A Comparative Analysis of Machine Learning Models for Early Detection of Alzheimer’s Disease

V S Manoj Kumar  
Department of CSE  
BNM Institute of TechnologyBengaluru, India  
23cse022@bnmit.in

Tanmay Anand  
Department of CSE  
BNM Institute of TechnologyBengaluru, India  
23cse009@bnmit.in

*Abstract*—Alzheimer’s disease (AD) is a progressive neurodegenerative disorder that affects millions of people worldwide. AD is one of the most extensively researched neurological disorders, yet there is no cure or proven method to slow or halt its progression. It is a dynamic ailment that decimates the mind’s memory and it’s general functioning. Early diagnosis of AD is crucial for effective disease management and therapeutic intervention. Currently, there is no definitive test to diagnose Alzheimer’s disease. Brain scans alone cannot serve as a conclusive indicator of the condition. At present, physicians often rely on reports from family members about changes in social behavior, combined with a review of the individual’s medical history, to determine whether the person may be experiencing AD.

Machine Learning (ML) models have emerged as promising tools for the early detection of AD. Machine learning methodologies offer promising avenues for early Alzheimer's detection, encompassing a spectrum of algorithmic approaches. By leveraging both traditional statistical classifiers and advanced computational techniques, researchers can develop nuanced diagnostic models capable of analyzing complex neurological data. This study explores an array of algorithmic strategies, including logistic regression, support vector machines, random forest implementations, Generalised Linear Model, and deep neural network architectures. Each model is systematically evaluated through comprehensive performance assessments, focusing on critical diagnostic metrics that illuminate their potential for accurate and early disease identification.

Keywords—Alzheimer’s Disease (AD), Machine Learning (ML), Healthcare, Logistic Regression, Generalised Linear Model, Neural Networks.

# Introduction

Alzheimer’s Disease (AD) is a progressive neurodegenerative disorder and one of the leading causes of dementia globally. Despite being one of the most extensively researched neurological conditions, there remains no cure or proven method to halt or slow its progression. AD is characterized by a gradual deterioration of memory, cognitive abilities, and overall brain function, significantly impacting the quality of life of patients and their caregivers. The disease predominantly affects individuals aged 65 and above, but early-onset cases can occur in younger populations. Early and accurate diagnosis is critical to enable effective management and therapeutic interventions, potentially improving patient outcomes and prolonging the quality of life.

The diagnosis of Alzheimer’s Disease presents a significant challenge due to its complex nature. At present, no single definitive test exists to diagnose AD. Brain imaging techniques, while advanced, cannot conclusively confirm the disease. Physicians often rely on a combination of medical history reviews, cognitive assessments, and input from family members regarding changes in social and behavioral patterns to make a clinical diagnosis. This reliance on subjective and fragmented data underscores the need for more objective and robust diagnostic tools.

Recent advancements in machine learning (ML) offer transformative potential in addressing these challenges. By leveraging large-scale datasets, such as neuroimaging data and electronic health records, ML models can analyze complex patterns and relationships that may elude traditional diagnostic methods. A diverse range of algorithmic approaches has been explored for AD diagnosis, including logistic regression, support vector machines (SVM), random forests, generalized linear models (GLM), and K-Nearest Neighbors (KNN).. These algorithms differ in their methodological frameworks and capabilities, from statistical classifiers to advanced architectures capable of capturing intricate, non-linear dependencies in data.

This paper aims to provide a comprehensive analysis of ML models applied to the early detection of Alzheimer’s Disease. By systematically evaluating the performance of various algorithms across diagnostic metrics, we seek to highlight their potential for improving accuracy and early disease identification. Section 2 delves into the theoretical foundations and practical applications of key ML techniques such as SVM, GLM, and KNN. Section 3 presents a comparative assessment of these methodologies, discussing their strengths, limitations, and implications for clinical practice.

