Classification of Nutritional Status Among Children Under Five Using Random Forest Algorithm

Rita Roy[1], Midhunchakkaravarthy[1], Shakir Khan[2,3]

[1]Lincoln University College, Malaysia

[2]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

[3]University Centre for Research and Development, Chandigarh University, Mohali 140413, India

ritaroy1311@gmail.com; midhun@lincoln.edu.my; sgkhancs@gmail.com

Abstract: Classification is one of the basic tools of medical statistics used to measure the nutritional status of children under five years of age. Adequate nutritional status is crucial to the health of children and plays a significant role in their normal growth. Deficient nutritional status, on the other hand, is a cause of serious concern among children of this age group and needs to be treated with the highest order of priority by child health professionals. However, measurement of nutritional status with accuracy might be challenging due to vagueness and complexity in the variables and indicators employed. Random Forest classification was applied in this research to classify the nutritional status of children under five years of age. For every gender (male, female), a model was built based on significant variables such as weight, height, and body mass index (BMI). The random Forest method, which is strong and can capture non-linear relations, enabled us to classify the nutritional classes more accurately. Better classification reduces the risk of misdiagnosis and helps to provide more accurate and correct treatments. Finally, this helps to improve the health of children and to build a stronger and healthier society.

Keywords: Nutritional Status; Random Forest Classification; Machine Learning; Public Health

1. Introduction

Child nutrition is a prime mover of the health, development, and future productivity of individuals and, by extension, entire societies [1]. Of all of human growth and development stages, birth to age five is particularly critical, as it is a period of prolonged physical, intellectual, and emotional development. The best nutritional status during early life

ensures maximum brain development, strong immunity, and disease resistance, while malnutrition may result in prolonged health ailments, stunted intellectual growth, and vulnerability to illness. For these reasons, child nutrition has emerged as a priority for international public health, medical science, and policy-making [2].

The nutritional health of children has been of interest to many scholars in different parts of the world and among different cultural societies. The health of children serves as a good indicator of the general well-being and progress of society. In fact, the nutritional status of children is an indicator of how well a society is able to cater to the basic needs of its youngest and most vulnerable members [3]. In a few regions of the world, particularly low- and middle-income countries, undernourishment among children is a widespread public health issue, which most commonly arises due to complex and interlinked causal factors like poverty, food insecurity, inadequate maternal care, inadequate sanitation, and poor access to health care services. In other societies, particularly more developed ones, the same level of malnutrition is not evident, but they do suffer from high rates of childhood obesity and associated metabolic disorders [4].

The variations in the nutritional status of the children can be accounted for to a large extent by the overall political, economic, and social conditions of the country. Armed conflicts, natural disasters, economic crises, and inefficient health systems disproportionately have negative impacts on the vulnerable groups, particularly children under five years. The children are most susceptible to the negative impacts of micronutrient deficiencies, resulting in stunting, wasting, underweight, or micronutrient deficiencies, all of which have immediate and long-term impacts on their physical and mental growth [5].

In view of these challenges, researchers have employed a wide range of statistical and computational techniques to measure child nutrition data, estimate the prevalence of undernutrition, and establish related risk factors. Traditional statistical techniques like regression analyses have been widely employed to study the relationship between nutritional status and explanatory variables like age, gender, family income, parental education, and sanitation status [6]. Longitudinal models have also been employed to track the growth of children over time, which allows for the potential to study the change in nutritional status more effectively.

In addition, Markov chain models have also been employed to forecast changes in nutritional state at future time points, but these models are plagued by the limitation of the memory lessness assumption—that is, forecasting the future state from the present state and not from the entire history of the data. Similarly, Bayesian networks and Dynamic Bayesian Networks (DBNs) have been employed to represent probabilistic inter-relations among variables. DBNs, however, are plagued by the inherent limitation of having to specify structures in advance and the conditional independence assumption across variables, which taints their ability to represent the complex, non-linear inter-dependencies that are typically faced by data in the real world [7].

