Classification Analysis of Stunting Data on Toddlers in Bekasi City Using Logistic Regression and Random Forest

1st Azka Nadhira
School of Computing
Telkom University
Bandung, Indonesia
azkanadhiraa@student.telkomuniversity.ac.id

2nd Putu Harry Gunawan
School of Computing
Telkom University
Bandung, Indonesia
phgunawan@telkomuniversity.ac.id

***Abstract*—The problem of stunting in toddlers is one of the pressing nutritional problems in Indonesia. This condition is a key indicator of maternal and infant health because it is caused by chronic malnutrition, especially in early life. Regular monitoring of toddler growth and development is necessary for stunting prevention, so developing accurate and efficient predictive models is essential. This study aims to analyze the performance of two Machine Learning models, namely Logistic Regression and Random Forest, in predicting stunting status in toddlers. The dataset used is toddler development data from the Kota Baru Health Center, Bekasi City. Due to the limited amount of data, the data split is done using k-fold cross-validation with k=5 to maximize the use of the dataset. Given the imbalanced data, with 39.4% stunting from the total data, an oversampling method was applied to the minority class. Logistic Regression was chosen for its simplicity and interpretability, while Random Forest was chosen for its ability to handle data complexity by combining predictions from multiple decision trees. The evaluation results show that Random Forest performs better, with an average accuracy of 88.07% and F1-score macro average of 87.32%, compared to Logistic Regression, which has an average accuracy of 85.48% and F1-score macro average of 84.73%. This research is expected to support stunting prevention in Indonesia by providing a predictive approach based on Machine Learning to improve the monitoring of the growth and development of toddlers.**

***Index Terms*—stunting, logistic regression, random forest**

I. Introduction

Stunting in toddlers is a major nutritional problem in Indonesia that can be a key indicator of maternal and infant health because this condition is caused by chronic malnutrition [1]. Stunting can cause inhibition of physical growth, disruption of motor skills, and suboptimal health development [2]. Based on the Indonesian Nutrition Status Survey for 2022, the prevalence of national stunting decreased from 24. 4% in 2021 to 21.6% in 2022. Bekasi City plays an active role in collecting data on toddler growth and development and running nutrition monitoring and counseling programs, based on data from www.kemkes.go.id, in line with the national target to reduce the prevalence of stunting to 14% by 2024.

One of the most effective ways to address stunting is through prevention to eliminate the issue before it arises, it is necessary to regularly monitor the growth and development of toddlers. Machine Learning can help identify stunting status efficiently and accurately based on weight, height, and age. In early 2024, research by Gustriansyah et al. [3] showed that Random Forest was the best method for predicting the nutritional status of toddlers with an accuracy of 97.37% using a confusion matrix and k-fold cross-validation, followed by a Support Vector Machine with an accuracy of 96.09%. Meanwhile, research by Rahman et al. [4] showed that Random Forest excels in stunting classification with 88.3% accuracy, followed by SVM with 88.1% accuracy and Logistic Regression with 87.7% accuracy.

Random Forest was employed to predict toddler stunting conditions, utilizing 10-fold cross-validation and achieving an average accuracy of 97.87% [5]. Then, research [6] used the oversampling method with SMOTE (Synthetic Minority Oversampling Technique) from the Python imblearn.over_sampling module, which successfully increased the accuracy of Random Forest from 98.83% to 99.77%. In addition, research [7] compared Machine Learning models, such as Naive Bayes, K-Nearest Neighbors, and Random Forest, to predict stunting using a limited dataset. Random Forest showed the best performance with an accuracy of 87.75% and F1-score of 92.2%, thus demonstrating this model's effectiveness in detecting stunting.

