Ingénierie des Systèmes d'Information
Vol. 30, No. 9, September, 2025, pp. 2487-2498
Journal homepage: http://iieta.org/journals/isi

Machine Learning with Neural Networks and Random Forest to Predict Nutritional Risk in Children Under 5 Years of Age

Inoc Rubio Paucar[1], Cesar Yactayo-Arias[2*], Laberiano Andrade-Arenas[3]

[1] Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima 15046, Peru
[2] Departamento de Estudios Generales, Universidad Continental, Lima 12001, Peru
[3] Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima 15304, Peru

Corresponding Author Email: cyactayo@continental.edu.pe

https://doi.org/10.18280/isi.300922

Received: 29 April 2025
Revised: 14 July 2025
Accepted: 23 August 2025
Available online: 30 September 2025

Keywords:
KDD methodology, machine learning, malnutrition, neural networks, predictive analysis, public health, random forest

ABSTRACT

Child malnutrition has a high prevalence and is associated with multiple risk factors. Although established treatments and methodologies exist, there is still a need for innovative approaches that enable early detection and more effective intervention. This study implemented a predictive model based on machine learning (ML), using algorithms such as neural networks and random forest. The Knowledge Discovery in Databases (KDD) methodology was applied to structure the analysis process and ensure accurate results. A dataset of 500 randomly selected patients was used, including anthropometric, clinical, and socioeconomic variables. The developed model effectively identified the presence of child malnutrition and the key factors associated with the condition. The analysis revealed that 89% of the cases presented some degree of malnutrition, indicating a significant risk pattern in the studied population. The application of predictive models based on advanced ML techniques offers an effective tool for the early identification of child malnutrition, potentially improving prevention and treatment strategies and strengthening the response to this critical public health issue.

1. INTRODUCTION

Child malnutrition represents one of the most persistent and alarming challenges worldwide according to the World Health Organization (WHO) [1]. Millions of children face this condition, not only as a consequence of lack of food, but also as a result of multiple social, economic and cultural factors according to their context. This problem, far from being isolated, reflects deep inequalities and requires urgent comprehensive attention [2].

Childhood is a crucial stage in the development of human beings, and adequate nutrition during these first years of life is fundamental to ensure healthy physical, cognitive and emotional growth. However, in many regions of the world, especially in underdeveloped countries, the lack of access to a balanced diet, basic medical care and decent living conditions has turned childhood malnutrition into a silent but devastating threat [3, 4]. From a medical point of view, this condition can have multiple causes, including recurrent infectious diseases such as diarrhea, respiratory infections or intestinal parasites, which hinder the proper absorption of nutrients. In addition, malnutrition is often associated with low birth weight, inadequate breastfeeding practices and deficiencies of essential micronutrients such as iron, zinc and vitamin A, which are essential for strengthening the immune system and cellular development. The risk factors associated with child undernutrition have motivated the search for specific criteria to identify effective alternatives to mitigate this problem [4, 5].

In this context, there is an urgent need to find reliable solutions that contribute to anticipate and address cases of malnutrition, especially in vulnerable populations. This research is justified precisely in the application of innovative technological tools, such as ML, which offer new possibilities to face this challenge from a preventive approach. The use of algorithms such as artificial neural networks and random forest represents a powerful alternative to traditional methods of analysis, since they allow processing large volumes of data and detecting complex patterns that could go unnoticed with conventional approaches [6, 7]. By applying these techniques to the study of child malnutrition in children aged 0-59 months, we seek not only to identify the determinants of its occurrence, but also to generate inputs to guide the formulation of more effective public policies focused on prevention and timely care.

To address this challenge, and in accordance with the proposed objective, we proposed the development of a predictive model using artificial neural networks and the random forest algorithm, implemented through the Python programming language. This tool allowed a rigorous analysis of the data and the generation of accurate statistical visualizations that facilitated the interpretation and understanding of the results obtained. In order to adequately structure each of the stages of the process, the KDD methodology was used, which provided a systematic approach for the selection, cleaning, transformation, mining and evaluation of the data extracted from the selected database [8, 9].

The methodological approach not only guaranteed the quality of the processed data, but also helped to optimize the performance of the algorithms used, increasing the accuracy in the detection of patterns associated with child malnutrition. The fact of using Python because of its wide ecosystem of libraries such as Scikit-learn, Pandas, NumPy and Matplotlib helped to efficiently implement the models, being possible to reproduce the experiments in a controlled and verifiable way [10]. Overall, this combination of statistical techniques, machine learning and exploratory data analysis laid the foundation for a robust technological solution applicable in real-world contexts.

The study aimed to develop a predictive model using neural network algorithms and random forest to identify risk factors associated with child malnutrition in children aged 0 to 59 months, in order to predict its occurrence in vulnerable populations. The achievement of the proposed objective, had as a contribution to make an early prevention of child malnutrition to take corrective measures for a balanced diet suitable for each child.

2. LITERATURE REVIEW

The present research focuses mainly on the development of algorithms based on neural networks and random forest, using Python tools and specialized libraries for data analysis. This section aims to perform a comprehensive analysis of several scientific studies conducted by experts in this field, highlighting their relevant contributions related to the topic in question.

A state-of-the-art study set out to subject reduced intergenic spacing (RIS) to a dissection study in which advanced machine learning techniques were employed, as stated by its authors [11, 12]. Four modeling strategies: logistic regression, classification and regression tree, monolayer hidden neural network and random forest, were carried out in order to study the relationship between RIL and the presence of multiple concurrent forms of childhood nutritional deficiency (MFCNI) or at least one form of nutritional deficiency (AFCNI) as well as adjusted odds ratio (AOR), Confidence Interval (CI) and Probability of margin of error (P). In addition, modulating variables of this correlation, such as female empowerment, educational level and socioeconomic status, were explored. Of the 4652 cases analyzed, 22.16% corresponded to infants born in short intergeneric intervals. After application of propensity score matching (PPS), it was observed that infants with RIL exhibited an increased risk of NFCM (AOR = 1.25; 95% CI = 1.02-1.56; p = 0.038) and NFCNI (AOR = 1.20; 95% CI = 1.01-1.42; p = 0.045). This correlation was intensified in specific subgroups, such as women with high autonomy and mothers who married before the age of 18 years.

On the other hand, another study proposed the use of machine learning models for the detection and prediction of risk factors linked to nutritional deficiency and infant mortality, using the same methodology in data collection, assessing the nutritional status of infants aged 0-59 months using three key anthropometric indicators: growth deficit (height-for-age), acute nutritional deficiency (weight-for-height) and the combination of both (height-for-age and weight-for-height). Four artificial intelligence-based approaches were applied: conventional machine learning models, tabular deep learning, water models (H_2O) and Automated Machine Learning (AutoML). However, it must be considered that factors such as mortality and morbidity led the researchers [13, 14] to go deeper into the subject and carry out studies such as the implementation of an artificial intelligence model based on machine learning. In this study, an artificial intelligence model based on machine learning was implemented, which took as a reference the application of advanced classification algorithms, including random forest, J48 and Naïve Bayes, which were trained with data collected in hospitals in Afghanistan. The results showed that the random forest model stood out as the most accurate, achieving 97.14% accuracy, while J48 performed acceptably with an accuracy of 94.51%. The concern about this situation led to consider new strategies that are seen as an opportunity to contribute to research related to the subject; an example of this is the research proposed by the author [15] which aimed to employ an approach based on the analysis of socioeconomic data extracted from various reliable data sources. It included ordinary least squares regression algorithms (OLSR), spatial models, machine learning algorithms and artificial intelligence explainable methods (IAEX). First, Recovery Movement Control Order allowed for the assessment of linear correlations between the selected variables, which resulted in findings that the study revealed a marked geospatial concentration and territorial disparities in child stunting [16, 17].

