Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 8, No. 3, November 2025, pp. 558 – 568
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Advanced Machine Learning Implementation for Early Detection and Prediction of Alzheimer's Disease

[1]Christian Petrus Silalahi, [2*]Sri Lestari
[1,2]Department of Informatics Engineering, Institut Informatika dan Bisnis Darmajaya, Indonesia
Email: [1]christian.2111010017@mail.darmajaya.ac.id, [2]srilestari@darmajaya.ac.id

Article Info
Article history:
Received Aug 13th, 2025
Revised Sep 05th, 2025
Accepted Sep 20th, 2025

Keyword:
Decision Tree
Deep Learning
Machine Learning
Naïve Bayes
Random Forest

ABSTRACT
Early detection of Alzheimer's disease is essential for more effective patient care. This study explores the application of Machine Learning (ML) algorithms in detecting Alzheimer's disease by analyzing influential factors, such as demographic profile, medical history, and clinical examination results. Five ML methods, namely Deep Learning, Random Forest, Decision Tree, Naïve Bayes, and Logistic Regression, are used to classify Alzheimer's disease cases. In addition, the study used RFE and BPSO methods for feature selection with the aim of improving model performance. The evaluation was conducted using cross-fold validation and split-validation techniques, with performance measured in terms of accuracy, precision, recall, and F1-score. The results showed that the Random Forest algorithm combined with BPSO achieved the best performance, with 99% accuracy and high precision and recall values, surpassing other methods. These findings demonstrate that integrating feature selection significantly improves classification quality and confirms the practical potential of ML models as reliable tools for the early detection of Alzheimer's disease, thereby assisting clinicians in diagnostic decision-making and enhancing patient care.

Corresponding Author:
Sri Lestari,
Department of Informatics Engineering,
Institute of Informatics and Business Damajaya
Z.A. Pagar Alam 93 Road, Gedong Meneng, Bandar Lampung, Lampung, Indonesia.
Email: srilestari@darmajaya.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v8i3.38004

1. INTRODUCTION

Alzheimer's disease is a progressive neurodegenerative disorder that significantly affects cognitive function and daily activities. Early detection of the disease is critical to slowing its progression and improving patient outcomes. Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have provided promising solutions in medical diagnosis, including Alzheimer's disease detection [1]. Currently, there is no single test that can definitively diagnose Alzheimer's disease or dementia. Diagnosis usually involves a combination of neurological examinations, cognitive tests, genetic tests, brain imaging (MRI, CT, PET), as well as biomarker analysis from cerebrospinal fluid or blood, combined with the patient's medical history. ML techniques can analyze large data sets and identify complex patterns that are undetectable by conventional methods. In the context of Alzheimer's disease, ML algorithms utilize imaging data, clinical records, genetic data, and cognitive assessments to improve diagnostic accuracy. This approach enables early detection and more effective intervention [2].

Some of them use ML models for disease classification, including Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), and Deep Learning (DL). These models are chosen for their ability to handle complex medical datasets and provide interpretable results [3]. Decision Trees is a ML model that can be described by recursively dividing data based on feature values, thus capturing non-linear interactions while Random Forest is a method that can handle high-dimensional data, missing data, and provide estimates of feature importance, making it effective in fields such as remote

sensing, genomics, and anomaly detection, with its ability to assess uncertainty being particularly useful in risk assessment tasks [4].

Research by RA Saputra to classify Alzheimer's disease by comparing several Decision Tree methods combined with feature selection using the Particle Swarm Optimization (PSO) algorithm on datasets from the Alzheimer's OASIS 2. The experimental result shows that the Random Forest algorithm achieves an accuracy of 91.15% without feature selection. After applying feature selection with PSO, the PSO-based Random Forest algorithm produces the highest accuracy of 93.56% with a kappa value of 0.884. It shows that feature selection using PSO can significantly improve the accuracy of Decision Tree algorithms. [5]. Research conducted by Matthew Velazquez and Yugyung Lee shows that the Random Forest model was able to predict conversion from EMCI to AD with high accuracy (93.6%) based on clinical features. In addition, this study emphasizes the aspect of explainability by evaluating the importance of each clinical feature. This model has the potential to be applied in clinical settings, both to predict the progression of Alzheimer's disease from the prodromal stage and to identify appropriate candidates for clinical trials.

Meanwhile, this study uses several classification methods, namely Naive Bayes, Logistic Regression, Decision Tree, Random Forest, and Deep Learning [6] to classify Alzheimer's disease. And perform feature selection to improve model performance using BPSO and RFE. Feature selection is a process that aims to eliminate irrelevant features in a dataset, where the algorithm used automatically selects features that have the most significant contribution to the desired predictor or output variable. The application of feature selection before the classification model training process can improve accuracy by reducing training time and minimizing potential prediction errors caused by excessive model complexity.

In addition to classical methods such as Decision Tree, Random Forest, and Logistic Regression, recent research has shown that Convolutional Neural Network (CNN)-based models are capable of identifying structural brain changes from MRI data with high accuracy. CNNs can extract complex spatial features from brain images that are difficult for traditional methods to recognize, making them widely used in neuroimaging-based diagnostic research. Meanwhile, ensemble algorithms such as Extreme Gradient Boosting (XGBoost) have been applied to clinical data to predict the conversion from Mild Cognitive Impairment (MCI) to Alzheimer's disease, yielding competitive results in terms of both accuracy and interpretability. Therefore, this research is situated within a rapidly evolving literature landscape, where ML approaches not only enhance diagnostic accuracy but also create opportunities for more sophisticated and integrated applications across diverse patient data types [7].