This study aims to determine the stunting status of toddlers by utilizing a dataset from the Kota Baru Health Center, Bekasi City, consisting of 193 physical measurement records of toddlers. The relatively small size of the dataset compared to other prediction datasets is a challenge, so this research applies validation techniques such as k-fold cross-validation to maximize the use of data and ensure reliable analysis results. The study was conducted by comparing two Machine Learning models, namely Logistic Regression and Random Forest, to predict the stunting status of toddlers in Bekasi City. This research aims to identify the best-performing model that can be applied practically. The results of this research are expected to contribute to stunting prevention efforts in Indonesia by improving accuracy and efficiency in identifying

stunting status in toddlers.

II. Methods

A. Research Design

This research starts with a literature review to get the background of the problem and ideas of methods that can be used. The next step is to collect data from Kota Baru Health Care. The collected data then went through the data understanding stage to understand each feature. After that, feature selection is carried out to select relevant features. The features that have been chosen then go through the data cleaning and data conversion stages to fit the required format. The processed data is visualized to make it easier to understand. Next, the data is balanced to ensure the dataset is balanced before applying the method. After preprocessing, the Logistic Regression and Random Forest methods are implemented on the data. The accuracy of the two methods is compared to get the best results. The comparison results are used to formulate a conclusion. Fig. 1 illustrates the flowchart of the research process.

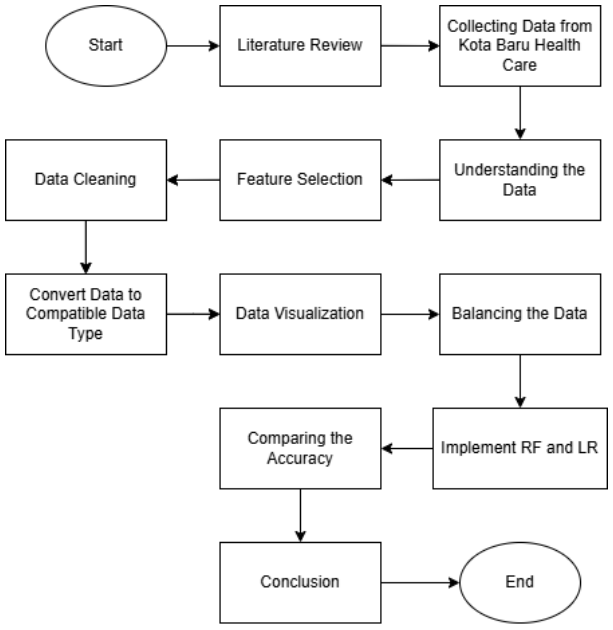

Fig. 1. Flowchart of Research

B. Logistic Regression

Logistic Regression is a simple algorithm with interpretability and the ability to handle continuous or more than two explanatory variables simultaneously [4] [8]. This model uses mathematical equations to model the relationship between the independent and dependent variables and sigmoid functions to generate probabilities [9]. Logistic Regression models the likelihood of an outcome based on individual characteristics [8]. The following is the sigmoid function formula:

$$P(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

Where $P(x)$ denotes a probability estimation function with a value between 0 and 1, x is the predicted value of the algorithm, and e is a mathematical constant valued at 2.71828 [10].

C. Random Forest

Random Forest is a diverse Decision Tree ensemble learning algorithm that excels at capturing non-linear relationships [11]. Its bootstrap method and random feature selection reduce variance and the risk of overfitting [12] [13] [14]. The final prediction is determined through majority voting of all Decision Trees, with the final result being the class that gets the most votes [15]. Algorithm 1 demonstrates an implementation of the Random Forest algorithm [16].

