Enhancing Early Stunting Detection: A Novel Approach using Artificial Intelligence with an Integrated SMOTE Algorithm and Ensemble Learning Model

A.A.G. Yogi Pramana
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia
aagdeyogipramana@mail.ugm.ac.id

Muhammad Fazil Maulana
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia
muhammadfazilmaulana@mail.ugm.ac.id

Melvin Cahyadi Tirtayasa
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia
melvincahyaditirtayasa@mail.ugm.ac.id

Dyah Aruming Tyas
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia
dyah.aruming.t@ugm.ac.id

***Abstract*—The detection of stunting is a prominent issue in Indonesian healthcare concerning the Sustainable Development Goal of good health and well-being. Stunting poses a significant risk towards children below the age of five. If not treated, stunting may cause symptoms such as a lower cognitive function, lower productivity, a weakened immune system, delayed nerve development, and degenerative diseases. Particularly in regions where stunting is widespread and welfare resources are low, the challenge of detecting children that require treatment is of great importance. Many problems often arise in the diagnostic process, such as the lack of skilled man power to tackle the issue, the lack of experience in medical workers, incompatible anthropometric equipment, and an inefficient medical bureaucracy. For this problem, the implementation of artificial intelligence provides a transformative tool in enhancing medical diagnostic technology. This paper employs and compares the precision, recall, and the f-1 scores of Random Forest, Ada Boost, and Bagging as the performance measure. From the experiment, it is obtained that SMOTE-ENN using Ada Boost classifier and SMOTE using Bagging classifier has the highest F1-Score for minority classes namely stunted and stunting.**

Keywords— Stunting, SDGs, SMOTE, Ensemble Learning.

I. INTRODUCTION

Stunting, as characterized by the World Health Organization (WHO), is the impaired growth and development that children experience from poor nutrition, repeated infection, and inadequate psychosocial stimulation. Stunting not only hampers individual potential but also has broader implications for societal health.

The condition emerges as a considerable risk for children under the age of five, with potential repercussions that extend across various facets of health and development. If left untreated, stunting can cause a variety of symptoms, including lower cognitive function, diminished productivity, a compromised immune system, delayed nerve development, and an increased susceptibility to degenerative diseases.

Stunting is a common problem in many countries globally and occurs in 161 million children between 0 and 5[1]. In Indonesia, the prevalence of stunting in 2022 was 21.6%, marking a slight improvement from the 24.4% reported in 2021[2]. However, despite this improvement, Indonesia still lies amongst the top 5 countries in the world for cases of stunting, ranking second highest in southeast Asia.[3]

The detection of stunting poses a significant challenge within the realm of Indonesian healthcare. Additionally, in regions where stunting is widespread and welfare resources are low, the challenge of detecting children that require treatment is of even greater importance. By harnessing the power of artificial intelligence, machine learning models can sift through vast datasets to identify subtle patterns and indicators associated with stunting. This approach not only enhances the accuracy and efficiency of stunting detection but also facilitates the timely implementation of targeted interventions.

As per researching this topic, there were only few scientific articles exploring the idea of imbedding SMOTE oversampling methods to datasets about stunting. One study [4] discusses the idea of stunting detection using the Random Forest classification algorithm, in which they also test their model with various k-fold cross validation iterations. The highest average evaluation score they achieved was a 97.9%. Though more recently, papers such as [5] begin to introduce the techniques utilized in this paper. The paper written by Eko Prasetyo [5], incorporates ensemble-based learning methods to improve different classification algorithms in a model about stunting. Although their model didn't have as high evaluation scores, there was still a notable improvement in their results after implementing the Bagging algorithm to their classifiers.

What sets this paper apart is that we complement the Ensemble Machine Learning Model with SMOTE algorithms in the data preparation. With this, we are able to apply our model to data sets that are imbalanced. Thus, in this paper, we are not only comparing Random Forest, Ada Boost, and Bagging algorithms, we are also comparing the effects of applying a few different SMOTE algorithms in the data preparation stage as well.

II. METHODOLOGY

This paper uses the CRISP-DM (Cross Industry Standard Process for Data Mining) approach [7] which consists of 6

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00098

phases starting from Business Understanding, Data Understanding Phase, Data Preparation Phase, Modeling Phase, Evaluation Phase, and Deployment Phase. The process of CRISP-DM can be seen in Figure 1 below [8].

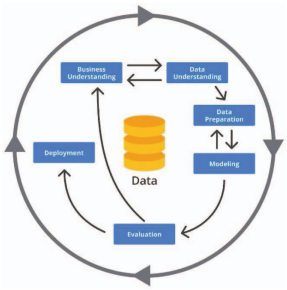

Figure 1. CRISP-DM Phase.

A. *Business Understanding*

The business objective of this research is to develop a more accurate and efficient approach for early stunting detection in children. The specific goals include improving accuracy, early detection and enhanced robustness.

