International Journal of Computer, Network Security and Information System (IJCONSIST)
Vol. 7 No. 1, September 2025

Application of the Random Forest Classifier Method in Grouping Patients with Intellectual Disabilities

Nuchaila Ainiyah
Data Science Study Program
Telkom University, Surabaya Campus
Surabaya, Indonesia
nuchailaainiyah@telkomuniversity.ac.id

Muhammad Afifudin
Departement of Informatics, Faculty of Computer Science
UPN "Veteran" Jatim
Surabaya, Indonesia
20081010187@student.upnjatim.ac.id

Reyhan Dela Masyhuri
Departement of Informatics, Faculty of Computer Science
UPN "Veteran" Jatim
Surabaya, Indonesia

Muhamad Hakam Fardana
Departement of Informatics, Faculty of Computer Science
UPN "Veteran" Jatim
Surabaya, Indonesia

Sischa Wahyuningtyas
Departement of Data Science, Faculty of Computer Science
UPN "Veteran" Jatim
Surabaya, Indonesia
sischa_wahyuningtyas.sada@upnjatim.ac.id

Awang Putra Sembada R
Departement of Data Science, Faculty of Computer Science
UPN "Veteran" Jatim
Surabaya, Indonesia
awang_putra.sada@upnjatim.ac.id

Muhamad Liswansyah Pratama
Departement of Data Science, Faculty of Computer Science
UPN "Veteran" Jatim
Surabaya, Indonesia
muhamad_liswansyah.sada@upnjatim.ac.id

***Abstract*—This research explores the effectiveness of the Random Forest Classifier method in grouping mental retardation patients based on their level of severity. Medical record data from mental hospitals is collected and processed to train a classification model. The preprocessing process is applied to ensure data quality before use. Model evaluation is carried out by measuring the accuracy of the scores. The research results showed that the Random Forest Classifier succeeded in classifying mental retardation patients with an accuracy of 84%. These findings show the potential of the Random Forest Classifier method as a clinical tool for doctors in determining appropriate interventions for mental retardation patients based on their level of severity**

Keywords— mental retardation; classification; random forest

I. INTRODUCTION

According to [1], approximately one in ten children worldwide experiences some form of disability. In Indonesia, although recent comprehensive data remain limited, [2] reported a child population of 82.8 million, with around 10% identified as having disabilities. Data from [3] further indicated that there were approximately 130,573 children with disabilities, of whom 30,460 were diagnosed with intellectual disabilities.

Intellectual disability, formerly known as mental retardation, is a developmental disorder that occurs before the age of 18 and affects an individual's cognitive functioning and adaptive abilities [4]. This condition results in limitations in reasoning, problem-solving, and daily living skills, including communication, learning, and social interaction appropriate to age and cultural context.

Diagnosis of intellectual disability is typically conducted through interviews with family members, clinical observations, and IQ testing to assess the individual's ability to adapt to their environment [5]. Although the condition is lifelong and cannot

be cured, appropriate therapeutic interventions can significantly improve independence and quality of life [3].

In the field of technology, machine learning algorithms such as Random Forest have been increasingly applied to assist in the classification and diagnosis of developmental disorders, including intellectual disabilities [6]. Random Forest operates by constructing multiple decision trees from different subsets of data to generate more accurate predictions and reduce the risk of overfitting [7,8]. This algorithm is recognized for its robustness, high accuracy, and ease of application across various classification tasks [6].

This study aims to evaluate the effectiveness of the Random Forest algorithm in classifying types of intellectual disabilities in psychiatric hospitals, with the goal of improving diagnostic accuracy and supporting more precise therapeutic planning.

[9] conducted a study on the classification of individuals with intellectual disabilities using the Classification and Regression Tree (CART) method. Based on their analysis, the classification accuracy using the Gini index splitter reached 72.7% with 15 terminal nodes, while the Twoing index splitter achieved 71.4% accuracy with 11 terminal nodes. Although the model using the Gini index yielded slightly higher accuracy, the model with the Twoing index demonstrated a more complex tree structure. These findings indicate that both approaches are effective for classifying individuals with intellectual disabilities.

