Enhancing Multiclass Classification of Child Nutritional Status Using KNN and Random Forest with SMOTE

1st Aqeela Fathya Najwa
CoE HUMIC, School of Computing
Telkom University
Bandung, Indonesia
aqeelafn@student.telkomuniversity.ac.id

2nd Indwiarti
CoE HUMIC, School of Computing
Telkom University
Bandung, Indonesia
indwiarti@telkomuniversity.ac.id

3rd Putu Harry Gunawan
CoE HUMIC, School of Computing
Telkom University
Bandung, Indonesia
phgunawan@telkomuniversity.ac.id

***Abstract*—This study investigates the application of SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance in children's nutritional status datasets, focusing on two indicators: BB/U (Weight-for-Age) and BB/TB (Weight-for-Height). The goal is to enhance the predictive performance of machine learning models, particularly in classifying underrepresented nutritional categories. K-Nearest Neighbors (KNN) and Random Forest were employed to evaluate SMOTE's effectiveness. The results reveal significant improvements in recall for minority classes. For KNN, testing accuracies reached 96.66% for BB/U and 93.58% for BB/TB, with enhanced recall values for minority categories. Random Forest demonstrated superior performance with cross-validation accuracies of 97.59% for BB/U and 94.79% for BB/TB, achieving balanced classification across major and minor classes. The dual use of BB/U and BB/TB as target columns proved crucial for a comprehensive assessment of nutritional status, as each captures different dimensions of child growth. Additionally, key features such as gender and prior weight status were found to significantly influence model predictions. By improving the ability to detect at-risk groups, this study offers actionable insights to support more precise and data-driven nutritional interventions. The findings provide valuable guidance for policymakers and healthcare professionals in Indonesia, contributing to more effective strategies to combat childhood malnutrition and promote equitable health outcomes. These results highlight the potential of machine learning techniques, when combined with SMOTE, to address public health challenges in a robust and scalable manner.**

***Index Terms*—nutrition status, machine learning, SMOTE**

I. Introduction

Nutritional problems in early childhood (0-60 months) are a critical public health issue because this period is a highly sensitive phase of growth. Nutrition plays a crucial role in supporting the growth and development of children [1]. Malnutrition, whether in the form of undernutrition or overnutrition, can significantly affect their development and increase the risk of chronic diseases in the future. The adverse effects of nutritional disorders during this early age can persist into adulthood, ultimately affecting their quality of life in the long term [2]. Therefore, ensuring optimal nutrition for young children is essential for supporting healthy development.

Although Indonesia has made significant progress in reducing the prevalence of undernutrition and malnutrition in children, these issues still represent a major challenge. Based on available data, the prevalence of undernutrition in children aged 0-5 years decreased from 31% in 1989 to 17.9% in 2010, and malnutrition decreased from 12.8% in 1995 to 4.9% in 2010 [1]. However, the prevalence of undernutrition and overnutrition in toddlers remains high, despite various interventions. The age period of 2-5 years is a critical transition period between infant and adult food, making children at this age particularly vulnerable to nutritional imbalances [3]. In 2005, the number of children aged 0-6 years in Indonesia reached 27.6 million, but only 25% had access to nutrition improvement programs [2].

One of the common methods used to monitor children's nutritional status is through anthropometric indicators such as Weight for Age (BB/U) and Weight for Height (BB/TB) [2]. Although these indicators are widely used, the inconsistency of BW/H measurements often makes it difficult for health workers to determine children's nutritional status and formulate appropriate interventions. Therefore, a more effective approach is needed to improve the accuracy of children's nutritional status classification.

The selection of machine learning algorithms, such as Random Forest (RF) and K-Nearest Neighbor (KNN), is based on their ability to handle complex and high-dimensional data. KNN, effective for datasets with multiple features, improves accuracy by selecting the optimal K value to avoid overfitting or underfitting [4]. RF, an ensemble method, aggregates results from multiple decision trees to enhance prediction accuracy and can handle large, imbalanced datasets with missing data [5], [6]. While both algorithms have been applied in classifying health issues, their use for classifying nutritional status with anthropometric indicators like WFA and WFH remains limited, making them suitable choices for this study.

This study aims to develop an artificial intelligence-based model that is faster, more accurate, and more objective in classifying children's nutritional status. The model will be trained using RF and KNN algorithms, with evaluation using accuracy and F1-score metrics. Through a multiclass classification approach, this model is expected to provide a more precise solution and support data-driven decision-making by health professionals and the government, in order to reduce the prevalence of undernutrition and malnutrition in toddlers in Indonesia.