B. *Data Understanding*

To be able to diagnose whether a child is stunted, stunting or normal, we look at how much the child's height/age deviates from the z-score for their age group. For *stunted* cases, their z-score falls more than 3 standard deviations below the median and for *stunting* cases, their z-score falls within 3 and 2 standard deviations from the median [6]. With this information, we can deduce what sorts of data is needed to draw conclusions for our model. The adequate data being the child's height, weight, age and gender.

The dataset utilized in this research is focused on the stunting status of children in Yogyakarta. This data collection process was done through collaboration with the local *posyandu* in which the organization provided key data in order to produce results using the Learning Model.

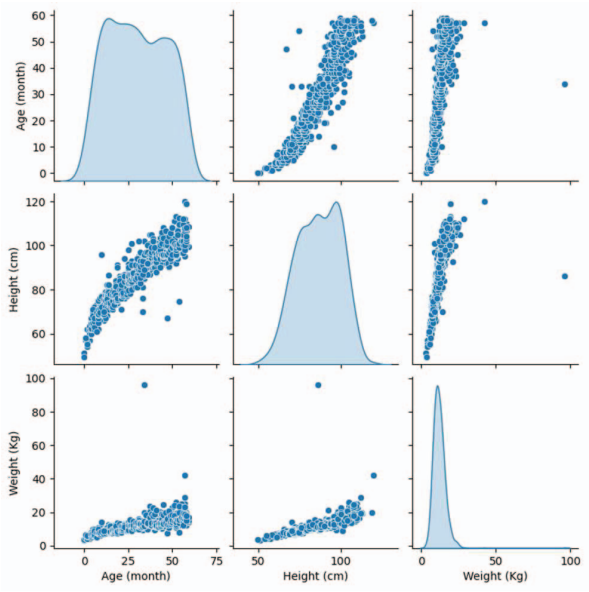

Figure 2.1. Visualization of the dataset using a pair-plot, comparing Heights, Weights and Age.

It is worth noting that in figures 2.1. there exist some outliers within the data. This outlier data will be removed from the dataset at the data preparation stage.

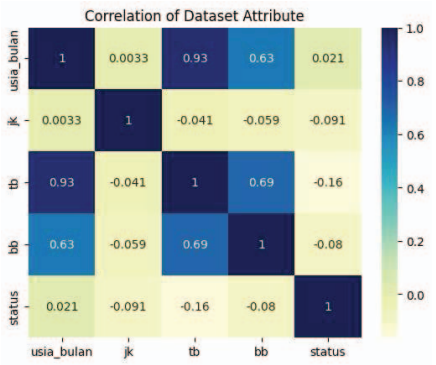

Figure 2.3. Correlation graph of every attribute in the dataset

Figure 2.3. shows the different correlations between the attributes found in the dataset. The strongest correlation in the dataset is the relation between the height (*tb*) and the age of the child in months (*usia_bulan*). The weakest correlation being the relation between height and stunting status (*status*).

C. *Data Preparation Phase*

	Age (months)	Gender	Height (cm)	Weight (kg)	Status
0	56	0	110.0	22.7	0
1	33	0	89.0	12.1	0
2	57	0	100.0	14.9	0
3	32	0	86.0	11.2	0.5
4	44	0	92.0	13.1	0.5
…	…	…	…	…	…
751	58	1	108.8	17.7	0

Figure 2.4. Table representation of the attributes in our data set

In figure 2.4 it can be seen that here are a number of attributes in the dataset that have numerical values instead of categorical data. Originally, the data was recorded as categorical, with the values of gender being 'male' and 'female' and the values for Status being 'normal', 'stunting' and stunted'. However, for the model to work optimally the data is encoded and assigned numerical values. Male is assigned as '0' and Female '1'. As for Status, normal is '0', stunting is '0.5' and stunted is '1'.

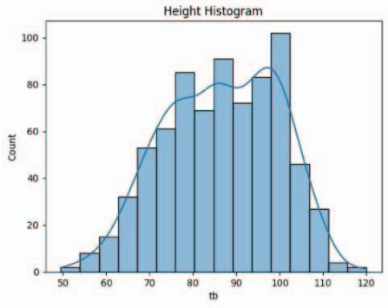

Figure 2.5. Height Histogram.