[10] analyzed the performance of the Backpropagation Neural Network (BPNN) algorithm with and without a momentum parameter in classifying intellectual disability disorders. The study found that adding a momentum value of 0.8 and a learning rate of 0.07 significantly improved accuracy to 100% with a 90:10 data split, whereas the model without momentum achieved only 84.62%. Thus, the use of momentum was proven to enhance the network's ability to recognize patterns and accelerate model convergence.

[11] conducted a study on early-stage diabetes prediction using Random Forest, Support Vector Machine (SVM), and Naïve Bayes classification algorithms. The performance of these algorithms was evaluated using Precision, Accuracy, Recall, and F-Measure metrics. The results showed that Random Forest achieved the highest accuracy of 97.88%, outperforming the other algorithms. These findings reaffirm the superiority of Random Forest in medical classification tasks that require high accuracy.

[12] further explored the classification of children with intellectual disabilities using the Backpropagation Momentum algorithm. The study utilized 127 data samples with 17 input variables obtained through interviews with psychologists regarding symptoms and severity levels of intellectual disability. The tested parameters included learning rate (0.01–0.07), momentum (0.1–0.9), and 17 hidden layer neurons. The results revealed that a combination of a 0.07 learning rate and 0.8 momentum produced the highest accuracy, reinforcing the effectiveness of the Backpropagation Momentum method for classifying intellectual disabilities.

II. METHOD

A. K-Fold Cross Validation

K-fold cross-validation is a cross-validation technique that divides the data into k equal-sized subsets. Then, the model is trained on k-1 subsets and tested on the remaining subsets. This process is repeated k times, with each subset used as a test set once. K-fold cross-validation is a technique for estimating the error rate of digital image processing tests. K-fold cross-validation works by grouping the training and test data into separate subsets, then repeating the testing process k times [13].

In k-fold cross-validation, we divide the dataset into k equal-sized subsets. Then, we use each subset as validation data and use the other subsets as training data. In this way, we can calculate the model's accuracy score on each subset and calculate the average accuracy score across all subsets.
Here are the steps for performing k-fold cross-validation:

A. Determine the value of k.
B. Divide the data into k equal-sized subsets.
C. For each iteration:
   a. Select one subset as the test set.
   b. Train the model on the other k-1 subsets.
   c. Evaluate model performance on the test set.
D. Calculate the average model performance across all iterations.

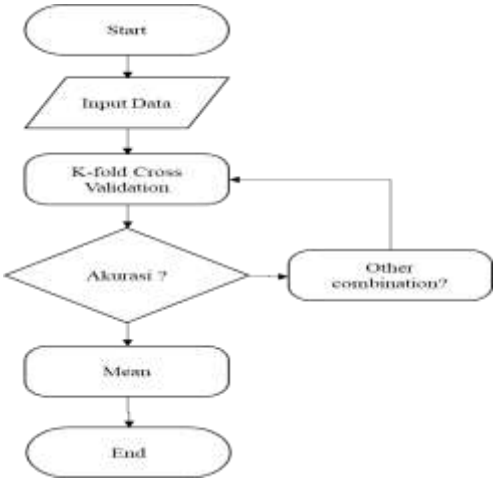
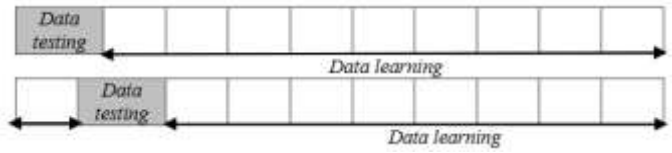


Fig 1. Flowchart of the K-Fold Cross Validation Method

For example, suppose we have a dataset of 1,000 samples. We can divide this dataset into 10 folds. Then, we train the model on 9 folds (900 samples) and validate on the remaining fold (100 samples). This process is repeated 10 times, with each fold serving as the validation set exactly once. Finally, we average the performance across all 10 folds to get a more robust estimate of the model's performance. Here's how it looks in the figure:



