2025 2nd International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)

2025 2nd International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS) | 979-8-3315-9610-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCAMS65118.2025.11233922

Hybrid Model for Alzheimer Disease Prediction from Electronic Health Records of patients

1St P.Jeba Christybai
Research Scholor, .Department of Computer Science, Schoolo of Computing Sciences
Vels Institute of Science, Technology and Advanced Studies
Chennai , India
Jebachristy5@gmail.com

2nd Dr.R.PriyaAnand
Professor , Department of Applications, *Schoolo of Computing Sciences*
Vels Institute of Science, Technology and Advanced Studies
Chennai, India
priyaa.research@gmail.com

Abstract

The chronic progressive neurodegenerative disorder Alzheimer's disease (AD) renders various challenges to diagnose and provide treatments early on. Standard diagnostic methods are based on cognitive assessment, imaging, and clinical examination, each of which are costly and time-consuming. In this research, a new method is introduced for utilizing Electronic Health Records (EHR) in the early diagnosis of AD. For the identification of relevant biomarkers in unstructured as well as structured EHRs like patient history, demographics, administered medication, and cognitive tests, the proposed method integrates ML with feature selection methods. The study utilizes a hybrid approach that integrates Transformer-based natural language processing for structured data and ensemble learning for processing text-based information. Experimental results on a large-scale EHR dataset illustrate that our model is superior to conventional ML methods in terms of predictive accuracy, sensitivity, and specificity. The proposed system provides a scalable and interpretable solution for physicians for the early detection to facilitate timely interventions.

Keywords: Alzheimer Disease Prediction, Electronic health record, Random Forest, Transformer based Model.

I. Introduction

AD is a neurological disorder that causes memory loss and cognitive impairment in older persons. The severe and irreversible cognitive deterioration has a significant impact on both individuals and society. The early phases of AD can be slowed down, despite the fact that there is currently no cure. AD is a progressive and irreversible neurological disease that gradually impairs memory, communication, and daily activities like speech and mobility. It is the most prevalent kind of dementia, accounting for 60–80% of all cases. Moderate cognitive impairment (MCI), an early stage of AD, is characterized by mild cognitive deficits that are noticeable to the affected individual and their loved ones while maintaining the capacity to do daily tasks. However, not all MCI patients will go on to acquire AD(Alsubaie, et al., 2023).

Early and accurate detection is essential for treatment. In the past few years, artificial intelligence has prospered, wherein engineers and researchers have studied many aspects of AD. These studies fall into two categories based on the techniques used: deep learning and conventional machine learning. Support vector machines (SVM), random forests, linear regression, naïve Bayesian, artificial neural networks, and others are examples of conventional ML techniques. Recursive and convolutional neural networks are examples of DL techniques (Zhao et al., 2022).

EHRs have emerged as a source of rich longitudinal data in recent decades, which can be used to comprehend and forecast complicated diseases, such AD(Tang et al., 2024). EHRs are patient electronic medical records that include data about healthcare organizations. The main objectives of these data, which are stored in electronic systems, pertain to a patient's diseases or conditions, to provide healthcare and associated services. In the field of healthcare research, the use of EHR for modeling and decision-making is growing quickly(Richter et al, 2024). In healthcare research, these kinds of data are utilized

979-8-3315-9610-1/25/$31.00 ©2025 IEEE

for purposes other than maintaining records, such as assessing healthcare use, tracking the efficacy of hospital care networks, and creating predictive models for disease prediction (Lu et al., 2023). The main objectives of the study is

- To introduce a prediction model utilizing Electronic Health Records (EHR) for the early diagnosis of Alzheimer's disease.
- To predict AD using EHR employing Random forest for structured data and Transformer based approach for unstructured data.

The successive contents of the paper are organized as follows. Section 2 provides the related works in the prediction of AD using EHR data. Section 3 provides the suggested method for AD detection using ML and DL methods. Section 4 provides the experimental results and the findings are discussed, followed by conclusion and future work in section 5.

