The Journal of Prevention of Alzheimer's Disease 12 (2025) 100169

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Original Article

Using machine learning and electronic health record (EHR) data for the early prediction of Alzheimer's Disease and Related Dementias

Sonia Akter[a,#], Zhandi Liu[b,#], Eduardo J. Simoes[c], Praveen Rao[a,b,*]

[a] Institute for Data Science and Informatics, University of Missouri, USA
[b] Department of Electrical Engineering and Computer Science, University of Missouri, USA
[c] Department of Biomedical Informatics, Biostatics and Medical Epidemiology, University of Missouri, USA

ARTICLE INFO

Keywords:
Alzheimer's disease
Dementias
Machine learning (ML)
Electronic health record data
Early prediction

ABSTRACT

Background: Over 6 million patients in the United States are affected by Alzheimer's Disease and Related Dementias (ADRD). Early detection of ADRD can significantly improve patient outcomes through timely treatment.
Objective: To develop and validate machine learning (ML) models for early ADRD diagnosis and prediction using de-identified EHR data from the University of Missouri (MU) Healthcare.
Design: Retrospective case-control study.
Setting: The study used de-identified EHR data provided by the MU NextGen Biomedical Informatics, modeled with the PCORnet Common Data Model (CDM).
Participants: An initial cohort of 380,269 patients aged 40 or older with at least two healthcare encounters was narrowed to a final dataset of 4,012 ADRD cases and 119,723 controls.
Methods: Six ML classifier models: Gradient-Boosted Trees (GBT), Light Gradient-Boosting Machine (LightGBM), Random Forest (RF), eXtreme Gradient-Boosting (XGBoost), Logistic Regression (LR), and Adaptive Boosting (AdaBoost) were evaluated using Area Under the Receiver Operating Characteristic Curve (AUC-ROC), accuracy, sensitivity, specificity, and F1 score. SHAP (SHapley Additive exPlanations) analysis was applied to interpret predictions.
Results: The GBT model achieved the best AUC-ROC scores of 0.809–0.833 across 1- to 5-year prediction windows. SHAP analysis identified depressive disorder, age groups 80–90 yrs and 70–80 yrs, heart disease, anxiety, and the novel risk factors of sleep apnea, and headache.
Conclusion: This study underscores the potential of ML models for leveraging EHR data to enable early ADRD prediction, supporting timely interventions, and improving patient outcomes. By identifying both established and novel risk factors, these findings offer new opportunities for personalized screening and management strategies, advancing both clinical and informatics science.

1. Introduction

Alzheimer's Disease and Related Dementias (ADRD) are a group of irreversible neurodegenerative disorders that progressively impair cognitive functions, memory, and the ability to perform daily activities [1,2]. Alzheimer's Disease (AD) was first described by a German psychiatrist Alois Alzheimer. It was observed in a patient called Auguste, who died in 1906 due to the loss of cognitive function [3]. Biologically, AD is defined as the pathological deposition of amyloid-beta (Aβ), tau proteins, and neurodegeneration in the brain [4–7]. These pathological changes often emerge 20 years before clinical symptoms appear [8]. As the disease advances, it progresses from Mild Cognitive Impairment (MCI) to severe dementia, with limited treatment options [9–11].

In 2022, over 6 million Americans aged 65 or older were living with AD, and it was the seventh leading cause of death [12]. ADRD imposes a significant burden on patients, families, and society [13,14]. By 2030, the number of AD patients is expected to exceed 75 million and double by 2050 [12,13,15]. In 2022, the treatment for AD and dementia cost $321 billion, along with an additional $271 billion in unpaid caregiving, with projected annual costs exceeding $1 trillion by 2050 [13,16].

Early detection of ADRD is essential, as it allows for intervention before major cognitive decline takes place. Despite extensive research efforts, nearly 99 % of clinical trials failed between 2002 and 2012 to develop successful treatments for ADRD [17]. However, diagnosis at the mild cognitive impairment (MCI) stage instead of late dementia could save the U.S. Healthcare System up to $7.9 trillion [12,18,19].

* Corresponding author at: Department of Electrical Engineering & Computer Science, University of Missouri, Columbia, USA.
E-mail address: praveen.rao@missouri.edu (P. Rao).
# Equal contribution.

https://doi.org/10.1016/j.tjpad.2025.100169
Received 29 January 2025; Received in revised form 4 April 2025; Accepted 7 April 2025

Since 1998, the U.S. Food and Drug Administration (FDA) has approved only six drugs to relieve ADRD symptoms: rivastigmine, galantamine, donepezil, memantine, a combination of memantine and donepezil, and aducanumab [20–22]. These FDA-approved medications are most effective in the early to middle stages of the disease, offering symptomatic relief but not halting disease progression [17,20–23]. The limited success of these treatments highlights the need for more effective therapeutic strategies.

The emergence of machine learning (ML) offers new hope for improving the early detection of ADRD. By leveraging vast datasets from electronic health record (EHR) systems, administrative claims, and neuroimaging, ML models can uncover patterns and insights that may not be apparent through traditional methods. Recent studies have demonstrated the potential of ML in predicting AD incidence and managing other neurological conditions [24–28]. Previous studies have demonstrated the significant adaptability of ML models for different datasets, highlighting their robustness and applicability to other critical conditions such as cardiovascular disease, breast cancer, and prostate cancer [29–31]. Furthermore, a deeper understanding of comorbidities, medication usage, and demographic factors for ADRD can inform public health interventions. For example, depression and cardiovascular disease have been consistently linked to increased ADRD risk [32,33]. Sleep disorders have been associated with impaired cognitive function and amyloid accumulation [34]. Understanding demographic variation in ADRD risk, such as stronger associations in older adults and differences by sex, can inform the design of age and gender-sensitive prevention strategies. Adjusting medication use such as avoiding anticholinergic drugs, which are associated with increased risk of dementia [35], can lead to improved patient outcomes. Identifying high-risk patients earlier can enable personalized interventions for ADRD patients [36].

In our study, we applied six different ML models including Gradient-Boosted Trees (GBT) [37], Light Gradient-Boosting Machine (LightGBM) [38], Random Forest (RF), eXtreme Gradient-Boosting (XGBoost) [39], Logistic Regression (LR) [40], and Adaptive Boosting (AdaBoost) [41] to classify ADRD using de-identified EHR data from the University of Missouri (MU) Healthcare, which is part of the National Patient-Centered Clinical Research Network (PCORnet). By optimizing the ML models for accuracy, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), sensitivity, and F1-scores, and by identifying key predictive features, we aimed for accurate and early diagnosis of ADRD. Our study incorporated a comprehensive range of EHR variables as predictors in the ML models. These variables included demographic factors (e.g., age, race, marital status, sex), vital signs (e.g., systolic and diastolic blood pressure), behavioral risk factors (e.g., smoking history), a broad spectrum of comorbidities, and medical diagnoses. The comorbidities encompassed metabolic conditions (e.g., diabetes, obesity), cardiovascular diseases (e.g., hypertension, heart disease), neurological disorders (e.g. stroke), and psychiatric disorders (e.g., depression, anxiety), and other health conditions (e.g., sleep disorders, kidney disease, vitamin D deficiency). A detailed list of these variables is provided in the Methods section. We tested different prediction windows spanning 1 to 5 years. Despite significant advances in healthcare and data analytics, there remains an opportunity to improve predictive ML models for the early diagnosis of ADRD. Current approaches face challenges such as limited integration of comprehensive EHR data and inadequate interpretability for clinical use. This study aims to address these gaps by employing advanced ML models and SHAP (SHapley Additive exPlanations) analysis to identify key risk factors and improve the precision of ADRD diagnosis.