Algorithm 1 Random Forest

Input: D: training dataset, c: number of decision trees.
Output: Classification result from the Random Forest.
Build c **decision trees:**
for $i = 1$ to c **do**
  Create bootstrap sample D_t from D.
  Initialize root node N_t with D_t.
  Call **BuildTree**(N_t).
end for

BuildTree(N)**:**
if all data in N belongs to the same class **then**
  Return.
else
  Randomly select $x\%$ of the features in N.
  Identify the feature F that maximizes information gain.
  Split N into f nodes $N_1, N_2, \ldots, N_f$, where f equals the possible values of F.
  for each child node N_i **do**
    Assign the subset of D matching F_i to N_i.
    Recursively call **BuildTree**(N_i).
  end for
end if

D. K-Fold Cross-Validation

K-fold cross-validation is a model evaluation technique that divides the dataset into k folds of equal size, where the model is trained on k-1 folds as training data and tested on the remaining folds until each fold becomes the test data once [5]. This process is repeated k times, resulting in an average of the evaluation metrics from all iterations for more stable and accurate performance estimation. This technique ensures that all data is optimally used for training and testing, making it incredibly compelling. This study, 5-fold cross-validation was used, with each subset of data alternately serving as test data and k-1 fold as training data. This method helps to reduce data splitting bias, optimize model parameters, and achieve lower empirical risk than simple data splitting [17].

E. Evaluation Metrics

Evaluation metrics such as accuracy, precision, recall, and F1-score are essential for assessing the performance of models predicting stunting in toddlers. These metrics provide insights into different aspects of model effectiveness in binary classification, such as distinguishing between stunted and non-stunted cases. By using their respective formulas, these metrics help evaluate the model's overall accuracy, ability to correctly identify stunted cases (precision), detect all stunted cases (recall), and balance precision and recall (F1-score). Here's a detailed explanation of these metrics, including their formulas and how they relate to predicting stunting [18].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\text{-}Score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right) \tag{5}$$

In these metrics, TP (True Positive) refers to correctly identified stunted cases, TN (True Negative) refers to correctly identified non-stunted cases, FP (False Positive) refers to non-stunted cases misclassified as stunted, and FN (False Negative) refers to stunted cases misclassified as non-stunted.

III. RESULTS AND DISCUSSIONS

A. Exploratory Data Analysis

The dataset used in this study consists of child growth and development data collected from Kota Baru Health Center, Bekasi City. This dataset includes 193 entries of measurements taken in August 2023 and January 2024. The features analyzed include Age, $Height$, $Weight$, $Gender$, $Birth\ Weight$, $Birth\ Height$, and $Height/Age$. $Birth\ Weight$ and $Birth\ Height$ are considered relevant factors due to their association with fetal growth and their role as early indicators of a child's growth potential. Low birth weight and short birth height are widely recognized as significant predictors of stunting, as they reflect suboptimal fetal growth and development. Meanwhile, $Gender$ was included as a variable to account for potential differences in growth patterns and stunting risk between males and females [19]. The $Height/Age$ variable was used as the target for model training, as it contains information on the classifying of a child's status.

In Fig. 2, the distribution of the target variable shows a class imbalance, where the 'Stunting' class has less data than the majority class, the 'Normal' class. This imbalance poses a serious challenge, as the model tends to be biased towards the majority class, reducing its predictive ability to detect the minority class accurately. Class imbalance could affect the overall performance of the model.

Furthermore, Fig. 3 shows the feature distribution analysis, where the $Weight$ feature shows extreme values at both boundaries, with one outlier at the upper boundary. Meanwhile, the $Height$ feature has an outlier at the lower boundary. Additionally, the $Birth\ Weight$ and $Birth\ Height$ features also exhibit outliers at both the upper and lower boundaries. These extreme values can introduce noise and affect the stability of the model's predictions, requiring special handling to ensure optimal performance. On the other hand, the Age feature has no outliers, indicating a more stable data distribution for the feature.

