just imaging, this method provides a cost-effective and accessible option for diagnosing AD. Adding genetic information improves the predictions, reflecting the important role of genetics in AD risk. This study highlights the potential of combining different types of data with advanced machine learning and LSTM to improve early AD diagnosis. Future research should explore more ways to combine data and test different machine learning models to further enhance diagnostic tools.**

***Keywords*—Alzheimer; dementia; LLMs; ChatGPT; LSTM**

I. Introduction

Alzheimer's disease (AD), the most common form of dementia, accounts for 60–80% of all dementia cases and is expected to become even more prevalent as populations age [1-7]. By 2050, it is estimated that 152 million people worldwide will be living with Alzheimer's and other forms of dementia [1]. AD is particularly significant among non-communicable diseases, which together account for approximately 70% of global deaths [8-10]. The disease primarily affects older adults, impairing memory, behaviour, and reasoning abilities [11, 12].

To improve the health and quality of life for older adults, it is increasingly important to develop effective treatments for AD [13-15]. Although several biomarkers for early diagnosis have been identified [4], accurate diagnosis still partly relies on clinical criteria, which can take up to six months as symptoms gradually become apparent [16, 17]. Even when symptoms are clear enough for a confirmed diagnosis, AD remains incurable [10, 18, 19].

The Global Deterioration Scale [20] outlines symptoms such as decreased work performance, increased forgetfulness, and frequent disorientation during the mild cognitive impairment (MCI) stage. MCI is characterized by a prolonged period of decline that can last for around seven years [18]. This has led medical professionals to focus on establishing criteria for early diagnosis to slow or potentially prevent disease progression [14, 18].

Current medical practices for diagnosing AD can be time-consuming. As a result, researchers are increasingly exploring approaches that combine medicine with computer science, particularly deep learning, to develop methods for earlier and more accurate diagnoses. This interdisciplinary focus has become a key area of ongoing research.

Various approaches have been employed to diagnose Alzheimer's disease (AD) as early as possible. These include skeleton-based human action evaluation [18], 3D CNN-based classification of sMRI and MD-DTI images to detect brain changes [8, 9, 19], and speech analysis [15, 17]. Monitoring dementia progression using 2D and 3D imaging techniques has become a sophisticated method, utilizing advanced medical imaging to track changes in brain structure and function over time [2, 7, 19, 21, 22]. Additionally, it is crucial to determine whether cognitive measures identified as predictive in research cohorts are also applicable in clinical memory clinics [2, 13, 23-25].

This study aimed to identify the most effective measures for predicting future AD dementia in clinical settings where expensive biomarkers may not be widely available. Early detection of AD severity is critical [5, 14, 19, 26, 27]. While neuroimaging and computer-assisted diagnostic tools can detect AD in its early stages, these methods often lack high accuracy [11]. Techniques like Computed Tomography (CT), Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI) significantly contribute to diagnosing brain disorders [28].

Advancements in medical imaging, supported by computer-aided diagnostic research, have greatly improved the ability to monitor and predict dementia progression. 2D and 3D imaging techniques play an essential role, offering a comprehensive view through structural, functional, and molecular imaging. These advances continue to improve our understanding and management of this complex condition, providing hope for better patient outcomes.

However, several challenges persist. Advanced imaging techniques such as MRI and PET are expensive, making routine monitoring financially difficult for many patients. Access to high-quality imaging facilities is often limited, particularly in rural or underserved areas. Additionally, technical limitations such as resolution and sensitivity may prevent the detection of small or early brain changes. Imaging data can also be affected by noise and artifacts, complicating interpretation. Health-related concerns include exposure to ionizing radiation from PET and CT scans, especially with repeated use, and discomfort or anxiety during MRI scans, particularly for patients with claustrophobia.

From a technical perspective, the large volume of data generated by 3D imaging requires robust data management and analysis systems, along with advanced computational tools and expertise for handling and interpreting big data. While these imaging techniques provide valuable insights, their limitations underscore the importance of a balanced approach.

Recent advancements have also identified blood-based biomarkers that can assist in monitoring dementia progression. These tests are less invasive than cerebrospinal fluid (CSF) tests and more practical for regular use. Key biomarkers include Amyloid Beta (Aβ), Tau Proteins, and Neurofilament Light Chain (NfL). Cognitive tests remain essential for assessing cognitive functions and tracking changes over time. Commonly used tests include the Mini-Mental State Examination (MMSE), Montreal Cognitive Assessment (MoCA), Clock Drawing Test, Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog), and neuropsychological testing. This paper consists of literature review, methodology, results, discussion and conclusion, where every section highlights related information.

II. Literature Review

Alzheimer's disease (AD) is a neurological disorder that progressively worsens over time. It is typically divided into three main stages, with the early stage being particularly important [29]. One challenge with Alzheimer's treatments is their limited effectiveness, especially during the disease's typical eight-year progression [29].

In the early stage, daily activities gradually become more difficult. The most common signs include short-term memory loss and trouble learning new things [12, 29]. Other symptoms may include low energy, a sad mood, and a lack of interest or motivation [30]. Notably, individuals diagnosed with mild cognitive impairment (MCI) are at a significantly higher risk of developing AD [29]. In many cases, MCI is classified as the early stage of AD [31].

Researchers have also observed that difficulties with language—such as trouble pronouncing or remembering complex words—alongside challenges in performing daily tasks, may serve as early indicators of AD [29, 31, 32]. These findings suggest that there are more affordable diagnostic tools for detecting dementia than costly imaging techniques. Many studies have applied statistical analysis alongside other types of data (excluding imaging) to explore alternative methods for diagnosing Alzheimer's disease (AD). For instance, human action evaluation has shown potential applications in areas such as assisted living, physical rehabilitation, sports activity scoring, and skills training [18]. This approach can also be applied to AD by using sequences of 3D skeletal joint data to assess the severity of the disease in patients. For example, Yu et al. [18] utilized a two-task graph convolutional network to analyze skeleton data for tasks involving abnormality detection and quality evaluation. Their method, evaluated using the UI-PRMD dataset, demonstrated accurate abnormality detection.

Taghvaei et al. [33] applied statistical analysis to investigate the relationship between white matter hyperintensities (WMH) tract disconnection and cognitive performance. Their study highlighted the significant role of demographic factors, such as education level, age, and sex, in influencing the relationship between WM tracts and cognitive scores. Similarly, Raj et al. [1, 34] emphasized the importance of genetics, which contributes to approximately 70% of the overall risk for AD. They introduced a system that combines text mining and machine learning to identify and prioritize candidate genes for AD, categorizing them into three association classes with corresponding weights.

Another cost-effective method for predicting AD involves analyzing patients' speech patterns [35]. Many studies have focused on acoustic and syntactic analysis of speech. For example, Haj Zargarbashi and Bagher [17] employed statistical and neural methods to classify audio signals into dementia and control groups, achieving an accuracy of 83.6%. Similarly, Vincze et al. [36] analyzed specific utterances using deep learning models, with and without demographic data, and found no significant differences between the models. Colla et al. [37] used large language models alongside N-grams and perplexity metrics to predict potential AD with an accuracy of 84%. Gómez-Zaragoza et al. [15] provided empirical evidence that punctuation and pauses in speech could reveal early signs of AD. A comprehensive review by [38] highlighted the high potential for deep learning to be utilized with medical data in future research.

Cai et al. [35] examined methods for detecting AD through speech analysis by transcribing audio into text and extracting audio features using the WavLM model. They tested pre-trained models and Graph Neural Networks (GNNs) with the DementiaBank Pitt dataset and applied fine-tuning techniques such as data augmentation (e.g., synonym replacement and GPT-based augmentation). Wang et al. [39] also used pre-trained language models with fine-tuning methods, achieving up to 89% accuracy. Finally, blood tests have shown promise in detecting AD. Certain biomarkers in blood may indicate the potential presence of the disease. Kim and Lee [40] demonstrated that complex interactions among blood proteins could predict the likelihood of AD development.

Several gaps have been identified in Alzheimer's disease (AD) research. For example, while the ADNI dataset has been extensively used in studies focusing on MRI images [29], it has only been partially utilized. Most studies have concentrated solely on MRI images, overlooking the wealth of other information in the dataset that could enhance the analysis of AD. A recent systematic literature review by Singh et al. [41] found that the majority of AD studies used ADNI and deep learning methods, a trend confirmed by Alwuthaynani et al. [2],

but with an exclusive focus on MRI data. Similarly, Essemlali et al. [42] highlighted that one of the key challenges in dementia prediction lies in distinguishing between MCI and AD, as well as between NC (normal cognition) and MCI. These are among the most difficult classification tasks and often require additional data, such as multi-modality or genetic information, to improve predictions.