In the remaining section of the paper, we will delve into the methodology, and the results, and discuss the significance of the findings.

2. Methods

We used de-identified EHR data from MU Healthcare, provided by the MU NextGen BMI. The data were modeled using the PCORnet Common Data Model (CDM) [42], containing longitudinal EHRs encompassing diverse patient characteristics, including demographics, diagnoses, medications, vital signs, and smoking history. This study was approved by the MU Institutional Review Board under protocol IRB2095682.

2.1. Study participants

This retrospective case-control study focused on predicting ADRD diagnosis in adults aged 50 years and older. We began with a cohort of 380,269 patients who met the following criteria (1): aged 40 years or older as of January 1, 2010 (the start date of records in the MU EHR system) (2), admitted between January 1, 2010, and December 31, 2023, and (3) had at least two recorded encounters in MU Healthcare.

2.2. Selection criteria for cases and controls

In this study, we defined ADRD cases using two main criteria for prediction. First, patients who had an ADRD diagnosis based on the International Classification of Diseases, 9th or 10th Revisions (ICD-9/ICD-10), which included codes 331.0, 290.0, 290.1, 290.2, 290.3, 290.4, 290.43, 331.82, 294.1, G30.0, G30.1, G30.8, G31.83, F00, F00.2, F01, F02, and F00.9. These ICD codes cover early onset, late onset, and confirmed ADRD cases (that do not specifically fit into early or late onset categories). We also included patients who were prescribed dementia-related medications commonly used for AD, such as rivastigmine, galantamine, donepezil, memantine, aducanumab, and brexpiprazole. Second, patients must have recorded at least two encounters in the MU EHR system.

Note that brexpiprazole was included in the study due to its FDA approval in 2023 as the first pharmacologic treatment for agitation associated with dementia due to Alzheimer's disease [43]. Although it serves as an adjunct treatment for major depressive disorder and schizophrenia, we included it because of its relevance in managing a common neuropsychiatric symptom in ADRD.

For the control group (non-ADRD), we selected individuals who (1) had no ADRD-related diagnoses based on ICD-9/10 codes or were not prescribed any dementia-related medications, and (2) had at least two recorded encounters in the MU EHR system. Our final dataset comprised of 123,735 unique patients; of these, 4012 were diagnosed with ADRD (cases), while 119,723 had no ADRD or ADRD-related diagnoses (controls).

2.3. Study design

In our study, we divided the time into two sections (1): an observation window and (2) a prediction window. We set an index date for each case as the earliest date of either an ADRD diagnosis or the first prescription of a dementia-related medication. We defined multiple prediction windows (1 year, 2 years, 3 years, 4 years, and 5 years), representing the time windows during which a case had its index date. The observation window spanned from January 1, 2010 (the start date of records in the MU EHR system) to December 31, 2018 for both case and control data. Data from the observation window were exclusively used for training the models, allowing us to evaluate the potential for forecasting ADRD at different time intervals.

We used a fixed control dataset [2010–2019) and different case datasets for each prediction window, as shown in Fig. 1. For the 1-year prediction window, the case data included those whose index date was during [2019–2020). For the 2-year prediction window, the case data included those whose index date was during [2019–2021) and so on. Our approach utilizes large amounts of control data to establish a strong "healthy" reference, in contrast to a shorter pre-onset case window.

Our dataset included variables captured by the EHR system, such as [1] demographic variables (age, race, marital status, and sex) [2], behavioral risk factors (e.g., smoking history), and [3] vital variables, namely, diastolic blood pressure (DBP) and systolic blood pressure

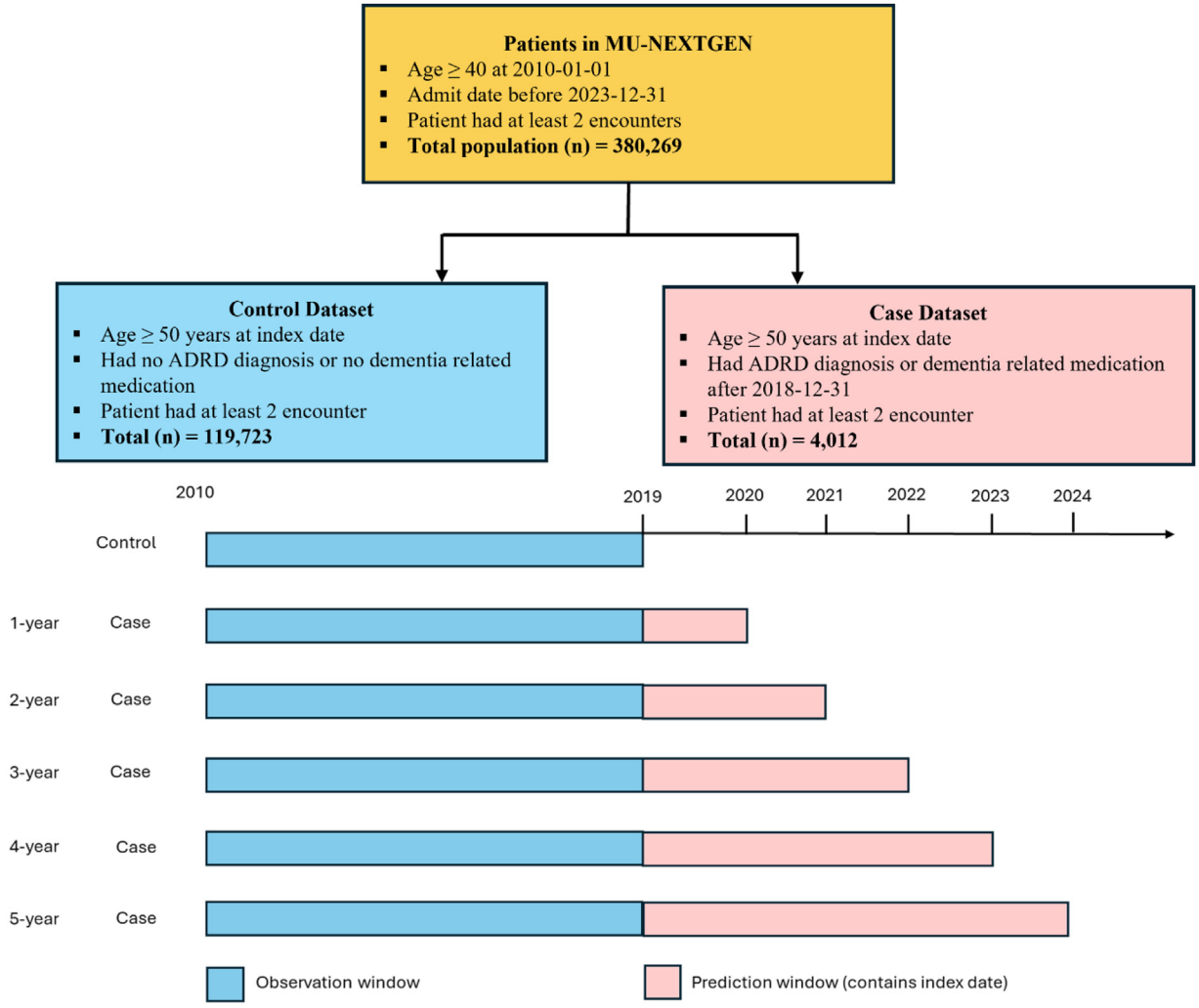

Fig. 1. Flowchart for preparing the case-control study.

