Predicting Toddler Stunting at Bandarharjo Health Center: Applications of K-Nearest Neighbors and Naïve Bayes Algorithms

1st Ashiva Prameswara
School of Computing
Telkom University
Bandung, Indonesia
prameswaraashiva@student.telkomuniversity.ac.id

2nd Putu Harry Gunawan
CoE HUMIC, School of Computing
Telkom University
Bandung, Indonesia
phgunawan@telkomuniversity.ac.id

***Abstract*—According to the 2021 Indonesian Nutrition Status Study (SSGI), the incidence of toddler stunting in the nation is 24.4 %, which is a significant public health problem. Stunting is characterized by having a height under standard caused by malnutrition and lack of healthcare before and after a birth which has negative impacts on the toddlers' physical growth and brain development. This study evaluates how imbalanced data influenced The K-Nearest Neighbors (K-NN) and Naïve Bayes in predicting stunting status. In the imbalanced dataset, K-NN shows high performance with an accuracy of 98.41% and a Macro F1-Score of 89.30%, while Naive Bayes recorded an accuracy of 90.76% and a Macro F1-Score of 70.00%. After applying SMOTE to balance data, K-NN stayed stable with 98.15% accuracy and Macro F1-Score increased to 89.90%. Besides, Naïve Bayes experienced a decline in accuracy to 87.14%, although Macro F1-Score increased to 87.20%. These findings highlight the advantages of deep K-NN handle data imbalance and its stability is better than Naïve Bayes. The increased F1-Score after applying SMOTE reflects improved precision and recall for identifying stunted cases, even though overall accuracy declined. This decline can be attributed to the algorithm's sensitivity to class distribution because the dataset became more balanced. Collectively, these results underscore the importance of using balanced datasets.**

***Index Terms*—stunting, K-Nearest Neighbors, Naïve Bayes, machine learning**

I. Introduction

Stunting is a significant nutritional issue in Indonesia, where many toddlers experience stunted growth due to chronic malnutrition and inadequate healthcare before and after birth [1]. This condition, marked by a child having a height below the expected standard for their age, is influenced by direct factors such as maternal malnutrition during pregnancy, preterm birth, lack of exclusive breastfeeding, and infections. Indirect factors include poor health services, education, sociocultural practices, and environmental sanitation [2] [3].

In 2020, Indonesia was ranked 115th out of 151 countries for stunting prevalence by the United Nations International Children's Emergency Fund (UNICEF) and the World Bank. Among Southeast Asian nations, Indonesia has the second highest prevalence of stunting after Cambodia [4]. This issue has been prioritized in the Sustainable Development Goals (SDGs) to eliminate hunger and malnutrition by 2030 and achieve food security in Indonesia [5].

Stunting impacts not only physical growth but also brain development, which can affect a child's academic performance, productivity, and creativity in the future [6]. The 2021 Indonesian Nutritional Status Study (SSGI) reported a national stunting prevalence of 24.4%, with Central Java at 20.9% and Semarang City at 21.3%. Specifically, the Bandarharjo Community Health Center in Semarang recorded 155 stunted toddlers out of 3,787 in February 2024.

Previous research has explored various machine learning models to predict stunting in toddlers. Hindratomo Hady Sutarno et al. (2021) used the K-Nearest Neighbors (K-NN) algorithm, achieving 97.31% accuracy with three k-fold cross-validation tests [7]. Monica Yoshe Titimeidara and Wiwien Hadikurniawati (2021) employed the Naïve Bayes method, resulting in 88% accuracy [8]. Nur Fitriyani Bahany et al. (2024) compared several models and found Naïve Bayes to have the highest performance with an accuracy of 98.57% [9].

This research aims to predict stunting in toddlers by using the K-Nearest Neighbors and Naïve Bayes algorithms, utilizing data from the Bandarharjo Community Health Center. K-NN is chosen for its simplicity and effectiveness in handling non-linear data relationships, making it suitable for this context where the characteristics of stunting may not follow a strict linear pattern [7] [10]. On the other hand, Naïve Bayes is selected for its efficiency and ability to perform well with smaller datasets, as well as its assumption of feature independence, which can help in quickly identifying the risk factors associated with stunting [9] [11]. By using these models, the study seeks to accurately identify toddlers at risk of stunting, contributing to efforts to reduce its prevalence in Indonesia.