This study leverages the powerful classification capabilities of large language models (LLMs), such as ChatGPT, to classify dementia stages. There is a notable lack of research utilizing LLMs with non-imaging data from the ADNI dataset to address classification problems. For instance, Agbavor and Liang [43] demonstrated that GPT-3 embeddings significantly improved Alzheimer's detection accuracy from spontaneous speech, outperforming traditional methods based on acoustic features. They used models such as support vector classifiers and logistic regression, achieving high classification performance. Another study [44] combined imaging and phenotype data from the ADNI dataset with LLMs, achieving state-of-the-art performance in classifying Alzheimer's and various stages of cognitive impairment. This highlights the effectiveness of integrating LLMs with diverse data types.

However, as discussed, this study focuses on utilizing LLMs with textual and numerical data rather than imaging data. This approach aims to provide a cost-effective solution suitable for clinical settings and requiring less computational power. The methodology is informed by the work of Feng et al. [44], excluding imaging data, to create a more accessible and efficient model for AD classification.

III. METHODOLOGY

A. ADNI Dataset

This study utilized the ADNI dataset due to its extensive use in recent research and its comprehensive collection of data necessary to achieve the goals of this study. The ADNI dataset, available at adni.loni.ucla.edu, has been widely referenced in studies related to Alzheimer's detection [41, 45].

The ADNI project was launched in 2003 by the National Institute on Aging (NIA), the Food and Drug Administration (FDA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), private non-profit organizations, and pharmaceutical companies [41, 45]. Its original purpose was to determine whether the combination of genetic, neuroimaging, biomarker, clinical, and neuropsychological data could be used to predict Alzheimer's disease.

The ADNI dataset is renowned for its longitudinal design, capturing data at multiple time points. Images and other data are collected at baseline and then at intervals of 6 months, 12 months, 24 months, and 48 months [42]. Several versions of the dataset, including ADNI2 and ADNI-Go, have been developed to expand its scope.

B. Feature Set

In addition to demographic data, such as sex, age, gender, and race, which have been shown to significantly contribute to predicting the clinical status of individuals [36], this study incorporates a variety of other features for analysis. Cognitive tests play a central role in assessing various aspects of memory, language, and executive function. These include the Mini-Mental State Examination (MMSE), a widely used measure of cognitive function, and the Montreal Cognitive Assessment (MOCA), another standard cognitive measure. The study also utilizes scores from the Rey Auditory Verbal Learning Test (RAVLT), including immediate recall, learning, forgetting, and percentage of forgetting, which assess different dimensions of memory. Other cognitive tests include the Logical Memory Delayed Recall test (LDELTOTAL), the Digit Span test (DIGITSCOR), which measures attention and working memory, and the Trail Making Test Part B (TRABSCOR), which evaluates executive function. The Functional Activities Questionnaire (FAQ) is also included to assess daily living capabilities.

Biomarkers represent another crucial component of the analysis. The Apolipoprotein E (APOE4) genotype, strongly associated with Alzheimer's disease risk, is used alongside imaging biomarkers from PET scans, including tracers such as FDG, PIB, AV45, and FBB. Additionally, cerebrospinal fluid protein levels of amyloid-beta (ABETA), tau (TAU), and phosphorylated tau (PTAU) are examined for their role in the disease's progression.

Clinical and diagnostic scores also contribute significantly to the analysis. These include the Clinical Dementia Rating – Sum of Boxes (CDRSB), which evaluates cognitive and functional performance, and various subscales from the Alzheimer's Disease Assessment Scale (ADAS), such as ADAS11, ADAS13, and ADASQ4. Reports of cognitive function from both patients and their study partners are included, with patient-reported scores covering memory, language, visuospatial ability, planning, organization, divided attention, and overall cognitive function. Study partner-reported scores assess the same domains, providing additional perspectives on cognitive performance.

Baseline data for all these measures are also incorporated to analyze changes over time. Baseline values for clinical scores such as CDRSB, ADAS, and MMSE, as well as cognitive tests like RAVLT, Logical Memory, and FAQ, are included. Baseline levels of biomarkers, such as ABETA, TAU, and PTAU, are also considered. Patient- and study partner-reported cognitive scores at baseline offer further context for tracking progression. Temporal variables, such as the number of years or months since the baseline visit, are included to provide additional detail about the timing of data collection.

In summary, the features included in this study's prediction models encompass cognitive assessments, genetic information, biomarkers, demographic data, and clinical details. These measures provide a comprehensive dataset for early detection of dementia, offering insights into cognitive decline and associated risk factors. By integrating this diverse set of features, the study aims to improve the accuracy and practicality of predictive models for Alzheimer's disease [27, 46, 47].

C. Models

Shah and Shah [12] explored the use of machine learning (ML) algorithms, particularly convolutional neural networks (CNNs), for the early diagnosis of Alzheimer's disease (AD)

through the analysis of medical imaging data, such as MRI scans and biomarkers. Their study compared various ML algorithms, including k-nearest neighbor (KNN) and support vector machines (SVM), and highlighted the superior accuracy and reliability of CNNs in detecting AD. The deep learning capabilities of CNNs enable them to extract subtle features from medical images, making them particularly effective for this application. Shah and Shah [12] demonstrated that CNNs outperformed other algorithms due to their deep architecture, which can handle complex data and identify patterns that simpler models may miss.

However, the authors also noted several challenges in using CNNs for AD diagnosis. These include the need for large, well-curated datasets, as CNN models are prone to overfitting when trained on small or imbalanced datasets. Additionally, they emphasized the importance of transparency and interpretability in ML models, especially in medical applications where clinicians need to understand the rationale behind a diagnosis. Despite these challenges, CNN-based models were identified as the most effective for early detection of AD, particularly when applied to MRI scans and biomarker data.

While Shah and Shah [12] provided a thorough investigation into the use of ML and deep learning methods for early AD detection, their work predominantly focused on image-based data, such as MRI scans. They did not explore the application of these methods to other types of data, such as textual or numerical information, nor did they consider the potential of large language models (LLMs). LLMs, with their ability to process both structured and unstructured text, could provide valuable insights and significantly enhance prediction models for AD. Additionally, the study did not examine the use of Long Short-Term Memory (LSTM) networks, which are particularly effective for analyzing sequential data, such as time-series health records or longitudinal datasets.

Building on their work, this study proposes the integration of both LSTM networks and LLMs alongside traditional machine learning methods to develop cost-effective and accurate prediction solutions for AD. By focusing on non-image data, such as cognitive assessments, biomarkers, and clinical records, this approach aims to expand the scope of predictive models and offer accessible diagnostic tools for clinical settings.

1) LSTM: Long Short-Term Memory (LSTM) networks are a specialized type of recurrent neural network (RNN) designed to learn and retain long-term dependencies in sequential data. They are particularly well-suited for tasks involving time-series or sequential data due to their unique architecture, which includes memory cells and gating mechanisms (input, output, and forget gates) to control the flow of information. This capability makes LSTM networks highly effective in addressing the vanishing gradient problem, a common challenge in traditional RNNs.

In this study, LSTM networks are employed to process clinical and biomarker data from the ADNI dataset. The input features, which consist of textual and numerical data, are well-suited to LSTM's architecture. The model begins with an input layer that accepts the data features, followed by one or more LSTM layers that process the sequences of observations. The final dense layer produces the output, classifying the stages of Alzheimer's disease based on the processed data. By leveraging LSTM's ability to handle sequences effectively, this study aims to improve classification accuracy for Alzheimer's diagnosis. Additionally, LSTM networks are prioritized due to their demonstrated effectiveness in managing textual and numerical data, which are key components of the ADNI dataset. The following equations illustrate the core functionality of LSTM networks in classification tasks:

The forget gate controls which information from the previous cell state (C_{t-1}) should be carried forward to the current cell state (C_t).

$$f_t=\sigma(W_f \cdot[h_{t-1}, x_t] + b_f)$$

Where f_t is the forget gate's activation vector; W_f and b_f are the weight matrix and bias for the forget gate; h_{t-1} is the previous hidden state; x_t is the current input and σ is the sigmoid activation function [48]. Input Gate [49] represented as follows:

$$i_t=\sigma(W_i\cdot[h_{t-1}, x_t] + b_i)$$

$$\dot{C}_t =tanh\ (W_C \cdot[h_{t-1}, x_t] + b_C)$$

where i_t is the input gate's activation vector; $\dot{C}_t$ is the candidate cell state, representing new information; and W_i, W_C b_i, b_C are the weight matrices and biases for the input gate and cell state.

Meanwhile; Cell State Update is implemented as follows

$$C_t=f_t*C_{t-1}+i_t* \dot{C}_t$$

And the output gate [50] is implemented as follows

$$o_t=\sigma(W_o\cdot[h_{t-1}, x_t]+b_o)$$

$$h_t=o_t*tanh(C_t)$$

where o_t is the output gate's activation vector; h_t is the current hidden state (which also serves as the output for classification).

