Article

Machine learning model for predicting fetal nutritional status

B. Selemani, **D. Machuve**, **N. Mduma**
Nelson Mandela African Institution of Science and Technology, Arusha Tanzania
E-mail: selemanib@nm-aist.ac.tz; dina.machuve@nm-aist.ac.tz, neema.mduma@nm-aist.ac.tz

Received 25 April 2023; *Accepted 20 July 2023*; *Published online 23 October 2023*; *Published 1 March 2024*

Abstract
Malnutrition tends to be one of the most important reasons for child mortality in Tanzania and other developing countries, in most cases during the first five years of life. This research was conducted todevelop machine learning model for predicting fetal nutritional status. Several machine learning techniques such as AdaBoost, Logistic Regression, Support Vector Machine, Random Forest, Naïve Bayes, Decision Tree, K-nearest neighbor and Stochastic Gradient Descent, were used to categorize the children in the test dataset as "malnourished" or "nourished". The accuracy, sensitivity, and specificity of these algorithms' prediction abilities were comparedusing performance measures such as accuracy, sensitivity, and specificity. Results show that malnutrition status can be predicted using Random Forest machine learning technique which was about 98% and brings positive impact to the society. The study findings indicated a need for more attention on nutrition to expected mothers and children under five to be well administered with the government and the society at large by putting relevance to the suggestion that cooperation between government organizations, academia, and industry is necessary to provide sufficient infrastructure support for the future society.

Keywords malnutrition; mobile application; machine learning; Tanzania.

Computational Ecology and Software
ISSN 2220-721X
URL: http://www.iaees.org/publications/journals/ces/online-version.asp
RSS: http://www.iaees.org/publications/journals/ces/rss.xml
E-mail: ces@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Malnutrition remains endemic in many poor nations, particularly in Sub-Saharan Africa and parts of Asia. According to UNICEF, WHO and the World Bank, Stunting affected an estimated 22.0% or 149.2 million children under five (5yrs) globally in 2020, over a third resided in Africa (World Health Organization, 2021). In Tanzania, around 3 million children under five years of age are estimated to be stunted due to inadequate nutrition. The World Health Report (2003, 2003) argued that in Mainland, the level of stunting was considered to be very high (≥30%) in 15 regions out of 26 (Tanzania Food and Nutrition Center, 2015). Tanzania National Nutrition Survey (2018, 2019) reported that stunting affect about 10.0 % of children countrywide. Whereas the most affected regions with a prevalence of stunting exceeding 40% were mentioned to be Ruvuma (41.0%),

Iringa (47.1%), Rukwa (47.9%), Kigoma (42.3%), Njombe (53.6%) and Songwe (43.3%). While in Zanzibar, stunting rates were ranging from 20.4% in Stone Town to 23.8% in Unguja North.

Proper diagnosis and treatment of malnutrition could reduce the risk of malnutrition. Various medical devices have been developed to assess children's nutritional status. The primary goal of the diagnostic procedures is to correctly predict the associated disease. Machine learning (ML), a scientific approach that combines artificial intelligence and statistical learning research, is a method for investigating large amounts of data in order to discover previously unknown relationships or patterns (Zhang, 2010; Zou et al., 2018; Alghamdi et al., 2017).

In medical research, various machine learning algorithms have been used to predict a wide range of diseases (Zhao et al., 2017; Ion-Mărgineanu et al., 2017). Algorithms such as random forests, support vector machines, and artificial neural networks have been used to define the status of diseases using common risk variables. Kilicarslan et al. (2021) and Anand et al. (2020) used machine learning algorithms to predict childhood anemia. (Lai et al., 2019) used machine learning techniques to create predictive models for diabetes mellitus. Machine learning approaches are used in Fenta et al. (2021), Kaushik et al. (2021) and Browne et al. (2021) to develop child malnutrition prediction models. Therefore, this study observed the needs to use machine learning techniques for predicting fetal nutritional status on eliminating malnutrition in Tanzania.

2 Materials and Methods

The study setting was divided into four different categories of people living in Tanzania mainland. Data were collected in Dar es salaam, Arusha, Tabora, Njombe, Mufindi, Mafinga, Makambako and Iringa. The selection of the study areas was based on malnutrition existence of over 40% according to Tanzania Nutrition Survey of 2018.

