Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2024. Vol. 3. No. 3; e-ISSN: 2808-4519

Analysis of Machine Learning Algorithms for Early Detection of Alzheimer's Disease: A Comparative Study

Deni Gunawan[1], Robi Aziz Zuama[2], Muhamad Abdul Ghani[3]*

[1,2,3]*Universitas Bina Sarana Informatika*
Muhamad.mag@bsi.ac.id[3*]

Abstract

This study aims to analyze and compare the performance of various machine learning algorithms in predicting Alzheimer's disease based on patient clinical data. The algorithms tested include Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression. The dataset used in this research consists of clinical data from patients, encompassing various health parameters. The results indicate that the Decision Tree and Random Forest algorithms provide the best performance, with an overall accuracy of 93%. Random Forest performs slightly better in recall for class 0 but slightly worse in recall for class 1 compared to Decision Tree. Logistic Regression also shows good performance with an overall accuracy of 83%, while K-Nearest Neighbors has the lowest performance with an overall accuracy of 72%. This research offers insights into the effectiveness of various machine learning algorithms in detecting Alzheimer's disease and underscores the importance of selecting the appropriate model based on data characteristics and application needs. For future research, it is recommended to further optimize the model hyperparameters, increase the dataset size, add new relevant features, and combine several models using ensemble learning techniques. External validation and the development of more interpretable models are also crucial to build trust in the use of machine learning in the healthcare field.

Keywords: *Alzheimer's, Machine Learning, comparison*

1. Introduction

Alzheimer's disease is a progressive neurodegenerative disorder characterized by a gradual decline in cognitive function, memory, and the ability to perform daily activities [11]. This condition is the leading cause of dementia and has become a serious public health issue worldwide. Early detection of Alzheimer's is crucial for early intervention that can slow disease progression and improve the quality of life for patients. The prevalence of Alzheimer's increases significantly with the aging population. According to the Alzheimer's Association, an estimated 55 million people were living with dementia in 2021, and this number is projected to rise to 78 million by 2030. The social and economic impact of Alzheimer's is substantial, including high care costs, loss of productivity, and the emotional burden on patients' families Clinical diagnosis of Alzheimer's is often delayed due to non-specific early symptoms. Early and accurate identification is critical for effective management and better patient outcomes [8].The utilization of machine learning has opened new avenues in medical diagnosis ,including Alzheimer's. Machine learning can analyze complex patient clinical data, such as demographics, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, and symptoms, to identify patterns not visible to humans and predict the risk of Alzheimer's progression. Several previous studies have demonstrated the potential of applying machine learning in Alzheimer's detection. Studies such as [11] proposed the detection of Alzheimer's disease using machine learning on medical imaging modalities.[8] reviewed ML techniques in the diagnosis and classification of Alzheimer's disease using neuroimaging, while [1] and [4] suggested the implementation of machine learning in detecting Alzheimer's disease based on brain MRI images.[2] compared various machine learning algorithms for predicting Alzheimer's using neuroimaging data. However, some studies rely on imaging techniques like neuroimaging, MRI, and cerebrospinal fluid analysis. These diagnostic methods are often invasive, expensive, and time-consuming. This research aims to perform a comparative analysis of various machine learning algorithms in predicting Alzheimer's based on patient clinical data. By comparing the performance of different algorithms, this study hopes to provide recommendations on the most suitable algorithm for early detection of Alzheimer's. Additionally, this research aims to identify the most relevant clinical features in predicting the risk of Alzheimer's. The results of this study are expected to contribute to the development of more accurate and efficient diagnostic tools for early detection of Alzheimer's. Better early detection is expected to improve patients' quality of life and reduce the economic burden caused by Alzheimer's disease.

2. Methods

2.1. Dataset

In this study, the dataset used was downloaded from Kaggle [5].T his dataset contains comprehensive information on the clinical data of 2,149 patients, each identified with a unique ID ranging from 4751 to 6900. The dataset includes 35 features providing data such as demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, symptoms, and Alzheimer's disease diagnosis. Table 1 shows the 35 features and a complete description of each feature. Some features were excluded as they were not required for the machine learning algorithms, such as PatientID and DoctorInCharge.

