Scientific Journal of Informatics
Vol. 13, No. 2, May 2026

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Predicting Stunting Risk in Toddlers Using Interpretable Machine Learning on Imbalanced Data with Random Forest

Utami[1], Amir Ali[2*], Eka Wilda Faida[3], Titik Kuntari[4]
[1, 4]Master of Public Health Study Program, Universitas Islam Indonesia, Indonesia
[2, 3]Department of Medical Record and Health Information, StikesYayasan RS Dr. Soetomo Surabaya, Indonesia

Abstract.
Purpose: The prevalence of stunting in toddlers is still significant at over 20%, according to the Indonesian Nutritional Status Survey (SSGI). By creating machine learning based early prediction models especially made to handle unbalanced datasets, this project seeks to hasten the elimination of stunting.
Methods: Three hybrid data balancing methods SMOTE, SMOTE + Tomek Links, and SMOTE + ENN were used in conjunction with a Random Forest algorithm. An anthropometry dataset comprising 40.444 toddlers was used to train and evaluate the models. Data cleaning, labeling, and imbalance handling comprised preprocessing. Accuracy, precision, recall, F1-score, and specificity metrics produced from a confusion matrix were then used for evaluation.
Findings: Four key features significantly influence classification: ZSTB/U (0.698961), Height (0.116384), Weight (0.102773), and LiLA (0.035833). The Random Forest algorithm, paired with SMOTE-based techniques, achieved near-perfect performance across all metrics (approaching 1.0). This demonstrates excellent capability in accurately distinguishing the nutritional status of toddlers.
Originality: This study offers a scientific explanation of the main stunting variables as well as a high-performance classification methodology. It offers a strong framework for early stunting diagnosis in Indonesia by successfully correcting class imbalance through SMOTE, SMOTE-Tomek, and SMOTE-ENN.

Keywords: Prevalence of Stunting, SMOTE, SMOTE-Tomek, SMOTE-ENN, Prediction models
Received February 2026 / **Revised** April 2026 / **Accepted** June 2026

INTRODUCTION

In Indonesia, stunting is still a serious public health issue. The frequency of stunting among toddlers was 27.7% in 2019, 24.4% in 2021, and 21.6% in 2022, according to the Indonesian Nutritional Status Survey (SSGI). This number is still significantly higher than the WHO's 20% criterion for public health issues, but declining.

Stunting affects future generations' productivity, cognitive development, and physical growth, endangering the caliber of Indonesia's people resources. For instance, as seen in figure 1 below, the stunting rate in Sidoarjo Regency rose from 14.8% to 16.1% in the 2022 Indonesian Nutritional Status Survey (SSGI).

*Corresponding author.
Email addresses: amir_ali@stikes-yrsds.ac.id (Ali)
DOI: 10.15294/sji.v13i2.42131

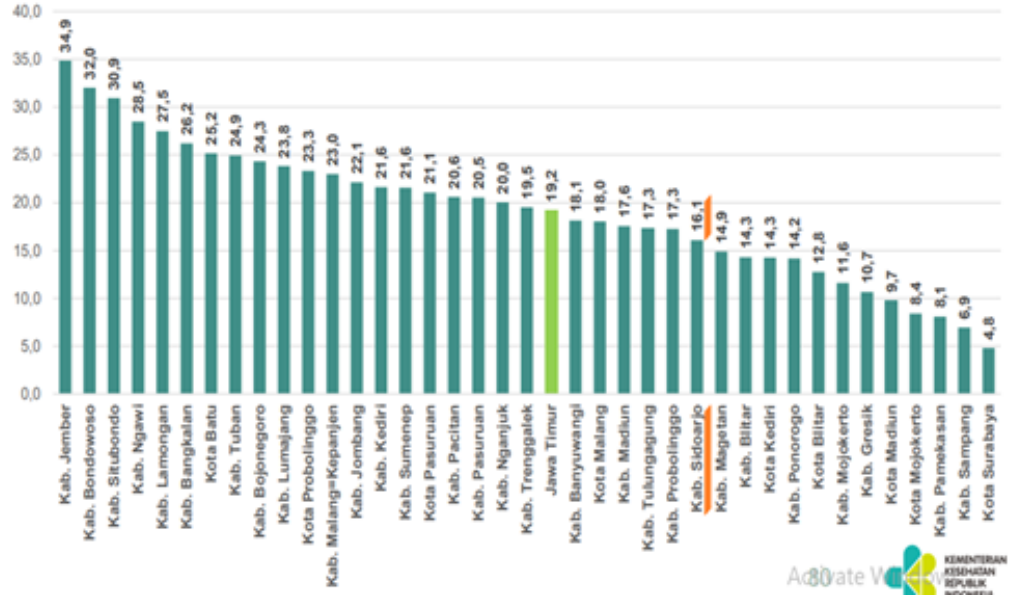

Figure 1. Prevalence of Stunting in Toddlers in East Java based on SSGI
(Source: 2022 Indonesian Nutritional Status Survey)

Furthermore, according to the 2022 Sidoarjo Regency health profile book, out of 85.114 children whose height was measured, 4.905 were short (5.8%). This suggests that stunting issues are still present at the district level.

Stunting is known to be caused by a number of variables, including feeding habits, infectious illnesses, birth weight, maternal nutritional status, and environmental and socioeconomic factors. To determine the main causes of stunting, data analytics is required. By 2024, the government hopes to reduce the frequency of stunting by 14% through Presidential Regulation No. 72/2021. However, without data-driven innovation, this accomplishment is still challenging.

Reducing the prevalence of stunting is one of the national measures in compliance with Presidential Regulation No. 72 of 2021's mandate for the acceleration of stunting reduction. Strengthening and expanding systems, data, information, research, and innovation through the creation of stunting prediction models is required to implement this national goal. This is anticipated to be a successful step in lowering the prevalence of stunting.

Several studies related to stunted toddlers using machine learning are as follows from a study by Qadrini [1] titled "Undersampling and K-Fold Random Forest for Imbalanced Class Classification" focused on improving recall accuracy values by implementing an undersampling resampling technique. The findings demonstrated that undersampling yielded superior accuracy compared to oversampling, achieving a recall accuracy of 83% for the K-Fold 10 Random Forest algorithm and 65% for the undersampling Random Forest. Ariyadi [2] conducted research on the "*Klasifikasi Balita Stunting Menggunakan Random Forest Classifier di Kabupaten Blitar*." While this study shared the goal of classifying toddler stunting status and measuring confusion matrix values, it differed by relying solely on standard confusion matrix calculations for its metrics. The model produced an accuracy of 90%, a precision of 71.4%, and a recall of 62.5%, leading to the conclusion that the model still required improvement for predicting toddler stunting cases effectively. In another geographical focus, Haris [3] explored the "Prediction of Stunting Prevalence in East Java Province with Random Forest Algorithm." Unlike other studies that focused on individual classification, this research aimed to predict overall prevalence values and identify contributing variables based on correlation. Out of 20 candidate factors analyzed, the study successfully isolated 12 suspected causes of stunting. Finally, Perdana [4] investigated "Prediksi Stunting Pada Balita dengan Algoritma Random Forest," sharing the core objective of using Random Forest to predict stunting in young children. A key differentiator in this research was the evaluation of model performance specifically through K-Fold cross-validation. The testing revealed a stable accuracy evaluation in the 97% range, concluding that the Random Forest algorithm is a highly capable tool for predicting stunting conditions in toddlers.

