SITEKNIK
Sistem Informasi, Teknik dan Teknologi Terapan
E-ISSN: 3032-3991 P-ISSN: 3090-1626
Vol. 2. No. 3 July 2025. Pages. 231-239

Optimization of Random Forest Algorithm Using Random Search for Alzheimer's Disease Detection

Hasyim Sri Wahyudi[2], Ferian Fauzi Abdulloh[2]

hasyimelmarusy@gmail.com[1], hasyimelmarusy@gmail.com[2]
[12] Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta, Indonesia

Keywords: Alzheimer's, Random Forest, Optimization, Random Search, Machine Learning.

Submitted: 16/03/2025
Revised: 21/03/2025
Accepted: 26/03/2025

Abstrak

Alzheimer's disease is a type of neurodegenerative disorder that causes a decline in cognitive function. Early detection is crucial to enable more effective interventions and slow the progression of the disease. However, the diagnosis of Alzheimer's disease often faces challenges, particularly in detecting the early stages of the disease from complex and diverse medical data. This study aims to optimize the Random Forest algorithm using the Random Search method for detecting Alzheimer's disease. The Random Forest algorithm was applied as the primary model in this research, while hyperparameter optimization was performed using the Random Search method to improve model performance. The results showed that the Random Forest model without optimization achieved an accuracy of 96%. After performing hyperparameter optimization, the model's accuracy increased to 97%. In conclusion, the application of hyperparameter optimization using the Random Search method successfully enhanced the performance of the Random Forest model. The resulting model provides more accurate predictions, making it a reliable tool for the early detection of Alzheimer's disease.

Corresponding Author:
Hasyim Sri Wahyudi
Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta, Indonesia
Jl. Ring Road Utara 55281 Condongcatur Daerah Istimewa Yogyakarta
Email: hasyimelmarusy@gmail.com

INTRODUCTION

Alzheimer’s disease is one of the most common neurodegenerative disorders among the elderly, characterized by a progressive decline in cognitive function and memory. According to the World Health Organization (WHO), the number of Alzheimer's patients is expected to increase significantly in the coming decades, making it one of the major global health challenges (World Health Organization, 2023). Early detection of Alzheimer’s is crucial as it can help slow disease progression and improve patients’ quality of life through timely medical intervention. However, conventional detection methods, such as brain imaging and genetic testing, are often costly and require complex procedures, highlighting the need for a more efficient approach (Klyucherev et al., 2022).

In recent years, machine learning technology has been increasingly utilized to analyze complex medical data for disease detection (Van Oostveen & De Lange, 2021), including Alzheimer’s. One commonly used algorithm is Random Forest, which is known for its advantages in handling large-feature datasets and imbalanced data distributions. This algorithm has been widely applied in biomarker analysis and brain imaging to improve diagnostic accuracy. However, the performance of Random Forest highly depends on selecting the right hyperparameters. Poor hyperparameter selection can lead to suboptimal model performance in terms of both accuracy and computational efficiency (Song et al., 2021).

One commonly used hyperparameter optimization method is Grid Search, which exhaustively evaluates all possible parameter combinations (Ismail, P, & Ali, 2023). However, this method has limitations as it requires significant computational resources and time, especially when dealing with large parameter spaces. Therefore, Random Search has emerged as a more efficient alternative, as it selects hyperparameter values randomly while still having a high probability of finding an optimal combination (Elgeldawi et al., 2021).

This study aims to optimize the Random Forest algorithm for Alzheimer’s disease detection using Random Search as a hyperparameter optimization method. The dataset used in this research is sourced from Kaggle, containing medical records of Alzheimer’s patients with various health-related variables (Widyantoro, Widhiastuti, & Atlantika, 2021). By implementing this optimization, the model's performance is expected to improve, particularly in early Alzheimer’s detection. The findings of this research are anticipated to contribute to the development of machine learning-based diagnostic methods, assisting the medical field in enhancing the effectiveness of neurodegenerative disease detection, particularly Alzheimer’s disease.

RESEARCH METHODS

Random Forest is a machine learning algorithm widely used for classification tasks. It consists of multiple decision trees that independently make predictions, and the final output is determined through majority voting (Dana et al., 2024).