Forms from physical and Reproductive and Child Health (RCH Cards for 0-59months were about 20,896 and about 18,286 mothers' data were collected from clinics. The data set was merged to have a complete data of a child to a mother and form a total of 14662 samples.

2.1 Data cleaning and analysis

The dataset which was combined with children under five and mothers were started to be learned from 18899 samples with 22 columns and reduced to 14662 data entries with 12 columns by imputing the missing data and removing other data which were not relevant in thisresearch such as Iodine intake, Vitamin A, Blood group, names of mothers and children, region, level of education, residence and occupation. The dataset was then remained with the following features Age of a child in month, Height of a child in cm, Weight of a child in kg, Sex of a child, child food intake per day, Food group consumed by a child, Age of mom, BMI of mom, Height of mom in centimeter, Weight of mom in kg, Education level of a mom, No of Children ever born, Mom Food intake per day.

2.2 Data analysis

The study used Python language on Jupiter Anaconda and open-source library. Data were analyzed using pandas and numpy libraries, while the matplotlib library was used in visualization of the dataset. The distribution of curves of BMI calculations, Weight/Height, Height/Age and Weight/Age all followed bell shaped curves. The standard of living, education status, areas of locality, social distribution behavior deviation for the distribution of Height/Age z-score was analyzed. The standard deviations of Weight/Height z-score and Weight/ Age z-score for the selected regions were inside the acceptable range of thequality data.

2.3 Machine learning models

2.3.1 Logistic Regression (LR)

The most common statistical model for classification problems is logistic regression, which uses the maximum likelihood estimation process to estimate the parameter of interest. Let's say there are n features denoted by. $A_{\mathrm{t}} = (A_{1,} \dots \dots \dots \dots, A_{\mathrm{n}})$. Also let $\beta_t = (\beta_1, \dots \dots \dots \dots \beta_n)$ denote the model parameters.

Then there is the logic regression model, which is defined as

$$\log\left(\frac{\alpha}{1-\alpha}\right) = \pi\theta + \beta_i A_{i\,,} \qquad i = 1,2 \dots \dots \dots n. \tag{1}$$

where α represents the chance of occurrence and event, and $(1-\alpha)$ denotes the probability of occurrence and event to the probability of occurrence and event to be shown as for the set.

2.3.2 Support Vector Machine (SVM)

The support vector Machine is a supervised machine learning approach that may be used to solve both classification and regression issues. It's a system that uses a hyperplane to best separate the two classes. It is based on the premise that the assistance will be provided. The importance of vectors alone cannot be overstating, whereas other training samples can be disregarded. Though this type of classifier is viable in high dimensional spaces, the Radio Basis Function (RBF) kernel was also modified during the trials.

2.3.3 Linear Discriminant Analysis (LDA)

Linea Discriminant Analysis is a supervised machine learning technique for extracting the most important features from a dataset. It is used to avoid overfitting the data as well as to keep computing cost to minimum. This is done by projecting a feature space onto a lower-dimensional space with the best class detachability. The axes that are responsible for maximizing the segment among the various classes are given more emphasis in Linear Discriminant Analysis.

2.3.4 K-nearest neighbors (k-NN)

The supervised Machine Learning family of algorithms includes k-nearest neighbors (k-NN), a robust and versatile classifier. Because it makes no explicit assumptions about the distribution of the dataset, k-NN is a non-parametric algorithm. This method saves every single available case and categorizes new cases using a similarity metric. A case is assigned to a class that is generally regular among its k nearest neighbors, as determined by a distance function, based on a majority of votes cast by its neighbors.

2.3.5 Random Forest (RF)

Random Forest is a classification technique, which depends on "growing" a troupe of tree structured classifiers. To classify another individual, features of this individual are utilized for classification utilizing every classification tree in the forest. The grown trees are assembled randomly, and each tree gives a classification (or "voting") for a class label. The decision depends on the majority votes over maximum trees in the forest.

2.4 Implementing the machine learning algorithms

In this study more than five machine learning (ML) algorithms (LR, SVM, LDA, k-NN and RF) were applied by utilizing a sample of 80% of the individuals in each group (training dataset, n=80%) and validated in the staying 20% (test dataset, n=20%). All models were trained based on 10-fold cross validation. 10-fold cross validation were utilized on the training set, and the performance was estimated on the testing set.

