SJER Scientific Journal of Engineering Research

futuristech PT. Teknologi Futuristik Indonesia

Date of publication April 1, 2025, date of current version April 1, 2025.
e-ISSN: 3109-1725
Digital Object Identifier 10.64539/sjer.v1i2.2025.9

Article

Comparison of Machine Learning Algorithms for Stunting Classification

Muhajir Yunus [1,*], Muhammad Kunta Biddinika [2], Abdul Fadlil [3]

[1] Universitas Muhammadiyah Gorontalo, Gorontalo, Indonesia; muhajiryunus@umgo.ac.id
[2] Master Program of Informatics, Universitas Ahmad Dahlan, Yogyakarta 55191, Indonesia
[3] Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta 55191, Indonesia
* Correspondence

The researchers would like to thank the Gorontalo Regency Government for permitting the use of data on the nutritional status of toddlers. Thank you also to the academic supervisors who helped complete this research.

Abstract: Indonesia is one of the countries with medium stunting data over the past decade, around 21.6%. Stunting prevention is a national program in Indonesia, and stunting reduction in children is the first of the six goals in the Global Nutrition Target for 2025. Based on SSGI data in 2022, the prevalence of stunting in Gorontalo Province is 23.8% and is in the high category. Stunting prevention is an early effort to improve the ability and quality of human resources. This study compared two Machine Learning algorithms for stunting classification in children, namely the Naive Bayes method and Decision Tree C4.5 using Python by dividing the training and testing data a total ratio of 80:20. The performance of each algorithm was evaluated using a dataset of child health information based on z-score calculation data with a total of 224 records, consisting of 4 attributes and 1 label, namely gender, age, weight, height and nutritional status. The results of the research that have been conducted show that the Decision Tree C4.5 algorithm achieves the highest accuracy in the classification of stunting events with an accuracy of 87% while for the Naïve Bayes algorithm produces a low accuracy of 71% so that for this study the Decision tree C4.5 algorithm is the best algorithm for the classification of stunting events. These findings suggest this algorithm can be a valuable tool for classifying children's stunting.

Keywords: Stunting Classification; Machine Learning; Decision Tree C4.5; Naïve Bayes

1. Introduction

Stunting, often called short is a condition of failure to thrive in children under five years old (toddlers) due to chronic malnutrition and repeated infections, especially in the first 1,000 days of pregnancy, namely from a fetus to a child aged 23 months [1, 2]. Some of the factors that influence stunting include lack of access to nutritious food, clean water, and poor sanitation, and lack of knowledge about nutrition and healthcare [3]. The problem of stunting in children under five is still a health problem, especially in developing countries. In 2019 seven regions had a very high prevalence of stunting, namely East Africa, Central Africa, South Africa, Oceania, West Africa, and Southeast Asia, where Indonesia is included [4].

Children classified as stunting have short-term and long-term consequences that affect the health and development of human resources [5]. In the short term, stunting can lead to increased mortality and morbidity, decreased cognitive, motor, and language skills, and increased costs of treatment for illness. In the long run, stunting can cause children to become short as adolescents, thereby increasing the risk of obesity and decreasing reproductive health [6].

Stunting prevention is an early effort to improve the ability and quality of human resources [7]. The existence of Machine Learning can be one of the solutions to predicting stunting in children. However, various types of Machine Learning algorithms make it difficult for some people to choose an accurate algorithm according to their needs [8].

Machine Learning is one technology that can facilitate data processing and analysis on a large scale [9]. Machine

learning can build predictive and classification models to quickly and accurately process data [10]. To identify patterns and trends hidden in very complex and large data. Some popular machine learning methods have been used in various applications, such as facial recognition, sentiment analysis, voice recognition, and more. However, some disadvantages of machine learning classification methods, such as requiring adequate and representative data to learn suitable patterns [11].

This study analyzed the performance of two machine learning classification methods against stunting prevalence data, the Naïve Bayes method and Decision Tree C4.5. These two methods are compared to determine which algorithm is more efficient with the highest accuracy.