The classification layer is implemented using this formula [51]:

$$y=softmax(W_y\cdot h_t+by)$$

where y is the predicted class probabilities; W_y and by are the weight matrix and bias for the classification layer.

finally; Loss Function for Classification (Cross-Entropy) [52] is presented by

$$L= -\sum_{i=1}^{N} y_i\ log(y'_i)$$

Where: y_i is the true label; y'_i is the predicted probability for class and N is the number of classes.

At each time step, the LSTM updates the cell state and hidden state using the forget, input, and output gates. The final hidden state after processing the entire sequence is passed to the classification layer, where the class probabilities are calculated. The loss function is used to optimize the model by comparing the predicted output with the actual labels. This process allows the LSTM to classify sequential data effectively.

2) Few-Shot: Few-shot learning is a machine learning technique that enables models to perform tasks with minimal examples or training data. Unlike traditional machine learning methods, which rely on large datasets and extensive training, few-shot learning allows for adaptability and flexibility with significantly reduced training overhead. This approach is particularly useful for large language models (LLMs) like OpenAI's ChatGPT-3.5-turbo, as it enables the model to learn effectively from a limited number of examples. The primary advantage of few-shot learning is its ability to achieve accurate predictions with less effort in data preparation and model training.

In this study, few-shot learning was applied to classify Alzheimer's disease stages using examples from the ADNI dataset. The process began by defining a small set of examples representing the desired outcomes: "CN" (cognitively normal), "Dementia," and "MCI" (mild cognitive impairment). A sample of 10 examples was provided, reflecting various cases in the dataset, including instances with missing values across specific groups of features. Next, a prompt was constructed to include these examples and the task to be performed, guiding the model toward the desired predictions. The model then used this context to generate predictions, which were subsequently extracted to retrieve the relevant outputs. This streamlined process demonstrates the efficiency of few-shot learning in handling limited data while maintaining accurate and meaningful results.

IV. Results

A. Demographic

The ADNI dataset contains numerous cases categorized under each stage of dementia. These cases include multiple rows for the same patient, corresponding to different visits over time. The distribution of cases across the stages of dementia is illustrated in Fig. 1. As shown, the majority of patients are at the MCI stage, followed by those at the CN stage, highlighting the prevalence of MCI cases in the dataset.

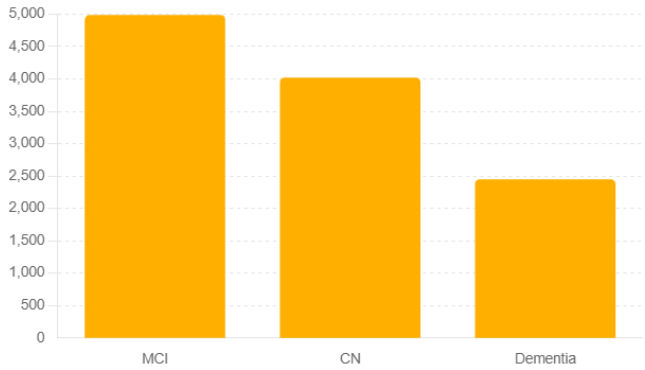

Fig. 1. Patients'stage in ADNI dataset.

The demographic data available in the ADNI dataset is presented in Fig. 2. The data shows a balanced distribution in terms of gender. However, there is an imbalance in race, with the white population significantly outnumbering other racial groups. The age range of participants spans from 55 to 90 years.

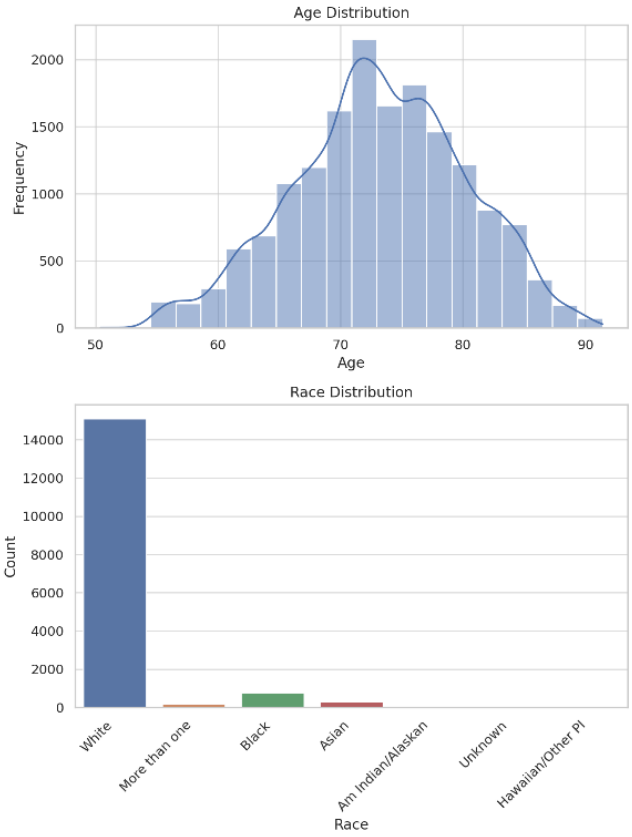

Fig. 2. Demographics related statistics.

B. Feature Selection and Machine learning Methods' Prediction

To identify the most important features for predicting the diagnosis label (DX), three approaches were implemented: (a) Correlation Analysis to examine relationships between features and DX, (b) Random Forest Feature Importance to rank features by their predictive significance, and (c) Recursive Feature Elimination (RFE) to select the most relevant features based on model performance.

The Correlation Analysis revealed that the features most strongly correlated with DX are: CDRSB (Cognitive Dementia Rating – Sum of Boxes) with a correlation of 0.751, EcogSPTotal (Total score of Everyday Cognition, Study Partner version) at 0.735, EcogSPMem (Everyday Cognition, Study Partner Memory score) at 0.732, FAQ (Functional Activities Questionnaire) at 0.730, and ADAS13 (Alzheimer's Disease Assessment Scale, 13-item) at 0.725. These results indicate that cognitive and functional assessments are highly correlated with Alzheimer's diagnosis. Using Random Forest Feature Importance, the top features for predicting DX were identified as EcogSPTotal (20.28%), ADAS13 (18.84%), ADAS11 (16.54%), EcogSPMem (12.64%), and EcogSPOrgan (9.45%). These findings reinforce the importance of cognitive and functional assessments in predicting Alzheimer's.

With Recursive Feature Elimination (RFE), the same top five features were selected: EcogSPTotal, EcogSPMem, ADAS13, ADAS11, and EcogSPOrgan. A Random Forest Classifier trained using only these features achieved an average

cross-validation accuracy of 38.75% (standard deviation: 0.89%) and a test accuracy of 39%. Class 2 (Dementia) was predicted with the highest recall (63%), but precision and recall for Class 0 (Cognitively Normal) and Class 1 (MCI) were low. The macro-average F1-score was 0.32, highlighting significant confusion between classes, particularly between Cognitively Normal and MCI.

To address these limitations, several strategies were explored. Class Imbalance Handling techniques, such as oversampling, undersampling, and class-weighted models, improved the weighted F1-score slightly to 0.37, with Class 2 (Dementia) achieving 63% recall. However, confusion between Cognitively Normal and MCI persisted. Feature Engineering was then applied, creating interaction features by combining cognitive scores and biomarkers. This approach marginally increased the overall accuracy to 40%, with a recall of 66% for Class 2. While improvements in classifying Cognitively Normal and MCI cases were observed, misclassification between MCI and Dementia remained significant.

Next, Advanced Models were tested, including Gradient Boosting and Support Vector Machines (SVM). Gradient Boosting achieved an overall accuracy of 43%, with Class 2 (Dementia) recall at 92%. However, predictions for Class 0 (Cognitively Normal) and Class 1 (MCI) were poor, with F1-scores close to zero. Similarly, SVM performed slightly better with an overall accuracy of 45% and a recall of 99% for Class 2, but almost no correct predictions for Classes 0 and 1. Both models struggled with class imbalance, favoring Dementia at the expense of distinguishing other classes.

Finally, Ensemble Modeling was implemented, leveraging techniques like Voting Classifiers (combining Random Forest, Gradient Boosting, and SVM), Stacking Classifiers (training multiple models and using their predictions as inputs for a meta-model), and Bagging (aggregating predictions from models trained on different data subsets). These ensemble methods aim to balance predictive performance across classes and address the challenges of class imbalance and overlapping features. Further evaluation and refinement of these approaches are ongoing to improve the overall diagnostic accuracy and robustness of the model.