(SBP). Additionally, we incorporated comorbidities as risk factors, identified through a thorough review of existing literature [23,44]. The comorbidities and medical diagnoses include diabetes, epilepsy, depression, obesity, stroke, anxiety, hypertension, hyperlipidemia, cardiovascular disease, sleep disorder, headache, periodontitis, concussion, heart disease, sleep apnea, insomnia, kidney disease, cholesterol, vitamin D deficiency, enlarged prostate, bone disease, and depressive disorder.

2.4. Data pre-processing

To address missing data, we excluded the feature variables with a missing rate of 30 % or higher. Patients with more than 20 % missing data were removed from the analysis, while those with less than 20 % missing data were imputed. Initially, our dataset included 4012 unique patients in the positive class (cases), and 232,795 unique patients in the negative class (controls), the final number of cases and controls is shown in Fig. 1.

In the data preprocessing phase, one-hot encoding was applied to all categorical variables. Continuous variables were handled according to their specific characteristics. For example, Age was categorized into five distinct categories: [50,60), [60,70), [70,80), [80,90), and above 90 years. DBP and SBP were categorized into three levels based on clinical thresholds: "normal," "high," and "critically high." Specifically, DBP was categorized as normal (<80 mmHg), high (80–90 mmHg), and critically high (>90 mmHg), while SBP was categorized as normal (<120 mmHg), high (120–140 mmHg), and critically high (>140 mmHg) [45].

The resulting feature vector consisted of binary values, where 0s and 1s indicated the absence or presence of each category. For medical diagnosis or comorbidities conditions, one-hot encoding was used to construct the feature vector. The smoking history was encoded into binary values: never smoker was mapped to 0, while all other smoking categories, including former and current smokers, were mapped to 1. This approach enabled a simplified distinction between non-smokers and those with any smoking history.

The dataset was divided into training and testing sets, with 80 % for training and the remaining 20 % for testing. The training set was used to develop models, while the testing set was employed to assess their performance.

2.5. Model validation and analysis

We trained and tested six different ML classification models: GBT, LightGBM, RF, XGBoost, LR, and AdaBoost to predict ADRD at an early stage. These models were chosen based on their demonstrated effectiveness in prior studies on ADRD prediction and related medical classification tasks. GBT, LightGBM, and XGBoost are widely recognized for their strong predictive performance and ability to handle complex, non-linear relationships in structured healthcare data, making them particularly effective for ADRD risk assessment [46]. RF was used in an Alzheimer's

Table 1
Descriptive Statistics in Control and Case.

		Control (n = 119,723)		Case (n = 4012)	
Variables	Sub-categories	n or mean	% or SD	n or mean	% or SD
Demographic					
Age		75.51	10.18	77.50	9.25
Sex					
	Female	64,072	53.5 %	2415	60.2 %
	Male	55,651	46.5 %	1597	39.8 %
Race					
	White	109,568	92.3 %	3778	94.3 %
	Black/African American	7289	5.2 %	169	4.2 %
	Asian	1058	0.9 %	19	0.5 %
	American Indian/Alaskan Native	208	0.2 %	3	0.1 %
	Native Hawaiian/Other Pacific Island	49	0.03 %	N/A	N/A
	Some Other Race	1103	1 %	24	0.6 %
	Unknown	448	0.5 %	12	0.3 %

study for image classification due to its robustness against overfitting and its ability to handle high-dimensional datasets [47]. LR serves as a baseline model due to its simplicity and interpretability, and it has been frequently applied in ADRD research for modeling disease progression using clinical data [48]. AdaBoost enhances classification performance by improving weak classifiers, particularly in imbalanced datasets, and has been successfully applied in ADRD prediction using deep learning methods [49].

A nested cross-validation approach was employed to optimize and evaluate these models. Each model was incorporated into a pipeline that included a StandardScaler for feature normalization, followed by the respective classifier.

Hyperparameter tuning was conducted using a 5-fold StratifiedKFold inner cross-validation loop with grid search (GridSearchCV). The optimal hyperparameters were then applied in an outer 5-fold StratifiedKFold cross-validation to assess model performance. Predicted probabilities were classified using a 0.5 threshold, and the model performance was measured using metrics such as accuracy, precision, sensitivity, F1-score, AUC-ROC, and specificity, with confusion matrices generated for each fold. Bootstrapping with 1000 iterations was applied to estimate point values. The model with the best performance across all metrics was selected and further evaluated on a hold-out test set to assess generalization.

We applied SHAP (Shapley Additive exPlanations) methods [50] to interpret the predictions of the ML models, we generated SHAP values for each of the five prediction windows (1, 2, 3, 4, and 5 years). These values were used to generate summary plots, providing insights into the model interpretability and highlighting risk factors over time. Specifically, SHAP bar plots and summary plots were generated for the top 12 risk factors in the best-performing model. Features with positive SHAP values were linked to a higher probability of ADRD, whereas those with negative values were associated with a lower risk. The magnitude of each SHAP value reflected the overall importance of that feature with the model.

3. Results

3.1. Sample characteristics

Table 1 shows the descriptive statistics of the case and control groups. Our data set includes 119,723 control individuals and 4012 cases, all aged 50 years and older, covering the period from January 1, 2010, to December 31, 2023. The analysis focused on the key demographic characteristics of both groups. The mean age in the case group (77.50 ± 9.25 years) was higher than in the control group (75.51 ± 10.18 years), indicating an older population in the case group. Female patients were more prevalent than male patients in both groups. Additionally, the White race was predominant in both groups, given the demographics of patients visiting MU Healthcare.

3.2. Performance evaluation of model prediction

We trained six different ML classification models, namely: LR, GBT, LightGBM, XGBoost, RF, and AdaBoost, by applying hyperparameter tuning on the training set. All models were trained using the complete set of predictors, as described earlier, without excluding any variables. These predictors included demographic factors, vital signs, behavioral risk factors, and comorbidities. The models were trained to predict ADRD incidence over 1-year, 2-year, 3-year, 4-year, and 5-year prediction windows. Using the best-performing model for each classifier, we classified the unseen test set into two classes: ADRD (case, positive class) and non-ADRD (control, negative class) patients.