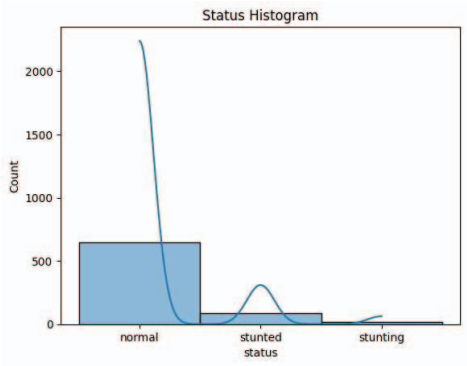

Figure 2.6. Status of Stunting Histogram

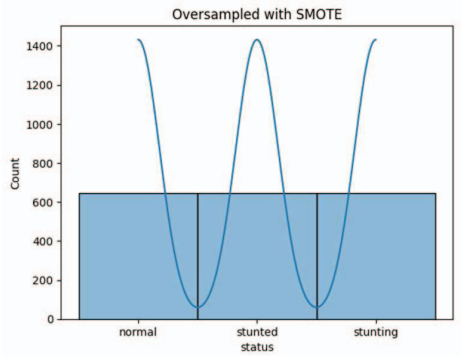

Figure 2.7. Result of Oversampled Data from SMOTE

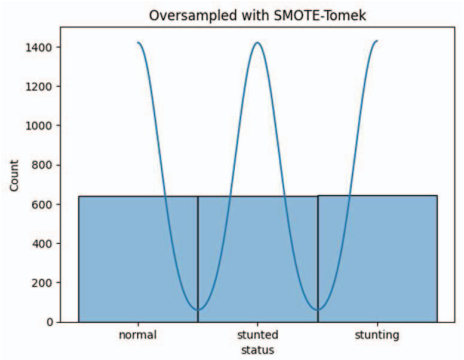

Figure 2.8. Result of Oversampled Data from SMOTE – Tomek

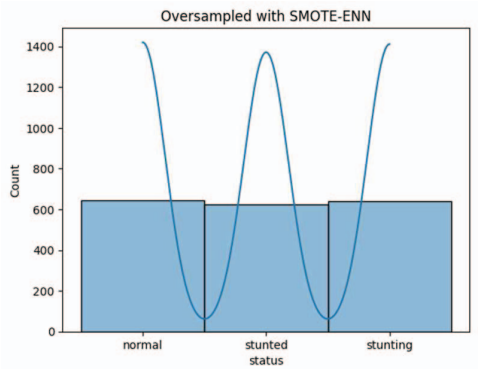

Figure 2.9. Result of Oversampled Data from SMOTE – ENN

As shown in Figure 2.4. our dataset consists of roughly 750 data records of children in Yogyakarta. The status of stunting in Figure 2.4. is indicated in the 'status' column in which a number of 0 means normal, 0.5 means stunting and 1 meaning stunted.

Based on the status histogram in Figure 2.6, we can see that the data distribution of the dataset is imbalanced. Table 1 shows the data distribution for the status class in more detail.

TABLE I. DATASET DATA DISTRIBUTION BEFORE SMOTE

Status	Count	Percentage
Normal	645	86%
Stunted	89	12%
Stunting	18	2%

TABLE II. DATASET DATA DISTRIBUTION AFTER SMOTE

Status	Count	Percentage
Normal	645	34%
Stunted	641	33%
Stunting	623	33%

Based on this imbalanced condition, it is necessary to adopt sampling techniques and data cleaning techniques to help solve the imbalanced data problem to increase the accuracy of the classifier.

SMOTE Synthetic Minority Over-sampling Technique (SMOTE) [10] is an over-sampling method. The fundamental concept of SMOTE algorithm is that for each minority class sample x, some more samples are randomly selected from their k-nearest neighbors, and a new sample is constructed according to Table I. In this way, new minority class samples will produce a new sample, it will result in a problem called sample overlap [11].

$$x_{new} = x_i + |x_i^{'} - x_i| \times \partial \qquad (1)$$

x_{new} is the new sample; x_i is the minority sample; $x_i^{'}$ is one of the k-nearest neighbors of x_i; ∂ is a random number and $\partial \in [0,1]$. **ENN** Wilson [12] developed the Edited Nearest Neighbor (ENN) algorithm in which **S** starts out the same as training data sets, and then each instance in S is removed if it does not agree with the majority of its k nearest neighbors (with k=3, typically) [13] . If a sample belongs to minority class, and there're two or more of its three nearest neighbors that belong to the majority class, then the sample will be removed, thereby leading to smoother boundaries between classes [11].

SMOTE+ENN SMOTE-ENN is a hybrid oversampling technique that combines the strengths of SMOTE oversampling and the built-in KNN classifier [14]. Firstly, the training data are over-sampled by using SMOTE. Secondly, each sample' three nearest neighbors are found in the training data. Thirdly, the samples that are misclassified are removed, producing cleaner data. In this way, not only can we balance the data distribution, but also boundaries between classes are clearer [11].

SMOTE – Tomek is a hybrid sampling method designed for addressing imbalances in datasets. It merges the Synthetic Minority Over-sampling Technique (SMOTE) with Tomek links under-sampling techniques. In the process, it initially

employs Tomek links under-sampling to eliminate noisy samples from the majority class. Subsequently, it applies SMOTE to generate synthetic samples for the minority class. This helps to balance the dataset while also reducing the noise in both the majority and minority classes [15].

Random Forests serve as ensemble learning methods suitable for both classification and regression problems. They consist of numerous decision trees generated from bootstrap samples [16]. **Bagging** is another ensemble learning method which uses multiple subsets of the training data that are created by randomly sampling with replacement. This means that some data points may be repeated in a subset while others may be omitted [17]. The third ensemble learning method is the **AdaBoost**. This method adapts and focuses on getting better at the areas where previous models struggled, gradually creating a strong learner from a series of weaker ones [18].
Voting algorithm refers to ensemble learning techniques where multiple individual models are combined to make predictions. [19]

D. Modeling Phase

To detect stunting in our dataset, a robust machine learning model was devised. However, the research began with the essential literature review and data understanding.