It is essential to emphasize that child nutritional deficiency is an important and very critical problem for society, affecting the development of children and maintaining a cycle of poverty and vulnerability. For this reason, the authors [18, 19] made a scientific contribution with a study, which aims to locate the determinants of anemia in children under five years of age, where the reference population is very vulnerable to anemia. For this, a technique in data mining was proposed, with improved pre-processing, cleaning and data reduction in a data lake architecture. Subsequently, several machine learning algorithms, namely decision tree, logistic regression and k-nearest neighbors, were implemented to determine the most accurate model of anemia [20]. For the selection of the most relevant variables, the analysis of variance (ANOVA) F-test and chi-square filters were used, reducing them to 10 key variables to improve the predictive capacity of the model. After evaluating the performance of the proposed algorithms, it was determined that Naive Bayes is the most efficient model for predicting anemia in infants under 5 years of age, reaching a recall of 74%, a precision of 43% and an accuracy of 70%, consolidating itself as the best alternative for early diagnosis and the implementation of public health intervention strategies [21, 22]. The development of the ensemble learning-based predictive model was performed with randomly selected data from a sample of 574 infants under 5 years of age, with the aim of building a highly accurate and reliable model, as proposed by the author [23]. The data preprocessing process included segmentation of the dataset, allocating 70% for training and the remaining 30% for testing. Nine machine learning algorithms were implemented, including: support vector machines (SVM), decision tree (DT), Naïve Bayes Gaussian (NBG), Naïve Bayes Bernoulli (NBB), Naïve Bayes complementary (NBC), logistic regression (LR), random forests (SF), AdaBoost and Extreme Gradient Boosting (XGBoost).

In many cases, children suffering from nutritional deficiencies or problems suffer serious health complications in their daily lives. Within this scenario and given this, the central objective or aim of this research was to conduct a nutritional study or analysis in children aged 1 to 5 years using the latest

in artificial intelligence techniques, as suggested by the authors [24]. The methodology on which this study was based is based on the selection of reliable data sources and the exhaustive review of scientific literature. For the development of the prediction model, four machine learning algorithms widely used in data classification were tested: random forests (BA), SVM, k-nearest neighbors (KVC) and LR. Once the performance of the prediction models was comparatively analyzed, random sticks and logistic regression were the best, with prediction accuracies of 91.11% for random forests and 89.88% for LR. In the training phase both reached an accuracy value of 1.000 for random forests and 0.847 for LR. Random forests demonstrated the best discriminative ability of all the algorithms presenting predictions more adjusted to the assessment of child nutritional status [25, 26]. In a different context, research suggests that estimation of severe and moderate acute nutritional deficiency in children is critical for rapid responses in emergency situations. In contrast, these two traditional measurement tools, mid-upper arm circumference (CMBS) and weight-for-height (WFT), are impractical in conflict and catastrophic disaster situations, as they require highly trained personnel, bulky equipment, and strict control [27]. Method for extreme and rapid observation of nutritional status (MOERN)accommodates this limitation by means of simple photographs of faces. Facial feature extraction is performed to predict body mass index (BMI) in adults and weight-for-height Z-score (WFTZ) in children under five years of age. MOERN correctly predicted the adult BMI category by 78%, while a variant of this model, trained on a sample of 3167 children in Kenya, correctly classified 60% of the cases [28, 29]. On the other hand, another research proposed by the author [30, 31] focused on the comparative analysis of the VGG16 and SqueezeNet models for the automatic detection of infant nutritional deficiency. For this purpose, a transfer learning-based approach was implemented, with the purpose of taking advantage of the knowledge previously acquired by these models in computer vision tasks. The study focused on the binary classification of infants into two groups: healthy and malnourished. The performance evaluation of both models showed that VGG16 and SqueezeNet have a high predictive potential in the identification of infant nutritional deficiency [32, 33]. The best performance was obtained with the VGG16 model, achieving an accuracy of 93.73%, an F1-score of 0.94, an accuracy of 0.92 and a recall of 0.95. In comparison, SqueezeNet achieved an accuracy of 90%, an F1-score of 0.89, an accuracy of 0.99, and a recall of 0.81. These findings demonstrate the effectiveness of the use of convolutional neural networks (CNNs) in the early detection of nutritional deficiency, which could contribute significantly to nutritional monitoring and intervention strategies.

The review of the scientific articles identified several limitations, such as the lack of a detailed analysis in each of the studies reviewed and the underuse of agile methodologies in their structuring. A proposal for improvement would be the incorporation of advanced predictive models based on machine learning algorithms, combined with other emerging technologies, such as natural language processing (NLP). This could contribute to the development of more accurate and robust models, improving decision making in complex contexts and providing more reliable solutions for risk prediction and intervention customization.

3. METHODOLOGY

3.1 KDD methodology

The KDD methodology is an orderly procedure, which consists of a series of phases that allow identifying and extracting patterns and models from the extensive analysis of data collected in an organizational database [34]. Through this ordered way of working, we can process and transform the information into executable knowledge for strategic decisions as shown in Figure 1. The KDD methodology allows iterations between its phases, i.e., to return to previous stages to debug, optimize and refine the data as the analysis develops. This feedback between the different phases of the KDD methodology ensures the validity and relevance of the results obtained, in the sense that the information obtained from its application ends up being valid and useful for data-driven decision making [35, 36].

3.1.1 Compilation

For this study, a database that collects information on children with possible indicators of child malnutrition was used in the Kaggle platform. The data include demographic, anthropometric, socioeconomic and access to nutritional intervention programs. Data were obtained from public health records, nutritional surveys, and government databases [37, 38]. Records of children aged 0-59 months were selected, ensuring a representative sample from diverse regions and ethnic groups. Table 1 shows how the variables are classified according to the type of statistical data and the description of each one of them. On the other hand, Table 2 shows the statistical description both as numerical and categorical.

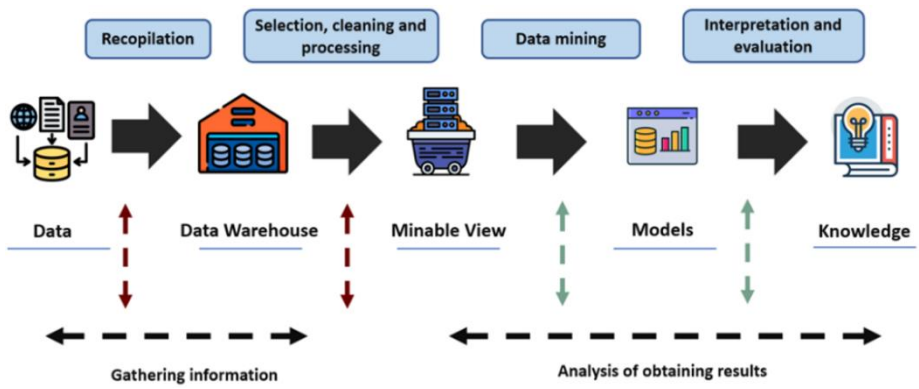

Figure 1. KDD methodology process

In terms of gender distribution, the sample is balanced, with 252 male and 248 female children. Regarding ethnicity, most of the data corresponds to the Xinca population (131), Garifuna (127), Maya (126) and Ladina (116). In addition, it was found that 273 families do not have access to health services, while 227 do. The data show that the average age of the children evaluated is 29.91 months, with an average weight of 11 kg and an average height of 79.51 cm. It has been identified that the monthly family income presents a mean of 272.26 dollars, with a considerable variability reflected in a standard deviation of 130.18 dollars.

