MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Salma & Majahar Ali. | Malaysian Journal of Fundamental and Applied Sciences, Vol. 21 (2025) 1529-1541

RESEARCH ARTICLE

Evaluating the Factors and Forecasting Childhood Anemia Through Machine Learning Algorithms

Nahid Salma[a,b] , **Majid Khan Majahar Ali**[a*]

[a]School of Mathematical Sciences, Universiti Sains Malaysia, Penang, Malaysia;
[b]Department of Statistics and Data Science, Jahangirnagar University, Savar, Dhaka, Bangladesh-1342

Abstract Anemia, characterized by insufficient hemoglobin levels, affects a significant portion of the global population, both in developed and developing nations, and is one of the most prevalent health conditions worldwide. With timely diagnosis and proper care, the risk of anemia can be reduced, potentially saving many lives. In this context, machine learning (ML) techniques can serve as valuable tools for disease diagnosis. Therefore, the objective of this study was to determine the most effective machine learning approach while considering the risk factors for childhood anemia in Bangladesh. Secondary data from the 2011 Bangladesh Demographic and Health Survey were analyzed, with both filter (chi-square test) and wrapper (Boruta algorithm) feature selection methods used to identify significant factors. The findings revealed that 52.11% of all Bangladeshi children suffered from varying degrees of anemia (mild: 29.72%**,** moderate: 21.60%**,** and severe: 0.8%**).** Nine key variables—children's fathers' education, child age, breastfeeding status, mother's age, mother's education, toilet type, water source, and the number of children under five years old—were found to be directly linked to anemia. Seven machine learning algorithms (KNN, NB, SVM, RF, Bagging, Gradient Boosting, and XGBoost) were compared based on model evaluation metrics, including accuracy, sensitivity, specificity, precision, Cohen's Kappa, F1-score, and AUC. The results showed that Gradient Boosting outperformed the other algorithms with 87.46% accuracy, 85.31% sensitivity, 96.56% specificity, 95.35% precision, 0.5713 Kappa**,** 0.8990 F1-score, and 0.9099 AUC. Random Forest followed closely with 83.13% accuracy, 87.36% sensitivity, 83.01% specificity, 84.10% precision, 0.3601 Kappa, 0.8555 F1-score, and 0.8531 AUC. Support Vector Machine (SVM) showed 84.46% accuracy, 0.3501 Kappa, 0.8046 F1-score, and 0.8264 AUC, while XGBoost demonstrated 75.99% accuracy, 75.87% sensitivity, 76.23% specificity, 82.76% precision, 0.3319 Kappa**,** 0.7913 F1-score, and 0.7599 AUC. These findings suggest that machine learning techniques—especially Gradient Boosting**—**can be highly effective for predicting anemia in Bangladeshi children, assisting medical professionals in early detection and intervention. The results of this study are expected to guide policymakers and healthcare providers in improving patient care and advancing Bangladesh's progress towards achieving the Sustainable Development Goal (SDG) related to health.

Keywords: Childhood-anemia, Boruta algorithm, machine learning algorithms, Bangladesh.

***For correspondence:**
majidkhanmajaharali@usm.my

Received: 07 April 2024
Accepted: 30 Dec. 2024

Introduction

Anemia, characterized by inadequate blood hemoglobin concentrations, is a global health challenge affecting a substantial portion of the population in both industrialized and developing countries. According to WHO recommendations [11], anemia is classified as a significant health concern when its prevalence in a susceptible population is ≥ 40%. Although anemia can occur across all age groups, it is

more common in young children and pregnant women. The condition profoundly impacts children's cognitive development, educational performance, motor skills, behavior, and immune function [4–7], with significant economic consequences [8]. In women, anemia can impair productivity and increase the risk of pregnancy-related complications, including miscarriage [38]. Alarmingly, anemia is the primary cause of death for 75% of individuals in parts of Asia and Africa [9]. In low- and middle-income countries (LMICs), anemia is a leading cause of disease and mortality among children, with approximately 293.1 million anemic children under five globally—43% of the total under-five population [35]. South Asia, in particular, has reported higher prevalence rates, such as 55.12% in 2016 [36] and 52% in 2019 [32].

While there may be a hereditary component to the onset of childhood anemia, it is not the sole cause [7]. Among many factors, nutritional deficiencies—especially iron deficiency, as well as deficiencies in folate, vitamins A and B12, and copper—are prominent non-genetic causes of childhood anemia [40]. Crucially, a significant amount of research has indicated that iron shortage is the primary cause of anemia in underdeveloped nations, with other causes including hemoglobin disorders, infections by parasites, nutritional deficiencies, and malaria [30-32]. In addition to these biological and environmental factors, socio-economic conditions play a significant role in the prevalence of childhood anemia. Children from lower socio-economic backgrounds are more likely to experience poor nutrition, limited access to healthcare, and higher exposure to environmental risks such as infections and malnutrition, all of which increase the likelihood of developing anemia. The prevalence of anemia varies according to age, sex, maternal age, household wealth status, occupation, and community wealth index [6, 31], alongside factors such as malnutrition (particularly stunting) [22, 32, 20], insufficient daily meals, recent fever, diarrhea, and worm infestations [2, 4, 5, 10, 23, 24, 30]. Incorporating socio-economic factors into predictive models can significantly enhance accuracy by capturing the indirect effects of these factors on health outcomes. Previous research has mostly focused on identifying the risk factors behind the rising frequency of pediatric anemia, but socio-economic factors, which are often intertwined with nutrition and healthcare access, must be carefully considered to create more precise models for anemia prediction. This is where machine learning (ML) prediction models prove invaluable. By leveraging their ability to analyze and integrate clinical, nutritional, and socio-economic data, ML models can uncover hidden patterns and relationships that traditional methods may overlook [44, 50].

