DOI: https://doi.org/10.52436/1.jutif.2024.5.6.4162
p-ISSN: 2723-3863
e-ISSN: 2723-3871

CLASSIFICATION OF TODDLER NUTRITIONAL STATUS USING SUPPORT VECTOR MACHINE AND RANDOM FOREST TECHNIQUES WITH OPTIMAL FEATURE SELECTION

Femmi Widyawati[1], Hanif Pandu Suhito[2], Warusia Yassin[3], Heru Agus Santoso[*4]

[1,4]Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia
[2]Semarang City Health Office, Indonesia
[3]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia
Email: [1]111202113487@mhs.dinus.ac.id, [2]mashanifps@gmail.com, [3]s.m.warusia@utem.edu.my,
[4]heru.agus.santoso@dsn.dinus.ac.id

(Article received: December 08, 2024; Revision: December 31, 2024; published: January 04, 2024)

Abstract

Nutritional problems in toddlers, such as stunting, wasting, being underweight, and obesity, are major challenges in monitoring toddler health in Indonesia because they can hurt toddler growth and development. Therefore, handling nutritional problems comprehensively, including prevention efforts and appropriate dietary interventions, is very important. This study aims to develop a toddler nutritional status classification model based on machine learning algorithms, namely Support Vector Machine (SVM) and Random Forest, by utilizing a toddler dataset obtained from Health Institutions in Indonesia containing 9,735 data. The model was designed using the Recursive Feature Elimination (RFE) technique for selecting relevant features and the Synthetic Minority Over-sampling Technique (SMOTE) to handle class imbalance. The results showed that the Random Forest algorithm performed best with 95% accuracy, 77% precision, 87% recall, and 81% f1-score. This study contributes to developing a machine learning-based approach to support a more effective nutritional monitoring system and enable more appropriate dietary interventions to address toddler health problems in Indonesia.

Keywords: *classification, nutritional status, prediction, random forest, support vector machine.*

1. INTRODUCTION

According to the World Health Organization (WHO), nutritional status is the main indicator for evaluating toddlers' growth and dietary needs, including measuring the child's weight and height with standard anthropometric benchmarks. Nutritional status can vary based on gender, age, weight, height, and head circumference [1]. In Indonesia, dietary problems are still a significant challenge in the health sector [2]. Based on the Decree of the Minister of Health of the Republic of Indonesia Number 1995/MENKES/SK/XII/2010, determining nutritional status in toddlers is guided by anthropometric indices, including Height for Age (H/A), Weight for Age (W/A), and Weight for Height (W/H). These indicators classify toddlers as malnutrition, stunting, wasting, or obesity [3]. Data from the 2022 Indonesian Nutritional Status Study (SSGI) shows that the prevalence of stunting in toddlers reached 21.6%, underweight toddlers were 17.1%, wasting was 7.7%, and toddlers who are obese were 3.5% [4]. The high prevalence rate shows the importance of monitoring and early intervention to improve the nutritional status of toddlers in Indonesia. Nutritional problems not only hurt physical growth but can also inhibit cognitive development and reduce productivity in the future [5]. Therefore, early intervention is very much needed to overcome this problem.

Machine learning-based classification methods are one of the relevant approaches to support this effort. Classification is a data analysis method that uses machine learning to predict the membership of data samples into predetermined classes and groups [6]. Machine Learning (ML) is an interdisciplinary field built on ideas from cognitive science, computer science, statistics, and optimization, among many other scientific and mathematical disciplines [7]. This field allows the processing of large amounts of data, provides deep insight into data behaviour, and supports more precise decision-making based on the resulting analysis [8]. In its application, machine learning algorithms can be categorized into four main types, namely supervised, unsupervised, semi-supervised, and reinforcement learning. Each category has different goals and approaches according to the data analysis needs [9]. Machine learning techniques can process large and complex data [10]. Thus providing an opportunity to produce a classification of toddler nutritional status with high accuracy.

Previous research by Khansa et al. [11] has utilized machine learning algorithms such as K-Nearest Neighbor (KNN) and Naive Bayes to classify the nutritional status of toddlers, with datasets obtained from the Bojongsoang Health Center using the SMOTE oversampling technique to handle data imbalance, with attributes used in the study including age in months, height, weight, Z-Score Weight by Age, Z-Score Height by Age, and Z-Score Weight by Height. After applying the oversampling technique, the study showed a significant improvement in the KNN and NB models. In particular, the KNN model showed superior performance, increasing the F1 Score from 67.20% to 95.62%, with an accuracy of 95.67%. The Naive Bayes model also improved, increasing the F1 Score from 71.22% to 95.62%, with an accuracy of 94%.

Another study by Gina et al. [12] used the K-Nearest Neighbor (KNN) and Neural Network (NN) algorithms. The nutritional status of toddlers in this study was measured using age, gender, weight (BB), and height (TB) data. The results of the study showed that the JST algorithm, KNN with k = 3 on the BB/A, BB/TB, and TB/A datasets, KNN with k = 5 on the TB/A dataset, KNN with k = 7 on the TB/A dataset had the most optimum accuracy value 99% with a small error value 0.007.

Previous studies conducted by Khansa et al. using the K-Nearest Neighbors (KNN) and Naive Bayes algorithms, as well as research by Gina et al. using the K-Nearest Neighbor (KNN) and Neural Network (NN) algorithms for toddler nutritional status classification, have shown very good accuracy. However, both studies have not utilized feature selection techniques to identify the most relevant features and improve model performance more efficiently.

Feature selection is an important technique that allows models to run faster, eliminate noisy data, remove redundancy, reduce overfitting, increase accuracy, and improve generalization ability on test data [13]. This study introduces a new approach that implements feature selection techniques. Feature selection has a major impact on various applications, such as building simpler ones, improving learning performance, and creating clean and understandable data [14]. This study uses the Recursive Feature Elimination (RFE) feature selection technique, which has been proven to have higher accuracy than other feature selection methods such as ANOVA, Lasso, and Random Forest Feature Importance.

In addition, the Synthetic Minority Over-sampling Technique (SMOTE) is used in this study to handle class imbalance, ensuring the model is not biased towards the majority class. This technique effectively improves classification performance on imbalanced datasets [15]. Balancing data between the majority and minority classes, SMOTE can improve accuracy in predicting less common statuses, such as obesity or overnutrition.