The study protocol was reviewed and approved by the Research Ethics Committee of the University of Sciences and Humanities under ACTA CEI No. 018 with approval code 004-2025, complying with national and international regulations for research involving human subjects. Furthermore, the tests performed were based on data extracted from a public database.

3.1.2 Selection, cleaning and processing

It is the value obtained by adding all the data and dividing the result by the total number of observations. It is used to represent a central value in the distribution of the data. As shown in Eq. (1) which is used to calculate the Mean.

$$X = \frac{\sum x_i}{n} \tag{1}$$

On the other hand, the dispersion of the data with respect to the mean indicates that the data are highly dispersed, while a low value indicates that the data are more clustered closer to the mean. In that sense, a high standard deviation, for the sample, indicates that the values are highly dispersed, while a low one indicates that the data are closer to the mean using Eq. (2).

$$\sigma = \sqrt{\frac{\sum(x_i - x)^2}{n-1}} \tag{2}$$

Represents the point at which 25% of the data is less than or equal to this value. It is used to analyze the distribution of the data and detect possible outliers in the lower part of the sample. This allows for the presence of outliers if there is a large difference between Q_1 and the median explained in Eq. (3).

$$Q_1 = x_{\frac{n+1}{4}} \tag{3}$$

The median is the central value of the sample. It divides the data set into two equal halves. If the number of observations is odd, the median is the middle value; if it is even, it is the average of the two central values. It is interpreted as representing data in distributions where the mean is a good indicator of the central value for both even and odd as indicated by Eqs. (4a) and (4b).

$$\text{Si n es impar } Q_2 = x_{\frac{n+1}{2}} \tag{4a}$$

$$\text{Si n es par } Q_2 = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \tag{4b}$$

Represents the value below which 75% of the data lies. It helps to identify the distribution of the upper part of the data and to detect outliers. The Q_3 value indicates that most of the values are high relative to the mean, to calculate these values the Eq. (5) is used.

$$Q_3 = x_{\frac{3(n+1)}{4}} \tag{5}$$

Maximum normalization is a method that adjusts the values of a variable so that they fall within a specific range, usually between 0 and 1. This allows different variables to be compared more equitably by establishing a uniform scale for their values, as indicated in Eq. (6).

$$X_{norm} = \frac{Xi - X_{min}}{X_{max} - X_{min}} \tag{6}$$

- **Data Inspection and Cleaning**

The dataset was reviewed to verify variable types (numerical, categorical, and target) and correct any misclassifications, such as numerical values stored as text.

- **Handling Missing Data**

The percentage of missing values was quantified for each variable. For numerical variables with less than 5% missing data, median imputation was applied; for categorical variables, mode imputation was used. Records with more than 20% missing data across all fields were excluded from the analysis.

- **Outlier Detection and Treatment**

Outliers in continuous variables were identified using either a z-score threshold of ±3 or the Interquartile Range (IQR) method. For anthropometric measures, extreme values were cross-checked against WHO growth standards. Outliers were treated through winsorization, replacing them with the nearest acceptable limit rather than removing them, in order to preserve sample size.

- **Encoding of Categorical Variables**

Non-ordinal categorical variables were transformed using one-hot encoding, while ordinal categorical variables were encoded using label encoding.

- **Feature Scaling**

For the Neural Network model, z-score standardization (mean = 0, standard deviation = 1) was applied to all continuous variables. For the random forest model, the original scales were preserved, as the algorithm is scale-invariant.

- **Class Imbalance Handling**

The class distribution of the target variable (malnourished / not malnourished) was assessed. In the presence of imbalance greater than 60-40%, the Synthetic Minority Oversampling Technique (SMOTE) was applied, or alternatively, class weights were adjusted (class_weight="balanced" in random forest and weighted loss functions in Neural Networks).

- **Data Splitting**

The dataset was partitioned into training (60%), validation (20%), and testing (20%) subsets using stratified sampling to maintain the proportion of classes in each split.

The following Figure 2 shows the variability of weight by gender. If the violins have similar shapes, the weights are similarly distributed in both genders; if one is wider in certain areas, it indicates greater concentration of individuals in that weight range. The inner lines represent quartiles, helping to visualize whether a gender has greater dispersion or outliers.

On the other hand, in Figure 3 the dispersion illustrates the relationship between weight and height, where a positive trend is expected: the greater the height, the greater the weight. If

the points are well aligned, it indicates a strong relationship; if they are scattered, it suggests greater variability. Differentiation by color makes it possible to analyze whether a gender tends to have a higher weight for the same height, and outliers can indicate possible cases of undernutrition or overweight. Differentiation by color according to gender makes it possible to observe specific patterns between males and females. For example, it can show if one gender has a slightly higher weight compared to the other for the same height, which may be influenced by physiological or contextual differences. In addition, the presence of outliers in the plot points that deviate significantly from the general pattern may point to possible cases of undernutrition, overweight or even errors in data collection. Identifying these outliers is key to performing a more detailed analysis, detecting unusual health conditions and making informed decisions on nutritional or medical interventions.

Table 1. Classification of variables

Data	Variable	Data Type	Description
Demographics	Date, Region, Age (months), Gender, Ethnicity	Categorical, Numerical	General information about the child
Anthropometrics	Weight (kg), Size (cm)	Numeric	Key indicators of nutritional status
Socioeconomic	Family Income, Parents' Education Level	Numeric Categorical	Economic conditions and level of education
Health Access	Access to Health Services	Categorical	Whether or not the child has access to
Intervention	Intervention Program	Categorical	Type of program received (Nutrition, Health, Education, None)

Table 2. Descriptive statistics

Statistic	Age (months)	Weight (kg)	Size (cm)	Monthly Family Income ($)
count	500.00	500.00	500.00	500.00
mean	29.91	11.00	79.51	272.26
std	16.98	5.38	23.28	130.18
min	0.00	2.05	40.00	50.11
25% (Q1)	16.00	6.32	59.88	158.39
50% (Q2, median)	30.00	10.39	78.70	274.52
75% (Q3)	45.00	16.25	98.69	386.71
max	59.00	19.99	119.93	499.98

