The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 3, June 2025
https://doi.org/10.33022/ijcs.v14i3.4364

Prediction of Alzheimer Patients with Machine Learning Algorithms

Eko Priyono[1,2]
14220040@nusamandiri.ac.id[1]
[1] Mandiri University, Jakarta, Indonesia
[2] Badan Meteorologi Klimatologi dan Geofisika, Indonesia

Article Information

Submitted : 9 Aug 2025
Reviewed: 25 Apr 2025
Accepted : 9 Jun 2025

Keywords

Alzheimer's Disease, Early Detection, Machine Learning Classification, Alzheimer's Prevalence, Neurodegeneration.

Abstract

Alzheimer's disease is a neurological illness that impacts mental and emotional functions functions, has become a global concern due to its increasing prevalence. While age is the primary risk factor, other factors such as the APOE ε4 gene, family history, and brain injury also play a role. To date, there is no effective treatment for Alzheimer's, making early detection crucial. This study aims to explore early detection methods for Alzheimer's using machine learning algorithms, including transformer techniques. The results indicate that the Random Forest algorithm with Transformer methods achieved the highest accuracy of 98.9%. These findings are expected to contribute to the development of more accurate and efficient early detection strategies and improve the management of developing Alzheimer's later on.

A. Introduction

One of the most prevalent neurological diseases, Alzheimer's affects millions of individuals globally. Along with notable behavioral changes, this condition is characterized by a deterioration in memory, cognitive function, and the capacity to carry out daily tasks. While age is the primary risk factor for Alzheimer's, genetic factors such as the APOE ε4 gene, family history, and environmental factors also contribute [1]. The prevalence of Alzheimer's disease continues to rise, particularly in high-income countries where it has become one of the leading causes of death. To date, there is no effective treatment to cure or halt the progression of this disease, making early detection crucial to minimize its broader impact on individuals and society [2],[3].

This area of research focuses on the development of early detection methods for Alzheimer's disease through the application of machine learning algorithms [4],[5],[6]. In this study, several classification techniques, including transformer techniques, were used to analyze and compare the performance of these algorithms in predicting early symptoms of Alzheimer's disease. This research lies at the intersection of computer science, particularly machine learning, and the field of healthcare, with the primary goal of identifying more effective methods for the early diagnosis of Alzheimer's [7].

The primary motivation for this research is the high demand for more accurate and efficient early detection methods for Alzheimer's disease, given the lack of a cure for this condition. The rising prevalence of Alzheimer's, particularly in high-income countries, underscores the urgency of this study [8]. By developing and evaluating various machine learning algorithms, this research aims to make a significant contribution to reducing the burden of Alzheimer's in society. Furthermore, the study seeks to advance detection methods through technological innovations such as transformer techniques, which are expected to improve diagnostic accuracy and ultimately support better care planning for Alzheimer's patients [9],[10],[11].

Therefore, the utilization of machine learning methods is crucial for predicting, assessing, and anticipating Alzheimer's by taking into account dietary patterns, physical activity, and other related attributes [12]. The goal is to provide support for healthcare professionals and public health workers, particularly those in high-income areas. Various studies, including the research conducted by Faizal and his colleague Mochammad Faizal Nazili et al., have employed machine learning techniques to estimate Alzheimer's risk. Researchers used data to categorize characteristics and calculate the likelihood of an individual developing Alzheimer's based on their level of physical activity. According to their findings, the random classifier yielded the best results, achieving an overall accuracy of 90%. Additional studies exploring this subject are detailed in Table 1.

Table 1. Compare the performance of the method with current studies.

Author(S)	Publication Year	Classifier	Performance (%Accuracy)
Mochammad Faizal Nazili et al. [13]	2023	CNN	90%
Deni Gunawan et al. [14]	2024	DT, RF, LG dan KNN	93%, 90%, 83% dan 72%
R. Aarthi et al. [15]	2024	SVM, DT, NB, KNN, RF	0.79, 0.73, 0.84, 0.71, 0.81
Chinnu Mary George et	2024	XGBoost	97.8%

al. [16]			
Tripti Tripathi et al. [17]	2024	XGBoost	75.59%

B. Research Method

The procedure used to obtain the findings of the predictive analysis for Alzheimer's Disease Classification is shown in Figure 1.