And so on until the 10th fold is formed as follows

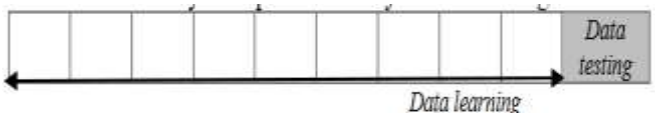


Fig 2. Visualization of Model Performance

B. Random Forest Classifier

The random forest classifier algorithm is a machine learning algorithm used for data classification. This algorithm works by constructing multiple decision trees and then combining the results from these trees to make predictions. Random Forest is a combination of each existing decision tree technique, which is then combined and combined into a single model [14].

Random Forest maps attributes from classes so that it can be used to find predictions for data that has not yet appeared [15]. Decision trees themselves are a "divide and conquer" approach to learning problems from a set of independent data depicted in a tree chart [16].

The Random Forest algorithm is a widely used machine learning technique for classification and regression. It works by constructing multiple individual decision trees ("forests"), then combining the predictions from all decision trees to produce a more accurate and stable final output. The Random Forest (RF) method is an extension of the Classification and Regression Tree (CART) method, which applies bootstrap aggregation (bagging) and random feature selection [17]. Random Forest is a method used for classification by building many classification trees. This method can improve accuracy results by generating child nodes for each node (the node above it) and selecting them randomly [18].

The following is the formula for Random Forest:

$$Entropy(Y) = -\sum p(c|Y)log^2\ \mathrm{p(c|Y)} \quad (1)$$

Description:

Y = Set of cases

P(c|Y) = Proportion of Y values to class c

$$\text{Information Gain (Y, a)} = \text{Entropy (Y)} - \sum_v \varepsilon Values(a) \frac{|Y_v|}{|Y_a|} Entropy\ (Y_v). \quad (2)$$

Description:

Values(a) = Possible values in the set of cases a.

Y_v = Subclass of Y with class v corresponding to class a.

Y_a = All values corresponding to a.

C. Confusion Matrix

A confusion matrix is a table that classifies the number of correct and incorrect test data [19]. A confusion matrix is a method used to calculate accuracy in data mining concepts. The confusion matrix evaluation is a prediction matrix that will perform testing to estimate correct and incorrect objects to produce accuracy, precision, and recall values. Precision or confidence is the proportion of predicted positive cases that are also true positives in the actual data. Recall or sensitivity is the proportion of true positive cases that are correctly predicted positive [20].

TABLE 1. CONFUSION MATRIX

Correct classification	Prediction Class	
	1	0
1	TP	FN
0	FP	TN

A Confusion Matrix is a table containing four different combinations of predicted and actual values. When measuring performance using a Confusion Matrix, four terms represent the results of the classification process. These four terms are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Negative (TN) represents the number of correctly detected negative data, while False Positive (FP) represents negative data detected as positive. True Positive (TP) represents positive data correctly detected. False Negative (FN) is the opposite of True Positive, meaning positive data detected as negative [21].

The calculation of Accuracy, Precision, Recall, and RMSE using the Confusion Matrix table is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recal = \frac{TP}{TP + FN} \quad (5)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - Y_j)^2}{n}} \quad (6)$$

Description:

Y_i = initial data (actual data)

Y_j = final data (estimated data)

n = number of data

In this study, the evaluation score Accuracy and RMSE will be used.

