Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26–28 February 2024; pp 197–206

Predicting Stunting in Under-Five Children in Low Socio-Demographic Index States of India: A Machine Learning Approach

Mukesh Vishwakarma[1], **Gargi Tyagi**[1], **and Pawan Kumar Dubey**[2]
[1]*Department of Mathematics and Statistics, Banasthali Vidyapith, Rajasthan.*
[2]*Monitoring & Evaluation, India Health Action Trust, Lucknow, Uttar Pradesh.*

Received: 10 August 2024; Revised: 01 September 2024; Accepted: 06 September 2024

Abstract

Stunting, the most prevalent form of child malnutrition, is characterized by a lack of height relative to age in children. Globally, 5.2 million children under five dies, with catastrophic stunting rates in central and southern Asia. Pakistan, India, and Afghanistan have the highest malnutrition rates in South Asia. India has about 270 million poor individuals and one-third of malnourished children. The National Family Health Survey (NFHS-5), 2019-21 found that 35.5% of Indian children under five are stunted. Factors influencing nutritional status include social, economic, educational, and maternal health issues. Despite efforts, India struggles to significantly curb child undernutrition, especially in states like Uttar Pradesh, Bihar, Rajasthan, Madhya Pradesh, Chhattisgarh, and Jharkhand, classified as low SDI states based on their Socio-Demographic Index (SDI) in the Global Burden of Disease study for 2019. The objective of this study is to train and evaluate machine learning (ML) classification algorithms on the National Family Health Survey (NFHS-5), 2019-21 dataset for predicting stunting among children under five years of age in states with a low socio-demographic index (LSDI). The machine learning models applied in this study include logistic regression (LR), random forest (RF), support vector classification (SVC), decision tree classifier (DTC), and gradient boosting classifier (GBC) algorithms. The performance of the ML algorithms are evaluated and compared using accuracy, recall, precision, F1-score, receiver operating curve (ROC) and recall curves on test dataset and 5-fold cross validation dataset. Important features of childhood stunting are also identified using Random Forest algorithm. It is observed that out of 82,158 children, 39% were stunted. Among the algorithms applied, the GBC algorithm achieved the highest accuracy in predicting stunting, with 65.5% on the testing data and 65 % $\pm$ 7% on the 5-fold cross validation data. In LSDI states of India, social structure and mother education are found to be major predictors of stunting in children under five, according to the random forest model for features importance. These results can aid in the swift diagnosis of stunting and the prompt development of preventive measures.

Key words: Stunting; Malnutrition; Child development; Healthy growth; Machine learning.

AMS Subject Classifications: 62K05, 05B05

Corresponding Author: Mukesh Vishwakarma
Email: vishwakarma.m96@gmail.com

1. Introduction

Stunting is one of the most serious health and welfare issues across the world with more than 149 million children, which accounts for 21% of all children under the age of five, suffer from stunted growth. Moreover, the majority, 91%, of these children reside in low- and middle-income countries (LMICs) like India (UNICEF, 2020). Stunting is a condition that occurs when children suffer from prolonged inadequate nutrition and is defined as weight-for-height < -2 SD (standard deviation) in the WHO Growth Standard median (WHO, 2019).

Between 2005–06 and 2015–16, India has made a modest decrease in the prevalence of stunting and underweight in children under five years, but the progress is insufficient compared to its economic growth. Although there was a moderate decline in child undernutrition during this period, over one-third of children under five years old remain stunted (Jose *et al.*, 2018). This situation of stunting in India is quite evident in highly populated regions, namely Uttar Pradesh, Bihar, Rajasthan, and Madhya Pradesh *etc.* These states also come at lower end of Socio demographic Index (SDI) paradigm. The Socio-demographic Index (SDI) is a composite index of development that is significantly associated with health impact. It represents the geometric mean of the indices ranging from 0 to 1 for mean education among individuals aged 15 or older (EDU15+), total fertility rate under 25 (TFU25), and lag distributed income (LDI) per capita. A location with an SDI of 0 has a theoretical minimal level of health-related development, whereas a location with an SDI of 1 has a theoretical maximum level (Global Burden of Disease Collaborative Network, 2020). The SDI quantiles are utilized for classification. Based on their SDI, the states Uttar Pradesh, Bihar, Rajasthan, Madhya Pradesh, Chhattisgarh, and Jharkhand fall into the category of LSDI states.