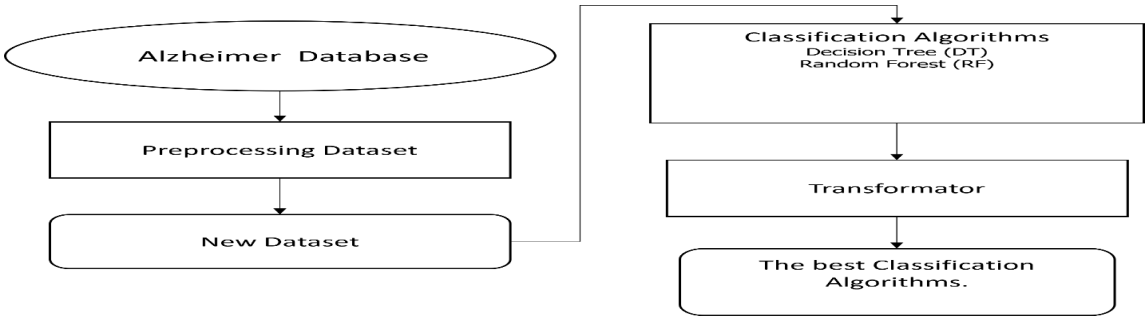

Figure 1. Research Methodology

A comprehensive analysis in this study was conducted using Python version 3.9.12. To support this analysis, several essential modules were integrated, including NumPy, Matplotlib, and Scikit-Learn. Python 3.9.12 utilizes NumPy, a library that provides extensive support for multidimensional arrays and matrices. NumPy not only offers efficient representation of numerical data but also provides a variety of high-level mathematical operations that can be applied to these arrays. These advantages are crucial in data analysis involving array manipulation and complex mathematical operations.

In the context of numerical analysis, Matplotlib acts as an extension of NumPy, specifically dedicated to data visualization and graph creation. As a plotting library for the Python programming language, Matplotlib enables clear and informative visual representation of analysis results, facilitating data understanding and interpretation.

To support the machine learning aspect of this research, Python 3.9.12 utilizes Scikit-Learn (formerly known as scikits learn or sklearn). Scikit-Learn is a free machine learning library specifically designed for use with Python. This module provides the algorithms and utility functions necessary for model training, performance evaluation, and the implementation of various machine learning techniques.

With the combination of Python 3.9.12 and these modules, comprehensive analysis can be performed efficiently and effectively. The use of NumPy for numerical data manipulation, Matplotlib for visualization, and Scikit-Learn for machine learning implementation ensures a robust and in-depth approach to data processing and interpretation.

Dataset

The source of the dataset was Kaggle ML, a Machine Learning Repository. This dataset is used for research related to Alzheimer's issues. The dataset is in CSV format. (https://www.kaggle.com/datasets/ananthu19/alzheimer-disease-

and-healthy-aging-data-in-us/code). The steps reported in the research include data collection, with data gathered through Kaggle ML, the Machine Learning Repository. The research focuses on Alzheimer's issues. Class imbalance handling involves the use of oversampling techniques, specifically SMOTE, to address class imbalance in the dataset. Attributes in the Alzheimer's dataset may include various features or variables measured or observed for each sample in the dataset. These attributes can provide information about patient or subject characteristics, Alzheimer's symptoms, or other relevant factors (see Table 2).

Table 2. Alzheimer's data attributes and descriptions

Attribute	Information
Class	Indicates a particular class or type of data related to Alzheimer's disease.
Data_Value	Data values are related to specific attributes or variables.
Data_Value_Alt	Alternative or backup data values.
Geolocation	Information about the geographic location or coordinates of the scene.
High_Confidence_Limit	High confidence limits are associated with data values.
LocationAbbr	Geographic location abbreviation.
LocationDesc	Full description of geographic location.
LocationID	Unique identification for location.
Low_Confidence_Limit	Low confidence limits are associated with data values.
Question	Questions or topics from surveys or related studies.
QuestionID	Unique identification for a question or topic.
Stratification1	Information about data stratification.
Stratification2	Additional information about data stratification.
StratificationCategory2	Additional categories of data stratification.
StratificationCategoryID2	Unique identification for additional categories of data stratification.
StratificationID1	Unique identification for data stratification.
StratificationID2	Unique identification for additional data stratification.
Topic	The topic or subject of the data or question.
TopicID	Unique identification for a topic or subject.
YearEnd	Year of the end of the data period.
YearStart	The initial year of the data period.

