Bulletin of Informatics and Data Science
Vol. 3 No. 2, November 2024, Page 54−62
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v3i2.93
https://ejurnal.pdsi.or.id/index.php/bids/index

Hybrid Gradient Boosting and SMOTE-ENN for Toddler Nutritional Status Classification on Imbalanced Data

Alfry Aristo Jansen Sinlae[1,*], Moh. Erkamim[2], Farid Fitriyadi[3], Lilik Suhery[4] , Rachmat Destriana[5]

[1] Computer Science Study Program, Faculty of Engineering, Universitas Katolik Widya Mandira, Kupang, Indonesia
[2] Smart City Information System Study Program, Universitas Tunas Pembangunan Surakarta, Surakarta, Indonesia
[3] Informatics Study Program, Faculty of Science, Technology, and Health, Universitas Sahid Surakarta, Surakarta, Indonesia
[4] Health Informatics Study Program, Faculty of Health and Science, Universitas Mercubaktijaya, Padang, Indonesia
[5] Faculty of Business Management and Information Technology, Universiti Muhammadiyah Malaysia, Padang Besar, Malaysia
Email: [1,*]alfry.aj@unwira.ac.id, [2]erkamim@lecture.utp.ac.id, [3]faridfitriyadi@gmail.com, [4]liliksuhery@gmail.com, [5]p52400016@student.umam.edu.my
Correspondence Author Email: alfry.aj@unwira.ac.id

Abstract

Stunting in toddlers remains a serious global health issue with long-term impacts on children's physical and cognitive development. One of the main challenges in classifying nutritional status is class imbalance, where the number of normal cases significantly exceeds that of minority classes such as stunted and severely stunted. This study aims to develop a hybrid approach by integrating the Gradient Boosting algorithm with the SMOTE-ENN (Synthetic Minority Oversampling Technique–Edited Nearest Neighbors) method to improve classification performance on imbalanced data. The dataset used was obtained from the Kaggle platform, consisting of 121,000 entries with four nutritional status categories. Data preprocessing included label encoding, numerical feature standardization, and stratified data splitting with an 80:20 ratio. The model was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The proposed hybrid model successfully increased the recall for the stunted class from 61.80% to 98.41%, and the F1-score from 71.93% to 83.58%. Overall accuracy improved from 92.39% to 93.35%, while the ROC-AUC score increased from 99.08% to 99.63%. These findings demonstrate that integrating Gradient Boosting with SMOTE-ENN is effective in enhancing sensitivity to minority classes and improving overall multi-class classification performance.

Keywords: Stunting Classification; Imbalanced Data; Gradient Boosting; SMOTE-ENN; Toddler Nutrition

1. INTRODUCTION

Stunting, a condition of impaired growth in children caused by chronic malnutrition, remains a significant global health concern. According to the latest report by UNICEF, WHO, and the World Bank (2023), approximately 148 million children under the age of five are affected by stunting worldwide, with the highest prevalence reported in South Asia and Sub-Saharan Africa [1]. Stunting affects not only physical development but also cognitive abilities, increases susceptibility to illness, and reduces long-term productivity [2]. Therefore, early detection of nutritional status, particularly in identifying stunting cases, is a crucial step in supporting data-driven health interventions.

Despite its importance, the process of classifying nutritional status presents several challenges. One of the main issues is the imbalance in class distribution, where the number of children with normal nutritional status significantly exceeds those who are classified as stunted or severely stunted [3]. In addition, the quality of survey data in many developing countries is often compromised by the presence of outliers, missing values, and inconsistencies in field data collection [4]. If not properly addressed, these conditions can lead to classification models that are biased toward the majority class and unable to detect important minority cases that require urgent attention.

With the advancement of artificial intelligence and data mining, many studies have explored machine learning approaches to tackle the challenges in stunting classification. For example, a study by Yunus et al. (2023) utilized the C4.5 algorithm and achieved an accuracy of 61.82 percent with an AUC of 0.584, but did not incorporate boosting or data balancing methods [5]. Alita et al. (2024) employed a Decision Tree algorithm and achieved 83.26 percent accuracy, although they did not explicitly address the class imbalance issue [6]. Another study by Azani et al. (2024) compared the performance of Naive Bayes Classifier (NBC) and Support Vector Machine (SVM), finding that SVM outperformed NBC with 91 percent accuracy compared to 80% [7]. However, this study only handled binary classification (stunted and not stunted), which does not represent the complexity of multi-class classification problems. In contrast, Ma'muriyah et al. (2024) proposed the use of XGBoost combined with imputation and outlier detection techniques and achieved an accuracy of up to 95% [8]. Nevertheless, their primary focus was on data quality enhancement rather than addressing class imbalance.

From the analysis of previous studies, it is evident that there is a gap in the integration of boosting algorithms with data balancing strategies for multi-class classification of toddler nutritional status. This research offers a different approach by combining Gradient Boosting with a data balancing method known as SMOTE ENN (Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbors) to address this challenge. Gradient Boosting is well regarded for its effectiveness in a wide range of classification tasks due to its capability to build a strong predictive model from a series of weak learners using a sequential learning process [9]. However, this algorithm often becomes biased toward the majority class when applied to imbalanced datasets because it minimizes overall prediction error without considering the distribution of class labels [10]. To address this limitation, this study employs SMOTE ENN, a resampling method that applies oversampling to minority classes and removes potentially noisy or ambiguous samples from the majority class using the Edited Nearest Neighbors technique. Unlike standard SMOTE, which only increases the number

Bulletin of Informatics and Data Science
Vol. 3 No. 2, November 2024, Page 54−62
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v3i2.93
https://ejurnal.pdsi.or.id/index.php/bids/index

of minority class samples, SMOTE ENN improves data distribution by simultaneously filtering out misleading examples, resulting in a cleaner and more balanced dataset [11]. This technique is considered appropriate for stunting classification tasks, which naturally involve imbalanced class distributions and a high risk of misclassification.