Over the years, classical statistical models have been utilised to discover characteristics that are autonomously linked to stunting in children under the age of five (Mzumara *et al.*, 2018; Rakotomanana *et al.*, 2017; Das and Gulshan, 2017). However, these methods are not reliable in instances where the number of covariates exceeds the number of observations and when there is multicollinearity among variables. In addition, these models adhere to stringent assumptions regarding the data and the method by which the data is generated. These assumptions include the distribution of errors and the linearity of parameters with linear predictors. However, it is important to note that these assumptions may not be valid in real-world scenarios (Rajula *et al.*, 2020). Machine learning approaches surpass traditional models by addressing the analytical difficulties associated with a large number of covariates and multicollinearity. They also require fewer assumptions and can handle high dimensional data, resulting in a more adaptable relationship between predictor and outcome variables (Iniesta *et al.*, 2016). These techniques have been utilised to forecast malnutrition by employing various datasets (Shahriar *et al.*, 2019; Jin *et al.*, 2020; Markos *et al.*, 2014; Talukder and Ahammed, 2020). Moreover, machine learning techniques have demonstrated their superiority over traditional statistical methods in solving categorisation difficulties.

In this study, we focused on training, evaluating, and selecting the optimal machine learning classifier to predict stunting in children under five years old in low sociodemographic index (LSDI) states of India. Utilizing data from the NFHS-5 (2019-21), we also aimed to identify key variables that contribute to stunting. This model is intended to lay the

groundwork for creating an intelligent system for diagnosing or predicting stunting. The identified predictors of stunting will be prioritized in the development of interventions aimed at preventing stunting among children under five in LSDI states of India.

2. Materials and methods

2.1. Data source

The data used in this study is obtained from the Children's Recode (KR) dataset of the National Family Health Survey (NFHS-5), conducted from 2019 to 2021. This dataset includes one record for every child born to interviewed women within the five years preceding the survey consisting 232920 children. Unit-level data is accessible through the Demographic Health Survey (DHS) data repository, and requests for access can be made via the DHS Program website (www.dhsprogram.com/data/). The unit of analysis in this research encompasses 424 districts from LSDI states of India, including a total of 104692 children under the age of five years at the time of the survey.

2.2. Data pre-processing

The target variable in this study was stunting, defined according to the WHO standard as a height-for-age Z-score (HAZ) of less than -2 standard deviations (SD) (WHO, 2019). Various socioeconomic, demographic, and environmental factors were considered as features in the analysis. Missing instances for each variable included in the study were excluded from the analysis. After removing the missing observations, we were left with 82,158 cases. Further, to prepare categorical features for machine learning, the traditional one-hot encoding technique was utilized, in which multicategorical variables are transformed into several binary feature vectors. The continuous variables in the study were standardized.

2.3. Feature selection

Random Forest (RF) feature selection was employed to identify the most significant features. As suggested by Talukder and Ahammed (2020) for the application of RF feature selection in constructing predictive models for malnutrition. This method assigns an importance score to each feature, and those with scores below the average were excluded from the model. Figure 2 illustrates the importance scores associated with each feature.

2.4. Model training

The data was divided into a training dataset including 70% of the total data, and a testing dataset comprising the remaining 30%. We have employed five commonly used machine learning classifiers, namely, logistic regression (LR), random forest (RF), decision tree (DT), support vector classifier (SVC), and gradient boosting (Géron, 2022). The algorithms are implemented using scikit-learn library in Python (Pedregosa *et al.*, 2011).

2.5. Evaluation of model performance

The model's performance on the test set was evaluated using metrics like as accuracy, precision, recall, area under the curve (AUC), and F1 score.

Suppose an output variable has two classes, namely, positive and negative classes. A confusion matrix is a square matrix that includes the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (PN) values. These values can be used to calculate the one-dimensional performance measures (Luque *et al.*, 2019).

Accuracy: It is the ratio of events that have been correctly classified to the total number of cases in the dataset.

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

Precision: It measures the ratio of true positive correct predictions by the classifier out of all positive predictions

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

Recall: It measures the percentage of real positive predictions out of all actual positive instances in the dataset.

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

F_1 **score:** The F_1 score is computed by taking the harmonic mean of the precision and recall.

$$F_1 \text{ score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Area under the Curve (AUC): The area under the curve (AUC) is the are under the receiver operating characteristic (ROC) curve. ROC curve is a graphical representation of the true positive rate (TPR) vs the false positive rate (FPR) at different thresholds. The AUC provides an assessment of the overall model performance at every potential classification threshold. It measures the degree of separability between positive and negative classes. The value of AUC lies between 0 and 1. The higher the value, the better a model can distinguish between positive and negative classes.

