LONTAR KOMPUTER VOL. 16, NO. 1 APRIL 2025 **p-ISSN 2088-1541**
DOI : 10.24843/LKJITI.2025.v16.i01.p01 **e-ISSN 2541-5832**
Accredited Sinta 2 by RISTEKDIKTI Decree No. 158/E/KPT/2021

Implementation of Random Forest Method with Information Gain Selection and Hyperparameter Tuning for Alzheimer's Disease Classification

Riska Nuril Fadhila[a1], Nurissaidah Ulinnuha[a2], Dian Yuliati[b3]

[a]Department of Mathematics, Faculty of Sains and Technology, UIN Sunan Ampel Surabaya
Jl. Ahmad Yani 117, Surabaya, Indonesia 50275
[1]riskanurilfadhila@gmail.com
[2]nuris.ulinnuha@uinsa.ac.id (Corresponding author)
[3]dian.yuliati@uinsa.ac.id

Abstract

Alzheimer's disease is one of the leading causes of decreased quality of life in the elderly aged 65 years and above. One of the problems facing Alzheimer's cases is the difficulty of making an early diagnosis to prevent disease progression, as early symptoms are often mistaken for senile dementia. Using the Random Forest method with information gain feature selection and hyperparameter tuning optimization, this study aims to determine the results of optimization with feature selection and hyperparameter tuning using Random Search and Grid Search to classify Alzheimer's medical record data consisting of 32 variables, including lifestyle factors, clinical measurements, cognitive and functional assessments, as well as symptoms that indicate Alzheimer's. The results showed that applying Information Gain and parameter optimization with the Grid Search method achieved the highest accuracy among all tested experiments. Random Forest with Information Gain and Grid Search gave an accuracy of 95.57%, sensitivity of 92.93%, and specificity of 96.99%, which showed better performance than the Random Search method. This indicates that parameter optimization has a vital role in improving model performance. This research contributes to assisting paramedics in determining whether a patient has Alzheimer's disease based on the characteristics derived from the data.

Keywords: *Alzheimer's, Hyperparameter Tuning, Information Gain, Random Forest*

1. Introduction

Alzheimer's is a disease that has a significant impact on the quality of life of older adults, especially those aged 65 years and above[1]. Alzheimer's cases account for 60% to 70% of all global dementia cases[2]. Over time, Alzheimer's causes a progressive decline in the memory ability and cognitive function of the sufferer, resulting in an increasingly poor quality of life[3]. Based on a report from the Ministry of Health's Online Hospital Information System, the number of new Alzheimer's cases from 2019 to 2023 was recorded at 83.5 thousand for outpatient care and 2.4 thousand for inpatient care[4].

The increasing trend of Alzheimer's cases in Indonesia is expected to continue, even projected to reach 4 million cases by 2050 [5]. Until now, no drugs or therapies have been found that can cure Alzheimer's disease altogether. Available treatments can only relieve the symptoms experienced by sufferers[6]. Symptoms of Alzheimer's disease are often interpreted as ordinary senile dementia, causing the disease to be poorly detected[7]. Therefore, early detection and diagnosis play a significant role in managing Alzheimer's [8].

Random Forest is one of the effective methods for data classification that uses ensemble learning principles and combines multiple decision trees to improve accuracy and minimize the possibility of overfitting [9]. This method is also resistant to outliers and effectively handles imbalanced data through bagging techniques and tuning optimization [10][11]. However, this method has several disadvantages, including requiring high computational resources on datasets with a large number of trees [12]. In addition, parameters set with fixed values can result in non-optimal performance

measurements, so optimization with Hyperparameter Tuning is needed to find the optimal value of a model's parameters [13].

Some studies show that random forests can accurately predict diseases, so this study will apply the random forests method to perform classification. Research [14] compared Random Forest, logistic regression, K-Nearest Neighbors (KNN), and Stochastic Gradient Descent (SGD) methods to predict liver disease. The results of this study showed that the Random Forest method obtained the highest accuracy of several other methods, which was 90%. In contrast, the accuracy of the logistic regression, KNN, and SGD methods was 68%, 82%, and 48%. Depari et al. [15] compared Decision Tree, Naive Bayes, and Random Forest methods with normation techniques for heart disease classification. The results of this study show that the classification performance evaluation of the Random Forest method is superior, with an accuracy of 0.75, compared to the accuracy of the Decision Tree and Naive Bayes methods of 0.71 and 0.72. Furthermore, previous research by Samad et al., [16] compared methods, namely Naive Bayes, K-Nearest Neighbors, Decision Trees, and Ensemble methods by optimizing using Spearman feature selection to predict Alzheimer's disease. The results of this study show that the highest accuracy of some of these methods is the ensemble method, which is 94.07% using 13 features, while the accuracy of the Naive Bayes, KNN, and Decision Tree methods is 89.47%, 86.84%, and 93.09%.

However, Random Forest also has some drawbacks. One of them is the need for relatively high computational resources, especially on datasets with a large number of trees, which can cause long processing times and high memory consumption[12]. To overcome these problems, feature selection methods can be used to select relevant features and eliminate unimportant or redundant features, thereby reducing errors in detection[17]. Feature selection is performed before the classification stage to select features of similar relevance that can improve the efficiency and speed of the classification algorithm, which in turn has the potential to improve the accuracy of the model[18].