Preprocessing

Data Preparation Stage, handling missing data includes checking for missing values and using the mean value to fill in missing data. Data duplication management involves preprocessing steps such as checking for and removing duplicate entries in the Alzheimer's dataset. Categorical data conversion to numeric involves changing categorical data into a numeric format to ensure the

entire dataset can be processed by models using Python. Data standardization is performed to avoid the dominance of certain attributes, with the standardization method using Min-Max Normalization. The data is processed and prepared for use in the machine learning modeling process.

The plot shows data points where the X-axis represents Data_value and the Y-axis represents High_confidence_limit. Linear pattern, where points tend to form a straight line, indicates a linear relationship between the two variables. This scatter plot demonstrates the relationship between Data_Value and High_Confidence_Limit. From the plot, it is evident that there is a positive linear pattern, suggesting that as Data_Value increases, High_Confidence_Limit also tends to increase. Overall, this analysis indicates a strong relationship between the two variables (see Figure 2).

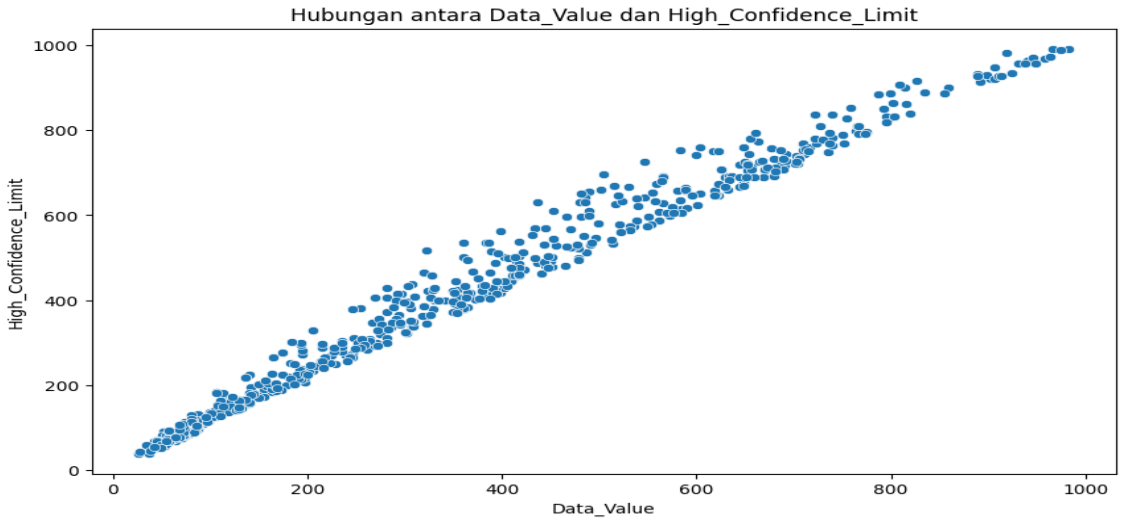

Figure 2. Visualization of the relationship between Data_Value and High_Confidence_Limit

Test Data

After completing the final preprocessing stage, the data is saved in CSV (Comma-Separated Values) format and used as input for the classification stage. The data is then split before entering the classification phase with two models: RF and DT. The dataset will be divided into training and testing data using the sklearn (scikit-learn) module [18], with Python 3.9.12. Therefore, we divide the data into 80% training data and 20% testing data. This split proportion is chosen randomly as it is a simple technique suitable for large datasets. The number of samples used in this study is 1,500, so with this split proportion, validation data is obtained.

