e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

EARLY STAGE ALZHEIMER DISEASE PREDICTION USING MACHINE LEARNING MODELS

Prof. Rajesh G Kolte*1, Prathama S Bhoir*2, Maryam M Haque*3

*1Head Of The Department, Data Science Usha Mittal Institute Of Technology Mumbai, Maharashtra, India.

*2,3Department, Data Science Usha Mittal Institute Of Technology Mumbai, Maharashtra, India.

DOI : https://www.doi.org/10.56726/IRJMETS52182

ABSTRACT

Alzheimer's disease is one of the most disabling diseases in the elderly. It affects reasoning and memory, and reduces the size of the brain overall, resulting in death. Developing more effective therapies for Alzheimer's disease depends on early detection of the condition. In this paper, authors suggest using machine learning techniques to detect early onset of Alzheimer's. They create a dataset based on the features that represent the early signs of AD. They obtain experimental results using Random Forest, Statistical Machine Learning (SVM), XGBoost, Naive Bayes, and other classifiers. They evaluate the experimental results using metrics such as Confusion Matrix, Precision and Sensitivity. On average, the XGBoost model provides 86% validation accuracy on AD test data, comparable to the established techniques in the literature. So, with the help of XGBoost Model we have built Early Onset Alzheimer prediction model using Python GUI.

Keywords: Prediction, Alzheimer's Disease (AD), Machine Learning, Feature Selection.

I. INTRODUCTION

Alzheimer's disease (AD) is a neurological condition that leads to short-term memory loss, agitation, and delusional thinking. It is sometimes misinterpreted as simply stress or a part of the aging process. Dementia is identified by a collection of symptoms marked by a progressive decline in memory and other cognitive abilities, rather than being a specific disease itself. Indeed, the elderly population is significantly impacted by this condition. Research indicates that 2% of individuals with dementia are under the age of 65. Dementia affects more than 50 million individuals on a global scale, with close to 10 million new cases diagnosed annually. The number of dementia patients is projected to surpass 75 million by 2030, with costs to society nearing $2 trillion [1]. Dementia exerts a more profound influence on healthcare in contemporary society than any other ailment. Moreover, the diagnosis of dementia poses a considerable challenge due to the absence of a universally accepted test for its early detection. Consequently, a cure for dementia remains elusive at present. Nonetheless, numerous interventions and treatments are accessible to enhance the well-being of both patients and their caregivers. [2][3].

In 2018, a staggering 50 million people were affected by Alzheimer's disease. Within the United States, Alzheimer's disease ranks as the sixth leading cause of death. While physical and neurological assessments can be employed to diagnose this condition, they can be both expensive and time-consuming. The symptoms of Alzheimer's typically manifest slowly and progressively worsen over time, significantly impeding daily activities. Nevertheless, identifying this illness in its early stages, before the majority of symptoms become apparent, poses a considerable challenge. [2],[4].

It is advisable to predict Alzheimer's disease during the pre-symptomatic stages in order to impede the progression of the disease. Currently, the diagnosis of Alzheimer's disease involves the utilization of the Multi Slice Multi Echo (MSME) score [5],[6]. Studying numerous brain tissue plates, a task that is both time-consuming and expensive, may be necessary. The early stages of Alzheimer's disease are challenging to predict. Administering therapy during the initial phases of the disease yields superior outcomes and minimal damage compared to treatment later on as the condition advances. Within the suggested approach, the authors introduce a fresh technique rooted in machine learning models that aids in the early detection or onset of Alzheimer's Disease. This proposed technique enables neurologists to forecast the illness before it manifests. Additionally, the suggested machine learning-based method plays a significant role in reducing mortality rates associated with Alzheimer's disease.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