Based on the analysis of previous studies, several key research gaps emerge regarding resampling techniques, validation methods, and location coverage. First, regarding the resampling technique gap & data imbalance, study (1) focused on undersampling but still yielded a recall value between 65% and 83%, while study (2) achieved a high accuracy of 90% but a low recall of 62.5%, indicating that the standard Random Forest model still struggles to recognize the minority "Stunting" class compared to the majority "Normal" class. Second, in terms of the validation & generalization gap, study (2) is at risk of overfitting as it only uses a regular Confusion Matrix without Cross-Validation, whereas study (4) implemented K-Fold Cross-Validation with 97% accuracy but failed to mention how the model handles imbalanced data. Lastly, concerning the location coverage gap & factors, study (3) conducted predictions at the provincial level (East Java) and identified causal factors, but it did not integrate in-depth model optimization techniques such as resampling or hyperparameter tuning to enhance specific prediction performance for individual toddlers.

The use of SMOTE (Oversampling), SMOTE-Tomek Links, and SMOTE-ENN in the Random Forest algorithm to address the imbalance of stunted toddler data is what makes this research novel. Furthermore, the combination of interpretative and predictive methods, in which machine learning models are employed not only to categorize nutritional status but also to identify the main causes of stunting incidence.

Our observations indicate that while the nutrition monitoring system (E-PPGBM) is only used for recording and reporting, the Health Office employs the E-PPGBM (Community-Based Electronic Nutrition Recording and Reporting) information system. Health cadres, community health centers, and local district health offices are among the relevant parties who can use the gathered data as a dataset to develop a stunting prediction model for early child stunting detection. It's crucial to remember that there are fewer datasets of stunted toddlers than there are of normal toddlers. The distribution of the data becomes unbalanced as a result.

Automated analysis techniques for predicting stunting in toddlers are complicated by this unbalanced data distribution, where the number of normal toddlers (the majority class) is significantly higher than the number of stunted toddlers (the minority class). As a result, stunting cases (the minority class) frequently go undetected. As a result, toddlers who are not identified as stunted do not receive the necessary therapies. Unbalanced data is a significant issue in health prediction. When it comes to stunting, the model tends to favor the majority class because there are far more normal toddlers than stunted toddlers. A number of algorithms can be categorized into three groups in order to address class imbalance. The first is data-level methods that use oversampling and undersampling algorithms to try to balance data distribution. The second is algorithm-level approaches, which modify existing algorithms to account for the significance of minor classes or by developing new algorithms. The third is a combination of algorithmic and data-level approaches [1]. An unbalanced data set between the majority and minority classes will affect the classification results. While the accuracy may be very high, the prediction results will be inaccurate if the minority class is predicted to be small. Therefore, data balancing is necessary using a resampling technique, where the data with the largest number of classes is adjusted to the number of minority classes. To overcome this, a resampling technique called Synthetic Minority Oversampling Technique (SMOTE) is used. SMOTE is performed by generating more synthetic data for the minority class from their k nearest neighbors. To do this, the minority class data is taken from their k nearest neighbors and synthetic samples are added to the line connecting one or all of the k nearest neighbors of the minority class, and the selection of the k nearest neighbors is done randomly [5]. Other data balancing techniques include SMOTE + Tomek Links and SMOTE + ENN (Edited Nearest Neighbors). These techniques aim to improve recall and F1-score, making the model more capable of detecting stunting cases (minority class).

Several previous studies have applied the Random Forest algorithm to analyze stunting cases, each with its own specific focus and limitations. A study by Lestanti et al. (2023) focused on the classification of stunted toddlers in Blitar Regency, but the resulting recall was still low at 62.5%, indicating that the model was not yet optimal in recognizing stunting cases. In contrast, Yudha et al. (2021) predicted stunting using Random Forest combined with K-Fold Cross-Validation and achieved a high accuracy of 97%; however, this research did not address the dominant risk factors or handle data imbalance. Meanwhile, Haris et al. (2023) shifted the scope to a broader level by predicting stunting prevalence in East Java; although their study successfully identified various risk factors, the analysis remained limited to correlation values and did not incorporate a data-balanced analysis [3].

This research attempts to answer the following research questions that's are:

1. How can hybrid sampling techniques such as SMOTE + Tomek Links or SMOTE + ENN improve recall and F1-score in identifying stunting cases?
2. How does the Random Forest algorithm perform in predicting stunting status on unbalanced data?
3. What factors in toddler anthropometric data contribute most to stunting status based on the interpretation of the Random Forest model?

METHODS

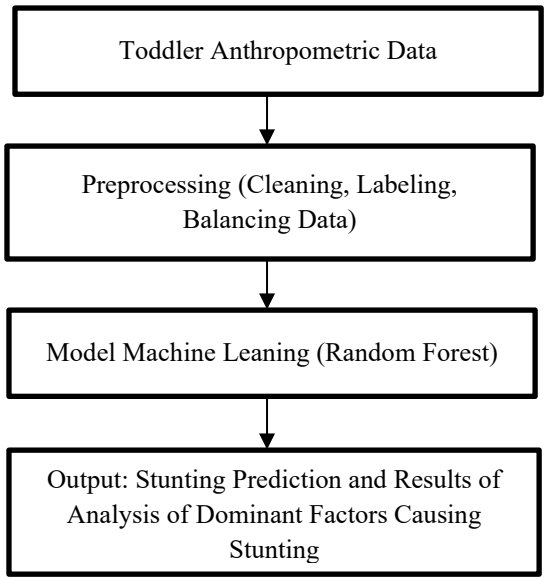

Figure 2. Research scheme

Toddler antropometric data

This study utilized data on stunted and normal toddlers collected from the Sidoarjo Regency Health Office through the E-PPGBM (Community-Based Electronic Nutrition Recording and Reporting) information system application. The data used totaled 6,883, representing toddlers born between 2020 and 2025 across eight sub-districts in Sidoarjo Regency. The data on stunted and normal toddlers obtained from the Sidoarjo Regency Health Office contains 34 important and detailed attributes. These attributes include Toddler's National Identification Number (NIK), Toddler's Name, Gender, Date of Birth, BBBirth, TBBirth, Parent's Name, Province, Regency/City, Sub-district, Community Health Center, Village/Sub-district, Integrated Health Post, RT, RW, Address, Age at Measurement, Date of Measurement, Weight, Height, Measurement Method, Upper Arm Circumference (LILA), Weight per Age (BB/U), Z-Score Weight per Age (ZSBB/U), Height per Age (TB/U), Z-Score Height per Age (ZSTB/U), Weight per Height (BB/TB), Z-Score Weight per Height (ZSBB/TB), Weight Gain, Additional Food Provision (PMT) received in (Kg), Amount of Vitamin A received (Amount of Vitamin A), Daily Care Practice Questionnaire (KPS A), Mother and Child Card (KI A).