Classification Function:

$$y = \frac{1}{n}\sum_{i=1}^{n} hi(x) \tag{1}$$

Where:

y is the final predicted result.

n is the number of trees in the forest.

$hi(x)$ is the prediction from the i-th decision tree for input (x).

Random Search is one of the widely used methods for hyperparameter optimization. This method works by randomly selecting combinations of parameter values. Random Search identifies the optimal values by defining the lower and upper bounds for each parameter (Pramudhyta & Rohman, 2024). Over time, this method has undergone various improvements, evolving from initially finding solutions at local minimum points to now being able to identify global minimum points through modifications in the search variable length within its algorithm.

With the formula:

$$\theta \sim P(\theta) \quad (2)$$

Where:

θ is model parameter vector.

$P(\theta)$ is probability distribution of parameter θ.

The research process in this study follows these steps:

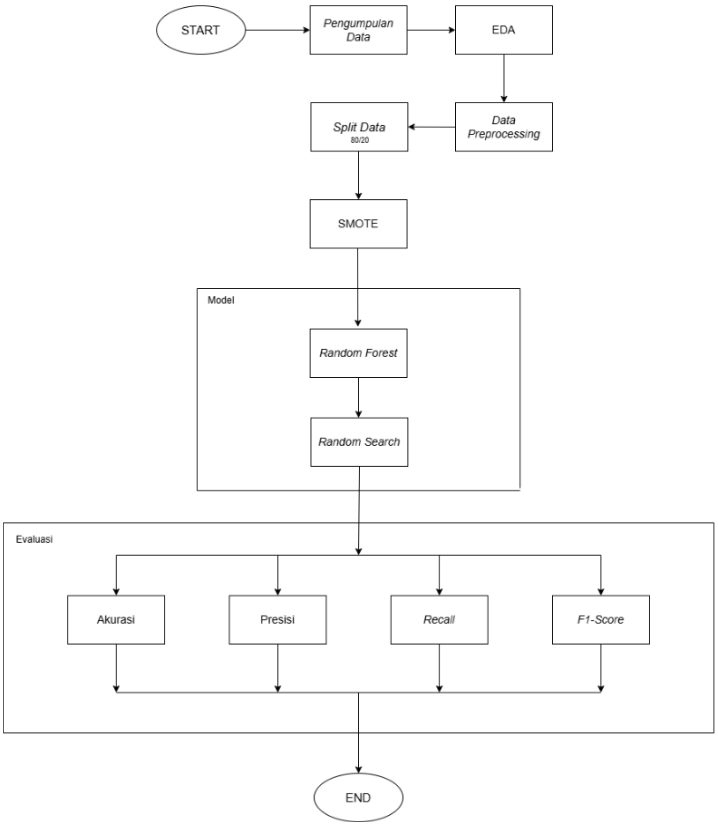


Figure 1. Research Flow

In this study, the research process consists of several stages, as follows:

Dataset Collection: The dataset used in this study was obtained from Kaggle, containing medical records related to Alzheimer's disease.

Exploratory Data Analysis (EDA): At this stage, an initial analysis was conducted to understand the dataset, identify missing values, and visualize the data distribution.

Data Preprocessing: The preprocessing phase involved handling missing values, encoding categorical data, and normalizing numerical features to ensure consistency and improve model performance (Beskopylny et al., 2022).

Data Splitting: After preprocessing, the dataset was divided into training and testing sets using an 80:20 ratio with a random state of 42 (Aprilliandhika & Abdulloh, 2024).

SMOTE (Synthetic Minority Over-sampling Technique): This technique was applied to balance the dataset by generating synthetic samples for the minority class, improving the model’s ability to classify both classes effectively (Kurniawan et al., 2023).

Model Development (Random Forest): The Random Forest algorithm was implemented to classify Alzheimer's disease based on the given features.