II. LITERATURE SURVEY

Early detection of Alzheimer's disease poses a significant challenge. R. Chaves et al., (2010) [7] introduced a classification strategy for diagnosing early Alzheimer's disease by utilizing association rule mining. Building on the insights from PET data, R. Chaves et al. (2012) [8][9] sought to enhance the predictive accuracy of AD identification through Apriori AR progression, as well as to develop novel treatments and monitor their effectiveness while reducing the computational burden and cost of clinical trials. Liu, Zhang, et al. (2012) [10] employed an ensemble sparse classification technique to assist in the early diagnosis of Alzheimer's disease. They utilized a Sparse representation-based classifier (SRC) to create local patch-based classifiers, which were then combined to offer more resilient and precise classification in the proposed research. The authors of [11] put forward an association rule-based system for diagnosing Alzheimer's disease. Martinez-Murcia et al. [12] explored AD data processing using deep convolutional autoencoders. [13] presented machine learning and deep learning methodologies. Prajapati R and et. al[14] suggested an efficient deep neural network approach for detecting Alzheimer's disease. The authors of [15][16] deliberated on modifiable risk factors, dementia prevention, and care. Chaves R and et.al.[17] and [18] have utilized mining-based techniques for Alzheimer's detection. Kavitha C and et. al[19] proposed a learning-based method for detecting Alzheimer's disease.

The literature has explored various techniques that utilize image data, such as PET and MRI scans, to predict AD. However, obtaining medical images is not always practical. Additionally, it has been observed that by the time patients are recommended for PET and MRI scans, AD has already reached an irreversible stage. Therefore, in order to decrease the mortality rate associated with AD, the authors suggest a new machine learning model that relies on symptoms and features for the early detection of AD.

III. PROPOSED METHODOLOGY

The authors of this paper suggest identifying the initial stages of AD by utilizing a range of innovative characteristics outlined in Table 1.

Table 1. Dataset Features

Feature	Description
Age	Age of the patient
Hand	Right and Left Hand of Patient
Education	Depends on Years on Education
SES	Socio Economic Status Level
MMSE	Mini Mental State Examination
CDR	Mini Mental State Examination
Type 2 Diabetes	Adults with Type 2 diabetes have ahigher risk of developing

The dataset comprises different attributes including age, gender, education level, cognitive scores, and brain imaging data. These attributes are utilized to construct a machine learning model capable of forecasting the likelihood of a patient developing Alzheimer's disease. Fig.1 illustrates the patient's dataset in a .csv file format.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

	A	B	C	D	E	F	G	H	I	J	K
1	ID	Group	Visit	M/F	Hand	Age	EDUC	SES	MMSE	CDR	Type 2 Diabetes
2	OAS1_0001	Nondemented	1	M	R	87	14	2	27	0	0
3	OAS1_0002	Nondemented	2	M	R	88	14	2	30	0	0
4	OAS1_0003	Demented	1	M	R	75	12		23	0.5	0
5	OAS1_0004	Demented	2	M	R	76	12		28	0.5	0
6	OAS1_0005	Demented	3	M	R	80	12		22	0.5	0
7	OAS1_0006	Nondemented	1	F	R	88	18	3	28	0	0
8	OAS1_0007	Nondemented	2	F	R	90	18	3	27	0	0
9	OAS1_0009	Nondemented	1	M	R	80	12	4	28	0	0
10	OAS1_0010	Nondemented	2	M	R	83	12	4	29	0.5	0
11	OAS1_0011	Nondemented	3	M	R	85	12	4	30	0	0
12	OAS1_0012	Demented	1	M	R	71	16		28	0.5	0
13	OAS1_0013	Demented	3	M	R	73	16		27	1	1
14	OAS1_0014	Demented	4	M	R	75	16		27	1	1
15	OAS1_0015	Nondemented	1	F	R	93	14	2	30	0	0
16	OAS1_0016	Nondemented	2	F	R	95	14	2	29	0	0
17	OAS1_0017	Demented	1	M	R	68	12	2	27	0.5	0
18	OAS1_0018	Demented	2	M	R	69	12	2	24	0.5	0
19	OAS1_0019	Demented	1	F	R	66	12	3	30	0.5	0
20	OAS1_0020	Demented	2	F	R	68	12	3	29	0.5	0
21	OAS1_0021	Nondemented	1	F	R	78	16	2	29	0	0
22	OAS1_0022	Nondemented	2	F	R	80	16	2	29	0	0
23	OAS1_0023	Nondemented	3	F	R	83	16	2	29	0	0
24	OAS1_0025	Nondemented	1	F	R	81	12	4	30	0	0
25	OAS1_0026	Nondemented	2	F	R	82	12	4	30	0	0

Fig.1 Generated Dataset .csv file

The methodology under consideration is segmented into three distinct stages. The schematic representation of this methodology can be found in Figure 2.