II. Model

A. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a classification method that classifies data based on its proximity to other data points

979-8-3315-0649-0/24/$31.00 ©2024 IEEE

in a higher-dimensional space [4]. In KNN, the training data is projected into a multi-dimensional space, with each dimension representing a feature of the data being analyzed. The classification process is done by finding the nearest neighbors of the new data to be classified, using Euclidean distance as the proximity measure [7].

The selection of K (the number of neighbors used for classification) is crucial, and its optimal value can be determined using techniques such as cross-validation to avoid overfitting or underfitting [4].

The Euclidean distance formula for two-dimensional data is as follows [8]:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

where x_1, y_1 and x_2, y_2 are the coordinates of the two points being compared. This formula is used for two-dimensional data.

B. Random Forest

Random Forest (RF) is an ensemble learning algorithm that uses a collection of decision trees, built randomly from a subset of the training data. Each tree is constructed using the bootstrap sampling technique [5], and the final classification decision is made by majority voting from all the trees.

RF is effective at handling large datasets, missing data, and imbalanced class data, with a key advantage in reducing overfitting due to the use of multiple trees [6]. At each tree node, RF selects features randomly to find the best split, unlike a single decision tree that uses all features [9].

C. Research Design

This study begins by defining the problem, which is to classify the nutritional status of children based on the BB/U and BB/TB indicators. The initial step includes planning the research, which involves setting objectives, selecting the analysis methods, and identifying the data sources. A literature review was conducted to understand relevant studies and methods, including algorithms such as Random Forest (RF) and K-Nearest Neighbor (KNN), as well as imbalanced data handling techniques, such as SMOTE.

The data used in this study were obtained from the Bandarharjo Health Center, imported from Excel files to DataFrame using the pandas library. After importing the data, the next step was to understand the data structure, clean it from duplicates, and analyze its distribution to identify the proportion of each nutritional status category, such as normal, malnourished, and other categories based on BB/U and BB/TB indicators.

Relevant features such as Body Weight (BW), Height (H), Age in Months, Gender, and Weight Gain Status were selected for analysis to determine whether these features affect children's nutritional status. The data is then visualized to identify patterns, distributions, and outliers that may affect the classification results. Afterward, the data is prepared for further analysis by splitting the dataset into features (X) and target (label), as well as dividing the data into training and testing sets.

For the application of the **Random Forest** and **K-Nearest Neighbor (KNN)** algorithms, the data is standardized using StandardScaler in the Random Forest model to ensure all features are on the same scale. This is important to prevent differences in feature scales from affecting the model's performance. For KNN, data standardization is performed because the KNN method is sensitive to the distance between data points and requires features with uniform scales.

Data balancing is checked to ensure the proportions between classes, particularly in the nutritional status category, are balanced. If imbalances are found, the **SMOTE** technique is applied to balance the data. The **Random Forest** and **K-Nearest Neighbor** algorithms are applied to build classification models. The accuracy of these models is compared using evaluation metrics such as accuracy, precision, recall, and F1-score. The model with the best performance is selected for further analysis.

The conclusion of this study includes a comparison of model performance and recommendations for child health interventions based on the research findings. The final report is prepared to provide valuable insights to the Bandarharjo Health Center and other relevant stakeholders to improve child health in the area. The **research flowchart** used in this study will be illustrated in Fig. 1.

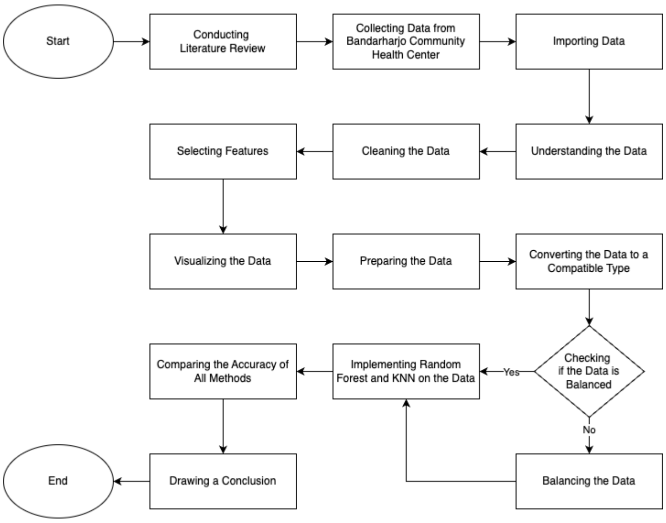
Start
Conducting Literature Review
Collecting Data from Bandarharjo Community Health Center
Importing Data
Understanding the Data
Cleaning the Data
Selecting Features
Visualizing the Data
Preparing the Data
Converting the Data to a Compatible Type
Checking if the Data is Balanced
Yes
No
Implementing Random Forest and KNN on the Data
Comparing the Accuracy of All Methods
Balancing the Data
Drawing a Conclusion
End

