Detecting Alzheimer’s Disease by The Decision Tree Methods Based On Particle Swarm Optimization

R A Saputra\textsuperscript{1*}, C Agustina\textsuperscript{1}, D Puspitasari\textsuperscript{2}, R Ramanda\textsuperscript{3}, Warjiyono\textsuperscript{1}, D Pribadi\textsuperscript{4}, Lisnawanty\textsuperscript{1}, K Indriani\textsuperscript{2}

\textsuperscript{1}Sistem Informasi Akuntansi, Universitas Bina Sarana Informatika, Indonesia
\textsuperscript{2}Sistem Informasi, Universitas Bina Sarana Informatika, Indonesia
\textsuperscript{3}Teknik Informatika, STMIK Nusa Mandiri, Indonesia
\textsuperscript{4}Ilmu Komputer, Universitas Bina Sarana Informatika, Indonesia

E-mail: rizal.rga@bsi.ac.id

Abstract. In this study aims to determine the classification of Alzheimer’s disease, this disease is a dangerous disease that can eliminate memory loss and can even result in a loss of ability to remember. For this reason, early detection of this disease is needed so that it can prepare for medical treatment. In this study the proposed method is to compare several decision tree methods with feature or attribute selection using the Particle Swarm Optimization (PSO) algorithm with the Alzheimer OASIS 2 dataset: Longitudinal Data from kaggle.com. The results of experiments with ten-fold cross validation, by testing the decision tree algorithm before the feature or attribute selection is performed, the highest accuracy value is obtained from the random forest algorithm with a value of 91.15%. The feature selection process is carried out using the PSO algorithm and the experiment is repeated using the Decision tree, the PSO-based random forest algorithm has the highest accuracy value of 93.56% with a kappa value of 0.884. Feature or attribute selection using the PSO algorithm is proven to be able to improve the accuracy of the decision tree algorithm, and is included in the algorithm with a very good range of values.

1. Introduction
Alzheimer’s disease is the most common trigger for dementia that occurs in older people [1], this disease attacks the nervous system of the brain which can irreversibly trigger the loss of neuron cells resulting in decreased memory [2] which results in a loss of ability to remember, difficulty in communication, thinking clearly, changes in behavior and ability to take care of themselves [1]. Memory loss in patients with Alzheimer’s occurs gradually which is experienced for three to nine years [2]. It is estimated that around 44 million people worldwide in 2015 have Alzheimer’s disease or dementia [3]. This disease is recognized as a fatal disease that leads to death and is the main cause of death ranked 7 in high-income countries [4]. In Indonesia, Alzheimer’s disease continues to increase from 1 million sufferers in 2013 to 1.2 million in 2015 and is predicted to continue to increase to 2.2 million by 2020. However, until now there has not been found a drug that can cure this disease [5]. Therefore, the importance of early detection to begin to plan for treatment and adequate medical needs as well as testing new drugs for the disease [6].

In diagnosing Alzheimer’s disease, one method that can be used is to apply data mining classification methods. The classification method is included in the supervised learning method...
used to find input attributes and target attributes [7]. Research on Alzheimer’s disease using data mining classification methods has been widely carried out, such as research [8] comparing the DA, NN, SVM, DT, KNN and TANN algorithms, obtained the TANN method has a high performance of 98%. Research [9] uses an algorithmic approach to detect Alzheimer’s disease from non-image data using an affiliation algorithm model with an accuracy of 87%. Research [10] applies the KNN algorithm with two tasks, namely the reduction of dimensions of the high dimensional feature vector and the classification that produces an adaptive class, and the KNN algorithm is proven to work well. Research [11] using three different feature extractions namely GMV, GLCM, Gabor and improvised by the SVM-RFE algorithm, proved that the algorithm was effective in improving classification performance. Research [2] using Fuzzy C-Mean for feature Segmentation, GLCM for feature extraction and SVM for classification, obtained 93.33% accuracy results.

Decision tree algorithm is the best algorithm in every decision making, where the decision tree algorithm has several models namely CHAID, C4.5, ID3 and Random Forest [12]. This algorithm has been widely used by researchers for the case of classification and prediction, but it has a weakness that often overlaps, especially in classes and criteria which are quite numerous, so it requires time and the amount of memory to increase [13][14].