II. Methods

A. Research Design

This research begins by reviewing relevant literature to establish the background of the problem and gather methodological insights to address stunting issues in toddlers. Subsequently, toddler data is collected from the Bandarharjo Community Health Center in Semarang City. The collected data is initially comprehended to understand each feature within the dataset. Relevant features are then selected for further processing and visualized to facilitate explanation and analysis. Prior to implementing chosen methods on the data, it undergoes cleaning to remove anomalies and is converted to compatible data types for processing. The data's balance is assessed, and if found imbalanced, it is adjusted accordingly. Upon completion of preprocessing, the dataset is divided into

training and testing sets. Next, the K-Nearest Neighbor (K-NN) and Naïve Bayes algorithms are employed to predict stunting status in toddlers, aiming to provide accurate and effective predictions. Figure 1 will show the flowchart used for the research.

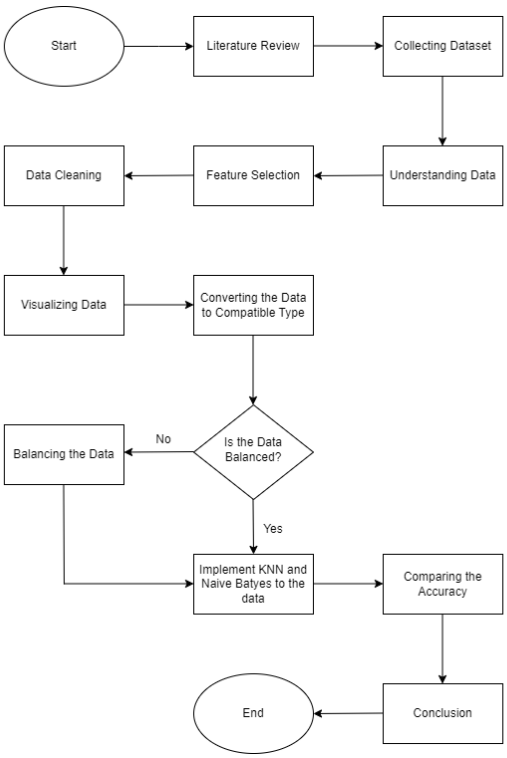

Fig. 1. Research Flowchart

B. K-Nearest Neighbor

K-Nearest Neighbors (K-NN) is a classification technique that determines the smallest distance between test and training data. Whenever new data or an example has to be categorized, the K-NN algorithm will look for a group of training data that has characteristics that are similar or close to the new instance that wants to be classified, and then determines the class label based on the majority of class labels from the nearest neighbors in question [12] [13]. To determine how close or the shortest distance is between the nearest neighbors in the training data, use the Euclidean Distance calculation using the following formula [14]:

$$d = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \tag{1}$$

Where d is an equation calculated as a Euclidean equation and n is the number of features of a data. Then *x* is training data and *y* is test data that will be evaluated against training data. The smaller the valued, the closer and more similar the two data are. The K-NN algorithm can be seen in Algorithm 1.

C. Naïve Bayes

The Naïve Bayes Classifier (NBC) algorithm is a statistical classification method used to predict the probability or possibility of class membership for data. Naïve Bayes performs probability calculations using Bayes' theorem and frequency values from data [7]. The calculation process is by

Algorithm 1 K-Nearest Neighbors (K-NN)

Input: X: Training data, Y: Class labels for X, K: number of nearest neighbors
Output: test sample class x

Start
Classify (X, Y, x)
1. for each sample x
Calculate the distance to each data on X using the Euclidean formula
end for
2. classification of each data x based on the results of the majority of the closest data
End

studying the probabilities of the classes in the training dataset. Then, for each given test data, the algorithm calculates the probability of each class based on its attributes using Bayes' Theorem. The calculation of Bayes' theorem can be calculated with the following formula on the other [15] [16]:

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)} \tag{2}$$

In Bayesian probability, the variable X represents data with an unknown class, while the hypothesis that data X belongs to a certain class is represented by H. The posterior probability $P(H|X)$ represents the probability of hypothesis H given condition X. The probability of X, given that hypothesis H is true, is represented as $P(X|H)$. The Naive Bayes approach may be seen in Algorithm 2 [17].