1) Voting classifier results: The model achieved an overall accuracy of 43%, with Class 2 (Dementia) having the highest recall at 90%. However, performance for Class 0 (Cognitively Normal) and Class 1 (MCI) was poor, with F1-scores of approximately 0.09 and 0.05, respectively. The macro-average F1-score was 0.24, highlighting that the model disproportionately favors Dementia while struggling to accurately classify Cognitively Normal and MCI cases. Significant confusion remains between the classes, with many CN and MCI cases misclassified as Dementia.

2) Stacking classifier results: The model achieved an overall accuracy of 45%, matching the performance of the Voting Classifier. However, the results reveal a significant bias, as Class 2 (Dementia) was the only class predicted, with a recall of 100%. Both Class 0 (Cognitively Normal) and Class 1 (MCI) had 0% recall and precision, indicating that no cases from these classes were correctly identified. The macro-average F1-score was 0.21, underscoring the model's heavy bias toward Dementia and its inability to distinguish between Cognitively Normal and MCI cases. All instances of CN and MCI were misclassified as Dementia.

3) Bagging: Using Bagging with the Random Forest model, which inherently trains on different subsets of data, the overall accuracy achieved was 41%. Class 2 (Dementia) had a recall of 73%, while Class 0 (Cognitively Normal) and Class 1 (MCI) showed lower recall and F1-scores. The macro-average F1-score was 0.31, indicating a moderate improvement in balancing predictions across classes compared to previous classifiers. Bagging demonstrated better differentiation between classes than the Voting and Stacking classifiers, though significant misclassifications remained between Cognitively Normal, MCI, and Dementia. Notably, Bagging performed slightly better in predicting minority classes.

In summary, the Voting Classifier achieved an accuracy of 43% but was heavily biased toward predicting Dementia, performing poorly on other classes. The Stacking Classifier achieved a slightly higher accuracy of 45%, but it failed to classify Cognitively Normal and MCI cases, predicting only Dementia. Bagging, with an accuracy of 41%, showed improved balance across the classes compared to Voting and Stacking but continued to struggle with distinguishing between Cognitively Normal and MCI cases.

Among the ensemble methods, Bagging demonstrated better overall balance in classifying multiple categories, although limitations remained. To further enhance performance, future work could focus on advanced feature engineering, fine-tuning model hyperparameters, or exploring other sophisticated techniques to address the challenges in distinguishing between these diagnostic categories.

C. Advanced Analyses

To further improve model performance, particularly in distinguishing between the different diagnostic classes (Cognitively Normal, MCI, and Dementia), several advanced techniques could be employed. Automated machine learning (AutoML) tools, such as TPOT, can automate the process of testing a range of algorithms and hyperparameter configurations, helping to identify the best model without requiring manual experimentation.

In this study, an AutoML framework was used to automatically test multiple models and optimize their hyperparameters. The results from the TPOT AutoML run provided valuable insights into the model selection and tuning process. The internal cross-validation (CV) accuracy, which TPOT uses to evaluate different pipelines during its evolutionary search, stabilized at approximately 47.7% across most generations. The final best pipeline achieved an internal CV score of 47.78%, indicating the highest performance based on TPOT's cross-validation evaluation.

TPOT selected the ExtraTreesClassifier as the best model. This ensemble method, similar to Random Forest, reduces variance by averaging predictions across multiple decision trees, often resulting in more stable outcomes. The

ExtraTreesClassifier was identified as the optimal model with the following key parameters:

- Bootstrap: True (bootstrap sampling was used).
- Criterion: Gini (used for measuring the quality of a split).
- Max features: 0.8 (80% of the features are considered when looking for the best split).
- Min samples leaf: 17 (minimum number of samples required to be at a leaf node).
- Min samples split: 5 (minimum number of samples required to split an internal node).
- n_estimators: 100 (number of trees in the forest).

The test set accuracy was 43.6%, meaning the final model achieved 43.6% accuracy on unseen data. While consistent with the performance of other models tested, this accuracy highlights the challenges in distinguishing between the diagnostic categories (Cognitively Normal, MCI, and Dementia). The consistency of accuracy between 43% and 45% across different models suggests that the feature set may require further refinement or that the inherent complexity of differentiating between these classes, particularly between Cognitively Normal and MCI, remains a significant challenge.

Since focusing on the top five features did not result in substantial improvements in accuracy, a new approach was tested by including DX_bl (the baseline diagnosis) as a key feature for predicting the final diagnosis (DX). This approach is logical, as the baseline diagnosis likely correlates strongly with the final diagnosis, and transitions between diagnostic categories (e.g., from MCI to Dementia) over time can provide valuable insights. Using a Random Forest model with DX_bl as a feature, the accuracy improved significantly to 83%, demonstrating that DX_bl is a strong predictor of the final diagnosis.

The model's performance metrics for each class are as follows:

Class 0 (Cognitively Normal - CN):

Precision: 89% (89% of cases predicted as CN are correct).

Recall: 92% (92% of actual CN cases were correctly identified).

F1-Score: 91% (indicating strong and balanced performance for this class).

Class 1 (MCI):

Precision: 100% (all predicted MCI cases were correct).

Recall: 51% (only 51% of actual MCI cases were identified correctly).

F1-Score: 67% (highlighting an imbalance, as many MCI cases are misclassified).

Class 2 (Dementia):

Precision: 75% (75% of Dementia predictions were correct).

Recall: 93% (93% of actual Dementia cases were identified).

F1-Score: 83% (indicating strong performance for this class).

The Macro Average F1-Score, which gives equal weight to each class, was 80%, showing good overall performance but highlighting some imbalance in the prediction of MCI. The Weighted Average F1-Score, which accounts for the number of instances in each class, was 82%, reflecting the model's strong performance for the larger classes (CN and Dementia).

Cognitively Normal (CN) and Dementia cases are well predicted, with high precision and recall. However, MCI remains the most challenging class for the model to classify, with high precision but low recall. This indicates that many MCI cases are misclassified as either CN or Dementia. While DX_bl is a highly predictive feature for the final diagnosis, the model still struggles to effectively differentiate MCI from the other categories.

D. Baseline Data Analysis and Prediction

Baseline data analysis can provide valuable insights and contribute to improving predictions. Among the diagnostic groups, Cognitively Normal (CN) and Alzheimer's Disease (AD) dominate the baseline dementia diagnoses. This analysis specifically compares these two groups with respect to key variables, such as CDRSB (Cognitive Dementia Rating Sum of Boxes) and PTAU_bl (Phosphorylated Tau at baseline). Using the Mann-Whitney U test, a non-parametric test suitable for comparing two independent samples, the following results were obtained:

CDRSB: The p-value was effectively 0, indicating a highly significant difference in cognitive scores between the CN and AD groups.

PTAU_bl: The p-value was 0.0001, also showing a highly significant difference in phosphorylated tau levels between the two groups. These findings suggest that both cognitive scores (CDRSB) and biomarker levels (PTAU_bl) are significantly different between CN and AD groups, consistent with established Alzheimer's research. The Mann-Whitney U test for ABETA levels yielded a p-value of <0.0001, further confirming a significant difference between CN and AD groups. This result aligns with the well-documented role of amyloid-beta in Alzheimer's disease pathology. Similarly, for MMSE (Mini-Mental State Examination), the p-value was <0.0001, indicating that AD participants had significantly lower MMSE scores, reflecting more severe cognitive impairment. The same was observed for MOCA (Montreal Cognitive Assessment), where the p-value was 0.0001, demonstrating a significant difference in MOCA scores between CN and AD groups.

The analysis of neuroimaging biomarkers, such as FDG (Fluorodeoxyglucose PET) and AV45 (Amyloid PET), also revealed significant differences. The p-value for FDG PET was 0.0001, indicating that brain glucose metabolism, as measured by FDG PET, is significantly lower in individuals with AD. Similarly, AV45 PET, which measures amyloid accumulation, showed a p-value of 0.0001, confirming higher amyloid levels

in the AD group. These findings underscore the significant differences in neuroimaging biomarkers between CN and AD groups.

Regarding the APOE4 genotype distribution, the results highlight the strong association between APOE4 and Alzheimer's risk. Among individuals with AD, 47.7% had one copy of the APOE4 allele, 21.2% had two copies, and 31.1% had no copies. In contrast, among CN individuals, 72.6% had no copies, 25.4% had one copy, and only 2% had two copies. These distributions confirm the well-established link between APOE4 and increased Alzheimer's risk, with individuals carrying one or two APOE4 alleles being significantly more likely to develop the disease.

Examining the impact of APOE4 status on biomarkers and cognitive scores provides additional insights into how this genetic risk factor influences Alzheimer's pathology. Individuals without APOE4 alleles had the highest average ABETA levels (990), while those with two alleles had the lowest (521), consistent with APOE4's role in promoting amyloid accumulation. Similarly, TAU and PTAU levels were progressively higher in individuals with more APOE4 alleles. Those with two alleles had the highest TAU (363) and PTAU (35), reflecting greater neurofibrillary pathology. Cognitive function, as measured by MMSE, decreased with the number of APOE4 alleles, with individuals carrying two alleles having the lowest average MMSE score (24.5). The CDRSB score, indicating cognitive impairment, increased with the number of APOE4 alleles.