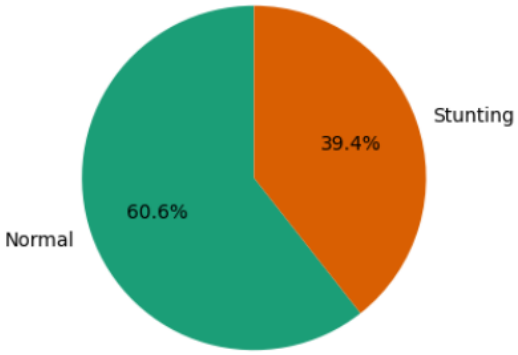

Fig. 2. Class Distribution on Target Feature

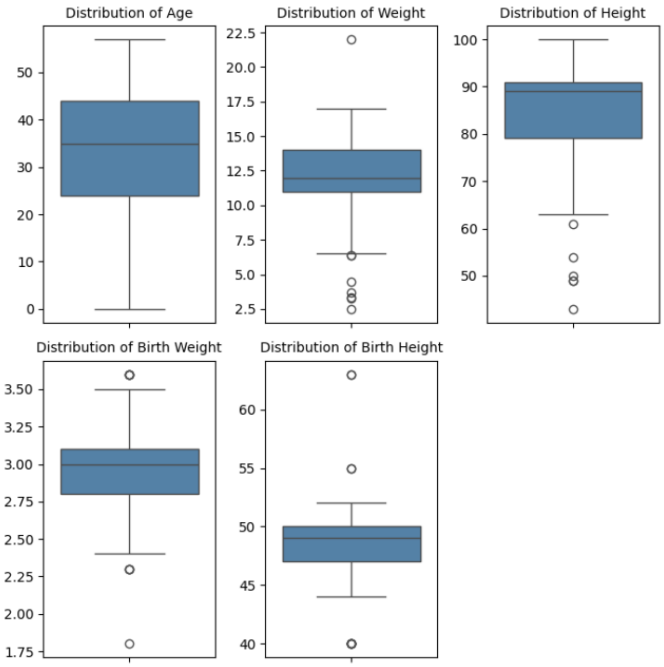

Fig. 3. Visualization of Numerical Feature Distribution Before Preprocessing

B. Data Preprocessing

The stunting dataset with 193 records was processed for analysis using seven out of 35 features, including the target feature. In the $Height/Age$ and $Gender$ columns, the data type was converted to integers using encoding, with the class 'Stunting' mapped to 1 and 'Normal' to 0 . Additionally, for the $Gender$ column, 'Male' was encoded as 0 and 'Female' as 1. The Age column was converted to age in months for

consistency. Missing values in the $Height$, $Weight$, and $Birth\ Height$ columns were resolved by calculating the average based on the 'Stunting' and 'Normal' categories.

To outliers in numerical features were handled using the Interquartile Range (IQR) method by replacing outlier values with the median, and Winsorization was applied to limit extreme values [20]. Feature selection was performed by reducing 35 to seven relevant features to improve model accuracy. The preprocessing results show that missing values and outliers have been effectively managed, as shown in Fig. 4.

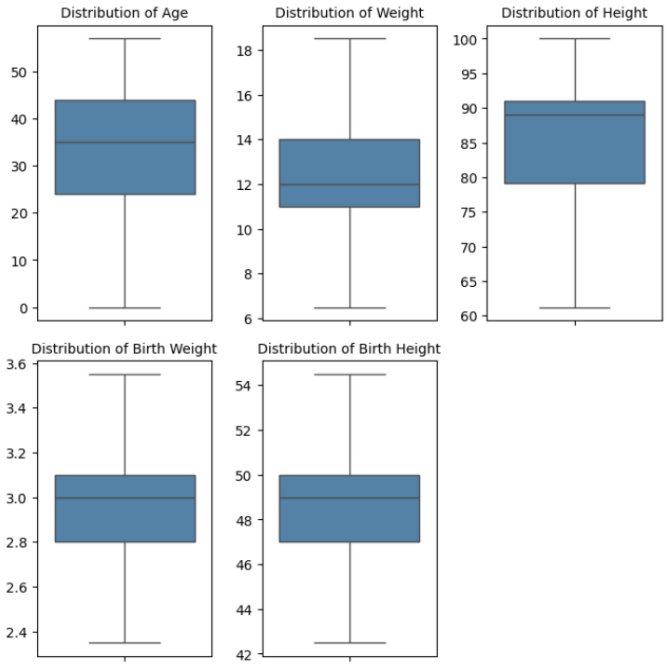

Fig. 4. Visualization of Numerical Feature Distribution After Preprocessing

Based on Fig. 3, the dataset shows a class imbalance with 117 cases of 'Normal' and 76 cases of 'Stunting'. This imbalance can cause the model to predict the majority class (Normal) more accurately but often misclassify the minority class (Stunting). To overcome this problem, the Synthetic Minority Over-sampling Technique (SMOTE) is used, which generates synthetic samples for the minority class so that the class distribution becomes balanced [21].