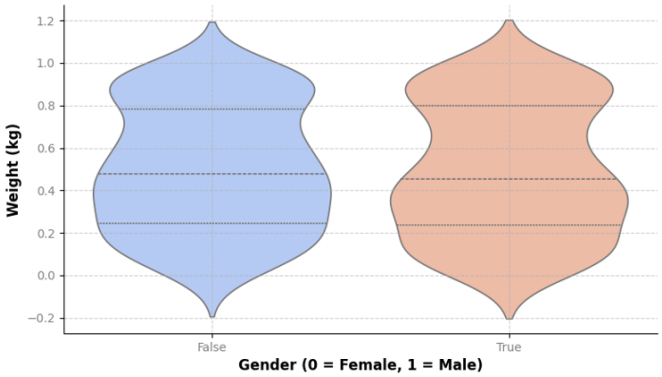

Figure 2. Weight distribution by gender

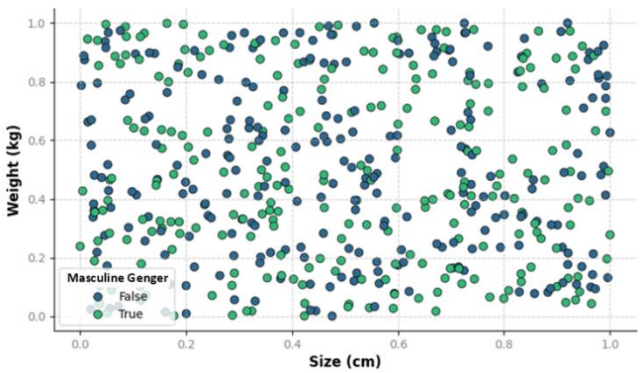

Figure 3. Weight distribution by size

The density curves show the distribution of age and weight in the population analyzed. If the curves are similar in shape and position, both values are homogeneously distributed; if one is wider, it indicates greater variability in that aspect as mentioned in Figure 4. The peaks represent the concentration of children in certain ranges, helping to identify growth patterns and possible nutritional deviations.

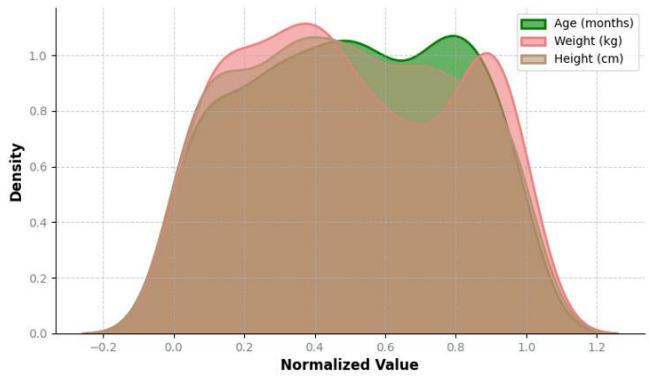

Figure 4. Age and weight density

3.1.3 Data mining

The implemented model corresponds to a Multilayer Perceptron (MLP) neural network with a feedforward architecture, specifically designed to address the binary classification of child malnutrition. This architecture consists of an initial dense input layer with 64 neurons and a ReLU activation function. It contains the previously normalized variables based on the collected data, thus enabling better convergence during training. Next, an intermediate hidden dense layer with 32 neurons and the same ReLU activation function is incorporated, responsible for capturing complex

nonlinear relationships between the child's characteristics and their socioeconomic and health environment (see Table 3). The network concludes with a dense output layer composed of a single neuron with a sigmoid activation function, which returns a continuous probability between 0 and 1, allowing for the determination of whether the child is malnourished or not. Regarding the training parameters, the Adam optimizer was used, known for its efficiency and fast convergence, along with the binary crossentropy loss function, which is ideal for binary classification tasks. Accuracy was adopted as the main evaluation metric to measure the proportion of correct predictions, with a batch size of 8 samples. Additionally, 20% of the dataset was reserved for validation during the training process, allowing for monitoring the model's behavior and preventing overfitting (see Table 4). This configuration demonstrates a robust and well-balanced structure for prediction tasks based on both physical and socioeconomic factors.

Table 3. Multilayer perceptron architecture

Layer	Type	Neurons	Activation Function	Comment
Input	DEMSA	64	ReLU	Receives the normalized variables from the dataset
Hidden Layer 1	DEMSA	32	ReLU	Extract non-linear features
Output Layer	DEMSA	1	Sigmoid	Returns the probability of malnutrition (0 or 1)

Table 4. Training parameters

Parameter	Value
Optimizer	Adam
Loss function	Binary cross-entropy
Metrics	Accuracy
Period	50
Batch size	8
Validation	20% del dataset

The Rectified Linear Unit (ReLU) function returns the input value if it is positive, and zero otherwise. It is used to introduce non-linearity into the model and to help learn complex patterns. Eq. (7) refers to the explanation provided above.

$$ReLU_{(x)} = max(0, x) \tag{7}$$

The sigmoid function transforms any real value into a value between 0 and 1, allowing the output to be interpreted as the probability that a child is malnourished (1) or not (0), as shown in Eq. (8).

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

This function penalizes predictions that deviate from the actual value. If the true value $\hat{y} = 1$, the model must maximize $\hat{y}; si\ y = 0$ must be minimized y. t is ideal for binary classification tasks like in this case, as shown in Eq. (9). It measures the percentage of times the model correctly predicted the outcome, as indicated in Eq. (10), for both positive and negative classes. In this case, the calculation is determined based on the area under the ROC curve (AUC), which assesses the true positive rate (TPR) versus the false positive rate (FPR), as shown in Eq. (11). An AUC of 1 indicates perfect classification, whereas an AUC of 0.5 is equivalent to random guessing.

$$L(y, \hat{y}\{y\}) = -[y \cdot log(\hat{y}) + (1 - y) \cdot log(1 - \hat{y})] \tag{9}$$

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ predictions} \tag{10}$$

$$AUC = \int_{FPR=0}^{FPR=1} TPR(FPR) dFPR \tag{11}$$

- **Training Accuracy vs. Validation Accuracy:** If validation accuracy starts to stabilize while training accuracy continues to improve, it may indicate overfitting. In this case, the curves appear to be aligned, suggesting that the model generalizes well.
- **Training Loss vs. Validation Loss:** Similar to accuracy, the training loss decreases over time, which is a sign of the model's learning progress. The validation loss also follows a downward trend, further supporting the idea that the model is not overfitted and is capable of generalizing to new data, as shown in Figure 5.

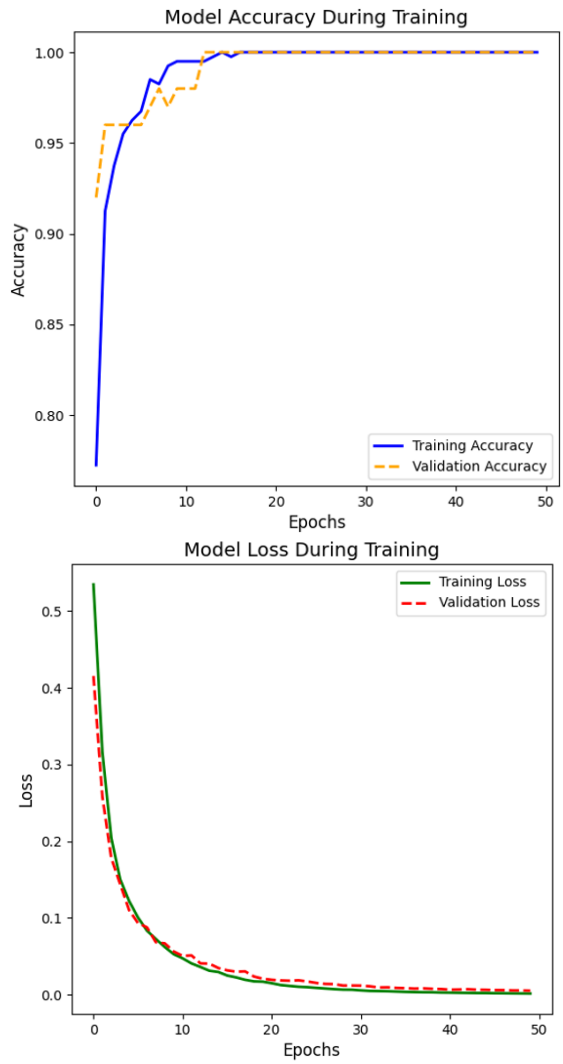

Figure 5. Training accuracy and loss curves

4. RESULT

4.1 Evaluation and interpretation

- **Baseline Model Performance**

The LR baseline achieved an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.914, as shown in Figure 6. This was substantially higher than the no-skill classifier (AUC = 0.500), indicating a strong ability to discriminate between malnourished and non-malnourished children. Despite its robust classification capability, both the random forest and Neural Network models outperformed the baseline in terms of AUC, accuracy, sensitivity, and specificity. These findings confirm that, while LR provides a solid foundation, advanced algorithms can offer additional predictive power for early detection of nutritional risk in children under five years of age.

