Social Science and
Everant Journal
Humanities Journal

ISSN: 2456-2653
Volume 09 Issue 01 January 2025

https://sshjournal.com/
Impact Factor: 2024: 6.576
2023: 5.731

DOI: https://doi.org/10.18535/sshj.v9i01.1456

Early Prediction of Alzheimer's Disease using Random Forest and E-SS Algorithm

Kevin Chaily[1, a] | Victor Valentino[2, b*] | Jimmy Tjen[3, c] | Genrawan Hoendarto[4, d]

[1, 2, 3, 4] Informatics, Widya Dharma Pontianak University, Pontianak, West Kalimantan, Indonesia
[a]chailykevin@gmail.com,[b]victorvalentino484@gmail.com,[c]jimmytjen@widyadharma.ac.id,
[d]genrawan@widyadharma.ac.id

Received* *07-12-2024
Revised 08-12-2024
Accepted* *31-12-2024
Published 03-01-2025

Abstract:

Alzheimer Disease (AD) is a neurogenerative disorder that progressively damages the nervous system. Early detection of AD is crucial, as it allows patients to receive therapy at an earlier stage, helping to slow the progression of the disease. This research proposes a model with improved effectiveness and accuracy by combining Random Forest with Entropy-based Subset Selection (E-SS). E-SS is used to identify subsets of parameters that correlate with each other based on entropy. The results show that the combination of Random Forest and E-SS outperforms traditional Random Forest, Decision Tree, SVM, and k-NN models, achieving an accuracy of 95.81% while reducing the number of parameters from 33 to 29. This demonstrates that the proposed algorithm could be applied in the medical field, improving predictive accuracy by eliminating parameters with weak correlations to the disease.

Keywords: Alzheimer Disease, Random Forest, Entropy-based Subset Selection, Machine Learning.

Introduction:

Alzheimer Disease (AD) is a neurodegenerative disorder characterized by short-term memory loss, weakened cognitive abilities, impaired spatial cognition, and diminished executive function [1]. It is predicted that by 2025, at least 1 in 85 people will be affected by AD [2]. This rise is largely driven by an aging population, highlighting age as a significant risk factor in the development of AD. Since aging is inevitable, one way to minimize the impact of AD is by addressing modifiable factors such as lifestyle changes and increased physical activity [3]. Early detection is crucial, as an accurate diagnosis can help those affected by AD make necessary adjustments in lifestyle, financial planning, and mental health preparation [4, 5]. One effective approach for early detection is through the use of machine learning (ML) techniques.

ML is commonly applied in disease diagnosis to enhance decision-making capabilities [6]. Random

Forest (RF) is one of the key algorithms in ML, specializing in ensemble methods that combine multiple decision trees. Each tree is constructed using a random subset of data and features, which contributes to the model's robustness. Random Forest has several advantages, including resistance to overfitting and noise, making it well-suited for real-world datasets that may contain errors or outliers. Additionally, RF includes a feature called variable importance, which measures the significance of each feature in the dataset [7, 8].

In [9], a study was conducted on the use of RF in predicting heart disease, concluding that RF could classify whether someone has heart disease with 93.3% accuracy. Another study [10] applied the RF algorithm to predict stroke using demographic and behavioral data, showing that RF outperformed Decision Tree and Logistic Regression, achieving an accuracy of 94.11%. Additionally, [11] investigated the use of the RF to detect Diabetes Mellitus, using three different seeds in WEKA to observe results. Seed 2 achieved 98.24% accuracy, while seeds 1 and 3 both attained 97.94%. Furthermore, [12] compared the effectiveness of Naïve Bayes, Decision Tree, and RF in analyzing Chronic Kidney Disease. The results demonstrated that Naïve Bayes had an accuracy of 97.50%, the Decision Tree with the J48 algorithm reached 98.33% accuracy, and RF achieved 100% accuracy.