These findings demonstrate that APOE4 is strongly associated with greater amyloid and tau pathology and more severe cognitive decline. The relationship between APOE4 status, biomarkers, and cognitive scores underscores the genetic influence on Alzheimer's disease progression.

E. Neural Network Prediction Models

The results of the models developed in this study, particularly the LSTM model, were evaluated using the ADNI dataset. Performance metrics included accuracy, precision, recall, and F1-score. The model achieved a Validation Loss of 0.369 and a Validation Accuracy of 84.5%, with an overall Accuracy of 84.5%. A detailed classification report is presented in Table I, while the confusion matrix is illustrated in Fig. 3, providing further insights into the model's performance across diagnostic categories.

TABLE I. PERFORMANCE MATRIX FOR LSTM-BASED MODEL

	precision	Recall	F1-score
CN	0.84	0.81	0.83
Dementia	0.79	0.80	0.80
MCI	0.86	0.87	0.87
Accuracy	0.84		
macro avg.	.83	.83	.83
Weighted avg	0.84	0.84	.84

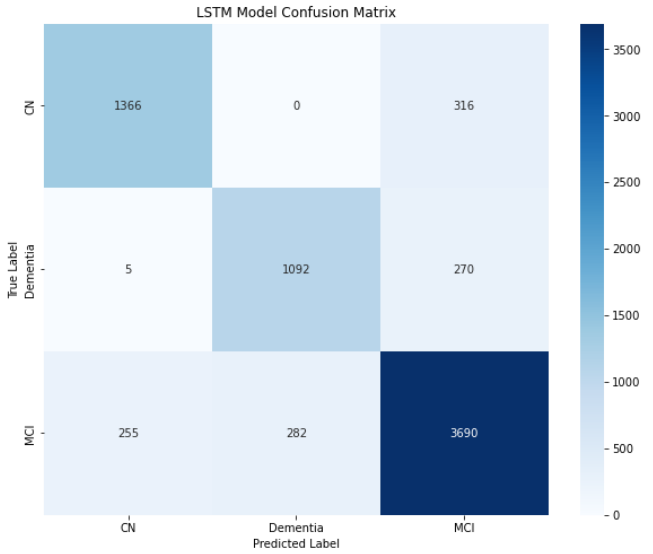

Fig. 3. Confusion matrix for LSTM-based model.

It was observed that the features illustrated in Fig. 4 contribute significantly more to predicting Alzheimer's disease compared to other features. Among these, MMSE_bl (baseline Mini-Mental State Examination) and MMSE_fu (follow-up Mini-Mental State Examination) showed the greatest contribution, highlighting their critical role in the predictive model.

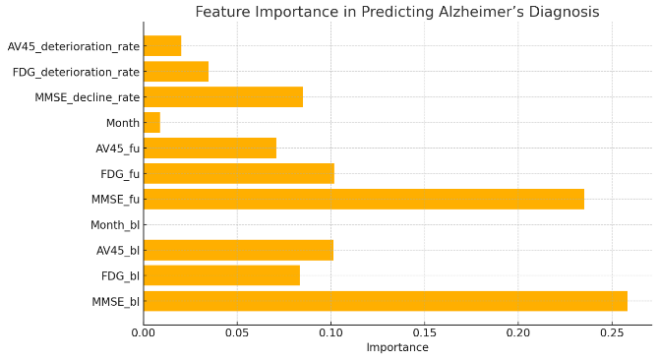

Fig. 4. The important features in predicting Alzheimer.

Additional models were explored using baseline information, including Gradient Boosting and Support Vector Machine (SVM). The Gradient Boosting model achieved an impressive 96% accuracy on the test set, with perfect precision, recall, and F1-scores for both the Cognitively Normal (CN) and Alzheimer's Disease (AD) groups. Similarly, the SVM classifier performed exceptionally well, achieving 95% accuracy on the test set and perfect scores across all performance metrics. These results demonstrate the strong predictive capabilities of both models when using baseline data.

F. Evaluating Model Performance Longitudinally

To assess longitudinal performance, the model's ability to predict changes in cognitive scores, imaging biomarkers, or diagnoses over time was evaluated. This involved two main aspects:

Predicting Cognitive Decline: The model was used to predict future cognitive scores based on baseline data and observed longitudinal trends.

Evaluating Model Stability: The model's predictive accuracy was tested across multiple visits for the same individuals to determine its consistency over time.

The analysis began with visualizing interactions among key features, followed by evaluating longitudinal performance. A heatmap (Fig. 5) illustrates the correlations between key features. Notably, the MMSE Decline Rate is negatively correlated with both MMSE at Baseline and FDG (Fluorodeoxyglucose PET), indicating that as cognitive function declines, brain metabolism also tends to decrease. In contrast, AV45 (Amyloid PET) shows a weaker correlation with other features, highlighting its specific role in amyloid accumulation, which is less directly tied to immediate cognitive decline. These findings underscore the nuanced relationships between cognitive decline, biomarkers, and brain function over time.

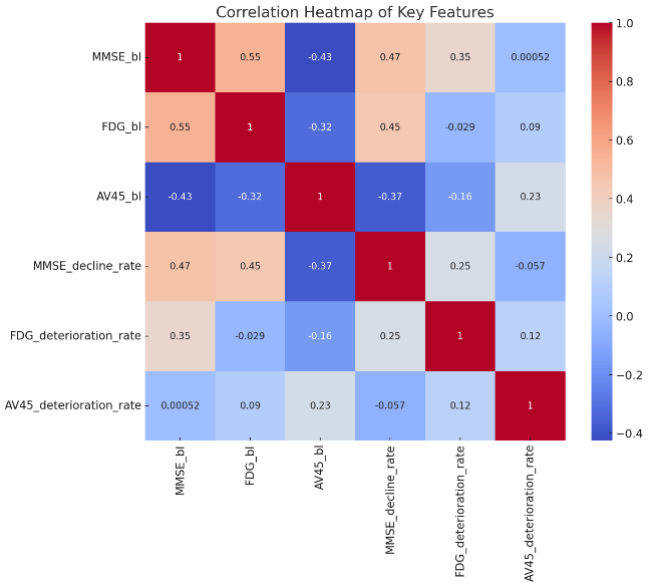

Fig. 5. Correlation of heatmap of key features (baseline features).

The model's ability to predict cognitive decline, such as changes in MMSE scores over time, was evaluated using baseline data. The assessment focused on how well the model predicts cognitive decline across multiple visits. A linear regression model, which used baseline features including MMSE, FDG, and AV45, achieved a Mean Squared Error (MSE) of approximately 0.01. This indicates that the model performed reasonably well in predicting cognitive decline over time. The low error suggests the model has potential for forecasting Alzheimer's progression, though further refinement and validation on larger datasets could improve its accuracy and robustness.

When a more advanced regression method, such as Gradient Boosting, was applied to refine the longitudinal model, it achieved an MSE of 0.0121. While the Gradient Boosting model captured some patterns in the data, its slightly higher MSE suggests that the simpler linear regression model performed better for this specific task.

Fig. 6 visualizes the relationship between predicted and actual MMSE decline rates. The scatter plot includes a red dashed line representing ideal predictions, where predicted values perfectly match the actual values. Most predictions were reasonably close to the actual values, with some variability, which is expected in longitudinal predictions. These findings demonstrate the model's promise in predicting cognitive decline over time while highlighting areas for further improvement in longitudinal performance. The same analysis used to predict cognitive decline was applied to forecast imaging deterioration over time, focusing on FDG (glucose metabolism) and AV45 (amyloid accumulation). Baseline features such as FDG_bl, AV45_bl, and MMSE_bl were utilized to predict changes in FDG and AV45 over time. Gradient Boosting Regression was employed for both tasks, demonstrating strong performance in predicting longitudinal imaging changes.

For FDG deterioration rate prediction, the model achieved a Mean Squared Error (MSE) of 0.000006, indicating excellent accuracy in forecasting changes in glucose metabolism over time. Similarly, for AV45 deterioration rate prediction, the MSE was 0.000028, showing the model's effectiveness in predicting changes in amyloid accumulation. These low error rates suggest that the model performs well for both imaging biomarkers, making it a valuable tool for tracking Alzheimer's progression longitudinally. Visualization of the results further supports the model's effectiveness. In Fig. 7, the scatter plot illustrates the predicted vs. actual FDG deterioration rates, with most points closely aligned with the ideal prediction line (red dashed line), indicating strong predictive performance. Similarly, Fig. 8 shows the predicted vs. actual AV45 deterioration rates, with most points clustering near the ideal line, demonstrating the model's capability to accurately forecast amyloid accumulation changes.