Age (months): The age distribution shows a concentration of cases among younger children, which is common in studies of child malnutrition, where younger children are more vulnerable.

Weight (kg): The weight distribution reflects variability in the values, with a slight skew toward lower values, which is consistent with malnutrition.

Height (cm): Similar to weight, the height distribution also shows variability, with more cases concentrated at shorter heights, which could indicate a correlation with malnutrition or stunted growth.

Monthly Family Income ($): The distribution of family income shows a concentration in the lower income ranges, which is relevant because lower income levels are correlated with a higher risk of malnutrition.

The predictions for the values "1" (malnutrition) and "0" (no malnutrition) are shown alongside the true values in Figure 7. An exact match between the predictions and the actual values suggests that the model is making accurate predictions, while any discrepancies could indicate that the model is not predicting correctly in some cases. Figure 8 presents the distribution of age, weight, height, and family monthly income for the study population. Age is evenly distributed across the 0–60 month range, while weight and height show natural variability. Family income varies widely, reflecting diverse socioeconomic conditions. These variables were selected due to their strong association with nutritional status in early childhood. Anthropometric measures such as weight and height are core indicators in the WHO Child Growth Standards for detecting acute and chronic malnutrition, while age contextualizes these measures. Socioeconomic status, represented by family income, influences dietary quality and food security, both of which are critical determinants of child nutrition. Their inclusion in the predictive model is therefore supported by both dataset patterns and prior literature, reinforcing their value in assessing nutritional risk in children under five years of age.

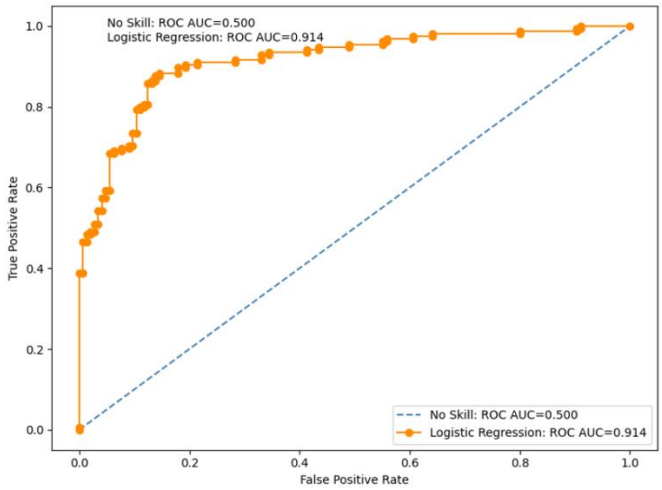

Figure 6. ROC and AUC curve

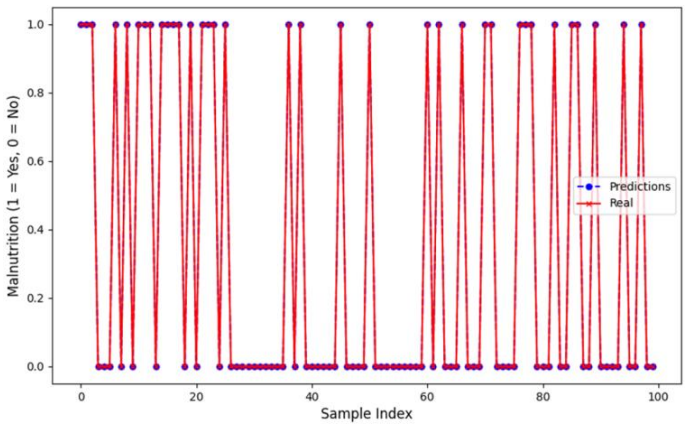

Figure 7. Graph of predictions vs. actual values

The feature importance graph, generated by a random forest model, reveals that weight is the most influential factor in the prediction, with an importance score of 0.23. It is followed in order of importance by height (0.18), age (0.15), and income (0.12). Other features such as maternal education, family size, and birth interval have moderate importance (0.10, 0.08, and 0.07 respectively), while factors like type of residence, vaccination, and access to water show relatively low importance (0.04, 0.02, and 0.01). In summary, basic anthropometric characteristics (weight, height, age) and the socioeconomic factor of income appear to be the most decisive criteria according to this model, as determined in Figure 9.