3. Results

Table 1 shows that the prevalence of stunting in rural areas is 40.15%, which is higher compared to 32% in urban areas. Children whose mothers have no education exhibit the highest prevalence of stunting at 46.9%, followed by those with only primary education at 42.9%. Households lacking improved sanitation facilities and those using unclean cooking fuel have stunting prevalences of 45.3% and 41.6%, respectively. The prevalence of stunting is highest among Muslim children (40.7%), followed by Hindu children (38.3%) and those of other religions (37.1%). In Scheduled Castes, the prevalence is 40.5%. The average age of the mothers having stunted child is 27 years, while the average weight of the stunted children in the study is 10 kg.

Various machine learning models have been applied to predict the likelihood of stunting and the predictive performance of each classifier on the test and 5-fold cross-validation datasets. The results have been presented in Table 2.

Table 1: Prevalence of stunting in children under 5 in low sociodemographic index states of India by characteristics; NFHS 2019-2021

Variable	Category	Frequency / Mean	Percentage (%) / SD
Place of residence	Urban	4933	32
	Rural	26735	40.1
Highest educational level	No education	11291	46.9
	Primary	4966	42.9
	Secondary	13003	35.5
	Higher	2409	24.5
Sanitation facility (Improved)	No	12128	45.3
	Yes	19541	35.3
Clean Cooking fuel	No	21816	41.6
	Yes	9852	33.2
Religion	Hindu	26964	38.3
	Muslim	4199	40.7
	Others	505	37.1
Caste	SC	8274	43.2
	ST	4086	40.8
	OBC	15484	37.8
	Other	3825	31.7
Media Exposure	No	13991	45.4
	Yes	17678	34.4
Wealth index	Poorest	12796	47.3
	Poorer	8095	40.4
	Middle	5077	36
	Richer	3485	29.8
	Richest	2216	23.9
Mother Anemia	No	11803	36.7
	Yes	19865	39.8
Birth Order	1	9558	34.6
	2	9629	37.2
	3	6159	41.6
	4 or more	6322	45.9
Sex of child	Male	16822	39.4
	Female	14846	37.6
Skilled Birth Attendant	No	5179	45.4
	Yes	26490	37.4
Institutional Births	No	5811	47
	Yes	25857	37
Delivery by Caesarean Section	No	28849	39.8
	Yes	2819	29.1
Size of child at birth	No	5072	37.8
	Yes	3695	43.2
Birth Weight less than 2.5 kg	No	25927	37.4
	Yes	5742	44.7
Infectious diseases in past 2 weeks	No	24944	38.4
	Yes	6725	39.3
Child immediately put on chest after the birth?	No	5473	37.3
	Yes	26196	38.8
Distance to health facility is a big problem	No	22449	37.6
	Yes	9220	41.1
Mother Age (in years)		27	5
Respondent's height (in cm)		149.9	5.99
Mother BMI		20.72	3.35
Mother haemoglobin level (g/dl)		11.24	1.58
Child's weight (in kg)		10.08	2.86

Table 2: Summary of classification model performance in predicting stunting

Model		Accuracy	Precision	Recall	f1-score	AUC
Logistic Regression	5-folds cross-validation	0.62±0.10	0.29±0.05	0.55±0.03	0.38±0.017	0.64±0.31
	Test data	0.633	0.298	0.560	0.389	0.658
Random Forest Classifier	5-folds cross-validation	0.65±0.10	0.39±0.12	0.58±0.14	0.47±0.12	0.68±0.11
	Test data	0.658	0.394	0.598	0.475	0.690
Support Vector Classifier	5-folds cross-validation	0.65±0.05	0.32±0.12	0.60±0.13	0.42±0.11	0.68±0.16
	Test data	0.657	0.332	0.617	0.432	0.682
Gradient Boosting Classifier	5-folds cross-validation	0.65±0.07	0.40±0.10	0.59±0.13	0.49±0.11	0.70±0.06
	Test data	0.665	0.412	0.608	0.492	0.702
Decision Tree	5-folds cross-validation	0.56±0.11	0.45±0.13	0.46±0.14	0.45±0.12	0.54±0.17
	Test data	0.575	0.473	0.459	0.466	0.557