Algorithm 2 Naïve Bayes

Input: Training dataset, testing dataset
Output: Classified train dataset

1. Training
For each feature
For each feature-value
For each class-label H
$P(X|H)$ = (Total number of occurrences of feature values with class labels) | (Total number of occurrences of class label)
Endfor
Endfor
Endfor
2. Testing
For each instance in test data
Calculate probability using Bayes's theorem formula
Endfor
3. Assign the class label with maximum probability to the test example

D. Evaluation Metrics

The evaluation stage to measure the performance of K-NN and Naïve Bayes is by using the Confusion Matrix. The Confusion Matrix shows the amount of correct and incorrect data that is classified and consists of four values produced as a representation of the results of the classification process, including True Positive (TP), True Negative (TN),

False Positive (FP), and False Negative (FN) . This metric is calculated as follows.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$F1 - score = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \tag{6}$$

III. Results and Discussions

A. Exploratory Data Analysis

The dataset used was taken from the Bandarharjo Community Health Center, Semarang City. This dataset contains personal information and health status of toddlers in February 2024. The dataset consists of 3787 records. The selected features are Age, Birth Weight, Birth Height, Weight, Height, Height/Age,Height/Age ZS,Weight/Height ZS. Besides that, the Height/Age column in dataset becomes the target class because it contains information on whether the child is average or has a stunting condition. Here are the distributions of some features to classify the dataset.

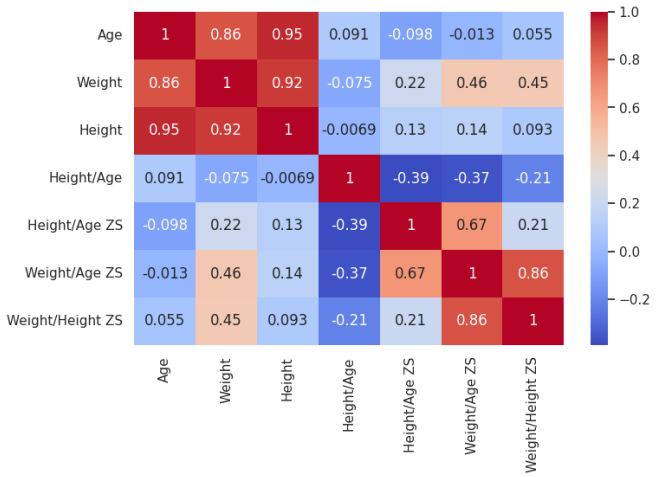

Fig. 2. Correlation Data

Figure 2 illustrates that the highest correlation was observed between age and height, with a correlation coefficient of 0.95. In addition, there is a strong correlation between age and body weight, with a coefficient of 0.86. Therefore, the target height/age variable shows a high correlation with age, namely 0.091.

Figure 3 illustrates stunting data, indicating that only 4.1% of the population falls into the 'short' and 'very short' categories, while 95.9% falls into the 'normal' category. This notable imbalance in class distribution is crucial to highlight because a dataset is generally considered imbalanced when there is a considerable disparity in the number of instances between the different classes [18]. Additionally, the Weight feature and all Z-Score features exhibit extreme values at both ends. These outlier values can introduce noise and lead to inaccurate results, highlighting the need for careful handling to optimize model performance and ensure reliable predictions. The boxplot of the stunting data distribution is shown in Figure 4.