In healthcare, ML has already demonstrated its potential by outperforming traditional models in diagnosing and predicting conditions like diabetes [39], severe appendicitis [16], and multiple sclerosis [41]. For anemia, ML models could significantly improve early detection, resource allocation, and treatment prioritization, ultimately reducing the disease burden on vulnerable populations [45]. Traditional diagnostic methods, although useful, are often constrained by factors such as cost, accessibility, and the need for laboratory testing, which can limit their widespread application, particularly in rural and underserved areas. Machine learning offers a unique opportunity to address these limitations by leveraging large and diverse datasets that incorporate clinical, socio-economic, and nutritional factors [47]. By doing so, ML models can provide more accurate, timely predictions of anemia risk, enabling early identification of high-risk children. These predictive models not only offer the potential for better-targeted interventions but can also help healthcare professionals make data-driven decisions that improve child health outcomes [46-48]. Moreover, by incorporating non-invasive data, such as dietary habits, household socio-economic conditions, and medical histories, ML can offer a more accessible and cost-effective solution, particularly in areas where traditional testing is scarce [49]. This innovative approach could lead to earlier detection, more personalized treatment, and ultimately, a significant reduction in anemia-related health issues in Bangladesh, supporting both healthcare providers and policymakers in implementing more effective public health strategies tailored to the country’s unique needs [50].

Despite the promising applications of machine learning (ML), its use in predicting childhood anemia in Bangladesh remains limited. While the country has made significant strides in achieving many health-related Millennium Development Goals (MDGs), anemia continues to be a major and persistent challenge, especially among young children [13, 15, 16, 19]. Bangladesh is one of the South Asian nations most at risk for anemia, with the prevalence of anemia among children under five rising alarmingly from 47% in 2004 to 68% in 2013, according to the National Nutrition Project (NNP) [5, 29]. These concerning statistics highlight the urgent need for innovative approaches to better understand and predict anemia risk among Bangladeshi children. To date, only a few studies have explored machine learning (ML) approaches for predicting childhood anemia in Bangladesh [19, 43]. According to the authors [19], the study considered a comprehensive set of machine learning models, including linear discriminant analysis (LDA), classification and regression trees (CART), k-Nearest Neighbors (k-NN), support vector machines (SVM), and random forest (RF), alongside the traditional logistic regression approach, allowing for a well-rounded comparison of predictive performance. As noted by the author [51], Logistic Regression (LR) and Random Forest (RF) have been employed as feature selection

techniques to determine the key risk factors associated with anemia. However, no research has comprehensively compared multiple ML techniques to identify the most effective model for forecasting anemia risk in this context. This gap underscores the need for further investigation into ML applications that account for Bangladesh's unique socio-economic, nutritional, and health-related factors. Given the complexity of anemia's causes—such as poverty, malnutrition, and limited healthcare access—a more nuanced approach using ML could offer more accurate and efficient predictions. This would complement traditional clinical methods, enabling earlier identification of at-risk children and allowing for targeted interventions. Moreover, ML models could help policymakers design more effective, evidence-based public health strategies [50]. The goal of this research is to identify the most appropriate ML approach for predicting childhood anemia in Bangladesh, addressing the country's specific challenges and providing a foundation for scalable solutions.

Materials and Methods

Data Description

This study made use of secondary data from the Bangladesh Demographic and Health Survey (BDHS-2011), which included up-to-date information on anemia. In this cross-sectional study, samples from the whole Bangladeshi people were chosen at random. As part of the demographic and health survey (DHS) program, the National Institute of Population Research and Training (NIPORT), ICF International (USA), and Mitra and Associates started partnering in tandem in 1993 to launch a nationally representative poll on a quarterly basis. For gathering samples for the investigation, a two-stage cluster sampling approach was adopted. 600 clusters were picked in the first session, which was with 207 clusters situated around urban areas and 393 clusters in the countryside. Employing a conventional systematic selection technique, 30 households (HHs) received consideration from each cluster in the subsequent stage. Through direct interviews, survey participants' demographic, socioeconomic, health, and nutritional data were gathered [19].

Anthropometric measures and blood tests were performed on participants (women and children) by a competent practitioner [18]. The HemoCue rapid diagnostic technique was utilized to determine the blood hemoglobin levels of under 5 years children from each third home included in the BDHS sample. 2278 youngsters were finally taken into consideration by the survey after missing values were corrected (of 6-59 months of age). The complete report of BDHS-2011 contains further information regarding the survey's methodology, data gathering, and indicators [23].

Outcome Variables

The outcome variable in the present study was the existence of anemia. In accordance with the WHO's recommendations, children aged 6-59 months who have Hb levels of less than 11.0 grammes per deciliter (g/dl) are classified as "anemic," while individuals without such levels are regarded as "non-anemic" [18]. Based on the severity of anemia, the anemic children were again divided into three subcategories: mild anemia (Hb level 10.0 to 10.9 g/dl), moderate anemia (Hb level 7.0 to 9.9 g/dl) and severe anemia (Hb level < 7.0 g/dl) [23].

Explanatory Variables

This study considered the factors of anemia found from previous literature [1, 18, 21] also added some more possible factors as explanatory variables. Table 1 lists the levelled factors that are taken into consideration.

Table 1. Summary of sleeted explanatory variables

Variables Name	Description	Categorization
mothers age	Mother's age in 5-year groups	(i) 6-24 (ii) 24-36 (iii) 36-59
mothers_edu	Mother's educational status	(i) No education i(i) Primary (iii) Secondary (iv) Higher
fathers_edu	Husband/partner's education level	(i) No education (ii) Primary (iii) Secondary (iv) Higher
work_mother	Respondent currently working	(i) No (ii) Yes
child_age	Child's age in months	None
breastfeeding	Currently breastfeeding	(i) Yes (ii). No
sex_child	Sex of child	(i)Male (ii)Female
toilet_typ	Type of toilet facility	Flush to piped sewer system Flush to septic tank Flush to pit latrine Flush to somewhere else Flush, don't know where Ventilated Improved Pit latrine Pit latrine with slab Pit latrine without slab/open pit No facility/bush/field Hanging toilet/latrine Not a dejure resident
water_source	Source of drinking water	(i) Home (ii) Hospitals (iii) Others
res_type	Type of place of residence	(i) Urban (ii) Rural
division	Division	(i). Dhaka (ii). Barisal (iii). Chittagong (iv). Rangpur (v). Rajshahi (vi). Khulna (vii). Sylhet
child_num	Number of living children	None
birth_size	Size of child at birth	Very Large, Larger than average, Average, Smaller than Average, Very Small, Do not Know
vitamin_A	Vitamin A in last 6 months	Yes, No, Do not Know
iron_pills	Taking iron pills, sprinkles or syrup	Yes, No, Do not Know
drug_intes	Drugs for intestinal parasites in last 6 months	Yes, No, Do not Know
mem_house	Number of household members	None
child5_num	Number of children 5 and under in household (de jure)	None

Data Pre-processing

The dataset was splitted into two parts, 70% was training set and rest 30% was considered to validate all models.