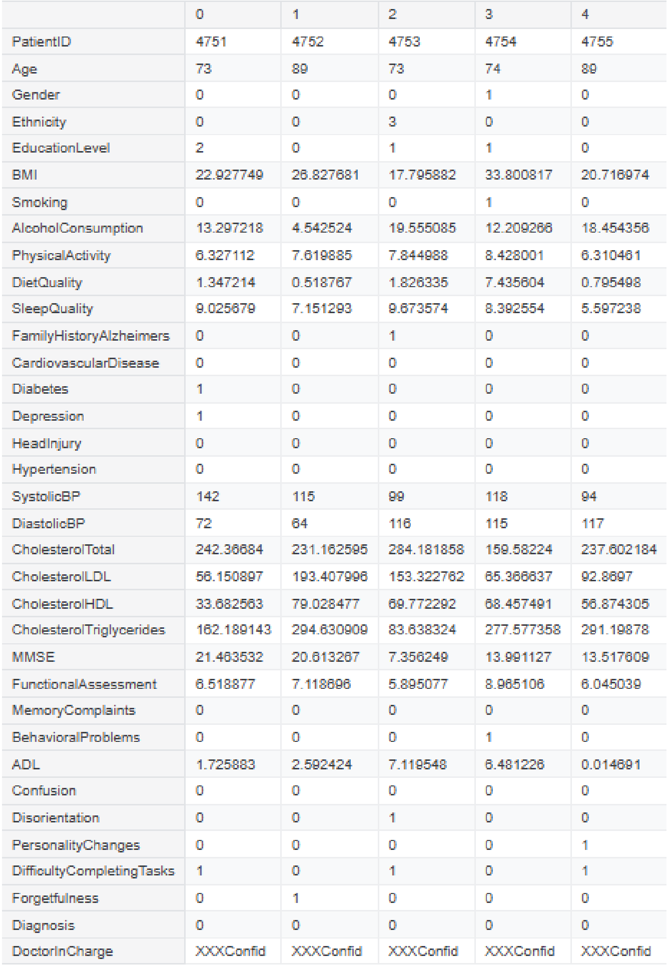
Hyperparameter Optimization (Random Search): To enhance the performance of the Random Forest model, Random Search was applied to optimize the hyperparameters and improve prediction accuracy.

Evaluation: The performance of the optimized model was assessed using accuracy, precision, recall, and F1-score to determine its effectiveness in detecting Alzheimer's disease.

RESULTS AND DISCUSSION

This study focuses on evaluating the effectiveness of the Random Forest algorithm in detecting Alzheimer's disease, incorporating hyperparameter optimization using the Random Search method. The research workflow includes data preprocessing steps such as label encoding for categorical variables, handling class imbalance with SMOTE, and applying Correlation Thresholding for feature selection. The dataset is divided into training and testing sets using an 80:20 split, and model performance is assessed based on accuracy, precision, recall, and F1-score.

At the beginning of this research, the dataset was sourced from Kaggle, a well-known open-access platform for datasets across various fields. The dataset used in this study is specifically related to Alzheimer's disease and contains multiple patient-related medical variables. It is structured in a CSV format and includes a sufficient number of samples to facilitate machine learning model training and evaluation.



	0	1	2	3	4
PatientID	4751	4752	4753	4754	4755
Age	73	89	73	74	89
Gender	0	0	0	1	0
Ethnicity	0	0	3	0	0
EducationLevel	2	0	1	1	0
BMI	22.927749	26.827681	17.795882	33.800817	20.716974
Smoking	0	0	0	1	0
AlcoholConsumption	13.297218	4.542524	19.555085	12.209266	18.454356
PhysicalActivity	6.327112	7.619885	7.844988	8.428001	6.310461
DietQuality	1.347214	0.518767	1.826335	7.435604	0.795498
SleepQuality	9.025679	7.151293	9.673574	8.392554	5.597238
FamilyHistoryAlzheimers	0	0	1	0	0
CardiovascularDisease	0	0	0	0	0
Diabetes	1	0	0	0	0
Depression	1	0	0	0	0
HeadInjury	0	0	0	0	0
Hypertension	0	0	0	0	0
SystolicBP	142	115	99	118	94
DiastolicBP	72	64	116	115	117
CholesterolTotal	242.36684	231.162595	284.181858	159.58224	237.602184
CholesterolLDL	56.150897	193.407996	153.322762	65.366637	92.8697
CholesterolHDL	33.682563	79.028477	69.772292	68.457491	56.874305
CholesterolTriglycerides	162.189143	294.630909	83.638324	277.577358	291.19878
MMSE	21.463532	20.613267	7.356249	13.991127	13.517609
FunctionalAssessment	6.518877	7.118696	5.895077	8.965106	6.045039
MemoryComplaints	0	0	0	0	0
BehavioralProblems	0	0	0	1	0
ADL	1.725883	2.592424	7.119548	6.481226	0.014691
Confusion	0	0	0	0	0
Disorientation	0	0	1	0	0
PersonalityChanges	0	0	0	0	1
DifficultyCompletingTasks	1	0	1	0	1
Forgetfulness	0	1	0	0	0
Diagnosis	0	0	0	0	0
DoctorInCharge	XXXConfid	XXXConfid	XXXConfid	XXXConfid	XXXConfid