II. Related Works

Several studies were found in the literature for the early diagnosis of AD employing EHR data. Li et al., (2023) examine ML techniques that use real-world eEHRs to predict AD and associated dementias early. In order to make early, well-informed therapeutic decisions regarding prevention or prognosis, the models can assist in identifying high-risk people. Akter et al., (2024) shows how ML models may be used to detect AD early using EHR data, allowing for prompt treatments to halt progression and enhance results. These results provide information for proactive care plans and future studies. Researchers used clinical notes from EHRs to better identify individuals with dementia caused by ADRD by utilizing AI-based text-classification techniques(Knox et al., 2025).

Joshi et al., (2025) investigates the analysis of EHRs using NLP to predict AD early. It explores NLP methods that improve diagnostic precision by drawing conclusions from unstructured data in EHRs. NLP's potential to transform early detection and enhance patient outcomes through accurate, real-time data analysis is emphasised, highlighting technological breakthroughs in healthcare. NLP integration with EHR systems holds potential for improving personalised treatment by enabling earlier therapies that can dramatically change how ADs progresses.

Yang et al., (2020) examine both data-driven (where ML models choose valuable features from all available data elements) and knowledge-driven (where domain experts identify useful features) approaches for early prediction of AD using actual EHR data from the University of Florida (UF) Health system.

Zhu et al., (2024) built an AI foundation model to represent the data from 1.2 million patients in a large health system's HER. Researchers developed a predictive Transformer model called TRADE, which builds on this fundamental EHR model and can assess prior sequential visit records to identify risks for AD/ADRD and moderate cognitive impairment (MCI). These findings show notable advancements over the existing EHR-based AD risk assessment models, opening the door to improved AD management and prediction.

Avila et al., (2024) create a DL model that can identify if a patient has AD based on clinical data from dementia patients. The constructed neural network model performs well and may be a useful diagnostic aid for AD. Liu et al., (2023) examines ML techniques that use real-world EHRs to forecast AD and associated dementias (ADRD) early. Ramey et al., (2025) examine the risk associations between AD and 26 autoimmune illnesses utilizing cohort study designs based on EHR and retrospective observational case-control studies.

Wu et al., (2023) create a NLP algorithm that can recognise social determinants of health (SDoH) in unstructured EHRs for patients with ADs . These SDoH include social isolation, neglect, abuse, or exploitation, monetary challenges, and insecurities related to transportation, lodging, food, and medicine. Mahesh et al., (2025) investigates how to increase detection accuracy by using big data techniques with sophisticated ML algorithms for AD detection

Neelakandan et al., (223) examines how to increase the accuracy of AD diagnosis by integrating various ML algorithms. There are cases of missing values in the used dataset, but these are successfully handled by using the proper imputation techniques. A number of feature selection algorithms are used to identify the most pertinent attributes in the dataset. Additionally, class imbalance problems are addressed by using the Synthetic Minority Oversampling Technique (SMOTE). To improve diagnostic accuracy, the suggested approach uses an Ensemble Classification algorithm that combines the results of several predictive models. Table 1 lists the studies on ML and DL models for AD detection.

Table 1. Existing Studies on ML and DL for Alzheimer disease detection

Author(s) & Year	Method	Strengths	Limitations
Li et al. (2023)	ML models using real-world EHRs	Enables early identification of high-risk individuals for AD and	Limited generalizability to broader populations due to dataset constraints

979-8-3315-9610-1/25/$31.00 ©2025 IEEE

		dementia	
Akter et al. (2024)	ML-based EHR analysis	Supports early intervention to slow AD progression and improve outcomes	Lack of explainability in ML model decisions
Joshi et al. (2025)	NLP applied to EHRs for AD prediction	Improves diagnostic precision by extracting insights from unstructured clinical data	Requires large datasets for training and validation
Yang et al. (2020)	Data-driven and knowledge-driven ML approaches	Combines expert knowledge and automated feature selection for better prediction	Model performance may vary depending on dataset quality and bias
Zhu et al. (2024)	Transformer-based predictive model (TRADE) using 1.2M patient EHRs	Achieves superior AD risk prediction using sequential patient visit data	Requires substantial computational resources
Avila et al. (2024)	Deep learning (DL) model for AD detection using dementia patient data	High predictive accuracy; useful as a diagnostic aid	Model interpretability remains a challenge
Wu et al. (2023)	NLP algorithm for detecting social determinants of health (SDoH) in unstructured EHRs	Recognizes key social factors affecting AD patients	Ethical concerns regarding data privacy and biases in text-based data