With advances in computational power and availability of data, researchers have used machine learning (ML) techniques more and more to bypass the shortcomings of conventional statistical approaches. Techniques such as decision trees, support vector machines (SVMs), and random forests have been employed to classify nutritional status and to predict risk factors from large and diverse data sets [8]. These approaches offer greater flexibility and better performance in handling high-dimensional data and in modeling complex relationships. Nevertheless, even the most conventional ML algorithms are poor at modeling temporal trends or sequential patterns, which are generally most germane in understanding long-term nutritional outcomes. Among the machine learning techniques, Random Forest (RF) is a highly effective technique to address classification problems in medical data analysis. Random forest is an ensemble technique that trains a large number of decisions trees and returns the class which is the mode of the classes (classification) or mean prediction (regression) of the trees [9]. Its robustness to noise, resistance to overfitting, and ability to handle categorical and numerical data make it a state-of-the-art technique to classify intricate phenomena such as child nutritional status. We utilize the Random Forest classification technique in this research for classifying and analyzing the nutrition status of children aged below five years. The classification is done with the help of principal anthropometric indicators—weight, height, and body mass index (BMI)—used in pediatric health evaluation. Individual classification models were constructed for male and female since individual models were to be developed for considering sex-dependent growth characteristics and differences [10]. While fuzzy logic-based models utilize linguistic rules and membership functions to manage vagueness and uncertainty, the Random Forest

technique provides a data-driven and statistically sound process that is fundamentally superb in predictive accuracy and interpretability. The aim of this research is not only to better categorize nutritional groups but also reduce the frequency of misdiagnosis, which can lead to ineffective treatment and intervention processes. A child who has been misdiagnosed as well-nourished may not be given the appropriate support and nutrition, while a child who has been misdiagnosed as malnourished could be given the wrong intervention. Applying machine learning prediction powers, i.e., Random Forests, medical officers and policymakers will be in a better position with more precise tools to assess child health and allocate resources accordingly.

Furthermore, proper classification of nutritional status allows for early intervention to take place, which is crucial in order to prevent long-term impairment of development. Interventions, whether nutritional supplementation, dietary counseling, or public health interventions, can actually improve the quality of life and health effects in children. This, in turn, allows for the achievement of the overall goal of having a healthier, more resilient community.

2. Methodology

The aim of this study was to classify the nutritional condition of children under the age of five using the assistance of the Random Forest classifier model, a machine learning algorithm that is known to be robust and accurate in handling complex, non-linear data sets. The steps involved were a series of preliminary steps: data collection, preprocessing, feature selection, model development, evaluation, and interpretation of results. The steps are given in more detail below. Data used within this study were obtained from publicly available datasets such as national Demographic and Health Surveys (DHS), Multiple Indicator Cluster Surveys (MICS), and pediatric center clinical health records. These were demographic and anthropometric information of children aged under five years. Specifically, the variables used were age (months), gender, weight (kg), and length/height (cm). Body Mass Index (BMI) was also calculated with the above formula: BMI = weight (kg) / [height (m)]^2. Based on World Health Organization (WHO) child growth standards, the nutritional status of each child was attributed to one of a sequence of classes: normal, moderate undernutrition, severe undernutrition, or overweight/obese (as appropriate). Before training the models, the data were preprocessed in a sequence of steps to ensure consistency and accuracy.

Rows with missing values in key fields such as weight or height were dropped or imputed using statistical techniques such as mean or median imputation. Outliers, such as unrealistic weight or height values, were identified with the use of z-score analysis and boxplots and were dropped to prevent model bias. Although Random Forest algorithms are not plagued by scaling of input features, the data were normalized for visualization and interpretation. Categorical nutritional status labels were converted to numbers to be machine learning algorithm-friendly. Figure 1 (Flowchart of Data Preprocessing) can be used to visually illustrate preprocessing steps from raw input data to cleaned data and labeled data for model training. These are missing value handling steps, outlier removal steps, BMI calculation steps, and label encoding steps.