Figure 2. The First Five Rows of the Dataset

Figure 2 shows The display of the first five rows of the dataset using df.head() shows that each column contains values consistent with the feature descriptions.

During the data preprocessing stage, irrelevant columns were removed, and feature selection was performed based on correlation. The PatientID and DoctorInCharge columns were dropped from the dataset using the df.drop(columns=['PatientID', 'DoctorInCharge']) method. These columns do not have a direct relationship with the target variable (Diagnosis) and serve only as administrative information. Feature selection was conducted by calculating the correlation matrix using df.corr(). Features with an absolute correlation value greater than 0.03 or less than -0.03 with the target variable (Diagnosis) were retained.

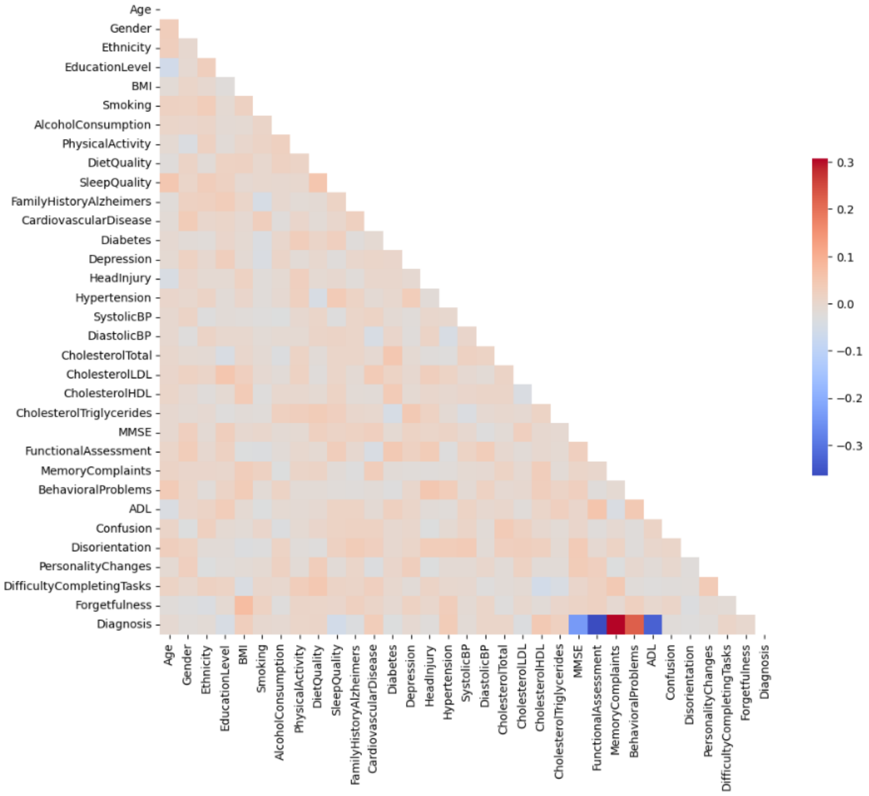


Figure 3. Heatmap of Feature Correlation Matrix

The selected features include EducationLevel, SleepQuality, Family History Alzheimers, Cardiovascular Disease, Diabetes, Hypertension, Cholesterol LDL, Cholesterol HDL, MMSE, Functional Assessment, Memory Complaints, Behavioral Problems, and ADL. This feature selection ensures that only relevant attributes are used in model training.