2. RESEARCH METHOD

2.1. Data Preparation

Data Collection in this study referred to the process of gathering relevant data from trusted sources to support further analysis. The dataset in this study was obtained from the Kaggle platform. it provided the data related to Alzheimer's disease patients based on the Oasis Longitudinal study. The dataset included 373 rows and 16 columns, with information such as patient demographics, medical examination results, and clinical conditions used to analyze factors contributing to Alzheimer's diagnosis. Data collection was done by accessing and downloading the dataset in tabular format, which was then analyzed to explore patterns and trends in disease progression. This dataset was designed to provide comprehensive insights into the characteristics of Alzheimer's patients and the factors influencing the prediction of the disease.

The ID feature served as a unique identification for each patient, enabling longitudinal tracking of the patient's condition over time. M/F records the patient's gender (Male/ Female), which can be used to analyze differences in disease prevalence by gender. Hand indicated the patient's dominant hand (right/left), although it is not directly related to Alzheimer's diagnosis. Age records the patient's age, which was a major risk factor in the development of the disease.

The EDUC feature indicated the number of years of formal education completed by the patient. It had an effect on one's cognitive reserve against neurodegenerative diseases. Socioeconomic Status (SES) described the patient's SES, which could also be a potential risk factor. MMSE (Mini-Mental State Examination) records the patient's cognitive test results, with lower scores indicating more severe cognitive impairment.

The Clinical Dementia Rating (CDR) column was used to measure the severity of the patient's dementia, with a scale of 0 (normal), 0.5 (MCI), 1 (mild dementia), 2 (moderate dementia), to 3 (severe dementia). Estimated Total Intracranial Volume (eTIV), Normalized Whole Brain Volume (nWBV), and Atlas Scaling Factor (ASF) were features to reflect the patient's brain imaging results, which can help in the analysis of brain structure changes due to Alzheimer's disease. With this dataset, in-depth analysis could be conducted to identify factors that influence the development of Alzheimer's, as well as build disease prediction models based on patient characteristics.

2.2. Data Preprocessing

In the data pre-processing process, feature selection was performed to select the most relevant columns in the analysis of Alzheimer's disease based on the Oasis Longitudinal dataset. The features considered include Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, and ASF, as these fields provide core information regarding the patient's age, education level, SES, dementia severity, and brain imaging results. M/F features were retained as gender differences could be a factor influencing Alzheimer's risk, while features such as Hand could be removed as they had no significant association with disease diagnosis. This feature selection aimed to ensure that only data that truly supports the purpose of the analysis is used, making the analytic process more efficient and focused.

In addition to these steps, the data preprocessing phase also included handling missing values in the SES and MMSE features using the mean imputation method to maintain data consistency. Numeric features such as Age, eTIV, nWBV, and ASF were then normalized to reduce bias caused by differences in scale across variables. Categorical features, such as M/F (gender), were converted into numeric form using one-hot encoding, while less relevant features, such as Hand, were removed from the dataset to prevent unnecessary model complexity. These steps ensured that the dataset was in optimal condition prior to the model training phase.

To clarify the research flow, a flowchart was employed to illustrate the process, starting from dataset collection, data preprocessing, feature selection using RFE and BPSO, implementation of classification algorithms (Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, and Deep Learning), and finally the evaluation stage using accuracy, precision, recall, and F1-score metrics. A detailed description of these stages is presented in Figure 1.

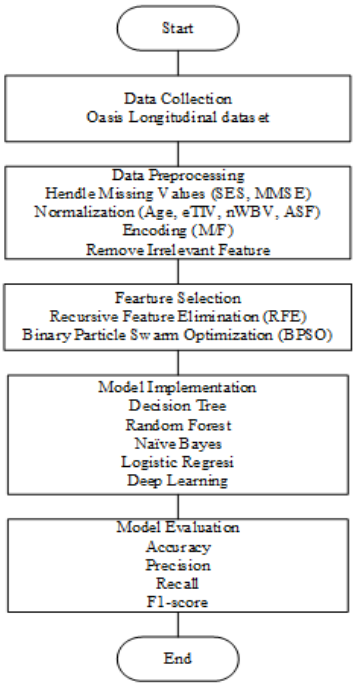

Figure 1. Research Stages

2.3. Implementation Of Algorithm

2.3.1. Decision Tree

A Decision Tree was a flowchart-like tree-shaped structure, where each internal node represented a feature or attribute, the branches indicated decision rules, and each leaf node described the final result [8]. Decision Tree algorithms have been widely used in various studies focusing on prognosis prediction. For example, Shamrat and his colleagues combined the Decision Tree algorithm with other supervised classification learning methods to predict the prognosis rate of kidney disease. Their research showed that the use of pre-processed datasets can effectively support disease prediction [9]. The process started with data pre-processing, where missing values were addressed, numerical features are normalized, and categorical variables were coded to make the data suitable for analysis. In addition, feature selection was applied to identify the most important variables that affect the likelihood of Alzheimer's, so that the model could focus more on the significant predictors. The core of the Decision Tree algorithm was its ability to recursively split

the data based on the features that provide the highest Information Gain or lowest Gini Impurity. Information Gain was calculated as the difference between the entropy of the initial dataset and the weighted sum of the entropies of the subsets generated by data sharing. Information Gain, Entropy, and Gain calculations are made using equations 1-3.

$$IG\ (D,A) = Entrophy\ (D) - \sum_{\mathrm{V}\ \in \mathrm{Values(A)}} \frac{|D|}{D}\ enthrophy\ (D_V) \quad (1)$$