Improving Model Accuracy Using Transformer Methods

The research data is taken from the same dataset. We select the same target variable and separate the features from the target variable. Our program includes a learning process using Random Forest Classifier, with adjustments in pre-processing using transformers. In this research, we use transformers in the form of Column Transformer from the scikit-learn library. This transformer allows different processing for numerical and categorical columns. Specifically, we apply

transformers such as StandardScaler for numerical features and OneHotEncoder for categorical features. This approach aims to enhance feature representation and ensure the model can learn effectively from different types of data.

Changes in the transformer include modifying the pre-processing approach by combining transformers for numerical and categorical columns into a single pipeline. This aims to improve code clarity and efficiency, as well as streamline the data processing process. Model Evaluation the model is tested using accuracy metrics on the test dataset, and the results are compared with the previous program. We evaluate whether changes in the pre-processing approach using transformers result in improvements in model accuracy.

Results the application of transformer methods in pre-processing successfully improved the accuracy of the Random Forest Classifier model. The model achieved an accuracy of 98.9% after changes, up from the previous 89.9%. This improvement demonstrates the significant potential of using transformer methods to enhance model performance.

Conclusion: The application of transformer methods in the pre-processing stage can bring significant benefits in improving the accuracy of predictive models for Alzheimer's disease. These findings provide a deeper understanding of the impact of transformer methods in the context of health data analysis. Future studies can further explore other types of transformers and more complex pre-processing configurations. A better understanding of the potential of transformers can guide the selection of optimal pre-processing methods to enhance model performance.

Machine learning classification methods

We used several common machine learning approaches, as explained in the sub-section below.

Random Forest (RF)

Ensembles of numerous individual decision trees, or “random forests,” are formed by using random data choices, often known as "bagging," in the RF machine learning technique. In addition to bagging, RF constructs trees by using random feature selection and random subsets of data. The most popular category is forecasted by the model, and each tree in the RF forecasts a category [19].

$$Entropy\ (S) = \sum -P1\ log2\ (\ )\ n \qquad (1)$$

i=1 = Number of partitions
S = Set of cases n, fraction of S to S = Pi

Decision Tree (DT)

DT, known as a tree diagram, is a graphical depiction of a series of choices or occurrences. It is used to assist discover the optimum course of action based on particular conditions or criteria by visualizing the processes in a decision-making process. DT often has practical significance that can be used to aid treatment decisions by drawing reasonable medical conclusions [20].

$$E(S) = \sum_{i=1}^{c} \text{Pi} - P_1 log_2 P_1 \quad (2)$$

S = stands for starting condition,
i = arrange a class on S,
Pi = likelihood or a node's share of class I

C. Result and Discussion

The goal of this project is to apply machine learning to create models for the diagnosis and management of Alzheimer's disease. The research techniques include data processing, model creation using six machine learning algorithms, and performance assessment with metrics such as accuracy, precision, recall, and F1 score. Based on the findings, Decision Tree (DT) achieved an accuracy of 97.9%, while Random Forest (RF) demonstrated superiority in diagnosing Alzheimer's, with the highest accuracy of 98.9% using Transformer methods. To create an ideal model, an automatic algorithm selection procedure was also implemented. With effective Alzheimer's detection methods, this research has important practical implications and could significantly impact public health practices, particularly in high-income areas with hedonistic lifestyles. The results indicate advancements and superiority of the proposed paradigm compared to previous research. Although further validation is needed, this study provides a strong foundation for future progress in Alzheimer's identification and management. The performance comparison of the machine learning techniques used in the research is shown in Table 3 and Table 4.

Table 3. Assess various methods.