Particle Swarm Optimization (PSO) is an effective optimization algorithm that can solve problems in the decision tree algorithm [15], because PSO can optimize overlapping datasets and can select a large number of attribute features [16]. This study aims to predict Alzheimer’s disease using the Alzheimer’s OASIS 2 dataset: Longitudinal Data obtained from the Kaggle site and published by Jacob Boysen, by finding the best accuracy value of the PSO-based Decision Tree algorithm and testing using 10-fold cross-validation.

2. Methods

2.1. C4.5 Algorithm

C4.5 algorithm is part of the decision tree algorithm group, this algorithm is one of the solutions to solving cases that are often used in solving classification model problems and including the 10 best most popular algorithms [17]. The C4.5 algorithm has two equations, the first calculates the entropy value and the second calculates the gain value, for the Gain value itself its function is to select attributes as a node, for the root node is chosen based on the highest Gain value of the existing attributes. Following the C4.5 algorithms are: Calculating the entropy value used by the formula:

\[
Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \left( \frac{|S_i|}{|S|} \right) \cdot Entropy(S_i)
\]  

Note: S as the set of cases; A as a feature; n as the partition number attribute A; Si as the proportion of Si to S; and S as the number of cases in S.

\[
Entropy(S) = -\sum_{i=1}^{n} P_i \log_2 P_i
\]

Note: S as the set of cases; A as an attribute; n as the number of partitions S and pi as the proportion of Si to S.

2.2. ID3 (Iterative Dichotomy 3) Algorithm

Iterative Dichotomy 3 (ID3) is a tree classification algorithm invented by J. Ross Quinlan (1979), by utilizing Shanon’s information theory or Information Theory [18]. Following is the equation
of the ID3 algorithm to find the value of Information Entropy and Information Gain:

$$\text{Entropy} (S) = \sum_{j=1}^{k} -P_j \log_2 P_j$$

(3)

Note: S as a dataset (dataset) of cases; k as many S partitions; pj as the probability obtained from Sum (Yes) divided by Total Cases. After getting the entropy value, the attribute selection is done by assessing the largest information gain.

$$\text{Gain} (A) = \text{Entropy} (S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \text{Entropy} (S_i)$$

(4)

Note: S as the sample data space used for training; A as an attribute; Si as the number of samples for the value of V; S as the sum of all data samples and Entropy (S) as entropy for samples that have a value of i.

2.3. CHAID (Chi-Square Automatic Interaction Detection) Algorithm

The CHAID algorithm is a classification tree technique that tests independent variables one by one, separating the categories in the variables used in the analysis phase, the arrangement is based on the level of statistical significance of chi-square of the dependent variable[19], three stages must be done in the analysis CHAID is a merger, separation, and termination. At the merge stage contingency tables are formed of the dependent variables and independent variables that have been categorized.

$$X^2 = \sum_{r=1}^{M_r} \sum_{k=1}^{n_k} \frac{(0_{rk} - e_{rk})^2}{e_{rk}}$$

(5)

Note: 0_{rk} and e_{rk} as the number of observations and expectation values in the ker row and k-column. The hypothesis testing criteria is to reject H0 if $X^2$ count > tabel$^2$ table = 0.05 with degree of freedom (k-1) (r-1).

2.4. Random Forest Algorithm

Random Forest is an ensemble learning method used to improve the accuracy of a data classification [20]. Increased accuracy is done through a combination of many sorters from a similar method and gets a final classification prediction through the voting process [21]. In Random Forest, many trees are used so that a forest or forest is formed and then each tree is analyzed continuously [22]. Equations for calculating the Gini Index in the classification, use the equation below [22]:

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2$$

(6)

Note: the value of pi is the probability of the object to be classified in a certain class/feature.

2.5. Particle swarm optimization (PSO) Algorithm

PSO is an optimization algorithm developed by Kennedy and Eberhart in 1995. This algorithm is inspired by the lifestyle of the movements of birds and fish. The advantage of the Particle Swarm Optimization method is that it has a simple concept, but is easy to implement, and is efficient in calculations than other mathematical algorithms[23]. Broadly speaking, Kennedy and Eberhart describe the PSO algorithm with the equation below:

$$v_t = v_i (t - 1) + \varphi C_1 (p_i - x_i (t - 1)) + \varphi C_2 (g - x_i (t - 1))$$

(7)
\[ x_i(t) = x_i(t-1) + v_i(t) \]  \hspace{2cm} (8)

Note: X\(_i\) as candidate for particle completion; P\(_i\) as experience of best particle position; g as information of best particle environment; X\(_i\)(t-1) as current particle position; \(\varphi\) as random vector with score range of 0.1; \(C_1\) as coefficient of cognitive learning and \(C_2\) as coefficient of social learning.