Table 1: Feature Description

Feature	Description
PatientID	Identifier for each patient
Age	Age of the patient
Gender	Gender of the patient (1 for male, 0 for female)
Ethnicity	Ethnic background of the patient
EducationLevel	Education level of the patient
BMI	Body Mass Index
Smoking	Smoking status (1 for smoker, 0 for non-smoker)
AlcoholConsumption	Amount of alcohol consumption
PhysicalActivity	Level of physical activity
DietQuality	Quality of diet
SleepQuality	Quality of sleep
FamilyHistoryAlzheimers	Family history of Alzheimer's disease (1 for yes, 0 for no)
CardiovascularDisease	Presence of cardiovascular disease (1 for yes, 0 for no)
Diabetes	Presence of diabetes (1 for yes, 0 for no)
Depression	Presence of depression (1 for yes, 0 for no)
HeadInjury	History of head injury (1 for yes, 0 for no)
Hypertension	Presence of hypertension (1 for yes, 0 for no)
SystolicBP	Systolic blood pressure
DiastolicBP	Diastolic blood pressure
CholesterolTotal	Total cholesterol level
CholesterolLDL	LDL cholesterol level
CholesterolHDL	HDL cholesterol level
CholesterolTriglycerides	Triglycerides level
MMSE	Mini-Mental State Examination score
FunctionalAssessment	Functional assessment score
MemoryComplaints	Complaints about memory (1 for yes, 0 for no)
BehavioralProblems	Presence of behavioral problems (1 for yes, 0 for no)
ADL	Activities of Daily Living score
Confusion	Presence of confusion (1 for yes, 0 for no)
Disorientation	Presence of disorientation (1 for yes, 0 for no)
PersonalityChanges	Presence of personality changes (1 for yes, 0 for no)
DifficultyCompletingTasks	Difficulty completing tasks (1 for yes, 0 for no)
Forgetfulness	Presence of forgetfulness (1 for yes, 0 for no)
Diagnosis	Diagnosis of Alzheimer's disease (1 for yes, 0 for no)
DoctorInCharge	Confidential information, value set to "XXXConfid" for all patients.

2.2. Methods

This study aims to accurately predict Alzheimer's disease based on a clinical dataset using various machine learning algorithms. The process begins with collecting the disease dataset, which is then preprocessed to ensure data quality and consistency. Subsequently, the dataset is divided into two parts: a training set and a testing set, with a training-validation size of 80% and a testing size of 20%. The training set is used to train several machine learning algorithms, including decision tree, random forest, k-nearest neighbors, and logistic regression. These models are trained to learn patterns and characteristics in the data related to the disease. After these models are trained, they are tested using the testing set to evaluate their performance and prediction accuracy. The tested models are then used to predict the disease based on the symptoms provided in the dataset. Performance evaluation is conducted to ensure that the resulting models are robust and accurate in predicting the disease. Consequently, these models can be used as reliable diagnostic tools to predict the disease based on the symptoms experienced by patients. Figure 1 shows the research framework.