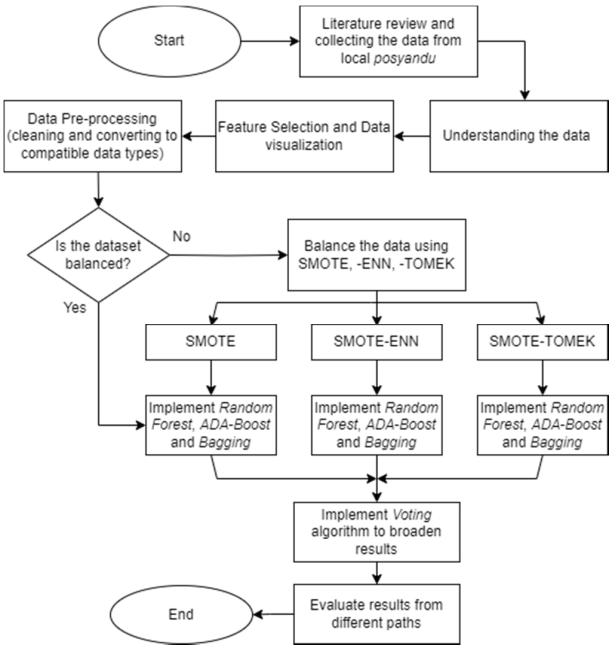

Figure 3. Machine Learning Modeling Flowchart

According to Figure 3, the research then continued by data pre-processing to check the missing value and doing exploratory data analysis (EDA). After the data is ready, the next step is to process the data with SMOTE, SMOTE-ENN, and SMOTE-Tomek Link techniques to balance the data. After the pre-process procedure, we split the data into training (80%) and testing (20%) types. We put the training data into the ensemble learning classifier namely Ada Boost, Random Forest, and Bagging.

The *Voting* algorithm is finally implemented in order to combine the different results of the three oversampling techniques.

True Label	Predicted Label		
	Normal	***Stunted***	***Stunting***
Normal	True Negative (TN)	False Positive (FP)	True Negative (TN)
Stunted	False Negative (FN)	True Positive (TP)	False Negative (FN)
Stunting	True Negative (TN)	False Positive (FP)	True Negative (TN)

TABLE III. CONFUSION MATRIX

The formulas for calculating this performance measure are given in (2), (3), and (4):

$$P(Precision) = \frac{TP}{TP + FP} \quad (2)$$

$$R(Recall) = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2PR}{P + R} \quad (4)$$

E. Evaluation Phase

For performance measure, precision, recall, and F1-value are taken. Confusion matrix is used and shown in *Table III.* The results of the over-sampling algorithm integrated with three ensemble learning algorithm namely Random Forest, Ada Boost, and Bagging as well as the groupings carried out in the testing field.

F. Deployment Phase

The dissemination of knowledge that will be generated based on the novel early stunting detection system will be used for helping government, healthcare facility, and parents to be more aware of these conditions.

III. RESULTS AND ANALYSIS

The proposed methodology leverages various SMOTE algorithm to address class imbalance, enhancing the model's robustness. The integration of ensemble learning further contributes to the overall predictive performance, offering a comprehensive and innovative solution to early stunting detection.

Our experiment was implemented on MacBook M1 Pro with 16.00G RAM. The results of using the three-over-sampling algorithm SMOTE, SMOTE-ENN, and SMOTE-Tomek Link integrated with ensemble learning model Random Forest, Ada Boost, and Bagging.

After preprocessing the data, we split the data into training and testing types. As it can be seen from *Fig.2.6*, the data distribution is imbalanced without any sampling techniques, there are many samples of Normal. In *Fig*. 2.7. – *Fig. 2.9*. it can be seen that the data distribution is more balanced with SMOTE, SMOTE-ENN, and SMOTE-Tomek technology, and number of stunting and stunted is increased.

The reported values for precision, recall and F1-value were gained by the classification report from scikit learn library. By analyzing Table IV – Table XV, it is apparent that SMOTE-ENN using Ada Boost classifier and SMOTE using Bagging classifier achieved the highest F1-value that other presented methods. Specifically, the F1-value for the minority classes (stunted and stunting) is 0.98 and 1.00.