This study uses the Support Vector Machine (SVM) and Random Forest algorithms to compare the performance of two machine learning methods in developing classification models with optimal results. SVM was chosen because of its ability to handle high-dimensional data [16]. At the same time, Random Forest was used because of its ability to produce reliable classification performance on datasets with high dimensions and variable complexity [17]. The performance is considered superior to other classification algorithms, such as K-Nearest Neighbor (KNN) and Neural Networks (NN), because with larger data sets, classification in KNN and NN takes longer. At the same time, SVM and Random Forest are more efficient in processing the data. In addition, SVM and Random Forest are easier, more practical, and faster to implement. [18]

This study aimed to develop a classification model of toddler nutritional status based on anthropometric indices based on machine learning algorithms. The classification includes good nutrition, poor nutrition, overnutrition, risk of overnutrition, and obesity. By implementing feature selection and data balancing techniques, this study is expected to contribute significantly to supporting a more effective toddler nutritional monitoring system and encouraging appropriate interventions in addressing dietary problems in Indonesia.

2. METHOD

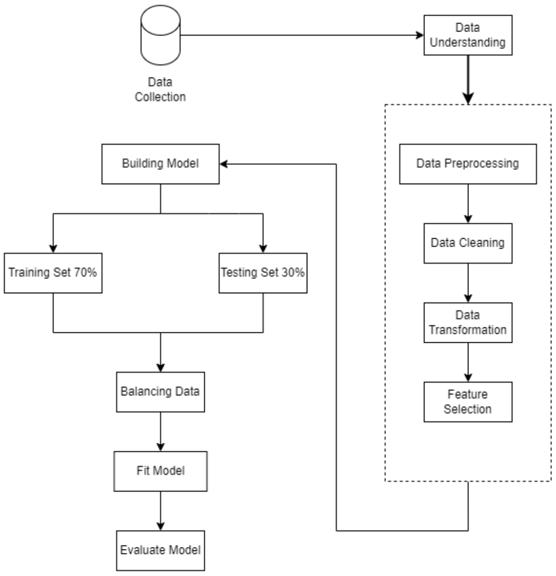


Figure 1. Flow Method Research

This study proposes a classification system for toddler nutritional status through structured stages. The research flow can be seen in Figure 1. The research process begins with data collection, followed by data preprocessing, which includes cleaning missing values, data scale transformation, categorical variable encoding, and feature selection using certain methods. Unlike previous studies, this study emphasizes feature selection to improve data relevance and model efficiency, reducing the risk of overfitting.

After the data is ready, the dataset is divided into 70% training and 30% testing data, with crossvalidation techniques to ensure consistent results. Data balancing is done using SMOTE, which is more effective in generating synthetic data for minority classes than conventional oversampling methods.

The next stage is the implementation of machine learning algorithms for toddler nutritional status classification, namely by implementing the SVM and Random Forest algorithms. In the final step, model evaluation is carried out to measure the algorithm's performance in terms of accuracy, precision, recall, and f1-score.

2.1. Data Collection

This study utilizes a toddler nutrition measurement dataset consisting of 9,735 records, sourced from a government agency in Indonesia. Of the total dataset, 8,731 records classify toddlers as well-nourished, 477 as undernourished, 345 as at risk of over-nutrition, 83 as over-nourished, 66 as malnourished, and 33 as obese. This dataset includes 19 attributes and one target class for nutritional status classification. Preliminary analysis revealed several data quality issues, such as missing values and outliers, across multiple attributes, which could affect the accuracy and robustness of the classification model. Therefore, a comprehensive data pre-processing stage is conducted to address these challenges, including imputation of missing values and treatment of outliers to maintain data integrity.

2.2. Data Understanding

This process aims to understand the characteristics of the dataset used deeply. This study aims to understand the characteristics of the toddler dataset used in classifying nutritional status. The dataset consists of various toddler information, such as age, gender, weight, height, and nutritional status labels, which include categories such as good nutrition, malnutrition, poor nutrition, risk of overnutrition, overnutrition, and obesity. At this stage, data exploration is carried out to determine the distribution of data, patterns of relationships between variables, and potential problems such as missing data or outliers. This understanding is the basis for ensuring that the data used is ready to be further processed in the preprocessing and analysis stages so that the resulting model can accurately classify toddler nutritional status.

2.3. Data Preprocessing

Ini merupakan contoh penggunaan sub-bab pada paper. Sub-bab diperbolehkan untuk dimasukkan pada semua bab, kecuali di kesimpulan

Data preprocessing is a crucial step in machine learning, involving modifications or encodings that make the data interpretable for computational models [19]. Data used in data mining processes often requires optimization to ensure that it is in a condition suitable for accurate and efficient processing. Common challenges, such as missing values, redundant data, outliers, and incompatible data formats, can significantly impact the outcomes and reliability of data mining results. Addressing these issues is essential for maintaining data quality and achieving accurate model predictions. To overcome these problems, a rigorous data pre-processing stage is necessary, which includes multiple techniques, such as data cleaning, normalization, encoding , and feature selection.

1. Data Cleaning

The data cleaning process is very important when there are missing values in the attributes because unhandled gaps can significantly affect the accuracy of the machine learning model. This process includes several techniques, such as missing data imputation and outlier handling, which are important for maintaining data quality.

One of the main problems in data is the presence of missing values, which can affect the quality of analysis and the machine learning model's performance. To handle missing values in numerical data, use the imputation approach by replacing empty entries with the average value in the same class. This approach helps maintain data balance and reduces potential bias affecting the analysis results. Meanwhile, sequence-based imputation methods handle missing values in categorical data, such as forward fill (filling with the previous value) or backward fill (filling with the next value). This method fills in missing values by considering the previous or next values in the data sequence, thereby maintaining the consistency and sequential nature of the analyzed data set.

In addition to handling missing data, outlier processing is an important part of data cleaning. The method used to detect outliers in numeric data is the Z-score. This method calculates how far the data value is from the average, measured in standard deviation units. Values with a Z-score greater than a certain threshold, such as 3 or -3, are considered outliers because they indicate that the data is more than three standard deviations from the average value. Outlier management is done by removing the data because it is considered irrelevant to improve the accuracy and performance of the machine learning model. The formula for this outlier using z-score is equetion (1):

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

Where:
Z : Value of z-score
x : The data value to be checked.
μ : The average (mean) of the data.
σ : Standard deviation of the data.