Through this study, we aim to contribute to the growing body of research on leveraging artificial intelligence for healthcare, with a particular focus on neurodegenerative diseases. The findings of this work have the potential to advance diagnostic practices for Alzheimer’s Disease, enabling earlier intervention and paving the way for improved patient care.

# Data

## Data Description

The Alzheimer’s Disease Prediction of Longitudinal Evolution (TADPOLE) challenge provided the dataset used in this study. It is the dataset that was collected from ADNI participants in North America. The goal of the multicenter ADNI project is to enhance the imaging, biochemical, genetic and clinical biomarkers for early AD diagnosis.

## Data Pre-Processing and Feature Selection

This study has made use of a subset of the TADPOLE dataset. Out of 1,737 individuals, 530 of them (106 from each AD stage) had their examination records chosen for trials utilizing various data mining methods for the classification of five distinct AD stages. Data completeness is crucial for accurate modeling in data mining approaches. On the other hand, the TADPOLE dataset is somewhat sparse. Consequently, we have chosen just the attributes with the greatest amount of data coverage. Complete data coverage is achieved when an attribute has a value for each of the 12,741 examination record instances. The 41 attributes with sufficient data have also undergone an examination to eliminate redundancy.

We chose particular features from the dataset (ADAS11, MMSE, RAVLT\_immediate, and CDRSB) to feed into the model. The target variable (DX\_bl), which stands for the diagnosis, is predicted using these columns.

Machine learning algorithms require numerical inputs, so these categorical variables are encoded using label encoding. This transforms categorical labels into integer values, making them compatible with the models.

To assess a model's capacity for generalization, the TADPOLE dataset was divided into training and testing sets. Usually, a 70-30 or 80-20 split is employed, where the testing set is utilized for evaluation and the training set is used for model training. In this instance, train\_test\_split() was used to separate the dataset into a 30% testing set and a 70% training set.

Class imbalances can occur in medical datasets like the TADPOLE dataset, where some classes (like control or MCI) are overrepresented in comparison to other classes (like Alzheimer's disease). To solve this problem and make sure the model doesn't start to favor the majority class, strategies like readjusting the model's class weights or oversampling the minority class (for instance, by employing SMOTE) might be employed.

According to prior knowledge of Alzheimer's disease progression, specific variables such as cognitive scores (MMSE, ADAS11, CDRSB), demographic data (e.g., age, gender), and clinical test findings were deemed most significant to illness diagnosis. Manual feature selection is based on domain knowledge to identify factors that are likely to have a substantial impact on the result.

Decision trees and random forests are two examples of machine learning algorithms that offer feature importance scores, which show the relative contribution of each feature to the model's predictions. While features with higher relevance scores are kept, those with lower important scores can be eliminated. This approach aids in determining the factors that have the greatest bearing on Alzheimer's disease prediction.

# model evaluation – confusion matrix

A confusion matrix is typically used to assess the classifier's performance. It is a particular table that shows the classifier's anticipated classes and true classes as well as the kinds of mistakes the classifier made. The confusion matrix uses four distinct terminology, which are listed below :

# proposed solution

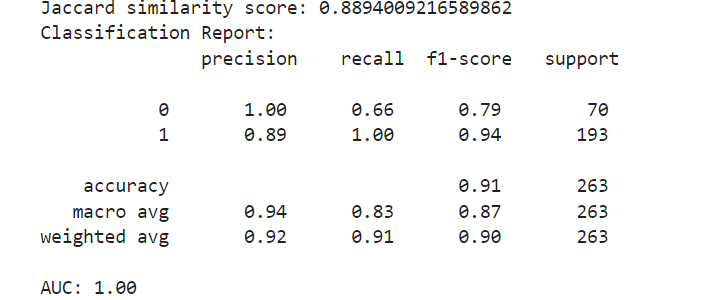
In order to categorize the five distinct stages of AD, this study applies five distinct machine learning and data mining algorithms to the ADNI dataset : K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes, Generalized Linear Model (GLM) and XGBoost. All of these methods are implemented in this study using Sci-kit Learn, a popular Python library to implement machine learning models.