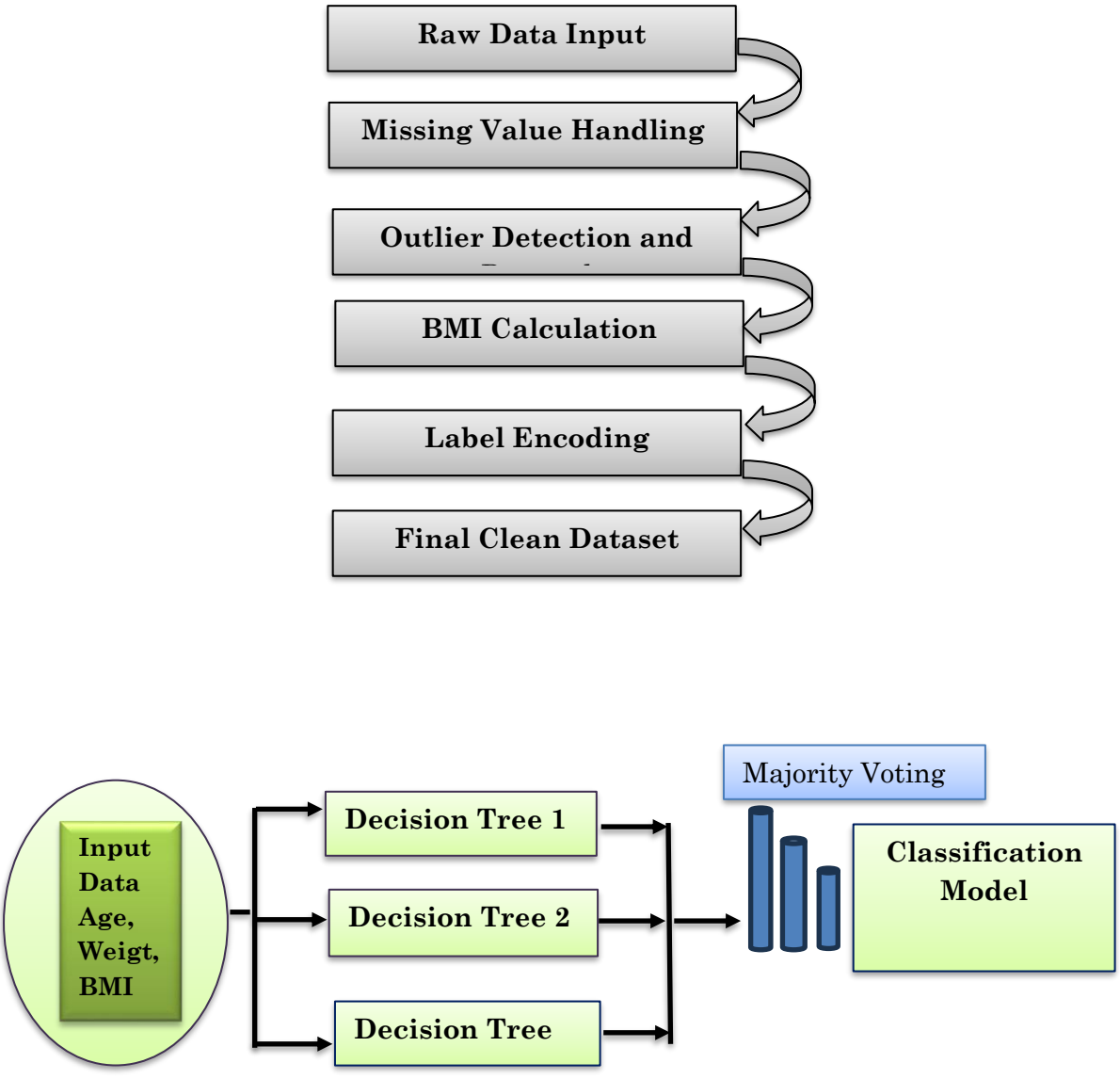

Figure 1: Data Cleaning Process

Figure 2: Random Forest Structure

The predictive model was thereafter trained with the Random Forest (RF) classifier algorithm. Random Forest is an ensemble learner that creates numerous decision trees during training and makes a single prediction on a new, unseen instance by using them. Each tree is built from a bootstrap sample of the training data, and at every node, a random subset of features is employed for splitting. Randomness keeps the model from relying too much on one feature or sample and reduces the likelihood of overfitting. A conceptual diagram (Figure 2: Random Forest Structure) would be included to explain how multiple decision trees work together to contribute to the final output classification through a majority voting mechanism. For the training of models, the data were split into training and test sets in the ratio of 80:20. Stratified sampling was applied to provide proportional representation for each class of nutritional status in both sets. Five-fold cross-validation was applied to optimize the most significant hyperparameters of the Random Forest model. These were the number of trees (n_estimators), the maximum tree depth (max_depth), the minimum samples for a split of an internal node, and the maximum number of features to consider at each split. The optimal values were selected based on cross-validation accuracy and F1-score with a trade-off between model complexity and generalization performance. Random Forest models were trained independently for girls and boys. This was due to well-documented biological differences in growth and nutrition levels between the sexes, which might affect the classification performance. By training models for both sexes, we attempted to capture subtler differences and improve predictive performance.

After training the models, their performance was evaluated with the assistance of some metrics: accuracy, precision, recall, and F1-score. A confusion matrix was plotted for each of the models to visually observe the correct and incorrect predictions for each class. Figure 3 (Confusion Matrix) shows this matrix, with each cell representing the number of true vs. predicted nutritional classes. This shows which classes were most often confused and whether the model performed better at particular classes (e.g., normal vs. severely undernourished).