Research conducted by Devia [19], compared three feature selection methods in the Random Forest algorithm, namely Information Gain, Recursive Feature Elimination with Cross-Validation (RFECV), and a combination of Mutual Information with Recursive Feature Elimination (MI-RFE) as a Hybrid method on the classification of the CIC-IDS-2018 dataset. The results of this study indicate that the feature selection method with Information Gain works well in determining essential features in the Random Forest algorithm with an accuracy of 99%. Furthermore, research conducted by Hasan [20], who applied the information gain method for feature selection to classify student study duration using the random forest algorithm. The results of this study indicate that the combination of Information Gain and Random Forest produces higher accuracy, which is 100%

Hyperparameter tuning is carried out to optimize parameters and improve the accuracy of a model, which can find the optimal value of model parameters. Hyperparameter tuning with Grid Search and Random Search techniques can automatically find optimal parameters in Random Forest [21]. The successful use of Hyperparameter Tuning is shown by research [22] in detecting emotions using the Random Forest method with Hyperparameter Tuning (Random Search) optimization. The results of this study showed an increase in accuracy from 0.85 to 0.86 for two classes and 0.73 to 0.76 for three classes. The Random Forest method with Hyperparameter Tuning (Grid Search) optimization was used to detect malware[23]. The results of this study show an increase in accuracy from 99.04% to 99.23%. The results show that Information Gain can be used to optimize Random Forest parameters.

The novelty of this research lies in random forest (RF) hyperparameter optimization for Alzheimer's disease classification based on medical record data, with feature selection using information gain (IG). The optimization results are then compared between Grid Search (GS) and Random Search (RS) to determine the most effective method. By applying Hyperparameter Tuning, this research is expected to improve the accuracy and efficiency of the RF model in detecting Alzheimer's disease.

2. Research Method

The method used in this research consists of several stages. The initial stage of the data preprocessing process is to normalize the data with different scales. Next, the division of testing

and training data is carried out. After that, the Random Forest classification process with hyperparameter tuning is continued with the evaluation of the results, which can be seen in Figure 1.

The first stage involves inputting medical record data from Alzheimer's patients obtained from Kaggle. The preprocessing stage normalizes numerical features using Min-Max Scaling to ensure a uniform data distribution. After preprocessing, the data is split into training and testing sets using the k-fold cross-validation method to allow for a more accurate and unbiased model evaluation. Next, feature selection is conducted using information gain. This step calculates the entropy for each feature to assess its importance. Features with information gain values exceeding a predetermined threshold are chosen for the classification process.

The next stage involves classification using the random forest algorithm. This technique creates multiple decision trees using bagging methods. Each tree is trained with randomly selected data samples, utilizing random sampling with replacement. Entropy calculations help determine the root nodes and the criteria for splitting. The predictions generated by all trees are then combined using the majority voting method, where the final decision is based on the most votes from all trees in the model. Once the model is trained, it is tested using the testing data to evaluate its performance. This evaluation uses a confusion matrix, which provides a detailed breakdown of the model's predictive accuracy.

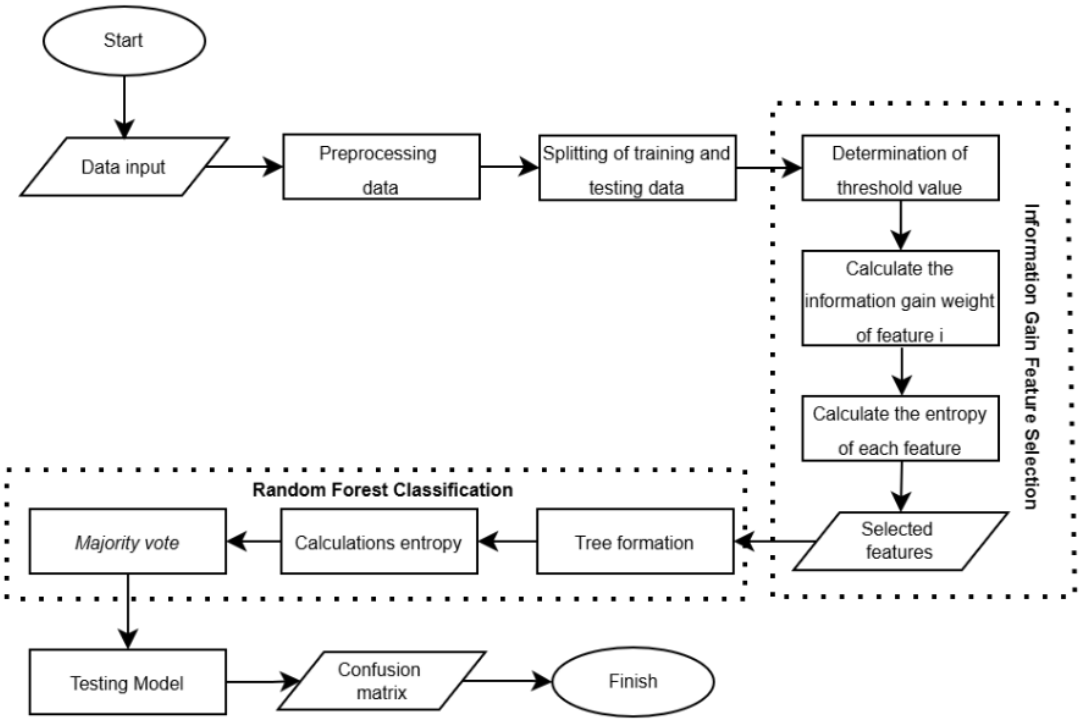

Figure 1. Flowchart of Random Forest

The data used in this study are secondary data of Alzheimer's patients obtained online from the Kaggle Dataset website published by Rabie El Kharoua (2024) under the CC BY 4.0 license [24]. This Alzheimer's dataset contains complete health information on 2,149 patients, consisting of 32 Independent and one dependent variable. Health information includes lifestyle factors, demographic details, medical history, cognitive and functional assessments, clinical measurements, symptoms, and diagnosis of Alzheimer's disease. 17 variables have categorical values, and 15 have numerical values, as shown in Table 1.