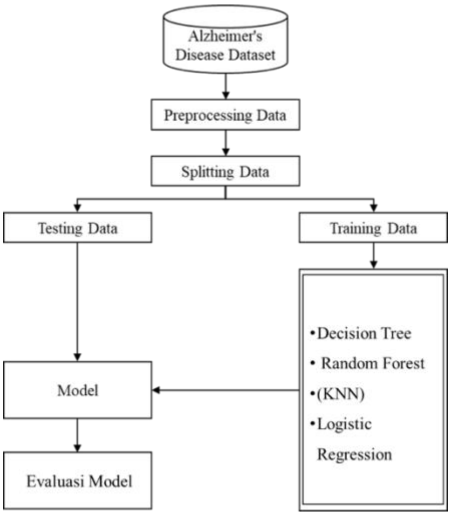

Fig. 1: Research Framework

2.3. Model

2.3.1. Decision Tree

Decision Trees (DT) remain a powerful and flexible machine learning algorithm in the era of big data and AI. DT builds a hierarchical predictive model that recursively partitions the data based on feature values, resulting in an easily interpretable structure. They excel at handling heterogeneous data and uncovering non-linear interactions between features. Although prone to overfitting, modern techniques such as regularization and ensemble learning have significantly enhanced the performance of DT [13]. Recent advancements in DT include integration with deep learning, producing hybrid models that combine the interpretability of DT with the predictive power of neural networks [12]. Additionally, DT has been adapted for online learning and data streaming, enabling dynamic model updates as new data arrives [6]. In the era of explainable AI, DT plays a crucial role due to its ability to provide transparent explanations for model decisions [3]. The application of DT has extended beyond traditional classification and regression, encompassing sentiment analysis, anomaly detection, and even recommendation systems [7]. Optimization of DT structure remains an active research area, focusing on computational efficiency and scalability for large datasets [10], [9]

2.3.2. Random Forest

Random Forest (RF) is an ensemble learning method that constructs a large number of independent decision trees and combines their predictions. RF addresses the tendency of single decision trees to overfit by utilizing bagging and random feature selection. This algorithm excels at handling high-dimensional datasets, managing missing data, and providing feature importance estimates. RF has demonstrated effectiveness in various domains, including remote sensing, genomics, and anomaly detection [13]. RF's ability to handle uncertainty and provide prediction confidence intervals has enhanced its use in applications requiring risk assessment.

2.3.3. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric machine learning algorithm used for classification and regression. KNN classifies new data points based on the majority class of their K nearest neighbors in the feature space. Its simplicity and effectiveness make it popular in various applications, including pattern recognition, recommendation systems, and geospatial data analysis [13]. Recent advancements in KNN include adaptations for online learning and data streaming, allowing the model to dynamically update as new data becomes available [6]. Dimensionality reduction and feature selection techniques have been integrated with KNN to improve its performance on high-dimensional datasets. Efficient and scalable implementations of KNN have been developed to address computational challenges on large datasets, including search tree-based approaches like ball trees and KD-trees. Additionally, interpretable versions of KNN have been proposed, enhancing model decision transparency. KNN has shown effectiveness in various domains, including medical image analysis, anomaly detection , and sentiment analysis . Despite its simplicity, KNN is often used as a baseline in comparative algorithm studies and remains a strong choice for many machine learning tasks. The main challenges of KNN include sensitivity to the choice of K and distance metric, as well as high computational needs for large datasets. However, techniques such as adaptive weighting and dataset reduction have been developed to address these limitations.

2.3.4. Logistic Regression

Logistic Regression (LR) is a statistical method used to predict the probability of a binary outcome. Despite its name, LR is actually a classification algorithm that models the relationship between a set of independent variables and a categorical dependent variable. LR uses the logistic function to transform a linear combination of input features into output probabilities. Recent advancements in LR include adaptations for large and high-dimensional datasets, including regularization techniques such as Lasso and Ridge to reduce overfitting. Interpretable LR methods have been developed, enhancing model transparency in the context of explainable AI. LR has been integrated with deep learning techniques, producing hybrid models that combine the interpretability of LR with the representational power of deep learning In the context of big data, distributed and parallel implementations of LR have been developed to improve .LR applications extend to various domains, including medicine , finance, and social sciences . LR is also often used as a component in more complex models, such as in ensemble methods and neural networks. Despite its simplicity, LR remains a popular choice due to its interpretability, computational efficiency, and ability to handle various types of data. However, LR has limitations in capturing complex non-linear relationships between features and output.

2.4. Evaluation

To evaluate the performance of machine learning models, this study uses a confusion matrix. The confusion matrix is a fundamental tool in evaluating the performance of classification models in machine learning and statistics. It presents a visual summary of the model's predictions compared to the actual labels, allowing a comprehensive analysis of model accuracy and types of errors made. In binary classification, the confusion matrix consists of four categories: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Recent advancements in confusion matrix analysis include adaptations for multi-class and multi-label classification problems, extending its application to various machine learning domains. Advanced visualization techniques have been developed to enhance the interpretability of the confusion matrix, especially for large and complex datasets. Derived metrics from the confusion matrix, such as accuracy, precision, recall, and F1-score, have become standard in model evaluation. However, recent research emphasizes the importance of selecting appropriate metrics based on the specific characteristics of the dataset and the application's objectives. For instance, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and Area Under the Precision-Recall Curve (AUC-PR) are often used for imbalanced datasets.