Preprocessing (cleaning, labeling, balancing data)

The above research begins with dataset preparation. The parameters in the dataset are toddler anthropometric data (gender, date of birth, birth weight, birth height, age, height, weight, Lila, height for age, Z-score height for age). Influential parameters will be the features used in the classification process for stunted toddlers. If there are empty data or missing values, they can be filled using an imputation algorithm. Data cleaning and labeling are also necessary. In addition, an unbalanced amount of data between the majority and minority classes will affect the classification results. The accuracy results obtained may be very high, but the prediction results will be incorrect if predicting a class with a small number (minority). Therefore, it is necessary to balance the data using a resampling technique where the data with the largest number of classes will be adjusted to the number of minority classes. This data balancing technique uses the SMOTE method.

Model machine learning with random forest

A branch of artificial intelligence called machine learning is extensively studied and applied to a variety of issues. Three machine learning categories—supervised learning, unsupervised learning, and reinforcement learning—are used to present reviews from diverse domains in the form of algorithms and problem-solving strategies[6]. The technique used by Supervised Learning is a classification method in which a dataset is fully labeled to classify unknown classes. Unsupervised Learning, on the other hand, is often called clustering because there is no need for labeling in the dataset and the results do not identify examples in predetermined classes [7]. Meanwhile, Reinforcement Learning is usually between Supervised Learning and Unsupervised Learning[8]. Specifically, reinforcement learning methods based on Markov decision-making processes (MDPs) include two types. First, model-based techniques like the SARSA algorithm, where reinforcement learning learns the model's knowledge before using it to choose the best course of action. Second, model-relevant techniques like the Q-learning algorithm and the Temporal Difference algorithm, which use reinforcement learning to determine the best course of action without the need for model knowledge[9]. One type of machine learning modeling uses the Random Forest algorithm. The Random Forest algorithm was chosen because of its advantages:

- It can handle data with many variables,
- It provides interpretation of risk factors through feature importance,
- It is more stable than a single decision tree.

A variation of the Decision Tree method, Random Forest employs several Decision Trees. Each Decision Tree is trained using separate samples, and each attribute is divided into trees chosen at random from a subset of attributes[10]. Among the many benefits of random forests include increased accuracy in the event of missing data, resistance to outliers, and effective data storage. Additionally, Random Forests improve the performance of classification models by extracting the best features through a feature selection procedure. Random Forests can efficiently handle big data sets with complicated characteristics because to this selection capability[11]. Leo Breiman initially presented Random Forest (RF) in 2001. RF is a technique that can increase the precision of randomly generated attributes for every node. RF is made up of a group of decision trees that are used to categorize data. The procedure of creating a decision tree using the RF approach is comparable to that of the Classification and Regression Tree (CART), with the exception that RF does not use pruning[12].

The Gini index is used to select features at each internal node of a decision tree. The Gini index value can be calculated as follows:

$$Gini\ (S_i) = 1 - \sum_{i=0}^{c-1} p_i^2 \tag{1}$$

With p_i being the relative frequency of class C_i in the set C_i being n classes for i = 1, ..., c-1, and c being the number of classes specified. The split quality of feature k into subsets S_i is the number of samples belonging to class C_i, then calculated as the sum of the Gini indices of the resulting subsets. The data can be calculated using the following formula:

$$Gini_{Split} = \sum_{i=0}^{k-1} \left(\frac{n_i}{n}\right)\ Gini\ (S_i) \tag{2}$$

Where n_i is the number of samples in subset S_i after splitting and n is the number of samples at a given node.

Data sharing activities were used to conduct this study, and the data was separated into training and test sets. Machine learning utilizing the random forest algorithm was applied to training and test data. 20% of the data was used for testing, while the remaining 80% was used for training. The child anthropometric data will be processed and examined using Python programming and Google Collaborations tools to identify any patterns or trends. A stunting prediction model is created by this modeling. In order to accommodate empty data (missing values), data preprocessing included imputation techniques. Additionally, a resampling strategy was utilized to balance the data, adjusting the data with the greatest number of classes to the number of minority classes in cases of stunted children. The SMOTE (Synthetic Minority Over-sampling Technique) approach was utilized to balance the dataset between majority and minority data, and the random forest algorithm was then applied. An evaluation matrix was used to assess the model's performance in order to assess the modeling findings.

Model performance evaluation

An evaluation matrix was used to evaluate the stunting prediction model's performance. Model performance can be evaluated using the idea of an evaluation matrix. Additionally computed were precision, accuracy, recall, and F1 score. In order to optimize prediction accuracy in this study, imputation techniques were used to address missing data, and resampling techniques were used for data balancing. In the case of stunted toddlers, the data with the greatest number of classes will be adjusted to the number of minority classes. The SMOTE (Synthetic Minority Oversampling methodology) method was employed in this data balancing methodology, and a random forest algorithm was subsequently applied for processing. A model performance evaluation matrix was used to assess the modeling findings.

The E-PPGBM (Community-Based Electronic Nutrition Recording and Reporting) system provided secondary data for this quantitative observational study. In order to solve data discrepancies that previously required manual computations, the research concentrated on creating a machine learning-based stunting risk prediction model (Random Forest).[13].

Toddler anthropometric data from the Sidoarjo Regency Health Office's E-PPGBM program is used in this study framework. Preprocessing (cleaning, labeling, and balancing) was done on the data. The random forest approach was then used to process it using a machine learning model.A stunting prediction model and an examination of the primary causes of stunting based on toddler anthropometric data are the results of this study.

Using Python tools and a machine learning algorithm technique, an intelligent system model is created as part of this study method's modeling process. The random forest algorithm is the machine learning algorithm that is employed. The random forest technique will be used to assess the anthropometric data of the input toddler. Trends in toddler anthropometric data patterns can serve as a foundation for decision-making by pertinent parties, in this case the health office and associated community health centers, to intervene with patients in order to suppress the stunting issue going forward and lower the number of stunted toddlers in the area.