2. Data Transformation

Data transformation is essential for preparing the dataset to be suitable for modeling, ensuring that the data is in a format compatible with machine learning algorithms [20]. The data transformation stage in this study includes feature standardization and feature coding to ensure that the features used in modeling have diverse scales and formats that machine learning algorithms can understand.

The standardScaler method is used for feature standardization, which is a data transformation process so that each feature in the dataset has an average (mean) of 0 and a standard deviation of 1. This process is carried out by calculating the difference between each data value and the average of the feature, then dividing it by the standard deviation of the feature in question. Thus, each feature will have a value centred around zero and have a uniform scale, which is important to ensure that all features contribute equally to the model. This standardization also helps speed up convergence in machine learning algorithms, reduces potential bias, and improves overall model performance. The StandardScaler equation can be seen in equation (2).

$$X_{Scaled} = \frac{x-\mu}{\sigma} \quad (2)$$

Where X is the original value of the feature, is the mean, and is the standard deviation.

The next transformation stage is categorical data encoding. Label coding is a process used for encode categorical values by changing each value is in a column as a number [21]. The label encoding method is applied at this stage to change the categorical features that were initially strings and cannot be directly used in machine learning algorithms because the model only accepts input in numeric form. Therefore, LabelEncoder is used to convert each unique category in the categorical attribute into a number so that the model can process it properly and allow the algorithm to work optimally without affecting the data distribution.

The formula for this LabelEncoder is equetion (3):

$$y_i = f(x_i) \quad (3)$$

Where is a mapping function that assigns an integer label to each unique category. This mapping is based on the frequency or occurrence order of the categories in the data.

3. Feature Selection

Feature selection in this study is conducted using the Recursive Feature Elimination (RFE) method, an effective approach for identifying the most relevant features to enhance model performance [22]. The Random Forest Classifier was applied to rank the features based on their contribution to predicting toddler nutritional status, ultimately selecting the top features that had the most significant impact. This selection process resulted in six key features: Head Circumference (HC), Mid-Upper Arm Circumference (MUAC), Age, Height, Weight-for-Age (WFA), and Weight, each demonstrating strong relevance in determining nutritional outcomes. This feature selection was followed by model training using only selected features, which resulted in a high accuracy of 95%. These results indicate that the selected features have predictive solid power and show the potential of the model to support nutritional status analysis more efficiently and accurately.

2.4. Building Model

In building a machine learning model, an important initial step is to divide the dataset into training and testing data. The goal is to ensure that the model built can learn patterns from the data well and generalize to new data that has never been seen before. Training Data is used to train the model. The model will learn the pattern of relationships between features (input) and labels (output) from this data. Testing Data evaluates the model's performance on new data not seen during the training process. This study divided the dataset into 70% for training data and 30% for testing data. This division was done so that the model had enough training and sufficient data for performance evaluation.

2.5. Balancing Data

In this study, the data imbalance issue is addressed using the SMOTE (Synthetic Minority Over-sampling Technique) method. SMOTE works by generating synthetic samples for minority classes, thereby balancing the dataset and improving the model's ability to learn underrepresented categories [23]. By creating new instances through interpolation between existing minority samples, SMOTE increases the minority class representation without directly duplicating data, which helps to reduce overfitting. This technique improves the model's sensitivity to the minority class, allowing for more accurate classification of less frequent categories, such as cases of malnutrition or obesity in toddlers. Furthermore, SMOTE is preferred over the ADASYN (Adaptive Synthetic Sampling) technique due to its more stable performance, as it avoids potential noise introduced by ADASYN's focus on hard-to-classify samples [24]. The balanced dataset generated through SMOTE improves the robustness of the model, ensuring that it is not overly biased towards the majority class. This ultimately contributes to higher accuracy and more reliable predictions, especially for rare courses critical in healthcare applications. The successful application of SMOTE in this study underscores its value in machine learning workflows where data imbalance is a critical challenge.

2.6. Fit Model

After the data is balanced, the next step is to train the machine learning model on it. In this study, the classification model was trained using two algorithms: Support Vector Machine (SVM) and Random Forest.

1. Support Vector Machine (SVM)

This study uses a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel to separate data into various classes. RBF kernel has superior capability in handling practical challenges, especially in non-linearly separated data, and requires fewer parameters to be adjusted than the polynomial kernel [25]. The RBF kernel can be formulated as follows :

$$K(x, x') = exp\left(-\gamma \left||x - x'|\right|^2\right) \tag{4}$$

Where:

- x and x' is two data vectors.
- $||x - x'||^2$ is the Euclidean squared distance between x and x'.
- γ is a kernel parameter that controls how much influence one data point has on another. A larger value of γ makes the model more sensitive to changes in the data, while a smaller value of γ makes the model more general.

2. Random Forest

In addition to SVM, this study utilizes the Random Forest algorithm, which offers several distinct advantages for classification tasks. Random Forest is known for its strong performance in accurately classifying data, its robustness in handling diverse types of datasets, and its processing speed, making it a popular choice in machine learning applications [26]. This algorithm operates by constructing an ensemble of decision trees, where each tree is trained on a subset of the data, allowing it to capture complex patterns while reducing the risk of overfitting. Random forest is formulated as:

$$\hat{y} = mode\,\{h_1(x), h_2(x), \dots, h_t(x)\} \tag{5}$$

Where:

- $\hat{y}$: The final prediction given by Random Forest for the input data x.
- $h_1(x), h_2(x), \dots, h_t(x)$: A set of predictions generated by t decision trees in an ensemble.
- $mode$: A function that takes the majority of votes from all predictions.

2.7. Evaluate Model

At this stage, an evaluation will be carried out to help determine whether the model can predict a classification well or not. Predictive ability is measured based on the value of the confusion matrix. The confusion matrix is a matrix that displays the actual classification prediction and the predicted classification .