Based on the identified problems, the objective of this study is to develop a hybrid classification model that integrates Gradient Boosting and SMOTE ENN in order to improve the accuracy and sensitivity of nutritional status classification for toddlers in imbalanced datasets. This research contributes to the field by introducing a combined approach of boosting and resampling methods, which has rarely been implemented together in the context of multi-class nutritional status classification.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This research follows a systematic framework designed to guide the researcher through the process of designing, implementing, and evaluating each stage of the study. A structured methodology is required to provide a clear workflow throughout the research process [12]. The sequence of activities can be illustrated in Figure 1.

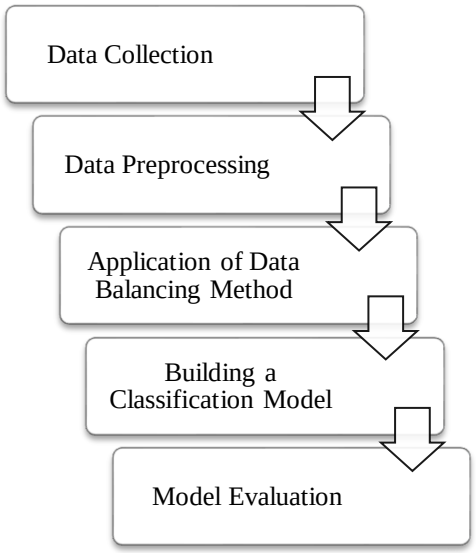


Figure 1. Research Flow

Based on Figure 1, the following provides a detailed explanation of each stage carried out in the study.

1) Data Collection
   The data collection phase marks the initial stage of the study, aimed at gathering data relevant to the analytical objectives. In this study, the dataset used is sourced from the Kaggle platform and titled "Stunting Toddler Detection" [13]. This dataset adopts a nutritional status classification approach based on z-score methods as recommended by the World Health Organization (WHO). The dataset primarily focuses on identifying stunting cases among children under five years old. In total, it contains 121,000 entries comprising three key features: age, gender, and height, along with a target label indicating the child's nutritional status. The nutritional status is categorized into four classes: stunted, severely stunted, normal, and tall.
2) Data Preprocessing
   Preprocessing was carried out to prepare the dataset in accordance with machine learning model requirements [14]. Categorical features such as Gender and Nutritional Status were converted into numerical format using label encoding, assigning values of 0 for female and 1 for male, and four numeric labels for nutritional status categories. Numerical features such as Age (in months) and Height (in cm) were standardized using the StandardScaler method to ensure a mean of zero and a standard deviation of one, thus preventing large-scale features from dominating the model and expediting the training process. Subsequently, the data was split into training and testing subsets in an 80:20 ratio using stratified sampling to maintain the proportional distribution of class labels in each subset. This ratio was chosen as it represents a common practice in machine learning, providing sufficient training data to build a representative model and enough test data for objective performance evaluation [15]. A bar chart visualization was used to depict the initial class distribution, illustrating the imbalance in sample counts across categories before any balancing method was applied.
3) Application of Data Balancing Method
   To address class imbalance, this study applied the SMOTE-ENN (Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbors) technique. SMOTE generates synthetic samples for minority classes by interpolating between nearest neighbors, while ENN removes samples from the majority class that are considered

Bulletin of Informatics and Data Science
Vol. 3 No. 2, November 2024, Page 54−62
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v3i2.93
https://ejurnal.pdsi.or.id/index.php/bids/index

noise or located near decision boundaries [11]. SMOTE-ENN is considered more effective than standard SMOTE because it not only increases the number of minority class samples but also reduces potential classification errors caused by ambiguous majority class data [16]. Applying SMOTE-ENN results in a more balanced class distribution, with increased representation of minority classes and reduced dominance of majority classes [17]. This enhances the model's ability to learn fairly across all classes.

4) Building a Classification Model
The main classification model employed in this study is the Gradient Boosting Classifier, an ensemble algorithm based on decision trees that incrementally builds a strong model from a set of weak learners by iteratively minimizing the loss function. Each new model in the Gradient Boosting sequence is trained to correct the prediction errors made by the previous model, thereby progressively increasing accuracy [18]. This model was chosen due to its ability to capture nonlinear relationships among features, its flexibility with various data types, and its competitive performance in classification tasks [19]. Furthermore, Gradient Boosting does not assume a particular data distribution and is robust when dealing with large datasets [20]. However, because the algorithm optimizes for overall prediction error, it tends to be biased toward the majority class when applied to imbalanced data [10]. To mitigate this limitation, the model is integrated with SMOTE-ENN as discussed in the previous stage to improve sensitivity to minority classes.
5) Model Evaluation
Evaluation was conducted to compare model performance on both imbalanced and balanced data (using SMOTE-ENN). The assessment began with the construction of a confusion matrix to display the number of correct and incorrect predictions for each class. From this matrix, key evaluation metrics were calculated, including accuracy, precision, recall, and F1-score per class. Precision measures the correctness of predictions for a specific class, recall measures the model's ability to identify actual instances of that class, and F1-score balances both, which is especially important in scenarios with class imbalance [21]. Additionally, the macro-averaged ROC-AUC score was used to evaluate the overall performance of the model in multi-class classification tasks, calculated as the average area under the ROC curve for each binarized class [22].