Statistical Analysis

Descriptive statistics were performed to summarize the basic demographic characteristics of the respondents. A Chi-square test was used to assess the association between childhood anemia and socio-demographic and anthropometric characteristics. Both filter (Chi-square test) and wrapper (Boruta algorithm) feature selection methods [3, 34, 8] were employed to extract the most important risk factors/features from the dataset. Seven popular machine learning models were specifically considered: Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), Bagging, Gradient Boosting, and XGBoost [33]. These models represent a combination of ensemble-based algorithms (RF, Bagging, Gradient Boosting, and XGBoost) and non-ensemble-based algorithms (SVM, NB, and KNN). The prediction performance of the models was evaluated using several metrics, including Cohen's Kappa, accuracy, sensitivity, specificity, precision, F1 score, and AUC. To improve the reliability and generalizability of the models, 10-fold cross-validation was applied, ensuring that the evaluation was robust and reduced the risk of overfitting. This approach allowed for a more

comprehensive assessment of each model's ability to predict anemia status in the dataset. For statistical analysis, the programming language R (Version 4.1.1) and IBM SPSS for Windows (Version 25.0) were used.

Results and Discussion

Descriptive and Bivariate analysis

Of the 2278 samples considered, 1187 (52.11%) children were found to be anemic, with 18 (0.8%) classified as severely anemic, 492 (21.60%) as moderately anemic, and 677 (29.72%) as mildly anemic. Additionally, 1091 (47.89%) children were not anemic (Figure 1, Table 2). This result aligns with previously published estimates of the country's anemia prevalence, which also reported a prevalence rate of 52% [1, 18, 19].

Table 2. Distribution of anemia level patterns

Anemia Level	Frequency
Severe	18
Moderate	492
Mild	677
Non-anemic	1091
Total	2278

Table 3 displays various exploratory factors' frequency distributions along with their connections with various anemia levels. Mothers between the ages of 20 and 25 account for the largest percentage of severely anemic children (33%) among all mothers. Additionally, 38% and 35% of women in these age groups have children who are moderately or mildly anemic. The-chi square score shows a strong association between the age of the mother and anemia. The proportion of anemic kids born to mothers in the 26–29 age spectrum is second highest at 22% (severe), 23% (moderate), and 27% (mild), respectively. A Study, conducted by [19] also found association between mom's age and their children's' anemia status. The degree of anemia in children is strongly correlated with the mother's educational attainment (p value<0.05). Mothers with higher education levels tend to have lower rates of anemic children (0 %), 5.1% (moderate), and 14% (mild). Conversely, the proportion of anemic children in all three categories: severe (33.3%), moderate (39.6%), and mild (43%), respectively is surpassing among mother with a secondary education. Remarkably, compared to mothers with education, women without any formal education have a significantly lower percentage of anemic children (71.9%). According to this study, children's anemia was also significantly influenced by the partner's education (p-value <0.001). Higher educated partners are more inclined to carry fewer anemic kids overall, only 30.9%. Yet, across the three categories, partners with primary education have the greatest proportion of children who are anemic. Though the working status of children's mother did not identify as significant factor from chi-square test but unexpectedly, a higher percentage of anemic children belong to the non-working mother group than to the working mother group. For example, the findings indicate that over 90% of children whose mothers do not work have anemia across all groups. The significant predictors extracted by this study was in line with the conclusions of other investigations [2, 4-7, 11, 13, 18-20].

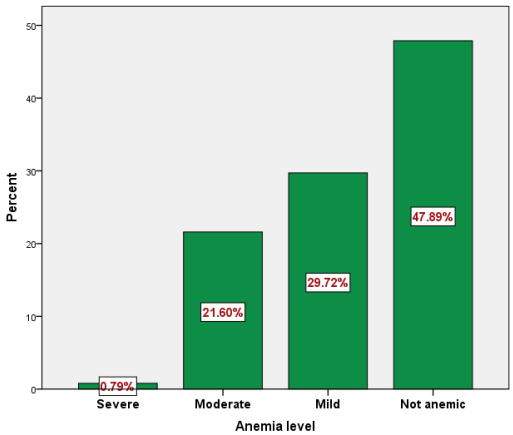

Figure 1. Distribution of anemia level

Anemia stages was observed to be linked with children's age in months (p<0.001). With a proportion of 30.7%, kids within the ages of 13 and 24 months were potential for being moderately anemic. Even though the results do not indicate a relationship between the kind of responder and their anemia status, rural inhabitants had a higher likelihood of being anemic, with 72%, 72%, and 73% of them classified as severe, moderate, and mild anemic citizens. Additionally, a child's breastfeeding status and anemia risk are related (p<0.000). According to the findings, over 70% of children who were not breastfed get varying degrees of anemia. Notably, the anemia of male children (severe: 56%, moderate: 55% and mild: 50.5%) is greater than that of female children (severe: 44%, moderate: 45% and mild: 39.5%). Anemia status was also shown to be substantially correlated with the types of toilets used by respondents and the medications given by the child for parasite infections over the past 6 months (p<0.001). About 62.4% moderately and 61% severely anemic child had not taken any drug for intestinal parasites during the previous six months. Additionally, it was discovered that the use of iron supplements was strongly linked to anemia. Remarkably, children who did not take iron supplements had anemia in 100% of cases, 96.9% in moderate cases, and 97.5% in mild cases. The anemia status was not shown to be substantially correlated with the division of respondents, the number of household members, the number of children still alive at birth, or the amount of vitamin A taken within the previous six months. The study's findings were in line with those of other studies [2, 29-31, 36, 38].