Next, we report on the ML model's performance. As shown in Table 2, the Area under the Curve (AUC) using the GBT model demonstrated the highest performance in predicting ADRD over 5 years. To ensure interpretability and identify the most important predictors influencing ADRD risk, we applied SHAP analysis. A detailed breakdown of feature contributions and their relationship with the model's predictions are presented in the subsection titled "*SHAP Analysis and Model Interpretability*".

The GBT model consistently outperformed the other model in terms of AUC and accuracy in all 5 prediction windows. The GBT model achieved the best AUC score of 0.833, 808, 0.822, 0.821, and 0.809 for predicting ADRD at 5-year, 4-year, 3-year, 2-year, and 1-year prediction windows, respectively. LightGBM and XGBoost also exhibited strong performance, with AUC scores of 0.831 and 0.829, respectively in the 5-year prediction window. In contrast, the LR model had the lowest AUC of 0.782 in the 5-year prediction window. The AUC score consistently increased as the prediction window extended from 1-year to 5-year. Similar results were obtained for other performance metrics, including accuracy, sensitivity, specificity, and F1 scores.

The AUC-ROC curves for the 5-year prediction window using six different ML models are displayed in Fig. 2 (right). For example, the GBT model achieved an AUC score of 0.833, followed by the LightGBM, XGBoost, RF, AdaBoost, and LR with an AUC score of 0.831, 0829, 0.822, 0.818, and 0.782, respectively. Additionally, Fig. 2 (left) displayed the AUC-ROC curves for predictions across 1-year, 2-year, 3-year, 4-year, and 5-year prediction windows using the best model, GBT. This upward trend from 1-year to 5-year suggests that the predictive accuracy of the GBT model helps with the inclusion of longitudinal data. The findings were consistent across other metrics, including specificity and sensitivity.

Table 2
Performance of ADRD Predictive Models ([a] best model according to AUC score).

Prediction Window	Model	Accuracy	AUC	Precision	Sensitivity	Specificity	F-1
1 Year	LR	0.696	0.775	0.970	0.696	0.695	0.801
	GBT	0.978	0.809[a]	0.970	0.978	0.999	0.968
	LightGBM	0.970	0.808	0.964	0.970	0.999	0.957
	XGBoost	0.979	0.799	0.975	0.979	1.000	0.970
	RF	0.978	0.792	0.978	0.978	1.000	0.967
	AdaBoost	0.978	0.782	0.969	0.978	1.000	0.968
2 Years	LR	0.684	0.787	0.965	0.684	0.682	0.789
	GBT	0.974	0.821[a]	0.969	0.974	1.000	0.962
	LightGBM	0.973	0.821	0.963	0.973	0.999	0.962
	XGBoost	0.976	0.818	0.972	0.976	0.999	0.967
	RF	0.974	0.816	0.974	0.974	1.000	0.961
	AdaBoost	0.974	0.796	0.964	0.974	0.999	0.962
3 Years	LR	0.688	0.784	0.961	0.688	0.687	0.789
	GBT	0.974	0.822[a]	0.971	0.971	0.999	0.965
	LightGBM	0.971	0.817	0.964	0.971	0.999	0.958
	XGBoost	0.974	0.798	0.970	0.974	0.998	0.966
	RF	0.971	0.807	0.972	0.971	1.000	0.958
	AdaBoost	0.970	0.800	0.941	0.970	1.000	0.955
4 Years	LR	0.698	0.763	0.956	0.698	0.698	0.795
	GBT	0.970	0.808[a]	0.964	0.970	0.999	0.957
	LightGBM	0.970	0.808[a]	0.964	0.970	0.999	0.957
	XGBoost	0.971	0.807	0.966	0.971	0.999	0.960
	RF	0.969	0.797	0.968	0.969	0.999	0.954
	AdaBoost	0.969	0.795	0.960	0.969	0.999	0.956
5 Years	LR	0.688	0.782	0.955	0.688	0.687	0.787
	GBT	0.970	0.833[a]	0.968	0.970	0.999	0.960
	LightGBM	0.969	0.831	0.966	0.969	0.999	0.958
	XGBoost	0.968	0.829	0.961	0.969	0.999	0.953
	RF	0.968	0.823	0.967	0.968	0.999	0.954
	AdaBoost	0.967	0.818	0.959	0.967	0.999	0.953

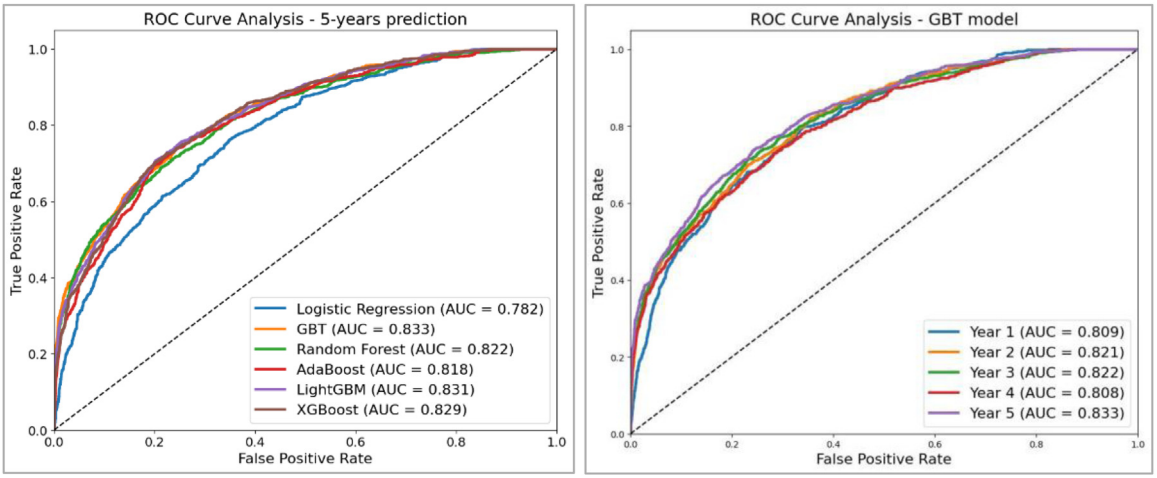

Fig. 2. Performance assessment of ML models in ADRD prediction. (Left) ROC curve analysis for the 5-year prediction window using six different ML models. (Right) ROC curve analysis for predictions across 1-, 2-, 3-, 4-, and 5-year windows. The GBT model, being the best performer, was used for the ROC plot.

3.3. SHAP analysis and model interpretability

We applied the SHAP to identify the key risk factors influencing ADRD prediction and their relationship with outcomes. Given the GBT model's excellent performance, SHAP values provided insight into the model's interpretability. Fig. 3 presents the SHAP analysis across the 1-year to 5-year prediction window, identifying the top 12 features that most influenced the model's predictions. Consistently, features such as a history of depressive disorder, higher age groups (70–80 yrs and 80–90 yrs), history of anxiety, history of sleep apnea, history of heart disease, history of headache, and high DBP were the most significant predictors of ADRD risk.