5. DISCUSSION

Previous researchers have considered the effect of the EIR on the nutrition of children in their early childhood through LR models, decision trees, a single-layer neural network, and decision forests. Although this research is very similar to previous ones in terms of the use of the same techniques, it distinguishes itself by employing APP, which has allowed them to reduce selection biases and obtain more robust causal estimates. Additionally, the mediators of women's empowerment and age at marriage were considered, which were not included in previous studies based on CPT [11, 12]. Previous research has primarily focused on estimating infant mortality from demographic and nutritional data, achieving high accuracy rates through the application of random forest, J48, and Naïve Bayes. In this research, the objective is not so much to achieve high prediction accuracy but to understand the causal mechanism between EIR and nutritional deficiencies, identifying sociocultural factors that exacerbate the relationship between EIR and malnutrition [13, 14]. Regarding the model used in this study, the following explains its choice: a deep neural network with a composite structure was developed, in which an initial layer with 64 neurons using the ReLU activation function was implemented, followed by a middle layer with 32 neurons also utilizing the ReLU activation, and finally, an output layer with a single neuron and a sigmoid activation function, designed to predict the probability of childhood malnutrition. The proposed model is able to adequately capture the non-linear relationships between the variables involved in both the socioeconomic and health domains. In contrast to the model proposed by the author [15], who applied Ordinary Least Squares Regression (OLSR), spatial models, and explainable AI algorithms to highlight geographical differences in growth delays, it was determined that while OLSR was useful for identifying linear correlations, the model in this research demonstrated better predictive performance, particularly when incorporating random forest and neural networks. This highlights the potential of adaptive, non-linear artificial intelligence techniques in complex contexts such as childhood malnutrition [16, 17]. In a second study, the authors [18, 19] examined anemia in children under 5 years old, starting with a deep understanding of data mining from a data lake environment. They optimized the preprocessing, cleaning, and variable reduction stages using statistical methods such as ANOVA F-test or chi-square, selecting 10 relevant variables and improving prediction performance [20]. By applying different machine learning algorithms, the Naive Bayes model was determined to be the most efficient for childhood anemia, achieving record (74%), precision (43%), and accuracy (70%) [21, 22]. In this regard, the random forest model developed in this research reached an accuracy of 89%, significantly surpassing these metrics. Additionally, the neural network developed, based on a three-layer architecture, achieved an effectiveness of 87%, highlighting the potential of deep learning and ensemble techniques in identifying childhood malnutrition. Random forest also identified the most determinant variables in the model (weight, height, age, and income), partially aligning with the variables selected by the authors of previous studies, but with greater explanatory power, positioning this model as a more solid alternative for contributing to early public health interventions. Furthermore, the previous studies also proposed a predictive model based on ensemble techniques, conducting the study on a representative sample of children under five years old, applying a standard 70% training and 30% test split, which was the methodology adopted in the present study. Comparing nine different algorithms (random forest, AdaBoost, and XGBoost) also showed the configurations with the highest predictive capacity [23]. In line with these findings, the random forest model applied in this research achieved an accuracy of 89%, standing out as the most effective compared to other individual models. The multilayer neural network used achieved 87%, demonstrating the potential of deep learning as a valid complement. Both studies emphasize the value of combining multiple artificial intelligence techniques to strengthen early diagnosis of childhood malnutrition and improve decision-making in public health. While both studies applied artificial intelligence algorithms to predict childhood malnutrition, this research incorporates a multilayer neural network in addition to classical models such as random forest, SVM, K-Nearest Neighbors, and Logistic Regression, used by other authors [24]. In terms of accuracy, random forest stood out in both studies, but with better performance in this research (89%). The neural network also showed competitive performance (87%), demonstrating that both ensemble models and deep learning are effective tools for early diagnosis of malnutrition. In this research, random forest showed the best performance with an accuracy of 89%, slightly outperforming other models like Logistic Regression (89.88%). The Neural Network achieved 87%, demonstrating its potential, although it did not surpass random forest. Random forest stood out for its ability to handle complex relationships in the data, while the simpler Logistic Regression was less accurate in complex situations. Overall, random forest proved to be the most robust option for predicting children's nutritional status [25, 26]. In the comparative analysis of predictive models for estimating nutritional deficiency in infants, random forest stood out as the most accurate model, achieving an accuracy of 89%. Its main advantage is its ability to handle complex data without the need for highly trained personnel, making it an ideal option in emergency contexts. Logistic Regression, with an accuracy of 89.88%, is simpler but limited by its inability to handle nonlinear interactions. On the other hand, the Neural Network, with a performance of 87%, has great potential for modeling complex relationships, though it requires more data and is harder to interpret. In summary, random forest proved to be the most suitable model for emergency situations, although each model has its strengths and weaknesses depending on the complexity of the context and available resources [27].

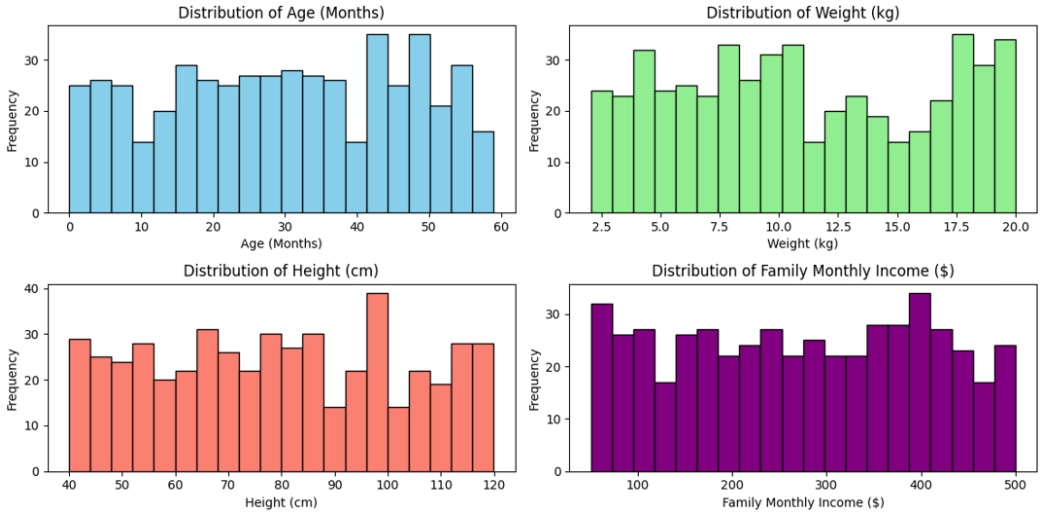

Figure 8. Distribution of numerical variables

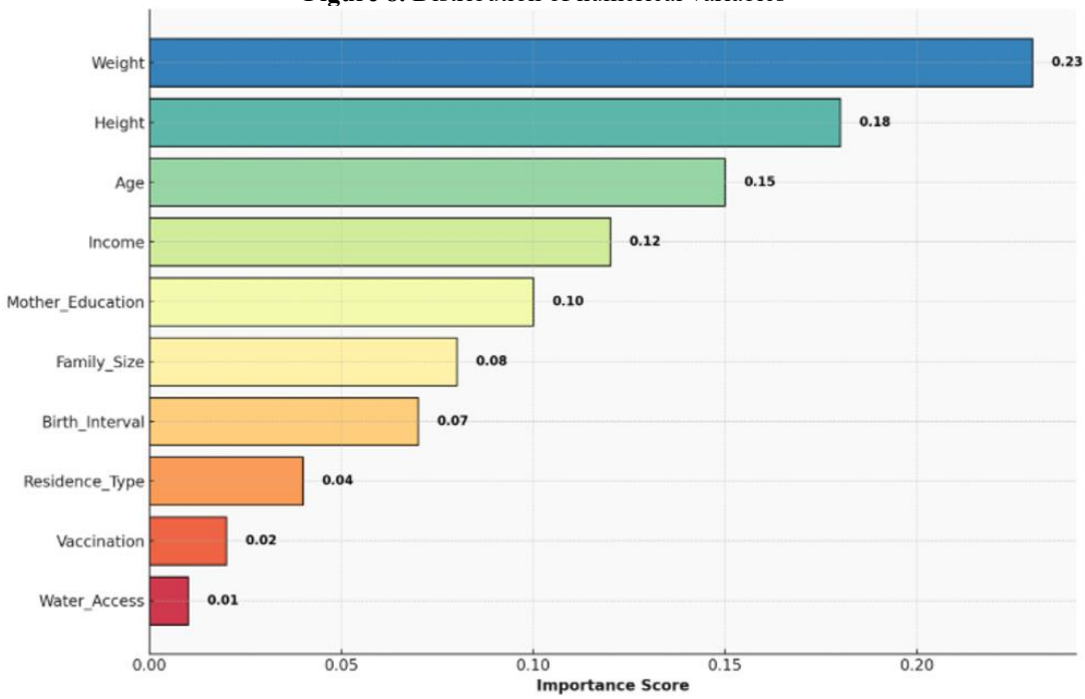

Figure 9. Importance of features using random forest

The method for extreme and rapid observation of nutritional status (MOERN) uses facial photographs to predict body mass index (BMI) in adults and weight-for-height Z-score (WAZ) in children under five years old. This model achieved an accuracy of 78% in classifying BMI in adults and 60% in a sample of 3167 infants in Kenya [28, 29]. Compared to my research, which used models like random forest and neural networks to predict childhood malnutrition, the use of facial images in MOERN offers an innovative and quick alternative, although its accuracy is still lower than traditional predictive models like random forest (89%) or neural networks (87%) applied in my study. On the other hand, the comparative analysis between VGG16 and SqueezeNet for automatic detection of childhood malnutrition in the cited study [30, 31] presents an approach focused on convolutional neural networks, which aligns with the use of neural networks in my research. However, the random forest model in my study achieved better performance in terms of accuracy (89%), suggesting that while neural networks are useful, ensemble models like random forest may offer better results in certain contexts, especially in the classification of childhood malnutrition.