Table 1. Confusion matrix

	Prediction	
Actual	**TRUE**	**FALSE**
TRUE	TP	FP
FALSE	FN	TN

This matrix consists of four main components, namely TP (True Positive), FP (False Positive), FN (False Negative), and TN (True Negative). The following is the explanation:

1) TP (True Positive) is the number of positive samples that are correctly classified.
2) TN (True Negative) is the number of negative samples that are correctly classified.
3) FP (False Positive) is the number of negative predictions that are incorrectly classified as positive.

Table 1. Shows the evaluation matrix that will be used to calculate the model performance, which can be computed using accuracy, precision, recall, and f1-score.

1. Accuracy

Accuracy is a metric that calculates the ratio between the number of correct predictions (both positive and negative) and the total amount of data. The accuracy formula is formulated in the following equation (6):

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

In the context of this study, accuracy describes how well the model can predict nutritional status from existing categories, such as good nutrition, poor nutrition, undernutrition, overnutrition, risk of overnutrition, and obesity. Although this metric provides a general overview of model performance, it has a weakness: sensitivity to uneven class distribution. Therefore, additional metrics are used to provide a more comprehensive evaluation.

2. Precision

Precision measures the level of accuracy of the model's positive predictions. This value indicates the proportion of correct positive predictions to all positive predictions. The precision formula is formulated in the following equation (7):

$$precision = \frac{TP}{TP+FP} \tag{7}$$

In this study, precision is important to assess how accurate the model is in identifying a particular nutritional status without producing many false positive predictions.

3. Recall

Recall, also known as sensitivity or True Positive Rate, indicates the model's ability to detect all positive data. The recall formula is formulated in the following equation (8):

$$recall = \frac{TP}{TP+FP} \tag{8}$$

Recall is relevant to this study because it describes how well the model can identify the true nutritional status category, especially in minority classes such as malnutrition or obesity.

4. F1-Score

F1-score is the harmonic mean of precision and recall, which provides a balance between the two. F1-score is very useful for evaluating models on datasets with imbalanced class distributions. The f1-score formula is formulated in the following equation (9):

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

The F1-score value ranges from 0 to 1, where values closer to 1 indicate that the model balances precision and recall. This is important to ensure that the model is accurate in its predictions and sensitive to the actual data.

3. RESULT

In this study, a classification model was developed to assess toddler nutritional status using two machine learning algorithms: Support Vector Machine (SVM) and Random Forest. These algorithms were selected for their effectiveness in handling complex classification tasks, with SVM providing precise boundary separation and Random Forest offering robust ensemble learning. This approach aims to accurately categorize nutritional statuses, supporting targeted interventions for toddler health.

3.1. Data Collection

The data used in this study were obtained from one of the Health Institutions in Indonesia, which contains information about toddlers for nutritional status analysis. This dataset consists of 9,735 entries equipped with 19 attributes that describe various demographic characteristics, health, and dietary measurements. These attributes include: Name, Date of Birth, Sex, District, Age of Measurement, Weight, Height, Upper Arm Circumference (UAC), Head Circumference, Measurement Method, W/H (Weight per Height), W/A (Weight per Age), H/A (Height per Age), Z-Score W/A, Z-Score H/A, Z-Score W/H, Parenting Pattern, Vit_A (Provision of Vitamin A), and Exclusive_ASI.

This dataset provides a comprehensive picture of the condition of toddlers, including aspects of physical growth, parenting patterns, and health interventions. The information in the dataset is very important in the analysis process because it helps understand the relationship between various factors that influence the nutritional status of toddlers.

Table 2. Dataset

...	Sex	Weight	Height	...	W_H
...	M	12,15	92,0	...	Gizi Baik
...	F	6,7	73,5	...	Gizi Buruk
...	M	8,45	78,3	...	Gizi Kurang
...	M	18,5	100,0	...	Gizi Lebih
...	F	22,65	100,8	...	Obesitas
...	...	...	...	...	...
...	...	...	...	...	...
...	F	14	90,5	...	Resiko Gizi Lebih

The dataset view is presented briefly in Table 2, showing the data structure for several attributes, including sex, weight, height and nutritional status category.

3.2. Data Understanding

The dataset used in this study consists of 9,735 entries with 19 attributes that describe the characteristics of toddlers from various aspects. The target label in this dataset is the nutritional status category, which consists of six classes: Good Nutrition, Undernutrition, Malnutrition, Overnutrition, Risk of Overnutrition, and Obesity. Initial analysis showed data imbalance, where most entries were in the Good Nutrition category, while the other categories had much fewer entries. To address this problem, this study used the SMOTE oversampling technique to balance the classes. In addition, the dataset has 2,932 missing values in the Upper Arm Circumference (UAC) column, 2,874 in the Head_Circumference column, 4,876 in the Parenting_Pattern column, 4,867 in the Vit_A column, and 4,867 in the Exclusive Breastfeeding column. The dataset also contains outliers in the columns of Upper Arm Circumference (UAC), Weight, Height, and Head Circumference. So, it is necessary to handle outliers so they do not affect model performance. In addition, attributes in the dataset have different scales, such as weight in kilograms and height in centimetres. This requires normalization so all features have a uniform scale before being used in a machine-learning model.

3.3. Data Preprocessing

Data preprocessing includes several processes, including:

1. **Data Cleaning**

The data cleaning stage is carried out to ensure the data quality used in the analysis. In this process, missing values and outliers in the dataset are handled. Missing values are identified in attributes such as Upper Arm Circumference (UAC), Head_Circumference, Parenting_Pattern, Vit_A, and Exclusive_Breastfeeding. Because the attributes are considered important for determining nutritional status, these attributes are not deleted, and the imputation method will be applied to overcome the missing values. The imputation method is the mean (average value) for numeric data type attributes and backwards and forward fill imputation for categorical type attributes. In addition, outlier detection is carried out using a z-score to identify data outside the normal limits. The detected outliers are then removed. This process is carried out carefully to maintain significant outliers in the domain.

2. Data Transformation

To ensure that all features in the dataset are on a uniform scale, a transformation using StandardScaler is performed. This transformation changes the value of each feature so that it has a mean of 0 and a standard deviation of 1. This process is important because the features in the dataset have different scales, such as Weight in kilograms and Height, Head Circumference, and Upper Arm Circumference (UAC) in centimetres, which can cause the machine learning algorithm to be biased towards features with large values. The results before and after the standard scaler can be seen in the Table 3 and Table 4.