Overall, the models effectively predict longitudinal changes in both FDG and AV45 imaging biomarkers, which are critical for monitoring Alzheimer's disease progression over time. These results highlight the utility of Gradient Boosting Regression in capturing complex patterns in imaging data.

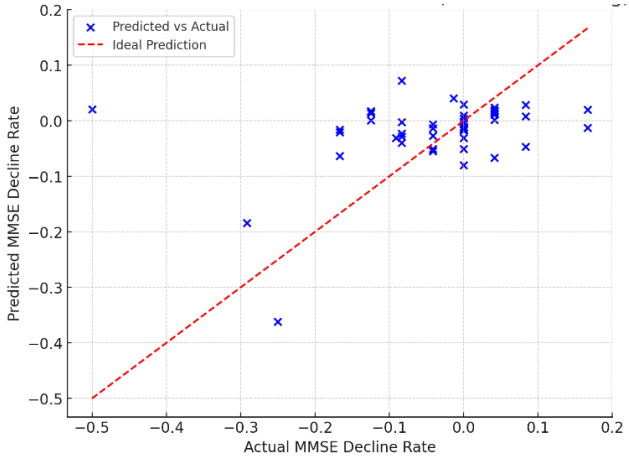

Fig. 6. Predicted vs. Actual MMSE Decline Rate (Gradient Boosting).

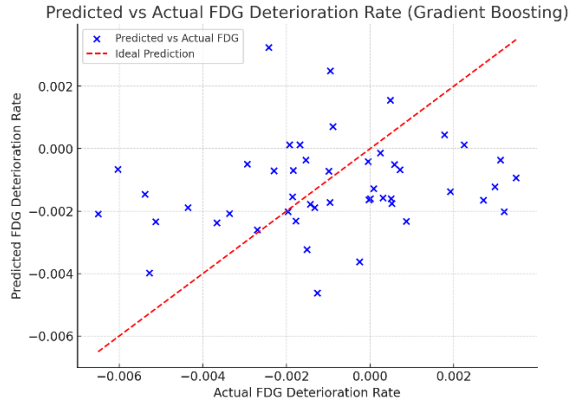

Fig. 7. Predicted vs. Actual FDG Deterioration Rate (using Gradient Boosting).

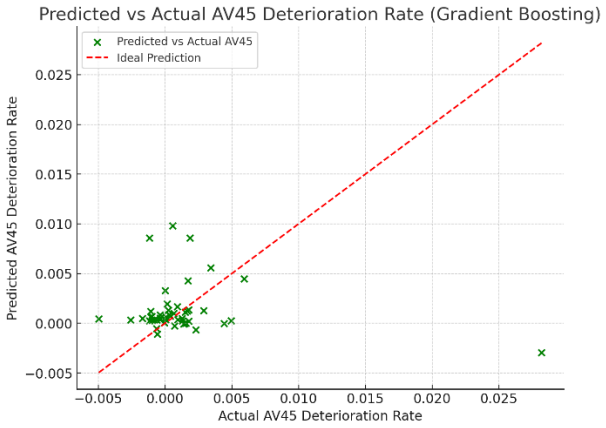

Fig. 8. Predicted vs. Actual AV45 Deterioration Rate (using Gradient Boosting).

We explored additional derived features by combining cognitive scores and biomarkers to create new indicators of Alzheimer's progression. Investigating interactions between longitudinal features and temporal patterns revealed deeper insights into disease progression. These derived features were designed to capture more complex relationships between cognitive and biomarker data. The newly created features included:

MMSE_TAU Interaction: A combination of baseline MMSE and FDG values to assess the relationship between cognitive function and brain metabolism.

ABETA_TAU Ratio: A ratio of amyloid (AV45) to glucose metabolism (FDG) to highlight the balance between amyloid accumulation and metabolic activity.

Combined Decline Rate: The sum of MMSE and FDG deterioration rates, providing a measure of overall cognitive and metabolic decline.

Using cross-validation, the performance of models with these derived features was evaluated. The Random Forest model achieved an accuracy of 97.46%, slightly outperforming the other models. The Gradient Boosting model achieved an accuracy of 96.83%, while the Support Vector Machine (SVM) reached 96.21%. All three models demonstrated strong predictive performance, with Random Forest showing a marginal advantage. These results highlight the potential of derived features in enhancing model performance by capturing intricate relationships between cognitive and biomarker data. This approach emphasizes the value of feature engineering in advancing the predictive capabilities of machine learning models for Alzheimer's progression.

G. *Few-shot LLMs*

Regarding the results of the experiment with few-shot training, we employed Large Language Models (LLMs), specifically ChatGPT 3.5, instead of LSTM models. The results demonstrated a significant improvement over LSTM, with the following performance metrics: Accuracy: 0.97, Precision: 0.97, Recall: 0.97, and F1-Score: 0.97. The confusion matrix is presented in Fig. 4. To evaluate the diagnostic performance of the model, we used Receiver Operating Characteristic (ROC) curves for various clinical conditions, as shown in Fig. 9. The ROC curve provides a graphical representation of the model's ability to discriminate between diagnostic categories by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) across different threshold settings. The model demonstrated excellent discriminatory ability across all diagnostic categories. The Area Under the Curve (AUC) for the ROC of Cognitively Normal (CN) subjects was 0.98, indicating a high level of accuracy in distinguishing CN individuals from those with cognitive impairment. Similarly, the ROC curve for dementia yielded an AUC of 0.99, reflecting outstanding performance in identifying subjects with dementia. For Mild Cognitive Impairment (MCI), the ROC curve achieved an AUC of 0.98, signifying robust capability in differentiating MCI from other conditions.

The high AUC values for CN, dementia, and MCI underscore the exceptional performance of our model in correctly classifying individuals into their respective diagnostic categories. These findings highlight the potential of our approach in supporting early and accurate diagnosis of Alzheimer's disease and related cognitive disorders. The results demonstrate that few-shot training with LLMs like ChatGPT 3.5 can provide significant advancements in diagnostic modeling, offering reliable and efficient tools for clinical applications.

V. Discussion

The ADNI dataset is regularly updated as more participants engage in studies on dementia progression. This work utilized the latest version of the dataset, published in 2024, which includes detailed information on assessments conducted during each patient visit. It was observed that machine learning models achieved high accuracy when predicting the baseline dementia stage using baseline information. However, their performance declined when tasked with predicting the dementia stage for each subsequent visit, highlighting the challenges associated with longitudinal predictions. When exploring neural networks and deep learning, the results of this study underscore the significant potential of integrating numerical and textual data from the ADNI dataset to develop highly accurate predictive models for Alzheimer's disease and related cognitive disorders. By leveraging the extensive cognitive assessments, biomarkers, and demographic data available in ADNI, our approach illustrates how comprehensive datasets can enhance diagnostic

accuracy. The inclusion of varied data types enables a multifaceted analysis, which is essential for understanding the complex progression of Alzheimer's disease. This integrative approach not only improves model performance but also provides deeper insights into the factors driving cognitive decline, reinforcing the value of holistic data utilization in Alzheimer's research.

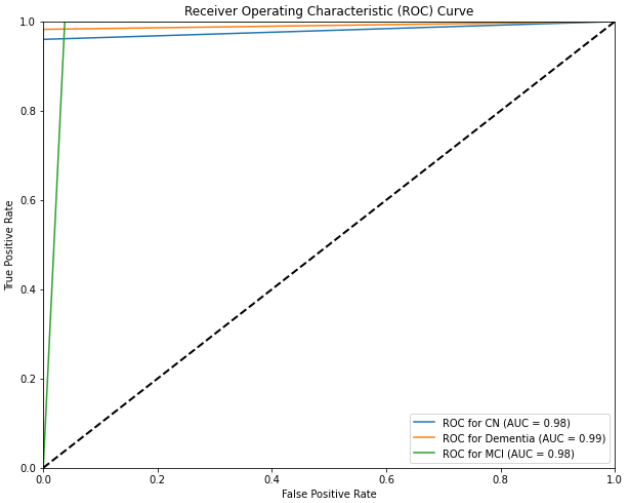

Fig. 9. ROC for the ChatGPT 3.5 after been few-shot training.

In the introduction, we highlighted the projected increase in Alzheimer's disease prevalence, with estimates suggesting that 152 million people globally will be affected by 2050. This alarming trend underscores the urgent need for effective diagnostic tools that facilitate early detection and intervention. The current study addresses this need by leveraging the extensive ADNI dataset, which includes diverse data such as cognitive test scores (e.g., MMSE, RAVLT), biomarker levels (e.g., amyloid-beta, tau proteins), and demographic information (e.g., age, gender, race, education). These features have been shown to significantly contribute to predicting clinical status, as noted by Banerjee (2020) and Tian et al. (2023).