Research related to stunting data has also been carried out by previous researchers such as a study conducted by Obvious Nchimunya Chilyabanyama et al, entitled "Performance of Machine Learning Classifiers in Classifying Stunting among Under-Five Children in Zambia" resulted Random Forest (highest) produces 79% accuracy and Naïve Bayes (lowest) produces 61.6% accuracy. Research has used four machine learning methods but has not used an artificial neural network approach [12].

The research conducted by Md. Merajul Islam, et al, entitled "Application of machine learning based algorithm for prediction of malnutrition among women in Bangladesh" resulted the random forest method provided an accuracy of 81.4% for underweight and an accuracy of 82.4% for overweight [13]. The research conducted by Fikrewold, et al, entitled "Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia" resulted xgbTree algorithm is a superior machine learning algorithm for predicting childhood malnutrition in Ethiopia compared to other machine learning methods [14].

The research conducted by Haile Mekonnen Fenta et al, entitled "A machine learning classifier approach for identifying the determinants of under-five child undernutrition in Ethiopian administrative zones" resulted the random forest algorithm was selected as the best ML model. In the order of importance; urban–rural settlement, literacy rate of parents, and place of residence were the major determinants of disparities of nutritional status for under-five children among Ethiopian administrative zones [15].

The research conducted by Elizabeth Harrison et al., entitled "Machine learning model demonstrates stunting at birth and systemic inflammatory biomarkers as predictors of subsequent infant growth a four-year prospective study" resulted in the random forest models identified HAZ at birth as the most crucial feature in predicting HAZ at 18 months. Of the biomarkers, AGP (Alpha-1-acid Glycoprotein), CRP (C-Reactive Protein), and IL1 (interleukin-1) were identified as subsequent solid growth predictors across both the classification and regressor models. [16].

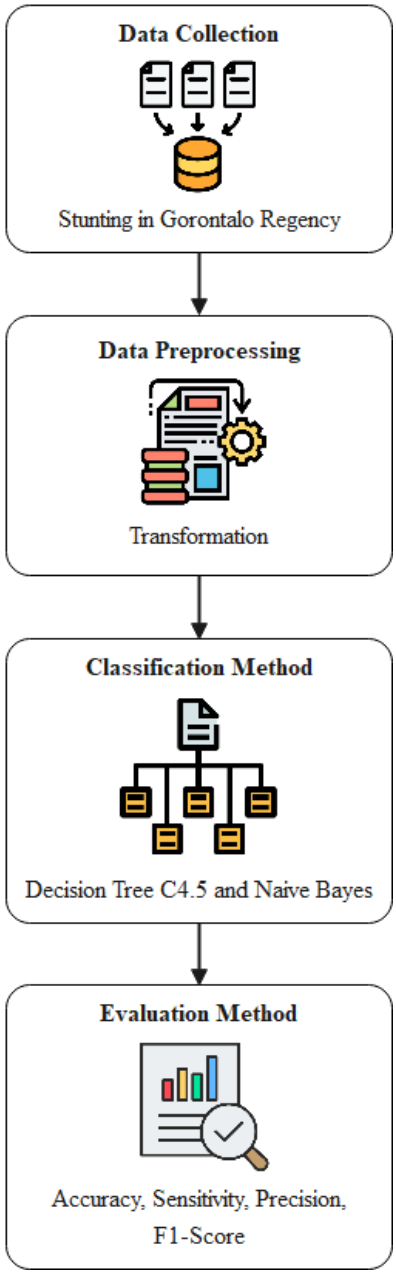

Figure 1. Research Stages.