In the data splitting stage, the dataset is divided into training data (80%) and testing data (20%). The split is performed while maintaining the target class proportion to ensure that the distribution between training and testing data remains representative. This process results in a total of 1,719 samples for training data and 430 samples for testing data for both variables (X and y).

Table 1. Number of Training and Testing Data Samples

Data	**Number of Samples**
Data *training* (X_train)	1719
Data *testing* (X_test)	430
Data *training* (y_train)	1719
Data *testing* (y_test)	430

In SMOTE, data imbalance is addressed using the Synthetic Minority Oversampling Technique (SMOTE). Data imbalance occurs when the number of samples in one target class is significantly smaller than in the other class.

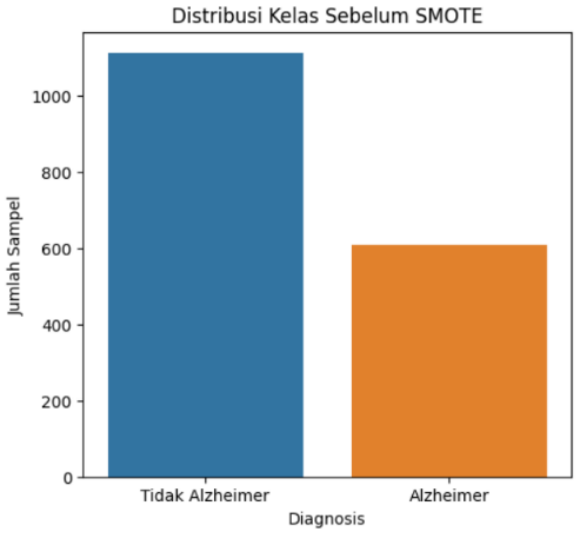


Figure 4. Class Distribution Before SMOTE

Based on the class distribution analysis before applying SMOTE, it can be seen that the 'Tidak Alzheimer' class has a significantly larger number of samples compared to the 'Alzheimer' class. This imbalance may cause the machine learning model to be more inclined to predict the majority class, thereby reducing its ability to accurately recognize the minority class.

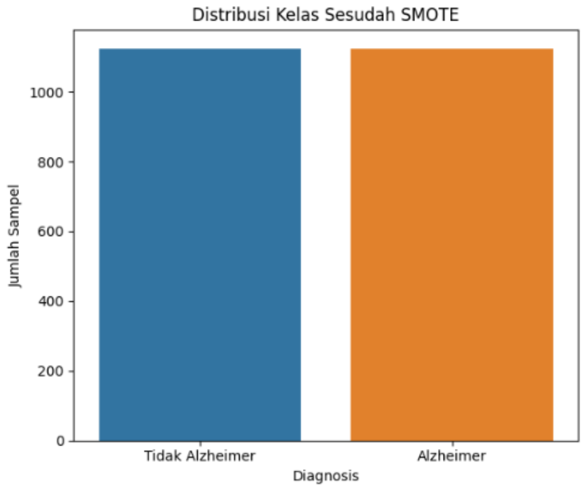
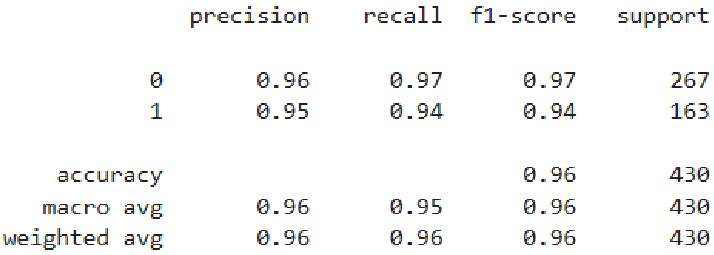


Figure 5. Class Distribution After SMOTE

After applying SMOTE, the class distribution becomes balanced. SMOTE works by generating synthetic data for the minority class by averaging several nearest neighbors to create new samples. This process increases the number of samples in the 'Alzheimer' class to match those in the 'Tidak Alzheimer' class. With a more balanced data distribution, the model is expected to learn both classes more effectively, thereby improving prediction performance, particularly for the minority class.