Based on the findings in [9, 10, 11, 12], RF consistently demonstrated the highest accuracy in each study. However, we believe there is still potential for improvement. Although RF includes its own feature selection mechanism, we propose combining RF with E-SS, an approach that has not yet been explored. E-SS identifies a subset of features based on entropy [13], with the goal of improving the algorithm's performance by reducing the number of features RF needs to evaluate, as these features will already be preselected by the E-SS algorithm. This research offers the following contributions:

1. A predictive algorithm that combines Random Forest and Entropy-based Subset Selection.

2. A model that can be used for the early prediction of Alzheimer's Disease.

This research paper is organized into five sections: Section 1 introduces the research topic, discusses related work, and outlines the methodology used. Section 2 details the preprocessing steps, and the research flow. Section 3 explains the employed algorithm, describes the dataset, and specifies the parameters utilized. Section 4 presents the research results using the proposed method and algorithm. Finally, Section 5 summarizes the findings and discusses potential future works.

Methodology:

This section of the article will explain the algorithm and research flow used to build a prediction model for AD based on Random Forest and E-SS. For a detailed understanding of the fundamentals of Random Forest, please refer to [8, 14] and for insights into E-SS, see [13].

Let H $= [\boldsymbol{y}\ X\ D]$, where $H \in \mathbb{R}^{m\times(n_1+n_2+1)}$ represents the dataset related to Alzheimer's Disease. Here, $y \in \{0,1\}^m$ indicates whether an individual has AD (e.g., $\boldsymbol{y} = 0$ means the patient does not have AD, while $\boldsymbol{y} = 1$ indicates the presence of AD). The matrix $X = [\boldsymbol{x_1}\ \boldsymbol{x_2} \ldots \boldsymbol{x_{n_1}}]$ consists of numeric parameters from the patients (e.g., MMSE, etc.), with $X \in \mathbb{R}^{m\times n_1}$, where n_1 is the number of numeric parameters. The matrix $D \in \mathbb{Z}^{m\times n_2}$ comprises discrete parameters from the patients (e.g., gender, and ethnicity), with n_2 representing the number of discrete parameters. In this context, m denotes the total number of samples in the dataset.

First Step : The dataset H is first used to create a prediction model using only the Random Forest algorithm. The performance of this model is then compared to the proposed method, which combines Random Forest with Entropy-based Subset Selection (E-SS), to assess whether the proposed method achieves higher accuracy. Let $H_v = [\boldsymbol{y}_v\ X_v\ D_v]$, where $H_v \in \mathbb{R}^{m_2 \times (n_1+n_2+1)}$ be the validation dataset, structured similarly to H, but not used during the training phase. Define $\%A_i$ as the prediction accuracy percentage of model i, calculated based on the validation dataset:

$$\%A_i = \frac{b}{m_2} \times 100\%. \qquad (1)$$

Where b_i represents the number of correct diagnoses made by the Random Forest-only prediction model, and m_2 is the total number of samples in the validation dataset H_v.

Second Step : In this step, the E-SS algorithm is applied to identify the parameters that exhibit correlations based on entropy. Let $\mathfrak{D}_i = [\boldsymbol{x_i}\ \boldsymbol{x_a}\ \boldsymbol{x_{a+1}} \dots \boldsymbol{x_{a^*}}]$, where $\mathfrak{D}_i \in \mathbb{R}^{m \times a^*}$, for $i = 1,2,3,\dots,n_1$, represents subset of X containing numeric parameters that are correlated based on entropy. Here, $x_a = [x_a(1)\ x_a(2) \dots x_a(m)]^{\intercal}$, and $x_a \in X$ represents the a-th parameters from X. In this step, $n_1 - 1$ subsets with entropy-based correlations will be identified. Each subset is represented as $H_i = [\boldsymbol{y}\ D\ \mathfrak{D}_i]$, where $H_i \in \mathbb{R}^{m \times (a^*+n_2+1)}$. This is the augmented version of $\mathfrak{D}_i$ paired with $\boldsymbol{y}$ dan D to provide complete information about the patients.

Third Step : In this step, a Random Forest model will be built for each data subset. For each H_i, suppose p_i is a Random Forest model built based on the parameters in H_i. Specifically :

$$\begin{aligned} &p_1 : y = F_{RF}(D, \mathfrak{D}_1), \\ &p_2 : y = F_{RF}(D, \mathfrak{D}_2), \\ &\vdots \\ &p_i : y = F_{RF}(D, \mathfrak{D}_i). \end{aligned} \qquad (2)$$

Where F_{RF} represents the Random Forest function [14].