The SHAP plot provides a detailed breakdown of how each feature affects the model's prediction, highlighting the model's interpretability and the complex relationships between different risk factors. Positive contributions, shown by the red segments, increased the likelihood of an ADRD prediction, while negative contributions, shown by the blue segments, decreased it.

4. Discussion

Our study demonstrated the feasibility of using de-identified EHR data and ML-based models to predict ADRD diagnoses up to 5 years in advance, representing advancements in both informatics and clinical science. A preliminary version of this study was presented as a poster [51] recently and demonstrated the effectiveness of ML for ADRD prediction given a 5-year window. In this study, we have substantially expanded our analysis through rigorous ML model comparison, the use of multiple prediction windows, and interpretability given via SHAP analysis. By comparing six ML models: GBT, LightGBM, RF, XGBoost, LR,

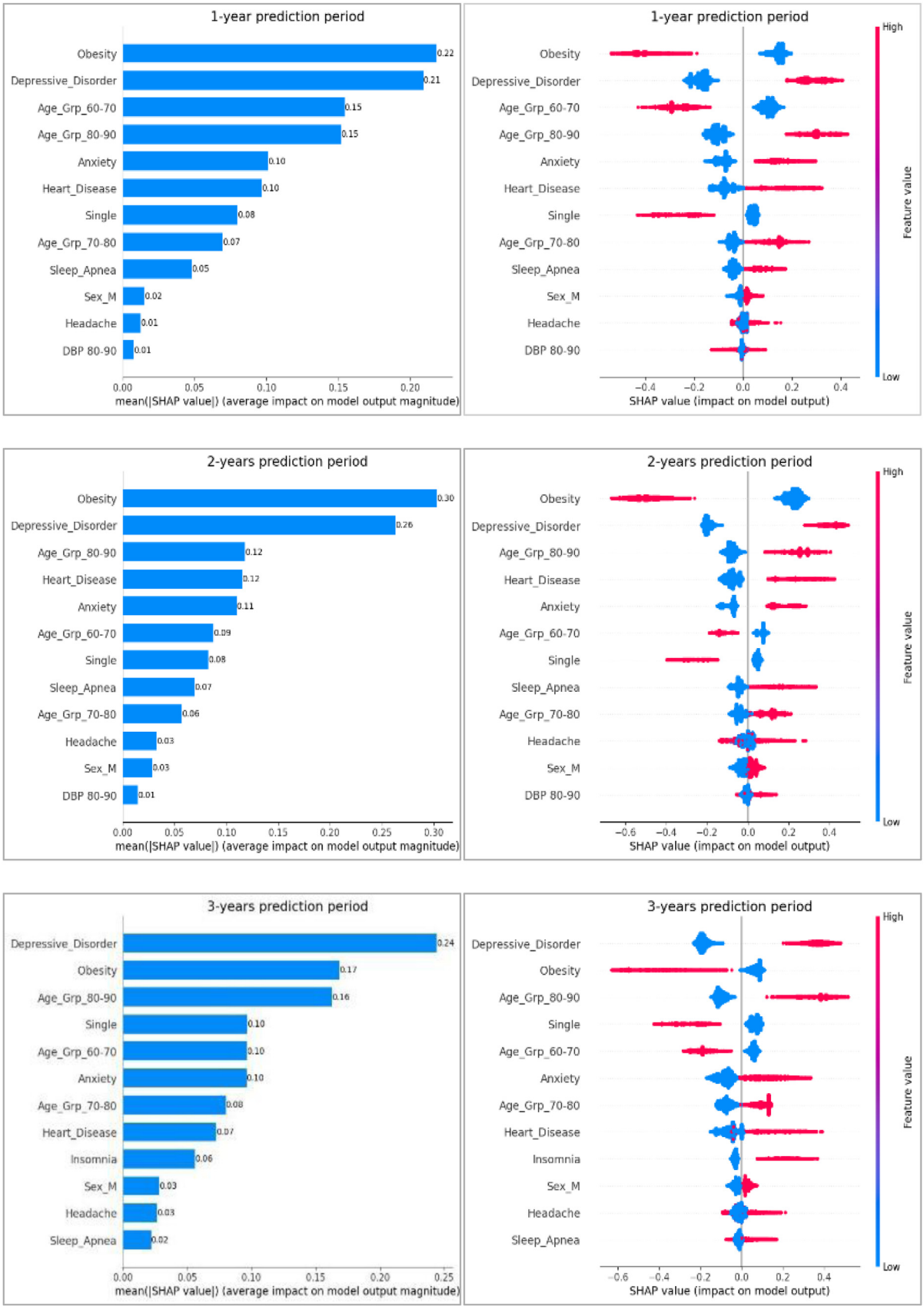

Fig. 3. SHAP plots of the top-12 features for the GBT models (1-year - 5-year prediction windows).