Where Entropy (D) was calculated as:

$$Entrophy\ (D) = \sum_{I=1}^{K} P_I\ \log_2(P_I) \quad (2)$$

Alternatively, the Gini Impurity was calculated as:

$$Gain = Entrophy\ (S) - \sum_{I=1}^{K} P_2^I \quad (3)$$

In order to evaluate the validity of the clinical Decision Tree algorithm, accuracy testing was performed which includes sensitivity, specificity, likelihood ratio, and predictive value. In a Decision Tree, each node that is not a leaf is used to test an attribute, while the branch of the node shows the result of the test. If the test results still included data from several classes, then additional test nodes were needed on the branch to continue the classification process until reaching the final result. [10]. Decision Tree was a multilevel model that combines a series of basic tests in an efficient and integrated manner, where each test compared a numerical feature with a certain threshold value [11].

One of the main advantages of the Decision Tree algorithm was its ease of interpretation. The tree structure can be visualized to understand which features were most influential in the decision-making process. In addition, hyperparameters such as tree depth and minimum number of samples per leaf node could be optimized to maintain a balance between model complexity and prediction accuracy. This made Decision Tree a flexible and easy-to-use algorithm in various cases of predictive analysis.

2.3.2. Random Forest

Random forest is an ML model used for classification and prediction tasks. To effectively train ML algorithms and AI models, high-quality and large amounts of data were required to optimize the data collection process [12]. Random Forest was one of the most popular ML algorithms. However, the Decision Trees that make up this algorithm might have low classification accuracy or a high degree of correlation between the trees, which in turn could affect the overall performance of Random Forest [13]. Each tree was trained using a random subset of data generated through bootstrap sampling, where data points were randomly selected with replacement. At each node in the tree, the algorithm evaluated which features to use for splitting based on criteria such as Gini Impurity or Information Gain. Gini Impurity was calculated using equation 4, and Information Gain using equation 5.

$$Entrophy\ (D) = \sum_{I=1}^{K} P_I\ \log_2(P_I) \quad (4)$$

Where pi represented the probability of each class in the dataset. Alternatively, Information Gain was calculated by comparing the entropy before and after splitting, using Equation 5.

$$IG\ (D,A) = Entrophy\ (D) - \sum_{\mathrm{V}\ \in \mathrm{Values(A)}} \frac{|D|}{D}\ enthrophy\ (D_V) \quad (5)$$

Gini Split was a measurement in Decision Trees to determine how well a feature (variable) divided data into classes. Gini Split measures the impurity of data division. The lower the Gini value, the better the feature was at separating data by class. Gini Split could be calculated using equation 6.

$$Gini\ Split = \sum_{i=1}^{c} \left(\frac{n_i}{n}\right) x\ Gini\ Indeks(S_i) \quad (6)$$

The Gini Split equation was used to measure the effectiveness of data division in a Decision Tree algorithm by calculating the impurity of the division. The function of this formula was to determine the extent to which a feature can separate data into more homogeneous classes. In this formula, the contribution of each class to the overall impurity was calculated based on the proportion of data in that class and the Gini index of each class. A lower Gini Split value indicated that the feature was more effective in separating the classes so that the algorithm could select the most informative features for the classification process.

2.3.3. Naïve Bayes

The Naive Bayes algorithm is a classification algorithm based on Bayes' theorem in statistics. It was used to estimate the likelihood of data belonging to a particular class. It was called "naive" because it assumed that each feature or attribute was independent of the other. Although this was not always true in reality. Despite this simple assumption, Naive Bayes often provided quite accurate results, especially in text classification problems such as spam filters and sentiment analysis. [14]. This theory was a fundamental statistical approach used in pattern recognition. This approach was based on measuring the balance between various classification decisions made using probabilities and the impact or consequences of these decisions [15].In the context of Alzheimer's diagnosis, Naïve Bayes calculated the probability of a patient being in the Alzheimer's class (1) or Non-Alzheimer's (0) based on features such as age, MMSE score, SES, ASF, and eTIV. This model used the following formula to determine the probability of a patient falling into a particular class. The Bayes Theorem equation uses equation 7.

$$p\,(c|x) = \frac{\mathrm{p\,(C|\,X).p(c)}}{p(x)} \tag{7}$$

Since Naïve Bayes assumed that each feature was independent, the likelihood could be calculated using equation 8.

$$(C|X) = (X1|C) \times (X2|C) \ldots . \times P(XN|C) \tag{8}$$

For continuous features, the probability $(X1|C)$ was often modeled using a Gaussian (Normal) distribution, the Gaussian (Normal) distribution could be calculated using equation 9.

$$p(c|x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{-(\mathrm{xi}-\mu)2}{2\sigma 2}} \tag{9}$$

Where μ and σ were the mean and variance of the features Xi in the class C, the classification decision was made by selecting the class C that produces the highest $(C|X)$ value, so that the resulting prediction was optimal based on a probabilistic approach.

2.3.4. Logistic Regression

Logistic Regression was a classical statistical method used to model binary outcomes and chosen in medical applications due to its high level of interpretability. Although ML methods often yield better performance on high-dimensional data, their complexity makes them more difficult to understand and explain. In addition, on low-dimensional data, the performance of ML methods tends to be comparable to Logistic Regression [16].