The study on the use of VGG16 and SqueezeNet for detecting childhood malnutrition showed that VGG16 achieved an accuracy of 93.73% and a recall of 0.95, making it highly effective for detecting malnutrition. In comparison, SqueezeNet had an accuracy of 90% and a lower recall (0.81). These results highlight the potential of convolutional neural networks for early detection of malnutrition [32, 33].

This study has certain limitations that should be considered. The dataset comes from a specific geographic and socioeconomic context, which may limit the generalizability of the results to other populations. Some potentially relevant variables, such as dietary diversity or healthcare access, were not available and therefore were not included in the analysis. Additionally, although random forest (89% accuracy) and the multilayer neural network (87% accuracy) showed strong performance using socioeconomic and anthropometric data, other approaches such as computer vision models (e.g., VGG16) can outperform them when photographic data is

available. However, such models are constrained by the need for high-quality images, which are not always feasible in public health contexts. Despite these considerations, the high accuracy achieved suggests that the proposed models retain strong predictive capacity within the studied context.

6. CONCLUSION

The issue of childhood malnutrition is a multifactorial problem deeply rooted in the social fabric, with its primary cause being low levels of nutritional literacy and unfavorable socioeconomic conditions, which determine access to and the quality of food in childhood. In the face of a thorough analysis of this phenomenon, the research conducted aimed to build a predictive model using machine learning techniques to determine whether children aged 0 to 5 years are malnourished or not. To address this issue, multilayer neural network algorithms and random forest were implemented, with the goal of identifying the most relevant characteristics associated with malnutrition in children. The results highlight that the most significant variables in the predictive model indicate that weight is the most important variable (with a weight of 0.23), followed by height (0.18), age (0.15), and family income (0.12). The data classification was structured into two possible classes: 1 for cases with malnutrition, and 0 for those without malnutrition. The trained model achieved an accuracy of 89%, which can be considered a high level of confidence in its predictive power, and it fully applied the KDD methodology for data processing. This led to the development of a procedure within the framework of the methodology, which included four phases: data selection, cleaning and preprocessing, application of data mining techniques, and evaluation and interpretation of results, according to what was deemed appropriate in each phase to ensure that the research added value to the study of the problem. At the same time, this research could also impact future work, as its applicability could be considered a starting point for combining the algorithms used with new technologies that strengthen the process of identifying variables related to malnutrition. Subjects such as the connection between Big Data and Data Warehouse systems are important in this sense, as the use of large volumes of information tends to lead to more precise and satisfactory results, ultimately contributing to better-informed and more effective decision-making in the fight against childhood malnutrition.

REFERENCES

[1] Gavilánez, R.A., Constante, D.T. (2024). Factores socio demográficos y alimenticios para la aparición de desnutrición infantil en sectores rurales. 593 Digital Publisher CEIT, 9(2): 194-204. https://doi.org/10.33386/593DP.2024.2.2312

[2] Calderón-Martínez, M.E., Taboada-Gaytán, O.R. (2024). Los patrones alimentarios y carencias sociales ponen en riesgo de desnutrición a los preescolares de zonas rurales. Agricultura, Sociedad y Desarrollo, 21(2): 222-240. https://doi.org/10.22231/asyd.v21i2.1589

[3] Merchán-Villafuerte, K.M., Sánchez-Pijal, K.D., Toala-Pincay, M.G. (2024). Impacto de la desnutrición en el desarrollo infantil de América Latina: implicaciones para la salud y el desarrollo integral. MQRInvestigar, 8(1): 3566-3586. https://doi.org/10.56048/MQR20225.8.1.2024.3566-3586

[4] Tadesse, G.A., Ferguson, L., Robinson, C., Kuria, S., et al. (2025). Forecasting acute childhood malnutrition in Kenya using machine learning and diverse sets of indicators. PLoS One, 20(5): e0322959. https://doi.org/10.1371/journal.pone.0322959

[5] García, M.C. (2023). El maíz transgénico como solución a la desnutrición infantil del estado de Hidalgo. Huella de la Palabra, 17: 8-19. https://doi.org/10.37646/HUELLA.V17I17.605

[6] Hernández Escalona, M.D.C., Aguilar, A., Kamenetzky, G.V. (2024). Efecto de un programa para mujeres embarazadas sobre las medidas de crecimiento de niños con desnutrición infantil. Revista ConCiencia EPG, 9(1): 90-105. https://doi.org/10.32654/CONCIENCIA.9-1.5

[7] Andrade-Arenas, L., Paucar, I.M.R., Yactayo-Arias, C. (2024). Expert systems and epidemiological surveillance for tuberculosis: Innovative tools for disease prevention and control. International Journal of Engineering Trends and Technology, 72(3): 72-90. https://doi.org/10.14445/22315381/IJETT-V72I3P108

[8] Rivera-Vásquez, J. (2024). Los primeros 1000 días de vida: Caracterización de la desnutrición infantil en Loja. Revista Económica, 12(1): 45-54. https://doi.org/10.54753/RVE.V12I1.1901

[9] Sinchiguano, J.M.S., Nieto, M.I.F. (2024). The impact of child malnutrition on the teaching-learning process of schoolchildren. Salud, Ciencia y TecnologÃa, 4: 4721. https://doi.org/10.56294/saludcyt2024721

[10] Iparraguirre-Villanueva, O., Guevara-Ponce, V., Roque Paredes, O., Sierra-Liñan, F., Zapata-Paulini, J., Cabanillas-Carbonell, M. (2022). Convolutional neural networks with transfer learning for pneumonia detection. International Journal of Advanced Computer Science and Applications (IJACSA), 13(9). https://doi.org/10.14569/IJACSA.2022.0130963

[11] Arya, P.K., Sur, K., Kundu, T., Dhote, S., Singh, S.K. (2025). Unveiling predictive factors for household-level stunting in India: A machine learning approach using NFHS-5 and satellite-driven data. Nutrition, 132: 112674. https://doi.org/10.1016/j.nut.2024.112674

[12] Jain, S., Khanam, T., Abedi, A.J., Khan, A.A. (2022). Efficient machine learning for malnutrition prediction among under-five children in India. In 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, pp. 1-10. https://doi.org/10.1109/DELCON54057.2022.9753080

[13] Momand, Z., Zarinkhail, M.S., Aryan, M.F. (2021). Machine learning based prediction of edematous malnutrition in Afghan children. In International Conference on Emerging Technologies and Intelligent Systems, Springer, Cham, pp. 235-245. https://doi.org/10.1007/978-3-030-85990-9_20

[14] Miranda, E., Aryuni, M., Zakiyyah, A.Y., Kurniawati, Y. E., Sano, A.V.D., Kumbangsila, M. (2024). An early prediction model for toddler nutrition based on machine learning from imbalanced data. Procedia Computer Science, 245: 263-271. https://doi.org/10.1016/J.PROCS.2024.10.251