The literature review emphasized the limitations of relying solely on imaging techniques for Alzheimer's diagnosis, such as the high cost and limited accessibility of MRI and PET scans. Previous studies, such as Balakrishnan et al. (2023), predominantly focused on MRI images, underutilizing the full spectrum of data available in ADNI. Our study addresses this gap by integrating numerical and textual data, offering a more holistic and cost-effective approach to Alzheimer's diagnosis. This aligns with findings by Feng et al. (2023), who demonstrated the efficacy of combining imaging and phenotype data with large language models (LLMs).

The application of LLMs, such as ChatGPT 3.5, significantly improved classification performance in this study. LLMs enable rapid processing and analysis of large datasets, achieving high accuracy in classification tasks. Our findings show that LLMs can accurately distinguish between cognitively normal individuals, those with mild cognitive impairment (MCI), and those with dementia. Specifically, the ROC curves for cognitively normal (CN) subjects, dementia, and MCI exhibited AUC values of 0.98, 0.99, and 0.98, respectively. These high AUC values highlight the robustness of LLMs in classifying different stages of cognitive impairment, thereby supporting early and precise diagnoses.

Genetics, a significant contributor to Alzheimer's disease risk (accounting for approximately 70% of overall risk), was also incorporated into our predictive models, as suggested by Raj et al. (2024). The importance of genetic data is underscored in this study, complementing other features. Furthermore, the literature review highlighted the effectiveness of speech analysis and text mining in detecting Alzheimer's disease. Studies by Agbavor and Liang (2022) and Colla et al. (2022) demonstrated the utility of LLMs in analyzing spontaneous speech and text data, aligning with our approach of utilizing LLMs to process numerical and textual data from ADNI.

The contribution of this work lies in demonstrating that LLMs provide not only a rapid and effective approach to classification tasks but also maintain high accuracy, making them valuable tools in clinical settings. This study fills a critical gap in existing research by focusing on the integration of textual and numerical data from ADNI, rather than relying solely on imaging data. By doing so, we offer a cost-effective alternative that reduces dependence on expensive and less accessible imaging techniques. The ability to utilize readily available data to achieve reliable diagnostic outcomes represents a significant advancement, paving the way for more accessible and scalable solutions in Alzheimer's disease detection.

Future research can expand the scope of Alzheimer's prediction beyond image analysis by incorporating a broader range of patient data, such as clinical notes, genetic information, and cognitive test results. This approach has the potential to lead to more comprehensive and accurate prediction models, facilitating earlier detection and enabling more personalized treatment strategies for patients with Alzheimer's disease.

In conclusion, this study highlights the transformative potential of LLMs in utilizing diverse datasets to enhance diagnostic accuracy for Alzheimer's disease. By integrating cognitive assessments, biomarkers, demographic data, and genetic information, our approach offers a comprehensive and efficient diagnostic tool. The findings emphasize the importance of multi-modal data integration and advanced machine learning techniques in addressing the growing challenge of Alzheimer's disease diagnosis and management.

VI. CONCLUSION

This study demonstrates the significant potential of integrating numerical and textual data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset to develop highly accurate predictive models for Alzheimer's disease and related cognitive disorders. By leveraging a comprehensive range of features, including cognitive assessments, biomarkers, demographic information, and genetic data, this approach provides a robust and holistic method for early diagnosis.

The findings underscore the utility of large language models (LLMs), such as ChatGPT 3.5, in processing and analyzing complex datasets. LLMs exhibited exceptional performance in classification tasks, achieving high accuracy rates and rapid processing times. Specifically, the ROC curves for cognitively

normal (CN) subjects, dementia, and mild cognitive impairment (MCI) yielded AUC values of 0.98, 0.99, and 0.98, respectively. These results highlight the efficacy of LLMs in distinguishing between different stages of cognitive impairment, thereby supporting early and precise diagnosis.

This study addresses a critical gap in existing research by focusing on the integration of numerical and textual data rather than relying solely on imaging data. This approach provides a cost-effective alternative, reducing dependence on expensive and less accessible imaging techniques. Utilizing readily available data to achieve reliable diagnostic outcomes represents a significant advancement, paving the way for more accessible and scalable solutions for Alzheimer's disease detection. Additionally, the inclusion of genetic information aligns with findings from previous studies that emphasize the importance of understanding the genetic basis of Alzheimer's disease. By incorporating diverse data types, the proposed models offer a more comprehensive analysis, improving prediction accuracy and supporting targeted interventions.

The transformative potential of combining multi-modal data with advanced machine learning techniques is a key contribution of this work. Integrating ADNI's rich dataset with LLMs offers a promising approach to enhancing diagnostic accuracy and efficiency. Beyond Alzheimer's disease, this work provides a framework for leveraging diverse datasets to address other complex medical conditions. Future research should focus on further integrating various data types and exploring advanced machine learning models to enhance diagnostic capabilities and improve patient outcomes.

REFERENCES

[1] S. Raj, A. Vishnoi, and A. Srivastava, "Classify Alzheimer genes association using Naïve Bayes algorithm," *Human Gene*, 2024, Art no. 201309, doi: https://doi.org/10.1016/j.humgen.2024.201309.

[2] M. Alwuthaynani, M, Z. Abdallah, S, and R. Santos-Rodriguez, "Transfer Learning and Class Decomposition for Detecting the Cognitive Decline of Alzheimer Disease," *arXiv*, 2023, doi: 10.48550/arXiv.2301.13504.

[3] K. Ong, Tzu-iunn *et al.*, "Evidence-empowered transfer learning for Alzheimer's disease," *techrxiv*, 2023, doi: 10.36227/techrxiv.22199635.v1.

[4] A. Aviles-Rivero, I, C. Runkel, N. Papadakis, Z. Kourtzi, and C.-B. Schönlieb, "Multi-Modal Hypergraph Diffusion Network With Dual Prior For Alzheimer Classification," presented at the 25th International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI, Singapore, Singapore, 2022.

[5] M. Memon, Hammad "Early Stage Alzheimer's Disease Diagnosis Method," presented at the 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, Chengdu, China, 2019.

[6] G. Lee, K. Nho, B. Kang, K.-A. Sohn, and D. Kim, "predicting Alzheimer's disease progression using multi-modal deep learning approach," *Scientific Report*, vol. 2019, no. 9, 2019, doi: https://doi.org/10.1038/s41598-018-37769-z.

[7] K. Aderghal, A. Khvostikov, A. Krylov, J. Benois-Pineau, K. Afdel, and G. Catheline, "Classification of alzheimer disease on imaging modalities with deep cnns using cross-modal transfer learning," presented at the 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS, 2018

[8] A. Khvostikov, K. Aderghal, J. Benois-Pineau, A. Krylov, and G. Catheline, "3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies," *ArXiv*, 2018.

[9] G. Awate, "Detection of Alzheimers Disease from MRI using Convolutional Neural Networks, Exploring Transfer Learning And BellCNN," *ArXiv*, vol. abs/1901.10231, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:59336290.

[10] Y. Gao, H. Huang, and L. Zhang, "Predicting Alzheimer's Disease Using 3DMgNet," *arXiv e-prints*, p. arXiv:2201.04370, 2022.

[11] M. Fareed, Muhammad, Sadiq *et al.*, "ADD-Net: An Effective Deep Learning Model for Early Detection of Alzheimer Disease in MRI Scans," *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3204395.

[12] 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, doi: https://doi.org/10.1016/j.abst.2024.08.004.

[13] P. Cao, X. Shan, D. Zhao, M. Huang, and O. Zaiane, "Sparse shared structure based multi-task learning for MRI based cognitive performance prediction of Alzheimer's disease," *Pattern Recognition*, vol. 72 no. 2017, pp. 219–235, 2017.

[14] M. El-Yacoubi, A, S. Garcia-Salicetti, C. Kahindo, A. Rigaud, S, and V. Cristancho-Lacroix, "From aging to early-stage Alzheimer's: uncovering handwriting multimodal behaviors by semi-supervised learning and sequential representation learning," *Pattern Recognition*, vol. 86, pp. 112–133, 2019.

[15] L. ı. Go´mez-Zaragoza´, S. Wills, C. Tejedor-Garcia, J. Mar´ın-Morales, M. Alcan˜iz, and H. Strik, "Alzheimer Disease Classification through ASR-based Transcriptions: Exploring the Impact of Punctuation and Pauses," presented at the Interspeech 2023, Dublin, Ireland, 2023.

[16] B. Dubois *et al.*, "Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria," *Lancet Neurol*, vol. 6, no. 8, pp. 734–746, 2007.

[17] S. Haj Zargarbashi, Soroush and B. Bagher, "A Multi-Modal Feature Embedding Approach to Diagnose Alzheimer Disease from Spoken Language," *arXiv*, 2019.