D. Flowchart

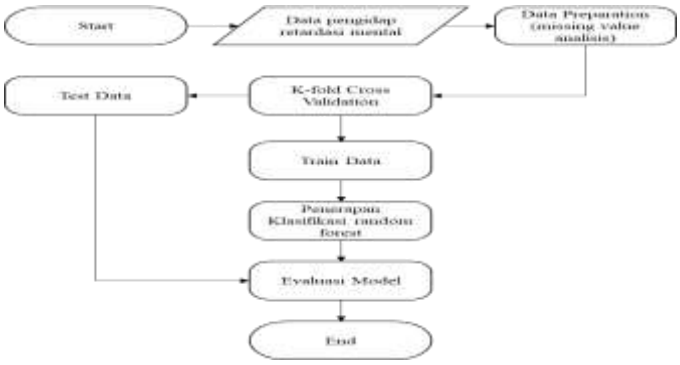


Fig 3. Flowchart for Data Collection of Mental Retardation Sufferers

This study begins with the collection of data on individuals with mental retardation. The collected data can be clinical data, such as the results of physical examinations, psychological examinations, and laboratory tests. This data is then cleaned to remove incomplete or irrelevant data. Next, the data preparation stage continues, namely preparing the data so that the data can be processed for training the random forest model. This includes data standardization and cleaning missing values. Then, the training data will be divided and the test data will be divided. To avoid overfitting, the K-fold cross validation method is used. After that, the model with the best performance quality will be selected from the evaluation process with a confusion matrix. The final stage is model implementation. This stage is carried out to apply the classification model to classify new data. The new data to be classified can be clinical data of new patients or data from other medical examinations.

E. Data Sources

The data used in this final project is secondary data obtained from a Mental Hospital. Data collection was conducted through the medical records of patients with mental retardation undergoing outpatient treatment. This process was carried out carefully and meticulously to ensure the accuracy and validity of the information. Patients were selected for inclusion in this study based on certain predetermined criteria, including patient age and relevant clinical characteristics, such as genetic factors, health factors, and environmental factors.
The research variables used consisted of a response variable (Y) and a predictor variable (X_n). The predictor variables used were factors influencing the classification of patients with mental retardation, determined based on previous studies and the availability of patient medical records at Menur Mental Hospital, East Java Province. The response variable (Y_n), consisting of mild, moderate, and severe types, was determined based on the mental condition experienced by patients with mental retardation, including etiological factors and psychiatric symptoms, as well as other variables that influence the classification.

III. Result and Discussion

In this section, the result derived from the computational process are presented and analyzed to evaluate their significance and alignment with the research objectives.

The total number of individuals diagnosed with mental retardation in the dataset is 124. Among these cases, 45 patients (approximately 36%) were classified as having mild severity, making it the most frequently observed category. Furthermore, 37 patients (30%) exhibited moderate severity, while 42 patients (around 34%) were categorized as severe. Overall, the distribution of severity levels among patients appears relatively balanced across the three categories. Based on the computational results, the model effectively captured these variations, demonstrating its ability to classify mental retardation severity levels with consistent accuracy across different categories. The detailed distribution can be observed in the following figure.

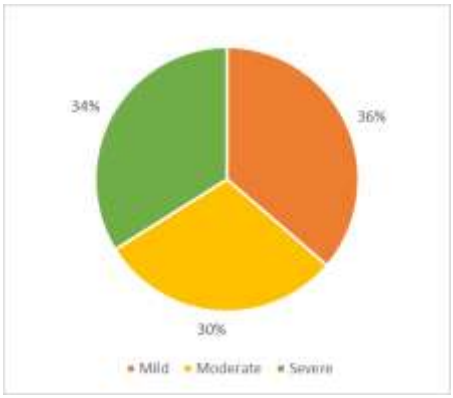


Fig. 4. Composition of Individuals with Mental Retardation

The comparison of the number and gender of individuals with mental retardation based on the severity level can be seen in the following figure.