Table 1. Description of Variables

Variable	Type	Range Value
Gender	Categorical	0 = Male, 1 = Female
Ethnicity	Categorical	0 = Caucasian, 1 = African American, 2 = Asian, 3 = Other
Education Level	Categorical	0 = None, 1 = High school, 2 = Bachelor's degree, 3 = > Bachelor's degree
Smoking, Family History Alzheimer, CVD, Diabetes, Depression, Head Injury, Hypertension, Memory Complaints, Behavioral Problems, Confusion, Disorientation, Personality Changes, Difficulty Completing Tasks, Forgetfulness	Categorical	0 = No, 1 = Yes
Age	Numeric	Min: 60, Max: 90, Mean: 74.90
BMI	Numeric	Min: 15, Max: 39.99, Mean: 27.65
Alcohol Consumption	Numeric	Min: 0, Max: 19.98, Mean: 10.03
Physical Activity	Numeric	Min: 0, Max: 9.98, Mean: 4.92
Diet Quality	Numeric	Min: 0, Max: 9.99, Mean: 4.99
Sleep Quality	Numeric	Min: 4, Max: 9.99, Mean: 7.05
Systolic BP	Numeric	Min: 90, Max: 179, Mean: 134.26
Diastolic BP	Numeric	Min: 60, Max: 119, Mean: 89.84
Cholesterol Total	Numeric	Min: 150.09, Max: 299.99, Mean: 225.19
Cholesterol LDL	Numeric	Min: 50.23, Max: 199.96, Mean: 124.33
Cholesterol HDL	Numeric	Min: 20, Max: 99.98, Mean: 59.46
Cholesterol Triglycerides	Numeric	Min: 50.40, Max: 399.94, Mean: 228.28
MMSE	Numeric	Min: 0, Max: 29.99, Mean: 14.75
Functional Assessment	Numeric	Min: 0, Max: 0.99, Mean: 5.08
ADL	Numeric	Min: 0, Max: 9.99, Mean: 4.98

This dataset has a class imbalance, with 1,389 patients (64.7%) in the normal class and 760 patients (35.3%) diagnosed with Alzheimer's disease. Despite the imbalance, Random Forest can handle imbalanced data through bagging techniques and hyperparameter tuning optimization[11].

2.1. Random Forest

Classification in a Random Forest combines several decision trees trained using available data samples [9]. Random Forest is an advancement of the Classification and Regression Tree (CART) method that uses bootstrap aggregating (bagging) techniques and the random selection of attributes at each node. CART is one of the decision tree methods for analyzing response variables, both Numeric and categorical [25]. Random Forest is an ensemble learning algorithm that constructs multiple decision trees sequentially and merges their results to provide more

accurate and reliable predictions [26]. Mathematically, the calculation of Random Forest in forming a tree is expressed as follows:

$$Entropy(S) = \sum_{i=1}^{n} = p_i . log_2\, p_i \quad (1)$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} . Entropy(S_i) \quad (2)$$

Where S is the number of cases with the number of partitions (n) and p_i represents the proportion of cases belonging to category i out of the total number of cases S. Next, the Gain value is calculated according to variable (A), where Si is the number of cases in the-i partition. The Gain calculation affects each node and internal node. If the Gain calculation reaches a value of 0, then the calculation process stops. However, if the result is not equal to 0, the calculation will continue to the next split node[27]. In classification, the result is used for majority vote, where the final prediction is the class most frequently selected by the decision tree[28]. Random Forest has a limitation, which is the low interpretability of the model. Random Forest consists of many decision trees working simultaneously, making it difficult to trace how the model arrived at a final decision.

2.2. Information Gain

Information gain is one of the methods in feature selection used to rank features by calculating the entropy of a class before and after observing features in the same data[17]. Feature determination using Information gain is done in three stages, namely: [20]

1. Determining the required threshold allows attributes with values equal to or higher than the threshold to be retained, while attributes below the threshold will be removed.
2. Calculate the gain value for each attribute in the dataset with Equation (1) and (2).
3. Removing irrelevant attributes.

2.3. Hyperparameter Tuning

Hyperparameter is the assignment of values to parameters before the learning process begins[29]. Hyperparameter tuning is used to find the most optimal combination of parameter values to enhance model performance, accuracy, and generalization ability[13]. Alternatives used in the Hyperparameter tuning process can be done by Random Search, which randomly samples from various hyperparameter combinations and evaluates their performance through cross-validation to find the optimal configuration[22]. Another alternative is GridSearchCV, which tries one parameter combination simultaneously and validates each combination. The difference between Random Search and Grid Search is that Grid Search tests all possible parameter combinations. In contrast, Random Search takes some random samples of the available parameter values and then combines them. This method focuses more on exploring parameter values that have a significant effect on model performance[30].

The hyperparameter tuning value combinations of the default value (n_estimator=100) used in the following are shown in Table 2.