2.6. Proposed Methods

Figure 1. The Proposed Methods

Based on the description of the method proposed in Figure 1, the flow of research can be explained through the following steps:

1. Starting from the selection of datasets that will be used in the research process, in this study using Alzheimer’s disease dataset from https://www.oasis-brains.org/OASIS
2. The next step is to divide the dataset into 10 parts using 10-fold cross-validations, the distribution of the dataset will be training and testing data.
3. After dividing the dataset, then selection of features or attributes using the PSO method.
4. The next step after obtaining the best attributes, then testing is done using a comparative Decision Tree algorithm (C4.5, ID3, CHAID and Random Forest).
5. Meanwhile, the testing data whose attributes are selected are validated by comparing the Decision Tree algorithm.
6. After the data transfer and data testing are validated using the Decision Tree comparison algorithm, the final step is the presentation of the results of the data classification comparison.

3. Result and Discussion

This stage explains the results of the experiment using an algorithm that is included in the decision tree with feature or attribute selection using PSO, following table 1 the results of the decision tree algorithm experiment using 10-fold cross-validation before using the feature or attribute selection:
Table 1. Experiment Result Comparison Algorithm Decision Tree

|               | C4.5   | ID3    | CHAID  | Random Forest |
|---------------|--------|--------|--------|---------------|
| Accuracy      | 88.45% | 89.54% | 89.54% | 91.15%        |
| Kappa         | 0.794  | 0.804  | 0.808  | 0.836         |

In table 1 it can be seen that the random forest algorithm has the highest accuracy of 91.15% with a kappa value of 0.836.

After that the feature or attribute selection is performed using PSO, the results of the feature or attribute selection that affect the classification of Alzheimer’s disease, MMSE and nWBV attributes have a value of 0, this shows that these attributes do not affect the results of the classification, the results of the selection of features or attributes can be seen in table 2 below:

Table 2. Result Feature Selection

| Attribute | Weight |
|-----------|--------|
| M/F       | 0.994  |
| Age       | 1      |
| EDUC      | 0.451  |
| SES       | 0.819  |
| MMSE      | 0      |
| CDR       | 0.985  |
| eTIV      | 0.162  |
| nWBV      | 0      |
| ASF       | 0.112  |

The results of this feature or attribute selection will be retested using a decision tree classification algorithm using 10-fold cross-validation. The following table 3 experimental results from the PSO-based decision tree algorithm:

Table 3. Experiment Result Comparison Algorithm Decision Tree based PSO

|               | C4.5+PSO | ID3+PSO | CHAID+PSO | Random Forest+PSO |
|---------------|----------|---------|-----------|-------------------|
| Accuracy      | 90.36%   | 89.55%  | 89.55%    | 93.56%            |
| Kappa         | 0.826    | 0.808   | 0.808     | 0.884             |

From table 3 above it can be explained, that the highest accuracy value obtained by PSO-based random forest algorithm is 93.56% with kappa value of 0.884, these results prove that feature or attribute selection using PSO algorithm can improve the performance of the accuracy value of the decision tree classification algorithm, it can be seen from table 1 and table 3 the results of comparison of decision tree algorithms before and after using PSO. And based on the Altman theory in 1991, the Kappa range of 0.81-1.00 is very good [29], so the kappa value in the PSO Based Random Forest Algorithm is very good, because it has a kappa value of 0.884.
4. Conclusion
The results of the comparison of decision tree algorithms using feature selection or PSO algorithm attributes, PSO-based random forest algorithm has the highest accuracy value of the other algorithms that is equal to 93.56% with a kappa value of 0.884, in other words that the PSO algorithm is able to improve the accuracy of the random forest algorithm, and all Decision tree algorithm has a kappa value which is categorized very good. This research can be developed by comparing the data classification method and optimizing feature selection or other attributes, so that it has better accuracy than the current research results.

Acknowledgments
The authors would like to convey their gratitude to Universitas Bina Sarana Informatika for giving a chance and big support in conducting this research project and for providing all the equipment facilities.