Data sharing was used in this study, and the data was separated into test and training sets. The random forest technique was used in machine learning to process the training and test data. 80% of the data was used for training, and the remaining 20% was used for testing. This anthropometric data will be processed and analyzed to find data trends in Sidoarjo Regency using Google Collaboration tools and Python programming. The anthropometric measurement data of toddlers will also be examined for data pattern patterns. A stunting prediction model is the result of this modeling.

RESULT AND DISCUSSION

Result

Identifying the dominant factors causing stunting from toddler anthropometric data

This study used data on stunted and normal toddlers obtained from the Sidoarjo Regency Health Office, along with representatives from the Balongbendo, Buduran, Candi, Gedangan, Jabon, Krian, Porong, Waru, and Barengkrajan health centers. The data covered 40,444 toddlers. The attributes used to predict stunting were gender, birth weight, birth height, age, weight, height, limb height, height for age, and height for age.

Table 1. Example of toddler anthropometric data

No	Gender	Weight Born	Height Born	Age	Weight	Height	Upper Arm Circumference	Height for Age	Z-Score height according to age
1	L	2.9	0	5.04	20.8	116.2	NaN	Normal	1.28
2	P	3.1	49	5.04	20.8	116.2	NaN	Normal	1.37
3	P	34	50	5.04	20.8	116.2	NaN	Normal	1.37
4	P	3	50	4.75	20.6	115.2	NaN	Normal	1.59
5	P	3	50	4.93	20.9	115.7	NaN	Normal	1.43
6	P	0	0	5.01	21	116.2	NaN	Normal	1.41
7	L	3.2	50	4.63	16	106	14	Normal	-0.34
8	P	2.8	47	4.93	17.5	109	15	Normal	0
9	P	2.7	49	4.78	23.5	114	17	Normal	1.28

To identify the dominant contributing factors to stunting in toddler anthropometric data, we used the importance feature of the random forest algorithm. The weighted values for each feature in this toddler anthropometric data are presented in Table 2 below.

Table 2. Feature weighting values in anthropometric data of stunted and normal toddlers in sidoarjo

Features	Weighting Values
ZSTB/U	0.698961
Height	0.116384
Weight	0.102773
Upper Arm Circumference	0.035833
Age	0.029897
Birth Height	0.008950
Birth Weight	0.005497
Gender	0.001704

The weighting values in Table 2 above were calculated using a random forest algorithm to weight the importance features. Based on the data in Table 2 above, it can be seen that the four features that have a widespread influence on the classification of stunted toddlers are ZSTB/U, Height and Weight, and Upper Arm Circumference. These features have weight values of 0.698961, 0.116384, 0.102773, and 0.035833, indicating that these features provide the greatest contribution to the classification model.

Data imbalance problem

The Random Forest algorithm was used to test the classification model using three methods for addressing data imbalance: SMOTE, SMOTE–Tomek, and SMOTE–ENN. Accuracy, Precision, Recall, and F1-score metrics—all pertinent to the classification of child nutritional status (H/U) with an uneven class distribution—were used to evaluate the model's performance.

The categorization results will be impacted if there is an imbalance in the data between the majority class of normal toddlers and the minority class of stunted toddlers. Predictions will be erroneous if the minority class is small, even though the accuracy may be extremely high. As a result, data balancing utilizing a resampling technique—in which the data with the greatest number of classes is modified to reflect the minority class—is required. The Synthetic Minority Oversampling Technique (SMOTE) is a resampling method used to solve this problem.

A statistical technique called SMOTE (Synthetic Minority Over-sampling Technique) is used to artificially enhance the number of samples in minority classes. In contrast to conventional oversampling techniques, which merely replicate preexisting data (which frequently results in overfitting), SMOTE generates new data that is comparable to but distinct from the original data.

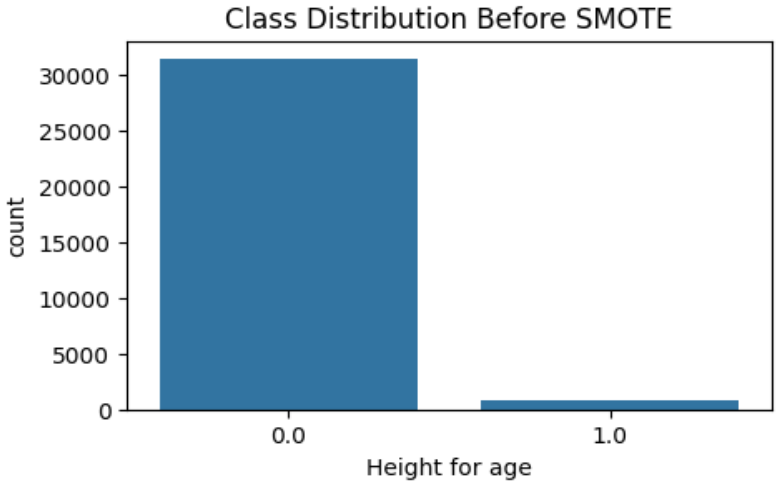

Figure 3. Class distribution before SMOTE

According to Figure 3, from the data the distribution of the number of stunted toddler data is 870 toddlers, while there are 31.485 normal toddlers.

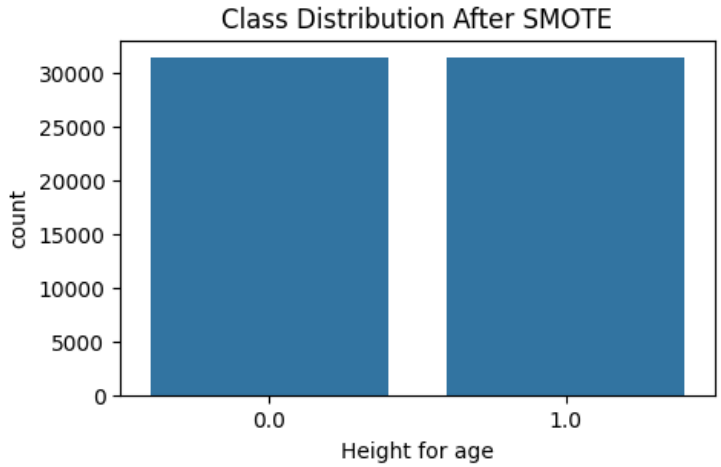

Figure 4. Class distribution after SMOTE

According to Figure 4, the distribution of the number of data for stunted toddlers and normal toddlers is 31.485 toddlers.

Apart from that, there are other data balancing techniques, namely SMOTE-TOMEK Links and SMOTE-ENN. SMOTE-Tomek Links is a hybrid method that combines oversampling (data addition) and undersampling (data cleaning) techniques. This method is very effective for handling imbalanced datasets with large amounts of overlapping data.