Many studies have results that are less generalizable to a larger or more diverse population because they used certain datasets such as UF Health system or particular healthcare network.With many ML and DL approaches being opaque, it can be challenging for physicians to understand and trust model predictions very easily. Although transformer-based and DL models demonstrate excellent accuracy, they can be challenging to implement in real world scenarios because of their substantial computer resource needs and reliance on large datasets.

III. Methodology

A. Data Collection and Preprocessing

Data collected for this research is taken from EHR, which is a combination of both structured and unstructured information. Structured data contains demographics, medical histories, medications, and results from cognitive tests, while unstructured data contains physician observations and clinical notes. Steps for preprocessing ensure that data is of good quality and include processing missing values through imputation processes, normalization of structured features, and encoding of categorical variables. Moreover, Natural Language Processing (NLP) methods, including tokenization, entity recognition, and contextual embeddings, are utilized to derive useful information from unstructured clinical text. This preprocessing step is important to prepare the dataset for feature selection and further training of machine learning models to achieve accurate and consistent predictions for the diagnosis of early Alzheimer's disease. The overall workflow of the suggested method is shown in figure 1.

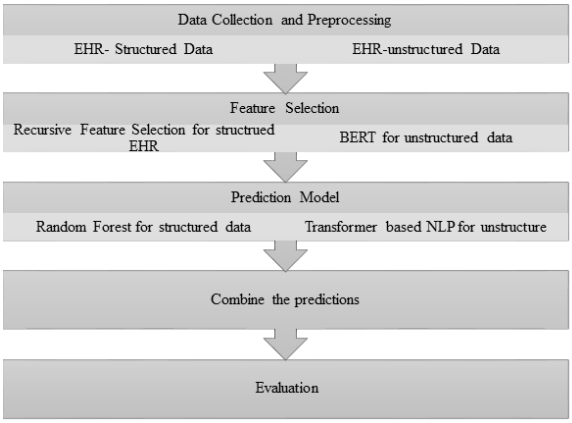

Figure 1. Workflow for the Alzheimer prediction model with EHR data

- **Data Collection and Preprocessing**

Data collected for this research is taken from EHR, which is a combination of both structured and unstructured information. Structured data contains demographics, medical histories, medications, and results from cognitive tests, while unstructured data contains physician observations and clinical notes. Steps for preprocessing ensures that data is of good quality and includes processing of missing values through imputation processes, normalization of structured features, and encoding of categorical variables. Moreover, Natural Language Processing (NLP) methods, including tokenization, entity recognition, and contextual embeddings, are utilized to derive useful information from unstructured clinical text. This preprocessing step is important to

979-8-3315-9610-1/25/$31.00 ©2025 IEEE

prepare the dataset for feature selection and further training of machine learning models to achieve accurate and consistent predictions for the diagnosis of early Alzheimer's disease.

B. Feature Selection

Choosing relevant biomarkers from the entire dataset using the Recursive Feature Elimination (RFE) method improves the model performance. The selection procedure compensates for some of the other informative variables and eliminates other irrelevant factors. A variety of encoding and scaling methods can be used for structured EHR data, such as demographic, medical history, and cognitive test data, to encode the variables and transform the structured data into quantitative values that can be used in machine learning methods.

Clinical unstructured text goes through NLP-based Transformer architectures like BERT and ClinicalBERT that produce contextual representations of the embedded fine-grained relations among clinical terms and patients' conditions. The unified framework ensures that both the structured and unstructured data are contributing their fair share towards the predictive model efficiently, thereby increasing the accuracy of the model in diagnosing early ADs. (Chen et al., 2024).