Table 3. Dataset before scaled

Index	Before Scaled				
	Weight	...	Height	UAC	...
7788	13,1	...	88,0	15,0	...
4535	11,7	...	81,8	17,0	...
8751	11,3	...	87,5	9,9	...
495	9,2	...	82,5	14,0	...
9089	16,3	...	100,0	15,0	...

Tabel 4. Dataset after scaled

Index	After Scaled				
	Weight	...	Height	UAC	...
7788	0,41023	...	0,11616	0,36747	...
4535	-0,00511	...	-0,35801	1,09356	...
8751	-0,12378	...	0,07792	-1,11376	...
495	-0,74679	...	-0,30447	0,07703	...
9089	1,35958	...	1,03392	0,36747	...

This study utilized a dataset of toddler anthropometry measurements and demographic to develop a classification model for toddler nutritional status. The dataset includes essential attributes that were carefully transformed and encoded to ensure compatibility with the classification algorithms. Categorical attributes were converted into numeric format using LabelEncoder, which preserves the data structure while making it readable and processable by the model. This transformation facilitates the algorithm's ability to interpret categorical data accurately, enhancing overall model performance. Tables 2 and 5 present a comparison of the dataset before and after encoding, illustrating the transformation's impact on data readiness for analysis.

Table 5. Dataset after encoding

...	Sex	Weight	Height	...	W_H
...	0	12,15	92,0	...	0
...	1	6,7	73,5	...	1
...	0	8,45	78,3	...	2
...	0	18,5	100,0	...	3
...	1	22,65	100,8	...	4
...	...	...	...	...	...
...	...	...	...	...	...
...	1	14	90,5	...	5

Table 5 displays the results of the encoding process. In the "Sex" attribute, the original categorical values of M (Male) and F (Female) have been converted to 0 and 1, respectively. Similarly, for the "W_H" attribute, which represents the nutritional status of toddlers, categories have been assigned numeric values: Good Nutrition is encoded as 0, Malnutrition as 1, Undernutrition as 2, Overnutrition as 3, Obesity as 4, and Overnutrition Risk as 5. These numerical codes allow the algorithm to efficiently recognize and process the variations in nutritional status categories without altering the underlying meaning of the original data. This encoding approach simplifies the data for the model while preserving the integrity of categorical distinctions. As a result, the transformation enhances algorithm compatibility and contributes to more accurate and streamlined processing of the dataset.

3. Feature Selection

Following the encoding of the dataset, the next step is feature selection. In this process, several feature selection techniques are applied to identify the features that contribute most significantly to accurately classifying toddlers' nutritional status. The methods used include Recursive Feature Elimination (RFE), Lasso Regression, Analysis of Variance (ANOVA), and Random Forest Feature Importance. Each technique is evaluated based on the accuracy it achieves, helping to determine the optimal combination of features for predicting toddlers' nutritional status. The accuracy results of each feature selection method are presented in Table 6, providing a basis for selecting the most relevant features to enhance the model’s overall performance. This selection process aims to streamline the dataset by focusing on the most impactful features, thus improving efficiency and prediction accuracy.

Table 6. Feature Selection Test Results

	RFE	Lasso	ANOVA	RF Feature Importance
Accuracy	95%	91%	93%	90%

As shown in Table 6, the Recursive Feature Elimination (RFE) method achieved the highest accuracy among the feature selection techniques, with an impressive accuracy of 95%. This result highlights the critical role of selecting relevant features in enhancing the performance of the toddler nutritional status classification model. The selected feature, Height and Weigh, demonstrated a significant contribution to accurate predictions of toddler nutritional status. These features not only improve the model's accuracy but also streamline the classification process by focusing on the most impactful indicators. This outcome underscores the importance of robust feature selection in building reliable and effective models for health-related applications. The dataset of the feature selection results can be seen in Table 7.

Table 7 . Dataset After Feature Selection

Weight	Height	W_H
12,15	92,0	0
6,7	73,5	1
8,45	78,3	2
18,5	100,0	3
22,65	100,8	4

...	...	...
...	...	...
14	90,5	5

3.4. Building Model

In this study, the model-building process was carried out to classify the nutritional status of toddlers based on anthropometric indices. The data used in this study were divided into two main parts, namely training data and testing data, with a proportion of 70% for training and 30% for testing. This division aims to ensure that the model has enough data to learn from existing patterns, as well as to test the model's ability to generalize to data that has never been seen before.

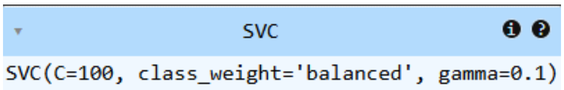
3.5. Balancing Data

To address the issue of class imbalance in the dataset, SMOTE was applied. SMOTE balances the distribution across nutritional status categories, including good nutrition, undernutrition, overnutrition, poor nutrition, risk of overnutrition, and obesity. By using this technique, the model is better equipped to handle class imbalance, which enhances its generalization ability and accuracy in classifying toddler nutritional status. Initially, the dataset displayed a significant imbalance, with the good nutrition category comprising 8,731 records, while other categories had substantially fewer entries malnutrition had 477 records, risk of overnutrition had 345 records, overnutrition had 83 records, and obesity had only 33 records. After implementing SMOTE, each category was balanced, with 6,133 records in each nutritional status class, ensuring equal representation across categories. This balanced distribution allows the model to learn from each category equally, leading to more reliable and accurate predictions and reducing the risk of model bias toward the majority class. The use of SMOTE significantly improves the model's robustness and effectiveness in practical applications related to toddler nutrition.

3.6. Fit Model

In this study, the models chosen were SVM and Random Forest, the following is an explanation:

1. **Support Vector Machine (SVM)**

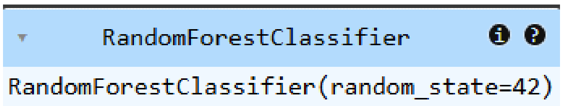


```
SVC(C=100, class_weight='balanced', gamma=0.1)
```

Figure 2. SVM Model Implementation

Figure 2 shows the implementation of the SVM model. The Support Vector Machine (SVM) model uses the SVC function from the sci-kit-learn library with a Radial Basis Function (RBF) kernel. The main parameters used include C=100, which provides regularization to control the balance between wide margins and misclassifications in the training data, and gamma=0.1, which determines the influence of a single data point on the model's decision. In addition, the class_weight='balanced' option is applied to handle class imbalance by giving proportional weights to the frequency of each class. The model is trained on data treated with the SMOTE method to improve the representation of the minority class and to ensure the model's generalization ability to non-linear patterns in the data.