[18] B. Yu, X, B, Y. Liu, K. Chan, C, C, Q. Yang, and X. Wang, "Skeleton-based human action evaluation using graph convolutional network for monitoring Alzheimer's progression," *Pattern Recognition*, vol. 119, p. 108095, 2021, doi: https://doi.org/10.1016/j.patcog.2021.108095.

[19] S. Doering *et al.*, "Deconstructing pathological tau by biological process in early stages of Alzheimer disease: a method for quantifying tau spatial spread in neuroimaging," *eBioMedicine*, 2024, doi: https://doi.org/10.1016/j.ebiom.2024.105080.

[20] B. Reisberg, S. Ferris, H, M. de Leon, J, and T. Crook, "The global deterioration scale for assessment of primary degenerative dementia," *Am. J. Psychiatry*, 1982.

[21] D. Agarwal, M. Berbís, Álvaro, A. Luna, V. Lipari, J. Ballester, Brito, and I. de la Torre-Díez, "Automated Medical Diagnosis of Alzheimer´s Disease Using an Efficient Net Convolutional Neural Network," *Journal of Medical Systems*, 2023, doi: https://doi.org/10.1007/s10916-023-01941-4.

[22] F. Alghamedy, H, M. Shafiq, L. Liu, A. Yasin, R. Khan, Ali, and H. Mohammed, Sobahi, "Machine Learning-Based Multimodel Computing for Medical Imaging for Classification and Detection of Alzheimer Disease," *Computational Intelligence and Neuroscience*, 2022, doi: https://doi.org/10.1155/2022/9211477.

[23] A. Beck, G. Emery, and R. Greenberg, *Anxiety Disorders and Phobias. A Cognitive Perspective*. New York: Basic Books, 1985.

[24] L. Chen, S. Ho, S, and M. Lwin, O "A meta-analysis of factors predicting cyberbullying perpetration and victimization: From the social cognitive and media effects approach," *New Media & Society*, vol. 19, no. 8, pp. 1-20, 2017.

[25] M. Eysenck, W and M. Keane, T, *Cognitive psychology: A student's handbook (6th ed.)*. New York: Psychology Press, 2010.

[26] E. Nichols and T. Vos, "The estimation of the global prevalence of dementia from 1990–2019 and forecasted prevalence through 2050: An analysis for the global burden of disease (GBD) study 2019," *Alzheimer's Dementia*, vol. 17, no. 10, pp. e105–e125, 2021.

[27] E. Goetzl, J, "Current Developments in Alzheimer's Disease: Developments in Alzheimer's Disease," *The American Journal of Medicine*, 2024, doi: https://doi.org/10.1016/j.amjmed.2024.08.019.

[28] A. Tufail, B, Y. Ma, -K, and Q. Zhang, -N, "Binary classification of Alzheimer's disease using sMRI imaging modality and deep learning," *J. Digit. Imag.*, vol. 33, no. 5, pp. 1073–1090, 2020.

[29] N. Balakrishnan, Bini, P. Sreeja, S, and J. Panackal, Jose, "Alzheimer's Disease Diagnosis using Machine Learning: A Review," *International Journal of Engineering Trends and Technology*, vol. 71, no. 3, pp. 120-129, 2023, doi: https://doi.org/10.14445/22315381/IJETT-V71I3P213.

[30] T. Tuan, Anh, T. Pham, Bao, J. Kim, Young, and J. Tavares, Manuel, R, S, "Alzheimer's diagnosis using deep learning in segmenting and classifying 3D brain MR images," *Int J Neurosci.*, vol. 132, no. 7, pp. 689-698, 2022, doi: 10.1080/00207454.2020.1835900.

[31] M. Dyrba *et al.*, "Robust Automated Detection of Microstructural White Matter Degeneration in Alzheimer's Disease Using Machine Learning Classification of Multicenter DTI Data," *Plos One*, vol. 8, no. 5, 2013, doi: 10.1371/journal.pone.0064925.

[32] J. Liu, M. Li, Y. Luo, S. Yang, W. Li, and Y. Bi, "Alzheimer's Disease Detection Using Depthwise Separable Convolutional Neural Networks," *Computer Methods and Programs in Biomedicine*, vol. 203, 2021, doi: https://doi.org/10.1016/j.cmpb.2021.106032.

[33] M. Taghvaei *et al.*, "Impact of white matter hyperintensities on structural connectivity and cognition in cognitively intact ADNI participants," *Neurobiology of Aging*, vol. 135, no. 2024, pp. 79-90, 2024, doi: https://doi.org/10.1016/j.neurobiolaging.2023.10.012.

[34] D. Hernández, S. Schlicht, Morgan , J. Clarke, Elli, M. Daniszewski, and C. Karch, M, "Generation of a gene-corrected human isogenic iPSC line from an Alzheimer's disease iPSC line carrying the PSEN1 H163R mutation," *Stem Cell Research*, vol. 79, 2024, Art no. 103495, doi: https://doi.org/10.1016/j.scr.2024.103495.

[35] H. Cai *et al.*, "Exploring Multimodal Approaches for Alzheimer's Disease Detection Using Patient Speech Transcript and Audio Data," *arXiv*, 2023, doi: https://doi.org/10.48550/arXiv.2307.02514.

[36] V. Vincze *et al.*, "Linguistic Parameters of Spontaneous Speech for Identifying Mild Cognitive Impairment and Alzheimer Disease," *Computational Linguistics*, vol. 48, no. 1, 2022, doi: https://doi.org/10.1162/COLI.a.00428.

[37] D. Colla, M. Delsanto, M. Agosto, B. Vitiello, and D. Radicioni, P, "Semantic coherence markers: The contribution of perplexity metric," *Artificial Intelligence in Medicine*, vol. 134, no. 102393, 2022, doi: https://doi.org/10.1016/j.artmed.2022.102393.

[38] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24-29, 2019, doi: 10.1038/s41591-018-0316-z.

[39] Y. Wang *et al.*, "Exploiting prompt learning with pre-trained language models for Alzheimer's Disease detection," *arXiv*, 2023, doi: https://doi.org/10.48550/arXiv.2210.16539.

[40] Y. Kim and H. Lee, "PINNet: a deep neural network with pathway prior knowledge for Alzheimer's disease," *Front. Aging Neurosci.*, vol. 15, 2023, doi: https://doi.org/10.3389/fnagi.2023.1126156.

[41] N. Singh, D. Patteshwari, N. Soni, and A. Kapoor, "Automated detection of Alzheimer disease using MRI images and deep neural networks- A review," *arXiv:2209.11282* 2022, doi: https://doi.org/10.48550/arXiv.2209.11282.

[42] A. Essemlali, E. St-Onge, M. Descoteaux, and P.-M. Jodoin, "Understanding Alzheimer disease's structural connectivity through explainable AI," presented at the Machine Learning Research, 2020.

[43] F. Agbavor and H. Liang, "Predicting dementia from spontaneous speech using large language models," *PLOS Digital Health* vol. 1, no. 12, 2022, Art no. e0000168, doi: https://doi.org/10.1371/journal.pdig.0000168.

[44] Y. Feng, J. Wang, X. Gu, X. Xu, and M. Zhang, "Large language models improve Alzheimer's disease diagnosis using multi-modality data," *arXiv:2305.19280* 2023, doi: https://doi.org/10.48550/arXiv.2305.19280.

[45] H. Musto, D. Stamate, I. Pu, and D. Stahl, "Predicting Alzheimer's Disease Diagnosis Risk over Time with Survival Machine Learning on the ADNI Cohort," *arXiv*, 2023, doi: https://doi.org/10.48550/arXiv.2306.10326.

[46] A. Banerjee, "Machine Learning for Health: Personalized Models for Forecasting of Alzheimer Disease Progression," Master, Department of Computing, Imperial College London, London, UK, 1, 2020.

[47] G. Tian, J. Hanfelt, J. Lah, J, and B. Risk, "Mixture of regressions with multivariate responses for discovering subtypes in Alzheimer's biomarkers with detection limits," *arXiv*, 2023, doi: 10.48550/arXiv.2303.00715.

[48] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: https://doi.org/10.1162/neco.1997.9.8.1735.

[49] K. Greff, R. Srivastava, Kumar , J. Koutník, B. Steunebrink, R, and J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222-2232, 2017.

[50] F. Gers, A, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM," presented at the Ninth International Conference on Artificial Neural Networks ICANN 99, Edinburgh, UK, 2000.

[51] I. Goodfellow, Y. Bengio, and A. Courville, "Chapter 6: Sequence Modeling," in *Deep Learning*: MIT Press, 2016.

[52] C. Bishop, M, "Chapter 4: Classification and Loss Functions," in *Pattern Recognition and Machine Learning*: Springer, 2006.