Table 3. The frequency distribution and connection among anemia level and other explanatory factors

Variables	Anemia Level			
	Severe n (%)	Moderate n (%)	Mild n (%)	p-value
Mothers age at birth in years				
15-19	4 (22)	86 (17.5)	94 (14)	
20-25	6(33)	185 (38)	235 (35)	0.004
26-29	4 (22)	113 (23)	180 (27)	
…	..	..	…	
45-49	0 (0)	1 (0.002)	2 (0.003)	
Mother's education				
No Education	6 (33.3)	99 (20.1)	125 (18.5)	
Primary	6 (33.3)	173 (35.2)	225 (33.2)	<0.001
Secondary	6 (33.3)	195 (39.6)	291 (43)	
Higher	0 (0)	25 (5.1)	96 (14.2)	
Partner's education				
No Education	6 (33.3)	128 (26)	201 (29.7)	
Primary	6 (33.3)	182 (37)	201 (29.7)	0.05*
Secondary	4 (22.2)	141 (29)	195 (28.8)	
Higher	2 (11.1)	41 (8)	80 (11.8)	
Mothers working status				
No	16 (89)	442 (90)	606 (90)	0.971
Yes	2 (11)	50 (10)	71 (10)	
Childs age in month				
06-12	6 (33.3)	131 (26.6)	93 (13.7)	
13-24	5 (27.8)	151 (30.7)	165 (24.4)	0.001**
25-36	4 (22.2)	67 (9.9)	146 (21.6)	
37-48	2 (11.1)	75 (11.1)	140 (20.7)	
49-59	1 (5.6)	68 (13.8)	133 (19.7)	
Residence type				
Urban	5 (28)	140 (28)	186 (27)	0.113
Rural	13 (72)	352 (72)	491 (73)	
Breastfeeding status				
Yes	5 (28)	110 (22.4)	208 (30.7)	0.001**
No	13 (72)	382 (77.6)	469 (69.3)	
Sex of Child				
Male	10 (56)	269 (55)	342 (50.5)	0.411
Female	8 (44)	223 (45)	335 (49.5)	
Types of toilet facility				
Piped into dwelling	2 (11)	13 (2.7)	26 (3.8)	
Tube well or borehole	11 (61)	395 (80)	544 (80.4)	0.001**
…	…	…	…	
Not a dejure resident	4 (22)	40 (8.1)	53 (7.8)	

Variables	Anemia Level			
	Severe n (%)	Moderate n (%)	Mild n (%)	p-value
Sources of drinking water				
Piped into dwelling	2 (11.1)	13 (2.64)	26 (3.84)	0.001**
Tube well or borehole	11 (61.1)	395 (80.3)	544 (80.4)	
…	…	…	…	
other	0	1 (0.2)	0	
Division				
Dhaka	3 (16.7)	71 (14.4)	102 (15.1)	
Barisal	3 (16.7)	64 (13)	104 (15.4)	
Chittagong	3 (16.7)	101 (20.5)	124 (18.3)	
Rangpur	0	75 (15.2)	100 (14.8)	0.101
Rajshahi	1 (5.5)	53 (10.8)	80 (11.8)	
Khulna	2 (11.1)	51 (10.4)	85 (12.6)	
Sylhet	6 (33.3)	77 (15.7)	101 (14.9)	
Number of living children				
1-2	13 (72.2)	303 (61.6)	429 (63.4)	0.336
…				
Size of child at birth				
Very Large	0	5 (1)	10 (1.5)	
Larger than average	4 (22.2)	63 (12.8)	93 (13.7)	
Average	12 (66.67)	334 (67.9)	460 (67.95)	0.909
…	…	…	…	
Do not know	0	0	1 (0.2)	
Taking Iron pills				
Yes	0	15 (3.1)	17 (2.5)	0.831
No	18 (100)	477 (96.9)	660 (97.5)	
Vitamin A in last 6 months				
Yes	11 (61.1)	273 (55.5)	403 (59.5)	0.260
No	7 (38.9)	216 (43.9)	268 (39.6)	
Do not Know	0	3 (0.6)	6 (0.9)	
Number of household members				
2-4				
…	6 (33.3)	153 (31.1)	188 (27.8)	0.037
Number of children under 5				
0	3 (16.7)	20 (4.1)	32	
1	8 (44.4)	274 (55.7)	409	
2	6 (33.3)	157 (31.9)	189	
3	1 (5.6)	23 (4.7)	30	0.001**
4	0	11 (2.2)	7	
5	0	5 (1)	7	
8	0	2 (0.4)	3	
Drugs for intestinal parasites in last 6 months				
No		307 (62.4)	345 (51)	0.001**
Yes	11 (61)	184 (37.4)	332 (49)	
Don't know	7 (39)	1 (0.2)	0 (0)	
	0 (0)			

Here, * and ** indicates the significance at 5% and 1% level of significance respectively

Feature Selection/Risk Factors Extraction

Children's fathers' education, child age, breastfeeding status, mother's age, mother's education, toilet type, water source, and number of children under 5 years are significantly associated with the likelihood of being anemic which was extracted by Boruta algorithm, as shown in Figure 2. The effectiveness of ML algorithms was assessed using these nine characteristics. The same variables were also found