2.5 Model evaluation

Several measures of parameters were taken into consideration as an evaluation some of them as shown here: -

2.5.1 Accuracy

The cornerstone for estimating the performance of any prediction model is accuracy. It calculates the proportion of correct predictions to total data points analyzed. The best accuracies acquired by multiple machine learning

algorithms after applying them are shown in this paper. The k-fold approaches, as well as feature selection Accuracy can be calculated mathematically as an illustrated here below:

$$\text{Accuracy} = \frac{True\ Positive + True\ negative}{True\ Positive + False\ negative + False\ positive + True\ Negative} \quad (2)$$

2.5.2 Sensitivity

The fraction of true positive cases that were projected as positive (or true positive) is known as sensitivity. Recall is another name for it. This indicates that there will be a fraction of genuine positive cases that are mistakenly forecasted as negative (the false negative). This can also be expressed as a rate of false negatives. Sensitivity can be calculated mathematically as follows:

$$\text{Sensitivity} = \frac{True\ Positive}{True\ Positive + False\ negative} \quad (3)$$

2.5.3 Specificity

The fraction of real negative cases that were correctly predicted as negative (or true negative) is known as specificity. This indicates that a proportion of true negative cases will be forecasted as positive, which could be referred to as false positives. This can also be expressed as a rate of false positives. Specificity can be calculated mathematically as follows:

$$\text{Specificity} = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad (4)$$

The model was deployed on Mobile application to ease access to the community for the use of it to help health workers and social welfare officers and the families at large. The system was developed using the android studio development environment and firebase technology for the database. Java was the main programming language with CSS for styling and mark-up respectively with the HTML. The machine learning model wasdeployed in a developed system and the development of the system followed an object-oriented approach through the dynamic system development methodology.

3 Results

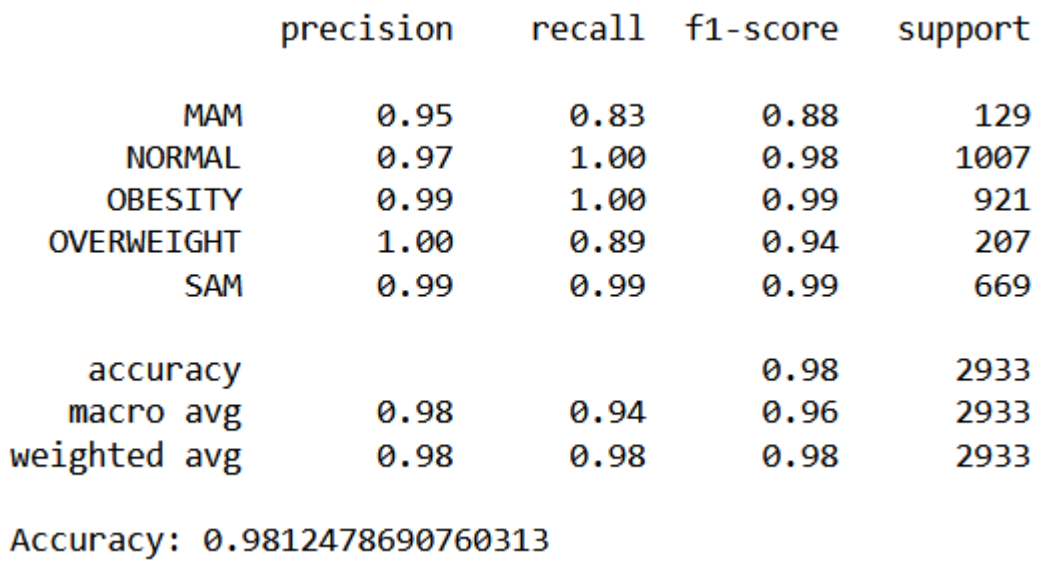
The results showed that children with severe acute malnutrition were frequently observed in Njombe and Iringa regions, whereas obesity was frequently observed in Dar es salaam and Arusha, and moderate malnutrition was mostlyobserved in Tabora region.

The findings for the machine learning algorithms showed that random forest performed well for both the training and test datasets with 98 percent of performance measuresas compared to other algorithms as presented in the Table 1.