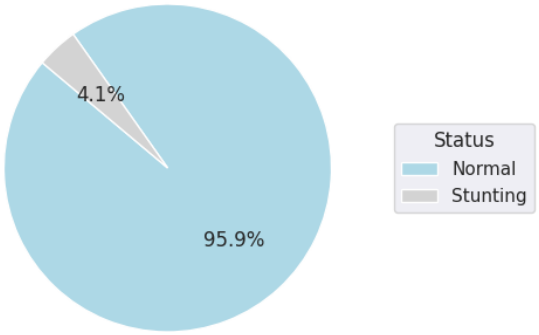

Fig. 3. Comparison of normal and stunting quantity

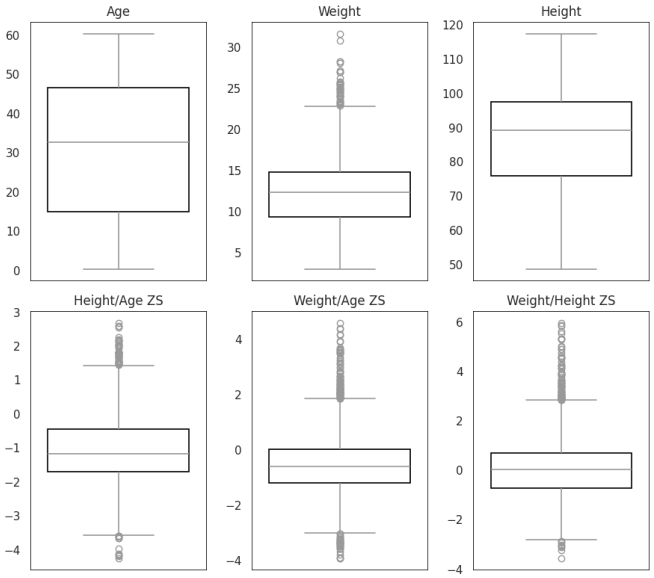

Fig. 4. Distribution of Stunting Data

B. Preprocessing

After completing the dataset collection process, the next stage is data preprocessing. At this stage, cleaning was carried out on the toddler dataset at the Bandarharjo Community Health Center in February 2024. This data cleaning was carried out because there was data in several rows in the toddler age column at the time of measurement considering that at that time, toddlers did not take measurements again, so the rows that had empty values are removed as a data cleaning step. Additionally, a thorough inspection of the dataset was performed by checking for missing values across all columns. The result showed that there were no missing values remaining in any column.

In this research, there is toddler data with non-numerical data types, so data transformation is required at the preprocessing stage. In this transformation stage, it is carried out by changing the categorical variables into numerical form with the Gender column "P" becoming "0" and "L" becoming "1". Then also change the TB/U status column "Normal" to "0", "Short" to "1", and "Very Short" to "2". Furthermore, the "Age at Measurement" column was converted into months to maintain consistency across the dataset.

According to the boxplots in Figure 5 shows the process of handling outliers in numerical features. Rare outliers were replaced with the median, while more frequent outliers were

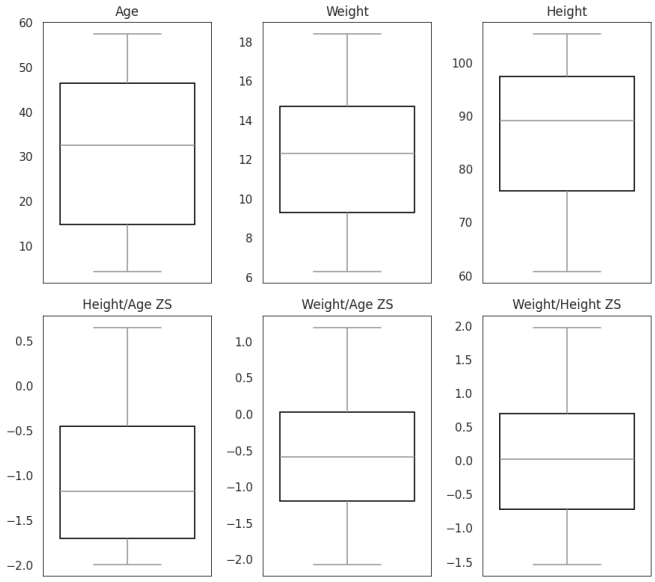

Fig. 5. Distribution After Preprocessing

capped using winsorization, a method that limits extreme values by capping them at specific percentiles closer to the data's center [19].

Based on Figure 3 unambiguously demonstrates the presence of a significant imbalance in the dataset. By using the SMOTE (Synthetic Minority Over-sampling Technique) Upsampling technique, we may augment the quantity of minority data, thereby addressing the problem of class imbalance in the dataset. Before applying SMOTE, the dataset was split into 60% for training and 40% for testing to preserve the authenticity of the test data. Initially, the training data consisted of 3632 normal cases and 155 stunted cases. After SMOTE was applied, the training data became balanced, with 3632 instances for both normal and stunted cases. This strategy aims to enhance the presence of underrepresented groups in the dataset by generating synthetic data that accurately reflects the distribution of minority data [20].

C. K-Nearest Neighbor Implementation

K-Nearest Neighbor is used for deep stunting classification this research. Implementation of K-NN using Python library called sci-kit-learn. There are two distinct datasets available, one for training and one for testing. The datasets are divided in a ratio of 60% for training and 40% for testing. Subsequently, the model underwent evaluation by cross-validation, employing a 5-fold division. The confusion matrix, shown in Table I, presents the K-NN results.

TABLE I
CONFUSION MATRIX FOR K-NN METHOD

		Predicted Values	
		Normal	Stunting
Actual Values	Normal	1438	2
	Stunting	22	53

D. Naïve Bayes Implementation

This research used the Naïve Bayes approach for the purpose of categorizing stunting. The Naïve Bayes implementation utilizes the Scikit-learn Python library. The dataset is divided between training and testing sets using a 60:40 split, which is comparable to the implementation of K-NN. Subsequently, the model underwent evaluation using cross-validation with a 5-fold division. The results of the Naïve Bayes method are shown in the confusion matrix of Table II.