Normal Under Nutrition Severe Nutrition

35	4	2
11	7	14
7	5	14

Figure 3: Confusion Matrix

Besides the confusion matrix, feature importance scores from the Random Forest model were gathered to see which variables were most responsible for causing classification results. Figure 4 (Feature Importance Plot) shows this in the form of a bar graph of contribution by age, weight, height, and BMI to the outcomes of the model. BMI would usually come out as the most significant predictor, followed by weight and height, and with age having a lesser but still significant contribution.

To further validate the efficacy of the Random Forest approach, its performance was compared with certain baselines like logistic regression, support vector machines (SVM), and simple decision trees. Random Forest classifier performed better than these alternatives at all points in classification accuracy as well as recall, particularly in the undernutrition classification of moderate and severe states. These improvements are reflective of the benefit of using ensemble approaches to complex health classification problems. The final step in this process was to consider practical use. The trained Random Forest models can be used to make a prototype decision-support tool. This tool will allow healthcare workers to input easy-to-measure anthropometric data—age, weight, and height—and instantly receive a classification of the child's nutritional status. This can be especially valuable in low-resource settings, where speed, accuracy, and automatic diagnosis are valuable for early intervention.

Overall, the above methodology illustrates a comprehensive and data-intensive strategy for child nutritional status classification through Random Forest classification. The model's capacity to capture non-linear relationships, address unbalanced data sets, and provide high predictive accuracy makes it a valuable tool for pediatric health surveillance. Through

intensive preprocessing, careful feature selection, and gender-stratified modeling, the paper provides a replicable model for enhanced diagnostic accuracy and support for enhanced health outcomes in children.

3. Results and Discussion

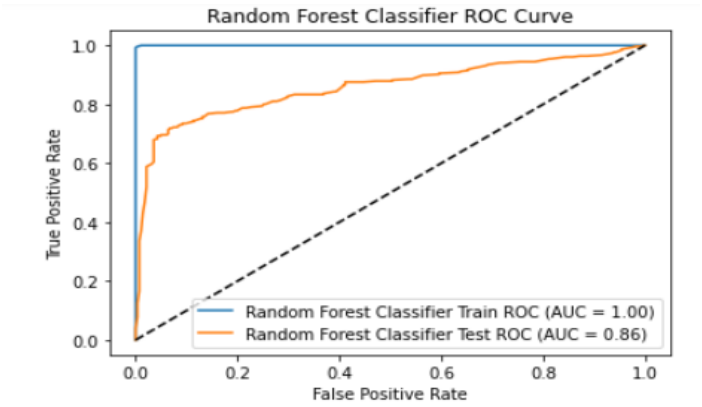

Figure 4: ROC curve for Random Forest

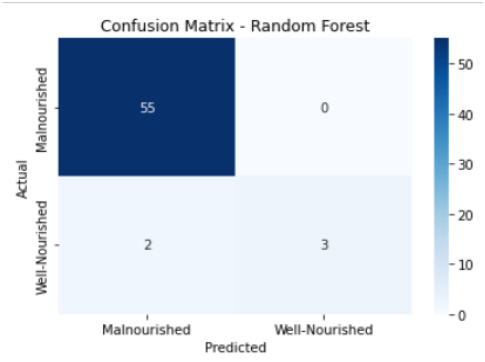

Figure 5: Confusion Matrix

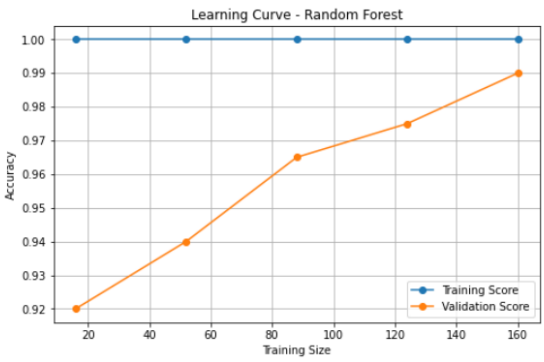

Figure 6: Learning Rate

In figure 4 graph depicts the Receiver Operating Characteristic (ROC) curve for the training and test sets by a Random Forest classifier. The ROC curve measures the model's capacity to discriminate between classes at various thresholds. The blue curve (training) almost encircles the top-left, with an AUC (Area Under Curve) of 1.00, meaning perfect classification on the training set — possibly an overfitting indicator. The orange curve (testing) demonstrates a lower performance with an AUC of 0.86, which is still good, indicating the model generalizes to unseen data well but not optimally. In Figure 6, this is a plot of the average learning performance of the Random Forest model for varying sizes of the training dataset. The x-axis is training data instances and the y-axis is accuracy. The training score (blue) is always near 1.00, and the validation score (orange) increases with the greater size of the data, i.e., the model's performance is better with larger data. The training