2.2 SMOTE-ENN Oversampling Technique

SMOTE-ENN (Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbors) is a hybrid data balancing method that combines both oversampling and undersampling strategies to address class imbalance in classification tasks [17]. This technique is designed to not only improve the representation of minority classes but also enhance the quality of data distribution by removing noise from the majority class [23].

It consists of two components: SMOTE and ENN. The first component, SMOTE, generates synthetic instances for the minority class by interpolating between an existing sample and one of its nearest neighbors. The synthetic sample x_{new} is calculated as shown in Equation (1).

$$x_{new} = x_i + \delta \times (x_{zi} - x_i) \quad (1)$$

where x_i is a randomly selected minority class sample, x_{zi} is one of its k-nearest neighbors, and δ is a random number between 0 and 1.

The second component, ENN (Edited Nearest Neighbors), serves as a cleaning mechanism. ENN removes any sample whose class label differs from the majority of its k nearest neighbors. This step aims to eliminate noisy or borderline samples, particularly from the majority class.

By combining these two steps, SMOTE-ENN results in a cleaner, more balanced, and more stable dataset that enables the classification model to better identify minority classes without being misled by irregularities from the majority class [24].

2.3 Gradient Boosting Method

Gradient Boosting is an ensemble learning algorithm that combines several weak learners, typically decision trees, in a sequential manner to produce a strong predictive model [25]. Introduced by Friedman (2001), this method builds upon boosting techniques by applying gradient descent to minimize a loss function iteratively [25]. Each successive model in the sequence is trained to correct the residual errors of the preceding model, thereby improving accuracy at each stage.

In general, Gradient Boosting works by fitting the model $F(x)$ to the target y through the minimization of a loss function $L\big(y, F(x)\big)$ using gradient descent. At the m^{th} iteration, the model is updated using Equation (2).

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (2)$$

where $F_{m-1}(x)$ is the model from the previous iteration, $h_m(x)$ is the new weak learner (such as a small decision tree) trained on the residuals, and γ_m s the learning rate or weight coefficient derived from minimizing the loss function.

The residuals $h_m(x)$, or gradient directions, are calculated using Equation (3).

$$r_{im} = -\left[\frac{\partial L\big(y_i, F(x_i)\big)}{\partial F(x_i)}\right]_{F(x)=F_{m-1(x)}} \quad (3)$$

Bulletin of Informatics and Data Science
Vol. 3 No. 2, November 2024, Page 54−62
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v3i2.93
https://ejurnal.pdsi.or.id/index.php/bids/index

di mana r_{im} adalah nilai residu atau arah gradien untuk sampel ke-i pada iterasi ke-mmm, $L(y_i, F(x_i))$ adalah fungsi kerugian (misalnya log-loss untuk klasifikasi atau MSE untuk regresi), y_i adalah nilai target aktual, $F(x_i)$ adalah prediksi model pada input x_i, $F_{m-1(x)}$ adalah output dari model sebelum iterasi ke- m.

Gradient Boosting mampu menangkap hubungan nonlinier yang kompleks dalam data dan beradaptasi dengan berbagai jenis fitur tanpa memerlukan transformasi khusus [26]. Keunggulannya terletak pada fleksibilitas, kemampuan generalisasi yang tinggi, dan performa yang kompetitif pada berbagai tugas klasifikasi maupun regresi.

3. RESULT AND DISCUSSION

The development of a nutritional status classification model for toddlers using the hybrid approach of Gradient Boosting and SMOTE-ENN began with the preparation of a relevant dataset. The dataset used was obtained from the Kaggle platform under the title “Stunting Toddler Detection” [13]. In total, the dataset consists of approximately 121,000 entries containing three primary features, namely age, gender, and height, along with one target label representing the nutritional status of toddlers. Nutritional status is categorized into four classes: stunted, severely stunted, normal, and tall.

Once the dataset was ready, the next step involved data preprocessing to ensure the dataset met the requirements of machine learning algorithms and to guarantee optimal input quality. The preprocessing stage started with the application of label encoding to categorical features such as gender and nutritional status, converting them into numerical representations. For instance, gender was encoded as 0 for female and 1 for male, while nutritional status was mapped into four numerical classes according to their respective categories. Numerical features such as age (in months) and height (in centimeters) were standardized using the StandardScaler method, which transforms the values to have a mean of zero and a standard deviation of one. This step is crucial to accelerate the training process and prevent dominance of larger scaled features in model learning. The distribution of numerical features before and after standardization is shown in Figure 2.