Table 1. Sample Dataset

No	Gender	Age	Weight	Height	Nutritional
1	Female	26	10	84	Normal
2	Male	26	9	81	Stunting
3	Female	39	11	87	Stunting
4	Male	21	9	81	Normal
5	Female	25	8.5	77	Stunting
6	Male	22	10	82	Normal
7	Female	19	7	73	Stunting
8	Male	13	8.9	74	Normal
9	Female	17	8.5	73	Stunting
10	Male	29	7.8	85	Normal
….	….	….	….	….	….
224	Male	15	7.5	72	Stunting

2. Materials and Method

This research consists of several stages, as shown in Figure 1.

Figure 1 describes several stages in this study, ranging from analyzing the problem, collecting data, processing data, and analyzing data to obtaining the expected output. The desired outcome is to compare the Decision Tree C4.5 and Naïve Bayes methods in children's classification of stunting. Based on the performance obtained, the method with the highest accuracy is expected to be known. Identifying problems in the performance of Machine Learning methods for stunting classification in children is how to determine the most effective method of classification stunting with a high level of accuracy and minimize overfitting of the data.

2.1. Data Collection

The dataset used in this study is a prevalence stunting in Gorontalo Regency based on the anthropometric data and calculation of Z-Score TB/U [17] with a total sampling data of 224 records. The data consists of 4 attributes and 1 label namely Gender, Age, Weight, Height and Nutritional Status. With details of 122 Normal and 122 Stunting data. The data types for each are shown in the following table 1.

2.2. Data Pre-processing

Data pre-processing helps improve the data's quality and reliability, removes inconsistencies and outliers, and makes the data compatible with the algorithms and models used [18]. At this stage, researchers transform the data by initializing 0 and 1 to data with categorical values such as gender attributes. Then, the researcher split the data by dividing the training and testing data with a total ratio of 80:20.

2.3. Decision Tree C4.5 Method

The C4.5 algorithm, discovered by John Ross Quinlan in 1986, is a development of the ID3 algorithm. In ID3, decision tree induction can only be done on categorical (nominal or ordinal) feature types, while numeric types (interval or ratio) cannot be used [19]. Unlike the ID3 algorithm, which can only be used for categorical (nominal or ordinal) feature types, the C4.5 algorithm, developed by John Ross Quinlan (1986), can be used for numeric data by building threshold values and sorting data into a number of intervals to obtain categorical values [20].

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (1)$$

Description:
S : Case set
A : Attribute
N : Number of partitions attribute A
|Si| : number of cases on partition to i
|S| : number of cases in S

Before getting the gain value is to find the entropy value. Entropy is used to determine how informative an attribute input is to produce an attribute. The basic formula of entropy is as follows [21]:

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2\, pi \quad (2)$$

Description
S : Case Set
n : Number of partitions S
pi : Proportion of Si to S

To calculate the Gain Ratio, you must first calculate the split information formulated as follows [22].

$$SplitInformation(S, A) \sum_{i=1}^{n} -\frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|} \quad (3)$$

Where S represents the data sample set, Si represents a subset of the data sample that is divided based on the number of value variations in attribute A. Next, the Gain Ratio is formulated as Information Gain divided by SplitInformation, which is:

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInformation(S, A)} \quad (4)$$

2.4. Naïve Bayes Method

The equation of Bayes' theorem is based on the following general formula [23].

$$P(H|\,X) = \frac{P(H|\,X) * P(H)}{P(X)} \quad (5)$$

Description

X	:	Data with unknown classes.
H	:	The X data hypothesis is a specific class.
P (H|X)	:	The probability of hypothesis H based on condition X (posteriori probability).
P (H)	:	The probability of hypothesis H (prior probability).
P (X|H)	:	Probability X based on conditions on hypothesis H.
P (X)	:	Probability X.

The basic idea behind Bayes' rule is that the outcome of a hypothesis or event (H) can be predicted based on some observed evidence (X). In general, naïve bayes for categorical type attributes are easy to calculate. However,

there is a special treatment for numerical (continuous) features before they are integrated into Naive Bayes, namely through the use of probability density functions [24].