In the context of Alzheimer's diagnosis, Logistic Regression calculated the probability of a patient being in the Alzheimer's (1) or Non-Alzheimer's (0) class based on features such as age, MMSE score, SES, ASF, and eTIV. This model used the following formula to determine the probability of a patient falling into a particular class. The probability of a class was calculated using a sigmoid function, as shown in Equation (10):

$$P(Y = 1|\mathrm{X}) = \frac{1}{1+e^{-Z}} \tag{10}$$

where z was defined as a linear combination of the input features using equation 11.

$$z = w0 + w1x1 + w2x2 + \cdots + wnxn \tag{11}$$

Classification decisions were made by setting a threshold, usually 0.5. If $(Y = 1|X)$ >0.5, then the patient was classified as Alzheimer's (1), while if $(Y = 1|X) > 0.5$, then classified as Non-Alzheimer's (0).

The advantage of Logistic Regression was its high interpretability, so it could be used to understand the relationship between clinical features and the likelihood of a person having Alzheimer's disease. The model could also be optimized with regularization such as L1 (Lasso) and L2 (Ridge) to control model complexity and improve prediction generalization.

2.3.5. Deep Learning

Deep learning was a method for training computers by utilizing artificial neural networks that mimic the way human neural networks work [17]. Deep learning enabled analysis of unstructured data as well as automatic identification of features. Recent developments in large-scale material databases had encouraged

the application of deep learning methods, particularly in atomistic prediction [18]. Deep Learning was an artificial neural network-based ML approach that could be used to diagnose Alzheimer's based on clinical features in the dataset. Compared to classical methods such as Logistic Regression or Naïve Bayes, Deep Learning was able to capture complex patterns in data thanks to its layered architecture that could perform non-linear modeling.

In the context of Alzheimer's diagnosis, Deep Learning used Artificial Neural Networks (ANN) consisting of several layers: input layers, hidden layers, and output layers. The model processed feature such as age, MMSE score, SES, ASF, and eTIV through various transformation layers to generate classification decisions.

The prediction process in artificial neural networks was based on a linear combination of given features with certain weights in each neuron; the neuron calculation could use equation 12.

$$z = WX + b \tag{12}$$

After linear transformation, the result was passed to the activation function to add non-linearity. ReLU (Rectified Linear Unit) for hidden layers could be calculated using equation 13.

$$(z) = (0, z) \tag{13}$$

During training, the model updated the weights W using an optimization algorithm such as Stochastic Gradient Descent (SGD) or Adam Optimizer by decreasing the value of the loss function, which in binary classification was often Binary Cross-Entropy. Binary Cross-Entropy can be calculated in equation 14.

$$Loss = \sum_{N}^{1}[(Y_{I \log\ (\widehat{Y_I})}] + (1 - Yi) \log\ (1 - \widehat{yi}) \tag{14}$$

Where yi was the actual label (Alzheimer or Non-Alzheimer) and $\hat{yi}$ was the model prediction.

Deep Learning enabled more complex models, such as CNN for brain image analysis from MRI, or Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) for analyzing patient longitudinal data. These models had the advantage of capturing non-linear and complex patterns, but they also required larger datasets and higher computation compared to classical methods.

2.4. Model Training

2.4.1. Before Feature Selection

The model was trained using all the features available in the dataset without first performing feature selection or filtering. In the context of the OASIS Longitudinal dataset used to predict Alzheimer's disease, the features used in model training include variables such as age, brain volume ratio (ASF), MMSE score, education level, and SES. The dataset initially consisted of 15 features, namely: Subject ID, MRI ID, Group, Visit, MR Delay, M/F, Hand, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, and ASF, before the feature selection or reduction process. The training process started with the data preparation stage, including data cleaning and transformation. Missing or incomplete data, such as blank values in the SES and MMSE columns must be handled by certain methods, for example by imputation of values or deletion of problematic rows. Some features in the dataset that had a categorical format, such as M/F (gender), needed to be converted to numerical form in order to be used in the model analysis. In addition, the uniform featured values or did not provide significant information to the prediction could be considered for deletion to reduce the complexity of the model.

The dataset was divided into two parts using a cross-validation technique, where the data was divided into K subsets with 10 folds. At each iteration, the model was trained on K-1 folds and tested on the remaining folds, thus ensuring a stronger model evaluation. Predictive models, such as Decision Tree or Random Forest were applied to the entire dataset without performing feature selection first. In this case, the model uses all available features, including features that might be less relevant than the main factors such as brain volume ratio or MMSE score in Alzheimer's diagnosis**.**

The training of the model was done by mapping the relationship between the existing features and the target variable, which was the patient's condition whether they have Alzheimer's or not. The model then learned from patterns in the training data, identified relationships between features and develops prediction rules. Using all features in the dataset without filtering might cause the model to over-rely on features that were less relevant for future predictions, thus reducing the model's ability to generalize to test or real-world data. Patterns detected in the training data might not appear in the new data, which can lead to a decrease in accuracy.

Including irrelevant or redundant features, such as age, which alone did not accurately indicate Alzheimer's without additional context reduced the model's performance by increasing complexity and limiting generalizability. Although age was a known risk factor, variables like MMSE scores or brain volume ratios provided stronger predictive value. Therefore, while training the model without feature selection gave an initial performance baseline, further evaluation was necessary to determine whether using all features improved accuracy or if feature selection simplified the model and enhanced its predictive effectiveness.