2. **Random Forest**



```
RandomForestClassifier(random_state=42)
```

Figure 3. Random Forest Model Implementation

Figure 3 shows the implementation of the random forest model. The Random Forest model was implemented using the RandomForestClassifier function from the scikit-learn library, with default parameters and random values specified via random_state=42 to ensure the reproducibility of the results. The algorithm works by constructing several decision trees from random subsets of the training data and aggregating them via majority voting for classification. The model was trained on training data processed using the SMOTE method to address class imbalance. This approach allows the model to capture complex patterns and improves generalization capabilities, especially on data with imbalanced class distributions.

3.7. Evaluate Model

The implementation process involves evaluating each algorithm's performance on the balanced dataset to determine its effectiveness in classifying toddler nutritional status. Key performance metrics, including accuracy, precision, recall, and f1-score are measured and compared for both algorithms to assess their strengths and limitations. The comparative results of the SVM and Random Forest models are displayed in Table 8, providing insights into each model's capability in handling this classification task. This analysis helps identify the more suitable algorithm for accurately predicting nutritional categories, ensuring a reliable and effective classification system.

Table 8. Results of testing the SVM and RF models

	SVM	**RF**
Accuracy	93%	95%
Precision	71%	77%
Recall	90%	87%
F1-score	79%	81%

Based on the results presented in Table 8, the test performance for the two classification algorithms, Support Vector Machine (SVM) and Random Forest (RF), shows promising results in classifying toddlers' nutritional status. The SVM algorithm achieved an accuracy of 93%, with a precision of 71% and a recall of 90%, indicating that while SVM is good at detecting positive cases, it struggles with false positives. On the other hand, the Random Forest algorithm slightly outperformed SVM, with an accuracy of 95%, precision of 77%, and recall of 87%. This indicates that Random Forest strikes a better balance between precision and recall than SVM. While both models performed well, Random Forest provided a higher F1 score of 81%, compared to SVM's F1 score of 79%, demonstrating its more balanced approach to handling false positives and false negatives. Overall, Random Forest is a more practical choice for accurately classifying the nutritional status of toddlers, particularly in handling complex data sets. Figure 4 displays the precision, recall, and F1-score values for each category of toddler nutritional status: "gizi baik" (good nutrition), "gizi buruk" (malnutrition), "gizi kurang" (undernutrition), "gizi lebih" (overnutrition), "obesitas" (obesity), and "resiko gizi lebih" (risk of overnutrition). The "gizi baik" (good nutrition) category demonstrates high precision, recall, and F1-score values.

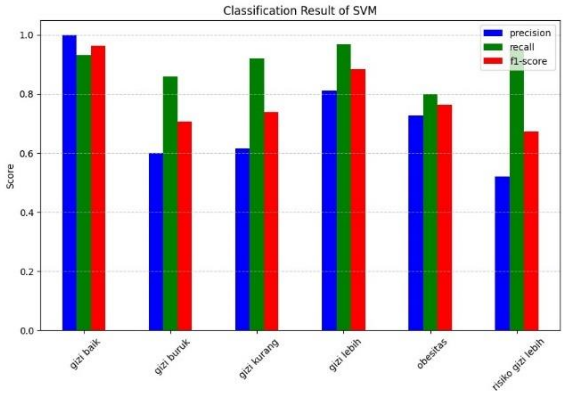


Figure 4. Classification Result of SVM

In Figure 4 shows that the SVM model performed very well in the ”gizi baik” (good nutrition) category, with a precision of 100%, a recall of 93%, and an F1-score of 96%, reflecting a very accurate classification. In the “gizi buruk” (malnutrition) category, the precision was 60%, while the recall was higher at 86%, with an F1-score of 71%. This indicates that the model could detect most cases of “gizi buruk” (malnutrition), although there were some false positives. For the “gizi kurang” (undernutrition) category, the precision was 61%, the recall was 92%, and the F1-score was 74%, indicating that the model was better at detecting cases (high recall) than providing accurate predictions. In the "gizi lebih" (overnutrition) category, the model performed strongly with a precision of 81%, a recall of 97%, and an F1-score of 88%, reflecting consistent accuracy in this category. For the "obesitas" (obesity) category, the precision was 73%, recall 80%, and F1-score 76%, indicating that the model was quite effective in identifying "obesitas" (obesity) cases but produced some false positives. In the "risiko gizi lebih" (risk of overnutrition) category, the precision was the lowest, at 52%, while the recall was very high, at 95%, resulting in an F1-score of 67%. This indicates that the model can detect cases in this category well, but its precision still needs to be improved. Overall, the SVM model achieved an accuracy of 93% for all categories. The macro average values were precision 71%, recall 90%, and F1-score 79%, while precision 96%, recall 93%, and F1-score 94%. This indicates that the SVM model works very well for the majority category but needs improvement in minority categories, such as “gizi buruk” (malnutrition) and "risiko gizi lebih" (risk of overnutrition), to improve prediction accuracy. The classification results from Random Forest can be seen in Figure 5.

