ISBN: **978-93-343-9865-6**

NSEI/2025/03

Evaluating Machine Learning Approaches for Early Detection of Alzheimer's Disease: Bridging Gaps in Rural Healthcare

Priyanjana Mitra[1], Tridib Chakraborty[2]. Aniruddha Roy[3], Shovon Chaudhuri[4]

[1,2] Assistant Professor of IT Department Guru Nanak Institute of Technology Kolkata

[3,4] Student of IT department Gurunanak Institute of Technology Kolkata

* Corresponding author email: priyanjana.mitra@gnit.ac.in, tridib.chakraborty@gnit.ac.in, ani321r@gmail.com, shovanchaudhuri91@gmail.com

Abstract-Now a days common people are going through erratic schedule and suffering from many neurological diseases. A progressive neurological disorder in older adults termed as Alzheimer's disease occurs because of abnormal accumulations of two harmful proteins: amyloid, which forms plaques around neurons, and protein tau, which creates tangles inside the brain cells. Although it appears to occur at old age, the disease develops at slow pace at early stage gradually leading to deterioration in memory, thinking capability, decision making and performing regular activities. Research to date indicates that there is no cure to stop the progression of Alzheimer's disease, which makes early diagnosis crucial for effective management and care. Early prediction is specifically tough in rural healthcare settings, due to limited access of neurologists, scarcity of diagnostic facilities and updated clinical resources leads to delay of disease detection and treatment. Considering these facts, contribution of Machine Learning/Deep Learning along with advanced sensors for monitoring the healthcare has proved to be incredible and offers a positive pathway to bridge the gap by mitigating these challenges. In this paper, we analyze patient data related to Alzheimer's disease using various Machine Learning techniques to predict the likelihood of the condition in suspected cases. Specifically, we employ algorithms such as Support Vector Machine (SVM), Decision Tree (DT), and XGBoost. The study emphasizes a comprehensive evaluation of Machine Learning methods, comparing different data processing approaches and assessing their accuracy, efficiency, and error rates, while also discussing the strengths and limitations of each algorithm. Furthermore, the proposed model aims to enable early-stage prediction of Alzheimer's disease, which can help reduce patient risk and lower the mortality rate associated with the condition and considers feasibility for deployment in rural healthcare infrastructure

Keywords- Alzheimer, Machine Learning, MMSE, SVM, XGBOOST, Decision Tree

I. INTRODUCTION

ISBN: **978-93-343-9865-6**

The causes of Alzheimer's disease arise from a combination of age-related brain changes along with genetic, environmental, and lifestyle factors [1]. The condition primarily affects individuals over the age of 65, though in some cases, it may develop earlier, during a person's 40s or 50s, which is referred to as early-onset Alzheimer's disease [2]. On average, patients live with this illness for 8 to 10 years, and approximately one in eight people over the age of 65 are affected. The Mini-Mental State Examination (MMSE) is widely recognized as a cognitive assessment method and plays a vital role in identifying Alzheimer's disease [3]. At the stage of MCI, individuals show higher levels of memory loss and cognitive challenges compared to what is usual for their age group. In this condition, Mild Cognitive Impairment (MCI) often leads to memory loss, and is likely to progress into dementia. Since Alzheimer's disease has no definitive cure, identifying it at the MCI stage is essential, as it enables individuals to adopt healthier lifestyle practices and make better plans for managing memory decline [4].We utilize SVM, Decision Tree, and XGBoost algorithms in this research to analyze their predictive efficiency for Alzheimer's disease. SVM is known by its good generalization ability and robustness to high dimensional data, like those used in this paper. Decision Trees have been applied in Alzheimer's detection, with new approaches comparing enhanced Decision Tree models against traditional versions. XGBoost, on the other hand, utilizes parallel tree boosting and is recognized as one of the most powerful machine learning libraries for tasks such as regression, classification, and ranking.

Organization of the paper is as follows: Section II discusses related literature, Section III describes the methodology, and Sections IV and V present results and conclusions.