References
[1] Birks J S and Harvey R J, 2009, Donepezil for dementia due to Alzheimer’s disease, in John Wiley Sons, .
[2] Novitasari D C R Pupitasari W T Wulandari P and Foeady A Z, 2018 Klasifikasi Alzheimer Dan Non Alzheimer Mengunakan Fuzzy C-Mean , Gray Level Co-Occurrence Matrix Dan Support Vector Machine J. Mat. “Mantik” 04, 02 p. 83–89.
[3] Cauwenberghs C V Van Broeckhoven C V and Sleeegers K, 2016 Open The genetic landscape of Alzheimer disease,clinical implications and perspectives 18, 5.
[4] Dailey C, 2016 The Impact of Alzheimer’s Disease -The Silent Killer JCCC Honor. J. 7, 2 p. 1–16.
[5] Friskahaja H Ilhamsyah and Barlian Y A, 2018 Perancangan Kampanye Pernmainan Teka Teki Silang Sebagai Pencegahan Penyakit Alzheimer Di Usia Dewasa in e-Proceeding of Art And Design 5, 3 p. 1827–1831.
[6] Samper-González J et al., 2018 Reproducible evaluation of classification methods in Alzheimer’s disease, Framework and application to MRI and PET data Neuroimage 183, March p. 504–521.
[7] Hendriyanto S, 2018 Algoritma Klasifikasi Data Mining Untuk Memprediksi 11, 3 p. 266–274.
[8] M. E F. A and H. M, 2016 Automatic Detection and Classification of Alzheimer’s Disease from MRI using TANNN Int. J. Comput. Appl. 148, 9 p. 30–34.
[9] Aditya C R and Pande M B S, 2016 An Algorithmic Approach for Alzheimer’s Disease detection from Non-Image Data Int. J. Curr. Eng. Technol. 6, 3 p. 784–787.
[10] Papakostas G A Savio A Graña M and Kaburlasos V G, 2014 A lattice computing approach to Alzheimer’s disease computer assisted diagnosis based on MRI data Elsevier 150, Part A p. 37–42.
[11] Xiao Z Ding Y Lan T Zhang C Luo C and Qin Z, 2017 Brain MR Image Classification for Alzheimer’s Disease Diagnosis Based on Multifeature Fusion Comput. Math. Methods Med. 2017 p. 1–13.
[12] Patel H H and Prajapati P, 2018 Study and Analysis of Decision Tree Based Classification Algorithms Int. J. Comput. Sci. Eng. 6, 10 p. 74–78.
[13] Cieslak D A Hoenis T R Chawla N V. and Kegelmeyer W P, 2012 Hellinger distance decision trees are robust and skew-insensitive Data Min. Knowl. Discov. 24, 1 p. 136–158.
[14] Longadge R Dongre S S and Malik L, 2013 Class Imbalance Problem in Data Mining: Review Int. J. Comput. Sci. Netw. 2.
[15] Arifin T and Herliana A, 2020 Optimasi decision tree menggunakan particle swarm optimization untuk identifikasi penyakit mata berdasarkan analisa tekstur J. Teknom. dan Sist. Komput. 8, 1 p. 59–63.
[16] Muzakkir I Syukur A and Dewi I N, 2014 Backpropagation Dengan Seleksi Fitur Particle Swarm Optimization Dalam Predikisi Pelanggan Telekomunikasi J. Pseudocode 1 p. 1–10.
[17] Han J and Kamber M, 2006 Data Mining Concepts and Techniques Second Edition San Francisco. Morgan Kaufmann.
[18] Ying-ying W Yi-bin L and Xue-wen R, 2017 Improvement of ID3 Algorithm Based on Simplified Information Entropy and Coordination Degree Algorithms 10, 4 p. 1526–1530.
[19] Helena K Susanti Y and Respatiwulan, 2019 Penerapan Metode Chi-Squared Automatic Interaction Detection (CHAID) dan Classification And Regression Trees (CART) pada Klasifikasi Status Kerja di Kabupaten Brebes in Seminar Nasional Sains dan Entrepeneurship IV.
[20] Breiman L, 2001 Random Forests Mach. Learn. 45, 1 p. 5–32.
[21] Ali J Khan R Ahmad N and Maqsood I, 2012 Random Forests and Decision Trees IJCSII (international J. Comput. Sci. Issues) 9, 5 p. 272–278.
[22] Dewi N K Syafitri U D Mulyadi S Y Statistika M D and Statistika D, 2011 Penerapan Metode Random Forest Dalam Driver Analysis Forum Stat. Dan Komputasi 16, 1 p. 35–43. [ 
[23] Herliana A Arifin T and Susanti S, 2018 Feature Selection of Diabetic Retinopathy Disease Using Particle Swarm Optimization and Neural Network in International Conference on Cyber and IT Service Management (CITSM 2018).
[24] Altman D G, 1991, Practical Statistics For Medical Research, (London: CRS Press), p. 277-300.