2.4.2. After Feature Selection

In ML modeling, selecting relevant features was essential to improve prediction accuracy and reduce model complexity. One commonly used method was Recursive Feature Elimination (RFE), which worked by gradually removing features that contributed the least to the model. The process began by training the model using all features, then removing features with the lowest feature importance values one by one until only the most significant features remained. Researchers often used RFE alongside algorithms such as Decision Tree or Random Forest, which were capable of assigning feature importance values. In this way, RFE helped improve model interpretation and maintained a balance between complexity and prediction accuracy. RFE selected the five most important features MMSE (rank 1), CDR (1), eTIV (1), nWBV (1), and ASF (1)—as features with rank 1 showed the greatest contribution to model accuracy. These results reflected the major role of cognitive and volumetric brain aspects in dementia classification, while other features were not selected because higher rankings indicated lower contribution, potential redundancy, or noise that could have degraded model performance if included. In addition to RFE, another method that researchers used for feature selection was Binary Particle Swarm Optimization (BPSO). BPSO was a swarm intelligence-based algorithm that adapted PSO to work in binary space, where each feature was represented as 0 (not selected) or 1 (selected).

The BPSO results identified an optimal feature subset, with MR Delay, EDUC, MMSE, and CDR emerging as the top features, and MR Delay ranked as the most critical. BPSO evaluated feature importance based on the impact of each feature's removal on model accuracy, assigning higher ranks to features whose absence caused the greatest accuracy decline. Each particle assessed model performance based on selected feature combinations and iteratively updated its position toward the best solution. Compared to elimination-based methods like RFE, BPSO explored a broader range of feature subsets, enabling more optimal selection. The combination of RFE and BPSO enhanced the feature selection process: RFE efficiently reduced irrelevant features, while BPSO refined the selection by identifying combinations that improved classifier performance.

2.5. Model Evaluation

Split validation was a common method in ML that divided the dataset into two subsets: a training set and a testing set, typically using ratios such as 80:20 or 70:30 [19]. The training set was used to fit the model, while the testing set evaluated the model's performance on unseen data. The main advantages of this approach were its simplicity and computational efficiency, as the model was trained only once. However, the method was sensitive to how the data was partitioned, potentially introducing performance bias and variability, particularly when the dataset was small or unevenly distributed.

As a more reliable alternative, cross-fold validation, commonly referred to as k-fold cross-validation, divided the dataset into k equally sized folds [20]. The model was trained and tested k times; in each iteration, k–1 folds served as the training set while the remaining fold was used for testing. This process continued until each fold had been used once as the testing set, and the final evaluation score was obtained by averaging the performance across all iterations. This method reduced variability in model assessment and provided more robust performance estimates than split validation. However, it required significantly more computational effort, particularly with larger k values, since the model needed to be trained and tested k times. Therefore, selecting a validation method required balancing dataset size, computational resources, and the desired accuracy of performance evaluation.

During the model evaluation stage, it tested the trained models using separate test data that the models had not encountered during training. They aimed to assess each model's ability to generate accurate predictions on unseen data. To evaluate performance, they employed several metrics, including accuracy, precision, recall, F1-score, Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) [22]. They used accuracy to measure the proportion of correct predictions relative to the total number of instances, as defined in Equation 15.

$$Accuracy = \frac{n\ of\ correct\ prediction}{Total\ data} \tag{15}$$

Precision measured the accuracy of the model in classifying positive instances, specifically indicating the proportion of true positive predictions among all predicted positives, as defined in Equation 16.

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Positive\ (FP)} \quad (16)$$

Meanwhile, recall measured the model's ability to detect all relevant positive data, using equation 17.

$$Recall\ = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Negative(FN)} \quad (17)$$

To provide a more balanced evaluation between precision and recall, the F1-score was used, representing the harmonic mean of the two metrics, as shown in Equation 18.

$$F-1\ Score = 2\ \times \frac{Precision \times\ Recall}{Precision\ +Recall} \quad (18)$$

In Decision Tree or Random Forest-based models, the evaluation also included analyzing the confusion matrix, which displayed the misclassifications between classes. When the model exhibited poor performance, hyperparameter tuning was applied to improve results by adjusting key parameters within the algorithm. Once the evaluation was completed and the model demonstrated satisfactory performance across the selected metrics, it was considered ready for deployment in further predictive tasks to support decision-making [23].

3. RESULTS AND ANALYSIS

This study was conducted using two scenarios, namely using the dataset directly in the model and using dataset whose important features had been selected using the Recursive Feature Elimination (RFE) and Binary BPSO methods. They compared algorithm performance by evaluating how effectively each model predicted the likelihood of an individual having Alzheimer's disease, using two validation methods: cross-fold validation and split validation. The researchers implemented Decision Tree and Random Forest algorithms for this purpose. They carried out a detailed comparison of evaluation metrics, including accuracy, precision, recall, and F1-score. In addition, they assessed the impact of feature selection using RFE and BPSO to determine how focusing on the most relevant features could enhance model performance in Alzheimer's disease detection.

3.1. Performance Comparison

The researchers conducted a performance evaluation of five ML algorithms—Decision Tree, Random Forest, Deep Learning, Naïve Bayes, and Logistic Regression—using two validation methods: cross-fold validation and split validation. They assessed each model using several evaluation metrics, including Accuracy, Recall, F1-score, RMSE, MAPE, and Precision. The analysis started with the Decision Tree algorithm, whose performance they evaluated both before and after applying feature selection using Recursive Feature Elimination (RFE). They then applied the same evaluation procedure to Random Forest, Deep Learning, Naïve Bayes, and Logistic Regression, allowing a comparative analysis of model performance before and after feature selection. The figure below presents the complete results of this evaluation.