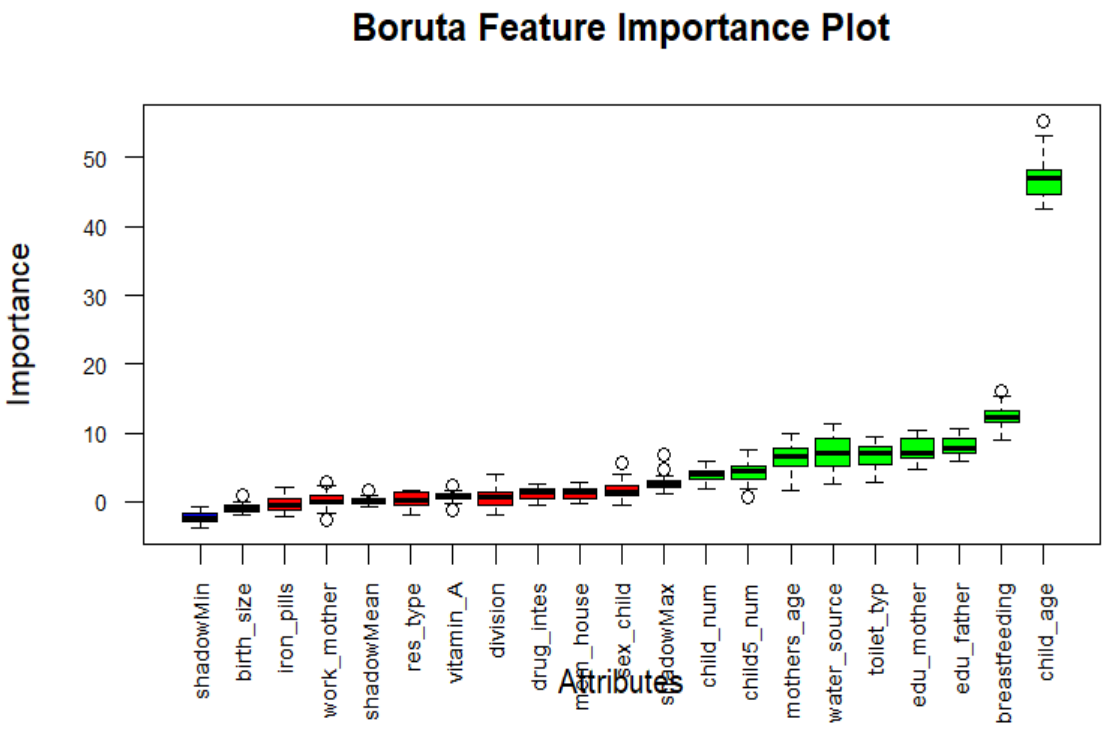

Figure 2. Features selection using the Boruta algorithm

According to the variable importance score (Table 4), several significant factors contribute to the likelihood of anemia in children, with the most influential being fathers' education, child age, breastfeeding status, mother's age, mother's education, toilet type, water source, and the number of children under 5 years. One of the most notable findings is that children whose fathers have completed higher education are less likely to be anemic. This is because parents with higher education levels typically possess a greater understanding of health, nutrition, and childrearing practices, which leads to better child care and healthier dietary habits [24-27]. Education also often correlates with higher income levels, enabling families to afford better healthcare and nutrition, both of which significantly contribute to the prevention of anemia. Regarding age, children between 1 and 2 years old are at a higher risk of moderate anemia, which is understandable as it can be challenging to provide balanced meals for children in this age group who are transitioning from breast milk to solid foods [2, 17]. The same reason is applicable for breastfeeding children, as those who are exclusively breastfed or breastfed for extended periods without adequate complementary feeding may be more prone to anemia [31]. Breastfeeding is widely acknowledged for its numerous health benefits, including providing infants with essential nutrients that support growth, development, and immune function. However, when it comes to anemia, breastfeeding alone may not provide sufficient iron for infants, especially as they grow [53]. While breast milk is rich in bioavailable iron, the total iron content is relatively low compared to other iron-rich foods [51]. For the first few months, exclusively breastfed infants typically receive enough iron from breast milk to prevent deficiency. However, as infants reach around 6 months, their iron stores, which were sufficient at birth, begin to deplete, and the body requires additional sources of iron. Without the introduction of iron-rich complementary foods like fortified cereals, meats, or legumes, the risk of iron deficiency anemia increases [52]. Therefore, while breastfeeding plays a protective role against anemia by promoting overall health and absorption of nutrients, its role in preventing iron deficiency anemia is limited unless complemented by appropriate weaning practices and a diversified diet that meets the child's growing nutritional needs [53, 54]. This is particularly important in low-resource settings, where access to iron-rich foods may be limited, and where exclusive breastfeeding for the first six months followed by timely introduction of solid foods is critical in preventing anemia [31].

Table 4. An overview of the Boruta algorithm's findings

Features	Feature Importance	Status
mothers_age	**0.968**	**Confirmed**
mothers_edu	**0.967**	**Confirmed**
fathers_edu	**1**	**Confirmed**
work_mother	0.065	Rejected
child_age	**1**	**Confirmed**
breastfeeding	**1**	**Confirmed**
sex_child	0.194	Rejected
toilet_typ	**0.964**	**Confirmed**
water_source	**0.959**	**Confirmed**
res_type	0	Rejected
division	0.032	Rejected
child_num	**0.83871**	**Confirmed**
birth_size	0	Rejected
vitamin_A	0	Rejected
iron_pills	0.0323	Rejected
drug_intes	0.064	Rejected
mem_house	0.032	Rejected
child5_num	**0.838**	**Confirmed**

Performance Evaluation of Different ML Techniques

Based on the potential risk factors identified as strongly linked through the chi-square test and the Boruta algorithm, seven different machine learning algorithms—KNN, NB, SVM, RF, Bagging, Gradient Boosting, and XGBoost—were utilized to classify children in the dataset into categories of "severe," "moderate," "mild," and "non-anemic." The study aimed to predict the anemia status of Bangladeshi children by using 70% of the observations for training and 30% for testing, with random seed 123. Additionally, to increase the model's performance, 10-fold Cross Validation (CV) was implemented. This technique helps in reducing overfitting and ensures that the model generalizes better to unseen data. The average performance metrics of the seven machine learning algorithms, including accuracy, sensitivity, specificity, precision, Cohen's Kappa, F1-score, and AUC, are summarized in Table 5. These metrics provide a comprehensive assessment of each algorithm's ability to predict the anemia status accurately and reliably, highlighting their strengths and limitations in handling imbalanced data and rare events [56].

The KNN algorithm achieved an accuracy of 72.72%**,** with a sensitivity of 82.76%**,** specificity of 77.08%**,** and precision of 75.38%**.** While KNN demonstrated a reasonable balance of sensitivity and specificity, its precision was somewhat higher than its specificity, suggesting that KNN predicted more positive instances accurately. The NB algorithm predicted the anemia status with 78.89% accuracy**,** 81.80% sensitivity, 83.61% specificity, and 81.19% precision. The NB algorithm performed relatively well, especially in terms of specificity, indicating it was effective in correctly identifying non-anemic instances.

The SVM with a linear kernel, RF, Bagging, and XGBoost achieved 84.46%**,** 83.13%**,** 78.29%**,** 87.46%**,** and 75.99% accuracy, respectively. Among these, SVM exhibited high sensitivity (83.37%) and a reasonable precision (77.95%), making it a strong performer in predicting both positive and negative cases. RF, with an accuracy of 83.13%**,** showed strong sensitivity (87.36%**)** but slightly lower specificity (83.01%). The RF algorithm exhibited robust predictive performance, particularly in identifying positive cases, though it showed some difficulty in distinguishing non-anemic cases. Bagging, with 78.29% accuracy, performed relatively evenly across all metrics but lagged behind RF and Gradient Boosting.