Before applying SMOTE, the data is divided using 5-fold cross-validation, where four subsets are used as training data and one subset as testing data in turn. Next, the features in the training data are scaled using StandardScaler and applied to the testing data to maintain consistency. Before oversampling, the average number of samples per class in the training data was 93 for 'Normal' and 62 for 'Stunting.' In contrast, after applying SMOTE, the class distribution balanced with an average of 93 samples per class in each fold. The slight variation between folds resulted from the different training data distribution in each fold.

C. Implementation of Logistic Regression

During training, Logistic Regression from the Scikit-learn library uses a gradient descent-based optimization method with a maximum number of iterations of 1000 to find the optimal parameters (coefficients) and ensure optimal model convergence. The model utilizes a logistic (sigmoid) activation function to estimate class probabilities, with the L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) solver, which was chosen for its efficiency in handling small to medium-sized datasets [21]. Before training, the dataset was processed using StandardScaler from Scikit-learn to scale the features, ensuring the variable distribution has a mean of 0 and a standard deviation of 1. In addition, the SMOTE (Synthetic Minority Oversampling Technique) method was applied to address the class imbalance between stunted and normal data.

Based on the confusion matrix presented in Table I, which represents the aggregated results across all folds of cross-validation, Logistic Regression implemented through Scikit-learn demonstrates good performance by utilizing linear relationships between features, making it an effective method for the initial classification of stunting cases.

TABLE I
CONFUSION MATRIX FOR THE LR METHOD

		Predicted Values	
		Stunting	**Normal**
Actual Values	**Stunting**	63	13
	Normal	15	102

D. Implementation of Random Forest

Random Forest captures non-linear patterns by forming many decision trees that collectively provide predictions based on the majority principle. The model was implemented using the Scikitlearn library with key parameters such as n_estimators=100, Gini splitting criteria, no maximum depth restriction (max_depth=None), and bootstrap=True technique to build each tree. The dataset was processed using StandardScaler to scale the features and SMOTE to handle class imbalance between stunted and normal. Random Forest builds 100 decision trees, where the final prediction result is an aggregation of the predictions of each tree.

Based on the confusion matrix evaluation results presented in Table II, which aggregates the results across all folds of cross-validation, Random Forest shows better performance than Logistic Regression with fewer misclassifications. These confusion matrix results demonstrate that Random Forest excels at recognizing complex patterns between features, making it more reliable for stunting data classification.

TABLE II
CONFUSION MATRIX FOR THE RF METHOD

		Predicted Values	
		Stunting	**Normal**
Actual Values	**Stunting**	61	15
	Normal	8	109

E. Model Evaluation

TABLE III
COMPARISON OF METHODS

Method	Accuracy	F-1 Score Macro Avg
Logistic Regression	85.48%	84.73%
Random Forest	88.07%	87.32%

Table III presents the evaluation results of the Logistic Regression and Random Forest models based on the accuracy and F1-Score Macro Average metrics. These two metrics are used to assess the model's ability to perform classification by considering the balance between precision and recall in each class. The evaluation results, aggregated across all folds of 5-fold cross-validation, show that Logistic Regression produces an accuracy of 85.48% and F1-Score Macro Average of 84.73%. The limitation of Logistic Regression in capturing non-linear patterns due to its linear approach makes this model less optimal for datasets with complex patterns. This observation is in line with previous research [22], which discusses the use of Logistic Regression in determining child nutritional status and stunting and mentions the limitations of this method in handling complex feature interactions.