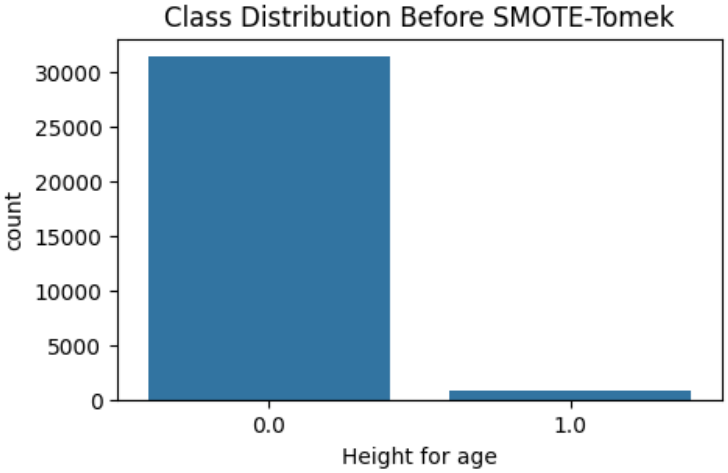

Figure 5. Class distribution before SMOTE-Tomek

According to Figure 5, from the data the distribution of the number of stunted toddler data is 870 toddlers, while there are 31.485 normal toddlers.

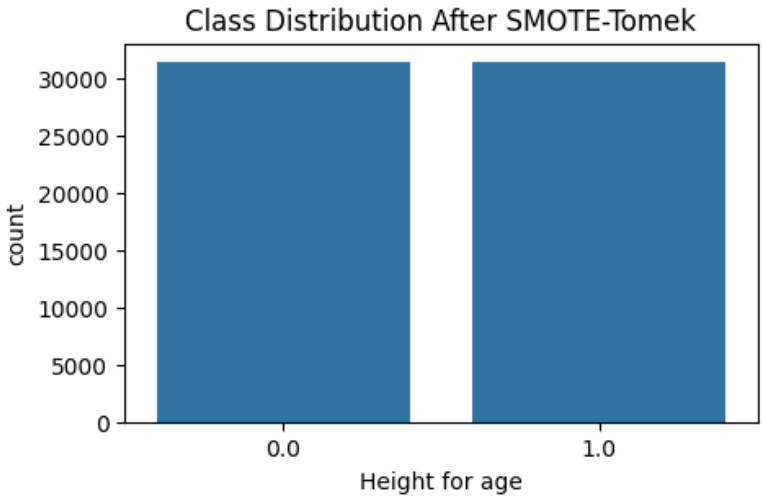

Figure 6. Class distribution after SMOTE-Tomek

According to Figure 6, the distribution of the number of data for stunted toddlers and normal toddlers is 31.470 toddlers.

SMOTE-ENN (Synthetic Minority Oversampling Technique - Edited Nearest Neighbors) is a hybrid method that combines oversampling (data augmentation) and undersampling (aggressive data cleaning).

While SMOTE-Tomek is considered a "gentle" cleaner, SMOTE-ENN is a "robust" cleaner that is highly effective on noisy datasets with significant overlap between classes.

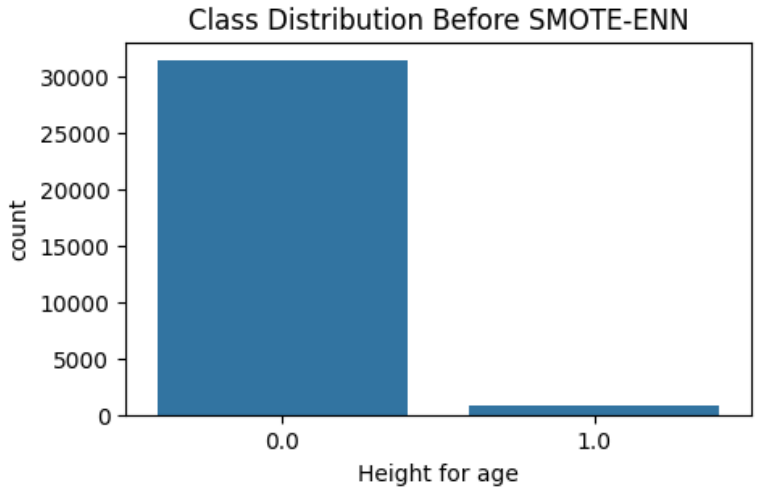

Figure 7. Class distribution before SMOTE-ENN

According to Figure 7, from the data the distribution of the number of stunted toddler data is 870 toddlers, while there are 31.485 normal toddlers.

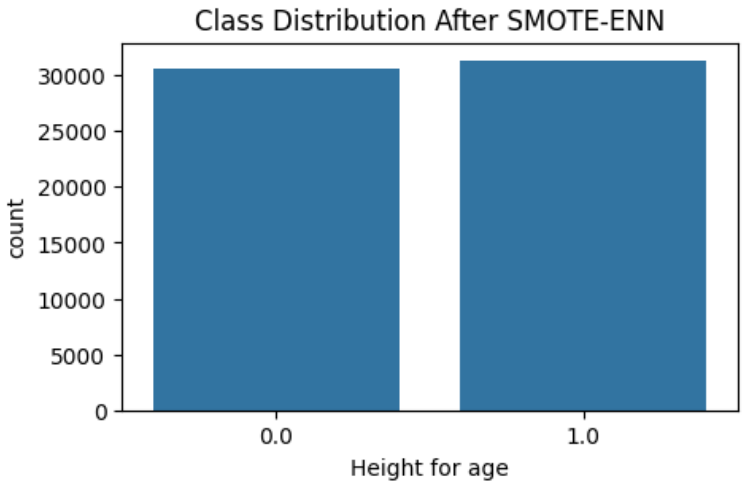

Figure 8. Class distribution after SMOTE-ENN

According to Figure 8, from the data the distribution of the number of stunted toddler data is 31,334 toddlers, while there are 30,579 normal toddlers

Building a stunning status prediction model

To build a stunting prediction model, data preprocessing is necessary, consisting of data cleaning and data transformation. Data cleaning is performed to ensure there is no missing or duplicate data. If missing data is found, it can be filled in using imputation techniques, one of which is the K-Nearest Neighbors (K-NN) imputation technique.

Table 3. Toddler anthropometric dataset before transformation and imputation

Gender	Birth weight	Birth height	Age at measurement	Heavy	Tall	Upper arm circumference	Height for age	Zscore Height for Age
L	2.9	0	5 Tahun - 0 Bulan - 15 Hari	20.8	116	NaN	Normal	1.28
P	3.1	49	5 Tahun - 0 Bulan - 14 Hari	20.8	116	NaN	Normal	1.37
P	34	50	5 Tahun - 0 Bulan - 13 Hari	20.8	116	NaN	Normal	1.37
P	3	50	4 Tahun - 9 Bulan - 0 Hari	20.6	115	NaN	Normal	1.59
P	3	50	4 Tahun - 11 Bulan - 4 Hari	20.9	116	NaN	Normal	1.43
P	0	0	5 Tahun - 0 Bulan - 5 Hari	21	116	NaN	Normal	1.41
P	3	50	4 Tahun - 9 Bulan - 2 Hari	20.6	115	NaN	Normal	1.58

As seen in table 3 above, there is empty data for the upper arm circumference column marked with the word NaN. This data needs to be preprocessed by filling it using the K-Nearest Neighbors (K-NN) imputation technique where the data will be filled by the value of the k nearest neighbors (k-nearest neighbors). The closeness is calculated from the distance of other non-empty features, where the imputed value = the average (mean) of the nearest neighbors.