Fig. 1. Research Flowchart.

III. Result and Discussion

A. Exploratory Data Analysis (EDA)

The dataset used in this study was obtained from Bandarharjo Public Health Center, encompassing health data of children under five years old. After a data cleaning process to remove duplicates, the final dataset consisted of 3,674 entries. The analyzed variables include the following main features:

- **Weight (BB):** The child's weight in kilograms.
- **Height (TB):** The child's height in centimeters.
- **Age in Months:** The child's age in months.
- **Gender:** Male or female.
- **Nutritional Status (BB/U):** Nutritional status category based on weight-for-age.
- **Nutritional Status (BB/TB):** Nutritional status category based on weight-for-height.
- **Weight Status:** Weight change categories, including increase, stable, or decrease.

The first step in the analysis involved visualizing feature distributions using boxplots, as shown in Fig. 2. The results revealed several extreme values in the *Weight (BB)* feature, particularly on the upper end of the distribution. Conversely,

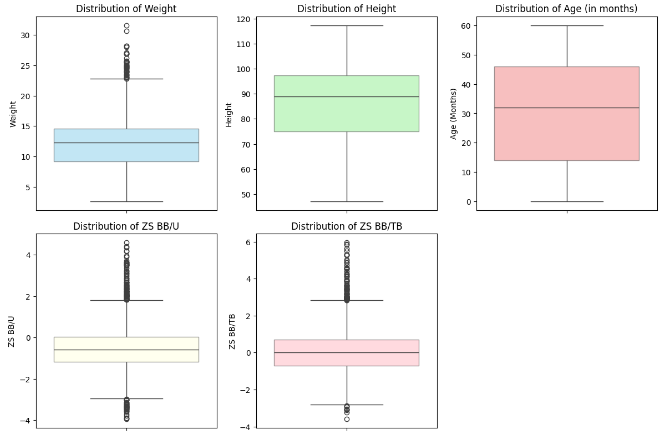

Fig. 2. Boxplot Distribution.

the distributions for *Height (TB)* and *Age in Months* were more consistent, with no significant outliers detected.

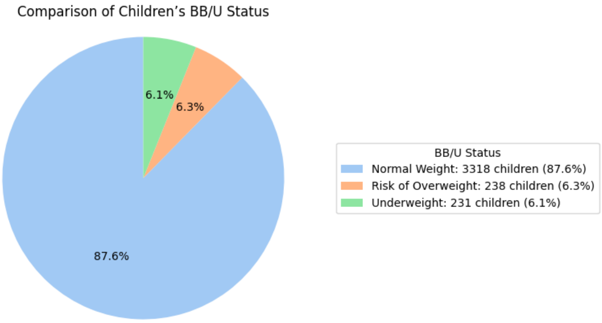

Fig. 3. Distribution of Children's BB/U Status.

The distribution of categories in nutritional status highlighted a significant imbalance. For *BB/U*, as presented in Fig. 3, 87.6% of children fell into the **Normal** category, with the **Risk of Overweight** and **Underweight** categories comprising only a small portion of the data.

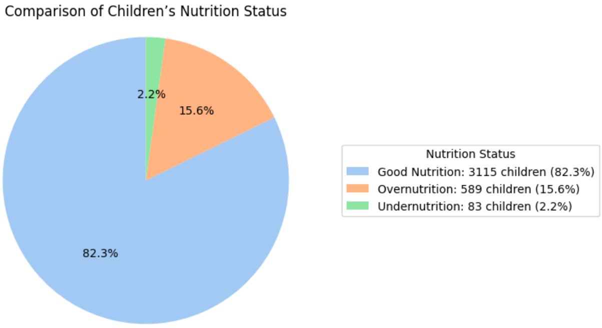

Fig. 4. Distribution of Children's Nutrition Status.