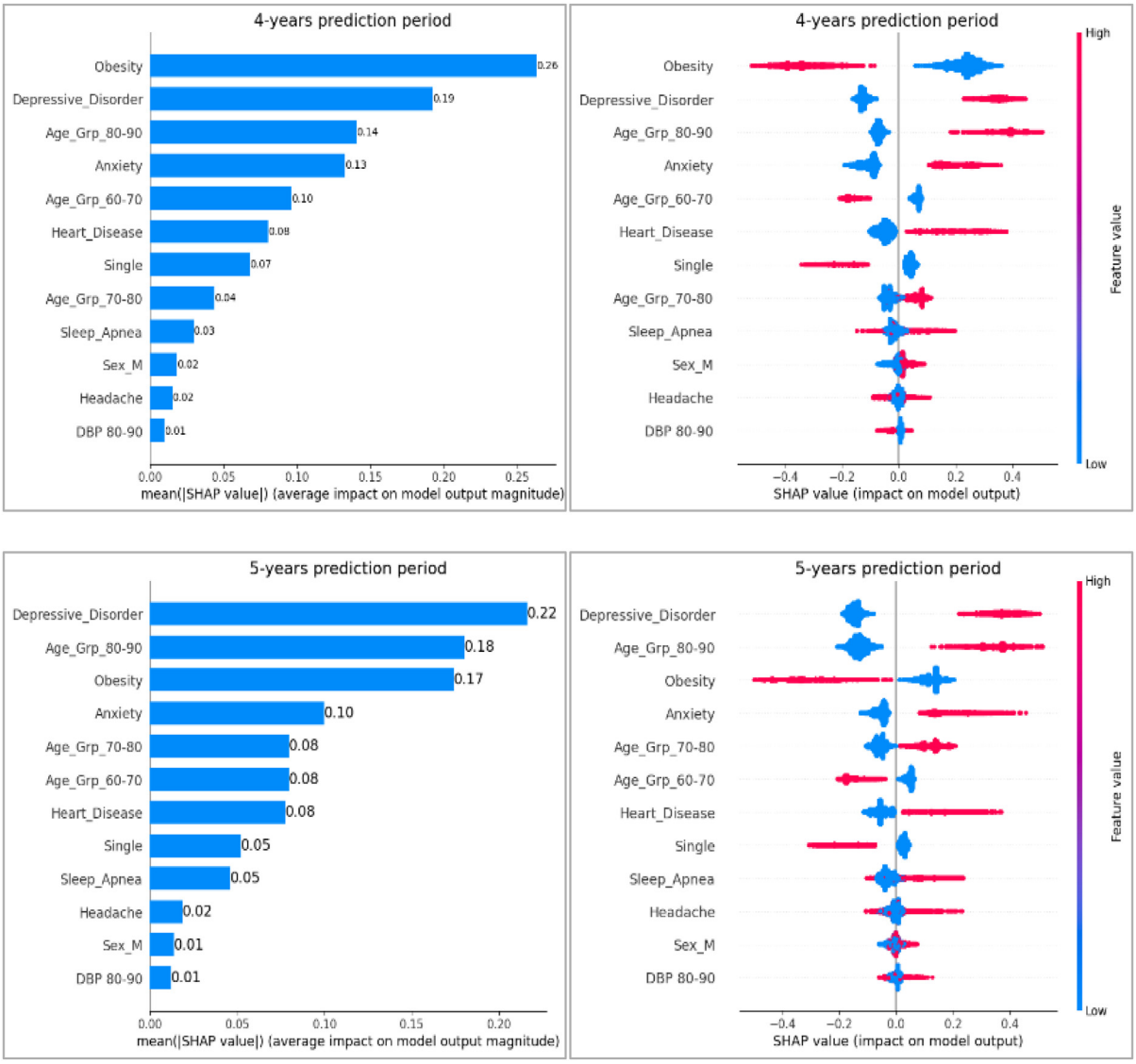

Fig. 3. Continued

and AdaBoost, we established that the GBT model consistently outperformed others across all prediction windows. Notably, the GBT model achieved its highest AUC-ROC score (0.833) for the 5-year prediction window, outperforming shorter prediction windows. While prior studies have shown improved performance with shorter windows due to the inclusion of recent data [23], our findings demonstrate that leveraging a broader historical dataset enhances long-term predicted accuracy. To our knowledge, this is the first study to employ a GBT model incorporating a comprehensive set of known ADRD risk factors to achieve robust prediction 5 years before clinical diagnosis. This novel contribution underscores the potential of ML models in advancing early detection and intervention for ADRD.

This study highlighted the potential of informatics-driven approaches to analyze longitudinal EHR data, addressing methodological gaps in previous studies. By maintaining a fixed dataset across prediction windows, we captured broader trends and reduced overfitting to short-term variations. The design supports scalability and generalizability, providing a framework for developing ADRD prediction models in other healthcare systems. Such innovations advance informatics by refining how longitudinal data can be leveraged for predictive modeling in chronic diseases.

Our study significantly contributes to clinical science by identifying key risk factors for ADRD through SHAP analysis. Established factors such as depressive disorder, heart disease, and age groups 80–90 yrs and 70–80 yrs were confirmed, consistent with their known associations with cognitive decline [52,53]. The SHAP analysis confirmed age as a key ADRD risk factor, with its impact varying across groups. The 80–90 yrs age group had the strongest positive influence, followed by 70–80 yrs age group, aligning with research showing ADRD risk accelerates beyond age 80 due to neurodegeneration, amyloid accumulation, and vascular dysfunction [54]. Cognitive reserve may help individuals in the 60–70 yrs age group compensate for early pathological changes, delaying ADRD diagnosis. The 60–70 yrs age group showed a more variable influence, suggesting age-related effects are not uniform. Some individuals may be in a preclinical ADRD stage, where pathological changes exist but remain asymptomatic. Additionally, differences in healthcare utilization and screening practices may contribute, as older adults (70+) undergo more frequent cognitive assessments, leading to earlier diagnosis. These findings highlight the importance of age-stratified risk assessments and the need for predictive models to account for age-related ADRD risk variations. Further research should explore how comorbidities and lifestyle factors modify ADRD risk at different life stages.

Depression reduces cognitive reserve, while heart disease impacts on vascular health, both of which contribute to ADRD progression. Additionally, our findings highlight sleep apnea and headache as novel predictors, underscoring the potential role of chronic pain and sleep disturbances in increasing ADRD risk. These insights challenge traditional assumptions and create new opportunities to improve clinical screening and intervention strategies.

Incorporating gender-specific risk factors in the study further enhances the clinical understanding of ADRD. For example, insomnia emerged as a top predictor among men, aligning with research linking sleep disturbances to cognitive impairment [55,56]. Additionally, gender differences in healthcare prevalence and their association with ADRD risk were emphasized, suggesting that sex-specific biological mechanisms may influence disease onset. These findings advocate for personalized approaches in ADRD prevention, moving beyond generalized strategies.

Interestingly, the absence of diabetes as a significant predictor in our cohort diverges from previous studies [57–59]. This discrepancy highlights the variability of risk factors across populations and emphasizes the importance of context-specific analysis. Our findings indicate that factors like depressive disorder and heart disease may dominate in some populations, prompting further exploration into how sociodemographic and clinical characteristics modulate ADRD risk.

Informatics advancements are further demonstrated through the integration of SHAP analysis, which enhances model interpretability by identifying feature importance. Unlike traditional black-box ML models, SHAP provides actionable insights into why specific predictions are made, bridging the gap between advanced analytical and clinical utility. For example, the identification of depressive disorder, sleep apnea, and headache as critical predictors underscores the potential of ML-driven tools to uncover previously overlooked relationships in EHR data. This transparency fosters trust among clinicians and paves the way for integrating ML models into routine care.

Clinically, early ADRD prediction enables proactive interventions to delay disease progression. Lifestyle modifications, such as improving sleep hygiene, managing cardiovascular risk factors, and addressing mental health issues, can significantly reduce ADRD risk. Our findings reinforce the multifactorial nature of ADRD, where psychological, vascular, metabolic, and demographic factors interact in complex ways. For instance, while obesity has traditionally been considered a risk factor for cognitive decline, its relationship with ADRD in our study was less straightforward, suggesting that the metabolic health role in dementia may involve unexamined mechanisms warranting further research.

Our findings have important implications for clinical and public health interventions. Depression and cardiovascular disease, identified as top predictors, have been consistently linked to increased ADRD risk, highlighting the need for integrated mental and physical health management in aging populations [32,33]. Similarly, sleep disorders such as sleep apnea, which emerged as a novel risk factor, have been associated with impaired cognitive function and amyloid accumulation [34]. These insights support the development of targeted screening programs and behavioral interventions. Additionally, understanding demographic variation in ADRD risk, such as stronger associations in older adults and differences by sex, can inform the design of age and gender-sensitive prevention strategies. Finally, integrating ML-based predictive tools into clinical decision support systems can help providers in many ways. For example, adjusting medication use such as avoiding anticholinergic drugs, which are associated with increased risk of dementia [35], can lead to improved patient outcomes. Also, identifying high-risk patients earlier can enable personalized interventions such as timely supportive care or supportive care for ADRD patients [36].