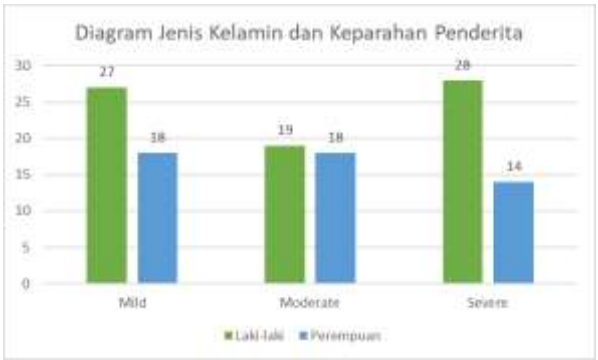


Fig. 5. Comparison of the Number and Gender of Individuals with Mental Retardation Based on Severity Level

From the dataset, it can also be observed that the composition of each gender varies across the different severity levels of mental retardation. When examined across the three categories—mild, moderate, and severe—it is evident that the number of male patients is higher than that of female patients in each severity group. This indicates a gender disparity in the prevalence of mental retardation, suggesting that males may

have a greater susceptibility or higher detection rate compared to females within the observed population.

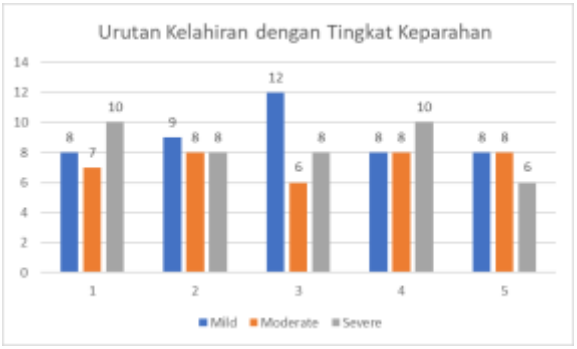


Fig. 7. Comparison of Individuals with Mental Retardation Based on Birth Order

In addition, the dataset provides information regarding the birth order indicator for each type of RM (mental retardation) case. The graph illustrates the relationship between birth order and the severity level of the disorder, where the x-axis represents birth order and the y-axis represents the severity level. The results show that first-born and fifth-born children tend to exhibit lower levels of severity compared to middle-born children (second, third, and fourth). This finding suggests that birth order may have a potential influence on the severity of RM, possibly related to prenatal or environmental factors affecting later born children.

Furthermore, statistical summaries indicate that approximately 40% of RM cases occur among second- and third-born children, while only around 25% are found among first-borns. This pattern may imply that increased maternal age, cumulative exposure to environmental risk factors, or variations in prenatal care could contribute to higher severity levels among later-born children. However, further analysis with a larger and more diverse dataset is required to validate these observations and to better understand the causal relationship between birth order and the severity of mental retardation. This relationship can be explained in detail in Table 2.

TABLE 2. DESCRIPTIVE STATISTICS OF THE AGE OF INDIVIDUALS WITH MENTAL RETARDATION

RM Type	Mean	Std. Deviation	Min	Max
Mild	8,11	5,50	0	17
Moderate	7,64	5,62	0	17
Severe	8,09	5,53	0	17

The table 2. presents the statistical values for each type of RM (mental retardation) along with the descriptive statistics of patients' ages, including the mean, standard deviation, and maximum values. The table consists of three rows representing the mild, moderate, and severe RM types. For mild RM, the mean value is 8.11, the standard deviation is 5.50, and the minimum and maximum values are 0 and 17, respectively. For moderate RM, the mean is 7.64, the standard deviation is 5.62, and the range is also between 0 and 17. Similarly, for severe RM, the mean value is 8.09, with a standard deviation of 5.53, and minimum and maximum values of 0 and 17.

Overall, the RM values across the three severity types exhibit an average close to 8 and a standard deviation around 6. This indicates that the data distribution is relatively normal, with most observations concentrated near the mean. The similarity in mean and variability across severity levels suggests that age differences among patients with varying RM severity are not substantially distinct, implying a consistent age distribution pattern within the dataset.

Based on the classification model developed using the Random Forest method, the evaluation results were obtained through two primary performance metrics: Accuracy Score and Root Mean Square Error (RMSE). These metrics were employed to assess the model's predictive capability and error rate. The Accuracy Score measures the proportion of correctly classified instances relative to the total number of cases, while the RMSE quantifies the average magnitude of prediction errors, indicating how closely the predicted values align with the actual data. The results of these evaluations are presented as follows, providing a comprehensive overview of the model's performance in classifying RM (mental retardation) severity levels.