Fourth Step : In this step, a search is conducted to determine which prediction model has the best accuracy by calculating the accuracy of each model using:

$$\%A_i = \frac{b_i}{m_2} \times 100\%. \qquad (3)$$

Where b_i, is the number of correct diagnoses based on the ith prediction model, and m_2 is the number of samples in the test dataset. The best prediction model is then used to perform the diagnosis, where p^* can be determined using :

$$p^* \coloneqq \{p_i : \underset{i}{\operatorname{argmax}}\, \%A_i\}. \qquad (4)$$

In other words, p_i with the highest accuracy. In this case, the parameters in $\mathfrak{D}^*$ (the numeric parameters from p^*) will be chosen as the most significant factors in the determining whether someone has AD or not. The entire algorithm for identifying the numeric parameters used using Entropy-Based Subset Selection is detailed in Algorithm 1 [15]

Algorithm 1: Identifying Parameters Based on Entropy

```
Input: Dataset described by matrix $X \in \mathbb{R}^{m \times n}$, variable to predict $j \in \boldsymbol{n}$, cardinality from the subset $n^* < n$
Output: Set of index $S \subset n, |S| = n^*$
Initialization:
$S \coloneqq \{j\}$
$n_s = \{i \subset n | r_{ij}^2 \geq r_{min}^2\}$
for $k = 1\colon \min(n^* - 1, n_s)$ do
$j^* = \underset{j \in n_s \setminus S}{\operatorname{argmin}} \widehat{H}(Z_{ij} | \beta_{ij})$
$S = S \cup j^*$
end for
```

Experimental Setup:

Dataset : The dataset used is obtained from the open-source website Kaggle [16]. This dataset contains 2,149 samples and 35 parameters related to Alzheimer’s Disease (AD). These parameters include both numeric variables.

Parameters : The 34 input parameters include : Patient ID, Age, Gender, Ethnicity, Education Level, BMI, Smoking, Alcohol Consumption, Physical Activity, Diet Quality, Sleep Quality, Family Historic Alzheimers, Cardio Vascular Disease, Diabetes, Depression, Head Injury, Hypertension, Systolic BP, Diastolic BP, Cholesterol Total, Cholesterol LDL, Cholesterol HDL, Cholesterol Triglycerides, MMSE, Functional Assessment, Memory Complaints, Behavioral Problem, ADL, Confusion, Disorientation, Personality Changes, Difficulty Completing Tasks, Forgetfulness, and Doctor In Charge. The output parameters is Diagnosis.

Pre-Processing : In this step, Patient ID and Doctor In Charge are removed because they do not contribute to determining whether or not someone has AD. Since E-SS is being used for parameter selection, the parameters are categorized into two types: numeric and discrete.

For the simulation, 1,504 samples (70% of the data) will be used for training, while the remaining 30% of the data will be used for testing to validate the accuracy of the proposed model using a confusion matrix.

Numeric Parameters : This section discusses the subset selection algorithm based on entropy and how it is used to create subsets of parameters that correlate with the presence of AD.

Let $X = [\boldsymbol{x_1}\ \boldsymbol{x_2} \ldots \boldsymbol{x_n}]$; $X \in \mathbb{R}^{m_t \times n}$ be a set of input parameters with numeric data type from Alzheimer’s dataset, where m_t represents the number of training data samples for diagnosis. For each $i = 1,2,\ldots n = 15$, Algorithm 1 will be performed i times to create subsets based on x_i parameters. These subsets are used in Random Forest model. Specifically, if $\boldsymbol{y} = [y(1)\ y(2) \ldots y(m_t)]^\intercal$ represents the diagnosis of AD, then the Random Forest function is expressed as :

$$y = f_{rf}(T\ \{x_i | i = 1,2,\cdots n;\ i \subset S_i\}\ . \tag{5}$$

Where S_i represents the subset of indices with a "good" entropy-based relationship with variable i and F_{rf} denotes the Random Forest function and T represents the number of trees created.