Algorithm	Accuracy	Precision	Recall	F1 Score
Decision Tree	**0.979**	0.991	0.987	0.989
Random Forest	0.899	0.913	0.884	0.897

Table 4. Model Accuracy Using the Transformer Method

Algorithm	Accuracy	Precision	Recall	F1 Score
Metode Transformer Model RF	*0.989*	*0.98*	*1.0*	*0.99*
Metode Transformer Model DT	*0.98*	*0.99*	*0.994*	*0.99*

The ROC Curve (Receiver Operating Characteristic Curve) is a graph that illustrates the performance of a classification model across various threshold values. The ROC Curve measures the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different thresholds. True Positive Rate (TPR): Also known as Sensitivity or Recall, TPR measures how well the model can identify positive cases. The formula is TPR = TP / (TP + FN), where TP is True Positive and FN is False Negative. False Positive Rate (FPR): FPR measures how often the model incorrectly classifies negative cases as positive. The formula is FPR

= FP / (FP + TN), where FP is False Positive and TN is True Negative. In the context of the explanation, an AUC of 0.93 indicates that the model has a good level of separation between positive cases (Alzheimer's patients) and negative cases (non-Alzheimer's). With an AUC value of 0.93, you have strong evidence that your classification model is very effective in distinguishing between individuals with Alzheimer's and those without it. Therefore, an ROC Curve with a high AUC value like this represents good model performance in the context of Alzheimer's disease detection (see Figure 3).

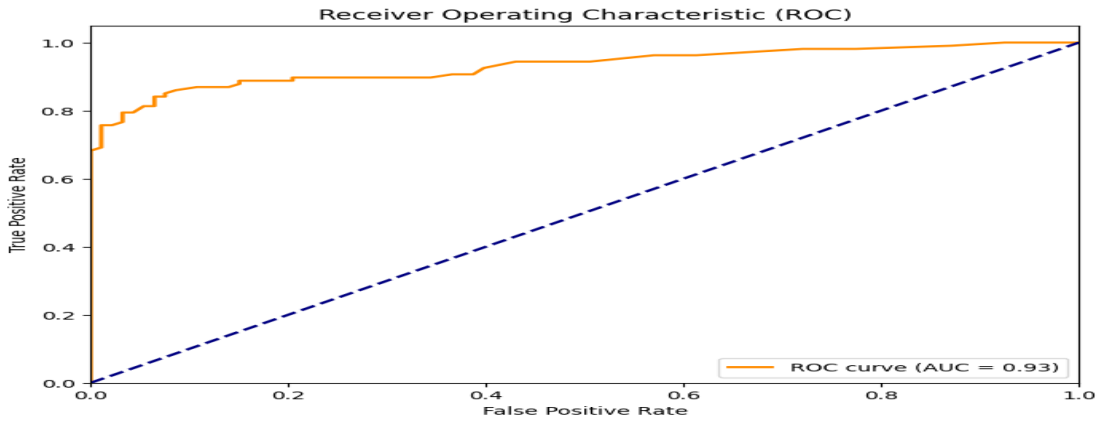

Figure 3 ROC Curve

Evaluation Metrics

Classification or prediction is one of the most controversial subjects that has garnered significant attention globally in the scientific community. Evaluating the performance of classification algorithms is crucial to determine whether a model performs well or not. In this study, we use performance evaluation metrics such as accuracy, precision, recall, and F1 score.

Our research results indicate that the Decision Tree (DT) classifier outperforms other classifiers, achieving an impressive accuracy of 97.9%. Additionally, Random Forest (RF) also demonstrates outstanding performance with an accuracy of 89.9%. However, the integration of DT, RF, and Transformer is particularly noteworthy, as it significantly enhances the effectiveness and accuracy of the model. The combination results in accuracy levels reaching 98% and 98.9%, revealing highly important risk variables.

It is important to note that the use of the proposed feature integration model with two classification algorithms on the dataset yields very satisfactory results. We compared our results with recent research, as shown in Table 1, and found that our accuracy significantly exceeds previous studies, which reported accuracy ranges between 71% and 93%.

By Mochammad Faizal Nazili et al. used CNN with 90% accuracy, which is 8.9% lower than the new model (98.9%). Research by Deni Gunawan et al. with Decision Tree (DT) achieved 93% (5.9% lower), Random Forest (RF) achieved 90% (8.9% lower), Logistic Regression (LR) achieved 83% (15.9% lower), and K-Nearest Neighbors (KNN) achieved 72% (26.9% lower). Research by R. Aarthi et al. with Support Vector Machine (SVM) achieved 0.79 (19.9% lower), Decision Tree (DT) achieved 0.73 (26.9% lower), Naive Bayes (NB) achieved 0.84 (14.9% lower), K-Nearest Neighbors (KNN) achieved 0.71 (27.9% lower), and Random Forest (RF)

achieved 0.81 (17.9% lower). Research by Chinnu Mary George et al. with XGBoost achieved 97.8% accuracy, which is only 1.1% lower than the new model. Research by Tripti Tripathi et al. with XGBoost achieved 75.59% accuracy, which is 23.31% lower than the new model.