Stage 1:

The initial step in the analysis of the Alzheimer's disease dataset involved preprocessing and cleaning using exploratory data analysis techniques. Subsequently, a comprehensive analysis was conducted. The datasets pertaining to Alzheimer's disease were found to be inconsistent and duplicated, which had an adverse effect on the accuracy of the algorithms. Prior to processing the data, any missing or NULL values were replaced.

Stage 2:

Prior to developing a machine-learning model, the data is divided into two segments - the training dataset and the testing dataset. The training data is utilized to construct/train the model. Sub sequently, after the model has been trained and fitted, the testing data is fed into the model. The model underwent training with the training dataset and was evaluated with the testing dataset using unseen data.

Stage 3:

In order to perform cross-validation, the dataset was divided into three distinct subgroups. Predictions are made using one subset of the data (test data), while the model's performance is assessed using the other data subsets (training and validation). The dataset was randomly split into 70% for training and 30% for testing.

A. Data Preparation.

During this stage, various data-pre-processing techniques were employed to cleanse and pre-process the data. The handling of missing values and retrieval of features were carried out. It was discovered that there were 9 rows with missing data in the SES column. To address this issue, two approaches were considered. The first and most direct option involved removing the rows with missing data. Alternatively, imputation, which involves replacing missing values with their corresponding values, was another method used to fill in the missing data. Given that there were 140 observations, it was determined that the model would perform better if we opted for imputation. In the case of the SES property, the 9 rows with missing values were eliminated and replaced with the median value

B. Preparation and splitting the data.

The following are the detailed steps of the data splitting stage:

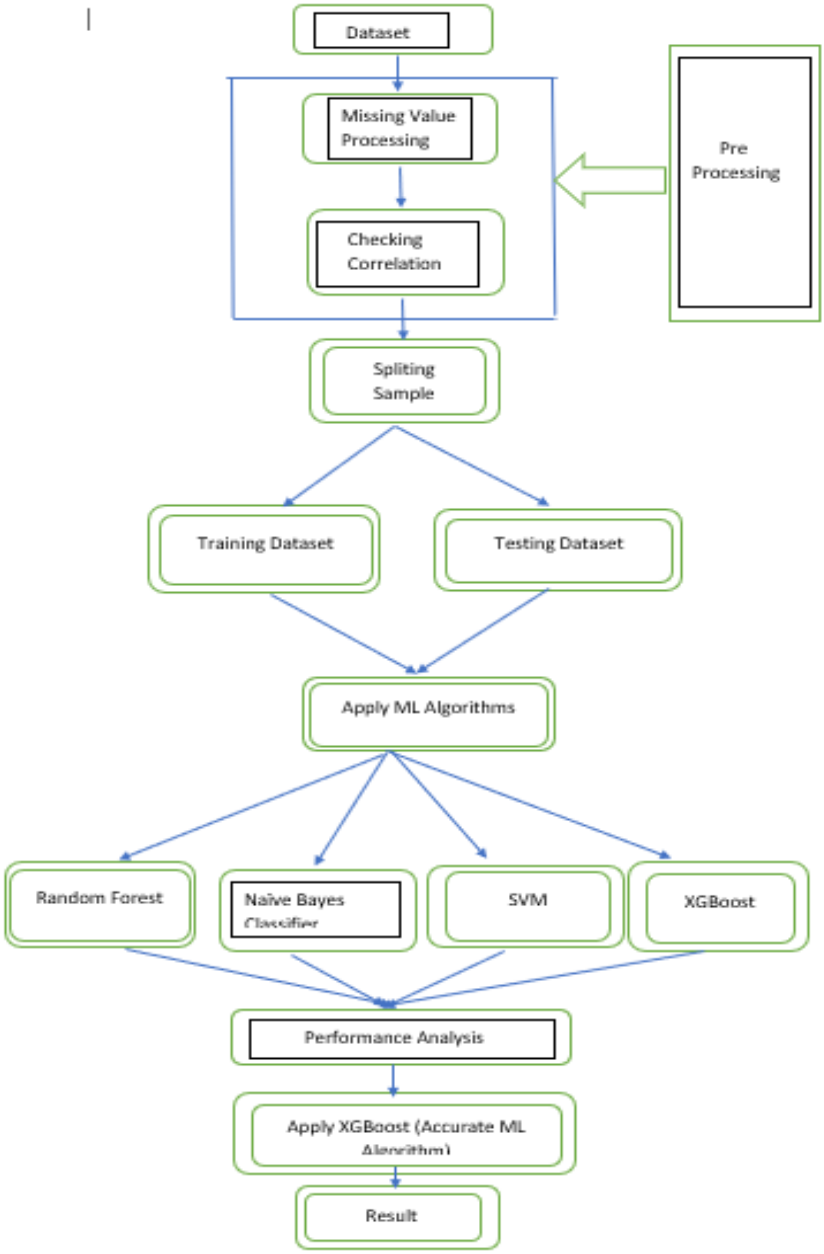
1) Enter the following data: Gender, Age, EDUC, SES,MMSE
2) Train _Data = round (0.5 n row(data)) [Choose 70% ofthe train data].