TABLE 3. MODEL EVALUATION RESULTS

Evaluation Metrics	Random Forest Evaluation Score
Accuracy Score	0,84
RMSE	0,84

The table presents the evaluation metrics of the Random Forest classification model, specifically the accuracy score and Root Mean Square Error (RMSE). The accuracy score represents the average performance of the Random Forest model in correctly predicting the target variable, while RMSE measures the square root of the average squared differences between the predicted and actual values. In this study, the model achieved an accuracy score of 0.84, indicating that the Random Forest algorithm successfully predicted the target variable with 84% accuracy. The RMSE value was also 0.84, suggesting that, on average, the prediction error of the model was relatively moderate.

From a statistical perspective, a higher accuracy score reflects a stronger classification capability, meaning that the model is effective in identifying patterns within the dataset. Conversely, a lower RMSE value indicates better model precision, as it implies that the predicted outcomes are closer to the actual observations. Although the achieved accuracy demonstrates strong predictive performance, the RMSE value suggests that some variability remains unexplained by the model. This residual error could be attributed to data noise, limited feature representation, or class overlap within the dataset.

The performance results highlight that the Random Forest algorithm is a robust and interpretable method for classifying

RM severity levels. Its ensemble learning nature allows it to minimize overfitting and improve generalization by aggregating multiple decision trees. Nevertheless, the presence of moderate error values suggests that the model's predictive capability could be enhanced through several strategies. Future studies could consider hyperparameter optimization (adjusting the number of estimators, maximum depth, and feature subset size), data balancing techniques such as SMOTE to handle class imbalance, and feature importance analysis to identify the most influential predictors. Additionally, comparing the Random Forest model with more advanced machine learning algorithms such as XG Boost, Light GBM, or Support Vector Machines (SVM) could provide deeper insights into model efficiency and robustness across various datasets.

IV. CONCLUSION

The application of the Random Forest method to classify RM (Repetitive Movement) disorders in children proved to be effective. Through its pattern recognition capability, the model successfully identified and classified RM disorders based on the learning process outcomes.

The model achieved an accuracy score of 0.84, indicating that it was able to predict the target variable with 84% accuracy. Additionally, the RMSE value of 0.84 reflects a relatively low average prediction error, suggesting good model performance. However, future research could explore the integration of feature selection or dimensionality reduction techniques to enhance model interpretability and reduce potential overfitting.

Furthermore, comparing the Random Forest approach with other advanced machine learning models such as XGBoost, LightGBM, or deep learning frameworks could provide deeper insights and potentially improve classification accuracy for RM disorder detection in children.

ACKNOWLEDGMENT

The authors would like to express sincere gratitude to the Faculty of Computer Science, UPN "Veteran" Jawa Timur, for the support and facilities provided throughout this research. Appreciation is also extended to the research team and mentors for their valuable guidance and contributions.

REFERENCES

[1] UNICEF, "The state of the world's children 2023: children with disabilities", 2023. [Online]. Available: https://www.unicef.org/reports/state-of-worlds-children-2023

[2] Central Bureau of Statistics, "Statistik kependudukan Indonesia 2007". Jakarta: Badan Pusat Statistik, 2007.

[3] Ministry of Social Affairs of the Republic of Indonesia, "Laporan perlindungan sosial untuk penyandang disabilitas", Jakarta: Kementerian Sosial RI, 2020.

[4] American Psychiatric Association, "Diagnostic and statistical manual of mental disorders, 5th ed., text rev. Arlington", VA: American Psychiatric Publishing, 2022.

[5] R. L. Schalock, S. A. Borthwick-Duffy, V. J. Bradley, W. H. Buntinx, D. L. Coulter, E. M. Craig, "Intellectual disability: definition, classification, and systems of supports, 11th ed. Washington", DC: American Association on Intellectual and Developmental Disabilities, 2010.