Table 1. Parameters Representation as a Variable

Parameters Name	Variable Representation	Parameters Name	Variable Representation
Age	$\boldsymbol{x_1}$	*Cholesterol Total*	$\boldsymbol{x_9}$
BMI	$\boldsymbol{x_2}$	*Cholesterol LDL*	$\boldsymbol{x_{10}}$
Alcohol Consumption	$\boldsymbol{x_3}$	*Cholesterol HDL*	$\boldsymbol{x_{11}}$
Physical Activity	$\boldsymbol{x_4}$	*Cholesterol Triglycerides*	$\boldsymbol{x_{12}}$
Diet Quality	$\boldsymbol{x_5}$	*MMSE*	$\boldsymbol{x_{13}}$
Sleep Quality	$\boldsymbol{x_6}$	*Functional Assessment*	$\boldsymbol{x_{14}}$
Systolic BP	$\boldsymbol{x_7}$	*ADL*	$\boldsymbol{x_{15}}$
Diastolic BP	$\boldsymbol{x_8}$		

Result and Discussion:

In this part, the result of the proposed model will be presented. For each model, the True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), False Negative Rate (FNR), and accuracy ($\%A$) will be shown using the following equation:

$TPR = \frac{TP}{TP + FN} \times 100\%,$	(6)
$TNR = \frac{TN}{TN + FP} \times 100\%,$	(7)
$FPR = \frac{FP}{TN + FP} \times 100\%,$	(8)
$FNR = \frac{FN}{FN + TP} \times 100\%,$	(9)
$\%A = \frac{TP + TN}{TP + TN + FP + FN}.$	(10)

With TP, TN, FN, and FP representing true positive, true negative, false positive, and false negative, respectively.

Regarding the cardinality or number of elements in each subset, it is determined that n* = 11. This is because 11 parameters achieved the highest level of accuracy among all tests conducted with n* ranging from 3 to 15.

Research Result. Table 2 shows the model predictive accuracy of the E-SS RF for diagnosing AD. Based on Table 2, the subset consisting of the parameters $x_{14}, x_{15}, x_5, x_{11}, x_3, x_7, x_9, x_8, x_1, x_4, x_{13} + D$ which includes Functional Assessment, ADL, Diet Quality, Cholesterol HDL, Alcohol Consumption, Systolic BP, Cholesterol Total, Diastolic BP, Age, Physical Activity, MMSE + D achieves the highest accuracy among all tested subsets when evaluated using the Random Forest model.

Table 2. Model Predictive Accuracy from Alzheimer Disease Diagnosis

Subset	%A	TPR	TNR	FPR	FNR
$x_2, x_{11}, x_{12}, x_{13}, x_8, x_{14}, x_5, x_9, x_1, x_3, x_{10} + D$	80.28	68.07	87.44	12.56	31.93
$x_3, x_2, x_{12}, x_8, x_{13}, x_9, x_5, x_{14}, x_1, x_{10}, x_4 + D$	77.95	71.10	81.46	18.54	28.90
$x_4, x_{12}, x_8, x_9, x_2, x_{14}, x_{13}, x_1, x_3, x_{10}, x_5 + D$	76.86	67.14	81.67	18.33	32.86
$x_5, x_4, x_9, x_{14}, x_2, x_8, x_7, x_1, x_3, x_{13}, x_{10} + D$	78.88	65.07	86.51	13.49	34.93
$x_6, x_7, x_3, x_{15}, x_4, x_8, x_9, x_2, x_{13}, x_{10}, x_5 \ + \ D$	76.71	63.51	83.14	16.86	36.49
$x_7, x_{12}, x_8, x_2, x_9, x_{14}, x_5, x_1, x_3, x_{10}, x_{13} + D$	80.90	70.34	85.98	14.02	29.66
$x_8, x_{11}, x_{12}, x_3, x_1, x_{14}, x_5, x_9, x_{13}, x_{10}, x_2 + D$	78.57	64.47	86.30	13.70	35.53
$x_9, x_{11}, x_{14}, x_5, x_1, x_3, x_{12}, x_8, x_{13}, x_{10}, x_2 + D$	78.57	69.16	83.69	16.31	30.84
$x_{10}, x_{11}, x_{14}, x_{12}, x_3, x_8, x_1, x_{13}, x_5, x_9, x_2 + D$	79.35	66.81	86.41	13.59	33.19
$x_{11}, x_2, x_{12}, x_{13}, x_8, x_{14}, x_5, x_9, x_1, x_3, x_{10} + D$	80.28	72.68	84.11	15.89	27.32
$x_{12}, x_{7,} x_8, x_2, x_9, x_{14}, x_5, x_1, x_3, x_{10}, x_{13} + D$	77.64	71.12	81.31	18.69	28.88
$x_{13}, x_7, x_4, x_8, x_9, x_2, x_{15}, x_5, x_3, x_1, x_{10} + D$	77.80	61.93	85.92	14.08	38.07
$x_{14}, x_{15}, x_5, x_{11}, x_3, x_7, x_9, x_8, x_1, x_4, x_{13} + D$	95.81	91.62	97.57	2.43	8.38
$x_{15}, x_{14}, x_5, x_{11}, x_3, x_7, x_9, x_8, x_1, x_4, x_{13} + D$	93.94	87.22	97.60	2.40	12.78