Similarly, for *BB/TB* based on BB/TB in Fig. 4, the **Good Nutrition** category dominated with 82.3% of the total data, while the **Poor Nutrition** category accounted for just 2.2%.

Gender distribution showed, as presented in Fig. 5, that 54.7% of the data were male, while 45.3% were female. Meanwhile, the distribution of *Weight Status* revealed that most children fell into the **Weight Increase** category (3,138 entries), followed by **Weight Decrease** (579 entries), **Weight Stable** (54 entries), and other categories (16 entries), as shown in Fig. 6.

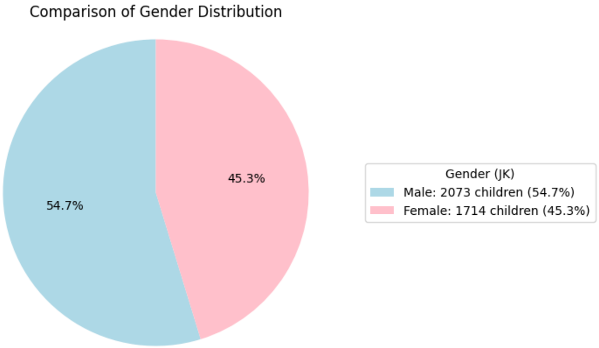

Fig. 5. Comparison of Gender Distribution.

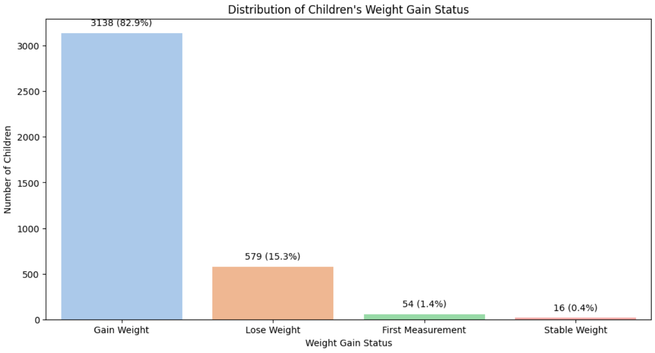

Fig. 6. Comparison of Weight Gain Status.

B. Data Preprocessing

The significant class imbalance in the data necessitated preprocessing steps before modeling. In this study, the *BB/U* and *BB/TB* columns were retained in multiclass format for classification analysis using *Random Forest* and *K-Nearest Neighbors (KNN)*. This approach allowed the model to identify each nutritional status category without converting it into binary classes.

To address class imbalance, the *Synthetic Minority Oversampling Technique (SMOTE)* was applied. Before SMOTE, the distribution ratio in the *BB/U* column indicated a dominant **Normal** class with a ratio of approximately 7:1 compared to the **Risk of Overweight** and **Underweight** classes. Similarly, for the *BB/TB* column, the **Good Nutrition** class dominated with a ratio of 5:1 compared to the **Overnutrition** and **Undernutrition** classes.

TABLE I
CLASS DISTRIBUTION OF BB/U BEFORE AND AFTER SMOTE

BB/U Nutrition Status	**Before SMOTE**	**After SMOTE**
Normal	3,269	2,611
Risk of Overweight	238	2,611
Underweight	231	2,611

After applying SMOTE, the class distribution became more balanced. For *BB/U*, as shown in Table I, each category (**Normal**, **Risk of Overweight**, **Underweight**) had the same number of entries, 2,611. A similar distribution occurred in

TABLE II
CLASS DISTRIBUTION OF BB/TB BEFORE AND AFTER SMOTE

BB/TB Nutrition Status	Before SMOTE	After SMOTE
Good Nutrition	3,068	2,448
Overnutrition	587	2,448
Undernutrition	83	2,448

TABLE III
KNN CLASSIFICATION REPORT FOR BB/U

BB/U Nutrition Status	Precision	Recall	F1-Score
Underweight	0.90	0.65	0.76
Normal	0.96	0.99	0.97
Risk of Overweight	0.94	0.68	0.79

BB/TB, where each category (**Good Nutrition**, **Overnutrition**, **Undernutrition**) had 2,448 entries, as shown in Table II. With this balanced distribution, the data was then split into *training* and *testing sets*, followed by normalization to ensure consistency and effectiveness during model training.