Table 1 Performance results of different Models. In each number (NO) of model type (type of model), percentage of accuracy result (Percent of accuracy results) are provided.

No.	Type of Model	Percent of accuracy result
1	Random forest	98
2	Support Vector Machine	88
3	AdaBoost	73
4	Stochastic Gradient Decent	73
5	Naïve Bayes	72
6	Logistic Regression	88

Precision which is positive predictive value showing the fraction of relevant instances among the retrieved instances, while recall which is known as sensitivity showing the fraction of relevant instances that were retrieved. Both precision and recall are captured based on relevance in Fig. 1.



	precision	recall	f1-score	support
MAM	0.95	0.83	0.88	129
NORMAL	0.97	1.00	0.98	1007
OBESITY	0.99	1.00	0.99	921
OVERWEIGHT	1.00	0.89	0.94	207
SAM	0.99	0.99	0.99	669
accuracy			0.98	2933
macro avg	0.98	0.94	0.96	2933
weighted avg	0.98	0.98	0.98	2933

Accuracy: 0.9812478690760313

Fig. 1 Classification report of random forest model. In this model precision, recall, f1-score and support were provided.

This confusion matrix was used to comparing predicted category labels to the true labels. We took in the list of actual labels, the list of predicted labels and an optional argument to specify the order of the labels. We have calculated the confusion matrix and get these results which have been displayed in Fig. 2.