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (6)$$

$$\sigma = \left[\frac{i}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2\right]^{0.5} \quad (7)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (8)$$

Description
μ : Mean
σ : Standard deviation
$f(x)$: Normal distribution

2.5. Evaluation Performance

Evaluating the performance of a machine learning model is crucial to understanding how well it is performing on unseen data. There are several commonly used metrics to evaluate the performance of classification models, namely the confusion matrix, as in Table 2. [25].

The formula used to calculate accuracy, Sensitivity, Precision, and F1-score [26].

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (9)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (10)$$

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$F1 - score = \frac{2 \times precision \times recall}{precision \times recall} \quad (12)$$

True Positive (TP) is a stunting positive class that is correctly predicted. False Positive (FP) is a stunting negative class that is predicted to be stunting positive. True Negative (TN) is a stunting negative class that is correctly predicted. False Negative (FN) is a stunting positive class that is predicted to be stunting negative.

3. Result and Discussion

This research begins with data collection, pre-processing, classification and performance testing. Data used in this study is a prevalence stunting dataset in Gorontalo Regency based on anthropometric data. The classification method of this research is Decision Tree C4.5 and Naive Bayes. The performance test uses a confusion matrix based on accuracy, sensitivity, precision, and score. Based on testing the Decision Tree C4.5 and Naive Bayes, the results were obtained as a confusion matrix, as shown in Figure 3 for the C4.5 method. Figure 4 for the results of the Naive Bayes method. The result of the comparison performance of the Decision Tree C4.5 and Naïve Bayes method is shown in Figure 2.