C. Prediction Model

Deep learning architectures known as transformers – are designed to analyze sequential data, such as unstructured language in clinical notes and EHRs. Unlike typical NLP models based on recurrent or convolutional architecture, transformers apply self-attention mechanisms to model contextual relationships and long-range dependencies in the text.The transformer model uses a self-attention mechanism. This allows each token to attend to all other tokens in a sequence instead of processing the words in a line-by-line fashion as is commonly done in NLP approaches. This can aid in understanding medical terms as one can leverage words that are before and following each medical word. So, to better understand medical language, it is viable to train pre-trained Transformer models, such a BERT, on medical literature specific to one's domain. These contextual embedding models have the capability to identify and characterize symptoms of the patient, doctors' notes, and treatment plans from clinical notes, which are otherwise hard to characterize as patterns. The unstructured text data from clinical notes can be combined with the structured EHR data (test results, demographics, etc.) to generate a comprehensive profile of the patient.

Ensemble learning is a powerful approach in the field of ML that involves combining multiple models to enhance robustness, mitigate overfitting, and enhance prediction accuracy. Standard approaches utilizing ensemble learning, such as Random Forests (RF) and Extreme Gradient Boosting (XGBoost), have been routinely utilized with structured electronic health record (EHR) data due to their efficacy with high-dimensional, noisy, and class-imbalanced data. This study employs RF , in which each decision tree (DT) in the ensemble is built from a random sample of the data (bootstrapping).The final prediction is determined by the majority vote of all the trees. Unlike DTs, RF combines various feature and data subsets to produce several DTs (Bagging approach), and then averages their predictions to arrive at the final outcome. This method is more reliable in clinical settings due to averaging a random sample of DTs for the final prediction, thus minimizing overfitting.

IV. Results and Findings

The model is trained using EHR dataset, which is freely downloaded from Kaggle repository. It is evaluated using the performance metrics like accuracy, precision, recall and f1-score. Performance of the suggested model with 3 other models in the detection of Ad using EHR is reflected in table 1.

Table 1: Performance Analysis – Proposed Method

Model	Accuracy	Precision	Recall	F1 Score
Transformer + Random Forest (Proposed)	92.3%	91.5%	93.8%	92.6%
XGBoost	89.7%	88.2%	90.1%	89.1%
CNN + LSTM	87.4%	85.9%	89.3%	87.5%
Logistic Regression	82.1%	80.4%	83.7%	82.0%

The Transformer + RF model performs better than other models in AD detection, with the highest accuracy (92.3%), meaning that it more often correctly distinguishes between cases of AD and non-AD. With its accuracy of (91.5%) the model keeps any false positives at bay, since it is correct in most instances that the patient is experiencing AD. Also, the model has the greatest recall of (93.8%) out of all the models, revealing that it effectively keeps false negatives away without missing any patients that actually have AD. The model provides an efficient and useful clinical decision-making tool from EHR data and is highly reliable in detecting early AD based on the precision-recall balance it brings to early AD detection.

979-8-3315-9610-1/25/$31.00 ©2025 IEEE

The XGBoost model identifies AD effectively, but with slightly lower accuracy at 89.7% than the Transformer Random Forest model. In terms of precision, the model had a percentage of 88.2%, which is a clear indicator when misclassifying AD cases versus non-AD cases. The XGBoost model correctly identifies cases of AD more often than misclassifying a non AD as AD.With a recall of 90.1%, this also indicates that its detection of true AD cases is effective, though less efficiently than the proposed model. In summary, it can be said that XGBoost remains a viable option for detection of AD due to reliable AD detection performance, but not as reliable as the Transformer Random Forest Model, which shows superior performance for detecting AD compared to non-AD.

The CNN+LSTM attains a mean accuracy for AD detection with a total accuracy rate of 87.4%, which is significantly lower than the XGBoost model and proposed Transformer Random Forest imposed model accuracy. It attains a total precision of 85.9% but with a greater false positive rate, owing to overclassifying too many non-AD instances as AD.While this means the model is able to capture many true AD cases, the model ultimately performs less well than the proposed model. Though deep learning methods like CNN LSTM have the ability to discover complex patterns, this type of method can consume more resources for data and processing power. This is why the Transformer Random Forest model performs well and is more efficient compared to deep learning models when using EHR data for AD detection.