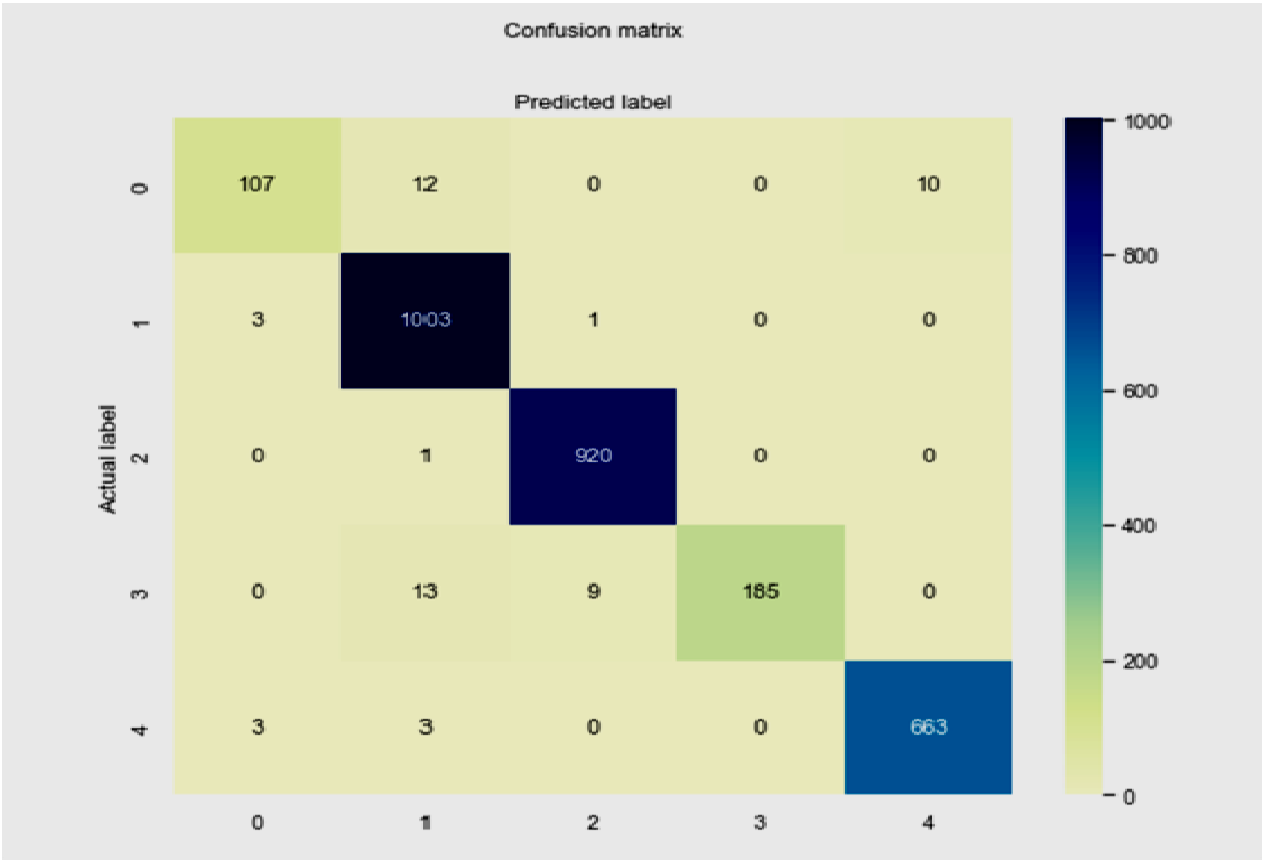


Fig. 2 Confusion Matrix which shows actual label against predicted label.

Nutrition status of several children taken from this dataset showed that there is a big problem to the coming society which shows some prediction of having people with obese and severe acute malnutrition. As it is revealed in thefigure 3, that more than 75% of the collected samples has malnutrition status either of overweight, obesity, severe acute malnutrition and moderate acute malnutrition status as shown in the Fig. 3.

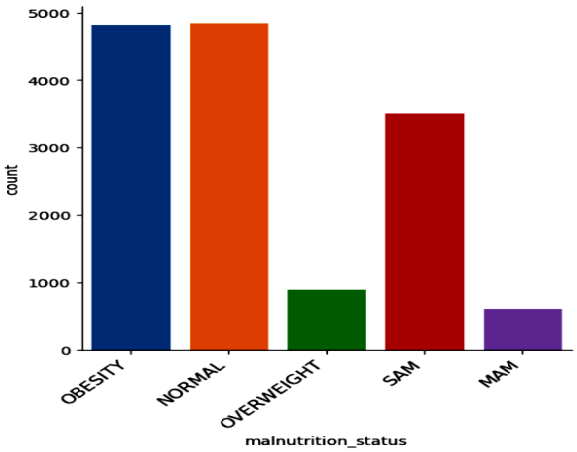


Fig. 3 Plot which shows number of children (count) with nutrition status of children (malnutrition status).

Also, findings of this study showed that nutritional status of mothers is showed at least some good progress of not having big number of mothers who are severe and moderate acute malnutrition (Khan et al., 2019). It predicts a big number of normal nutrition status but it has some prediction of having obesity and overweight mothers. This shows that the government should put some emphasize on balancing these nutrition behavior or policies to the society as presented in the Fig. 4.

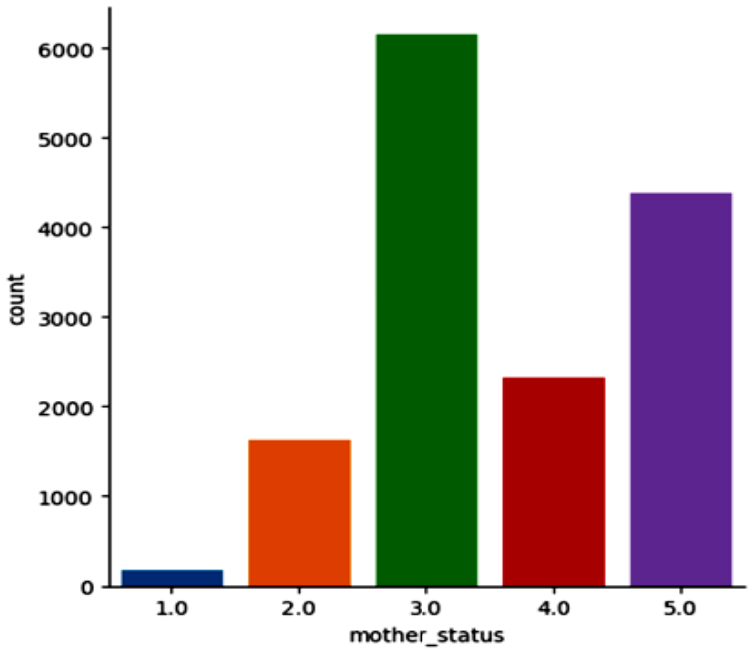


Fig. 4 Plot which shows Mother nutrition status (mother_status) against the number of mothers in the sample (count).