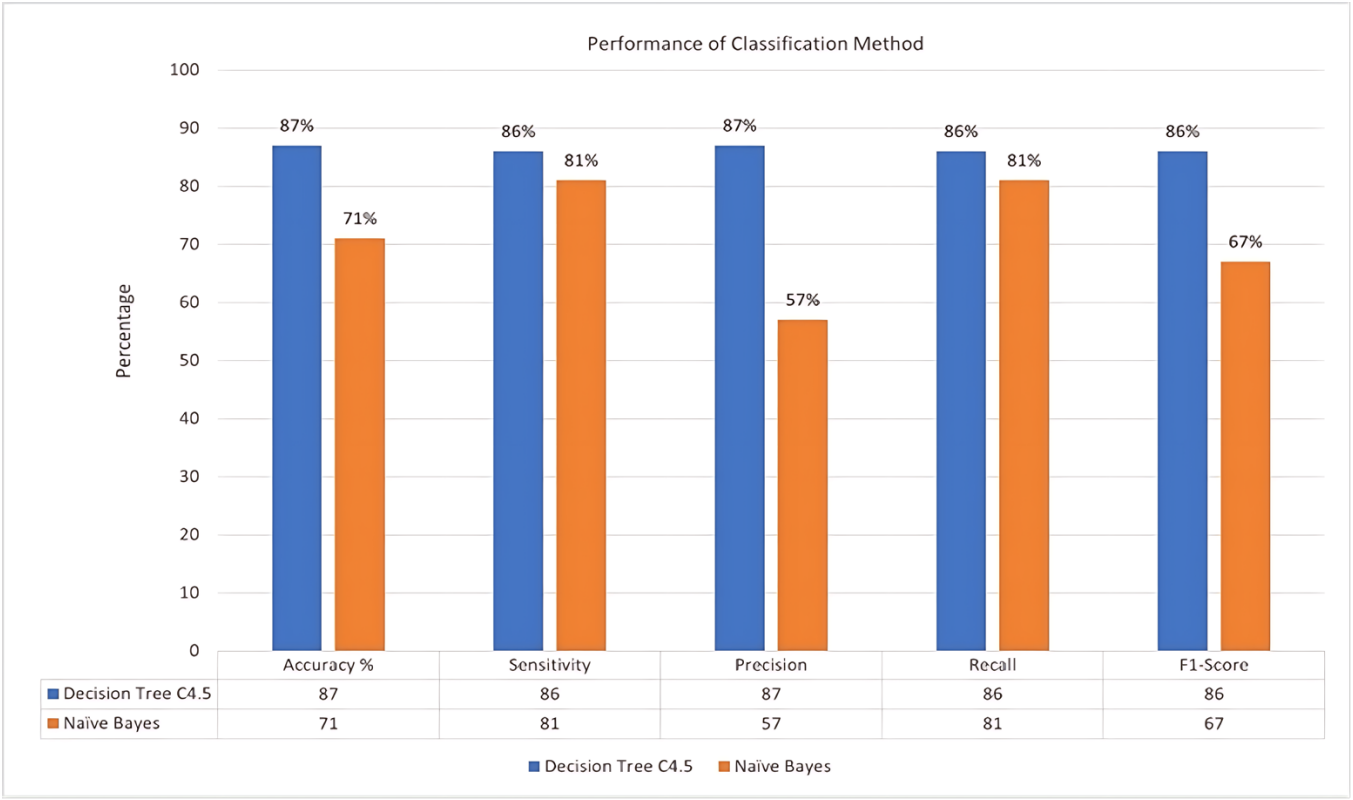

Figure 2. Result Performance of Classification Method.

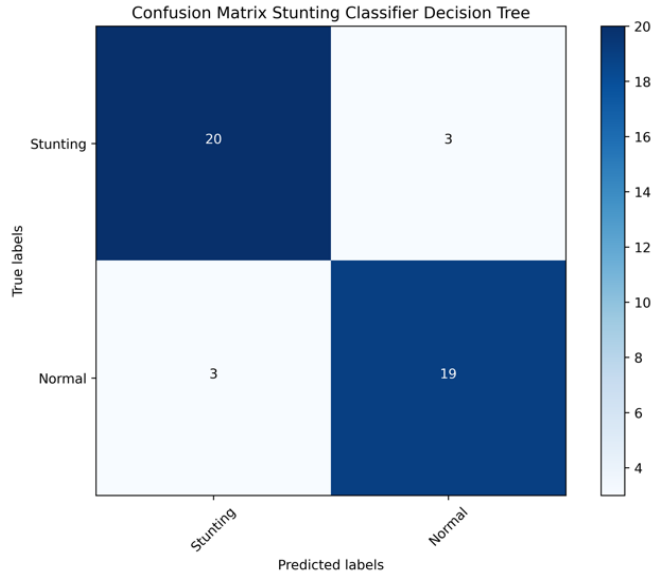

Figure 3. Result in Confusion Matrix of Decision Tree C4.5.

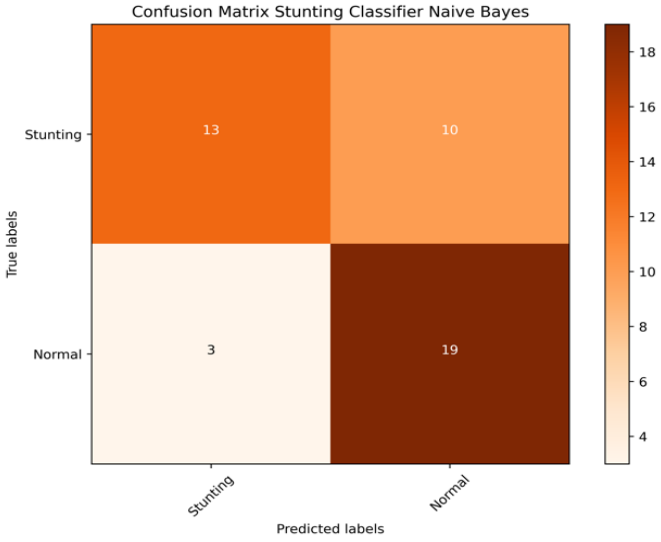

Figure 4. Result in Confusion Matrix of Naïve Bayes.