In contrast, Random Forest showed superior performance with an accuracy of 88.07% and F1-Score Macro Average of 87.32%, also calculated from the aggregated results across all folds of 5-fold cross-validation. Random Forest captures complex patterns and interactions between features more effectively by utilizing an ensemble of 100 decision trees. The model was also more accurate in recognizing the minority class (stunting). This finding is consistent with previous research [3], which shows that Random Forest excels in capturing complex data patterns. Based on these results, Random Forest is better than Logistic Regression for stunting dataset classification because it produces more substantial generalization and higher accuracy.

Using 5-fold cross-validation to evaluate both models is a correct and valid practice. This process helps produce a more stable and reliable estimation of model performance. Each fold provides accuracy and F1-score values, which are then averaged to give an overall picture of the model's performance. Previous research [23] also emphasizes the importance of cross-validation in producing more stable and reliable accuracy estimates.

IV. CONCLUSION

Stunting is a significant public health problem in Indonesia, influenced by poor nutrition, poor sanitation, and socioeconomic disparities. This study uses data from Bekasi City to develop a machine learning model that supports early detection of stunting, with evaluation through 5-fold cross-validation and application of SMOTE technique to overcome class imbalance. The results showed that the Random Forest model performed better than Logistic Regression, with an average accuracy of 88.07% and a macro average F1-score of 87.32%, outperforming Logistic Regression, which achieved an accuracy of 85.48% and F1-score of 84.73%. Random Forest excels because it can better capture complex patterns between features and handle class imbalance. The health department can implement this Random Forest model to prioritize interventions, such as supplementary nutrition programs or health monitoring in children at high risk of stunting. In addition, this research is expected to support stunting prevention efforts in Indonesia by providing a Machine Learning-based predictive approach that can be used to improve monitoring of the growth and development of children under five. Future research is recommended to explore more advanced algorithms, such as Gradient Boosting Machines or Neural Networks.

REFERENCES

[1] S. A. Mashar, S. Suhartono, and B. Budiono, "Faktor-faktor yang mempengaruhi kejadian stunting pada anak: Studi literatur," *Jurnal Serambi Engineering*, vol. 6, no. 3, 2021.

[2] N. Latifah, F. Fajrini, N. Romdhona, D. Herdiansyah, E. Ernyasih, and S. Suherman, "Systematic literature review: Stunting pada balita di indonesia dan faktor yang mempengaruhinya," *Jurnal Kedokteran dan Kesehatan*, vol. 20, no. 1, pp. 55–73, 2024.

[3] R. Gustriansyah, N. Suhandi, S. Puspasari, and A. Sanmorino, "Machine learning method to predict the toddlers' nutritional status," *JURNAL INFOTEL*, vol. 16, no. 1, pp. 32–43, 2024.

[4] S. J. Rahman, N. F. Ahmed, M. M. Abedin, B. Ahammed, M. Ali, M. J. Rahman, and M. Maniruzzaman, "Investigate the risk factors of stunting, wasting, and underweight among under-five bangladeshi children and its prediction based on machine learning approach," *Plos one*, vol. 16, no. 6, p. e0253172, 2021.

[5] A. Y. Perdana, R. Latuconsina, and A. Dinimaharawati, "Prediksi stunting pada balita dengan algoritma random forest," *eProceedings of Engineering*, vol. 8, no. 5, 2021.

[6] A. A. Dhani and W. J. Pranoto, "Perbaikan akurasi random forest dengan anova dan smote pada klasifikasi data stunting," *Teknika*, vol. 13, no. 2, pp. 264–272, 2024.

[7] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, "Analisis perbandingan algoritma machine learning untuk prediksi stunting pada anak: Comparative analysis of machine learning algorithms for predicting child stunting," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 257–265, 2024.