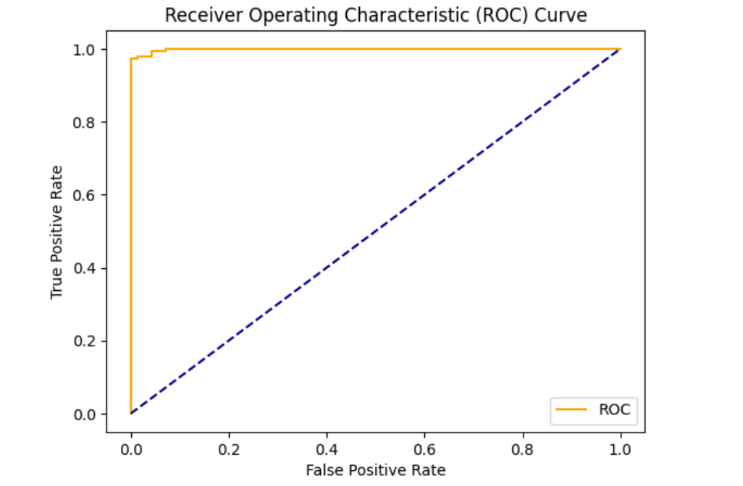
## Support Vector Machine (SVM)

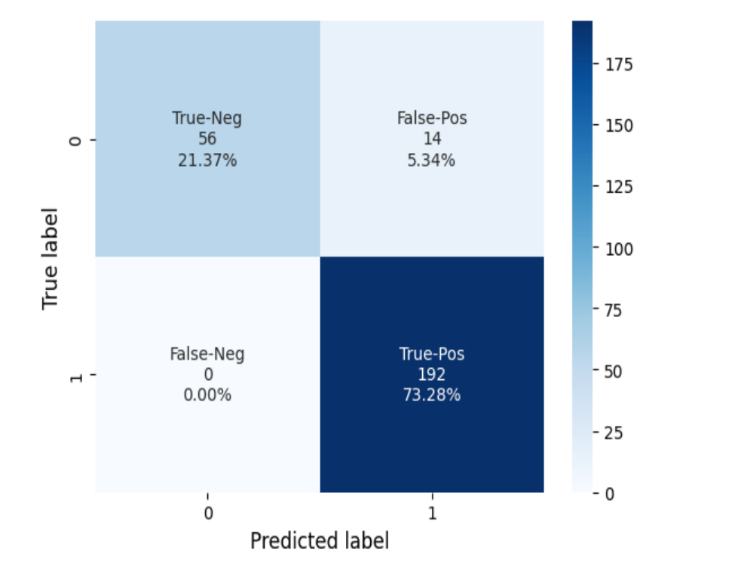
A popular supervised learning technique for classification and regression applications is the Support Vector Machine (SVM). It’s foundation is the idea of determining the appropriate hyperplane to divide data points into distinct classes. SVMs work well in high-dimensional spaces and are resistant to overfitting, particularly when there are more dimensions than data points.

The trade-off between minimizing the model complexity and attaining a low error on training data is managed by the SVM's C parameter. Better generalization may result from a simpler decision boundary, which is encouraged by a lower C.

The Radial Basis Function kernel was used by the SVM, to handle complex non-linear interactions by performing classification in a higher dimensional space. StratifiedKFold approach was also implemented to ensure that the folds are balanced in terms of the distribution of the target variable.