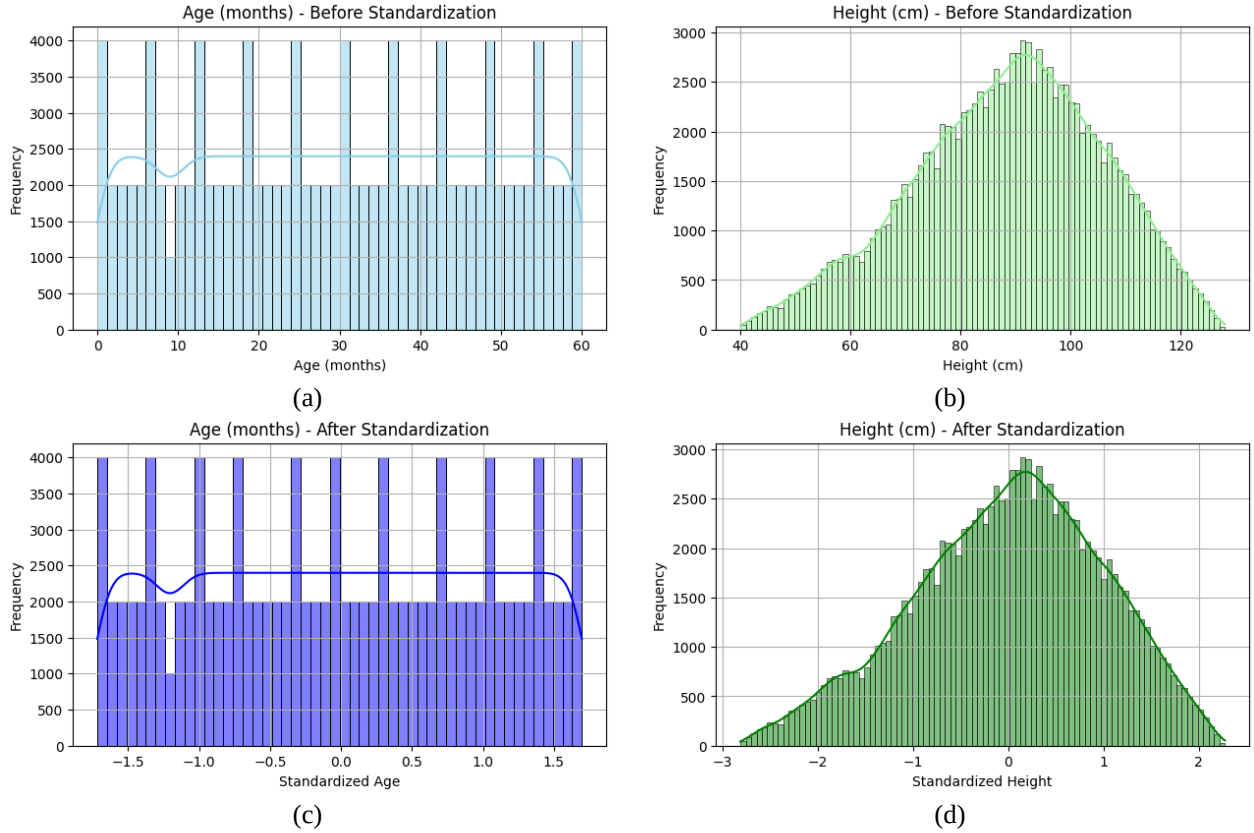


Figure 2. Distribution of Numerical Features Before and After Standardization: (a) Age Before, (b) Height Before, (c) Age After, and (d) Height After Standardization Using StandardScaler

The figure illustrates the distribution of age and height before and after standardization. Prior to standardization, age exhibits a uniform distribution due to monthly data recording, whereas height displays a near-normal distribution. After standardization, both features share a common scale while retaining their original distribution patterns. This ensures fair contribution of each feature in the training process.

The next step was to analyze the class distribution of the target variable to determine whether the data was balanced. This analysis is essential to understand the proportion of each nutritional status category, as imbalance can influence model bias and overall performance. It also serves as a foundation for selecting appropriate data balancing techniques such as oversampling or hybrid resampling. The class distribution for the target variable is shown in Figure 3.

Bulletin of Informatics and Data Science
Vol. 3 No. 2, November 2024, Page 54−62
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v3i2.93
https://ejurnal.pdsi.or.id/index.php/bids/index

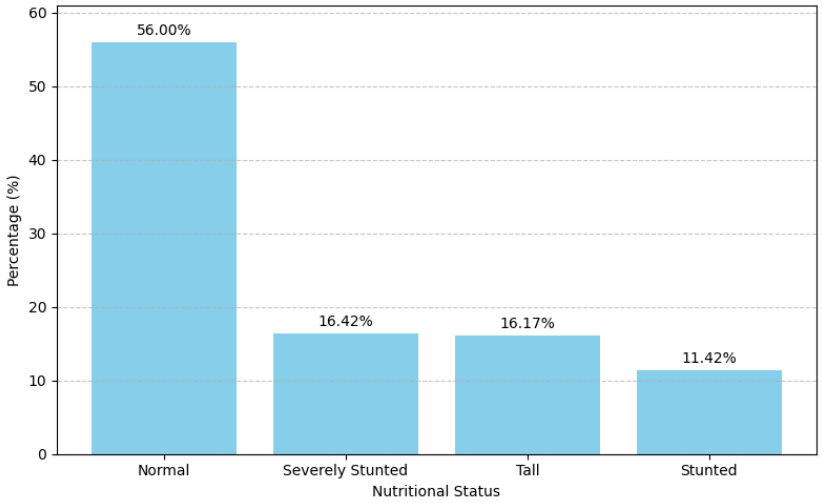


Figure 3. Distribution of Nutritional Status Classes in the Dataset

Figure 3 shows that the "Normal" class dominates the dataset with 56 percent, while the "Severely Stunted," "Tall," and "Stunted" classes are significantly underrepresented. This imbalance underscores the importance of applying data balancing methods to ensure fair and accurate model training. To illustrate the manual calculation of SMOTE ENN, a simplified example is used with only two classes: Normal (majority) and Stunted (minority), along with two numerical features: age (months) and height (cm). The initial sample is shown in Table 1.