A. *Random Forest*

TABLE IV. WITHOUT SMOTE

	Precision	Recall	F1-score	Support
Normal	0.89	1.00	0.94	127
Stunted	0.50	0.21	0.30	19
Stunting	1.00	0.20	0.33	5
Accuracy			0.87	151
Macro avg	0.80	0.47	0.52	151
Weighted avg	0.85	0.87	0.84	151

TABLE V. SMOTE

	Precision	Recall	F1-score	Support
Normal	0.99	0.97	0.97	139
Stunted	0.95	0.98	0.97	124
Stunting	1.00	0.99	1.00	124
Accuracy			0.98	387
Macro avg	0.98	0.98	0.98	387
Weighted avg	0.98	0.98	0.98	387

TABLE VI. SMOTE - ENN

	Precision	Recall	F1-score	Support
Normal	0.96	0.96	0.96	137
Stunted	0.93	0.92	0.93	117
Stunting	0.98	0.98	0.98	128
Accuracy			0.96	382
Macro avg	0.95	0.95	0.95	382
Weighted avg	0.96	0.96	0.96	382

TABLE VII. SMOTE - TOMEK

	Precision	Recall	F1-score	Support
Normal	0.97	0.98	0.98	142
Stunted	0.96	0.97	0.96	122
Stunting	1.00	0.98	0.99	122
Accuracy			0.98	386
Macro avg	0.98	0.98	0.98	386
Weighted avg	0.98	0.98	0.98	386

Based on the tables III, IV and V, we can deduce that the balancing algorithm with the highest average precision for the random forest is tied between SMOTE and SMOTE – Tomek score of 0.98. Its worth noting that the SMOTE – ENN algorithm also has a precision score that's considerably lower than the other two, with a precision score of 0.93.

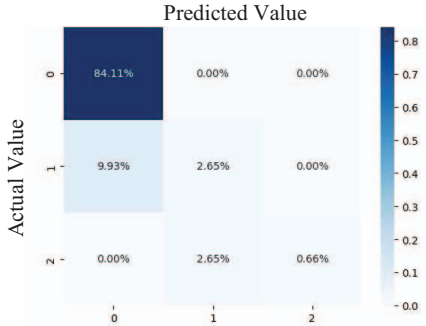

Figure 4.1. Ada Boost Classifier WITHOUT SMOTE confusion matrix

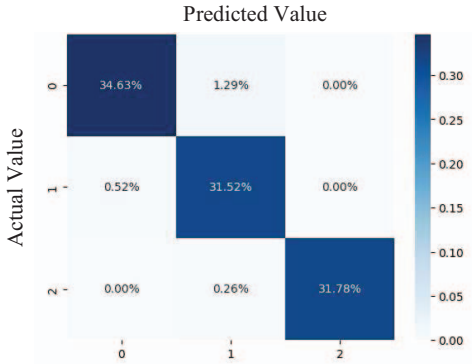

Figure 4.2. Random Forest Classifier with SMOTE confusion matrix

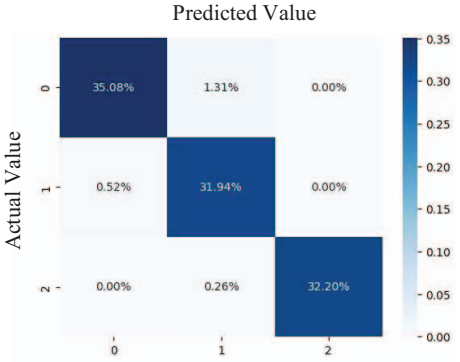

Figure 4.3. Random Forest Classifier with SMOTE – ENN confusion matrix

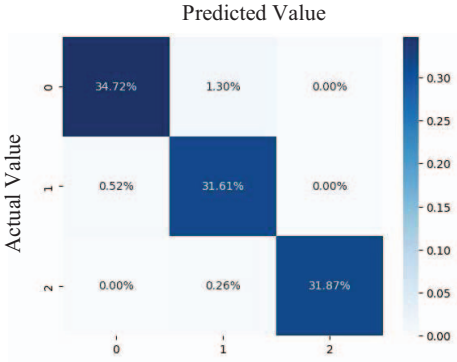

Figure 4.4. Random Forest Classifier with SMOTE - Tomek confusion matrix

Figures 4.2, 4.3 and 4.4 are confusion matrixes derived from the Random Forest algorithm implemented in the learning model developed in this paper. From the confusion matrixes for the Random Forest, there is a clear trend in areas of confusion. At points 0,1 (normal – stunted), 1,0 (stunted – normal) and 1,2 (stunted – stunting) there exists variable measures of confusion ranging up to 1.30%.

B. *Ada Boost*

TABLE VIII. WITHOUT SMOTE

	Precision	Recall	F1-score	Support
Normal	0.86	0.94	0.50	127

Stunted	0.27	0.16	0.20	19
Stunting	1.00	0.40	0.57	5
Accuracy			0.82	151
Macro avg	0.71	0.50	0.56	151
Weighted avg	0.79	0.82	0.80	151

TABLE IX. SMOTE

	Precision	Recall	F1-score	Support
Normal	0.99	0.96	0.97	139
Stunted	0.95	0.98	0.97	124
Stunting	1.00	0.99	1.00	124
Accuracy			0.98	387
Macro avg	0.98	0.98	0.98	387
Weighted avg	0.98	0.98	0.98	387

TABLE X. SMOTE - ENN

	Precision	Recall	F1-score	Support
Normal	0.99	0.99	0.99	137
Stunted	0.98	0.98	0.98	117
Stunting	1.00	0.99	1.00	128
Accuracy			0.99	382
Macro avg	0.99	0.99	0.99	382
Weighted avg	0.99	0.99	0.99	382

TABLE XI. SMOTE - TOMEK

	Precision	Recall	F1-score	Support
Normal	0.97	0.98	0.98	142
Stunted	0.96	0.97	0.96	122
Stunting	1.00	0.98	0.99	122
Accuracy			0.98	386
Macro avg	0.98	0.98	0.98	386
Weighted avg	0.98	0.98	0.98	386

Based on the tables, we can deduce that the balancing algorithm with the highest average precision for the Ada Boost is the SMOTE – ENN algorithm with a score of 0.99. The SMOTE – ENN algorithm also has the highest recall score of 0.99.