TABLE II
CONFUSION MATRIX FOR NAÏVE BAYES METHOD

		Predicted Values	
		Normal	Stunting
Actual Values	Normal	1341	99
	Stunting	7	68

E. SMOTE Implementation

To handle class imbalance in the dataset stunting, the Synthetic Minority Over-sampling technique is applied Technique (SMOTE). SMOTE is used to increase the number of examples from the minority class by generating synthetic data through the interpolation of existing data. Thus, SMOTE aims to improve class balance and provide better representation for minority classes in the dataset [20]. The classification results of the K-NN and Naïve Bayes models after SMOTE are carried out are displayed in the form of a confusion matrix in Table III and Table IV.

TABLE III
CONFUSION MATRIX FOR SMOTE K-NN METHOD

		Predicted Values	
		Normal	Stunting
Actual Values	Normal	1428	12
	Stunting	14	61

TABLE IV
CONFUSION MATRIX FOR SMOTE NAÏVE BAYES METHOD

		Predicted Values	
		Normal	Stunting
Actual Values	Normal	1252	188
	Stunting	5	70

F. Accuracy Comparison

The method used to classify stunting produce varying performance. The results will compare imbalanced data and balanced data. The experiment will be conducted over ten iterations. The results of imbalanced data can be seen in Table V.

TABLE V
ACCURACY AND F-1 SCORE FOR IMBALANCED DATASET

Method	**Accuracy**	**F-1 Score Macro**
K-Nearest Neigbhor	98.41%	89.30%
Naïve Bayes	90.76%	70.00%

Based on Figure 3, there is still an imbalance in the data between stunting and normal. Therefore, the following are the results after carrying out SMOTE to improve class balance.

Table VI shows that although the accuracy obtained from balanced data is not as high as imbalanced data, the F1-Score macro actually increases for each method. This indicates that classification is more effective and representative when using balanced data, because it is able to better capture model performance in minority classes.

TABLE VI
ACCURACY AND F-1 SCORE FOR BALANCED DATASET

Method	Accuracy	F-1 Score Macro
K-Nearest Neighbor	98.15%	89.90%
Naïve Bayes	87.14%	87.20%

IV. CONCLUSION

This research reveals the significant impact of data imbalance on the performance of classification models, especially K-Nearest Neighbors (K-NN) and Naïve Bayes. Imbalanced data shows high accuracy for both models, with K-NN achieving an accuracy of 98.41% and a Macro F1-Score of 89.30%, and Naïve Bayes with an accuracy of 90.76% and a Macro F1-Score of 70.00%. KNN shows very good ability to detect the majority class on imbalanced datasets, while Naïve Bayes also shows good results although not as high as K-NN. After applying the SMOTE method to balance the data, the model performance underwent significant changes. K- NN still shows consistent results with an accuracy of 98.15% and the Macro F1-Score increases to 89.90%, experiencing only a small decrease in accuracy compared to imbalanced data. In contrast, Naïve Bayes experienced a decrease in accuracy to 87.14%, but the Macro F1-Score increased to 87.20%, indicating an improvement in model performance in the minority class after the data was balanced.

The results of this study emphasize the importance of considering data balance in selecting a classification model. Although imbalanced data can provide high accuracy, balanced data can improve the F1-Score which is more representative in assessing model performance in minority classes. K-NN proved to be the most consistent and effective method in various dataset conditions, while Naïve Bayes showed improvements in handling balanced data although with reduced accuracy.

High accuracy means that a model predicts a large percentage of total instances correctly. However, in cases of imbalanced datasets, a model can achieve high accuracy simply by predicting the majority class well. In contrast, the F1-Score combines precision (how many of the predicted positive cases were actually positive) and recall (how many actual positive cases were correctly predicted). It provides a better measure of a model's ability to identify minority classes. Based on the results of this research, future research should explore additional machine learning algorithms. Expanding the dataset with features like socioeconomic factors, maternal health, and environmental conditions could further enhance prediction accuracy in recognizing stunting.

REFERENCES

[1] A. Daracantika, A. Ainin, and B. Besral, "Pengaruh negatif stunting terhadap perkembangan kognitif anak," *Jurnal Biostatistik, Kependudukan, Dan Informatika Kesehatan*, vol. 1, no. 2, p. 113, 2021.

[2] L. Kurniati *et al.*, "Edukasi dan konseling gizi terhadap kenaikan berat badan bayi bblr & balita stunting di klinik konsultasi gizi rsud h. boejasin pelaihari," *TEMU ILMIAH NASIONAL PERSAGI*, vol. 5, no. 1, 2023.