It can be seen from the Table 2 that the Decision Tree classifier was the least accurate model on both the test and cross-validation sets, with accuracies of 57.5% and 56 ± 11%, respectively. In contrast, the Gradient Boosting Classifier was the most effective model, achieving accuracies of 66.5% on the test set and 65 ± 7% on the cross-validation set. The Gradient Boosting model also had the highest F1 score at 49.2%, while Logistic Regression had the lowest F1 score in both the testing and 5-fold cross-validation. The Random Forest and Support Vector Machine classifiers had accuracy scores of 65.8% and 65.7%, respectively, on the test dataset, with F1 scores of 47.5% and 43.2%, respectively (Table 2). The area under the curve (AUC) was highest in the Gradient Boosting model at 70%, followed by the Random Forest and Support Vector Classifier at 69% and 68%, respectively, and was the lowest in the Decision Tree model.

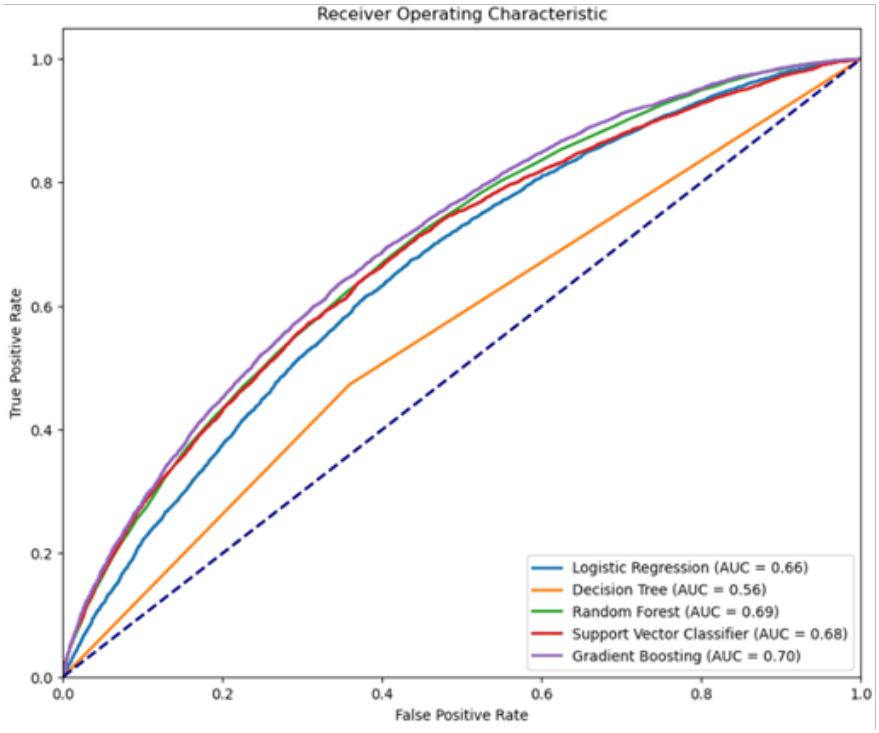

Figure 1: Receiver operating characteristic curve of stunting model

The ROC curve analysis revealed that the Gradient Boosting model achieved the highest Area Under the Curve (AUC) at 0.70, indicating superior performance in distinguishing between the positive and negative classes, followed by the Random Forest (AUC =

0.69) and Support Vector Classifier (AUC = 0.68). Logistic Regression and Decision Tree had lower AUC values of 0.66 and 0.56, respectively. The Decision Tree model performed the worst showing the lowest precision and recall values, indicating it was less effective in handling imbalanced datasets. These results underscore the robustness of the Gradient Boosting model in providing a balanced trade-off between precision and recall, making it the most reliable model among those evaluated (Figure 1).

Further, the importance scores of the determinants in the Random Forest model are presented in Figure 2.