This preprocessing step ensured that the model could accurately and fairly identify all nutritional status categories, reducing potential biases or misclassifications often caused by extreme class imbalances. Consequently, methods such as *Random Forest* and *KNN* could be utilized optimally to produce reliable models.

C. Random Forest and K-Nearest Neighbor Implementation

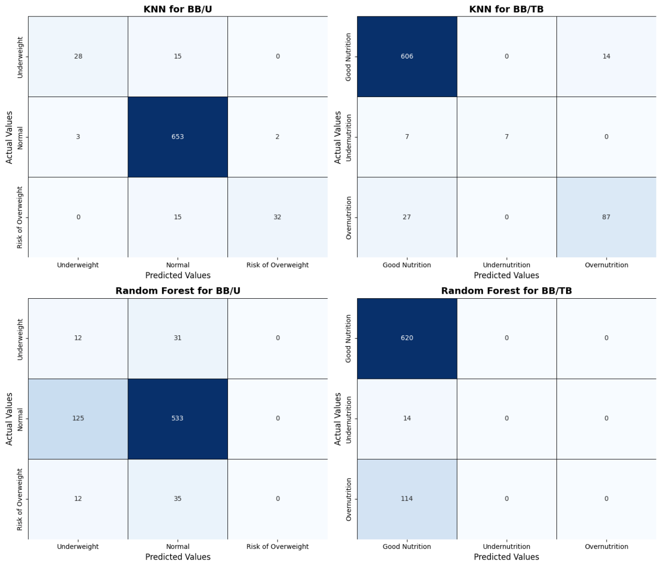

Fig. 7. KNN and RF Confusion Matrix.

This study implements two machine learning algorithms, namely K-Nearest Neighbors (KNN) and Random Forest (RF), without using the Synthetic Minority Over-sampling Technique (SMOTE), to classify two target variables, namely BB/U (weight-for-age) and BB/TB (weight-for-height). The features used include weight, height, age in months, gender, and weight gain. Before implementing the models, the data is divided into two parts: 80% for training data and 20% for testing data. To ensure optimal model performance, data standardization is performed using the StandardScaler to ensure uniform scaling across all features. The implementation of the models is carried out on the scaled data for each algorithm, KNN and RF.

1) K-Nearest Neighbor (KNN): In the KNN implementation, as presented in Table III, hyperparameter tuning is performed using grid search with 5-fold cross-validation to find the best parameters. A total of 16 parameter combinations were tested for each dataset, resulting in 80 fitting models. For the BB/U dataset, the best parameters obtained are **`n_neighbors = 3`**, **`p = 2`** and **`weights = 'distance'`**, which results in a model accuracy of 95.32%. In the classification report, the **Normal** class shows the best performance with a precision of 96%, recall of 99%, and f1-score of 97%. However, the **Underweight** and **Risk of Overweight** classes show lower performance, with recall values of only 65% and 68%, respectively. The confusion matrix shows that many misclassifications occur in the **Underweight** class, which is often classified as **Normal**.

TABLE IV
KNN CLASSIFICATION REPORT FOR BB/TB

BB/TB Nutrition Status	Precision	Recall	F1-Score
Good Nutrition	0.95	0.98	0.96
Undernutrition	1.00	0.50	0.67
Overnutrition	0.86	0.76	0.81

For the BB/TB dataset, as presented in Table IV, the best parameters obtained are the same as those for BB/U, with a model accuracy of 93.58%. The **Good Nutrition** class dominates with a precision of 95% and recall of 98%, but for the **Undernutrition** and **Overnutrition** classes, model performance significantly declines, especially in the recall of the **Undernutrition** class, which only reaches 50%. The confusion matrix indicates that many instances in the **Undernutrition** and **Overnutrition** classes are misclassified as **Good Nutrition**. Overall, the KNN implementation without SMOTE shows good performance in the majority classes, such as **Normal** for BB/U and **Good Nutrition** for BB/TB, but its performance is suboptimal for the minority classes.

TABLE V
RANDOM FOREST CLASSIFICATION REPORT FOR BB/U

BB/U Nutrition Status	Precision	Recall	F1-Score
Underweight	0.08	0.28	0.12
Normal	0.89	0.81	0.85
Risk of Overweight	0.00	0.00	0.00

2) Random Forest (RF): Similar to KNN, the implementation of Random Forest also involved hyperparameter tuning using grid search and 3-fold cross-validation to find the best parameter combinations. A total of 162 parameter combinations were tested for each dataset, resulting in 486 model fits.