Table 2. Confusion Matrix

Actual	Predicted	
	Positive	**Negative**
Positive	TP	FP
Negative	FN	TN

In Figure 3, the Decision Tree C4.5 method correctly classified the positive class (TP) as many as 20 stunting and the negative class classified incorrectly (FP) as many as 3 stunting. While the correctly classified negative class (TN) is 19 normal, and the incorrectly classified positive class (FN) is 3 normal.

In Figure 4, the Naïve Bayes method correctly classified the positive class (TP) is 13 stunting children and incorrectly classified the negative class (FP) in 10 stunting children. The correctly classified negative class (TN) is 19 normal and the incorrectly classified positive class (FN) is 3 normal.

Based on Figure 2, the performance of the C4.5 decision tree method is based on accuracy, sensitivity, precision and F1-score. In the original dataset, the Decision Tree C4.5 method has 87% accuracy, 86% sensitivity, 87% precision, 86% recall and 86% f1-score. The Naive Bayes method has an accuracy of 71%, a sensitivity of 81%, a precision of 57%, a recall of 81% and f1-score of 67%.

4. Conclusion

This study compares the accuracy performance of each machine learning method, namely the Decision Tree C4.5 and Naïve Bayes methods. Based on testing these two methods, the Decision Tree C4.5 method has a higher accuracy of 87%. Meanwhile, the lowest accuracy was obtained in the Naïve Bayes method, with an accuracy of 71%. Precision is significant in improving the Accuracy and F1-score performance of the method Decision Tree C4.5.

5. Conflicts of Interest

The authors declare no conflicts of interest.

6. References

[1] J. Liu, J. Sun, J. Huang, and J. Huo, "Prevalence of Malnutrition and Associated Factors of Stunting among 6–23-Month-Old Infants in Central Rural China in 2019," *Int J Environ Res Public Health*, vol. 18, no. 15, p. 8165, Aug. 2021, doi: 10.3390/ijerph18158165.

[2] M. S. Islam, A. N. Zafar Ullah, S. Mainali, Md. A. Imam, and M. I. Hasan, "Determinants of stunting during the first 1,000 days of life in Bangladesh: A review," *Food Sci Nutr*, vol. 8, no. 9, pp. 4685–4695, Sep. 2020, doi: 10.1002/fsn3.1795.

[3] A. D. Laksono, N. E. W. Sukoco, T. Rachmawati, and R. D. Wulandari, "Factors Related to Stunting Incidence in Toddlers with Working Mothers in Indonesia," *Int J Environ Res Public Health*, vol. 19, no. 17, p. 10654, Aug. 2022, doi: 10.3390/ijerph191710654.

[4] T. Huriah and N. Nurjannah, "Risk Factors of Stunting in Developing Countries: A Scoping Review," *Open Access Maced J Med Sci*, vol. 8, no. F, pp. 155–160, Aug. 2020, doi: 10.3889/oamjms.2020.4466.

[5] M. Ponum *et al.*, "Stunting diagnostic and awareness: Impact assessment study of sociodemographic factors of stunting among school-going children of Pakistan," *BMC Pediatr*, vol. 20, no. 1, 2020, doi: 10.1186/s12887-020-02139-0.

[6] S. Khan, S. Zaheer, and N. F. Safdar, "Determinants of stunting, underweight and wasting among children < 5 years of age: evidence from 2012-2013 Pakistan demographic and health survey," *BMC Public Health*, vol. 19, no. 1, p. 358, Dec. 2019, doi: 10.1186/s12889-019-6688-2.

[7] E. Yunitasari, R. Pradanie, H. Arifin, D. Fajrianti, and B.-O. Lee, "Determinants of Stunting Prevention among Mothers with Children Aged 6–24 Months," *Open Access Maced J Med Sci*, vol. 9, no. B, pp. 378–384, May 2021, doi: 10.3889/oamjms.2021.6106.

[8] T. Antoniou and M. Mamdani, "Evaluation of machine learning solutions in medicine," *Can Med Assoc J*, vol. 193, no. 36, pp. E1425–E1429, Sep. 2021, doi: 10.1503/cmaj.210036.