In the next stage, a Random Forest model was built and evaluated without hyperparameter optimization. The model was trained using the preprocessed training data and tested on the testing data to measure its prediction performance.

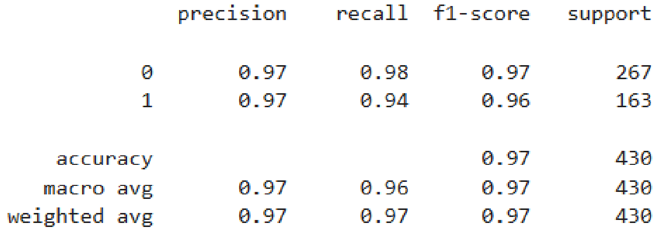


```
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       267
           1       0.95      0.94      0.94       163

    accuracy                           0.96       430
   macro avg       0.96      0.95      0.96       430
weighted avg       0.96      0.96      0.96       430
```

Figure 6 Evaluation Results of the Random Forest Model Without Optimization

Based on the evaluation results shown in Figure 1.6, the model demonstrates good performance in detecting Alzheimer's disease. The evaluation results indicate that the model achieves an accuracy of 96%, with precision, recall, and F1-score values of 95%, 94%, and 94%, respectively, for the 'Alzheimer' class. For the 'Tidak Alzheimer' class, the model has a precision of 96%, recall of 97%, and F1-score of 97%. The high values of these metrics indicate that the model is capable of handling both classes effectively.

After evaluating the Random Forest model without optimization, the next step is to perform hyperparameter optimization using the Random Search method. This optimization aims to find the best hyperparameter combination to enhance the model's performance in detecting Alzheimer's disease.



```
              precision    recall  f1-score   support

           0       0.97      0.98      0.97       267
           1       0.97      0.94      0.96       163

    accuracy                           0.97       430
   macro avg       0.97      0.96      0.97       430
weighted avg       0.97      0.97      0.97       430
```

Figure 7. Evaluation Results of the Random Forest Model After Optimization

The evaluation metrics in Figure 1.7 indicate an improvement in performance compared to the previous model. The model achieves an accuracy of 97%, with precision, recall, and F1-score of 97%, 94%, and 96%, respectively, for the 'Alzheimer' class. Meanwhile, for the 'Tidak Alzheimer' class, the precision and recall values are 97% and 98%, respectively, with an F1-score of 97%. The average metric values also show consistent results with high accuracy across all classes.

During the evaluation stage, an analysis and comparison of the Random Forest model's performance before and after hyperparameter optimization using Random Search were conducted. The evaluation includes metrics such as accuracy, precision, recall, and F1-score to assess the model's effectiveness in detecting Alzheimer's disease. The evaluation results are summarized in Table 2 below.

Table 2 Comparison of Model Evaluation Results Before and After Optimization

Metric	Before Optimization	After Optimzation	Increase
Accuracy	96%	97%	1%
Precision	95%	97%	2%
Recall	94%	94%	0%
F1-score	94%	96%	2%

From Table 2 above, it can be seen that hyperparameter optimization using Random Search improves the model's performance. The accuracy increased from 96% to 97%, precision improved to 97%, recall remained at 94%, and the F1-score increased to 96%. This evaluation demonstrates that hyperparameter optimization with Random Search successfully enhances the overall model performance, particularly in reducing prediction errors for both classes. Thus, the optimized model can be used to detect Alzheimer's disease with higher accuracy and reliability.

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the results of this study, it can be concluded that hyperparameter optimization using the Random Search method successfully improves the performance of the Random Forest algorithm in detecting Alzheimer's disease. The initial Random Forest model, without optimization, demonstrated relatively good performance with an accuracy of 96%, along with high precision, recall, and F1-score for both classes, particularly after applying the SMOTE method to address data imbalance.

After performing hyperparameter optimization using Random Search, the model's performance improved, with accuracy increasing to 97% and enhancements in precision and F1-score, specifically for the 'Alzheimer' class, reaching 97% and 96%, respectively.