For the **BB/U** dataset, as presented in Table V, the tuning results indicated that the best parameter combination was **`'bootstrap': False`**, **`'max_depth': 20`**, **`'min_samples_leaf': 1`**, **`'min_samples_split': 2`**, and **`'n_estimators': 200`**. This model achieved an accuracy of 72.86%, but its performance on minority classes remained low. According to the classification report, the **Normal** class showed better performance with a precision of 89% and a recall of 81%. In contrast, the **Underweight** class had a recall of only 28%

and a very low precision of 8%. The **Risk of Overweight** class could not be classified well, with precision, recall, and f1-score values all at 0%. The confusion matrix revealed that most misclassifications occurred in the **Underweight** and **Risk of Overweight** classes, which were often misclassified as the majority class, **Normal**.

TABLE VI
RANDOM FOREST CLASSIFICATION REPORT FOR BB/TB

BB/TB Nutrition Status	Precision	Recall	F1-Score
Good Nutrition	0.83	1.00	0.91
Undernutrition	0.00	0.00	0.00
Overnutrition	0.00	0.00	0.00

For the **BB/TB** dataset, as presented in Table VI, the best parameter combination was **`'bootstrap': True`**, **`'max_depth': None`**, **`'min_samples_leaf': 1`**, **`'min_samples_split': 2`**, and **`'n_estimators': 200`**. This model demonstrated higher accuracy at 82.89%. However, the model's performance on minority classes remained problematic. The **Good Nutrition** class dominated the classification results with a precision of 83% and a recall of 100%. In contrast, the **Undernutrition** and **Overnutrition** classes could not be classified at all, with precision, recall, and f1-score values all at 0%. The confusion matrix showed that all instances from the **Undernutrition** and **Overnutrition** classes were misclassified as **Good Nutrition**.

The results of the Random Forest implementation without SMOTE indicate that the model tends to be biased toward the majority class, such as **Normal** for the BB/U dataset and **Good Nutrition** for the BB/TB dataset. Although the overall accuracy appears relatively high, the model fails to recognize patterns in the minority classes, resulting in imbalanced classification performance.

D. SMOTE Implementation

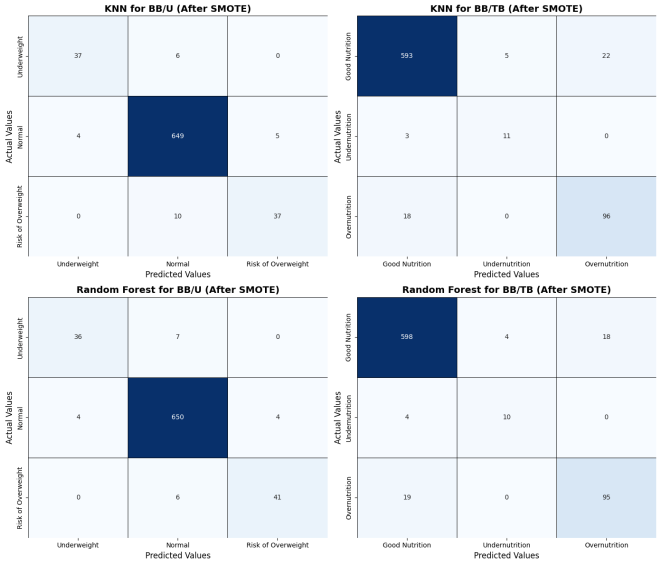

Fig. 8. KNN and RF Confusion Matrix using SMOTE.

To address the class imbalance issues identified in the BB/U and BB/TB datasets, the SMOTE (Synthetic Minority Over-sampling Technique) technique was applied. SMOTE is an over-sampling technique that generates synthetic samples based on the existing data [10]. The primary goal of applying SMOTE is to improve the representation of minority classes in the training dataset, allowing the model to better predict these classes [11]. The identified minority classes in the BB/U dataset are **Risk of Overweight** and **Underweight**, while in the BB/TB dataset, they are **Undernutrition** and **Overnutrition**. This technique works by increasing the number of samples in the minority classes through interpolation of existing data, resulting in synthetic data that better represents these classes.