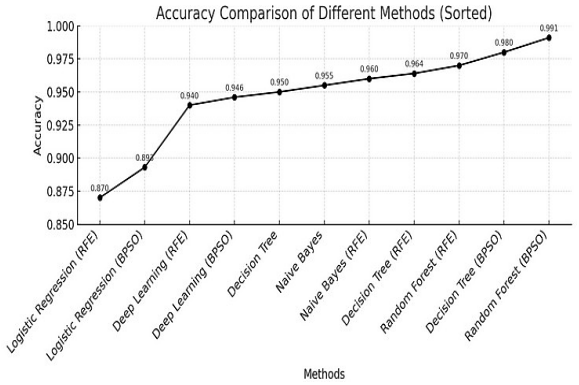

Figure 2. Results of the Performance Evaluation Accuracy the Split Validation

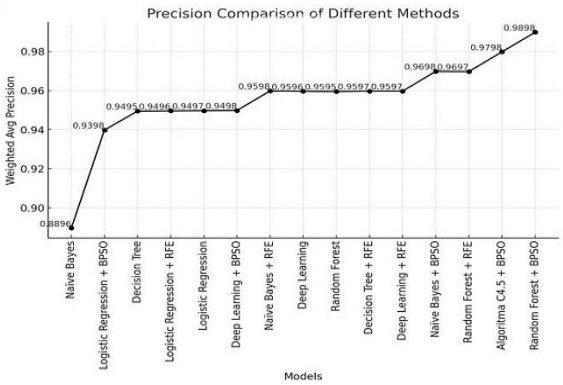

Figure 3. Results of the Performance Evaluation Precision of the Split Validation

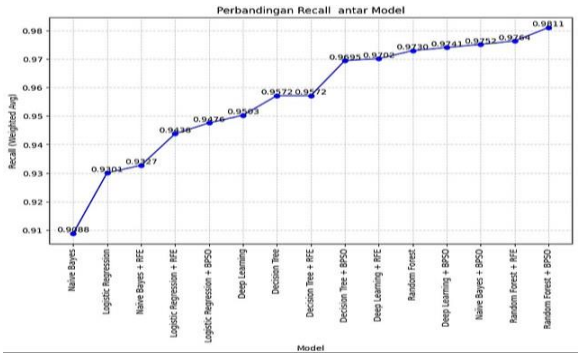

Figure 4. Results of the Performance Evaluation Recall of the Split Validation

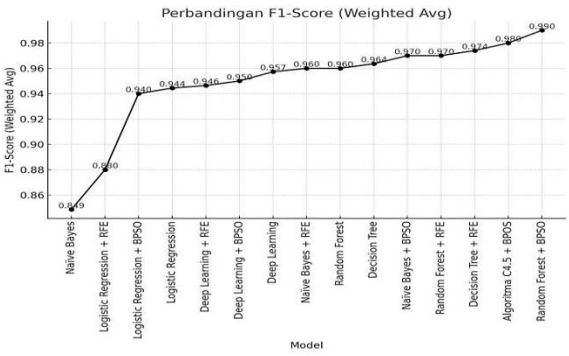

Figure 5. Results of the Performance Evaluation F1 Score of the Split Validation

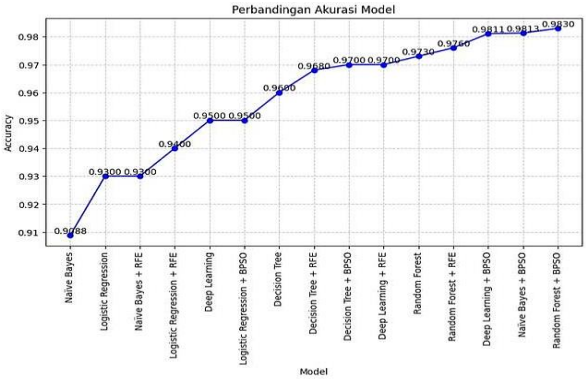

Figure 6. Results of the Performance Evaluation Accuracy of the Cross-Fold

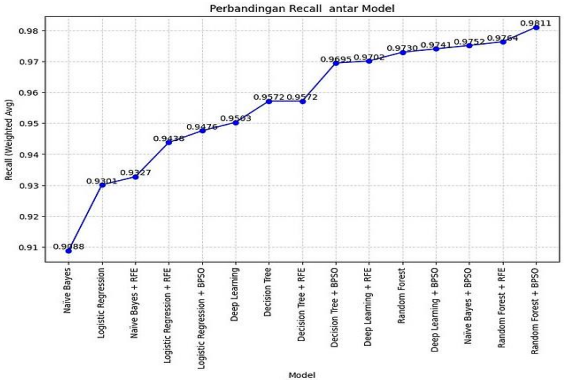

Figure 7. Results of the Performance Evaluation Recall of the Cross-Fold Validation