Our research achieves higher accuracy even with the application of transformers. The superior performance of our model compared to previous studies not only demonstrates the clinical relevance potential of integrating machine learning, particularly with Random Forest (RF) algorithms and transformer techniques. This improvement impacts not only the knowledge in predicting Alzheimer's but also highlights the effectiveness of our methodology in achieving higher accuracy levels.

Our research results have significant implications for improving the understanding of the disease, contributing to better clinical decision-making, and potentially enhancing patient outcomes. The integration of machine learning with transformers in our study opens doors to a deeper understanding of Alzheimer's complexity, which could help guide more precise care strategies and interventions. Thus, our research not only focuses on scientific advancement but also offers added value in the clinical context to drive significant improvements in Alzheimer related health practices.

D. Conclusion

This research aims to evaluate the effectiveness of classification algorithms in detecting Alzheimer's disease, a neurodegenerative disorder that affects cognitive function and requires early detection. Two classification algorithms, Decision Tree (DT) and Random Forest (RF), were assessed for detection accuracy using a dataset with diverse clinical and biological features related to Alzheimer's. The results show that Decision Tree achieved the highest accuracy of 97.9%, followed by Random Forest with an accuracy of 89.9%. The use of transformer methods improved the accuracy to 98.9%. These findings provide valuable insights into the performance of classification algorithms for early detection of Alzheimer's disease and serve as a basis for developing more efficient detection methods. In the context of classification, we adopted performance evaluation using metrics such as accuracy, precision, recall, and F1 score. The Decision Tree excelled in accuracy, while the integration of DT, RF, and Transformer achieved an outstanding level of accuracy, revealing important risk variables. The significance of these findings lies in the potential clinical relevance of integrating machine learning, particularly with RF and transformer techniques. Our research contributes vital information for better clinical decision-making and has the potential to improve patient outcomes. By using feature integration models and six classification algorithms, our results significantly exceed previous research findings, underscoring substantial progress. The integration of machine learning with transformers not only reflects scientific advancement but also offers added value in a clinical context, opening doors to a deeper understanding of Alzheimer's complexity. Its implications could guide more precise care strategies and interventions, strengthening practices related to Alzheimer's health.

E. References

[1] K. Woźniak *et al.*, "Alzheimer's Disease - A Comprehensive Review," *J. Educ. Heal. Sport*, vol. 56, pp. 195–209, 2024, doi: 10.12775/jehs.2024.56.013.

[2] K. Wahlberg *et al.*, "People get ready! A new generation of Alzheimer's therapies may require new ways to deliver and pay for healthcare," *J. Intern. Med.*, vol. 295, no. 3, pp. 281–291, 2024, doi: 10.1111/joim.13759.

[3] S. Sellappan, S. P. Anand, F. D. Shadrach, B. Krishnasamy, R. Karra, and U. Annamalai, "A survey of Alzheimer's disease diagnosis using deep learning approaches," *J. Auton. Intell.*, vol. 7, no. 3, pp. 1–18, 2024, doi: 10.32629/jai.v7i3.660.

[4] E. Priyono, T. Al Fatah, S. Ma'mun, and F. Aziz, "Tubercolusis Segmentation Based on X-ray Images," *J. Med. Informatics Technol.*, pp. 101–104, 2023, doi: 10.37034/medinftech.v1i4.22.

[5] P. C. Muhammed Raees and V. Thomas, "Automated detection of Alzheimer's Disease using Deep Learning in MRI," *J. Phys. Conf. Ser.*, vol. 1921, no. 1, 2021, doi: 10.1088/1742-6596/1921/1/012024.