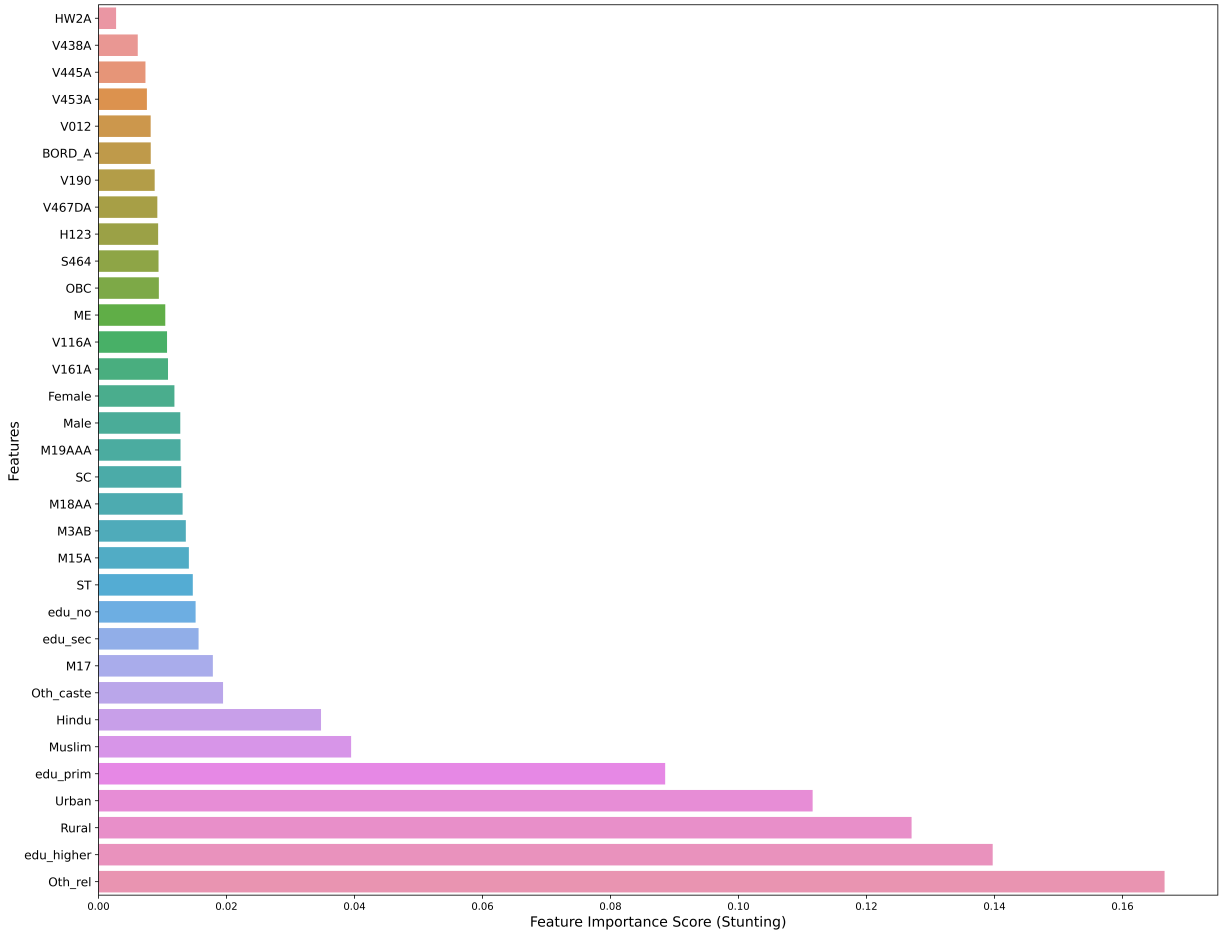

Figure 2: Features importance score

From the Figure 2, the key determinants of stunting can be identified, with the top variables being "Other" Religion (0.16), "higher education of mother" (0.13), and "Rural" place of residence (0.13). Other significant factors include "mother primary education" (0.10), and "Other caste" (0.07). These results highlight the critical roles of education level, religious affiliation, and place of residence in influencing stunting outcomes. Understanding these variables can guide targeted interventions to address and reduce stunting in affected populations.

4. Discussion

Our study found that the Gradient Boosting Classifier had the highest predictive accuracy and precision for stunting among children under five years in low socio-demographic index states of India, using the fifth round of NFHS data. The Gradient Boosting Classifier achieved the largest Area Under the Curve (AUC), suggesting its superior ability to distinguish between positive and negative classes. The Random Forest also performed well, although not as well as the Gradient Boosting Classifier.

In addition, our study has revealed significant factors for predicting stunting in children under the age of five in low socio-demographic index states of India. Some of the identified top features include religions other than Hindu and Muslim, mother higher education level, and a rural place of residence. Other significant factors include mother primary education and other castes (other than SC/ST and OBC). These were identified commonly as predictors of stunting in other studies also (Mzumara *et al.*, 2018; Shahriar *et al.*, 2019; Mediani, 2020; Bwalya *et al.*, 2015).