[9] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput Sci*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.

[10] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP J Adv Signal Process*, vol. 2016, no. 1, p. 67, Dec. 2016, doi: 10.1186/s13634-016-0355-x.

[11] S. Ray, "A Quick Review of Machine Learning Algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, IEEE, Feb. 2019, pp. 35–39. doi: 10.1109/COMITCon.2019.8862451.

[12] O. N. Chilyabanyama *et al.*, "Performance of Machine Learning Classifiers in Classifying Stunting among Under-Five Children in Zambia," *Children*, vol. 9, no. 7, Jul. 2022, doi: 10.3390/children9071082.

[13] M. M. Islam *et al.*, "Application of machine learning based algorithm for prediction of malnutrition among women in Bangladesh," *International Journal of Cognitive Computing in Engineering*, vol. 3, pp. 46–57, Jun. 2022, doi: 10.1016/j.ijcce.2022.02.002.

[14] F. H. Bitew, C. S. Sparks, and S. H. Nyarko, "Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia," *Public Health Nutr*, vol. 25, no. 2, pp. 269–280, Feb. 2022, doi: 10.1017/S1368980021004262.

[15] H. M. Fenta, T. Zewotir, and E. K. Muluneh, "A machine learning classifier approach for identifying the determinants of under-five child undernutrition in Ethiopian administrative zones," *BMC Med Inform Decis Mak*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12911-021-01652-1.

[16] E. Harrison *et al.*, "Machine learning model demonstrates stunting at birth and systemic inflammatory biomarkers as predictors of subsequent infant growth – a four-year prospective study," *BMC Pediatr*, vol. 20, no. 1, Dec. 2020, doi: 10.1186/s12887-020-02392-3.

[17] M. de Onis *et al.*, "Prevalence thresholds for wasting, overweight and stunting in children under 5 years.," *Public Health Nutr*, vol. 22, no. 1, pp. 175–179, Jan. 2019, doi: 10.1017/S1368980018002434.

[18] S.-A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Data preprocessing in predictive data mining," *Knowl Eng Rev*, vol. 34, p. e1, Jan. 2019, doi: 10.1017/S026988891800036X.

[19] J. R. Quinlan, *Induction of Decision Trees*, vol. 1, no. 1. Machine Learning, 1986. doi: 10.1023/A:1022643204877.

[20] Y. Lu, T. Ye, and J. Zheng, "Decision Tree Algorithm in Machine Learning," in *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, IEEE, Aug. 2022, pp. 1014–1017. doi: 10.1109/AEECA55500.2022.9918857.

[21] J. Jiang, X. Zhu, G. Han, M. Guizani, and L. Shu, "A Dynamic Trust Evaluation and Update Mechanism Based on C4.5 Decision Tree in Underwater Wireless Sensor Networks," *IEEE Trans Veh Technol*, vol. 69, no. 8, pp. 9031–9040, Aug. 2020, doi: 10.1109/TVT.2020.2999566.

[22] Y. Wang, "Prediction of Rockburst Risk in Coal Mines Based on a Locally Weighted C4.5 Algorithm," *IEEE Access*, vol. 9, pp. 15149–15155, 2021, doi: 10.1109/ACCESS.2021.3053001.

[23] Y.-C. Zhang and L. Sakhanenko, "The naive Bayes classifier for functional data," *Stat Probab Lett*, vol. 152, pp. 137–146, Sep. 2019, doi: 10.1016/j.spl.2019.04.017.

[24] S. Ruan, H. Li, C. Li, and K. Song, "Class-Specific Deep Feature Weighting for Naïve Bayes Text Classifiers," *IEEE Access*, vol. 8, pp. 20151–20159, 2020, doi: 10.1109/ACCESS.2020.2968984.

[25] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Inf Sci (N Y)*, vol. 507, pp. 772–794, Jan. 2020, doi: 10.1016/j.ins.2019.06.064.

[26] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.