II. Literature Review

A deep learning approach was proposed in [5] which integrates brain network analysis with clinically significant information such as the subject's age and gender for aimed at identifying Alzheimer's disease in its early stages. In this method, data obtained from resting-state functional MRI (R-fMRI) scans is used to estimate functional connectivity across different brain regions. These functional networks are then modeled using an autoencoder, which helps in identifying detailed patterns connected to early Alzheimer's symptoms and the transition into MCI . For this work, the dataset was sourced from the ADNI database. For early diagnosis, raw R-fMRI data first undergoes preprocessing. From this, time-series data in the form of a 90 × 130 matrix is generated, representing blood oxygen levels across brain regions over time. This information is then transformed into a 90 × 90 correlation matrix to construct brain networks. A three-layer autoencoder is applied to extract meaningful network features, simulating neural system growth and capturing essential attributes of the brain. Given the restricted size of the dataset, the k-fold cross-validation technique was used to minimize overfitting issues. Additionally, a multistage classifier using machine learning algorithms such as Support Vector Machine (SVM) was explored to improve classification performance across Na¨ıve Bayes and K-Nearest Neighbor (KNN) methods were proposed in [6] for classifying different subject groups. In addition, a Particle Swarm Optimization (PSO) technique was applied to identify the most relevant biomarkers indicating Alzheimer's disease (AD) or Mild Cognitive Impairment

ISBN: **978-93-343-9865-6**

(MCI). The dataset was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI), and the MRI scans underwent preprocessing before further analysis. Then feature is selected in terms of volumetric and thickness measurements. To distinguish between subjects, algorithms such as Gaussian Na¨ıve Bayes (GNB), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) have been employed. At the initial stage of classification, the GNB classifier was used to separate subjects into Alzheimer's disease (AD), Mild Cognitive Impairment (MCI), and Normal Control (NC) groups. In later stages, SVM and KNN were applied to further analyze the subjects based on the outcomes of the initial classification. In [7] the authors proposed a model that utilizes consecutive MRI scans to track disease progression over time, enabling more precise diagnosis. Specifically, MRI brain images captured at six different time intervals (each six months apart) from the ADNI database were used as input. A 2×2×2 convolutional operation was applied to extract features, followed by pooling of neurons through linear combinations. The architecture included fully connected layers that transmitted data through nonlinear activation functions, culminating in a softmax layer. This softmax function assigned probabilities (ranging from 0 to 1) to each class, ensuring that the sum of outputs equaled 1. It was optimized using backpropagation to predict class probabilities accurately. This method allowed dimensional and longitudinal features to be integrated for distinct identification. In [8] the authors introduced a multimodal framework where medical images were used for training two independent convolutional neural networks (CNNs). One CNN processed MRI scans, while the other handled PET scans. The CNN structure included a sampling layer for downscaling, convolutional and fully connected layers, pooling layers, and finally an output layer. The advantage of this approach is its ability to combine clinical neuropsychological data with neuroimaging results, thereby enhancing diagnostic accuracy.

III. METHODOLOGY

In this project, three Machine Learning (ML) algorithms are applied to predict Alzheimer's disease at an early stage. The following subsections provide a detailed discussion of these algorithms.

I. *Machine Learning Algorithm*

In this paper we have used three ML algorithms are discussed below:

- *Support Vector Machine:* In n-dimensional space, there may be several lines or decision boundaries used to separate the classes, but we must choose the optimum decision boundary that best aids in classifying the data points. The hyper plane of SVM is a name for this optimal boundary [9]. The dataset's features determine the hyper plane's dimensions, therefore if there are just two features (as in the example image), the hyper plane will be a straight line. Additionally, if there are three features, the hyper plane will only have two dimensions. We always build a hyper plane with a maximum margin, or the greatest possible separation between the data points.By presenting an example, the SVM algorithm's operation may be better understood. Consider a dataset with two tags (green and blue), two features (x1 and x2), and two tags. We're looking for a classifier that can classify the 14 pair of coordinates (x1, x2) as either green or blue.The SVM technique may be used