[6] D. Yates and M. Z. Islam, "FastForest: Increasing random forest processing speed while maintaining accuracy," arXiv preprint arXiv:2010.12302, 2020. [Online]. Available: https://arxiv.org/abs/2010.12302.

[7] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[8] A. Liaw and M. Wiener, "Classification and regression by randomForest," R News, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/

[9] P. M. Khasanah and D. Rahmawati, "Analisis CART untuk klasifikasi penderita retardasi mental berdasarkan gejala psikiatrik," Jurnal Sains dan Informatika, vol. 8, no. 3, pp. 245–252, 2022

[10] N. Yanti, A. Ramadhan, and W. A. Sari, "Pengaruh momentum pada Backpropagation Neural Network untuk klasifikasi retardasi mental," Jurnal Teknologi Informasi dan Komputer (JTIK), vol. 7, no. 2, pp. 167–175, 2021

[11] W. Apriliah, R. Susanto, and F. Nugraha, "Prediksi diabetes tahap awal menggunakan algoritma klasifikasi Random Forest, SVM, dan Naive Bayes," Jurnal Ilmiah Informatika dan Komputer (JIKO), vol. 6, no. 4, pp. 321–330, 2021

[12] N. Yanti, W. A. Sari, and M. Hidayat, "Klasifikasi retardasi mental anak menggunakan metode backpropagation momentum," Jurnal Ilmu Komputer dan Aplikasi (JIKA), vol. 10, no. 1, pp. 45–53, 2022

[13] C. Rozikin, A. Susilo Yuda Irawan, dan U. Singaperbangsa Karawang, "Klasifikasi jenis buah mangga dengan metode backpropagation," Techné : Jurnal Ilmiah Elektroteknika, vol. 20, no. 1, hlm. 1–12, Mar 2021, doi: 10.31358/TECHNE.V20I1.231.

[14] [14] M. R. Adrian, M. P. Putra, M. H. Rafialdy, dan N. A. Rakhmawati, "Perbandingan metode klasifikasi random forest dan SVM pada analisis sentimen PSBB," Jurnal Informatika Upgris, vol. 7, no. 1, 2021. doi: 10.26877/JIU.V7I1.7099.

[15] A. U. Zailani dan N. L. Hanun, "Penerapan algoritma klasifikasi random forest untuk penentuan kelayakan pemberian kredit di koperasi mitra sejahtera," Infotech: Journal of Technology Information, vol. 6, no. 1, pp. 7–14, 2020. doi: 10.37365/JTI.V6I1.61.

[16] I. H. Witten, E. Frank, dan M. A. Hall, "Data mining: practical machine learning tools and techniques, 3rd ed. Burlington", MA: Morgan Kaufmann, 2011. doi: 10.1016/C2009-0-19715-5.

[17] L. Breiman, "Random forests," Machine Learning, vol. 45, 2001.

[18] G. A. Sandag, "Prediksi rating aplikasi app atore menggunakan algoritma random forest," CogITo Smart Journal, vol. 6, no. 2, pp. 167–178, Dec. 2020. doi: 10.31154/COGITO.V6I2.270.167-178.

[19] D. Normawati dan S. A. Prayogi, "Implementasi naïve bayes classifier dan confusion matrix pada analisis sentimen berbasis teks pada twitter," J-SAKTI (Jurnal Sains Komputer dan Informatika), vol. 5, no. 2, pp. 697–711, 2021. doi: 10.30645/J-SAKTI.V5I2.369.

[20] Herdiawan, "Analisis sentimen terhadap telkom indihome berdasarkan opini publik menggunakan metode improved k-nearest neighbor," Skripsi, Universitas Komputer Indonesia, 2016. [Online]. Available: http://elib.unikom.ac.id/gdl.php?mod=browse&op=read&id=jbptunikompp-gdl-herdiawann-33861

[21] F. Endah dan I. S. Encis, "Evaluasi dan prediksi penguasaan bahasa inggris maritim menggunakan metode decision tree dan confusion matrix (studi kasus di Universitas Maritim Amni)," Tidak dipublikasikan.