Table 5. Accuracy, sensitivity, specificity, precision and Kappa statistic of considered ML algorithms (Using 10-fold CV)

Methods	Accuracy (%) (Average)	Precision (%) (Average)	Specificity (%) (Average)	Sensitivity (%) (Average)	Kappa	F1 score	AUC
k-NN	72.72	75.38	77.08	82.76	0.3026	0.7901	0.7986
NB	78.89	81.19	83.61	81.80	0.3467	0.8143	0.8264
SVM (Linear)	84.46	77.95	82.04	83.37	0.3501	0.8046	0.8264
RF	83.13	84.10	83.01	87.36	0.3601	0.8555	0.8531
Bagging	78.29	77.92	81.43	77.68	0.3332	0.7780	0.7962
Gradient Boosting	87.46	95.35	96.56	85.31	0.5713	0.8990	0.9099
XGBoost	75.99	82.76	76.23	75.87	0.3319	0.7913	0.7599

The Gradient Boosting algorithm outperformed all other classifiers with 87.46% accuracy, 95.35% precision, 96.56% specificity, and 85.31% sensitivity. This model demonstrated exceptional precision and specificity, successfully predicting both positive and negative cases with a high degree of accuracy. The higher precision indicates that Gradient Boosting was particularly effective in minimizing false positives, a valuable attribute when forecasting rare events such as anemia. On the other hand, the k-NN algorithm showed the weakest performance, with 72.72% accuracy, 82.76% sensitivity, 77.08% specificity, and 75.38% precision. While k-NN's sensitivity was relatively high, its overall performance was significantly lower than other models in terms of predictive accuracy and specificity.

Cohen's kappa statistic, which measures the agreement between the observed and predicted classes, indicated that all the models had "fair" discriminative power. However, Gradient Boosting stood out with the highest kappa statistic of 0.5713**,** suggesting it had the most reliable classification performance compared to the other models [55]. The kappa values for the other models, though indicating fair performance [55], were lower, with the k-NN algorithm achieving 0.3026 and the RF model 0.3601.

The detailed performance metrics for all models after 10-fold cross-validation are summarized in Table 5. As shown, the Gradient Boosting model not only achieved the highest accuracy but also displayed the best overall performance in terms of precision, specificity, sensitivity, and F1 score, making it the most robust algorithm for anemia prediction in this study. This result contrasts with previous findings by [18, 19], who compared six machine learning algorithms (SVM, LR, k-NN, RF, LDA, and CART) and concluded that RF was the strongest predictive model. In contrast, our study found that RF was the second-best model, with Gradient Boosting outperforming it. This is consistent with research showing that Gradient Boosting handles unbalanced datasets more effectively and can predict rare events with greater accuracy than RF [56, 12], as demonstrated by its higher precision and specificity. Moreover, Gradient Boosting's ability to handle complex relationships between features contributes to its superior performance in comparison to other algorithms like RF and SVM.

Conclusions

In conclusion, anemia in children aged 6 to 59 months is associated with various relevant variables, including the father's educational background, the child's age, breastfeeding status, mother's age, mother's education, toilet type, and water source. Given these key risk factors, we compared seven machine learning prediction models to determine the likelihood of a child having anemia. Among these, Gradient Boosting emerged as the best-performing model, demonstrating superior classification accuracy in predicting anemia within the Bangladeshi population. This study underscores the potential of machine learning techniques in enhancing the prediction of disease status, particularly by leveraging common risk factors. Additionally, by identifying children at risk for anemia, our research can assist policymakers and healthcare professionals in implementing targeted interventions and improving patient care. Ultimately, a model based on these shared risk factors could be a valuable tool in the prevention and management of childhood anemia. These findings align with the Sustainable Development Goal (SDG) 3, which aims to ensure healthy lives and promote well-being for all at all ages. Improving the prediction and management of childhood anemia, particularly in resource-limited settings like Bangladesh, can help reduce child mortality, ensure universal access to health services, and promote overall health and well-being. However, to fully realize this potential in such environments, machine learning models must be adapted to local contexts and resources. Furthermore, incorporating longitudinal data can significantly enhance the performance of machine learning models for anemia prediction. Longitudinal data captures temporal patterns by tracking individuals over time, offering a comprehensive view of anemia's progression. It facilitates the identification of early risk factors and supports predictions of future outcomes, paving the way for more effective interventions and

personalized treatments. Unlike cross-sectional data, longitudinal data reflects the dynamic nature of anemia, reduces bias, and provides a nuanced understanding of each child's health history. Its inclusion also addresses challenges such as class imbalance and enhances predictive accuracy, making it an invaluable resource for advancing anemia prediction models. To maximize the impact of machine learning in resource-limited settings, models must be tailored to local contexts and constraints. Simpler, lightweight algorithms are particularly advantageous, as they require less computational power and can operate on low-spec devices common in these environments. Prioritizing interpretability and designing user-friendly decision support systems can further enhance usability. Offline functionality is especially critical in areas with limited internet access, ensuring that these tools remain reliable and accessible. By leveraging machine learning, incorporating longitudinal data, and adapting models to local needs, we can significantly improve the prediction and management of childhood anemia. This approach not only transforms healthcare delivery in low-resource settings like Bangladesh but also contributes to global efforts to ensure healthy lives and well-being for children.

Conflicts of Interest

The authors affirm that they have no financial, personal, or professional conflicts that could have influenced the research, analysis, or conclusions presented in this study. All findings and interpretations are solely based on objective analysis and are free from any competing interests.

Acknowledgement

I am deeply grateful to Universiti Sains Malaysia for granting me the opportunity to study at this esteemed institution. I also extend my sincere appreciation to the exceptional administrative team at the School of Mathematical Sciences, whose unwavering support helped me navigate and overcome the challenges faced during this research.

References

[1] Abdullah, M., & Al-Asmari, S. (2016). Anemia types prediction based on data mining classification algorithms. In *Communication, management and information technology* (pp. 629–636). CRC Press.

[2] Afroja, S., Kabir, M. R., & Islam, M. A. (2020). Analysis of determinants of severity levels of childhood anemia in Bangladesh using a proportional odds model. *Clinical Epidemiology and Global Health, 8*(1), 175–180.

[3] Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta–a system for feature selection. *Fundamenta Informaticae, 101*(4), 271–285.

[4] Ayoya, M. A., Ngnie-Teta, I., Séraphin, M. N., Mamadoultaibou, A., Boldon, E., Saint-Fleur, J. E., *et al.* (2013). Prevalence and risk factors of anemia among children 6–59 months old in Haiti. *Anemia*, 2013.