ISBN: **978-93-343-9865-6**

to identify the optimal line or decision boundary; this line or area is known as a hyper plane. Which point between the two classes is closer is determined using the SVM algorithm. These areas of attention are known as help vectors [10]. The margin is the separation of the vectors from the hyper plane. The goal of SVM is to maxim this margin. The hyper plane with the largest margin is the ideal hyper plane. SVM is a powerful controlled computation that performs well on smaller yet more complicated datasets. Support Vector Machine, often known as SVM, may be used for both relapse and order tasks, although they typically perform well in characterization-related problems. They were invented in the 1990s and quickly became quite popular. They are now the most often used method for developing algorithms that work well with little adjustments. Depending on how many characteristics you have, you can decide between SVM and logistic regression. SVM performs better with a condensed and compact dataset. The best way to test out logistic regression is often to use it. SVM can be used without a kernel if it doesn't

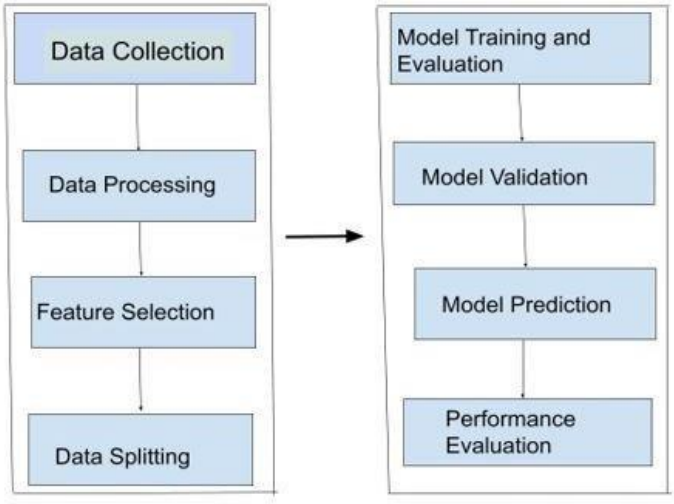


Fig. 1. Proposed workflow

perform well [11]. Logistic regression and SVM both operate similarly in the absence of a kernel, however your features may affect how efficiently they work.SVM is characterized to such an extent that it is characterized as far as the help vectors just, we don't need to stress over different perceptions since the edge is made utilizing the focuses which are nearest to the hyper plane (support vectors), while in calculated relapse the classifier is characterized over every one of the places. As a result, SVM enjoys some inherent speedups.

- *Decision Tree:* Decision Trees are structured like trees, where data is recursively partitioned based on selected feature values. Each split divides instances into subsets, where the intermediate subsets form internal nodes and the final outcomes are represented by leaf nodes [12] .Decision Trees are especially useful for identifying significant interactions between features and the target variable.They are called "Decision Trees" because their

ISBN: **978-93-343-9865-6**

structure resembles a tree: starting from a root node, the tree expands into branches and eventually reaches leaf nodes that represent decisions or outcomes [13].In this structure, there are two node types: Decision Nodes, responsible for making splits and branching, and Leaf Nodes, which denote final outcomes with no further branching.The process of testing or making decisions is carried out using the dataset provided, and the model is often illustrated using graphical representations. This visual format helps to outline all possible solutions or decisions that can be derived under given conditions. To construct a Decision Tree the CART (Classification and Regression Tree) algorithm is commonly used. This algorithm works by asking a series of questions, typically in a binary format (Yes/No), and based on the answers, the tree is split into different subtrees. Decision Trees resemble the way humans make decisions, which makes them intuitive and easy to interpret [14]. The underlying logic is straightforward because of the tree-like structure: the procedure starts at the root node, where the algorithm examines the value of a chosen attribute. Depending on the evaluation, it proceeds along the corresponding branch to the next node. At each subsequent node, the algorithm performs additional comparisons with other attributes, continuing this process until a final classification or decision is reached at the leaf nodes

- *XGBoost:* XGBoost is a powerful and scalable gradient boosting library designed for building machine learning models efficiently. As an ensemble technique, it integrates the outputs of several simpler models to generate a stronger and more precise outcome. The name “XGBoost” stands for “Extreme Gradient Boosting,” and it has become extremely popular in the machine learning community due to its ability to handle large datasets effectively and deliver high performance across a variety of tasks, such as classification and regression [15]. One of XGBoost’s key strengths is its ability to handle missing values naturally, allowing it to process real-world datasets without extensive preprocessing. Additionally, it supports parallel computation, which makes training on large datasets faster and more practical.