As it can also be seen in figure 2 above, there are categorical attributes that need to be converted into numeric form to make them easier to process by machine learning algorithms. The label encoding method can be used to convert categorical values, such as the gender columns 'L' and 'P', into numbers, as in the gender and Height for Age columns.

In this study, the attributes that underwent transformation included gender and height for age. For the gender column, gender L is given a code of 1 while gender P is given a code of 0. For the height for age column, the information data for Normal and tall is given a code of 0 while Short and very short is given a code of 1.

Data division, which separates the data into training and test data, comes after the data transformation and cleaning procedure. The random forest technique was used in machine learning to process the training and test data. Training data made up 80% of the total, while test data made up 20%.

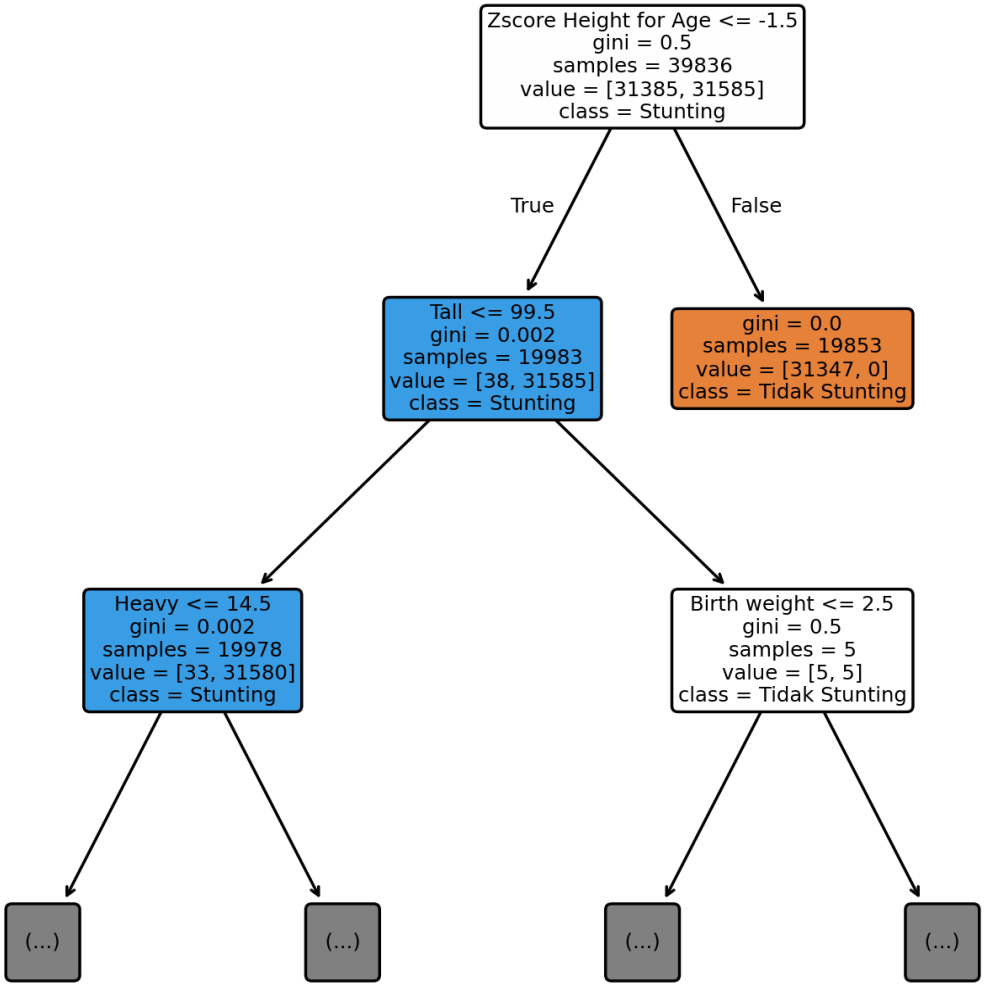

Figure 9. Visualization of stunting prediction modeling results

Based on Figure 9, the decision tree visualization shows that the height-for-age (H/A) Z-score is the most dominant variable in determining stunting status. Children with H/A values ≤ -1.5 are consistently classified as stunted, especially when reinforced by low body weight and small mid-upper arm circumference, which represent chronic malnutrition.

Testing the stunting status prediction model

The results of testing the stunting status prediction model with imbalanced data using SMOTE, SMOTE Tomek and SMOTE ENN are as follows:

Table 4. Comparison of the confusion matrix results of SMOTE, SMOTE Tomek and SMOTE ENN

No	Method	Accuracy	Precision	Recall	F1-Score
1.	SMOTE	0.999011	0.999046	0.999011	0.999020
2.	SMOTE-Tomek	0.998887	0.998932	0.998887	0.998898
3.	SMOTE-ENN	0.998887	0.998932	0.998887	0.998898

Based on Table 4, all methods produce very high performance values, with all metrics approaching the value of 1. This shows that the Random Forest algorithm has very good classification capabilities in distinguishing the nutritional status of toddlers.