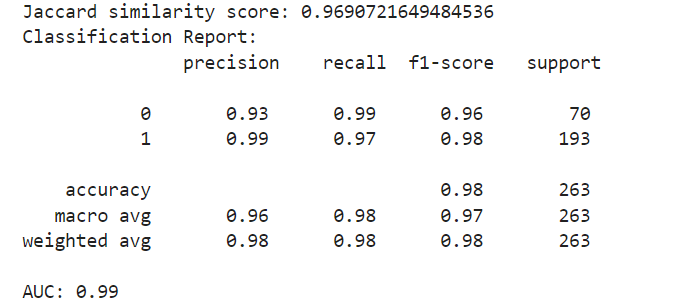


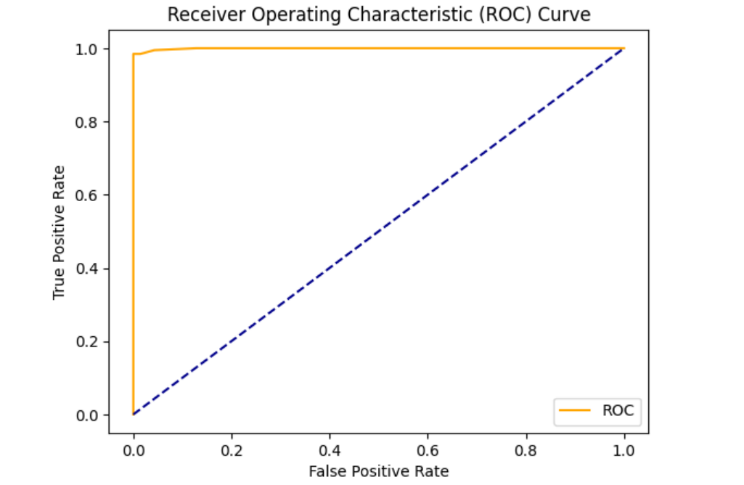


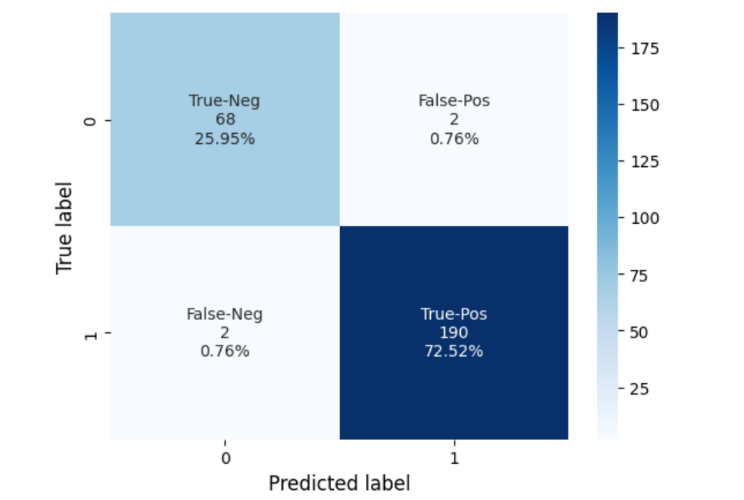
## K-Nearest Neighbors (KNN)

A straightforward data mining method for classification and regression issues is the K-nearest neighbor algorithm. Based on the majority classes of its K neighbors, the KNN classification algorithm places an object in a certain class. The number of neighbors is taken into account for polling is defined by the positive integer K. We employed a 5-fold cross-validation technique to assess model performance across different splits of the dataset. The features considered for classification were CDRSB, ADAS11, MMSE, and RAVLT\_immediate, while the target variable DX\_bl was encoded into numerical values using LabelEncoder.

Here, the choice of K = 5 was arbitrarily chosen, and different values of K can be experimented with to determine the optimal number of neighbors.