TABLE VII
KNN CLASSIFICATION REPORT FOR BB/U (AFTER SMOTE)

BB/U Nutrition Status	Precision	Recall	F1-Score
Underweight	0.90	0.86	0.88
Normal	0.98	0.99	0.98
At Risk of Overweight	0.88	0.79	0.83

1) KNN Implementation with SMOTE: In the previous implementation, the KNN model showed high accuracy for the majority class but performed suboptimally on the minority classes, as presented in Table VII. The application of SMOTE successfully enhanced the model's ability to recognize patterns in the minority classes. The results of hyperparameter tuning for the BB/U dataset showed the following optimal parameters: **`metric='euclidean'`**, **`n_neighbors=1`**, and **`weights='uniform'`**, with the highest cross-validation accuracy of 99.20%. The model achieved an accuracy of 96.66% on the test data. After applying SMOTE, recall for the minority classes such as **Underweight** and **Risk of Overweight** increased to 86% and 79%, respectively, while the majority class maintained high performance with precision and recall of 98% and 99%, respectively. The confusion matrix showed a reduction in classification errors for the minority classes compared to the implementation without SMOTE.

TABLE VIII
KNN CLASSIFICATION REPORT FOR BB/TB (AFTER SMOTE)

BB/TB Nutrition Status	Precision	Recall	F1-Score
Good Nutrition	0.97	0.96	0.96
Undernutrition	0.69	0.79	0.73
Overnutrition	0.81	0.84	0.83

Similarly, for the BB/TB dataset, as presented in Table VIII, with the same parameters, the KNN model achieved a cross-validation accuracy of 98.19% and a test accuracy of 93.58%. Significant improvements were also observed in the **Undernutrition** and **Overnutrition** classes, with recall rates of 79% and 84%, indicating that the model was able to more accurately recognize samples from the minority classes.

TABLE IX
RANDOM FOREST CLASSIFICATION REPORT FOR BB/U (AFTER SMOTE)

BB/U Nutrition Status	Precision	Recall	F1-Score
Underweight	1.00	0.84	0.91
Normal	0.98	0.99	0.99
At Risk of Overweight	0.91	0.85	0.88

2) Random Forest Implementation with SMOTE: In the Random Forest model, which previously exhibited suboptimal performance due to data imbalance, the application of SMOTE yielded very positive results. SMOTE helped improve the distribution of samples across classes and enhanced the model's ability to recognize patterns in the minority classes, as presented in Table IX.. After applying SMOTE and performing hyperparameter tuning, the best parameters for the BB/U dataset were: **`bootstrap=False`**, **`max_depth=20`**, **`min_samples_leaf=1`**, **`min_samples_split=5`**, and **`n_estimators=200`**, resulting in a cross-validation accuracy of 97.59%. The model achieved an accuracy of 97.85% on the test data, with a significant increase in precision and recall for the minority classes such as **Underweight** and **Risk of Overweight**. Precision for both minority classes increased to 100%, while recall rates reached 84% and 85%, respectively.

TABLE X
RANDOM FOREST CLASSIFICATION REPORT FOR BB/TB (AFTER SMOTE)

BB/TB Nutrition Status	Precision	Recall	F1-Score
Good Nutrition	0.97	0.97	0.97
Undernutrition	0.75	0.86	0.80
Overnutrition	0.86	0.84	0.85

For the BB/TB dataset, as presented in Table X, the optimal parameters were **`bootstrap=False`**, **`max_depth=20`**, **`min_samples_leaf=1`**, **`min_samples_split=2`**, and **`n_estimators=50`**, achieving a cross-validation accuracy of 94.79%. The model achieved an accuracy of 94.79% on the test data, with significant improvements in the minority classes **Undernutrition** (recall 86%) and **Overnutrition** (precision 86% and recall 84%).

The application of SMOTE improved both the KNN and Random Forest models, especially in predicting minority classes. Before SMOTE, both models had high accuracy for the majority class but low recall for the minority classes. After applying SMOTE, recall for the minority classes increased significantly. Random Forest with SMOTE outperformed KNN, showing higher accuracy and better ability to identify patterns in minority classes without affecting performance on the majority class, making it a more effective solution for data imbalance in this study.