Table 1. Initial Data Sample

ID	Age (months)	Height (cm)	Nutritional Status
1	12	75	Normal
2	14	77	Normal
3	13	76	Normal
4	24	65	Stunted
5	23	66	Stunted

Table 1 shows 5 samples with a class distribution of 3 data points labeled as Normal and 2 data points labeled as Stunted. This indicates that the data is imbalanced. To address this case study, the first step is to use SMOTE to generate synthetic data for the Stunted class through interpolation between nearest neighbors, where:

$x_i = (24,65)$
$x_{zi} = (23,66)$
$\delta = 0.6$ (random value between 0 and 1)

Calculate the synthetic sample using Equation (1), so the calculation is as follows:

$$x_{new} = (24,65) + 0.6 \times \big((23,66) - (24,65)\big)$$
$$= (24,65) + 0.6 \times (-1,1) = (23.6,65.6)$$
$$= (23.6,65.6)$$

After the SMOTE process generates new synthetic samples, for example, (23.4, 65.6, Stunted), the next step is to apply the Edited Nearest Neighbors (ENN) method. ENN is implemented after oversampling to filter out majority class instances that are considered inconsistent based on their local neighborhood. Specifically, ENN removes a sample from the majority class if the majority of its nearest neighbors belong to the minority class. For example, consider sample ID 2, which belongs to the Normal class and has three nearest neighbors: ID 3 (Normal), ID 5 (Stunted), and ID 6 (Stunted, which is a synthetic sample generated by SMOTE). Since two out of the three neighbors belong to the Stunted class, ID 2 is identified as ambiguous and is removed by the ENN procedure. This process helps to reduce noise around the decision boundary between classes and improves the overall quality of the training dataset. The final result after applying the SMOTE combined with ENN technique is presented in Table 2.

Table 2. Data Samples After Oversampling Using SMOTE-ENN

ID	Age (months)	Height (cm)	Nutritional Status
1	12	75	Normal
3	13	76	Normal
4	24	65	Stunted
5	23	66	Stunted
6	23.4	65.6	Stunted (synthetic)

Bulletin of Informatics and Data Science
Vol. 3 No. 2, November 2024, Page 54−62
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v3i2.93
https://ejurnal.pdsi.or.id/index.php/bids/index

The table shows that ENN successfully removed ambiguous majority-class instances while reinforcing the representation of the minority class. This produces a cleaner and more balanced dataset ready for classification model training. The updated class distribution after applying SMOTE ENN is visualized in Figure 4.

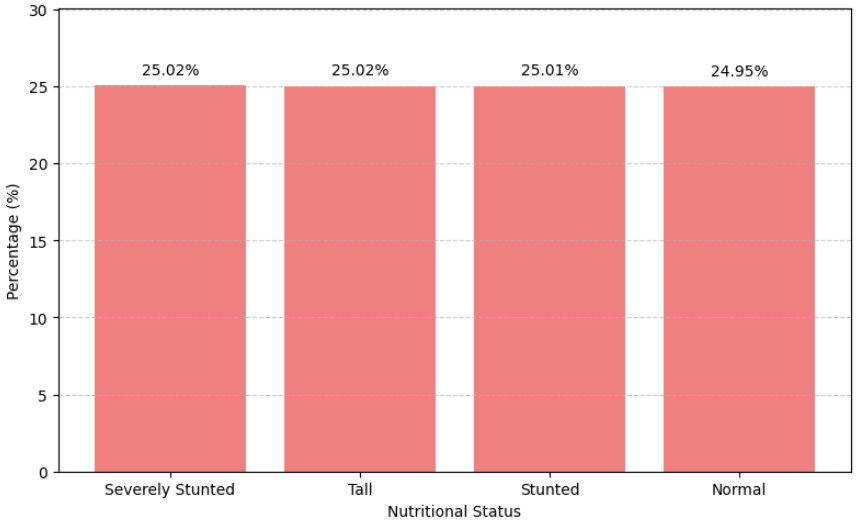
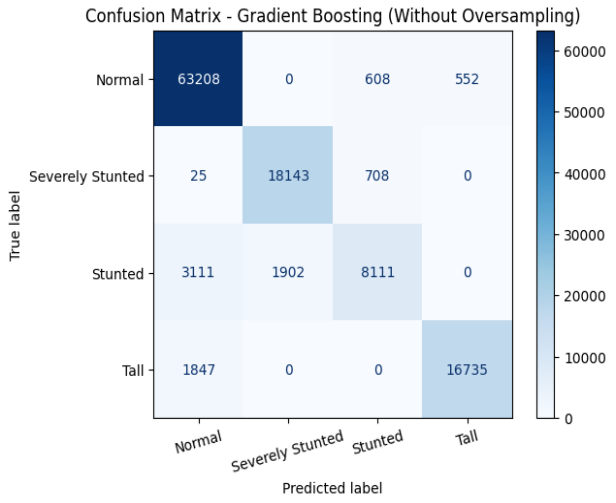