From this study it was observed that, nutrition status of mothers in their number of counts against age shows that, Nutrition status of mothers in their ages of giving birth to our dataset 1.0(SAM) Severe Acute Malnutrition, 2.0(MAM) Moderate Acute Malnutrition, 3.0-Normal, 4.0-Overweight, 5.0-Obesityas shown in Fig. 5. This implies that we have a good number of mothers from the age of 19 to 25 years of age who are having normal nutrition status, then it shows that most of them from age of 26 to 39 are having overweight and obesity mothers.

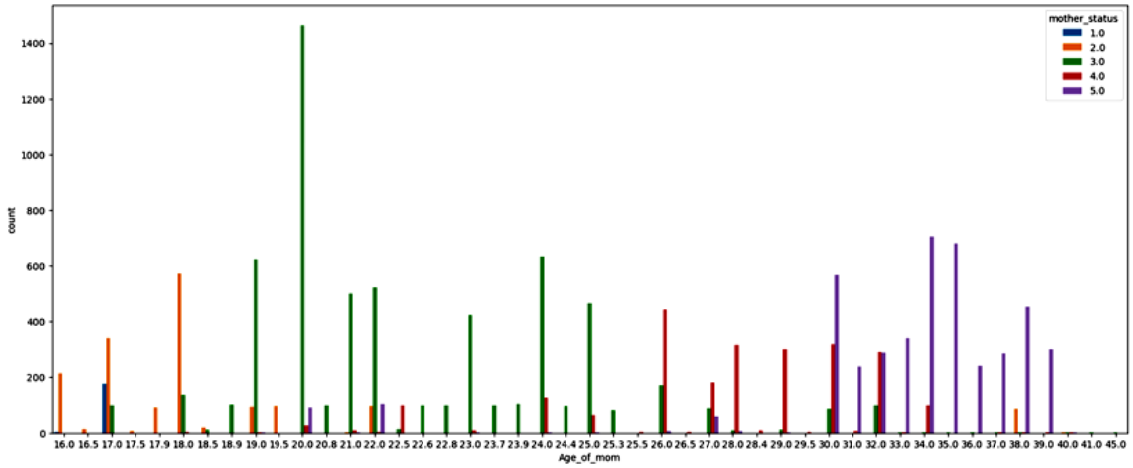


Fig. 5 Plot which shows the age of mothers (Age_of_mom) against their number (count) in the dataset with their nutrition status.

4 Discussion

This study, however, shows that machine learning algorithms and the use of mobile application can be used to predict malnutrition based on common risk factors and life style, which can aid in the creation of human behavior and treatments to avoid malnutrition among Tanzania's children and expected mothers.

On the age of giving birth the result shows that there is a big chance to streamline and make change to the society on nutrition behaviors as to make sure that may be the government should start giving nutritional supplements to ladies of such ages all over the country.

The results on children nutrition status indicated that we are going to have a society with a big problem in nutrition so we need to make sure those who are having severe and moderate malnutrition status are helped to overcome that situation by providing them with needed balanced diet, while those who are obese and overweight were needed to follows the advice from nutritionist.

The government has to give power to social welfare officers to collaborate with nutritionist to overcome this situation and if possible, some bylaws should be put in for the parents who will refuse to follow nutrition instruction to their children.

There is need to change behavior of nutrition situation in our society. Nutritionist and social welfare officers have the mandate to spread and streamline behavioral aspect to overcome some nutrition problems which are caused by some culture and behavior.

The main goal of this study is to predict malnutrition in children under the age of five. Several well-known machine learning (ML) methods were used to achieve our goal: AdaBoost, k-nearest neighbors (k-NN), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF), Decision Tree. These machine learning methods were tested on a sample of 80% of the dataset, (training dataset, n=11,649) and approved on the remaining 20% (test dataset, n=2,913). 10-fold cross validation was used to train all of the models. On the training set, we used 10-fold cross validation, and on the testing set, we estimated performance.

The accuracy, sensitivity, and specificity of these several ML algorithms were compared using three performance parameters, such as accuracy, sensitivity, and specificity.