[15] Zhang, X., Usman, M., Irshad, A.U.R., Rashid, M., Khattak, A. (2024). Investigating spatial effects through machine learning and leveraging explainable AI for child

malnutrition in Pakistan. ISPRS International Journal of Geo-Information, 13(9): 330. https://doi.org/10.3390/ijgi13090330

[16] Paucar, I.R., Yactayo-Arias, C., Andrade-Arenas, L. (2025). Random forest model based on machine learning for early detection of diabetes. International Journal of Advanced Computer Science & Applications, 16(6): 1051-1063. https://doi.org/10.14569/ijacsa.2025.01606103

[17] 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

[18] Francke, P., Acosta, G. (2021). Impacto del programa de alimentación escolar Qali Warma sobre la anemia y la desnutrición crónica infantil. Apuntes, 48(88): 151-190. https://doi.org/10.21678/apuntes.88.1228

[19] Van, V.T.S., Antonio, V.A., Siguin, C.P., Gordoncillo, N.P., Sescon, J.T., Go, C.C., Miro, E.P. (2022). Predicting undernutrition among elementary schoolchildren in the Philippines using machine learning algorithms. Nutrition, 96: 111571. https://doi.org/10.1016/j.nut.2021.111571

[20] Jouzi, Z., Leung, Y.F., Nelson, S. (2024). Characterizing the association between child malnutrition and protected areas in sub-Saharan Africa using unsupervised clustering. Journal of Environmental Studies and Sciences, 14(2): 300-312. https://doi.org/10.1007/s13412-023-00880-3

[21] Usman, M., Kopczewska, K. (2022). Spatial and machine learning approach to model childhood stunting in Pakistan: Role of socio-economic and environmental factors. International Journal of Environmental Research and Public Health, 19(17): 10967. https://doi.org/10.3390/ijerph191710967

[22] Ghosh, A., Freda, P.J., Shahrestani, S., Boyke, A.E., et al. (2025). Preoperative anemia is an unsuspecting driver of machine learning prediction of adverse outcomes after lumbar spinal fusion. The Spine Journal, 25(8): 1596-1607. https://doi.org/10.1016/j.spinee.2025.01.031

[23] Andrade-Arenas, L., Yactayo-Arias, C. (2025). Comparative evaluation of machine learning models for diabetes prediction: A focus on ensemble methods. Ingénierie des Systèmes d'Information, 30(7): 1795-1803. https://doi.org/10.18280/isi.300712

[24] Kebede Kassaw, A., Yimer, A., Abey, W., Molla, T.L., Zemariam, A.B. (2023). The application of machine learning approaches to determine the predictors of anemia among under five children in Ethiopia. Scientific Reports, 13(1): 22919. https://doi.org/10.1038/s41598-023-50128-x

[25] Biradar, V.G., Ankalaki, S., N, K., G, K., Somadas, S. (2024). AI model-based prediction of malnutrition among children using deep learning models with transfer learning. In International Conference on Information and Communication Technology for Intelligent Systems, Singapore: Springer Nature Singapore, pp. 145-154. https://doi.org/10.1007/978-981-97-6681-9_13

[26] Asare, J.W., Brown-Acquaye, W.L., Ujakpa, M.M., Freeman, E., Appiahene, P. (2024). Application of machine learning approach for iron deficiency anaemia detection in children using conjunctiva images. Informatics in Medicine Unlocked, 45: 101451. https://doi.org/10.1016/j.imu.2024.101451

[27] Shi, H., Yang, D., Tang, K., Hu, C., et al. (2022). Explainable machine learning model for predicting the occurrence of postoperative malnutrition in children with congenital heart disease. Clinical Nutrition, 41(1): 202-210. https://doi.org/10.1016/j.clnu.2021.11.006

[28] Garnica, D.C., Cota-Aguiar, R., Vaca-Palomares, I., Fernandez-Ruiz, J., Hevia-Montiel, N., Perez-Gonzalez, J. (2023). Machine learning-based classification of children affected by malnutrition using multimodal SMRI and DTI Brain Images. In 2023 19th International Symposium on Medical Information Processing and Analysis (SIPAIM), Mexico City, Mexico, pp. 1-4. https://doi.org/10.1109/SIPAIM56729.2023.10373493

[29] Asare, J.W., Appiahene, P., Donkoh, E.T. (2023). Detection of anaemia using medical images: A comparative study of machine learning algorithms–A systematic literature review. Informatics in Medicine Unlocked, 40: 101283. https://doi.org/10.1016/j.imu.2023.101283

[30] Zhang, F., Yang, J., Wang, Y., Cai, M., Ouyang, J., Li, J. (2023). TT@ MHA: A machine learning-based webpage tool for discriminating thalassemia trait from microcytic hypochromic anemia patients. Clinica Chimica Acta, 545: 117368. https://doi.org/10.1016/j.cca.2023.117368

[31] Sarsam, S.M., Al-Samarraie, H., Alzahrani, A.I., Shibghatullah, A.S. (2022). A non-invasive machine learning mechanism for early disease recognition on Twitter: The case of anemia. Artificial Intelligence in Medicine, 134: 102428. https://doi.org/10.1016/j.artmed.2022.102428

[32] Memmolo, P., Aprea, G., Bianco, V., Russo, R., et al. (2022). Differential diagnosis of hereditary anemias from a fraction of blood drop by digital holography and hierarchical machine learning. Biosensors and Bioelectronics, 201: 113945. https://doi.org/10.1016/j.bios.2021.113945

[33] Elsabbagh, E., Grossfeld, L., Foote, H. (2025). Machine learning-based pre-transplant risk stratification in severe aplastic anemia. Transplantation and Cellular Therapy, 31(2): S309-S310. https://doi.org/10.1016/j.jtct.2025.01.471

[34] Huerta, C.M., Atahua, A.S., Guerrero, J.V., Andrade-Arenas, L. (2023). Data mining: Application of digital marketing in education. Advances in Mobile Learning Educational Research, 3(1): 621-629. https://doi.org/10.25082/amler.2023.01.011

[35] Al-Alawi, L., Al Shaqsi, J., Tarhini, A., Al-Busaidi, A.S. (2023). Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. Education and Information Technologies, 28(10): 12407-12432. https://doi.org/10.1007/s10639-023-11700-0

[36] Aguagallo, L., Salazar-Fierro, F.A., García-Santillán, J., Posso-Yépez, M., Landeta-López, P.A., García-Santillán, I.D. (2023). Analysis of student performance applying data mining techniques in a virtual learning environment. International Journal of Emerging Technologies in Learning, 18(11): 175-195. https://doi.org/10.3991/ijet.v18i11.37309

[37] Cevallos, J.C., Escobar, M.C., Falcones, J.E., Cevallos, W.J. (2021). Modelado laboral de los egresados de la Facultad de Ciencias Informáticas de la Universidad Técnica de Manabí (Ecuador). Información tecnológica,

32(6): 111-122. https://doi.org/10.4067/S0718-07642021000600111
[38] Dimauro, G., Griseta, M.E., Camporeale, M.G., Clemente, F., Guarini, A., Maglietta, R. (2023). An intelligent non-invasive system for automated diagnosis of anemia exploiting a novel dataset. Artificial Intelligence in Medicine, 136: 102477. https://doi.org/10.1016/j.artmed.2022.102477