Table 2. Combination of Hyperparameter Tuning values for Random Forest

Hyperparameter	Range of Search	Default Value
n_estimator	[50, 100, 200]	[100]
max_depth	[5, 10, 15, 20, 25]	
min_samples_split	[2, 4, 6, 8, 10]	
min_samples_leaf	[1, 2, 4]	

Table 2 shows the parameters used to build and run the model. Parameter values will be randomly selected to get the best results.

Table 3. Random Forest Hyperparameter Function

Parameters	Function

n_estimator	Determine the number of models used in Random Forest[31].
max_depth	Maximum depth of each decision tree that can help curtail overfitting and improve the model's ability to generalize to data[32].
min_samples_split	Determines the number of samples needed to split the internal nodes in the decision tree, which can help control model complexity and prevent overfitting[32].
min_samples_leaf	Determines the number of samples required to form a leaf on the decision tree that prevents overfitting[31].

Table 3 shows the parameters used to specify the model built. The parameters used to test the Random Forest model are n_estimator, max_depth, min_samples_split, and min_samples_leaf.

In addition, finding the value of the explanatory variables (Mtry) aims to determine the number of variables to be randomly selected at each split. This process provides diversity to the decision trees in the forest, thus helping to reduce overfitting. The following is the calculation formula according to Breiman[33]:

$$Mtry1 = \frac{1}{2}\left|\sqrt{p}\right| \quad (3)$$

$$Mtry2 = \left|\sqrt{p}\right| \quad (4)$$

$$Mtry3 = 2 \times \left|\sqrt{p}\right| \quad (5)$$

Where p is the total number of variables.

Although hyperparameter tuning through techniques such as Grid Search and Random Search can enhance the model's accuracy, it also carries the risk of overfitting. This is particularly true if the number of trees is excessively high or if the model is overly complex. Overfitting may lead to excellent performance on training data but poor predictive capabilities on new data. To mitigate this issue, cross-validation is employed during the tuning process. This technique helps ensure the model performs well on the training data and generalizes effectively to the test data. By incorporating cross-validation, the risk of overfitting can be minimized, even in hyperparameter tuning.

2.4. K-Fold Cross Validation (CV)

K-Fold CV is a method to assess the performance of an algorithm model by dividing the data into subsets. This technique randomly separates the data into k-folds, where one group is used as testing data and the other as training data[27]. Using k-fold aims to test 'k' times, resulting in a more accurate assessment of model performance and reducing the risk of bias from missed data.

2.5. Evaluation

Confusion matrix is a method where the efficiency or success rate of the classification process can be measured[34]. The confusion matrix represents the feasibility level of the model against the classification process performed[35]. The confusion matrix generates key metrics such as accuracy, precision, specificity, and sensitivity[36]. The following is an illustration of the confusion matrix shown in Table 4.

Table 4. Confusion Matrix

		Prediction Class	
		Negative	Positive
Actual Class	Negative	TN	FP
	Positive	FN	TP

where:

a) True Positive (TP) is the actual data on Alzheimer's class, which is classified as Alzheimer's class, where the patient is correctly classified as having Alzheimer's disease.
b) True Negative (TN) is the actual data on the normal class classified as the normal class, where TN is the number of normal patients (without Alzheimer's) who are correctly classified.
c) False Positive (FP) is the actual data in the normal class that is classified as the Alzheimer's class, where normal patients (without Alzheimer's) are classified as Alzheimer's.
d) False Negative (FN) is the actual data of the Alzheimer's class classified as a normal class, where Alzheimer's patients are classified as normal patients (without Alzheimer's).

From the Confusion matrix, the classification results are evaluated with the following calculation formula [37].

a) Accuracy describes the system's success rate in classifying patients with Alzheimer's.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

b) Sensitivity, describes the suitability of a classification system in detecting patients with Alzheimer's.

$$Sensitivity = \frac{TP}{TP+FN} \quad (7)$$

c) Specificity, describes the suitability of the value in the classification of a system in detecting normal patients (without Alzheimer's) who are not diagnosed with Alzheimer's.

$$Specificity = \frac{TN}{TN+FP} \quad (8)$$

Confusion matrix and result evaluation provide a clear picture of the strengths and weaknesses of the model in classifying Alzheimer's medical record data by mapping the positive (Alzheimer's patients) and negative (non-Alzheimer's) classes. Through the accuracy, sensitivity, and specificity values, we can assess how effective the model is in detecting truly positive or negative patients and identify misclassifications, such as false positives and false negatives.

3. Result and Discussion

3.1. Result

Examples of data used in this study are shown in Table 5. The numerical features were normalized using Min-Max Scaling to ensure a more uniform data distribution, as shown in Table 6. In addition, the data was divided into training and testing sets using the k-fold cross-validation method (k=10) to improve the accuracy of model evaluation.

Table 5. Sample Research Data

No.	Age	Gender	Ethnicity	Education	BMI	· · ·	Forgetfulness	Class
1	73	0	0	2	22.9	· · ·	0	0
2	89	0	0	0	26.8	· · ·	1	0
3	73	0	3	1	17.7	· · ·	0	0
4	74	1	0	1	33.8	· · ·	0	0
5	89	0	0	0	20.7	· · ·	0	0
6	86	1	1	1	30.6	· · ·	0	0
7	68	0	3	2	38.3	· · ·	1	0
...	...	...	...	...	...	...	...	...
2148	78	1	3	1	15.2	· · ·	1	1
2149	72	0	0	2	33.2	· · ·	1	0

Table 6. Data After Normalization

No.	Age	Gender	Ethnicity	Education	BMI	· · ·	Forgetfulness	Class
1	0.43	0	0	2	0.31	· · ·	0	0
2	0.96	0	0	0	0.47	· · ·	1	0
3	0.43	0	3	1	0.11	· · ·	0	0
4	0.46	1	0	1	0.75	· · ·	0	0
5	0.96	0	0	0	0.22	· · ·	0	0
6	0.86	1	1	1	0.62	· · ·	0	0
7	0.26	0	3	2	0.93	· · ·	1	0
...	...	...	...	...	...	...	...	...
2148	0.60	1	3	1	0.01	· · ·	1	1
2149	0.40	0	0	2	0.73	· · ·	1	0

This research conducted four experimental models: Random Forest with default parameters, feature selection with Information Gain, Grid Search, and Random Search. Performance measures were calculated and compared to get the best results.