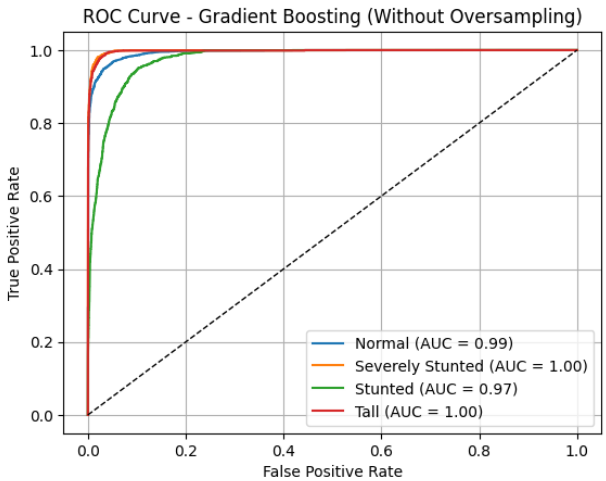
Figure 4. Distribution of Target Data After Oversampling Using SMOTE-ENN

Figure 4 shows that all four nutritional status categories, namely Severely Stunted, Tall, Stunted, and Normal, are now nearly equally represented, with each class comprising approximately 25 percent of the dataset. This demonstrates that SMOTE ENN was effective in balancing the data by adding synthetic samples to the minority classes and eliminating noisy instances from the majority class located near the decision boundaries. A balanced dataset enables the classification model to learn more equitably without favoring any particular class.

The next stage was to build a classification model using the Gradient Boosting algorithm for both training and testing. The dataset was split into training and testing sets using an 80 to 20 ratio with stratified sampling to preserve class proportions in both subsets. The Gradient Boosting model was then trained using the training data to identify patterns between features and target classes. Gradient Boosting is an ensemble learning technique that builds a strong predictive model by combining multiple weak learners, typically shallow decision trees. It operates in a forward stage-wise manner, where each subsequent tree is constructed to minimize the errors made by the previously built model. Specifically, the algorithm fits each new tree to the negative gradient of the loss function, which corresponds to the residual errors from the prior iteration. Through this iterative process, the model progressively improves its predictions by focusing more on the samples that were previously misclassified. The process continues until either the prediction error reaches a minimal value or the maximum number of iterations is completed. Model performance was evaluated using several metrics including confusion matrix, classification report, and ROC AUC score to assess its capability in multi-class classification. The confusion matrices and ROC curves for both scenarios (without oversampling and with SMOTE ENN) are shown in Figure 5.

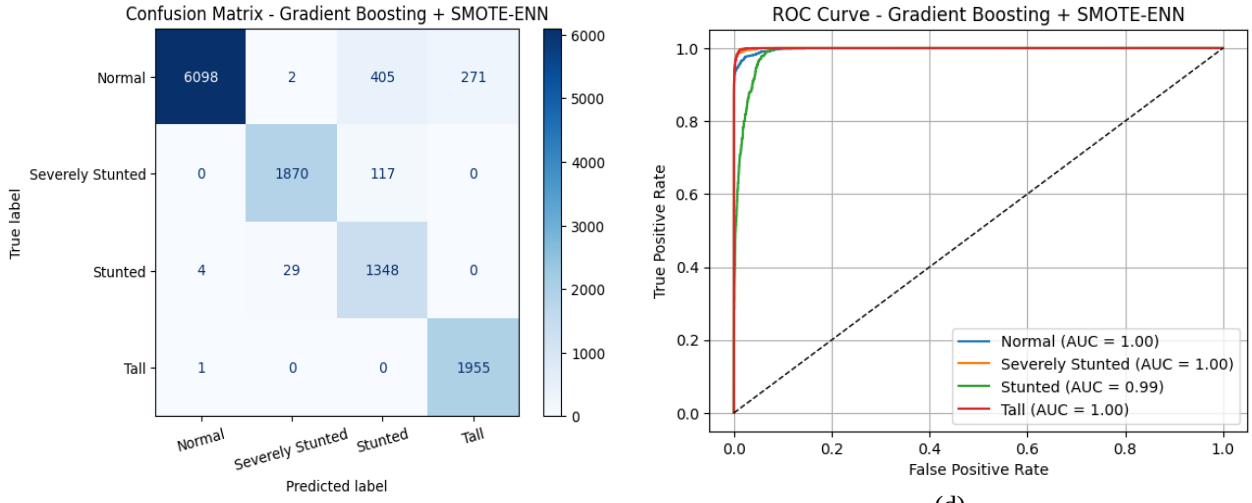


(a)



(b)

Bulletin of Informatics and Data Science
Vol. 3 No. 2, November 2024, Page 54−62
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v3i2.93
https://ejurnal.pdsi.or.id/index.php/bids/index



(c) (d)

Figure 5. (a) Confusion Matrix Without Oversampling, (b) ROC Curve Without Oversampling, (c) Confusion Matrix With SMOTE-ENN, and (d) ROC Curve With SMOTE-ENN

Figure 5 illustrates that SMOTE ENN improved the model's ability to classify minority classes such as Stunted and Severely Stunted. The confusion matrix in Figure 5(a) shows predictions are mostly concentrated in the majority classes (Normal and Tall), while in Figure 5(c), the predictions are more evenly distributed. Additionally, the ROC curve in Figure 5(d) shows improvement in AUC scores, particularly for the Stunted class, which increased from 0.97 to 0.99, reflecting improved model sensitivity.