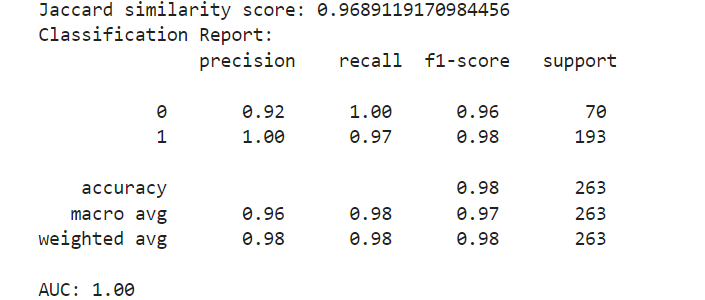


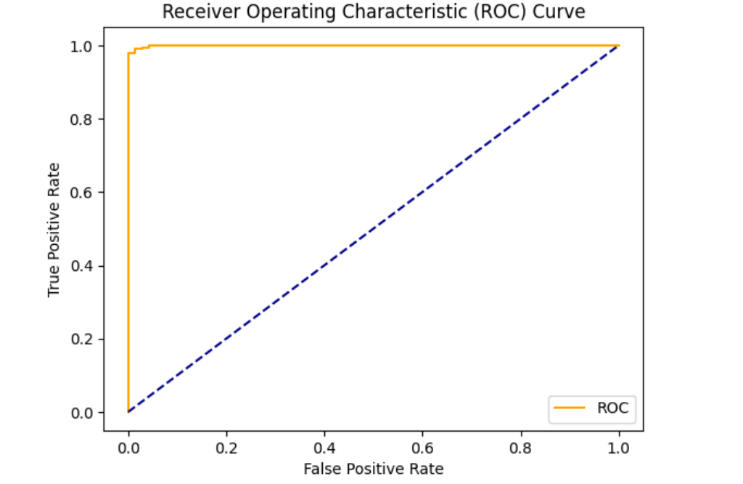


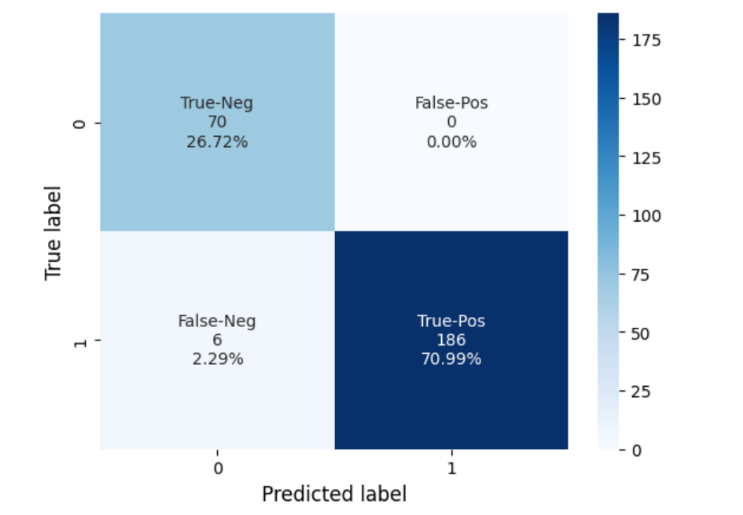


## Naïve Bayes

Based on Bayes' Theorem, Naive Bayes is a probabilistic supervised learning method that is frequently used for classification tasks. Although it can handle continuous features (as in the Gaussian form), it works best with datasets that contain categorical features. It assumes independence among features, which simplifies computation. A Stratified K-Fold cross-validation is used to ensure the class distribution remains consistent across splits. The GaussianNB classifier is trained on the training folds and used to predict on the testing folds.