This improvement indicates that selecting optimal hyperparameters enhances the model's ability to recognize data patterns more effectively, especially in detecting the minority class (Alzheimer).

Thus, this study confirms that hyperparameter optimization using Random Search significantly enhances the performance of Random Forest in the early detection of Alzheimer's disease. The findings of this study are expected to serve as a reference for developing machine learning-based diagnostic methods to assist medical professionals in detecting Alzheimer's more quickly and accurately.

Suggestion

Based on the results of this study, several recommendations can be considered for further development and future improvements. In addition to Random Forest, other algorithms such as Gradient Boosting, XGBoost, or Deep Learning can be explored to compare their performance in detecting Alzheimer's disease. This comparison may provide additional insights into which algorithm is more effective for this type of data. Furthermore, hyperparameter optimization in this study was conducted using Random Search. Future research could explore alternative approaches such as Grid Search or Bayesian Optimization to comprehensively and efficiently identify the best hyperparameter combinations. Additionally, future studies could incorporate model interpretability analysis using techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) to better understand which features have the most significant impact on the model's predictions.

REFERENCE

World Health Organization. (2023, March 15). Dementia. Retrieved from https://www.who.int/news-room/fact-sheets/detail/dementia

Dana, A. R., Kristananda, R. V., Wibowo, M. B. S., & Prasetya, D. A. (2024, September). Perbandingan Algoritma Decision Tree dan Random Forest dengan Hyperparameter Tuning dalam Mendeteksi Penyakit Stroke. In *Prosiding Seminar Nasional Informatika Bela Negara* (Vol. 4, pp. 66-75).

Klyucherev, T. O., Olszewski, P., Shalimova, A. A., Chubarev, V. N., Tarasov, V. V., Attwood, M. M., ... & Schiöth, H. B. (2022). Advances in the development of new biomarkers for Alzheimer's disease. *Translational neurodegeneration*, *11*(1), 25.

van Oostveen, W. M., & de Lange, E. C. (2021). Imaging techniques in Alzheimer's disease: a review of applications in early diagnosis and longitudinal monitoring. *International journal of molecular sciences*, *22*(4), 2110.

Song, M., Jung, H., Lee, S., Kim, D., & Ahn, M. (2021). Diagnostic classification and biomarker identification of Alzheimer's disease with random forest algorithm. *Brain sciences*, *11*(4), 453.

Ismail, W. N., PP, F. R., & Ali, M. A. (2023). A meta-heuristic multi-objective optimization method for Alzheimer's disease detection based on multi-modal data. *Mathematics*, *11*(4), 957.

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021, November). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. In *Informatics* (Vol. 8, No. 4, p. 79). MDPI.

Widyantoro, W., & Atlantika, A. P. (2021). Hubungan antara demensia dengan activity of daily living (ADL) pada lanjut usia. *Indonesian Journal for Health Sciences*, *5*(2), 77-85.

Pramudhyta, N. A., & Rohman, M. S. (2024). Perbandingan Optimasi Metode Grid Search dan Random Search dalam Algoritma XGBoost untuk Klasifikasi Stunting. *J. MEDIA Inform. BUDIDARMA*, *8*(1), 19.

Beskopylny, A. N., Stel'makh, S. A., Shcherban', E. M., Mailyan, L. R., Meskhi, B., Razveeva, I., & Beskopylny, N. (2022). Concrete strength prediction using machine learning methods CatBoost, k-nearest neighbors, support vector regression. *Applied*

Sciences, *12*(21), 10864.

Aprilliandhika, W., & Abdulloh, F. F. (2024). Comparison Of K-Nearest Neighbor And Support Vector Machine Algorithm Optimization With Grid Search CV On Stroke Prediction. *Jurnal Teknik Informatika (Jutif)*, *5*(4), 991-1000.

Kurniawan, I., Hananto, A. L., Hilabi, S. S., Hananto, A., Priyatna, B., & Rahman, A. Y. (2023). Perbandingan Algoritma Naive Bayes Dan SVM Dalam Sentimen Analisis Marketplace Pada Twitter. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *10*(1), 731-740.