The implications of this study extend to healthcare policy and practice. Implementing automated prediction models in clinical settings could facilitate earlier referrals to specialists, enabling timely diagnosis and personalized care plans. These tools also provide researchers with a robust method of identifying high-risk individuals for clinical trials, enhancing the efficiency of recruitment and the reliability of trial outcomes. Moreover, combining short-term and long-term prediction models could optimize patient management by addressing different stages of ADRD development, offering a comprehensive framework for disease monitoring.

This study also sets the stage for advancing clinical science by providing evidence for novel risk factors, such as sleep apnea and headache, which were not prioritized in previous studies [23]. Incorporating these findings into future research could lead to the development of targeted interventions, such as therapies for sleep disorders or chronic pain management, potentially mitigating ADRD risk. Additionally, the ability to generalize our model across prediction windows without expanding the dataset offers a replicable approach for other institutions seeking to utilize EHR data for chronic disease prediction.

By incorporating diverse risk factors of ADRD, this study provides critical insights into the multifaced nature of disease progression. The findings reinforce the importance of using predictive models to guide early detection and intervention for further research aimed at enhancing the precision and application of ML models in clinical practice.

5. Conclusion

This study aimed to develop and validate ML models to predict ADRD using de-identified EHR data from MU Healthcare in a retrospective case-control study design. Among the models evaluated, the GBT model consistently demonstrated superior performance across prediction windows, achieving robust AUC-ROC scores and providing reliable predictions, with its highest performance for the 5-year prediction window. To our knowledge, this is the first study to leverage a GBT model incorporating a comprehensive set of ADRD risk factors to achieve such long-term predictive accuracy.

SHAP analysis provided critical insights into key risk factors, including established predictors such as depressive disorder, age groups 80–90 yrs and 70–80 yrs, heart disease, and anxiety, as well as novel contributors like sleep apnea, and headache. These findings emphasize the multifactorial nature of ADRD risk and highlight the potential of ML models to aid clinicians in identifying high-risk patients. By enabling proactive and targeted interventions, these models can improve patient outcomes, enhance quality of life, and support personalized care strategies.

This approach, by enabling early detection, could optimize treatment strategies, reduce healthcare costs, and advance the support available to both patients and caregivers. Future work will focus on multi-center studies to validate these models in diverse populations, integrate them into routine clinical workflows, and further develop personalized screening and management strategies for ADRD.

Declaration of competing interest

The authors have no conflicts of interest to disclose.

CRediT authorship contribution statement

Sonia Akter: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Zhandi Liu:** Writing – review & editing, Validation, Software, Methodology, Formal analysis, Data curation. **Eduardo J. Simoes:** Writing – review & editing, Funding acquisition. **Praveen Rao:** Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition, Methodology, Investigation, Formal analysis, Conceptualization.

Data availability

We cannot share the raw patient data. The NextGen BMI is responsible for the original data and the MU Hospital does not allow for sharing of the data because of HIPAA regulations. The code is available on GitHub at https://github.com/MU-Data-Science/JPAD-2025.

Funding

Research reported in this publication was partly supported by the National Institute of Diabetes And Digestive And Kidney Diseases of the National Institutes of Health under Award Number P30DK092950 and the University of Missouri. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and the University.

Acknowledgements

We thank MU NextGen Biomedical Informatics (BMI) for providing the data for this study.