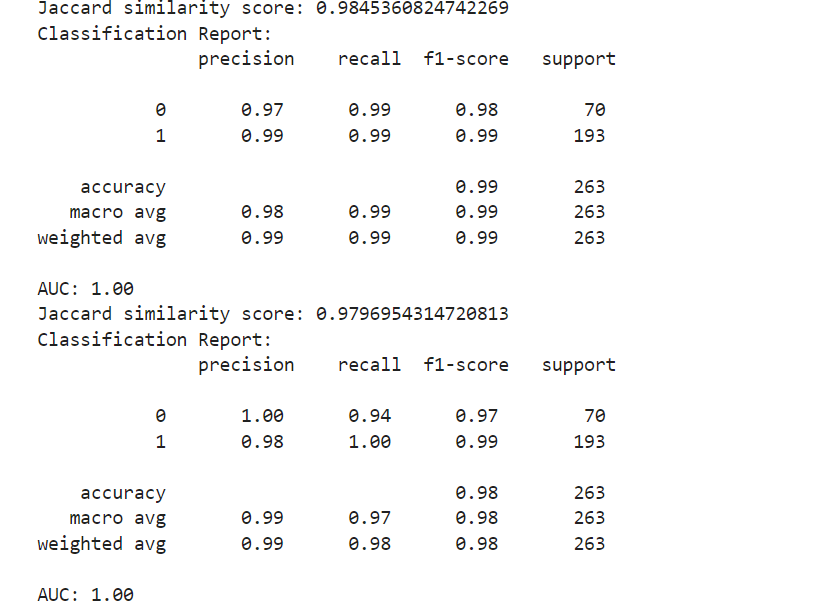


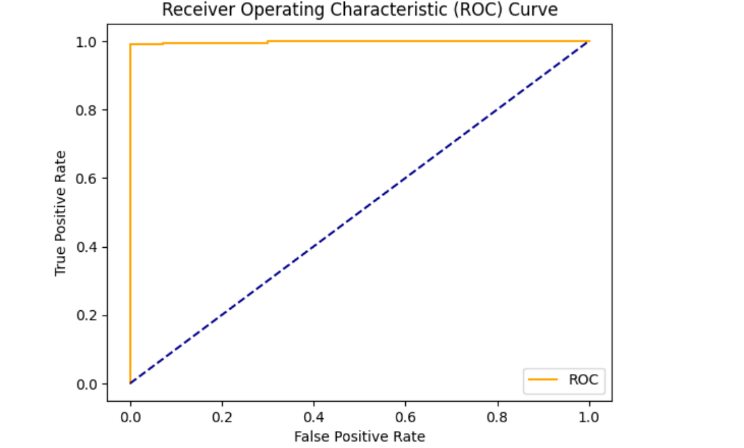


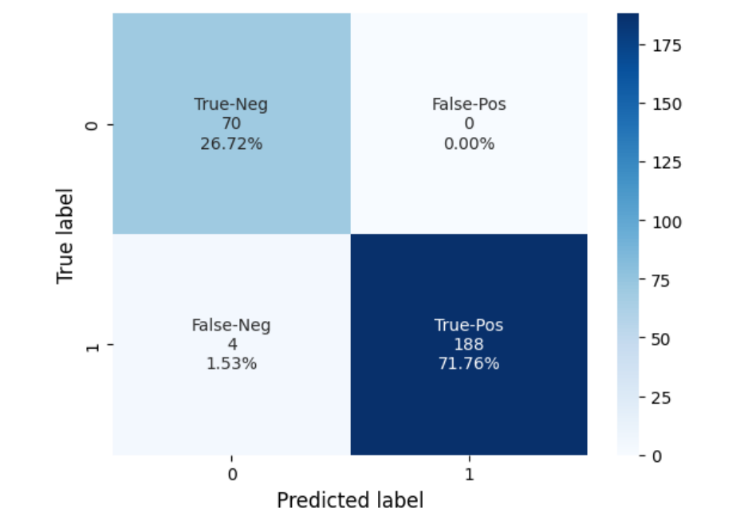
## XGBoost

Extreme Gradient Boosting, or XGBoost, is a gradient-boosting library that has been enhanced for speed and efficiency. To reduce classification error, it iteratively refines a collection of weak learners (decision trees). Boosting is a sequential process where models are trained iteratively, with each new model correcting errors of the previous one. It employs gradient-based changes to improve the predictions of each decision tree, which is a weak learner. All the final predictions are made by combining the output of all trees using a weighted sum of their predictions.