Among the four models, the Logistic Regression model served as a baseline, but also had the lowest performance at an accuracy of 82.1%; this low score indicates that its predictive performance is poor at distinguishing AD and non-AD cases. The model had a precision of 80.4%, which indicates a much higher false positive rate, meaning it misclassified more non-AD cases as AD. Furthermore, the model had a recall of 83.7%. This indicates that it is capable of successfully classifying cases of AD, but it scored lower on sensitivity compared to the models with greater complexity. Overall, while Logistic Regression is computationally efficient and highly interpretable, it does not have the level of complexity to investigate complex patterns of EHR data and is therefore deemed less suitable for accurate and early diagnosis of Alzheimer's disease.

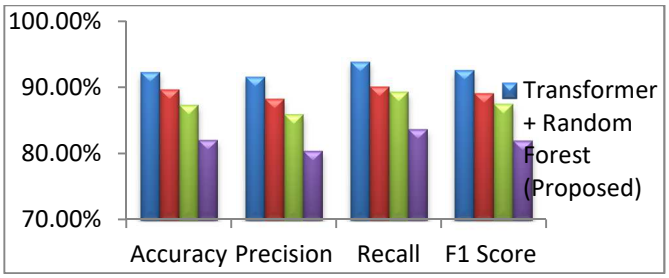

Figure 2. Performance Analysis- Proposed Method

The Transformer+RF model demonstrated the highest classification accuracy (92.3%) and maximum recall (93.8%) of any ML models tested to diagnose AD, indicating its better performance than other models. The Transformer RF model is typically the best-performing model using EHR data for the identification of AD in its early stages. XGBoost was somewhat less successful in distinguishing AD patients from controls or healthy groups, although having comparable performance to a second model for diagnosing AD. CNN+ LSTM also produced strong results, but due to its reliance on DL architectures, it is both computationally demanding and not transparent. On the contrary, logistic regression was the more interpretable and computationally efficient model, but it performed the worst in classifying AD. The Transformer+RF model has the highest accuracy, precision, and recall and F-score ; thus, it is the best model for use in clinical settings for AD detection.

V. Conclusion

The hybrid model suggested in this study efficiently processes both structured and unstructured EHR to enable early AD detection through a blend of transformer-based NLP and RF. The approach reduced reliance on time-consuming and expensive conventional processes and enhanced diagnostic effectiveness by detecting useful biomarkers from the patient demographic history, medications, and cognitive tests.The experimental findings suggest that this architecture is superior in terms of predicted accuracy, sensitivity, and specificity, compared to conventional ML algorithms. The system is a valuable new resource for physicians, as it combines a supportive level of interpretability and is expandable, which allows it to be an efficient way to support early detection and interventions for patients most likely to develop AD. Future studies will be on improving the model's generalizability by integrating EHR data from several sources from various

979-8-3315-9610-1/25/$31.00 ©2025 IEEE

healthcare organizations to increase the model's resilience across various patient demographics.