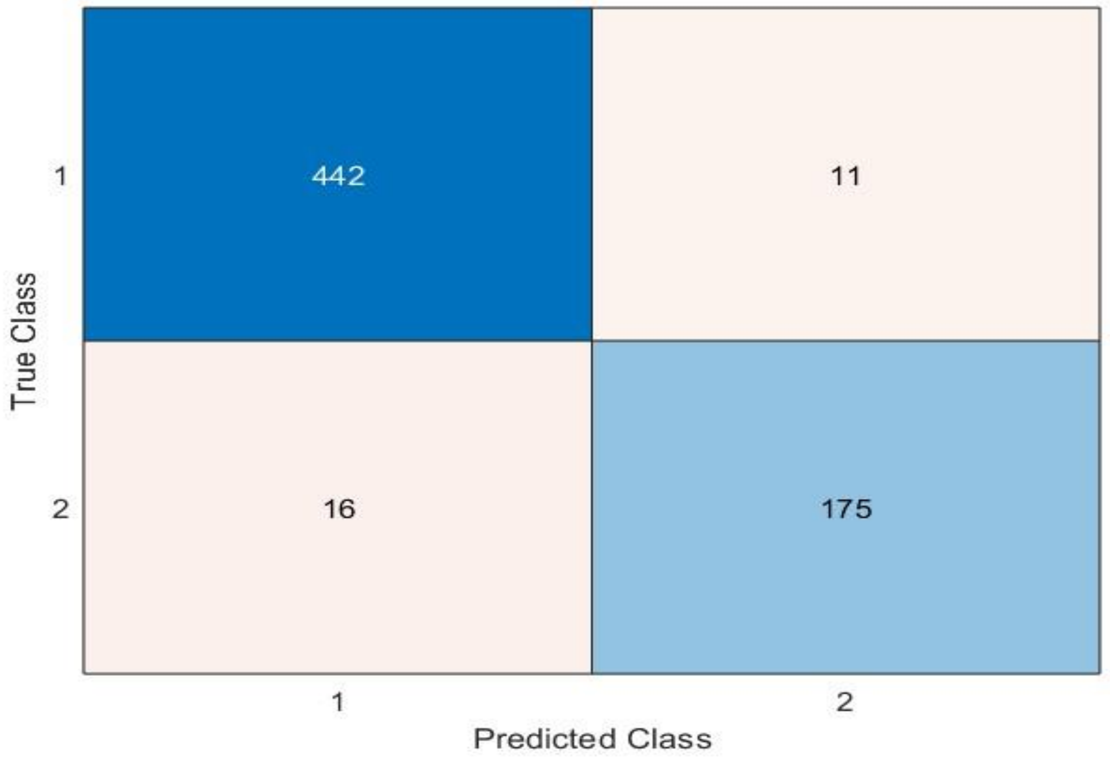

Figure 1. Confusion Matrix for $x_{14}, x_{15}, x_5, x_{11}, x_3, x_7, x_9, x_8, x_1, x_4, x_{13} + D$ model

Based on Fig. 1, The result of the test data, which included 644 samples (30% of the total data), using the subset with the highest accuracy correctly predicted 175 positive cases and 442 negative cases of AD, but fails to correctly predict 11 positive cases and 16 negative cases.