The best results were achieved by the Random Forest algorithm, which had an accuracy of 97.92 percent, a sensitivity of 94.66 percent, and a specificity of 69.76 percent based on various performance parameters. Furthermore, Random Forest classification revealed a high level of discriminative ability. Along these lines, we can assume that the RF algorithm predicts nutritional status among Tanzania's under-five children moderately better than any of the other Machine Learning algorithms used in this study. However the Random Forest algorithm has the best prediction power when it comes to predicting childhood anemia (Khare et al., 2017). (Khan et al., 2019) identified several key traits. This could be because of the dataset they utilized. Finally, our findings suggest that when malnutrition prediction is a primary concern in Tanzania, random forest classification with random forest feature selection should be used.

5 Conclusions

We compared several machine learning algorithms for identifying whether a child is malnourished based on some risk variables in this study. The Random Forest algorithm outperformed the other algorithms in terms of classification accuracy for predicting malnutrition in Tanzania children. This study looks at the utility of Machine Learning calculations and mobile application as well as the importance of using common socio-demographic and health-related characteristics to predict nutritional status. Furthermore, our findings would be useful for subsequently identifying children at risk of malnutrition, providing policymakers and

medical service providers with a tool to make vital interventions and enhance care practices. As a result, a model based on the key risk factors could help prevent and control child malnutrition.

References

Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. 2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PloS One, 12(7): e0179805

Anand P, Gupta R, Sharma A. 2020. Prediction of Anaemia among children using Machine Learning Algorithms. International Journal of Electronics Engineering, 11(2): 469-480

Browne C, Matteson DS, McBride L, Hu L, Liu Y, Sun Y, Wen J, Barrett CB. 2021. Multivariate random forest prediction of poverty and malnutrition prevalence. PloS One, 16(9): e0255519

Fenta HM, Zewotir T, Muluneh EK. 2021. A machine learning classifier approach for identifying the determinants of under-five child undernutrition in Ethiopian administrative zones. BMC Medical Informatics and Decision Making, 21(1): 1-12

Ion-Mărgineanu A, Kocevar G, Stamile C, Sima DM, Durand-Dubief F, Van Huffel S, Sappey-Marinier D. 2017. Machine learning approach for classifying multiple sclerosis courses by combining clinical data with lesion loads and magnetic resonance metabolic features. Frontiers in Neuroscience, 11: 398

Kaushik H, Singh D, Kaur M, Alshazly H, Zaguia A, Hamam H. 2021. Diabetic retinopathy diagnosis from fundus images using stacked generalization of deep models. IEEE Access, 9: 108276–108292

Khan JR, Chowdhury S, Islam H, Raheem E. 2019. Machine learning algorithms to predict the childhood anemia in Bangladesh. Journal of Data Science, 17(1): 195-218

Khare S, Kavyashree S, Gupta D, Jyotishi A. 2017. Investigation of Nutritional Status of Children based on Machine Learning Techniques using Indian Demographic and Health Survey Data. Procedia Computer Science, 115: 338-349. https://doi.org/10.1016/j.procs.2017.09.087

Kilicarslan S, Celik M, Sahin Ş. 2021. Hybrid models based on genetic algorithm and deep learning algorithms for nutritional Anemia disease classification. Biomedical Signal Processing and Control, 63: 102231

Lai H, Huang H, Keshavjee K, Guergachi A, Gao X. 2019. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders, 19(1): 1-9

Tanzania Food and Nutrition Center, T. 2015. Tanzania National Nutrition Survey 2014. Tanzania

Tanzania National Nutrition Survey. 2018. Final Report. 2019. Tanzania Food and Nutrition Centre, National Bureau of Statistics, Tanzania

The World Health report. 2003. Shaping the Future. 2003. WHO, Switzerland

World Health Organization. 2021. The State of Food Security and Nutrition in the World 2021: Transforming Food Systems For Food Security, Improved Nutrition and Affordable Healthy Diets For All (Vol. 2021). Food & Agriculture Organization, Rome, Italy

Zhang WJ. 2010. Computational Ecology: Artificial Neural Networks and Their Applications. World Scientific, Singapore

Zhao Y, Healy BC, Rotstein D, Guttmann CR, Bakshi R, Weiner HL, Brodley CE, Chitnis T. 2017. Exploration of machine learning techniques in predicting multiple sclerosis disease course. PloS One, 12(4): e0174866

Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. 2018. Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in Genetics, 9: 515. https://doi.org/10.3389/fgene.2018.00515