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

3) Train Data_ indices = sample (1: n row(data), Train Data). [Vector is made up of random in deices].
4) Training the ML = data [Train Data_ indices,][Thedataset for training has been created].
5) Split Formula= Gender + Age + EDUC + SES + MMSE
6) N = 5
7) Splitting the data = n Way Cross Validation (n row(data),n). [5-fold cross validation is produced].

C. Machine Learning Classifier Models Used:

The prediction efficiency of four commonly used machine learning models is calculated through the implementation of the proposed methodology.



1. Random Forest (RF)

Random forest models consist of multiple decision trees, which often yield superior results compared to individual decision trees. This ensemble approach generates predictions by aggregating the outputs of each individual decision tree using the majority voting technique, thereby minimizing overfitting while preserving the predictive power of each tree [20]. In the context of AD prediction, the Random Forest classifier is employed, achieving a prediction accuracy of 72.095 through the utilization of the Random Forest Algorithm.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

2. Support Vector Machine (SVM)

The methodology employed in this approach relies on the identification of data point classes within a multidimensional space. It is centered around the identification of appropriate hyperplanes, specifically targeting a hyperplane that effectively separates the data points belonging to two neighboring clusters of vectors. The support vectors, which are the points closest to the hyperplane, play a crucial role in this technique [20]. The authors were able to achieve an accuracy of 69.76% for their proposed methodology by utilizing the Support Vector Machine Algorithm.

3. XGBoost

XGBoost, short for extreme Gradient Boosting, utilizes gradient-boosted decision trees to achieve optimal speed and performance. The model is trained sequentially, resulting in improved efficiency. XGBoost has been shown to provide superior speed and performance, as evidenced by an accuracy rate of 86.04% in the proposed methodology [20].

4. Naive Bayes Classifier

The Naive Bayes Classifier is an algorithm used for supervised learning. It utilizes Bayes theorem to address classification problems. This technique is particularly useful in creating efficient machine learning models that can make swift predictions. The classifier predicts outcomes based on the probability of an item. It is considered a probabilistic classifier. By employing the Naive Bayes Classifier Algorithm, the proposed methodology attained an accuracy of 83.72%.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$Sensitivity = \frac{TN}{TP + FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Table 2: Performance measure of machine learning models

Model	Accuracy	Sensitivity
Naive Bayes Classifier	**83.72**	**81.81**
XGBoost	**86.04**	**86.36**
SVM	**69.76**	**81.81**
Random Forest	**72.09**	**81.81**

The XGBoost model attains the utmost accuracy and sensitivity, surpassing other techniques like Random Forest, Naïve Bayes Classifier, and SVM, which exhibit similar performance. The proposed technique predicts the early onset of Alzheimer's disease using non-image data, specifically relying on symptom-based features derived from diverse health factors. Experimental analysis substantiates the suitability of the XGBoost algorithm for Alzheimer's disease. Medical practitioners can employ the proposed technique to promptly identify AD patients, enabling timely treatment before the disease progresses. In this prediction model we need to add patients' data values like Gender, SES, MMSE, Education and model will detect the Alzheimer disease on the bases of given data.