References

[1] Mattson MP. Pathways towards and away from Alzheimer's disease. Nature 2004;430(7000):631–9.
[2] Kavitha C, Mani V, Srividhya SR, Khalaf OI, Tavera Romero CA. Early-stage Alzheimer's disease prediction using machine learning models. Front Public Heal 2022;10(March):1–13.
[3] Šerý O, Povová J, Míšek I, Pešák L, Janout V. Molecular mechanisms of neuropathological changes in Alzheimer's disease: a review. Folia Neuropathol 2013;51(1):1–9.
[4] Jack CR, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA Research Framework: toward a biological definition of Alzheimer's disease. Alzheimer's Dement [Internet] 2018;14(4):535–62 Disponible a. doi:10.1016/j.jalz.2018.02.018.
[5] Hammond TC, Xing X, Wang C, Ma D, Nho K, Crane PK, et al. β-amyloid and tau drive early Alzheimer's disease decline while glucose hypometabolism drives late decline. Commun Biol [Internet] 2020;3(1):1–13 Disponible a. doi:10.1038/s42003-020-1079-x.
[6] Chang CH, Lin CH, Lane HY. Machine learning and novel biomarkers for the diagnosis of alzheimer's disease. Int J Mol Sci 2021;22(5):1–12.
[7] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dement 2011;7(3):280–92.
[8] Villemagne VL, Burnham S, Bourgeat P, Brown B, Ellis KA, Salvado O, et al. Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. Lancet Neurol [Internet] 2013;12(4):357–67 Disponible a. doi:10.1016/S1474-4422(13)70044-9.
[9] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dement 2011;7(3):263–9.
[10] Petersen RC, Caracciolo B, Brayne C, Gauthier S, Jelic V, Fratiglioni L. Mild cognitive impairment: a concept in evolution. J Intern Med 2014;275(3):214–28.
[11] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dement 2011;7(3):270–9.
[12] Alzheimer's Association Report. 2022 Alzheimer's disease facts and figures. Alzheimer's Dement 2022;18(4):700–89.
[13] Alzheimer's Association Report. 2023 Alzheimer's disease facts and figures. Alzheimer's Dement 2023;19(4):1598–695.
[14] Hurd MD, Martorell P, Delavande A, Mullen KJ, Langa KM. Monetary costs of dementia in the United States. N Engl J Med 2013;368(14):1326–34.
[15] Zhang R, Simon G, Yu F. Advancing Alzheimer's research: a review of big data promises. Int J Med Inform [Internet] 2017;106:48–56 June 2016Disponible a. doi:10.1016/j.ijmedinf.2017.07.002.
[16] Wimo A, Guerchet M, Ali GC, Wu YT, Prina AM, Winblad B, et al. The worldwide costs of dementia 2015 and comparisons with 2010. Alzheimer's Dement 2017;13(1):1–7.
[17] Cummings JL, Morstorf T, Zhong K. Alzheimer's disease drug-development pipeline: few candidates, frequent failures. Alzheimers Res Ther 2014;6(4):37.
[18] Hampel H, Shaw LM, Aisen P, Chen C, Lleó A, Iwatsubo T, et al. State-of-the-art of lumbar puncture and its place in the journey of patients with Alzheimer's disease. Alzheimer's Dement 2022;18(1):159–77.
[19] Alzheimer's Association Report 2020. Alzheimer's disease facts and figures 2020. Alzheimer's Dement 2020;16(3):391–460.
[20] Barthold D., Joyce G., Ferido P., Drabo E.F., Marcum Z.A., Gray S.L., et al. Pharmaceutical treatment for Alzheimer ' s disease and related dementias: utilization and disparities. 2020;76:579–89.
[21] Yiannopoulou KG, Papageorgiou SG. Current and future treatments for Alzheimer's disease. Ther Adv Neurol Disord 2013;6(1):19–33.
[22] Alzheimer's Association Report 2021. Alzheimer's disease facts and figures 2021. Alzheimer's Dement 2021;17(3):327–406.
[23] Li Q, Yang X, Xu J, Guo Y, He X, Hu H, et al. Early prediction of Alzheimer's disease and related dementias using real-world electronic health records. Alzheimer's Dement 2023:1–13 (January).
[24] Falahati F, Westman E, Simmons A. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. J Alzheimer's Dis 2014;41(3):685–708.
[25] et al. 2011 Xuefeng Chen, Reiter PL, McRee AL. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage 2017;145:137–65.
[26] Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24(1):198–208.
[27] Park JH, Cho HE, Kim JH, Wall MM, Stern Y, Lim H, et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. npj Digit Med [Internet] 2020;3(1) Disponible a. doi:10.1038/s41746-020-0256-0.
[28] Nori VS, Hane CA, Crown WH, Au R, Burke WJ, Sanghavi DM, et al. Machine learning models to predict onset of dementia: a label learning approach. Alzheimer's Dement Transl Res Clin Interv 2019;5:918–25.
[29] Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. J Am Coll Cardiol 2018;71(23):2668–79.
[30] Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Med Inform Decis Mak 2019;19(1):1–17.
[31] Cuocolo R, Cipullo MB, Stanzione A, Ugga L, Romeo V, Radice L, et al. Machine learning applications in prostate cancer magnetic resonance imaging. Eur Radiol Exp 2019;3(1).
[32] Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet. 2020;396(10248):413–46.
[33] Stern Y. What is cognitive reserve? Theory and research application of the reserve concept. Psychol Aging 2017:35–47.
[34] Yaffe K, Falvey CM, Hoang T. Connections between sleep and cognition in older adults. Lancet Neurol 2014;13(10):1017–28.
[35] Gray SL, Anderson ML, Dublin S, Hanlon JT, Hubbard R, Walker R, et al. Cumulative use of strong anticholinergics and incident dementia: a prospective cohort study. JAMA Intern Med 2015;175(3):401–7.
[36] Wang L, Sha L, Lakin JR, Bynum J, Bates DW, Hong P, et al. Development and validation of a deep learning algorithm for mortality prediction in selecting patients with dementia for earlier palliative care interventions. JAMA Netw Open 2019;2(7):E196972.
[37] Friedman J. Greedy function approximation: a gradient boosting machine author [(s): jerome H. Friedmansourcee:thee Annals of Statistics,publishedd by: institute of Mathematical Statisticsstablee URL. Ann Stat [Internet] 2001;29(5):1189–232. Disponible a https://www.jstor.org/stable/2699986 .
[38] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017:3147–55 2017-Decem(Nips).
[39] Chen T., Guestrin C. XGBoost: a scalable tree boosting system. Proc ACM SIGKDD Int Conf Knowl Discov Data Min. 2016;13-17-Augu:785–94.
[40] R.A. Fishers, Sc.D. FRS. The use of multiple measurements in taxonomic problems. 1954;1(1):1–8. Disponible a: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x
[41] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997;55(1):119–39.
[42] National Patient-Centered Outcomes Research Network. Common data model (CDM) specification, version 3 . 0 1 [Internet]. 2015. p. 1–87. Disponible a: https://pcornet.org/data/
[43] U.S. Food and Drug Administration. FDA approves first drug to treat agitation symptoms associated with dementia due to Alzheimer's Disease [Internet]. 2023. Disponible a: https://www.fda.gov/news-events/press-announcements/fda-approves-first-drug-treat-agitation-symptoms-associated-dementia-due-alzheimers-disease
[44] Matthew Baumgart, Heather M, Snyder Maria CCarrillo, Fazio Sam, Kim Hye, Johns Harry. Summary of the evidence on modifiable risk factors for cognitive decline and dementia: a population-based perspective. Alzheimer's Dement Diagnosis, Assess Dis Monit 2015;11(6):718–26.
[45] Schwartz GL, Sheps SG. A review of the Sixth Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Curr Opin Cardiol [Internet] 1999;14(2):161–8. Disponible a https://www.ncbi.nlm.nih.gov/books/NBK9633/table/A32/ .
[46] Grueso S, Viejo-Sobera R. Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. Alzheimer's Res Ther 2021;13(1).
[47] Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Front Aging Neurosci 2017;9(OCT):1–12.
[48] Kumar S, Oh I, Schindler S, Lai AM, Payne PRO, Gupta A. Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review. JAMIA Open 2021;4(3):1–10.
[49] Javeed A, Dallora AL, Berglund JS, Anderberg P. An intelligent learning system for unbiased prediction of dementia based on autoencoder and adaboost ensemble learning. Life 2022;12(7).
[50] Lundberg Scott M, Lee Su-In. A unified approach to interpreting model predictions. Nips 2017;16(3):426–30.
[51] Akter S., Liu Z., Simoes E.J., Rao P. Machine learning for early prediction of Alzheimer's disease and related dementias using electronic health record (EHR) data. En Pittsburg, Penn: American Medical Informatics Association (AMIA) Informatics Summit, March 2025, Pittsburgh, PA (Poster).
[52] Zlokovic BV. Neurovascular pathways to neurodegeneration in Alzheimer's disease and other disorders. Nat Rev Neurosci 2011;12(12):723–38.
[53] Corrada MM, Brookmeyer R, Paganini-Hill A, Berlau D, Kawas CH. Dementia inci-

dence continues to increase with age in the oldest old the 90+ study. Ann Neurol 2010;67(1):114–21.
[54] Stern Y. Cognitive reserve in ageing and Alzheimer's disease. Lancet Neurol [Internet] 2012;11(11):1006–112 Disponible a. doi:10.1016/S1474-4422(12)70191-6.
[55] Dzierzewski JM. Insomnia and subjective cognitive decline in older adults: avenues for continued investigation and potential intervention. Sleep 2022;45(11).
[56] Ni Y, Yu M, Liu C. Sleep disturbance and cognition in the elderly: a narrative review. Anesthesiol Perioper Sci [Internet] 2024;2(26) Disponible a. doi:10.1007/s44254-024-00066-2.
[57] Huang CC, Chung CM, Leu HB, Lin LY, Chiu CC, Hsu CY, et al. Diabetes mellitus and the risk of Alzheimer's disease: a nationwide population-based study. PLoS One 2014;9(1).
[58] Doroszkiewicz J, Mroczko J, Winkel I, Mroczko B. Metabolic and immune system dysregulation: unraveling the connections between Alzheimer ' s Disease, diabetes. Inflam Bowel Dis, Rheumatoid Arthri 2024.
[59] Kumar V, Kim SH, Bishayee K. Dysfunctional glucose metabolism in Alzheimer's Disease onset and potential pharmacological interventions. Int J Mol Sci 2022;23(17).