II. *Dataset Description*

This dataset comprises a longitudinal collection of 150 individuals whose ages range between 60 and 96 years. Each participant underwent multiple scans, with at least one year between each visit, resulting in a total of 373 recorded imaging sessions. For each participant, the dataset provides three to four separate T1-weighted MRI scans, each obtained during a single scanning session. The participant pool consists of both men and women who are all right-handed. Among the participants, 72 individuals consistently exhibited no signs of dementia over the course of the study. Additionally, 64 participants were

ISBN: **978-93-343-9865-6**

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	83.92%	0.81	0.85	0.83
Decision tree	79.46%	0.74	0.83	0.79
Support vector machine	77.67%	0.71	0.88	0.79

Fig. 2. Comparison of Performance of different ML model

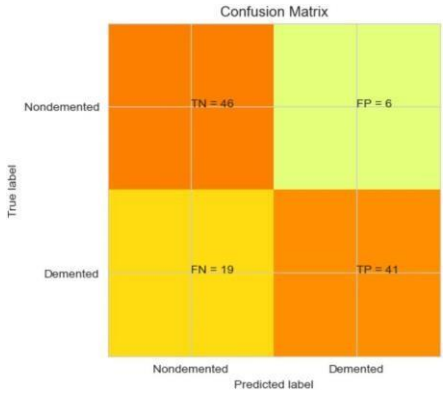


Fig. 3. Confusion Matrix of SVM

initially diagnosed with dementia during their initial assessments and continued to exhibit dementia in subsequent scans. This group consisted of 51 participants who had been identified as having mild to moderate Alzheimer's disease. Furthermore, 14 participants initially presented as no demented but were subsequently diagnosed with dementia during later visits.

III. *Data Preprocessing*

In this phase, various data mining techniques are employed to clean and preprocess the dataset. This includes handling missing values, extracting relevant features, and transforming them as needed. Missing values can be addressed in two primary ways. The simplest approach is to remove the rows containing missing data. Alternatively, missing values can be filled using imputation methods. In this study, nine rows with missing values in the SES attributes were removed, and the median value was used to perform imputation for the remaining missing entries.

IV. Prediction Model Design and Result Analysis

As mentioned earlier, we developed a predictive model utilizing three different machine learning techniques: Support Vector Machine (SVM), Decision Tree, and XGBoost. The workflow of the prediction system is presented in Fig. 1. The performance of the resulting models is compared by accuracy, recall, precision, and F1 scores. Fig. 3, Fig. 4, and Fig. 5 represent the confusion matrices of

ISBN: **978-93-343-9865-6**

SVM, Decision Tree, and XGBoost respectively. Table 1 presents the evaluation metrics—accuracy, recall, precision, and F1 scores—for the three predictive approaches.

V. CONCLUSION

Alzheimer's disease is a significant public health concern. Since there is currently no cure, efforts are focused on reducing risk, providing early interventions, and accurately diagnosing symptoms at an early stage. As highlighted in the literature, numerous studies have attempted to detect Alzheimer's disease using various machine learning algorithms and microsimulation techniques. However, identifying the most relevant features for early detection remains a challenging task. Future work will aim to extract and analyze new features that could enhance the early detection of Alzheimer's, while also removing redundant or irrelevant features from existing datasets to improve predictive accuracy. By incorporating metrics such as MMSE scores and education levels into our model, it will be better equipped to differentiate between healthy individuals and those with