In Information Gain, research was conducted to see the best parameters of variables that significantly impact the model's performance. The results of the experiment are shown in Figure 2.

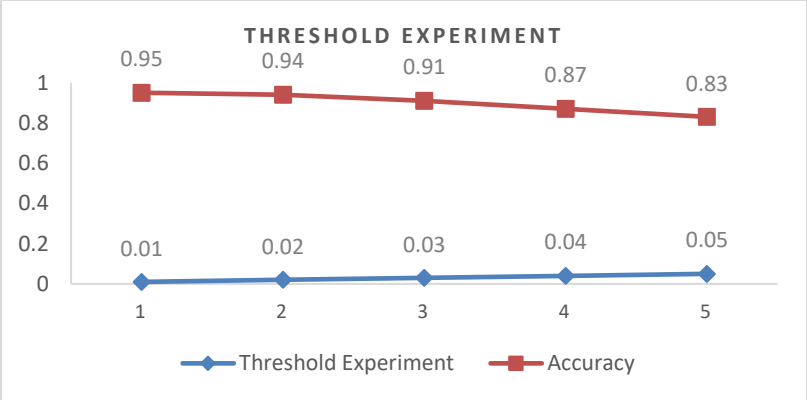

Figure 2. Threshold Experiment Value

From Figure 2, there is an orange line that states the accuracy value of each trial and that the best threshold value is at a value of 0.01, with the details of the selected variables as follows.

Table 7. Selected parameters from the threshold value experiment

Threshold Value	Selected Variables
0.01	Functional Assessment, MMSE, ADL, Memory Complaints, Forgetfulness, Ethnicity, Alcohol Consumption, Physical Activity, Behavioral Problems, Cholesterol HDL

Threshold Value	Selected Variables
0.02	Functional Assessment, MMSE, ADL, Memory Complaints, Behavioral Problems
0.03	Functional Assessment, MMSE, ADL, Behavioral Problems, Age
0.04	Functional Assessment, MMSE, ADL, Memory Complaints
0.05	Functional Assessment, MMSE, ADL

A lower threshold value of 0.01 enables the model to incorporate more relevant features into its predictions. This enhances the model's sensitivity by capturing additional information that can contribute to improved accuracy. In contrast, a higher threshold value of 0.05 may limit the number of features considered, resulting in a more selective model. This selectivity could lead to a decrease in accuracy, as some essential features, despite their small contributions, may be overlooked even though they are related to the prediction. Therefore, employing a threshold of 0.01 in this experiment yields the best accuracy, as it allows the model to account for more relevant features.

Based on Table 7, the 10 selected variables will enter the hyperparameter tuning testing stage with a threshold value of 0.01. Furthermore, in the optimization experiment using hyperparameter tuning with a literature review of similar research, the results are presented below in Table 8. The results in Table 8 indicate that Grid Search (GS) found a more optimal combination of hyperparameters for Random Forest (RF) compared to Random Search (RS), with a higher number of estimators. Combined with using Information Gain (IG) for feature selection, this improvement enhanced model performance by focusing on the most relevant features and optimizing parameter selection.

Table 8. Output parameters from the experiment

Experiment	*n_estimator*	*min_samples_split*	*min_samples_leaf*	*max_depth*	*max_features*
IG + RF + RS	100	10	4	15	6
IG + RF + GS	200	10	4	15	6

All four experiments used the same tests and parameters. Table 5 shows the parameters used for each classification with hyperparameter tuning to get optimal results. The Random Search processing time is 31m 33s with 30 iterations, while the Grid Search processing time is 34h 39m. Furthermore, the results of the three experiments are shown in Table 9.

Table 9. Experiment results

Experiment	Accuracy (%)	Sensitivity (%)	Specificity (%)
RF	93.90	87.52	97.42
IG + RF	95.11	91.38	97.12
IG + RF + RS	95.53	92.63	96.96
IG + RF + GS	**95.57**	**92.93**	**96.99**

Table 9 shows increased accuracy in random forest method optimization using information gain feature selection and hyperparameter optimization, namely in hyperparameter tuning with the Grid Search method. A comparison of the performance evaluation results of these two methods can be seen in Figure 3.

Figure 3 compares the experimental results between regular Random Forest (RF) and Random Forest combined with Information Gain (RF+IG), which shows an increase in accuracy from 93.90% to 95.34%. The experiment's results with hyperparameter tuning optimization obtained an increase in accuracy from the optimal parameters with random search (IG+RF+GS) of 95.57%.