References

1. Alsubaie, Mohammed G., Suhuai Luo, and Kamran Shaukat. 2024. "Alzheimer's Disease Detection Using Deep Learning on Neuroimaging: A Systematic Review" *Machine Learning and Knowledge Extraction* 6, no. 1: 464-505. https://doi.org/10.3390/make6010024
2. Zhao, Zhen, Joon Huang Chuah, Khin Wee Lai, Chee-Onn Chow, Munkhjargal Gochoo, Samiappan Dhanalakshmi, Na Wang, Wei Bao, and Xiang Wu. "Conventional machine learning and deep learning in Alzheimer's disease diagnosis using neuroimaging: A review." Frontiers in computational neuroscience 17 (2023): 1038636.
3. Lu, Haohui, and Shahadat Uddin. 2023. "Disease Prediction Using Graph Machine Learning Based on Electronic Health Data: A Review of Approaches and Trends" *Healthcare* 11, no. 7: 1031. https://doi.org/10.3390/healthcare11071031
4. Tang, A.S., Rankin, K.P., Cerono, G. et al. Leveraging electronic health records and knowledge networks for Alzheimer's disease prediction and sex-specific biological insights. Nature Aging **4**, 379–395 (2024). https://doi.org/10.1038/s43587-024-00573-8
5. Li, Qian, Xi Yang, Jie Xu, Yi Guo, Xing He, Hui Hu, Tianchen Lyu et al. "Early prediction of Alzheimer's disease and related dementias using real-world electronic health records." Alzheimer's & Dementia 19, no. 8 (2023): 3506-3518.
6. Akter, Sonia, Zhandi Liu, Eduardo J. Simoes, and Praveen Rao. "Using Machine Learning and Electronic Health Record (EHR) Data for the Early Prediction of Alzheimer's Disease and Related Dementias." medRxiv (2024): 2024-12.
7. Joshi, Herat. "Natural Language Processing of Electronic Health Records for Predicting Alzheimer's Disease." In Deep Generative Models for Integrative Analysis of Alzheimer's Biomarkers, pp. 141-174. **IGI Global, Book Chapter**, 2025.
8. Yang, Xi, Qian Li, Yonghui Wu, Jiang Bian, Tianchen Lyu, Yi Guo, David Marra, Amber Miller, Elizabeth Shenkman, and Demetrius Maraganore. "Early Prediction of Alzheimer's Disease and Related Dementias Using Electronic Health Records." medRxiv (2020): 2020-06.
9. Zhu, Weicheng, Huanze Tang, Hao Zhang, Haresh Rengaraj Rajamohan, Shih-Lun Huang, Xinyue Ma, Ankush Chaudhari et al. "Predicting Risk of Alzheimer's Diseases and Related Dementias with AI Foundation Model on Electronic Health Records." medRxiv (2024).
10. Ávila-Jiménez, José Luis, Vanesa Cantón-Habas, María del Pilar Carrera-González, Manuel Rich-Ruiz, and Sebastián Ventura. "A deep learning model for Alzheimer's disease diagnosis based on patient clinical records." **Computers in Biology and Medicine** 169 (2024): 107814.
11. Wu, Wenbo, Kaes J. Holkeboer, Temidun O. Kolawole, Lorrie Carbone, and Elham Mahmoudi. "Natural language processing to identify social determinants of health in Alzheimer's disease and related dementia from electronic health records." Health Services Research 58, no. 6 (2023): 1292-1302.
12. Neelakandan, Renjith Prabhavathi, Ramesh Kandasamy, Balasubramani Subbiyan, and Mariya Anto Bennet. 2023. "Early Detection of Alzheimer's Disease: An Extensive Review of Advancements in Machine Learning Mechanisms Using an Ensemble and Deep Learning Technique" *Engineering Proceedings* 59, no. 1: 10. https://doi.org/10.3390/engproc2023059010
13. Gale SA, Heidebrink J, Grill J, Graff-Radford J, Jicha GA, Menard W, Nowrangi M, Sami S, Sirivong S, Walter S, Karlawish J. Preclinical Alzheimer Disease and the Electronic Health Record: Balancing Confidentiality and Care. Neurology. 2022 Nov 29;99(22):987-994. doi: 10.1212/WNL.0000000000201347. Epub 2022 Sep 30. PMID: 36180237; PMCID: PMC9728033.
14. Chen, Zhaoyi, Hansi Zhang, Xi Yang, Songzi Wu, Xing He, Jie Xu, Jingchuan Guo et al. "Assess the documentation of cognitive tests and biomarkers in electronic health records via natural language processing for Alzheimer's disease and related dementias." International journal of medical informatics 170 (2023): 104973.
15. Richter-Laskowska, Monika, Ewelina Sobotnicka, and Adam Bednorz. "Cognitive performance classification of older patients using machine learning and electronic medical records." *Scientific Reports* 15, no. 1 (2025): 6564.

979-8-3315-9610-1/25/$31.00 ©2025 IEEE