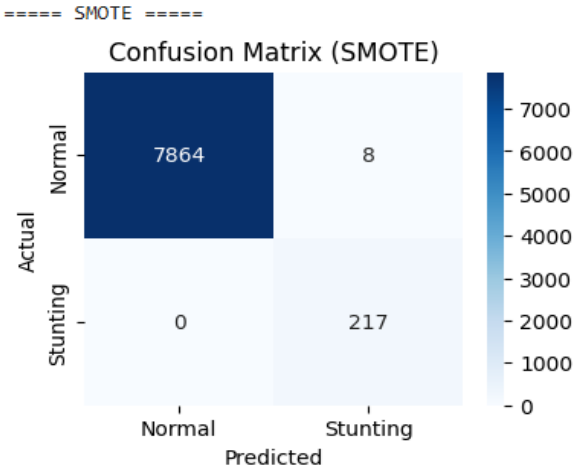

Figure 10. Confusion matrix SMOTE

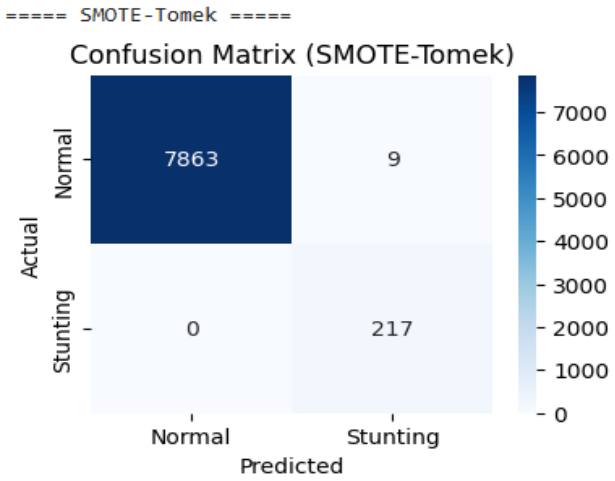

Figure 11. Confusion matrix SMOTE-Tomek

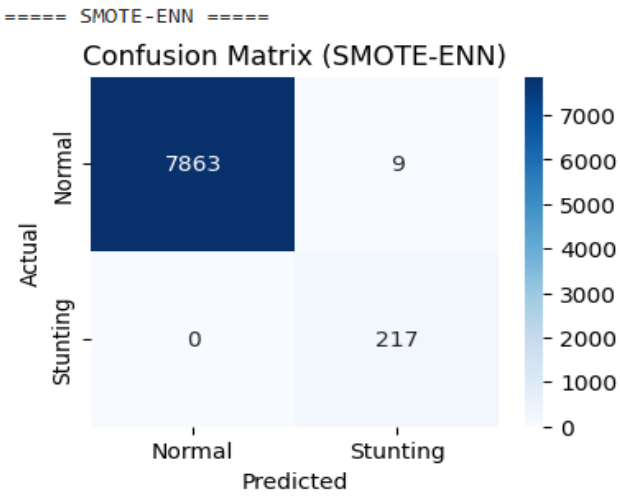

Figure 12. Confusion matrix SMOTE-ENN

Discussion

The study's findings show that SMOTE, SMOTE-Tomek, and SMOTE-ENN applications resulted in comparable classification performance. This result is consistent with the study of Chawla et al[14], which found SMOTE to be effective in increasing minority class representation without eliminating majority class data. SMOTE-Tomek and SMOTE-ENN are developments of SMOTE that combine oversampling with data cleaning using Tomek Links and Edited Nearest Neighbor. According Gustavo et al [15], this hybrid method aims to reduce overlap between classes and remove noise. However, its effectiveness is highly dependent on the characteristics of the dataset used. Research by Fernandez et al [16] demonstrates that performance differences between resampling algorithms are typically negligible in datasets with low noise levels and distinct class separation. This is in line with the study's findings, which showed that the three approaches yielded performance levels that were almost equal. Based on the data in Table 4, it can be seen that All three methods (SMOTE, SMOTE-Tomek, and SMOTE-ENN) produced nearly identical and very high scores, above 0.998. This indicates that your Random Forest model has a very high level of precision and reliability in distinguishing the nutritional status of toddlers. Technically, there are subtle differences between the three methods:

- SMOTE (Basic Method): Gives the marginally highest result (0.999011).
- SMOTE-Tomek & SMOTE-ENN (Hybrid Method): Both produce exactly the same value (0.998887).

The SMOTE-Tomek and SMOTE-ENN techniques are typically used to remove noise at the boundary between the majority and minority classes. Since the results are slightly (but very slightly) lower than those of pure SMOTE, this indicates that your original data likely already has very well-separated features. The selection of SMOTE-Tomek as the best method was based on methodological considerations, as this method not only balances the data but also reduces borderline data to the best selection method in the dataset balancing process. Although the SMOTE method produces slightly higher accuracy values, the performance difference between SMOTE-Tomek and SMOTE-ENN is very small and not practically significant. This approach aligns with the recommendation of Garcia [17] who stated that hybrid resampling methods often provide the best balance between improved performance and model stability, especially in academic research and real-world applications

In addition to evaluating classification performance, this study also conducted a feature importance analysis on the Random Forest model to identify the most influential variables in determining toddler nutritional status based on the Height-for-Age indicator. Feature importance analysis is crucial because it not only assesses prediction accuracy but also provides interpretive insight into the dominant factors causing stunting. The weighting values in Table 2 above were calculated using a random forest algorithm to weight the importance features. Based on the data in Table 2 above, it can be seen that the four features that have a widespread influence on the classification of stunted toddlers are ZSTB/U, Height and Weight, and Upper Arm Circumference. These features have weight values of 0.698961, 0.116384, 0.102773, and 0.035833, indicating that these features provide the greatest contribution to the classification model. To achieve this weighting value, the random forest feature importance is used. This feature is a model for assessing feature ranking that adopts the concept of a decision tree. Each branch of the tree created by this model is analyzed to determine the crucial features in the dataset[18].

An ensemble method called Random Forest is renowned for its strong resistance to unbalanced input. Breiman [19] explains that Random Forest builds multiple decision trees from different subsets of data and features, thereby reducing model bias and variance. Furthermore, research by Chen, Liaw, and Breiman [20] shows that Random Forest is able to maintain stable classification performance despite imbalanced class distributions. This explains why the various data balancing techniques used in this study did not significantly impact model performance. This finding is also supported by García [21] who stated that in many cases, model performance is more influenced by the quality of the features and classification algorithm than by the resampling method used.

In health research, recall metrics play a crucial role because they directly relate to the system's ability to detect all correct cases. Saito and Rehmsmeier[22] emphasized that in classification problems with critical minority classes, recall is more relevant than accuracy alone. The model is able to identify almost all instances of stunting when the recall value is close to 1 across all techniques. This is important since stunting is a long-term dietary issue that can affect cognitive development, physical growth, and the caliber of human resources [23]. Mothers who had stunted growth as children are more likely to give birth to children with

comparable issues since stunting is a cyclical phenomena. This leads to a difficult-to-break intergenerational cycle of poverty and a decrease in human capital[24].

According to Breiman (2001), Random Forest is able to measure the relative contribution of each feature based on the reduction in impurity (Gini Importance) generated during the decision tree formation process. A feature's role in the model's decision-making process increases with its significance value.

The results of the feature importance analysis show that several variables have a dominant contribution compared to others. These features are generally related to:

1. Child characteristics (e.g., age, weight, or growth history)
2. Maternal factors (e.g., maternal education or health status)
3. Environmental factors and health services

These findings align with research by Victora [25]and Black [26], which states that stunting is a multifactorial condition influenced by a combination of individual, family, and environmental factors.

This study's primary benefit is the combination of predictive and interpretive methods, wherein machine learning algorithms are employed to both categorize nutritional status and determine the primary causes of stunting. This approach provides added value compared to previous research, which generally focused solely on model accuracy without interpretability analysis.