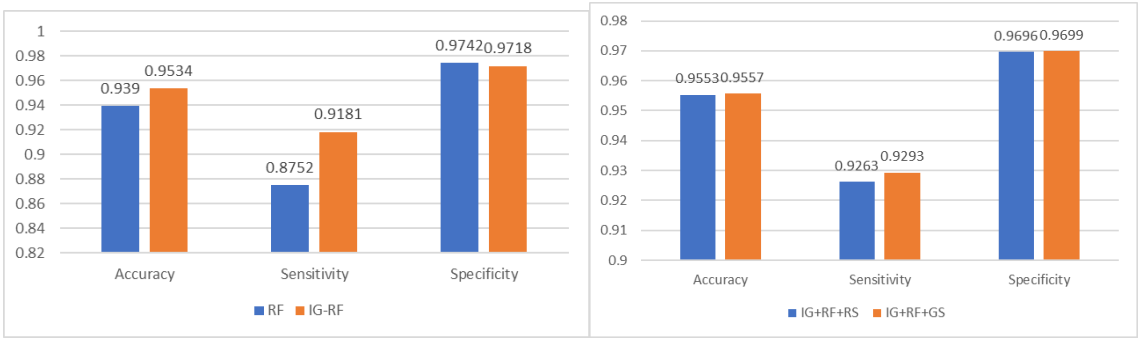

Figure 3. Comparison of Evaluation Result

3.2. Discussion

In the context of Random Forest (RF) optimization experiments utilizing Information Gain (IG) for feature selection, the authors concluded that IG is a viable method for enhancing model performance. Improved accuracy occurs using IG, which helps select the most influential features by assessing their contribution to reducing uncertainty. Focusing on more informative features makes the model more efficient, eliminates irrelevant data dimensions, and reduces the risk of overfitting. As such, the decision tree formed becomes more optimized, improving the model's overall accuracy. During the IG process, threshold experiments were conducted within a value range of 0.01 to 0.05, with the optimal results achieved at a threshold value of 0.01 leading to the selection of ten variables that proceeded to the hyperparameter tuning stage.

The combination of IG, RF, and Grid Search (GS) achieved an optimal accuracy of 95.57%, whereas the combination of IG, RF, and Random Search (RS) reached an accuracy of 95.53%. The GS method is more effective for optimization than RS despite requiring a longer computation time. This is attributed to GS's comprehensive testing of all combinations of predefined parameters, enabling it to identify the optimal configuration consistently. In contrast, RS only explores a subset of the parameter space, which may result in the best solution not being discovered. The model can achieve enhanced and more stable classification performance by integrating feature selection through IG, utilizing the RF algorithm, and optimizing parameters via GS.

This dataset was previously used in a study by Samad[16] to diagnose Alzheimer's disease using several algorithms, with the best results achieved through the Ensemble method with Spearman's algorithm for feature selection. The study achieved the highest accuracy of 94.07%, which was lower than the accuracy of the research model (IG-RF-GS), which reached 95.57%.

The results show that the Random Forest model with hyperparameter optimization can help in the early diagnosis of Alzheimer's disease with high accuracy in classifying patients based on risk factors and symptoms. However, this model still has limitations in handling ambiguous cases, such as patients with mild symptoms or cognitive scores within normal limits. The model should not be used as the sole decision tool but rather as clinical decision support to improve the accuracy of diagnosis and help clinicians consider other factors that affect the patient's condition. In addition, this study has not included real-world testing due to limited access to more extensive clinical data. To increase clinical relevance and reduce potential bias, further research should test the model with data from hospitals or clinics to assess its performance in more diverse and complex patient conditions.

4. Conclusion

This study presents a system that employs Random Forest and Information Gain methods to identify Alzheimer's disease, utilizing hyperparameter tuning optimization to achieve optimal outcomes. The feature selection process applied a threshold value of 0.01, successfully identifying 10 variables that contributed to model formation, which were subsequently used in the identification process. The results of the experiments indicate that the combination of Random Forest (RF) classification with Information Gain (IG) and Grid Search (GS) yielded the best performance. This IG+RF+GS classification achieved an accuracy of 95.57%, with a sensitivity of

92.93% and a specificity of 96.99%. The optimal parameters identified were n_estimator=200, max_depth=15, min_samples_split=10, min_samples_leaf=4, and max_features=6. Therefore, GS optimization significantly enhances performance and is superior in determining the best combination of hyperparameters. However, it necessitates more computational time and resources than Random Search (RS). Future research aims to continue applying Random Forest and Information Gain with hyperparameter tuning optimization to detect Alzheimer's disease, and the findings of this study can serve as a helpful reference. Moreover, the results could also be further explored using alternative decision tree methodologies, such as XGBoost, to yield even better results.

References

[1] B. I. Nabila, W. E. Kurniawan, and M. Maryoto, “Gambaran Tingkat Demensia Pada Lansia Di Rojinhome Ikedaen Okinawa Jepang,” *Jurnal Studi Keperawatan*, vol. 3, no. 2, pp. 1–8, 2022, doi: 10.31983/j-sikep.v3i2.8410.

[2] A. Rosyida and T. B. Sasongko, “Early Detection of Alzheimer’s Disease with the C4.5 Algorithm Based on BPSO (Binary Particle Swarm Optimization),” *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 12, no. 3, pp. 341–349, 2023, doi: 10.32736/sisfokom.v12i3.1716.

[3] Jan Sudir Purba, “Potential Implication of Treatments for Alzheimer’s Disease: Current and Future,” *Medicinus*, vol. 36, no. 1, pp. 3–10, 2023, doi: 10.56951/medicinus.v36i1.112.

[4] D. Arlinta, “Alzheimer Masih Terabaikan, Banyak Kasus Tidak Terdeteksi,” *Kompas.id*, 2024. https://www.kompas.id/baca/humaniora/2024/09/08/alzheimer-masih-terabaikan-banyak-kasus-tidak-terdeteksi (accessed Sep. 29, 2024).