XGBoost is initialized with specific parameters such as use\_label\_encoder to prevent warnings related to label encoding, eval\_metric to specify the evaluation metric to optimize during training and random\_state to ensure reproducibility. After training on a fold, predictions are made on the test subset. Predictions include both class labels predict() and probabilities.

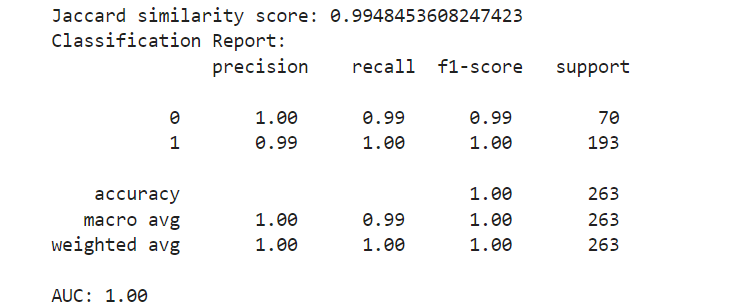


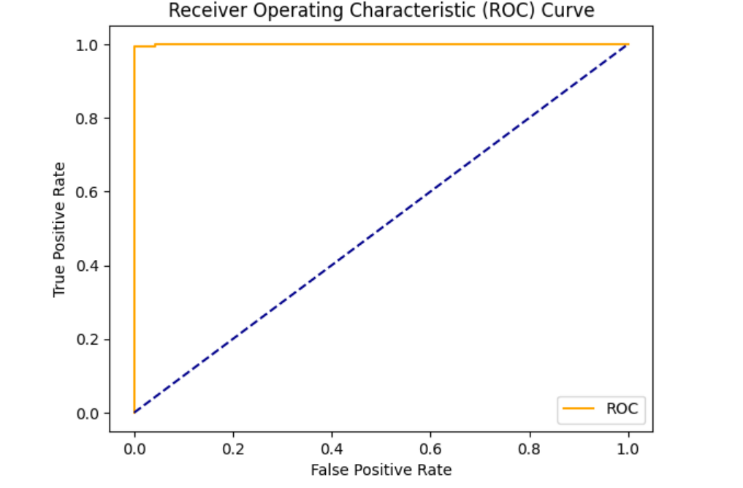


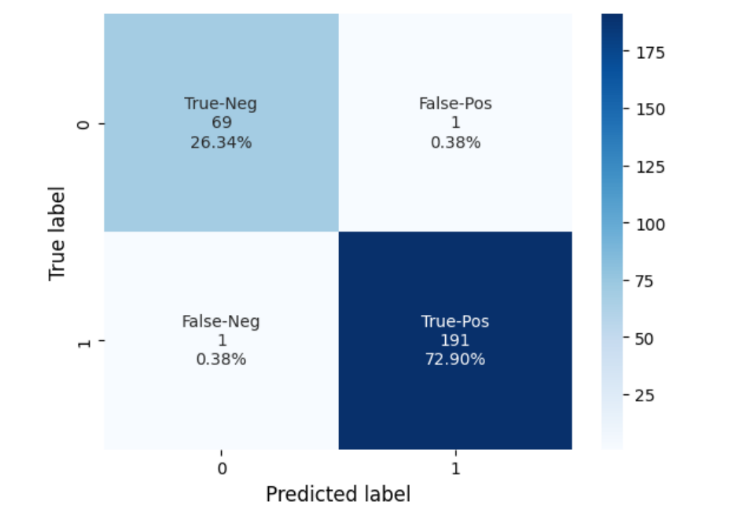


## Generalized Linear Model (GLM)

To manage a range of response variables, a Generalized Linear Model (GLM) expands on linear regression. One popular GLM for binary classification issues is logistic regression. It generates probabilities confined between 0 and 1 by modeling the likelihood of a target class using the logistic function. During cross-validation, logistic regression is set up and trained for every fold. The target class's probability is estimated by fitting a logistic curve. In order to optimize the likelihood of the observed data, the model computes coefficients for the features. The weights of each iteration are updated to minimize the cost function and the magnitude of the coefficients reflects the importance of the corresponding features. While naturally binary, logistic regression is capable of handling multi-class problems through techniques like one-vs-rest (OvR).