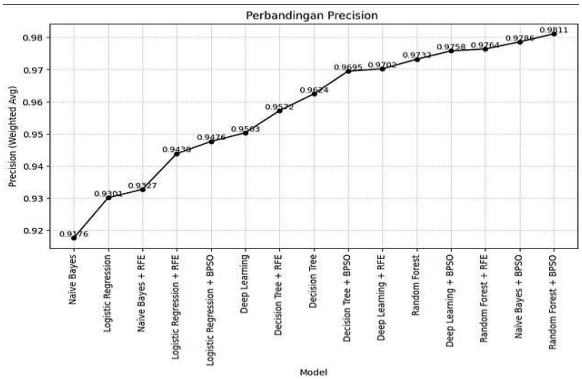

Figure 8. Results of the Performance Evaluation Precision of the Cross-Fold

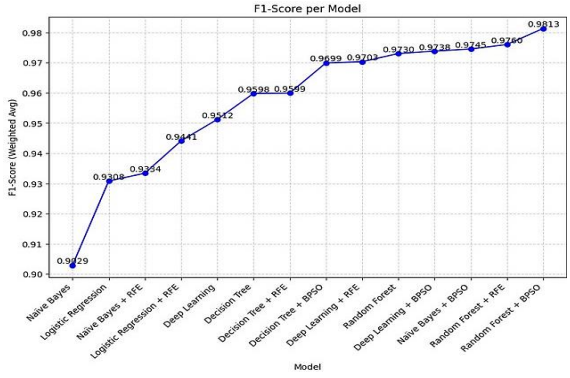

Figure 9. Results of the Performance Evaluation F1 Score of the Cross-Fold Validation

Figure 2-5 presents the evaluation results of five methods, namely Naïve Bayes, Logistic Regression, Deep Learning, Random Forest, and Decision Tree, using the split validation approach across three research scenarios, namely without feature selection, feature selection using RFE, and feature selection using BPSO. The evaluation indicated that the best performance was achieved through BPSO-based feature selection, with the Random Forest algorithm attaining the highest accuracy at 99%. Compared to the other methods, Random Forest consistently outperformed the other methods by effectively handling complex datasets and maintaining a strong balance between bias and variance. Although Deep Learning demonstrated strong potential for processing large-scale data, it demanded significantly more computational resources and involved more complex parameter tuning. Meanwhile, Naïve Bayes and Logistic Regression delivered lower performance compared to the other models in this scenario.

Similar findings were also shown in Figure 6-9. It evaluated the five methods with a cross-fold validation approach using the same three research scenarios. The results showed that BPSO feature selection

provided the most optimal results, and Random Forest was the best method with the highest accuracy. Random Forest's superiority in producing stable and accurate results in various conditions so it a superior choice compared to the other methods in this study.

In addition to comparisons based on accuracy, precision, recall, and F1-score, this study also considered computational time. The results showed that the Random Forest algorithm was relatively more efficient in training and testing compared to Deep Learning, which required longer computation times due to the complexity of its network architecture. This difference is particularly important in practical applications, where fast response times are essential in medical decision support systems.

Further analysis indicated that Random Forest's superiority over Deep Learning was related to the characteristics of the dataset. The OASIS Longitudinal dataset is relatively small, with a limited number of features, and consists largely of tabular data. Under these conditions, Random Forest performs well due to its efficiency in handling tabular data with limited variation, whereas Deep Learning requires a much larger dataset to achieve optimal performance. Thus, although Deep Learning has significant potential for analyzing large-scale data such as MRI images, in this study, Random Forest proved to be more stable, efficient, and accurate.

4. CONCLUSION

The evaluation results indicate that Random Forest consistently outperforms Decision Tree, Naïve Bayes, Logistic Regression, and Deep Learning across various evaluation metrics. After applying feature selection, particularly through the BPSO approach, Random Forest's performance improved significantly, achieving an accuracy of 99% with high precision, recall, and F1-score values. These findings demonstrate that the integration of Random Forest with BPSO provides an effective approach for the early detection of Alzheimer's disease in tabular datasets.

However, this study has several limitations. The relatively small dataset (373 samples) may restrict the model's generalizability to a broader population. In addition, the study relied solely on tabular data from the Oasis Longitudinal dataset, without incorporating multimodal data such as MRI images or genetic biomarkers, which could have enriched the analysis.

Future research should employ larger and more diverse datasets, including multimodal data (MRI, PET scans, biomarkers, and clinical records). Moreover, exploring other models such as XGBoost or CNN-based Deep Learning for brain imaging data may provide further insights. Ultimately, future studies are expected to produce more comprehensive, accurate, and clinically applicable models to support early detection and decision-making in Alzheimer's disease.

REFERENCES

[1] Franzmeier, N., Koutsouleris, N., Benzinger, T., Goate, A., Karch, C. M., Fagan, A. M., McDade, E., Duering, M., Dichgans, M., Levin, J., Gordon, B. A., Lim, Y. Y., Masters, C. L., Rossor, M., Fox, N. C., O'Connor, A., Chhatwal, J., Salloway, S., Danek, A., ... Ewers, M. (2020). Predicting sporadic Alzheimer's disease progression via inherited Alzheimer's disease-informed machine-learning. *Alzheimer's & Dementia*, *16*(3), 501-511. https://doi.org/10.1002/alz.12032

[2] Cardinali, L., Mariano, V., Rodriguez-Duarte, D. O., & Tobón Vasquez, J. A. (2025). Early detection of Alzheimer's disease via machine learning-based microwave sensing: An experimental validation. *Sensors, 25*(9), 2718. https://doi.org/10.3390/s25092718

[3] Gelir, F., Akan, T., Alp, S., Gecili, E., Bhuiyan, M. S., Disbrow, E. A., Conrad, S. A., Vanchiere, J. A., Kevil, C. G., The Alzheimer's Disease Neuroimaging Initiative (ADNI), & Bhuiyan, M. A. N. (2024). Machine learning approaches for predicting progression to Alzheimer's disease in patients with mild cognitive impairment. *Journal of Medical and Biological Engineering, 45*(1), 63–83. https://doi.org/10.1007/s40846-024-00918-z

[4] Gunawan, D., Zuama, R. A., & Ghani, M. A. (2024). Analysis of Machine Learning Algorithms for Early Detection of Alzheimer's Disease: A Comparative Study. *Journal of Artificial Intelligence and Engineering Applications*, 3(3). https://ioinformatic.org/15thJune2024.