[5] N. Nurbaiti, S. G. Sutoro, E. Supriyaningsih, S. W. Wiyanti, and I. Maesaroh, “Edukasi untuk Deteksi Dini dan Perawatan Lansia dengan Alzheimer di Masa Pandemi Covid-19,” *Jurnal Kreativitas Pengabdian Kepada Masyarakat*, vol. 6, no. 7, pp. 2887–2895, 2023, doi: 10.33024/jkpm.v6i7.10093.

[6] A. G. M. Sianturi, “Stadium, Diagnosis, dan Tatalaksana Penyakit Alzheimer,” *Majalah Kesehatan Indonesia*, vol. 2, no. 2, pp. 39–44, 2021, doi: 10.47679/makein.202132.

[7] N. S. Riasari, D. Djannah, K. Wirastuti, and M. Silviana, “Faktor-Faktor yang Mempengaruhi Penurunan Fungsi Kognitif pada Pasien Prolanis Klinik Pratama Arjuna Semarang,” *Jurnal Pendidikan Tambusai*, vol. 6, pp. 3049–3056, 2022, doi: https://doi.org/10.31004/jptam.v6i1.3345.

[8] A. Arfina, “Pengaruh Edukasi Terhadap Pengetahuan Masyarakat Tentang Deteksi Dini Alzheimer Di Kelurahan Labuh Baru Pekanbaru,” *Health Care: Jurnal Kesehatan*, vol. 10, no. 01, pp. 256–261, 2021, doi: 10.36763/healthcare.v10i2.170.

[9] L. Ratnawati and D. R. Sulistyaningrum, “Penerapan Random Forest untuk Mengukur Tingkat Keparahan Penyakit pada Daun Apel,” *Jurnal Sains dan Seni ITS*, vol. 8, no. 2, 2020, doi: 10.12962/j23373520.v8i2.48517.

[10] R. N. Alifah *et al.*, “Perbandingan Metode Tree Based Classification untuk Masalah Klasifikasi Data Body Mass Index,” *Indonesian Journal of Mathematics and Natural Science*, vol. 47, no. 1, p. 2024, 2024, doi: https://doi.org/10.15294/m2k97436.

[11] M. A. Abubakar, M. Muliadi, A. Farmadi, R. Herteno, and R. Ramadhani, “Random Forest Dengan Random Search Terhadap Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung,” *Jurnal Informatika*, vol. 10, no. 1, pp. 13–18, 2023, doi: 10.31294/inf.v10i1.14531.

[12] Y. A. Saadoon and R. H. Abdulamir, “Improved Random Forest Algorithm Performance for Big Data,” *Journal of Physics Conference Series*, vol. 1897, no. 1, 2021, doi: 10.1088/1742-6596/1897/1/012071.

[13] H. A. and S. A. Ludwig, “Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization,” *IEEE Congress on Evolutionary Computation (CEC)*, pp. 1551–1559, 2021, doi: 10.1109/CEC45853.2021.9504761.

[14] K. W. Kayohana, “Klasifikasi penyakit hati menggunakan random forest dan knn,” vol. 8, no. 4, pp. 7924–7929, 2024, doi: https://doi.org/10.36040/jati.v8i4.10457.

[15] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, “Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung,” *Informatik : Jurnal Ilmu Komputer*, vol. 18, no. 3, p. 239, 2022, doi: 10.52958/iftk.v18i3.4694.