# Results and discussion

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Generalized Linear Model | 99.48 |
| XGBoost | 98.45 |
| Naïve Bayes | 96.89 |
| KNN | 96.90 |
| SVM | 88.94 |

The table compares the accuracy percentages of different machine learning classifiers. Here is a summary :

* With the highest accuracy of \*99.48%\*, the \*Generalized Linear Model (GLM)\* exhibits good predictive performance in the evaluated task.
* Next in line is \*\*XGBoost\*, a powerful ensemble learning technique with an accuracy of \*98.45%.
* \*K-Nearest Neighbors (K-NN)\* and \*Naive Bayes\* have nearly identical accuracies, at \*96.89%\* and \*96.90%\*, respectively.
* This dataset shows that \*Support Vector Machine (SVM)\* is less successful than the other models, with the lowest accuracy of \*88.94%\*.

# Conclusion

In this study, we used the TADPOLE dataset to examine the effectiveness of several machine learning models, such as Support Vector Machines (SVM), XGBoost, and Logistic Regression (GLM), in detecting the early development of Alzheimer's disease. To guarantee a reliable and objective performance evaluation, each model was thoroughly trained and assessed using k-fold cross-validation. Key performance indicators like Jaccard Similarity Score, ROC-AUC, classification accuracy, and confusion matrix analysis were the focus of the validation and testing stages.

Because of its capacity to manage intricate relationships and produce precisely calibrated probability estimates, our results show that while all three models performed competitively, XGBoost was the most successful in terms of predicted accuracy and area under the ROC curve. Despite its simplicity, logistic regression provided competitive performance and interpretable findings, making it a good choice in situations where model explainability is crucial. Strong classification abilities were shown by SVM, especially when it came to spotting subtle patterns in the data.

These findings highlight how machine learning may help with early Alzheimer's disease detection, allowing for prompt treatments and individualized treatment. To further improve predicted accuracy, future research could build on our work by adding other biomarkers, longitudinal data, and ensemble approaches. Furthermore, the therapeutic application of these findings would be strengthened by including domain expertise into feature selection and model interpretation.

The revolutionary impact of AI and machine learning in healthcare is demonstrated by this study, opening the door for creative answers to urgent medical problems like neurodegenerative diseases.

##### References

1. P. Kishore, Ch. Usha Kumari,, M.N.V.S.S. Kumar c, T. Pavani. Detection and analysis of Alzheimer’s disease using various machine learning algorithms
2. Nahush Kulkarni, Gayatri Bangar, Sameer Arun Mahajan. Machine Learning Algorithms for Classification of Various Stages of Alzheimer's Disease: A review. 2395-0056
3. Lin Liu , Shenghui Zhao , Haibao Chen , Aiguo Wang. A New Machine Learning Method for Identifying Alzheimer’s Disease. <https://doi.org/10.1016/j.simpat.2019.102023>
4. Clifford R. Jack Jr., David A. Bennett, Kaj Blennow, Marilyn S. Albert, Steven J. Janssen, et al. NIA-AA Research Framework: Toward a Biological Definition of Alzheimer’s Disease. <https://doi.org/10.1016/j.jalz.2018.02.018>
5. Sara Pujol, Nick Weiner, et al. Hippocampal Subfield Analysis with Machine Learning Models for Alzheimer’s Disease Detection. <https://doi.org/10.1109/ACCESS.2021.3071843>
6. Rosa I. Arrieta-Cruz, Juan Carlos Martinez-Bautista, José M. Ramirez-Pedraza, et al. Evaluating Machine Learning Techniques for Alzheimer’s Disease Detection. DOI : s00500-021-05918-9