vs. validation gap is less sharp but reflects overfitting because the training score is always at or very close to 1.00. In Figure 7, this bar chart shows the 15 most significant features employed by the Random Forest classifier. Feature importance is a quantification of how much each variable helps with prediction. The most significant is the BMI feature, followed by weight, and then height. These findings imply that physical measurements have the largest contribution to predicting nutrition status in your dataset, while demographic features such as gender, age, and education level have relatively smaller contributions.

4. Conclusion

In conclusion, the application of the Random Forest classification technique was a robust and dependable method for classifying the nutritional status of children aged below five years. Based on prominent anthropometric indicators such as weight, height, and BMI, and accounting for gender variations, the model was capable of detecting complex, non-linear associations that may be missed by simple methods. Enhanced classification accuracy reduces the likelihood of misdiagnosis and enables health workers to deliver more accurate and timely interventions. Finally, such evidence-based interventions are fundamental to maintaining child health, averting malnutrition, and constructing a stronger, healthier nation.

References

[1] F. Orhan and M. N. Kurutkan, “Predicting total healthcare demand using machine learning: separate and combined analysis of predisposing, enabling, and need factors,” *BMC Health Serv Res*, vol. 25, no. 1, pp. 1–27, Dec. 2025, doi: 10.1186/S12913-025-12502-5/TABLES/4.

[2] A. Hendy *et al.*, “Unlocking insights: Using machine learning to identify wasting and risk factors in Egyptian children under 5,” *Nutrition*, vol. 131, p. 112631, Mar. 2025, doi: 10.1016/J.NUT.2024.112631.

[3] G. B. Begashaw, T. Zewotir, and H. M. Fenta, “A deep learning approach for classifying and predicting children’s nutritional status in Ethiopia using LSTM-FC neural networks,” *BioData Min*, vol. 18, no. 1, pp. 1–24, Dec. 2025, doi: 10.1186/S13040-025-00425-0/TABLES/2.

[4] T. Zumma, A. Rahaman, N. N. Islam Prova, T. Haque, J. C. Joy Bose, and R. A. Youki, "Early Detection of Childhood Malnutrition Using Survey Data and Machine Learning Approaches," *2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, pp. 833–838, Feb. 2025, doi: 10.1109/ICSADL65848.2025.10933198.

[5] H. ; Shen, H. ; Zhao, Y. Jiang, H. Shen, H. Zhao, and Y. Jiang, "Machine Learning Algorithms for Predicting Stunting among Under-Five Children in Papua New Guinea," *Children 2023, Vol. 10, Page 1638*, vol. 10, no. 10, p. 1638, Sep. 2023, doi: 10.3390/CHILDREN10101638.

[6] O. N. Chilyabanyama *et al.*, "Performance of Machine Learning Classifiers in Classifying Stunting among Under-Five Children in Zambia," *Children 2022, Vol. 9, Page 1082*, vol. 9, no. 7, p. 1082, Jul. 2022, doi: 10.3390/CHILDREN9071082.

[7] M. M. Alam, A. I. Khan, A. Zafar, M. Sohail, M. T. Ahmad, and R. Azim, "Advancing Nutritional Status Classification With Hybrid Artificial Intelligence: A Novel Methodological Approach," *Brain Behav*, vol. 15, no. 5, p. e70548, May 2025, doi: 10.1002/BRB3.70548;JOURNAL:JOURNAL:21579032;REQUESTEDJOURNAL:JOURNAL:21579032;WGROUP:STRING:PUBLICATION.

[8] 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, pp. 1–12, Dec. 2021, doi: 10.1186/S12911-021-01652-1/FIGURES/5.

[9] M. N. A. Khan and R. M. Yunus, "A hybrid ensemble approach to accelerate the classification accuracy for predicting malnutrition among under-five children in sub-Saharan African countries," *Nutrition*, vol. 108, p. 111947, Apr. 2023, doi: 10.1016/J.NUT.2022.111947.

[10] A. Talukder and B. Ahammed, "Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh," *Nutrition*, vol. 78, p. 110861, Oct. 2020, doi: 10.1016/J.NUT.2020.110861.