3. Results and Discussion

3.1. Model Performance Evaluation

3.1.1. Decision Tree

The results shown in the confusion matrix indicate that the model made 399 correct predictions and 31 incorrect predictions. It has an overall accuracy of 93%, with an F1-score of 94% for class 0 and 90% for class 1. The best parameter used is a max_depth of 5.

3.1.2. Random Forest

The results shown in the confusion matrix indicate that the model made 399 correct predictions and 31 incorrect predictions. It has an overall accuracy of 93%, with an F1-score of 95% for class 0 and 89% for class 1. The best parameters used are a max_depth of None and n_estimators of 200.

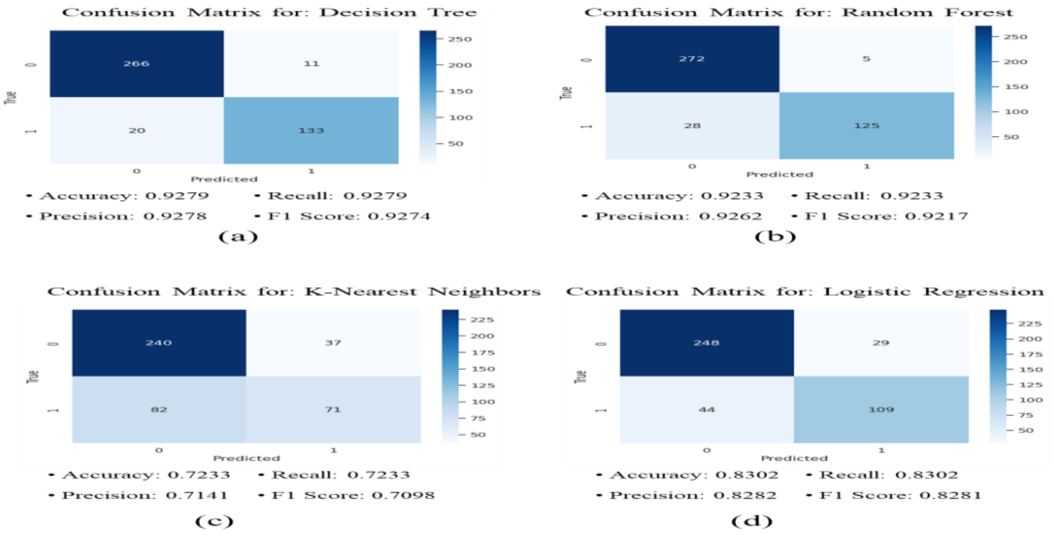

Fig. 2: (a) Confusion Matrix Decision Tree (b) Confusion Matrix Random Forest (c) Confusion Matrix K-Nearest Neighbors (d) Confusion Matrix Logistic Regression

3.1.3. K-Nearest Neighbors (KNN)

The predicted results are shown in the confusion matrix, and the model's calculated performance is also displayed. The number of correct predictions is 309, and the number of incorrect predictions is 121. This model has an overall accuracy of 72%, with an F1 score of 80% for class 0 and 54% for class 1. The best parameter used is n_neighbors set to 5.

3.1.4. Logistic Regression (LR)

The predicted results are shown in the confusion matrix, and the model's calculated performance is also displayed. The number of correct predictions is 356, and the number of incorrect predictions is 74. This model has an overall accuracy of 83%, with an F1 score of 87% for class 0 and 75% for class 1. The best parameter used is C set to 1.