[8] S. Sperandei, "Understanding logistic regression analysis," *Biochemia medica*, vol. 24, no. 1, pp. 12–18, 2014.

[9] S. Lonang, A. Yudhana, and M. K. Biddinika, "Analisis komparatif kinerja algoritma machine learning untuk deteksi stunting," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 4, pp. 2109–2117, 2023.

[10] G. Ambrish, B. Ganesh, A. Ganesh, C. Srinivas, K. Mensinkal *et al.*, "Logistic regression technique for prediction of cardiovascular disease," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 127–130, 2022.

[11] E. Watanabe, S. Noyama, K. Kiyono, H. Inoue, H. Atarashi, K. Okumura, T. Yamashita, G. Y. Lip, E. Kodani, and H. Origasa, "Comparison among random forest, logistic regression, and existing clinical risk scores for predicting outcomes in patients with atrial fibrillation: A report from the j-rhythm registry," *Clinical Cardiology*, vol. 44, no. 9, pp. 1305–1315, 2021.

[12] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal*, vol. 20, no. 1, pp. 3–29, 2020.

[13] S. Han, B. D. Williamson, and Y. Fong, "Improving random forest predictions in small datasets from two-phase sampling designs," *BMC medical informatics and decision making*, vol. 21, pp. 1–9, 2021.

[14] M. Velazquez, Y. Lee, and A. D. N. Initiative, "Random forest model for feature-based alzheimer's disease conversion prediction from early mild cognitive impairment subjects," *Plos one*, vol. 16, no. 4, p. e0244773, 2021.

[15] A. Govindu and S. Palwe, "Early detection of parkinson's disease using machine learning," *Procedia Computer Science*, vol. 218, pp. 249–261, 2023.

[16] H. Guo, H. Nguyen, D.-A. Vu, and X.-N. Bui, "Forecasting mining capital cost for open-pit mining projects based on artificial neural network approach," *Resources Policy*, vol. 74, p. 101474, 2021.
[17] W. Wijiyanto, A. I. Pradana, S. Sopingi, and V. Atina, "Teknik k-fold cross validation untuk mengevaluasi kinerja mahasiswa," *Jurnal Algoritma*, vol. 21, no. 1, pp. 239–248, 2024.
[18] P. Handayani, A. C. Fauzan, and H. Harliana, "Machine learning klasifikasi status gizi balita menggunakan algoritma random forest," *KLIK Kaji. Ilm. Inform. dan Komput*, vol. 4, no. 6, pp. 3064–3072, 2024.
[19] I. Rahmadiani, A. I. Fibriana, and M. Azam, "Low birth weight is related to stunting incidents: Indonesian nutrition status survey data analysis," *medRxiv*, pp. 2024–06, 2024.
[20] F. ZUBEDI, B. SARTONO, and K. A. NOTODIPUTRO, "Implementation of winsorizing and random oversampling on data containing outliers and unbalanced data with the random forest classification method," *Jurnal Natural*, vol. 22, no. 2, pp. 108–116, 2022.
[21] O. N. Chilyabanyama, R. Chilengi, M. Simuyandi, C. C. Chisenga, M. Chirwa, K. Hamusonde, R. K. Saroj, N. T. Iqbal, I. Ngaruye, and S. Bosomprah, "Performance of machine learning classifiers in classifying stunting among under-five children in zambia," *Children*, vol. 9, no. 7, p. 1082, 2022.
[22] M. Ohyver, J. V. Moniaga, K. R. Yunidwi, and M. I. Setiawan, "Logistic regression and growth charts to determine children nutritional and stunting status: a review," *Procedia computer science*, vol. 116, pp. 232–241, 2017.
[23] R. R. R. Arisandi, B. Warsito, and A. R. Hakim, "Aplikasi naïve bayes classifier (nbc) pada klasifikasi status gizi balita stunting dengan pengujian k-fold cross validation," *Jurnal Gaussian*, vol. 11, no. 1, pp. 130–139, 2022.