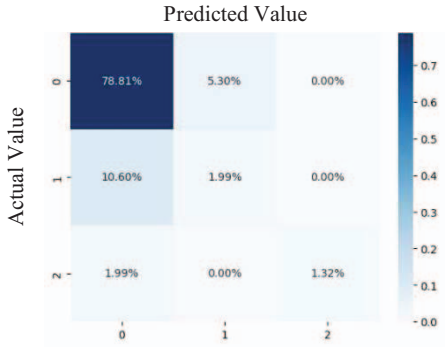

Figure 4.5. Ada Boost Classifier WITHOUT SMOTE confusion matrix

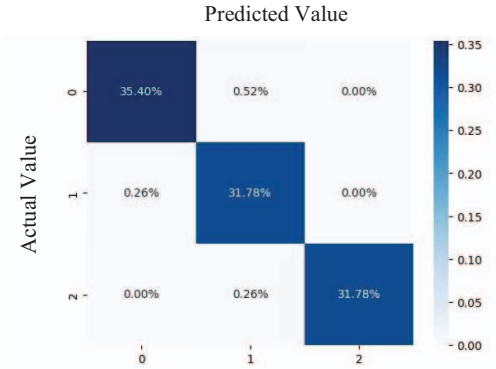

Figure 4.6. Ada Boost Classifier with SMOTE confusion matrix

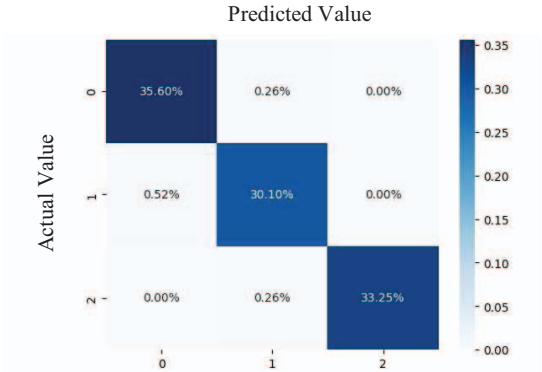

Figure 4.7. Ada Boost Classifier with SMOTE - ENN confusion matrix

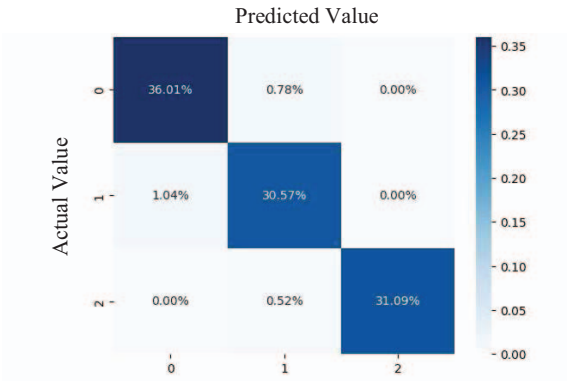

Figure 4.8. Ada Boost Classifier with SMOTE - Tomek confusion matrix

Figures 4.6, 4.7 and 4.8 are confusion matrixes derived from the Ada Boost algorithm implemented in the learning model developed in this paper. From the confusion matrixes for the Ada Boost, there is a similar trend in areas of confusion to the confusion matrixes for the Random Forest. The greatest percentage for confusion in the Ada Boost is found at point 0,1 (normal – stunting) for the SMOTE – Tomek algorithm.

C. Bagging

TABLE XII. WITHOUT SMOTE

	Precision	Recall	F1-score	Support
Normal	0.91	1.00	0.95	127
Stunted	0.60	0.32	0.41	19
Stunting	0.50	0.20	0.29	5
Accuracy			0.89	151
Macro avg	0.61	0.51	0.55	151
Weighted avg	0.86	0.89	0.86	151

TABLE XIII. SMOTE

	Precision	Recall	F1-score	Support
Normal	0.99	0.98	0.99	139
Stunted	0.97	0.99	0.98	124
Stunting	1.00	0.99	1.00	124
Accuracy			0.99	387
Macro avg	0.99	0.99	0.99	387
Weighted avg	0.99	0.99	0.99	387

TABLE XIV. SMOTE - ENN

	Precision	Recall	F1-score	Support
Normal	0.99	0.99	0.99	137
Stunted	0.97	0.98	0.98	117
Stunting	0.99	0.99	0.99	128
Accuracy			0.99	382
Macro avg	0.99	0.99	0.99	382
Weighted avg	0.99	0.99	0.99	382

TABLE XV. SMOTE - TOMEK

	Precision	Recall	F1-score	Support
Normal	0.98	0.96	0.97	142
Stunted	0.94	0.96	0.95	122
Stunting	0.98	0.98	0.98	122
Accuracy			0.97	386
Macro avg	0.97	0.97	0.97	386
Weighted avg	0.97	0.97	0.97	386

For the Bagging algorithm we can deduce that there is another tie in highest average precision for the balancing algorithms, the tie between SMOTE and SMOTE – ENN. These algorithms both have a average precision score of 0.99 and similarly with their their recall scores being 0.99.