[3] H. Janawisuta, P. H. Gunawan *et al.*, "Early detection of stunting in indonesian toddlers: A machine learning approach," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE, 2024, pp. 12–16.

[4] N. E. O. Aditia, M. Mitra, A. R. Abidin, Y. Priwahyuni, and C. V. G. Purba, "Factors associated with stunting in children under five years," *Jurnal kesehatan komunitas (Journal of community health)*, vol. 9, no. 1, pp. 122–131, 2023.

[5] H. Arifuddin, H. Arifuddin, A. Arifuddin, and A. F. Nur, "The risk factors of stunting children aged 0-5 years in indonesia: A multilevel analysis," *Healthy Tadulako Journal (Jurnal Kesehatan Tadulako)*, vol. 9, no. 1, pp. 109–120, 2023.

[6] A. A. Fadlilah and A. I. Fibriana, "Kejadian stunting pada balita di wilayah kerja puskesmas poncol," *HIGEIA (Journal of Public Health Research and Development)*, vol. 7, no. 2, pp. 293–302, 2023.

[7] H. H. Sutarno, R. Latuconsina, and A. Dinimaharawati, "Prediksi stunting pada balita dengan menggunakan algoritma klasifikasi k-nearest neighbors," *eProceedings of Engineering*, vol. 8, no. 5, 2021.

[8] M. Y. Titimeidara and W. Hadikurniawati, "Implementasi metode naive bayes classifier untuk klasifikasi status gizi stunting pada balita," *Jurnal Ilmiah Informatika*, vol. 9, no. 01, pp. 54–59, 2021.

[9] N. F. Sahamony, T. Terttiaavini, and H. Rianto, "Analisis perbandingan kinerja model machine learning untuk memprediksi risiko stunting pada pertumbuhan anak: Analysis of performance comparison of machine learning models for predicting stunting risk in children's growth," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 413–422, 2024.

[10] A. T. A. Sibuea, P. H. Gunawan *et al.*, "Classifying stunting status in toddlers using k-nearest neighbor and logistic regression analysis," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE, 2024, pp. 6–11.

[11] C. Fannany, P. H. Gunawan, and N. Aquarini, "Machine learning classification analysis for proactive prevention of child stunting in bojongsoang: A comparative study," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE, 2024, pp. 1–5.

[12] G. A. F. Khansa and P. H. Gunawan, "Predicting stunting in toddlers using knn and naïve bayes methods," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE, 2024, pp. 17–21.

[13] A. Argina, "Application of the k-nearest neighbor classification method on a dataset of diabetes patients," *Indones. J. Data Sci*, 2020.

[14] A. Almomany, W. R. Ayyad, and A. Jarrah, "Optimized implementation of an improved knn classification algorithm using intel fpga platform: Covid-19 case study," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3815–3827, 2022.

[15] S. K. P. Loka and A. Marsal, "Perbandingan algoritma k-nearest neighbor dan naïve bayes classifier untuk klasifikasi status gizi pada balita: Comparison algorithm of k-nearest neighbor and naïve bayes classifier for classifying nutritional status in toddlers," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 1, pp. 8–14, 2023.

[16] S. S. Bafjaish, "Comparative analysis of naive bayesian techniques in health-related for classification task," *Journal of Soft Computing and Data Mining*, vol. 1, no. 2, pp. 1–10, 2020.

[17] A. Gayathri, J. Aswini, and A. Revathi, "Classification of spam detection using naive bayes algorithm over k-nearest neighbors algorithm based on accuracy," *NVEO-Natural Volatiles & Essential Oils Journal— NVEO*, pp. 8516–8530, 2021.

[18] T. R. Hoens and N. V. Chawla, "Imbalanced datasets: from sampling to classifiers," *Imbalanced learning: Foundations, algorithms, and applications*, pp. 43–59, 2013.

[19] F. Zubedi, B. Sartono, and K. A. Notodiputro, "Implementation of winsorizing and random oversampling on data containing outliers and unbalanced data with the random forest classification method," *Jurnal Natural*, vol. 22, no. 2, pp. 108–116, 2022.

[20] A. F. Amida, S. E. Permana, D. Pratama, K. Anam, and A. R. Rinaldi, "Prediction of stunted toddlers using k-nearest neighbor algorithm in kamarang lebak village," *Instal: Jurnal Komputer*, vol. 15, no. 02, pp. 345–355, 2023.