[5] Bangladesh Bureau of Statistics. (2004). *Anemia prevalence survey of Urban Bangladesh and Rural Chittagong Hill Tracts 2003*. Dhaka, Bangladesh: Bangladesh Bureau of Statistics, Statistics Division, Ministry of Planning, Government of the People's Republic of Bangladesh UNICEF.

[6] Balarajan, Y., Ramakrishnan, U., Özaltin, E., Shankar, A. H., & Subramanian, S. V. (2011). Anaemia in low-income and middle-income countries. *Lancet, 378*, 2123–2135. https://doi.org/10.1016/S0140-6736(10)62304-5

[7] Chowdhury, M. R. K., Khan, M. M. H., Khan, H. T., Rahman, M. S., Islam, M. R., Islam, M. M., & Billah, B. (2020). Prevalence and risk factors of childhood anemia in Nepal: A multilevel analysis. *PLOS ONE, 15*(10), e0239409.

[8] Leong, L. K., & Abdullah, A. A. (2019, November). Prediction of Alzheimer's disease (AD) using machine learning techniques with Boruta algorithm as feature selection method. In *Journal of Physics: Conference Series* (Vol. 1372, No. 1, p. 012065). IOP Publishing.

[9] Dutta, M., Bhise, M., Prashad, L., Chaurasia, H., & Debnath, P. (2020). Prevalence and risk factors of anemia among children 6–59 months in India: A multilevel analysis. *Clinical Epidemiology and Global Health*, 0–1. https://doi.org/10.1016/j.cegh.2020.02.015

[10] Desai, M. R., Terlouw, D. J., Kwena, A. M., Phillips-Howard, P. A., Kariuki, S. K., Wannemuehler, K. A., *et al.* (2005). Factors associated with hemoglobin concentrations in preschool children in western Kenya: Cross-sectional studies. *American Journal of Tropical Medicine and Hygiene, 72*, 47–59. https://doi.org/10.4269/ajtmh.2005.72.47

[11] Faruk, A. (2000). Anaemia in Bangladesh: A review of prevalence and aetiology. *Public Health Nutrition, 3*(4), 385–393.

[12] Fafalios, S., Charonyktakis, P., & Tsamardinos, I. (2020). Gradient Boosting Trees. *Gnosis Data Analysis PC*, 1–3.

[13] General Economics Division (GED). (2015). *Millennium Development Goals: Bangladesh Progress Report 2015*. Planning Commission, Government of the People's Republic of Bangladesh.

[14] Helen Keller International. (2006). The burden of anemia in rural Bangladesh: The need for urgent action.

Nutrition Surveillance Project Bulletin, 16.
[15] Horton, S., & Ross, J. (2003). The economics of iron deficiency. *Food Policy, 28*, 51–75. https://doi.org/10.1016/S0306-9192(02)00070-2
[16] Hsieh, C. H., Lu, R. H., Lee, N. H., Chiu, W. T., Hsu, M. H., & Li, Y. C. J. (2011). Novel solutions for an old disease: Diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. *Surgery, 149*(1), 87–93.
[17] International Centre for Diarrheal Disease Research Bangladesh (icddr,b), United Nations Children's Fund (UNICEF), Global Alliance for Improved Nutrition (GAIN), & Institute of Public Nutrition. (2013). *National Micronutrients Status Survey 2011–12: Final Report*. Dhaka, Bangladesh: Centre for Nutrition and Food Security, icddr,b.
[18] Khan, J. R., Awan, N., & Misu, F. (2016). Determinants of anemia among 6–59 months aged children in Bangladesh: Evidence from nationally representative data. *BMC Pediatrics, 16*, 1–12.
[19] Khan, J. R., Chowdhury, S., Islam, H., & Raheem, E. (2019). Machine learning algorithms to predict childhood anemia in Bangladesh. *Journal of Data Science, 17*(1), 195–218.
[20] Moschovis, P. P., Wiens, M. O., Arlington, L., Antsygina, O., Hayden, D., Dzik, W., *et al.* (2018). Individual, maternal and household risk factors for anaemia among young children in sub-Saharan Africa: A cross-sectional study. *BMJ Open, 8*, 1–14.
[21] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung Journal of Medical Sciences, 29*(2), 93–99.
[22] Ntenda, P. A. M., Nkoka, O., Bass, P., & Senghore, T. (2018). Maternal anemia is a potential risk factor for anemia in children aged 6–59 months in Southern Africa: A multilevel analysis. *BMC Public Health, 18*, 1–13.
[23] National Institute of Population Research, Training (Bangladesh), Mitra and Associates (Firm), & Macro International. (2011). *Bangladesh demographic and health survey*. National Institute of Population Research and Training (NIPORT).
[24] Leal, L. P., Batista Filho, M., Lira, P. I. C. D., Figueiroa, J. N., & Osório, M. M. (2011). Prevalence of anemia and associated factors in children aged 6–59 months in Pernambuco, Northeastern Brazil. *Revista de Saúde Pública, 45*, 457–466.
[25] Rahman, M. S., Mushfiquee, M., Masud, M. S., & Howlader, T. (2019). Association between malnutrition and anemia in under-five children and women of reproductive age: Evidence from Bangladesh Demographic and Health Survey 2011. *PLOS ONE, 14*(7), e0219170.
[26] Rashid, M., Flora, M. S., Moni, M. A., Akhter, A., & Mahmud, Z. (2010). Reviewing anemia and iron folic acid supplementation program in Bangladesh—a special article. *Bangladesh Medical Journal, 39*(3).
[27] Rawat, R., Saha, K. K., Kennedy, A., Rohner, F., Ruel, M., & Menon, P. (2014). Anaemia in infancy in rural Bangladesh: Contribution of iron deficiency, infections and poor feeding practices. *British Journal of Nutrition, 111*(1), 172–181.
[28] Sanap, S. A., Nagori, M., & Kshirsagar, V. (2011, December). Classification of anemia using data mining techniques. In *International Conference on Swarm, Evolutionary, and Memetic Computing* (pp. 113–121). Springer, Berlin, Heidelberg.
[29] Sunuwar, D. R., Singh, D. R., Pradhan, P. M. S., Shrestha, V., Rai, P., Shah, S. K., & Adhikari, B. (2023). Factors associated with anemia among children in South and Southeast Asia: A multilevel analysis. *BMC Public Health, 23*(1), 1–17.
[30] Stevens, G. A., Finucane, M. M., De-Regil, L. M., Paciorek, C. J., Flaxman, S. R., Branca, F., *et al.* (2013). Global, regional, and national trends in haemoglobin concentration and prevalence of total and severe anaemia in children and pregnant and non-pregnant women for 1995–2011: A systematic analysis of population-representative data. *Lancet Global Health, 1*(1), 16–25. https://doi.org/10.1016/s2214-109x(13)70001-9
[31] Stevens, G. A., Finucane, M. M., De-Regil, L. M., Paciorek, C. J., Flaxman, S. R., Branca, F., *et al.* (2013). Global, regional, and national trends in haemoglobin concentration and prevalence of total and severe anaemia in children and pregnant and non-pregnant women for 1995–2011: A systematic analysis of population-representative data. *Lancet Global Health,* 1, 16–25.
[32] Stevens, G. A., Finucane, M. M., De-Regil, L. M., Paciorek, C. J., Flaxman, S. R., Branca, F., *et al.* (2022). National, regional, and global estimates of anaemia by severity in women and children for 2000–19: A pooled analysis of population-representative data. *Lancet Global Health, 10*, e627–e639. https://doi.org/10.1016/S2214-109X(22)00084-5
[33] Mahesh, B. (2020). Machine learning algorithms—a review. *International Journal of Science and Research (IJSR)*, 9(1), 381–386.
[34] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16–28.
[35] World Health Organization. (2001). *Iron deficiency anaemia: Assessment, prevention and control: A guide for programme managers*. World Health Organization. https://doi.org/10.1136/pgmj.2009.089987
[36] World Health Organization. (2016). South Asia - Prevalence of anemia. Retrieved from https://www.indexmundi.com/facts/south-asia/prevalence-of-anemia
[37] Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & DATA, M. (2005, June). Practical machine learning tools and techniques. In *Data Mining* (Vol. 2, No. 4).
[38] Yusuf, A., Mamun, A. S. M. A., Kamruzzaman, M., Saw, A., Abo El-fetoh, N. M., Lestrel, P. E., & Hossain, M. (2019). Factors influencing childhood anaemia in Bangladesh: A two-level logistic regression analysis. *BMC Pediatrics, 19*(1), 1–9.
[39] Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making, 10*(1), 1–7.
[40] Zhang, Q., Ananth, C. V., Li, Z., & Smulian, J. C. (2009). Maternal anaemia and preterm birth: A prospective