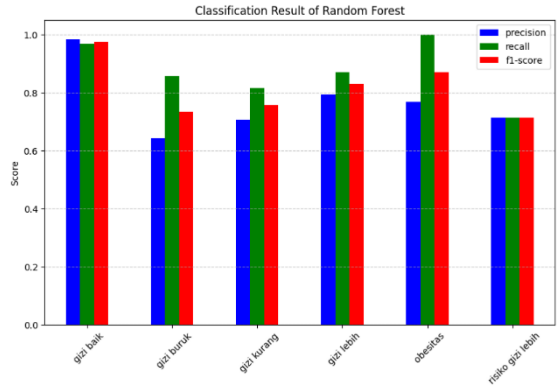


Figure 5. Classification Result of Random Forest

In Figure 5 shows that the Random Forest model performed very well in the ”gizi baik” (good nutrition) category, with a precision of 98%, a recall of 97%, and an F1-score of 98%, reflecting the model’s ability to classify this category accurately. In the “gizi buruk” (malnutrition) category, the precision reached 64%, a higher recall of 86%, and an F1-score of 73%, indicating that the model was quite good at detecting “gizi buruk” (malnutrition) cases despite producing some false optimistic predictions. For the “gizi kurang” (undernutrition) category, the precision was 71%, a recall of 82%, and an F1-score of 76%, indicating that the model was better at detecting cases than making precise predictions. In the "gizi lebih" (overnutrition) category, the precision reached 79%, a recall of 87%, and an F1-score of 83%, indicating good and consistent performance in classifying this category. In the "obesitas" (obesity) category, the model achieved a precision of 77%, a recall of 100%, and an F1-score of 87%, indicating that the model was very good at detecting all obesity cases (high recall) but produced some false positives. For the risiko gizi lebih" (risk of overnutrition) category, precision and recall were 71%, with an F1-score of 71%, indicating that the model performed moderately in classifying this category, possibly due to more complex data patterns or limited data size. Overall, the Random Forest model achieved 95% accuracy across all categories. The macro averages were 77%

precision, 87% recall, and 81% F1-score, indicating good performance for the minority category, although there is still room for improvement. Model errors tend to occur in categories with fewer data and patterns that are difficult to distinguish from other categories, such as “gizi buruk” (malnutrition) and "gizi lebih" (overnutrition). These results indicate that the Random Forest model can classify most nutritional status categories with high accuracy compared to the SVM model. Figure 6 shows a comparison graph of the performance of the SVM and Random Forest models.

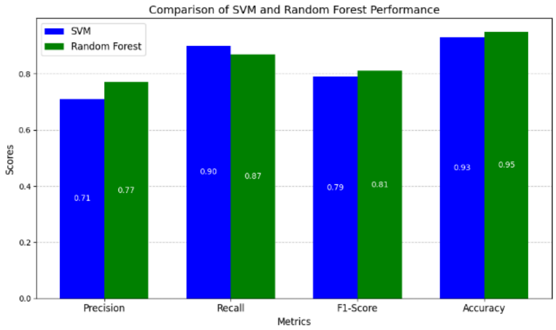


Figure 6. Comparison of SVM and Random Forest Performace

Based on the performance comparison graph of SVM and Random Forest shown in Figure 6 and from the explanation of the research results. Overall, Random Forest proved to be more effective in providing a more balanced performance and higher accuracy for classifying the nutritional status of toddlers in this study.

4. DISCUSSION

This study developed a classification of toddler nutritional status by comparing the performance of the Support Vector Machine (SVM) and Random Forest algorithms. The results showed that the Random Forest algorithm was superior, with an accuracy of 95%, compared to SVM. This difference can be explained by the nature of the Random Forest algorithm, which utilizes ensemble learning to build decisions based on many trees, making it better able to handle high data complexity and overcome data imbalance. The Recursive Feature Elimination (RFE) selection technique also helps identify relevant features, making the model training process more focused. On the other hand, SVM, although strong in handling data with clear margins, may face challenges in handling datasets with high dimensions and imbalanced class distributions, even though SMOTE has done oversampling. When compared to a similar study by Anamisa et al. [27], which also used SVM for the classification of malnutrition status on Madura Island, the results of this study show a significant difference. In the study [27], SVM with a polynomial kernel achieved an accuracy of 89.76%, while in this study, SVM produced lower accuracy than Random Forest. This may be due to differences in dataset characteristics, size, and class distribution. This study used a larger dataset (9.735 toddler data) with more diverse nutritional status classes compared to the study [27], which used 694 data. With larger data, Random Forest can show superior performance because of its nature, allowing it to capture data variations better than SVM.

This research also significantly impacts monitoring the nutritional status of toddlers in Indonesia. With the high accuracy achieved by Random Forest, this model can be implemented in a technology-based system to predict nutritional status automatically to help health workers make faster and more accurate decisions. The SMOTE oversampling technique used also shows potential in overcoming the problem of data imbalance, which often occurs in public health data in Indonesia.

However, this study has several limitations. Although oversampling with SMOTE successfully improves model performance, this technique also has several limitations that must be considered. SMOTE is effective on numeric data, but its application to text data is less than optimal. In addition, this study only used two classification algorithms, namely SVM and Random Forest, so it cannot provide insight into how other algorithms, such as Gradient Boosting or Neural Networks, will perform in the same dataset. In the future, further research can explore other algorithms or integrate hybrid methods to improve model performance. In addition, research can focus on testing data from various regions in Indonesia to understand whether the resulting model has good generalization to the entire population.

5. CONCLUSION

This study compares the Support Vector Machine (SVM) and Random Forest algorithms to classify the nutritional status of toddlers, with Random Forest outperforming SVM. Random Forest achieved 95% accuracy, higher than SVM, which achieved 93% and showed better balance with 77% precision, 87% recall, and 81% F1 score, making it more effective in handling imbalanced data. Random Forest's superior performance can be attributed to its ability to handle high-dimensional and complex data and its resilience to overfitting, a common problem in models dealing with smaller or imbalanced datasets. On the other hand, SVM struggled with minority class classification due to its focus on optimal splitting, which does not always consider the data distribution, leading to low precision and high false positives in that class. Although SMOTE has been applied for data balancing, SVM and Random Forest still face difficulties accurately classifying the minority class. This suggests that class imbalance is still challenging, and additional techniques may be needed to address this issue further.

Meanwhile, feature selection plays an essential role in improving model performance. Relevant feature selection can help reduce model complexity and enhance model performance to focus on the most critical attributes, ultimately improving classification accuracy and efficiency. Feature selection using Recursive Feature Elimination (RFE) effectively improves model performance. This study highlights how proper feature selection is critical to improving model performance in classifying toddler nutritional status.

For further research, it is recommended to explore deep learning-based algorithms, which can be more adaptive to data variations and better handle data complexity with a more significant number of features. In addition, to improve the model's ability to classify the minority class, it is recommended to use other oversampling techniques such as ADASYN or Borderline-SMOTE. These techniques can help generate synthetic samples that focus more on areas that are harder to predict, thereby improving the model's performance in minority class classification. In addition, testing the model in other regions with different dataset characteristics can also help ensure the generalizability of the research results.