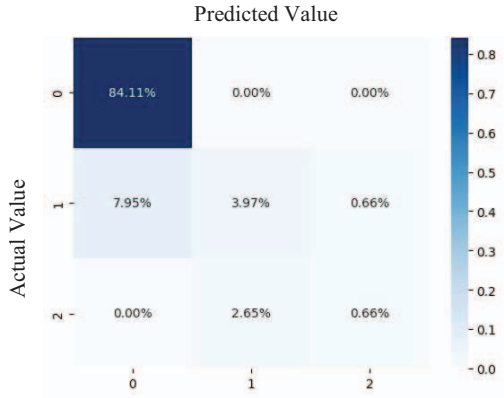

Figure 4.9. Bagging Classifier WITHOUT SMOTE confusion matrix

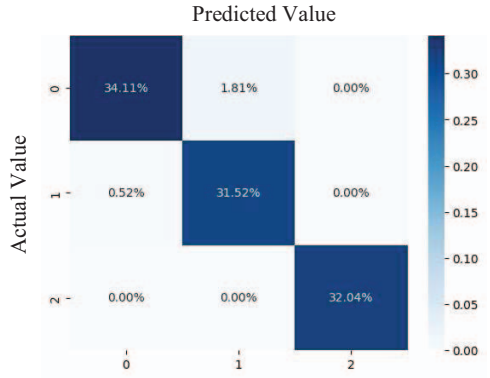

Figure 4.10. Bagging Classifier with SMOTE confusion matrix

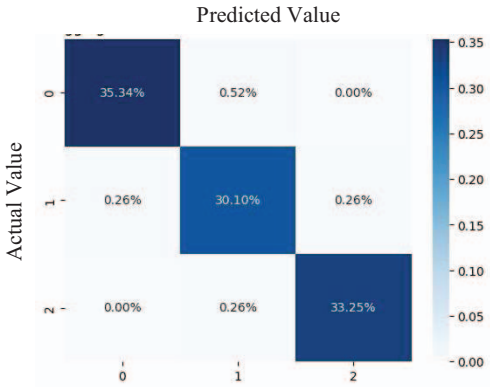

Figure 4.11. Bagging Classifier with SMOTE - ENN confusion matrix

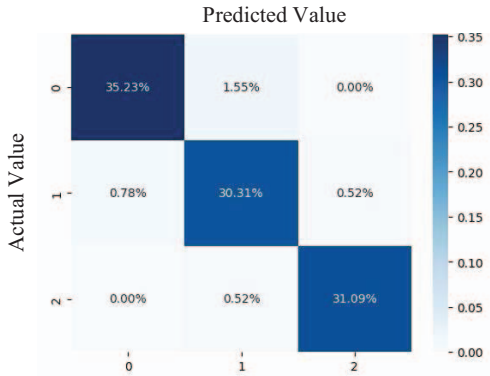

Figure 4.12. Bagging Classifier with SMOTE - Tomek confusion matrix

Figures 4.10, 4.11 and 4.12 are confusion matrixes derived from the Bagging algorithm implemented in the learning model developed in this paper. From the confusion matrixes for the Bagging algorithm, there is a slightly different trend in areas of confusion. There is are 4 points in which the learning model gets confused. Points 0,1 (normal – stunted), 1,0 (stunted – normal), 1,2 (stunted – stunting) and additionally point 2,1 (stunting – stunted).

D. Voting Results

TABLE XVI. TABLE OF RESULTS FOR VOTING ALGORITHM

	Accuracy	F-1 Macro-Avg
Voting without SMOTE	91%	72%
Voting with SMOTE	88%	65%
Voting with SMOTE-ENN	87%	62%
Voting with SMOTE-Tomek	88%	65%

For the results of the voting, we found that it unintentionally causes a reduction in the accuracy of the model with SMOTE oversampling methods. With future research and exploration, we intend increase the accuracy of the voting method and to understand what the cause of this reduction is.

IV. CONCLUSION

In this paper, multiple Ensemble Learning algorithm is employed, namely Random Forest, Ada Boost, and Bagging. In the pre-process procedure the data is labelled as three classes which are normal, stunted, and stunting. For the imbalanced data problem in the dataset, multiple oversampling technique is applied namely SMOTE, SMOTE-ENN, and SMOTE-Tomek to process the dataset and solve the problems of data imbalance and sample overlap. We also take precision, recall, and F1-value as the evaluation. Experimental results from the stunting dataset indicate that the proposed model can result in better prediction of minority classes than using traditional machine learning algorithm. As for the future plan of this research, we aim to deploy said learning model into an application that enables the user to accurately determine the condition of their growth by inputting required data.