Based on the confusion matrix and ROC results, classification metrics and ROC AUC scores were calculated for both models. The full comparison is presented in Table 3.

Table 3. Comparison of Performance Metrics Between Gradient Boosting and the Proposed Hybrid Model (Gradient Boosting + SMOTE-ENN)

Model	Class	Precision	Recall	F1-Score	Accuracy	ROC-AUC Score
Gradient Boosting	Normal	92.69%	98.20%	95.37%	92.39%	99.08%
	Severely Stunted	90.51%	96.12%	93.23%		
	Stunted	86.04%	61.80%	71.93%		
	Tall	96.81%	90.06%	93.31%		
Gradient Boosting + SMOTE-ENN	Normal	99.97%	90.17%	94.82%	93.35%	99.63%
	Severely Stunted	98.73%	94.21%	96.42%		
	Stunted	72.64%	98.41%	83.58%		
	Tall	87.98%	99.90%	93.56%		

Table 3 shows that the hybrid model significantly improves performance, especially for the Stunted class, where recall increased from 61.80 percent to 98.41 percent and F1-score from 71.93 percent to 83.58 percent. This demonstrates that SMOTE ENN enhances the model's sensitivity to underrepresented classes.

The main advantage of the hybrid model is its effectiveness in addressing class imbalance. SMOTE ENN not only increases minority-class representation but also eliminates noisy samples from the majority class, producing a cleaner and more balanced dataset. This contributes to improved overall accuracy from 92.39 percent to 93.35 percent and increased ROC AUC score from 99.08 percent to 99.63 percent, indicating better multi-class classification performance.

Nevertheless, the hybrid model shows a decline in precision for the Stunted class (from 86.04 percent to 72.64 percent). This could be due to the generation of overly similar synthetic samples, which may lead to false positives. To improve this, future research can explore advanced post-oversampling filtering techniques or use alternative boosting algorithms such as LightGBM or CatBoost with class weight adjustments. Intensive hyperparameter tuning may also help improve overall model performance.

4. CONCLUSION

This study successfully developed a hybrid model by combining Gradient Boosting with the SMOTE-ENN data balancing technique, which proved to be effective in enhancing the classification performance of toddler nutritional status on imbalanced data. The proposed hybrid model significantly improved the recall for the Stunted class from 61.80% to 98.41%, and increased the F1-score from 71.93% to 83.58%, indicating greater sensitivity toward minority classes.

Additionally, the overall accuracy of the model increased from 92.39% to 93.35%, accompanied by an improvement in ROC-AUC score from 99.08% to 99.63%. These findings demonstrate that the integration of SMOTE-ENN not only improves data distribution but also enhances the model's ability to classify all categories more fairly. However, the decrease in precision for the minority class suggests a potential increase in false positives due to the introduction of synthetic data. Therefore, future research is encouraged to explore alternative method combinations such as class weighting, adaptive undersampling, or the use of other boosting algorithms like LightGBM or CatBoost with more intensive hyperparameter optimization to achieve more stable and optimal results.

REFERENCES

[1] M. A. Ahmed, F. S. Duale, M. A. Ali, and A. M. Ibrahim, "Prevalence of Stunting and Associated Factors Among Under Five Years Children in Galkaio Town, Puntland, Somalia 2023: A cross-sectional study design," *J. Drug Deliv. Ther.*, vol. 14, no. 7, pp. 83–96, 2024, doi: 10.22270/jddt.v14i7.6699.

[2] B. Soetono and A. S. Barokah, "Trends in Stunting Prevalence Reduction: an Examination of Data Toward Achieving the 2024 Target in Indonesia," *Soc. Perspect. J.*, vol. 3, no. 1, pp. 51–68, 2024, doi: 10.53947/tspj.v3i1.795.

[3] E. Miranda, M. Aryuni, A. Y. Zakiyyah, Y. E. Kurniawati, A. V. D. Sano, and M. Kumbangsila, "An early prediction model for toddler nutrition based on machine learning from imbalanced data," *Procedia Comput. Sci.*, vol. 245, pp. 263–271, 2024, doi: 10.1016/j.procs.2024.10.251.

[4] S. Syahrial, R. Ilham, Z. F. Asikin, and S. S. I. Nurdin, "Stunting Classification in Children's Measurement Data Using Machine Learning Models," *J. La Multiapp*, vol. 3, no. 2, pp. 52–60, 2022, doi: 10.37899/journallamultiapp.v3i2.614.

[5] M. Yunus, M. K. Biddinika, and A. Fadlil, "Classification of Stunting in Children Using the C4.5 Algorithm," *JOIN (Jurnal Online Inform.*, vol. 8, no. 1, pp. 99–106, 2023, doi: 10.15575/join.v8i1.1062.

[6] D. Alita, I. Ahmad, R. J. Rumandan, M. Erkamim, and W. Widyasmoro, "Stunting Classification in Toddlers: Implementation and Evaluation of the Decision Tree Algorithm," in *International Conference on Information Technology and Computing (ICITCOM)*, IEEE, 2024, pp. 207–212. doi: 10.1109/ICITCOM62788.2024.10762254.

[7] N. W. Azani and M. Afdal, "Implementation of Naïve Bayes Classifier and Support Vector Machine for Stunting Classification," *Indones. J. Comput. Sci.*, vol. 13, no. 4, pp. 5079–5087, 2024, doi: 10.33022/ijcs.v13i4.4040.