cohort study. *International Journal of Epidemiology, 38*(5), 1380–1389.
[41] Zhao, Y., Healy, B. C., Rotstein, D., Guttmann, C. R., Bakshi, R., Weiner, H. L., *et al.* (2017). Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PloS One, 12*(4), e0174866.
[42] 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 Exerclse Testing (FIT) project. *PloS One, 12*(7), e0179805.
[43] Islam, M. M., Rahman, M. J., Roy, D. C., Islam, M. M., Tawabunnahar, M., Ahmed, N. F., & Maniruzzaman, M. (2022). Risk factors identification and prediction of anemia among women in Bangladesh using machine learning techniques. *Current Women's Health Reviews, 18*(1), 118–133.
[44] Tesfaye, S. H., Seboka, B. T., & Sisay, D. (2024). Application of machine learning methods for predicting childhood anaemia: Analysis of Ethiopian Demographic Health Survey of 2016. *PloS One, 19*(4), e0300172.
[45] Saputra, D. C. E., Sunat, K., & Ratnaningsih, T. (2023, February). A new artificial intelligence approach using extreme learning machine as the potentially effective model to predict and analyze the diagnosis of anemia. In *Healthcare, 11*(5), 697. MDPI.
[46] Gebeye, L. G., Dessie, E. Y., & Yimam, J. A. (2024). Predictors of micronutrient deficiency among children aged 6–23 months in Ethiopia: A machine learning approach. *Frontiers in Nutrition, 10*, 1277048.
[47] Qasrawi, R., Sgahir, S., Nemer, M., Halaikah, M., Badrasawi, M., Amro, M., *et al.* (2024). Machine learning approach for predicting the impact of food insecurity on nutrient consumption and malnutrition in children aged 6 months to 5 years. *Children, 11*(7), 810.
[48] Qasrawi, R., Badrasawi, M., Al-Halawa, D. A., Polo, S. V., Khader, R. A., Al-Taweel, H., *et al.* (2024). Identification and prediction of association patterns between nutrient intake and anemia using machine learning techniques: Results from a cross-sectional study with university female students from Palestine. *European Journal of Nutrition*, 1–15.
[49] Salma, N., Al-Rammahi, A. H. M., & Ali, M. K. M. (2024). A novel feature selection method for ultra-high dimensional survival data. *Malaysian Journal of Fundamental and Applied Sciences, 20*(5), 1149–1171.
[50] Reza, T. B., & Salma, N. (2024). Prediction and feature selection of low birth weight using machine learning algorithms. *Journal of Health, Population and Nutrition, 43*(1), 157.
[51] Dalili, H., Baghersalimi, A., Dalili, S., Pakdaman, F., Rad, A. H., Kakroodi, M. A., *et al.* (2015). Is there any relation between duration of breastfeeding and anemia? *Iranian Journal of Pediatric Hematology and Oncology, 5*(4), 218.
[52] Meinzen-Derr, J. K., Guerrero, M. L., Altaye, M., Ortega-Gallegos, H., Ruiz-Palacios, G. M., & Morrow, A. L. (2006). Risk of infant anemia is associated with exclusive breast-feeding and maternal anemia in a Mexican cohort. *The Journal of Nutrition, 136*(2), 452–458.
[53] Kramer, M. S., & Kakuma, R. (2012). Optimal duration of exclusive breastfeeding. *Cochrane Database of Systematic Reviews, (8)*.
[54] Marques, R. F., Taddei, J. A., Lopez, F. A., & Braga, J. A. (2014). Breastfeeding exclusively and iron deficiency anemia during the first 6 months of age. *Revista da Associação Médica Brasileira, 60*, 18–22.
[55] Rau, G., & Shih, Y. S. (2021). Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of English for Academic Purposes, 53*, 101026.
[56] Thölke, P., Mantilla-Ramos, Y. J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., *et al.* (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage, 277*, 120253.