Table 3. Model Prediction Accuracy Comparison Between Random Forest and other Algorithm

Model	%A	TPR	TNR	FPR	FNR
Random Forest Classic	93.94	87.40	97.78	2.22	12.60
Random Forest + E-SS	95.81	91.62	97.57	2.43	8.38
Decision Tree Classifier	91.77	87.82	94.08	5.91	12.18
SVM	82.60	69.74	90.15	9.85	30.25
k-NN	57.43	26.47	75.62	24.38	73.53

Based on Table 3, The accuracy of multiple models is presented, as analyzed by the author. It can be seen that Random Forest with E-SS algorithm achieves the highest accuracy of 95.81%, compared to other models such as Random Forest Classic, which has the second highest of 93.94%, Decision Tree Classifier with the third highest accuracy of 91.77%, SVM with 82.60% accuracy, and k-NN, which has the lowest accuracy of 57.43%. The higher accuracy Random Forest + E-SS compared to Random Forest

Classic indicates that the E-SS algorithm enhances the model's performance by increasing its accuracy.

Discussion:

Based on Table 1, the total number of parameters initially was 33. These parameters were divided into two subsets: discrete and numeric. The discrete subset includes Gender, Ethnicity, Education Level, Smoking, Family Historic Alzheimers, Cardio Vascular Disease, Diabetes, Depression, Head Injury, Hypertension, Functional Assessment, Memory Complaints, Behavioral Problem, Confusion, Disorientation, Personality Changes, Difficulty Completing Tasks, Forgetfulness. The numeric subsets includes Age, BMI, Alcohol Consumption, Physical Activity, Diet Quality, Sleep Quality, Systolic BP, Diastolic BP, Cholesterol Total, Cholesterol LDL, Cholesterol HDL, Cholesterol Triglycerides, MMSE, Functional Assessment, ADL.

The numeric subset was then used with the E-SS algorithm to reduce the number of parameters by creating smaller subsets that combine parameters with better correlation. This resulted in a reduction from 15 to 11 parameters used from the numeric subset which is 4 fewer than previously. Therefore, the total number of parameters to be used in the model is 29 parameters (down from the original 33 parameters), The Random Forest model was trained on 70% of the data, with the remaining 30% used for testing. The model's performance was evaluated using confusion matrix (TPR, TNR, FPR, FNR) and accuracy metrics. The results were as follows: TPR at 91.62%, indicating the model correctly predicting positive cases, TNR at 97.57%, indicating the model correctly predicting negative cases, FPR at 2.43%, showing the model's incorrect predictions of positive cases. FNR at 8.38%, showing the model's incorrect prediction of negative cases. The prediction accuracy was 95.81%.

The combination of parameters with better correlation led to an accuracy of 95.81%, approximately 2% higher than that of Random Forest Classic in diagnosing Alzheimer's Disease. This suggests that the proposed model outperforms Random Forest Classic while utilizing fewer parameters.

Among the 11 numeric parameters selected from the best-performing subset, Age, Functional Assessment, and ADL (Activities of Daily Living) were key example used to predict the likelihood of Alzheimer's Disease. Previous research supports this, showing that age is a significant factor in the risk of developing AD, with older individuals having a higher likelihood of onset [3, 17]. Moreover, individuals with AD typically exhibit a reduced ability to perform daily tasks such as feeding, bathing, and dressing compared to the general population [18].

The high accuracy of this model in making correct predictions indicates its reliability in the early prediction of Alzheimer's Disease. As a result this model also has potential to assist in the medical field by being developed into software for reliable early diagnosis Alzheimer's Disease.

Conclusion:

Based on the research conducted, the use of E-SS algorithm resulted in a reduction of parameters down to 29 (initially 33) by selecting those with better correlation. The chosen parameter include Functional Assessment, ADL, Diet Quality, Cholesterol HDL, Alcohol Consumption, Systolic BP, Cholesterol Total, Diastolic BP, Age, Physical Activity, MMSE. This combination of parameters when applied with the Random Forest algorithm, achieves a prediction accuracy of 95.81%, indicating the proposed model outperforms the Random Forest Classic, Decision Tree Classifier, SVM, and k-NN using the same dataset with all the parameters. This improvement in accuracy suggest that the E-SS algorithm enhances predictive performance and holds potential for practical application in the medical field to help with early diagnosis of Alzheimer's Disease. Future research could incorporate additional data sources such as larger and more diverse datasets, and explore other machine learning algorithm to be combined with the E-SS algorithm.