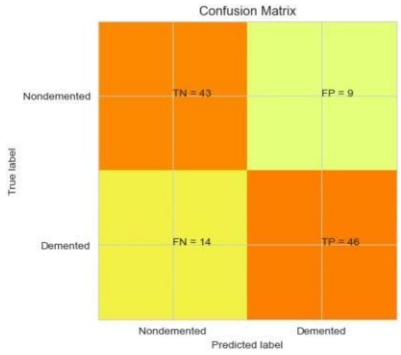


Fig. 4. Confusion Matrix of Decision Tree

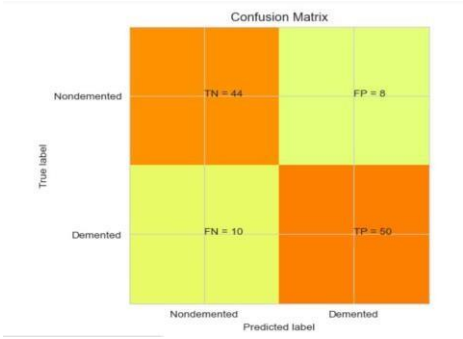


Fig. 5. Confusion Matrix of XGBoost

Alzheimer's. This study compares and evaluates different machine learning algorithms, including pre-processing and prediction steps for Alzheimer's detection. In the proposed approach, missing values were addressed by deleting columns in the SES and MMSE attributes. The experimental findings show that the XGBoost algorithm surpasses the other classifiers, obtaining the highest accuracy of 83

REFERENCES

ISBN: **978-93-343-9865-6**

1. Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert. Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimer's disease. *Medical Image Analysis*, 48:117–130, 2018.
2. Aunsia Khan and Muhammad Usman. Early diagnosis of alzheimer's disease using machine learning techniques: A review paper. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 01, pages 380–387, 2015.
3. Baiying Lei, Mengya Yang, Peng Yang, Feng Zhou, Wen Hou, Wenbin Zou, Xia Li, Tianfu Wang, Xiaohua Xiao, and Shuqiang Wang. Deep and joint learning of longitudinal data for alzheimer's disease prediction. *Pattern Recognition*, 102:107247, 2020.
4. Tong Tong, Katherine Gray, Qinquan Gao, Liang Chen, and Daniel Rueckert. Multi-modal classification of alzheimer's disease using nonlinear graph fusion. *Pattern Recognition*, 63:171–181, 2017.
5. Ronghui Ju, Chenhui Hu, pan zhou, and Quanzheng Li. Early diagnosis of alzheimer's disease based on resting-state brain networks and deep learning.
6. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(1):244–257, 2019.
7. K.R. Kruthika, Rajeswari, and H.D. Maheshappa. Multistage classifier-based approach for alzheimer's disease prediction and retrieval. *Informatics in Medicine Unlocked*, 14:34–42, 2019.
8. J. Neelaveni and M.S.Geetha Devasana. Alzheimer disease prediction using machine learning algorithms. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 101–104, 2020.
9. Fan Zhang, Zhenzhen Li, Boyan Zhang, Haishun Du, Binjie Wang, and Xinhong Zhang. Multi-modal deep learning model for auxiliary diagnosis of alzheimer's disease. *Neurocomputing*, 361:185–195, 2019.
10. Ragavamsi Davuluri. A survey of different machine learning models for alzheimer disease prediction. *International Journal of Emerging Trends in Engineering Research*, 8:3328–3337, 07 2020.
11. Theodoros Evgeniou and Massimiliano Pontil. Support vector machines: Theory and applications. volume 2049, pages 249–257, 09 2001.
12. Yongli Zhang. Support vector machine classification algorithm and its application. In Chunfeng Liu, Leizhen Wang, and Aimin Yang, editors, *Information Computing and Applications*, pages 179–186, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
13. Dana AL-Dlaeen and Abdallah Alashqur. Using decision tree classification to assist in the prediction of alzheimer's disease. In *2014 6th International Conference on Computer Science and Information Technology (CSIT)*, pages 122–126, 2014.
14. Mingwu Jin and Weishu Deng. Predication of different stages of alzheimer's disease using neighborhood component analysis and ensemble decision tree. *Journal of Neuroscience Methods*, 302:35–41, 2018. A machine learning neuroimaging challenge for automated diagnosis of Alzheimer's disease.
15. Morshedul Bari Antor, A. H. M. Shafayet Jamil, Maliha Mamtaz, Mohammad Monirujjaman Khan, Sultan Aljahdali, Manjit Kaur, Parminder Singh, and Mehedi Masud. A comparative analysis of machine learning algorithms to predict alzheimer's disease. *Journal of Healthcare Engineering*, 2021:9917919, Jul 2021.
16. Laboni Akter and Ferdib-Al-Islam. Dementia identification for diagnosing alzheimer's disease using xgboost algorithm. In *2021 International Conference*