Furthermore, feature importance analysis can support data-driven decision-making in public health. Information regarding dominant features can be used by policymakers to:

- Prioritizing nutrition interventions
- Directing stunting prevention programs more precisely
- Supporting evidence-based health policy planning

This approach aligns with WHO recommendations (2018), which emphasize the importance of utilizing data and analytical technology to support stunting reduction programs.

Thus, this study not only produces a highly performing classification model but also provides a scientific contribution in the form of a machine learning-based interpretation of dominant stunting factors, a key novelty in this study. The Random Forest algorithm, which uses SMOTE, SMOTE-Tomek, and SMOTE-ENN, can classify toddlers' nutritional status with very good performance, according to the findings and discussion. Variations in data balancing methods had little effect on model performance, suggesting that the Random Forest algorithm's stability and feature quality are more important.

CONCLUSION

The implementation of data balancing approaches can successfully address the issue of class imbalance in toddler nutritional status data, enabling the successful detection of stunting cases as a minority class, according to the research findings. The high recall value, which shows a low number of stunting cases incorrectly labeled as normal, illustrates this. The Random Forest algorithm-based stunting status prediction model performed exceptionally well, with almost flawless F1-score, recall, accuracy, and precision scores. This shows that Random Forest is a dependable and efficient algorithm for categorizing toddlers' stunting status based on anthropometric information. Through interpretive analysis of the Random Forest model, this study not only achieved good predictive performance but also successfully identified major elements contributing to stunting. In order to assist more focused decision-making and nutritional intervention planning, the interpretation of feature importance offers a deeper understanding of anthropometric characteristics and related factors that significantly influence stunting status. Other classification methods, like as the SVM algorithm, KNN, and logistic regression, can be utilized in later studies to forecast toddler stunting instances. When a dataset is unbalanced, it can be balanced using the Smote, Smote Tomek, and Smote ENN procedures.

REFERENCES

[1] L. Qadrini, "Undersampling dan K-Fold Random Forest Untuk Klasifikasi Kelas Tidak Seimbang," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 4, Mar. 2023, doi: 10.47065/bits.v4i4.3141.

[2] M. R. Akbar Ariyadi, S. Lestanti, and S. Kirom, "Klasifikasi Balita Stunting Menggunakan Random Forest Classifier di Kabupaten Blitar," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 6, pp. 3846–3851, Jan. 2024, doi: 10.36040/jati.v7i6.7822.

[3] M. S. Haris, M. Anshori, and A. N. Khudori, "Prediction of Stunting Prevalence in East Java Province with Random Forest Algorithm," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 1, pp. 11–13, Feb. 2023, doi: 10.52436/1.jutif.2023.4.1.614.

[4] Aditya Yudha Perdana, "Prediksi Stunting Pada Balita dengan Algoritma Random Forest," Universitas Telkom, Bandung, 2021.

[5] N. P. Y. T. Wijayanti, E. N. Kencana, and I. W. Sumarjaya, "Smote: Potensi Dan Kekurangannya Pada Survei," *E-Jurnal Matematika*, vol. 10, no. 4, p. 235, 2021, doi: 10.24843/mtk.2021.v10.i04.p348.

[6] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian Journal on Computer and Information Technology)*, vol. 5, no. 1, pp. 75–82, 2020, doi: 10.31294/ijcit.v5i1.7951.

[7] R. Thupae, B. Isong, N. Gasela, and A. M. Abu-Mahfouz, "Machine learning techniques for traffic identification and classifiacation in SDWSN: A survey," *Proceedings: IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, pp. 4645–4650, 2018, doi: 10.1109/IECON.2018.8591178.

[8] Financial Stability Board, "Artificial Intelligence and Machine Learning in Financial Services - Market Developments and Financial Stability Implications," *Financial Stability Board*, no. November, p. 45, 2017.

[9] Q. Wang and Z. Zhan, "Reinforcement learning model, algorithms and its application," *Proceedings 2011 International Conference on Mechatronic Science, Electric Engineering and Computer, MEC 2011*, no. 1, pp. 1143–1146, 2011, doi: 10.1109/MEC.2011.6025669.

[10] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis : Jurnal Ilmiah Ekonomi dan Bisnis*, vol. 13, no. 2, pp. 67–75, 2020, doi: 10.51903/e-bisnis.v13i2.247.

[11] S. Devella, Y. Yohannes, and F. N. Rahmawati, "Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 7, no. 2, pp. 310–320, 2020, doi: 10.35957/jatisi.v7i2.289.

[12] Suci Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, vol. 4, no. 3, pp. 121–127, 2022, doi: 10.35580/variansiunm31.

[13] D. Yudha, Adiyta Perdana. Latuconsina, Roswan. Ashri, "Prediksi stunting pada balita dengan algoritma random forest," *e-Proceeding of Engineering*, vol. 8, 2021.

[14] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique," vol. 16, pp. 321–357, 2002.

[15] Ronaldo. Monard. Maria Carolena, Gustavo E. A. P. A. Prati, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data".

[16] A. Fernández, S. García, M. Galar, and R. C. Prati, *Learning from Imbalanced Data Sets*. 2018.

[17] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE*, vol. 21, no. 9, pp. 1263–1284, 2009.

[18] L. D. Yulianto, E. H. Hermaliani, and L. Kurniawati, "Penerapan Machine Learning Dalam Analisis Stadium Penyakit Hati Untuk Proses Diagnosis dan Perawatan," *Media Online*, vol. 3, no. 4, pp. 170–180, 2023.

[19] L. Breiman, "Random Forest," *Kluwer Academic Publishers*, pp. 5–13, 2001, doi: 10.1007/978-3-030-62008-0_35.

[20] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," 2004.

[21] V. García, J. S. Sánchez, and R. A. Mollineda, "Knowledge-Based Systems On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," vol. 25, pp. 13–21, 2012, doi: 10.1016/j.knosys.2011.06.013.

[22] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, pp. 1–21, 2015, doi: 10.1371/journal.pone.0118432.
[23] A. Soliman *et al.*, "Early and Long-term Consequences of Nutritional Stunting : From Childhood to Adulthood," *Acta Biomed*, vol. 92, no. 4, pp. 1–12, 2021, doi: 10.23750/abm.v92i1.11346.
[24] A. J. P. & J. H. Humphrey, "The stunting syndrome in developing countries," *Taylor & Francis*, vol. 9047, 2014, doi: 10.1179/2046905514Y.0000000158.
[25] C. G. Victora *et al.*, "Maternal and child undernutrition : consequences for adult health and human capital," *Lancet*, vol. 371, pp. 340–357, 2008, doi: 10.1016/S0140-6736(07)61692-4.
[26] R. E. Black *et al.*, "Maternal and child undernutrition and overweight in low-income and middle-income countries," *Lancet*, 2011, doi: 10.1016/S0140-6736(13)60937-X.