IV. CONCLUSION

The application of the SMOTE (Synthetic Minority Oversampling Technique) method has proven effective in addressing class imbalance in the child nutritional status dataset, which includes BB/U (Weight for Age) and BB/TB (Weight for Height). This technique successfully improved the performance of both the K-Nearest Neighbors (KNN) and Random Forest models, particularly in increasing recall for the minority classes, which were previously underrepresented. The KNN model showed significant recall improvements, with recall reaching 86% for the ”Underweight” class and 79% for the ”Risk of Overweight” class in the BB/U dataset, and 79% for the ”Undernutrition” class and 84% for the ”Overweight” class in the BB/TB dataset. The test accuracy of KNN after applying SMOTE was 96.66% for BB/U and 93.58% for BB/TB.

After applying SMOTE, the Random Forest model showed more optimal performance compared to KNN, with significant improvements in precision and recall, particularly in the BB/U dataset, and 86% for the ”Undernutrition” class and 84% for the ”Overweight” class in BB/TB. The cross-validation accuracy for Random Forest reached 97.59% for BB/U and 94.79% for BB/TB, demonstrating the model's excellent ability to balance the classification of both majority and minority classes.

The use of two target columns (BB/U and BB/TB) provided a more comprehensive picture of children's nutritional status, as each column measures a different aspect of nutritional status, thus improving the accuracy of malnutrition detection. Additionally, features such as gender and previous weight status also play an important role in predicting children's nutritional status. Gender affects physical growth patterns, while previous weight status serves as a key indicator for predicting future malnutrition risk.

Overall, the application of SMOTE successfully addressed class imbalance and improved the model's ability to detect minority classes without reducing the accuracy of majority classes. The Random Forest model proved more effective in balancing classification between majority and minority classes, producing excellent precision and recall, as well as high cross-validation accuracy. Therefore, the use of two target columns, BB/U and BB/TB, along with appropriate feature selection, can improve model accuracy in classifying children's nutritional status and support more targeted nutritional interventions. The SMOTE technique, combined with relevant feature selection, enhances the model's performance in classifying children's nutritional status, ultimately supporting more effective nutritional intervention programs.

REFERENCES

[1] A. Amirullah, A. T. A. Putra, and A. A. D. Al Kahar, “Deskripsi status gizi anak usia 3 sampai 5 tahun pada masa covid-19,” *Murhum: jurnal pendidikan anak usia dini*, vol. 1, no. 1, pp. 16–27, 2020.

[2] U. Ramlah, “Gangguan kesehatan pada anak usia dini akibat kekurangan gizi dan upaya pencegahannya,” *Ana'Bulava: Jurnal Pendidikan Anak*, vol. 2, no. 2, pp. 12–25, 2021.

[3] I. Susanti, R. Pambayun, and F. Febry, “Description of nutritional status and many factors that influence the nutritional status of children who have 2-5 years old in farmer's family at pelangki muaradua, south oku,” *Jurnal Ilmu Kesehatan Masyarakat*, vol. 3, no. 2, pp. 96–107, 2012.

[4] A. Ahmad, A. Latief *et al.*, “Perbandingan metode knn dan lbph pada klasifikasi daun herbal,” *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 5, no. 3, pp. 557–564, 2021.

[5] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, “Prediction of covid-19 confirmed, death, and cured cases in india using random forest model,” *Big Data Mining and Analytics*, vol. 4, no. 2, pp. 116–123, 2021.

[6] Y. Religia, A. Nugroho, W. Hadikristanto *et al.*, “Analisis perbandingan algoritma optimasi pada random forest untuk klasifikasi data bank marketing,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 187–192, 2021.

[7] A. M. Argina, “Penerapan metode klasifikasi k-nearest neigbor pada dataset penderita penyakit diabetes,” *Indonesian Journal of Data and Science*, vol. 1, no. 2, pp. 29–33, 2020.

[8] S. Diansyah, “Klasifikasi tingkat kepuasan pengguna dengan menggunakan metode k-nearest neighbors (knn),” *Jurnal Sistim Informasi Dan Teknologi*, pp. 7–12, 2022.

[9] R. Chairunisa, W. Astuti *et al.*, “Perbandingan cart dan random forest untuk deteksi kanker berbasis klasifikasi data microarray,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 5, pp. 805–812, 2020.

[10] M. Mukherjee and M. Khushi, “Smote-enc: A novel smote-based method to generate synthetic data for nominal and continuous features,” *Applied system innovation*, vol. 4, no. 1, p. 18, 2021.

[11] A. J. Mohammed, M. Muhammed Hassan, and D. Hussein Kadir, “Improving classification performance for a novel imbalanced medical dataset using smote method,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3161–3172, 2020.