References:

[1] E. Area-Gomez and E. A. Schon, "Alzheimer disease," *Adv Exp Med Biol*, vol. 997, no. 1, pp. 149–156, 2017, doi: 10.1007/978-981-10-4567-7_11.

[2] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the global burden of Alzheimer's disease," *Alzheimer's and Dementia*, vol. 3, no. 3, pp. 186–191, 2007,

doi: 10.1016/j.jalz.2007.04.381.

[3] R. A. Armstrong, "Risk factors for Alzheimer's disease," *Folia Neuropathol*, vol. 57, no. 2, pp. 87–105, 2019,

doi: 10.5114/fn.2019.85929.

[4] R. Green, RC; Clarke, VC; Thompson, NJ; Woodard, JL; Letz, "Early detection of Alzheimer disease: methods, markers, and misgivings," *Alzheimer Dis Assoc Disord*, vol. 11, no. 5, p. S1, 1997,

doi: 10.2217/fnl.10.31.

[5] J. Rasmussen and H. Langerman, "Alzheimer's Disease – Why We Need Early Diagnosis," *Degener Neurol Neuromuscul Dis*, vol. Volume 9, pp. 123–130, 2019, doi: 10.2147/dnnd.s228939.

[6] I. Konenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif Intell Med*, vol. 23, no. 1, pp. 89–109, 2001, [Online]. Available: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed5&NEWS=N&AN=2001260608

[7] M. Pal, "Random forest classifier for remote sensing classification," *Int J Remote Sens*, vol. 26, no. 1, pp. 217–222, 2005,

doi: 10.1080/01431160412331269698.

[8] L. Breiman, "Random Forests," 2001.

[9] M. Pal and S. Parija, "Prediction of Heart Diseases using Random Forest," *J Phys Conf Ser*, vol. 1817, no. 1, 2021,

doi: 10.1088/1742-6596/1817/1/012009.

[10] O. Shobayo, O. Zachariah, M. O. Odusami, and B. Ogunleye, "Prediction of Stroke Disease with Demographic and Behavioural Data Using Random Forest Algorithm," *Analytics*, vol. 2, no. 3, pp. 604–617, 2023,

doi: 10.3390/analytics2030034.

[11] M. T. Islam, M. Raihan, F. Farzana, N. Aktar, P. Ghosh, and S. Kabiraj, "Typical and Non-Typical Diabetes Disease Prediction using Random Forest Algorithm," *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*, pp. 1–6, 2020,

doi: 10.1109/ICCCNT49239.2020.9225430.

[12] R. S. Walse, G. D. Kurundkar, S. D. Khamitkar, A. A. Muley, P. U. Bhalchandra, and S. N. Lokhande, "Effective Use of Naïve Bayes, Decision Tree, and Random Forest Techniques for Analysis of Chronic Kidney Disease," in *Information and Communication Technology for Intelligent Systems*, T.

Senjyu, P. N. Mahalle, T. Perumal, and A. Joshi, Eds., Springer Singapore, 2021, pp. 237–245.

[13] A. Smarra; Francesco; Tjen, Jimmy; D'Innocenzo, "Learning methods for structural damage detection via entropy-based sensors selection," *International Journal of Robust and Nonlinear Control*, vol. 32, no. 10, pp. 6035–6067, 2022, doi: 10.1002/rnc.6124.

[14] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.

[15] J. Tjen, "Identifikasi Parameter Kualitas Bahan Pangan dengan Metode *Entropy-Based Subset Selection* (E-SS) (Studi Kasus: Minuman Anggur)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, pp. 47–54, 2024, doi: 10.25126/jtiik.20241116850.

[16] R. El Kharoua, "Alzheimer's Disease Dataset," 2024, *Kaggle*.

[17] R. A. Armstrong, "What causes Alzheimer's disease?," *Folia Neuropathol*, vol. 51, no. 3, pp. 169–188, 2013, doi: 10.5114/fn.2013.37702.

[18] M. Kamiya, A. Osawa, I. Kondo, and T. Sakurai, "Factors associated with cognitive function that cause a decline in the level of activities of daily living in Alzheimer's disease," *Geriatr Gerontol Int*, vol. 18, no. 1, pp. 50–56, 2018,

doi: 10.1111/ggi.13135.