An important advantage of this study is its use of the NFHS dataset. The NFHS employs a robust sampling procedure, providing a reliable representation of the under-five population in both rural and urban areas of low socio-demographic index states in India. Additionally, the model's accuracy was assessed using test data, and 5-fold cross-validation was employed to prevent overfitting. However, there are still certain possible constraints that need to be considered. Consequently, the interpretability of the findings remains constrained. Furthermore, despite our efforts to incorporate a wide range of covariates, we were unable to eliminate the possibility of residual confounding resulting from unmeasured factors, such as the mother's height and weight. Additionally, certain details about the children were obtained from their mothers' memories, such as instances of diarrhoea and fever in the past two weeks. However, it is possible that there is a bias in these recollections.

5. Conclusion

In this study, several machine learning models have been employed to predict the prevalence of stunting in children under 5 years of age in LSDI states on India. The performance of the models was compared using various performance metrics. The results suggest that the Gradient Boosting Classifier model has highest predictive accuracy for stunting compared to the other applied models in this study. Feature importance is also studied using the random forest model. It suggests that social structure, mother education are the important predictors of stunting among the children under five in low socio demographic index states of India.

References

Bwalya, B. B., Lemba, M., Mapoma, C. C., and Mutombo, N. (2015). Factors associated with stunting among children aged 6-23 months in Zambian: evidence from the 2007 Zambia demographic and health survey. *International Journal of Advanced Nutritional and Health Science*, **3**, 116–31.

Das, S. and Gulshan, J. (2017). Different forms of malnutrition among under five children in Bangladesh: a cross sectional study on prevalence and determinants. *BMC Nutrition*, **3**, 1–12.

Géron, A. (2022). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Global Burden of Disease Collaborative Network (2020). *Global Burden of Disease Study 2019 (GBD 2019) Socio-Demographic Index (SDI) 1950-2019*. Institute for Health Metrics and Evaluation, Seattle, USA.

Iniesta, R., Stahl, D., and McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, **46**, 2455–2465.

Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., and Qiang, B. (2020). RFRSF: Employee turnover prediction based on random forests and survival analysis. In Huang, Z., Beek, W., Wang, H., Zhou, R., and Zhang, Y., editors, *Web Information Systems Engineering – WISE 2020*, pages 503–515. Springer International Publishing, Cham.

Jose, S., Bheemeshwar, R. A., and Agrawal, M. (2018). Child undernutrition in India. *Economic and Political Weekly*, **53**, 63–70.

Luque, A., Carrasco, A., Martín, A., and de Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, **91**, 216–231.

Markos, Z., Doyore, F., Yifiru, M., and Haidar, J. (2014). Predicting Under nutrition status of under-five children using data mining techniques: The Case of 2011 Ethiopian Demographic and Health Survey. *Journal of Health & Medical Informatics*, **5**, 152.

Mediani, H. S. (2020). Predictors of stunting among children under five year of age in Indonesia: a scoping review. *Global Journal of Health Science*, **12**, 83–95.

Mzumara, B., Bwembya, P., Halwiindi, H., Mugode, R., and Banda, J. (2018). Factors associated with stunting among children below five years of age in Zambia: evidence from the 2014 Zambia demographic and health survey. *BMC Nutrition*, **4**, 51.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, **12**, 2825–2830.

Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., and Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*, **56**, 455.

Rakotomanana, H., Gates, G. E., Hildebrand, D., and Stoecker, B. J. (2017). Determinants of stunting in children under 5 years in Madagascar. *Maternal & Child Nutrition*, **13**, e12409.

Shahriar, M. M., Iqubal, M. S., Mitra, S., and Das, A. K. (2019). A Deep Learning Approach to Predict Malnutrition Status of 0-59 Month's Older Children in Bangladesh. In *2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pages 145–149. IEEE.

Talukder, A. and Ahammed, B. (2020). Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. *Nutrition*, **78**, 110861.

UNICEF (2020). *Nutrition, for Every Child UNICEF Nutrition Strategy 2020-2030*. UNICEF, New York.

WHO (2019). Nutrition Landscape Information System (NLIS) country profile indicators: interpretation guide. Technical report, World Health Organization.