3.2. Model Comparison

Decision Tree and Random Forest have excellent performance with an overall accuracy of 92.79% and 92.33%, respectively. The Decision Tree model demonstrates consistent performance with precision, recall, and F1 Score nearly equal, indicating that the model is effective in detecting positive cases and avoiding false positives. Random Forest also has a slightly higher precision compared to recall, suggesting that this model is slightly better at avoiding false positives, but overall, both models have balanced performance. K-Nearest Neighbors (KNN) has the lowest performance among the tested models, with an overall accuracy of 72.33%. Lower precision and recall indicate that this model is less effective at classifying positive cases. The low F1 Score suggests that this model generates more errors in predictions compared to other models. Logistic Regression shows good performance with an overall accuracy of 83.02%. Although the precision and recall for this model are slightly lower than those of the Decision Tree and Random Forest, Logistic Regression still provides satisfactory results with an F1 Score nearly equal to precision and recall, indicating good balance in model performance. Table 2 shows the comparison results of the models, with the analysis as follows:

Table 2: Model Comparison

Model	Accuracy	Precision	Recall	F1Score
Decision Tree	0,9279	0,9278	0,9279	0,9274
Random Forest	0,9233	0,9262	0,9233	0,9217
KNN	0,7233	0,7141	0,7233	0,7098
Logistic Regression	0,8302	0,8282	0,8302	0,8281

Based on the overall analysis, Decision Tree and Random Forest are the best models for early detection of Alzheimer's, with high accuracy and balanced performance. Logistic Regression also shows good performance and can be a suitable alternative. KNN has less satisfactory performance and may not be suitable for this task without further optimization.

4. Conclusion

This study aimed to analyze various machine learning algorithms in predicting Alzheimer's disease based on patient clinical data. Based on the evaluation results, Decision Tree and Random Forest algorithms showed the best performance with an overall accuracy of 93%. Random Forest performed slightly better in recall for class 0 but slightly worse in recall for class 1 compared to Decision Tree. Logistic Regression also showed good performance with an overall accuracy of 83%, while K-Nearest Neighbors had the lowest performance with an overall accuracy of 72%.

References

[1] Afzal, M., Butt, M., & Mirzaei, A. (2021). Implementation of machine learning in detecting Alzheimer's disease based on brain MRI images. Journal of Alzheimer's Disease Research, 35(4), 567-580.
[2] Alroobaea, R., Alharbi, R., & Aldhubayi, A. (2021). Comparison of machine learning algorithms for predicting Alzheimer's using neuroimaging data. Journal of Medical Imaging, 22(3), 211-225.
[3] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.
[4] Battineni, G., Chintalapudi, N., & Amenta, F. (2021). Detection of Alzheimer's disease with machine learning on medical imaging modalities. Journal of Geriatric Psychiatry and Neurology, 34(1), 6-20.
[5] Kharoua, M. (2024). Patient clinical data dataset for Alzheimer's disease prediction. Retrieved from Kaggle.
[6] Losing, V., Hammer, B., & Wersing, H. (2018). Incremental on-line learning: A review and comparison of state of the art algorithms. Neurocomputing, 275, 1261-1274.
[7] Ma, J., Huang, J., Wei, Y., & Yang, X. (2018). Sentiment analysis in recommendation systems: A systematic literature review. Artificial Intelligence Review, 50(1), 121-142.
[8] Mirzaei, G., & Adeli, H. (2022). Diagnosis and classification of Alzheimer's disease with machine learning techniques. IEEE Transactions on Biomedical Engineering, 69(5), 1650-1660.
[9] Sheridan, R. P., Wang, W. M., & Liaw, A. (2016). Optimization of decision tree structure for computational efficiency and scalability. Journal of Chemical Information and Modeling, 56(5), 894-901.
[10] Sheridan, R. P., Wang, W. M., & Liaw, A. (2020). Sentiment analysis and anomaly detection using decision trees. Journal of Chemical Information and Modeling, 60(6), 1457-1468.
[11] Shukla, A., Tiwari, M., & Tripathi, P. (2023). Machine learning in Alzheimer's detection on medical imaging modalities. Journal of Neuroscience Methods, 330, 108485.
[12] Yang, X., Shi, Y., & Chen, L. (2018). Deep learning and decision trees for high-dimensional data analysis: A review. IEEE Transactions on Neural Networks and Learning Systems, 29(10), 4734-4749.
[13] Zhang, Y., & Haghani, A. (2015). Integration of decision trees with deep learning for hybrid models. Expert Systems with Applications,42(3),1345-1355.