V. ACKNOWLEDGEMENTS

This work was partially supported by the *Department of Computer Science and Electronics*, *Universitas Gadjah Mada* under the Publication Funding Year 2024.

REFERENCES

[1] R. W. Fonseka et al., “Measuring the impacts of maternal child marriage and maternal intimate partner violence and the moderating effects of proximity to conflict on stunting among children under 5 in post-conflict Sri Lanka,” SSM - Population Health, vol. 18, pp. 1–9, Jun. 2022.

[2] BKPK, H. Angka stunting Tahun 2022 Turun Menjadi 21,6 persen - badan kebijakan Pembangunan Kesehatan: BKPK Kemenkes. Badan Kebijakan Pembangunan Kesehatan | BKPK Kemenkes. https://www.badankebijakan.kemkes.go.id/angka-stunting-tahun-2022-turun-menjadi-216-persen/ (2023, January 25).

[3] HumbangHasundutankab.go.id, “Indonesia peringkat 5 di dunia, stunting Disebut Bukan hanya urusan pemerintah,” HumbangHasundutankab.go.id, https://humbanghasundutankab.go.id/main/index.php/read/news/828 (accessed Nov. 15, 2023).

[4] Perdana, A. Y., Latuconsina, R., Dinimaharawati, A. (n.d.). “Prediksi Stunting Pada Balita Dengan Algoritma Random Forest”. Proceeding of Engineering : Vol.8, No.5 Oktober 2021.

[5] Prasetyo, E., Nugroho, K. (n.d.). ”Optimasi Klasifikasi Data Stunting Melalui Ensemble Learning pada Label Multiclass dengan Imbalance Data Optimizing Stunting Data Classification Through Ensemble Learning on Multiclass Labels with Imbalance Data”. Techno.COM, Vol. 23, No. 1, Februari 2024.

[6] Minister of Health Regulation. “Peraturan Menteri Kesehatan Republik Indonesia, Standar Antropometri” Anak January (2020).

[7] D. Astuti, “Penentuan Strategi Promosi Usaha Mikro Kecil dan Menengah (UMKM) Menggunakan Metode CRISP-DM dengan Algoritma K-Means Clustering,” J. Informatics, Inf. Syst. Softw. Eng. Appl., vol. 1, no. 2, pp. 60–72, 2019.

[8] A. Purbasari, F. R. Rinawan, A. Zulianto, A. I. Susanti, and H. Komara, “CRISP-DM for Data Quality Improvement to Support Machine Learning of Stunting Prediction in Infants and Toddlers,” in 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), pp. 1–6, 2021.

[9] N. Mirantika, “Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Penyebaran Covid-19 di Provinsi Jawa Barat,” Nuansa Inform., vol. 15, no. 2, pp. 92–98, 2021.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, no. Sept. 28, pp. 321–357, 2002

[11] T. Lu, Y. Huang, W. Zhao, and J. Zhang, “The Metering Automation System based Intrusion Detection Using Random Forest Classifier with SMOTE+ENN,” Proc. IEEE 7th Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2019, pp. 370–374, 2019,

[12] D. L. Wilson, “Asymptotic Properties of Nearest Neighbor Rules Using Edited Data,” IEEE Trans. Syst. Man Cybern., vol. 2, no. 3, pp. 408–421, 1972,

[13] R. Alejo, J. M. Sotoca, R. M. Valdovinos, and P. Toribio, “Edited Nearest Neighbor Rule for Improving Neural,” pp. 303–310.

[14] B. F. Wee, H. H. Hwong, S. Sivakumar, K. H. Lim, and W. K. Wong, “Deep Learning and SMOTEENN-based Univariate Feature Selection Approaches for Diabetes Classification,” 2023 Int. Conf. Digit. Appl. Transform. Econ., pp. 1–5, 2023,

[15] M. R. Kumar, N. Natteshan, J. Avanija, K. R. Madhavi, N. S. Charan and V. Kushal, "SMOTE-TOMEK: A Hybrid Sampling-Based Ensemble Learning Approach for Sepsis Prediction," 2023 2nd International Conference on Edge Computing and Applications (ICECAA), Namakkal, India, pp. 724-729, 2023.

[16] D. P. Mohandoss, Y. Shi and K. Suo, "Outlier Prediction Using Random Forest Classifier," 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), NV, USA, pp. 0027-0033, 2021.

[17] Leo Breiman, “Bagging predictors,” Machine Learning, vol. 24, pp.123-140, 1996.1.

[18] Robert E. schapire, “The boosting approach to machine learning: An overview,” in MSRI Workshop on Nonlinear Estimation and Classification, 2002.

[19] Kumar, A., Sushil, R., Tiwari, A. K “Classification of Breast Cancer using User-Defined Weighted Ensemble Voting Scheme”. IEEE Region 10 Annual International Conference. (2021).