[8] N. Ma'muriyah, E. Noersasongko, P. Purwanto, S. Winarno, and M. I. Ashiddiq, "XG Boost Based Data Imputation and Outlier Detection Methods for Classification of Stunting," in *International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, 2024, pp. 812–817. doi: 10.1109/ISRITI64779.2024.10963432.

[9] Abdullah-All-Tanvir, I. Ali Khandokar, A. K. M. Muzahidul Islam, S. Islam, and S. Shatabda, "A gradient boosting classifier for purchase intention prediction of online shoppers," *Heliyon*, vol. 9, no. 4, p. e15163, 2023, doi: 10.1016/j.heliyon.2023.e15163.

[10] M. H. L. Louk and B. A. Tama, "Revisiting Gradient Boosting-Based Approaches for Learning Imbalanced Data: A Case of Anomaly Detection on Power Grids," *Big Data Cogn. Comput.*, vol. 6, no. 2, 2022, doi: 10.3390/bdcc6020041.

[11] X. Wang *et al.*, "Diabetes mellitus early warning and factor analysis using ensemble Bayesian networks with SMOTE-ENN and Boruta," *Sci. Rep.*, vol. 13, no. 1, pp. 1–15, 2023, doi: 10.1038/s41598-023-40036-5.

[12] Parjito, I. Ahmad, R. I. Borman, A. D. Alexander, and Y. Jusman, "Combining Extreme Learning Machine and Linear Discriminant Analysis for Optimized Apple Leaf Disease Classification," in *International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, IEEE, 2024, pp. 138–143. doi: 10.1109/ICE3IS62977.2024.10775844.

[13] J. Dasilva, "Diabetes Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/johndasilva/diabetes

[14] R. I. Borman, F. Rossi, Y. Jusman, A. A. A. Rahni, S. D. Putra, and A. Herdiansah, "Identification of Herbal Leaf Types Based on Their Image Using First Order Feature Extraction and Multiclass SVM Algorithm," in *International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, IEEE, 2021, pp. 12–17.

[15] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, pp. 331–339, 2024, doi: 10.14569/IJACSA.2024.0150235.

[16] M. Muntasir Nishat *et al.*, "A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset," *Sci. Program.*, vol. 2022, pp. 1–17, 2022, doi: 10.1155/2022/3649406.

[17] J. Wang, "Prediction of postoperative recovery in patients with acoustic neuroma using machine learning and SMOTE-ENN techniques," *Math. Biosci. Eng.*, vol. 19, no. 10, pp. 10407–10423, 2022, doi: 10.3934/mbe.2022487.

[18] P. Nie, M. Roccotelli, M. P. Fanti, Z. Ming, and Z. Li, "Prediction of home energy consumption based on gradient boosting regression tree," *Energy Reports*, vol. 7, pp. 1246–1255, 2021, doi: 10.1016/j.egyr.2021.02.006.

[19] F. H. Yagin, I. B. Cicek, and Z. Kucukakcali, "Classification of stroke with gradient boosting tree using smote-based oversampling method," *Med. Sci. Int. Med. J.*, vol. 10, no. 4, p. 1510, 2021, doi: 10.5455/medscience.2021.09.322.

[20] D. A. Setyarini, A. A. M. D. Gayatri, C. S. K. Aditya, and D. R. Chandranegara, "Stroke Prediction with Enhanced Gradient Boosting Classifier and Strategic Hyperparameter," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 23, no. 2, pp. 477–490, 2024, doi: 10.30812/matrik.v23i2.3555.

[21] R. Rusliyawati, K. Karnadi, A. M. Tanniewa, A. C. Widyawati, Y. Jusman, and R. I. Borman, "Detection of Pepper Leaf Diseases Through Image Analysis Using Radial Basis Function Neural Networks," in *BIO Web of Conferences*, 2024, pp. 1–10. doi: 10.1051/bioconf/202414401005.

[22] Y. Liu, Y. Li, and D. Xie, "Implications of imbalanced datasets for empirical ROC-AUC estimation in binary classification tasks," *J. Stat. Comput. Simul.*, vol. 94, no. 1, pp. 183–203, Jan. 2024, doi: 10.1080/00949655.2023.2238235.

[23] M. Lamari *et al.*, "SMOTE–ENN-Based Data Sampling and Improved Dynamic Ensemble Selection for Imbalanced Medical Data Classification BT - Advances on Smart and Soft Computing," in *Advances on Smart and Soft Computing*, F. Saeed, T. Al-Hadhrami, F. Mohammed, and E. Mohammed, Eds., Singapore: Springer Singapore, 2021, pp. 37–49.

[24] S. Dhanalakshmi, S. Das, and R. Senthil, "Speech features-based Parkinson's disease classification using combined SMOTE-ENN and binary machine learning," *Health Technol. (Berl).*, vol. 14, no. 2, pp. 393–406, 2024, doi: 10.1007/s12553-023-00810-

x.
[25] M. S. Hosen and R. Amin, "Significant of Gradient Boosting Algorithm in Data Management System," *Eng. Int.*, vol. 9, no. 2, pp. 85–100, 2021, doi: 10.18034/ei.v9i2.559.
[26] G. W. Cha, H. J. Moon, and Y. C. Kim, "Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables," *Int. J. Environ. Res. Public Health*, vol. 18, no. 16, 2021, doi: 10.3390/ijerph18168530.