ACKNOWLEDGMENT

This research received support from various parties. Therefore, the author would like to thank the Supervisor, who provided input and suggestions during the researcher's research and work on the Journal. Then, thanks to both parents, who always provide encouragement and prayers to the researcher. Finally, thanks to fellow strugglers, who always accompany the researcher during his education.

REFERENCES

[1] C. L. Nelms *et al.*, "Assessment of nutritional status in children with kidney diseases — clinical practice recommendations from the Pediatric Renal Nutrition Taskforce," pp. 995–1010, 2021, doi: https://doi.org/10.1007/s00467-020-04852-5.

[2] Kementerian Kesehatan Indonesia, *Menuju Solusi Gizi Seimbang: Tantangan dan Langkah-langkah Konkrit di Indonesia*. 2021.

[3] F. I. Shiyam *et al.*, "Relationship Between Mother ' s Knowledge Level In Utilizing The Maternal and Child Health Book With The Nutritional Status Of Toddlers," no. 3, 2024, doi: https://doi.org/10.62951/ijph.v1i3.72.

[4] B. Kebijakan, P. Kesehatan, and K. K. Ri, *Status Gizi SSGI 2022*.

[5] A. Soliman *et al.*, "Early and long-term consequences of nutritional stunting: From childhood to adulthood," *Acta Biomedica*, vol. 92, no. 1, 2021, doi: 10.23750/abm.v92i1.11346.

[6] D. Kurniasari, R. Nurul Hidayah, and R. Khoirun Nisa, "CLASSIFICATION MODELS FOR ACADEMIC PERFORMANCE: A COMPARATIVE STUDY OF NAÏVE BAYES AND RANDOM FOREST ALGORITHMS IN ANALYZING UNIVERSITY OF LAMPUNG STUDENT GRADES," *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 5, pp. 1267–1276, 2024, doi: 10.52436/1.jutif.2024.5.5.2066.

[7] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, 2020, doi: 10.1016/j.neucom.2019.10.118.

[8] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Machine learning towards intelligent systems: applications, challenges, and opportunities," *Artif Intell Rev*, vol. 54, no. 5, 2021, doi: 10.1007/s10462-020-09948-w.

[9] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," 2021. doi: 10.1007/s42979-021-00592-x.

[10] A. D. Milan Picard, Marie-Pier Scott-Boyer, Antoine Bodein, Olivier Périn, "Integration strategies of multi-omics data for machine learning analysis," *Comput Struct Biotechnol J*, vol. 19, pp. 3735–3746, 2021, doi: https://doi.org/10.1016/j.csbj.2021.06.030.

[11] G. A. F. Khansa and P. H. Gunawan, "Predicting Stunting in Toddlers Using KNN and Naïve Bayes Methods," *International Conference on Data Science and Its Applications (ICoDSA)*, pp. 17–21, 2024, doi: 10.1109/ICoDSA62899.2024.10651676.

[12] G. P. I. I. Y. Sri Rahmawati, "Implementation of KNN and ANN to the classification of the nutritional status of toddlers based on anthropometric indices," *CoSciTech*, vol. 4, no. 2, pp. 385–393, Aug. 2023, doi: https://doi.org/10.37859/coscitech.v4i2.5079 .

[13] E. O. Abiodun, A. Alabdulatif, O. I. Abiodun, M. Alawida, A. Alabdulatif, and R. S. Alkhawaldeh, "A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities," 2021. doi: https://doi.org/10.1007/s00521-021-06406-8.

[14] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare,"

IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3001149.

[15] A. Sakho, E. Scornet, and E. Malherbe, "Theoretical and experimental study of SMOTE: limitations and comparisons of rebalancing strategies," 2024. [Online]. Available: https://hal.science/hal-04438941v1

[16] N. Cesario, D. Lewis, C. Rosales, F. Antolini, R. Stojanovic, and L. Vandenberg, "Ransomware Detection Using Opcode Sequences and Machine Learning: A Novel Approach with t-SNE and Support Vector Machines," Oct. 22, 2024. doi: 10.36227/techrxiv.172963142.20817264/v1.

[17] J. Quist, L. Taylor, J. Staaf, and A. Grigoriadis, "Random forest modelling of high-dimensional mixed-type data for breast cancer classification," *Cancers (Basel)*, vol. 13, no. 5, 2021, doi: 10.3390/cancers13050991.

[18] B. E. , Yeboah, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K -Nearest-Neighbor , Support Vector Machine , Random Forest and Neural Network : A Review," *Journal of Data Analysis and Information Processing*, vol. 8, no. 4, 2020, doi: https://doi.org/10.4236/jdaip.2020.84020.

[19] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3104357.

[20] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3053759.

[21] M. G. Karthik, "Detecting Internet of Things Attacks Using Post Pruning Decision Tree-Synthetic Minority Over Sampling Technique," *International Journal of intelligent Engineering & System*, vol. 14, no. 4, pp. 105–114, 2021, doi: 10.22266/ijies2021.0831.10.

[22] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00327-4.

[23] J. H. Joloudari, M. Abdolreza, M. Ali Nematollahi, and S. Sunday Oyelere, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Applied Sciences*, vol. 13, no. 6, 2023, doi: https://doi.org/10.3390/app13064006.

[24] J. Brandt and E. Lanzén, "A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification," *Dissertation*, 2021, [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1519153

[25] X. Wu, C. S. Lai, C. Bai, L. L. Lai, Q. Zhang, and B. Liu, "Optimal kernel ELM and variational mode decomposition for probabilistic PV power prediction," *Energies (Basel)*, vol. 13, no. 14, 2020, doi: 10.3390/en13143592.

[26] D. Yates and M. Z. Islam, "FastForest: Increasing random forest processing speed while maintaining accuracy," *Inf Sci (N Y)*, vol. 557, pp. 130–152, 2021, doi: https://doi.org/10.1016/j.ins.2020.12.067.

[27] D. R. Anamisa, A. Jauhari, and F. A. Mufarroha, "PERFORMANCE TEST OF NAIVE BAYES AND SVM METHODS ON CLASSIFICATION OF MALNUTRITION STATUS IN CHILDREN," *Communications in Mathematical Biology and Neuroscience*, vol. 2024, 2024, doi: 10.28919/cmbn/8429..