[6] Q. Lin, C. Che, H. Hu, X. Zhao, and S. Li, "A Comprehensive Study on Early Alzheimer's Disease Detection through Advanced Machine Learning Techniques on MRI Data," *Acad. J. Sci. Technol.*, vol. 8, no. 1, pp. 281–285, 2023, doi: 10.54097/ajst.v8i1.14334.

[7] B. Health *et al.*, "Deep Learning-Assisted Diagnosis of Alzheimer ' s Disease from Brain Imaging Data," vol. 4, no. 1, pp. 36–44.

[8] M. Ghosh, P. Chejor, M. Baker, and D. Porock, "A Systematic Review of Dementia Research Priorities," *J. Geriatr. Psychiatry Neurol.*, vol. 0, no. 0, pp. 1–12, 2024, doi: 10.1177/08919887241232647.

[9] S. S. Kumar, V. N. Sasi, and V. Murali, "Alzheimer's Patient Support System Based on IoT and ML," *J. Electron. Electr. Eng.*, 2024, doi: 10.37256/jeee.3120244607.

[10] S. D. Machado, J. E. D. R. Tavares, and J. L. V. Barbosa, "Technologies for monitoring patients with Alzheimer's disease: A systematic mapping study and taxonomy," *J. Ambient Intell. Smart Environ.*, vol. 16, no. 1, pp. 3–22, 2024, doi: 10.3233/AIS-220407.

[11] F. Bermejo-Pareja and T. del Ser, "Controversial Past, Splendid Present, Unpredictable Future: A Brief Review of Alzheimer Disease History," *J. Clin. Med.*, vol. 13, no. 2, 2024, doi: 10.3390/jcm13020536.

[12] E. Priyono and S. Ma, "Effects of Diet and Physical Activity on Coronary Heart Disease Risk Among Badminton Players," pp. 55–59, doi: 10.37034/medinftech.v2i2.36.

[13] M. F. Nazil, A. B. Firmansyah, and R. Purbaningtyas, "Klasifikasi Keparahan Demensia Alzheimer Menggunakan Metode Convolutional Neural Network pada Citra MRI Otak," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 1–7, 2023, doi: 10.57152/malcom.v3i1.200.

[14] D. Gunawan, R. A. Zuama, and M. A. Ghani, "Analysis of Machine Learning Algorithms for Early Detection of Alzheimer ' s Disease : A Comparative Study," vol. 3, no. 3, pp. 2–6, 2024.

[15] D. Chitradevi and S. Prabha, "Analysis of brain sub regions using optimization techniques and deep learning method in Alzheimer disease," *Appl. Soft*

Comput. J., vol. 86, no. February, pp. 3175–3180, 2020, doi: 10.1016/j.asoc.2019.105857.

[16] C. M. George and S. Menon, "Machine Learning for Alzheimer Detection: a Comprehensive Approach," *J. Theor. Appl. Inf. Technol.*, vol. 102, no. 4, pp. 1389–1397, 2024.

[17] T. Tripathi and R. Kumar, "ML-Based Quantitative Analysis of Linguistic and Speech Features Relevant in Predicting Alzheimer's Disease," *Adv. Distrib. Comput. Artif. Intell. J.*, vol. 13, pp. 1–20, 2024, doi: 10.14201/adcaij.31625.

[18] G. Sam, P. Asuquo, and B. Stephen, "Customer Churn Prediction using Machine Learning Models," *J. Eng. Res. Reports*, vol. 26, no. 2, pp. 181–193, 2024, doi: 10.9734/jerr/2024/v26i21081.

[19] G. James, "Sciences Analysis of support vector machine and random forest models for predicting the scalability of a broadband network," vol. 6, pp. 1–10, 2024.

[20] A. Gaballah, A. E. B. Abu-Elanien, and A. I. Megahed, "A Decision Tree Based Ultra-high-speed Protection Scheme for Meshed MMC-MTDC Grids with Hybrid Lines," *J. Electr. Eng. Technol.*, vol. 19, no. 2, pp. 887–900, 2024, doi: 10.1007/s42835-024-01808-9.