[5] Saputra, R. A., Agustina, C., Puspitasari, D., Ramanda, R., Warjiyono, D., Pribadi, D., Lisnawanty, K., & Indriani, K. (2020). Detecting Alzheimer's Disease by the Decision Tree Methods Based on Particle Swarm Optimization. *Journal of Physics: Conference Series, 1641*(1), 012025. https://doi.org/10.1088/1742-6596/1641/1/012025

[6] Velazquez, M., & Lee, Y. (2021). Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects. *PLOS ONE, 16*(4), e0244773. https://doi.org/10.1371/journal.pone.0244773

[7] Alshamlan, H., Omar, S., Aljurayyad, R., & Alabduljabbar, R. (2023). Identifying effective feature selection methods for Alzheimer's disease biomarker gene detection using machine learning. *Diagnostics (Basel), 13*(10), 1771. https://doi.org/10.3390/diagnostics13101771

[8] Naswin, A., & Wibowo, A. P. (2023). Performance analysis of the decision tree classification algorithm on the pneumonia dataset. *International Journal of Artificial Intelligence in Medical Issues, 1*(1). https://doi.org/10.56705/ijaimi.v1i1.83

[9] S.K. Opoku, A. Y. Obeng, and M. O. Ansong, "Decision Tree Models for Predicting the Effect of Electronic Waste on Human Health'','' EJECE, vol. 7, pp. 28–34, 2023

[10] AL-Dlaeen, D., & Alashqu, A. (2014, March). *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). https://doi.org/10.1109/CSIT.2014.6805989

[11] Jijo, B. T., & Abdulazeez, A. M. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(1), 20–28. ISSN: 2708-0757

[12] Random Forest Algorithm Overview (H. A. Salman, A. Kalakech, & A. Steiti , Trans.). (2024). *Babylonian Journal of Machine Learning*, *2024*, 69-79. https://doi.org/10.58496/BJML/2024/007

[13] Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2023). An improved random forestbased on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications*, *222*, 121549. https://doi.org/10.1016/j.eswa.2023.121549

[14] Sobah, R., Fauzi, C., Arfida, S., Mutiara, S., & Nurlaila, S. (2022, December 15). *Naïve Bayes Classifier Algorithm for Predicting Non-Participation of Elections in Lampung Province*. Proceeding International Conference on Information Technology and Business, 1–9. https://darmajaya.ac.id

[15] Artaye, K. (2015, August 20–21). *Implementation of Naïve Bayes Classification Method to Predict Graduation Time of IBI Darmajaya Scholar*. International Conference on Information Technology and Business (ICITB), 284. ISSN 2460-7223.

[16] Lazarova, S., Grigorova, D., & Petrova-Antonova, D. (2023). Detection of Alzheimer's disease using logistic regression and clock drawing errors. *Brain Sciences, 13*(8), 1139. https://doi.org/10.3390/brainsci13081139

[17] ratama, R., Kurniawan, R., Rosandi, T., & Nisar. (2023). The application of the convolution neural network method uses a webcam to analyze the facial expressions of problematic students in the counseling guidance unit (Case study at SMAN 1 Penengahan Lampung Selatan). *Proceedings of the 9th International Conference on Information Technology and Business (P-ICITB)*. IIB Darmajaya. https://icitb.darmajaya.ac.id

[18] Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., Park, C. W., Choudhary, A., Agrawal, A., Billinge, S. J. L., Holm, E., Ong, S. P., & Wolverton, C. (2022). Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, *8*, Article 59. https://doi.org/10.1038/s41524-022-00734-6

[19] Jeong, S., Shivakumar, M., Jung, S.-H., Won, H.-H., Nho, K., Huang, H., Davatzikos, C., Saykin, A. J., Thompson, P. M., Shen, L., Kim, Y. J., Kim, B.-J., Lee, S., & Kim, D. (2025). Addressing overfitting bias due to sample overlap in polygenic risk scoring. *Alzheimer's & Dementia*, *21*(4), e70109. https://doi.org/10.1002/alz.70109

[20] Osterman, M. D., Song, Y. E., Lynn, A., Miskimen, K., Wheeler, N. R., Bartlett, J., Farrer, L. A., & the Alzheimer's Disease Sequencing Project (ADSP). (2024). Examining the performance of polygenic risk scores for Alzheimer's disease within and across populations using k-fold cross-validation. *Neurology: Genetics*, *10*(6). https://doi.org/10.1212/NXG.0000000000200198

BIBLIOGRAPHY OF AUTHORS

Christian Petrus is a student at the Institute of Technology and Business Darmajaya with a strong interest in the field of Artificial Intelligence. His research focuses on developing intelligent technologies that can provide innovative solutions across various sectors. He hopes that this study can offer tangible benefits to many people and contribute positively to the advancement of science and technology in Indonesia. For correspondence, the author can be reached via email at christian.2111010017@mail.darmajaya.ac.id.

Sri Lestari, obtained her Doctorate (Dr from the Electrical Engineering Doctor Program, Universitas Gadjah Mada, Yogyakarta, Indonesia in 2019. She is a lecturer in the Department of Computer Science, Institute of Informatics and Business Darmajaya, Bandar Lampung, Indonesia. Her research interests include artificial intelligence, recommendation systems, cf, data mining, decision support systems, and software engineering. She can be contacted at email: srilestari@darmajaya.ac.id.