[16] A. Samad and E. Samet Aydı, "Rapid Alzheimer's Disease Diagnosis Using Advanced Artificial Intelligence Algorithms," *International Journal of Innovative Science and Research Technology*, vol. 9, no. 6, pp. 1760–1768, 2024, doi: 10.38124/ijisrt/ijisrt24jun1915.
[17] A. Bijaksana, P. Negara, H. Muhardi, and I. M. Putri, "Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes Dan Seleksi Fitur Information Gain Sentiment Analysis on Airlines Using Naïve Bayes Method and Feature Selection Information Gain," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 3, pp. 599–606, 2020, doi: 10.25126/jtiik.202071947.
[18] M. Frananda Adiezwara Ramadhan, I. Rizal Setiawan, and A. Asriyanik, "Klasifikasi Hoax Dan Fakta Menggunakan Algoritma Shallow Neural Network Pada Berita Politik Pemilihan Presiden Indonesia 2024," *JATI (Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 4, pp. 8006–8013, 2024, doi: 10.36040/jati.v8i4.10621.
[19] A. Devia and B. Soewito, "Analisis Perbandingan Metode Seleksi Fitur untuk Mendeteksi Anomali pada Dataset CIC-IDS-2018," *Jurnal Teknologi Dan Sistem Informasi Bisnis-JTEKSIS*, vol. 5, no. 4, p. 572, 2023, doi: https://doi.org/10.47233/jteksis.v5i4.1069 Abstract.
[20] I. K. Hasan, R. Resmawan, and J. Ibrahim, "Perbandingan K-Nearest Neighbor dan Random Forest dengan Seleksi Fitur Information Gain untuk Klasifikasi Lama Studi Mahasiswa," *Indonesian Journal of Applied Statistics*, vol. 5, no. 1, p. 58, 2022, doi: 10.13057/ijas.v5i1.58056.
[21] P. R. Togatorop, M. Sianturi, D. Simamora, and D. Silaen, "Optimizing Random Forest using Genetic Algorithm for Heart Disease Classification," *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, vol. 13, no. 1, p. 60, 2022, doi: 10.24843/lkjiti.2022.v13.i01.p06.
[22] Z. K. Nur, R. Wijaya, and G. S. Wulandari, "Optimizing Emotion Recognition with Wearable Sensor Data : Unveiling Patterns in Body Movements and Heart Rate through Random Forest Hyperparameter Tuning," *Jurnal Media Informatika Budidarma*, vol. 8, no. 3, pp. 1–12, 2024, doi: https://doi.org/10.30865/mib.v8i3.7761.
[23] I. Muhamad Malik Matin, "Hyperparameter Tuning Menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware," *Multinetics*, vol. 9, no. 1, pp. 43–50, 2023, doi: 10.32722/multinetics.v9i1.5578.
[24] R. El Kharoua, "Alzheimer's Disease Dataset," *Kaggle*, 2024. https://www.kaggle.com/dsv/8668279 (accessed Aug. 01, 2024).
[25] W. Agwil, H. Fransiska, and N. Hidayati, "Analisis Ketepatan Waktu Lulus Mahasiswa Dengan Menggunakan Bagging Cart," *FIBONACCI : Jurnal Pendidikan Matematika dan Matematika*, vol. 6, no. 2, p. 155, 2020, doi: 10.24853/fbc.6.2.155-166.
[26] Y. Yuliani, "Algoritma Random Forest Untuk Prediksi Kelangsungan Hidup Pasien Gagal Jantung Menggunakan Seleksi Fitur Bestfirst," *Infotek : Jurnal Informatika dan Teknologi*, vol. 5, no. 2, pp. 298–306, 2022, doi: 10.29408/jit.v5i2.5896.
[27] R. Tuntun, K. Kusrini, and K. Kusnawi, "Analisis Perbandingan Kinerja Algoritma Klasifikasi dengan Menggunakan Metode K-Fold Cross Validation," *Jurnal Media Informatika Budidarma*, vol. 6, no. 4, p. 2111, 2022, doi: 10.30865/mib.v6i4.4681.
[28] G. A. Pradipta and Putu Desiana Wulaning Ayu, "Kombinasi Inisial Filtering Oversampling dengan Metode Ensemble Classifier pada Klasifikasi Data Imbalanced," *Jurnal Sistem dan Informatika*, vol. 17, no. 2, pp. 137–145, 2023, doi: 10.30864/jsi.v17i2.591.
[29] T. A. E. Putri, T. Widiharih, and R. Santoso, "Penerapan Tuning Hyperparameter Randomsearchcv Pada Adaptive Boosting Untuk Prediksi Kelangsungan Hidup Pasien Gagal Jantung," *Jurnal Gaussian*, vol. 11, no. 3, pp. 397–406, 2023, doi: 10.14710/j.gauss.11.3.397-406.
[30] M. Fajri and A. Primajaya, "Komparasi Teknik Hyperparameter Optimization pada SVM untuk Permasalahan Klasifikasi dengan Menggunakan Grid Search dan Random Search," *Journal of Applied Informatics Computing*, vol. 7, no. 1, pp. 14–19, 2023, doi: 10.30871/jaic.v7i1.5004.
[31] Hajiar Yuliana, "Hyperparameter Optimization of Random Forest for 5G Coverage Prediction," *Buletin Pos dan Telekomunikasi*, vol. 22, no. 1, pp. 75–90, 2024, doi: 10.17933/bpostel.v22i1.390.
[32] I. Afdhal, R. Kurniawan, I. Iskandar, R. Salambue, E. Budianita, and F. Syafria, "Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia," *Jurnal Nasional Komputasi dan Teknologi Informasi*, vol. 5, no. 1, pp. 122–130, 2022, doi: https://doi.org/10.32672/jnkti.v5i1.4004.

[33] C. J. S. Leo Breiman, Jerome Friedman, R.A. Olshen, *Classification and Regression Trees*. 2022. doi: https://doi.org/10.1201/9781315139470.
[34] I. Wardhana, Musi Ariawijaya, Vandri Ahmad Isnaini, and Rahmi Putri Wirman, "Gradient Boosting Machine, Random Forest dan Light GBM untuk Klasifikasi Kacang Kering," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 1, pp. 92–99, 2022, doi: 10.29207/resti.v6i1.3682.
[35] S. Lestari, A. Akmaludin, and M. Badrul, "Implementasi Klasifikasi Naive Bayes Untuk Prediksi Kelayakan Pemberian Pinjaman Pada Koperasi Anugerah Bintang Cemerlang," *PROSISKO Jurnal Pengembangan Riset dan Observasi Sistem Komputer*, vol. 7, no. 1, pp. 8–16, 2020, doi: 10.30656/prosisko.v7i1.2129.
[36] Y. Farida, N. Ulinnuha, S. K. Sari, and L. N. Desinaini, "Comparing Support Vector Machine and Naïve Bayes Methods with A Selection of Fast Correlation Based Filter Features in Detecting Parkinson's Disease," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 14, no. 2, p. 80, 2023, doi: 10.24843/lkjiti.2023.v14.i02.p02.
[37] P. R. Undersampling, "Effect of Random Under sampling , Oversampling , and SMOTE on the Performance of Cardiovascular Disease Prediction Models terhadap Kinerja Model Prediksi Penyakit Kardiovaskular," *Jurnal Matematika Statistika dan Komputasi*, vol. 21, no. 1, pp. 88–102, 2024, doi: 10.20956/j.v21i1.35552.