IV. CONCLUSION

Alzheimer's disease is a devastating neurological condition. Rather than focusing solely on finding a cure, it is crucial to prioritize reducing the risk and accurately identifying early symptoms. In their study, the authors present an advanced machine learning approach for detecting the early onset of Alzheimer's disease using specific features. Additionally, they conduct a comprehensive analysis comparing various machine learning techniques suitable for early detection. The results of the performance analysis demonstrate that the XGBoost technique offers superior predictive capabilities. Neurologists can effectively employ this proposed technique to identify the early stages of Alzheimer's disease. So, with the help XGBoost Model we have built Early Onset Alzheimer prediction model using Python GUI.

V. REFERENCES

[1] Prince, M., Wimo, A., Guerchet, M., Ali, G., Wu, Y., Prina, M., et al. (2015). The global impact of dementia. World Alzheimer Report, pages 1–82.

[2] World Health Organization (2020a). Alzheimer's disease fact sheet.

[3] World Health Organization (2020b). Dementia.

[4] Patterson, C. (2018). The state of the art of dementia research: New frontiers. World Alzheimer Report.

[5] Janghel, R. and Rathore, Y. (2020). Deep convolution neural network-based system for early diagnosis of Alzheimer's disease. IRBM.

[6] Kumar, U. (2019). Applications of machine learning in disease pre-screening. In Pre-Screening Systems for Early Disease Prediction, Detection, and Prevention, pages 278–320. IGI Global.

[7] Chaves, R., J. Ramírez, et al., 2010. Effective Diagnosis of Alzheimer's Disease by Means of Association Rules. Hybrid Artificial Intelligence Systems, Springer, 1, 452- 459.

[8] Chaves, R., J. Ramirez, et al. (2012). FDG and PIB biomarker PET analysis for the Alzheimer's disease detection using Association Rules. Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), IEEE.

[9] Liu, M., D. Zhang, et al. (2012). Ensemble sparse classification of Alzheimer's disease. Neuroimage 60(2), 1106-1116

[10] Chaves, R., J. Ramírez, et al. (2013). Integrating discretization and association rule-based classification for Alzheimer's disease diagnosis. Expert Systems with Applications 40(5), 1571-1578.

[11] Martinez-Murcia FJ, Ortiz A, Gorrie JM, Ramirez J, Castillo-Barnes D. Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders. IEEE J Biomed Health Inform. (2020) 24:17–26.doi: 10.1109/JBHI.2019.2914970

[12] Khan P, Kader MF, Islam SR, Rahman AB, Kamal MS, Toha MU, et al. Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances. IEEE Access. (2021) 9:37622–55. Doi: 10.1109/ACCESS.2021.3062484

[13] Prajapati R, Khatri U, Kwon GR. "An efficient deep neural network binary classifier for Alzheimer's disease classification," In: International Conference on Artificial Intelligence in Information and Communication (ICAIIC). (2021), p. 231–234.

[14] Yaffe K. Modifiable risk factors and prevention of dementia: what is the latest evidence. JAMA Intern Med. (2018) 178:281–2.doi: 10.1001/jamainternmed.2017.7299

[15] Livingston G, Sommer lad A, Ortega V, Costa Freda SG, Huntley, D, et al. Dementia prevention, intervention, and care. The Lancet. (2017) 390:2673– 73. Doi: 10.1016/S0140-6736<17>31363-6

[16] Chaves, R., J. Goris, et al. (2011). Efficient mining of association rules for the early diagnosis of Alzheimer's disease. Physics in medicine and biology 56(18): 6047.

[17] Chaves, R., J. Ramírez, et al. (2012). Association rule-based feature selection method for Alzheimer's disease diagnosis. Expert Systems with Applications 39(14): 11766-11774.

[18] Kavitha C, Mani V, Srividya SR, Khalaf OI and Tavera Romero CA (2022) Early-Stage Alzheimer's Disease Prediction Using Machine.

[19] Siv Akani GA, Ansari R. Machine learning framework for implementing Alzheimer's disease. Int Conference Commune Signal Process. (2020) 12:588– 92. Doi: 10.1109/ICCSP48568.2020.9182220