Epidemiology and Health System Journal
doi:10.34172/ehsj.26563 2025 Autumn;12(4):177-190 http://ehsj.skums.ac.ir

Original Article

Determining Risk Factors of Anemia Among Under-Five Children in Gujarat, India: A Comparative Analysis of Feature Selection Methods

Nuriye Sancar[1*], Anjana Vasthava[1]

[1]Department of Mathematics, Near East University, 99138 Nicosia, Cyprus

*Corresponding Author: Nuriye Sancar, Email: nuriye.sancar@neu.edu.tr

Received: August 5, 2025
Revised: November 17, 2025
Accepted: November 18, 2025
ePublished: December 30, 2025

Abstract

Background and aims: In Gujarat, the prevalence of anemia among children under 5 years of age (U5C) is higher than the national average for the entire Indian population. Accordingly, this study aimed to identify risk factors for anemia among U5C in Gujarat using various feature selection methods and to compare their performance.

Methods: This cross-sectional study used National Family Health Survey-5 (2019–21) data of 8,058 children aged 6–59 months, selected through a stratified two-stage sampling design. Stepwise, backward, forward, correlation-based, and least absolute shrinkage and selection operator (LASSO) methods were applied to identify the most significant factors affecting anemia. Accuracy, recall, precision, F1-score, deviance, and the area under the curve (AUC) were utilized as performance metrics to assess feature selection performance.

Results: Performance metrics varied across the stepwise, backward, forward, and correlation-based methods. The accuracy, recall, precision, F1-score, AUC, and deviance were 0.612–0.895, 0.124–0.857, 0.419–0.650, 0.194–0.739, 0.651–0.685, and 10803.0–11221.4, respectively. The LASSO method outperformed all others (accuracy = 0.945, recall = 0.915, precision = 0.783, F1 = 0.843, AUC = 0.747, deviance = 7610.4). Key variables identified by LASSO included higher maternal education, improved sanitation, breastfeeding, vitamin A supplementation, and antenatal visits as protective factors, whereas unprotected drinking water and diarrhea treatment increased anemia risk. Ultimately, wealth index and cooking fuel type demonstrated significant associations.

Conclusion: Overall, targeting these modifiable factors substantially reduces the anemia burden in Gujarat, demonstrating the need for integrated public health and social interventions that effectively address maternal education, environmental health, and nutritional support to combat anemia.

Keywords: Anemia, Child, Preschool, India, Logistic models, Risk factors

Introduction

Anemia is a condition in which the body does not have enough healthy red blood cells or hemoglobin to transfer oxygen to its tissues.[1] It is clinically diagnosed based on hemoglobin concentration, with thresholds (e.g., < 11.0 g/dL) defined by the World Health Organization (WHO).[1] In addition, anemia is recognized as a critical public health problem, especially in children under 5 years old (U5C), due to its severe implications for physical and cognitive development.[2] Research has shown that anemia in early childhood negatively affects neurodevelopment, leading to impaired cognitive function, reduced attention span, lowered immunity, and poor growth.[3] According to the WHO (2021), about 40% of children are affected by anemia worldwide, with the condition being more prevalent in low-income and middle-income countries.[4] This burden is particularly high in South Asia, where childhood anemia is strongly linked to malnutrition, infectious diseases, and poor maternal health.[5]

Anemia is still a critical issue in India, and approximately 67% of Indian U5C suffer from anemia, according to the National Family Health Survey-5 (NFHS-5). The prevalence of anemia varies, according to trends from previous NFHS surveys. The prevalence was 74.3% in NFHS-2 (1998–99), but it dropped to 69.5% and then 58.5% in NFHS-3 (2005–06) and NFHS-4 (2015–16), respectively. However, the NFHS-5 (2019-2021) data published after the downward trend demonstrate that the prevalence increased to 67.1%. Therefore, examining the risk factors affecting anemia is essential to taking the necessary precautions. A similar trend has been observed in Gujarat, where the prevalence of childhood anemia has remained persistently high. NFHS-2 (1998-99) reported a prevalence of 74.5%, which slightly declined to 69.7% in NFHS-3 (2005-06) and further to 62.6% in NFHS-4 (2015-16). Nonetheless, the latest NFHS-5 (2019-21) data show a concerning increase to 79.7%, making Gujarat one of the worst-affected states in India.[6] Despite various government programs (e.g., the Anemia Mukt Bharat initiative), the prevalence of anemia remains alarmingly high in Gujarat, highlighting gaps in the efficacy of existing interventions and the requirement for region-specific policies.[7] Accordingly, understanding and addressing modifiable risk factors are crucial for developing effective interventions to reduce the prevalence of anemia among children, especially U5C.[8]

The WHO (2021) emphasizes that inadequate dietary iron intake, poor maternal nutrition, and exposure to infections (e.g., malaria and helminth infestations) considerably contribute to the persistence of childhood anemia, particularly in low-income and middle-income countries.[4,8] Several studies in India have also indicated that maternal anemia, antenatal care (ANC) utilization, breastfeeding practices, poor dietary diversity, recurrent infections, household environmental factors (e.g., access to clean water and sanitation), and lack of iron-rich complementary feeding have a role in high anemia rates among young children.[9-11]

To develop effective strategies to prevent anemia, it is important to understand factors that contribute to this health problem. Feature selection in medical research is vital because it increases the accuracy and validity of the research and helps correctly identify disease risk factors, thereby enabling early diagnosis and the development of effective preventive strategies.[12] To the best of our knowledge, most studies have employed direct multivariable logistic regression models without applying or comparing different feature selection methods.[13-15] In contrast, our study fills this methodological gap by systematically comparing stepwise forward, backward, correlation-based filtering, and LASSO feature selection approaches using the NFHS-5 Gujarat dataset. Therefore, this study seeks to determine the risk factors for anemia among U5C in Gujarat with greater reliability and sensitivity by comparing feature selection methods based on various criteria.

Materials and Methods

The current study employed a cross-sectional design and secondary data from the NFHS-5 (2019-2021), focusing on Gujarat's U5C. The required data were collected through NFHS-5 structured questionnaires via computer-assisted personal interviewing from June 17, 2019, to January 30, 2020, and from January 2 to April 30, 2021. The NFHS-5 is a national survey conducted in India that provides detailed information on demographics, health, nutrition, and related topics. The dataset includes variables related to child characteristics, health and nutritional factors, maternal factors, socioeconomic and household factors, reproductive and delivery history, and environmental and cultural factors. The survey was performed by the International Institute for Population Science under the guidance of the Ministry of Health and Family Welfare, Government of India. The study extracted data for Gujarat's U5C from the 'Children's Recode (KR) File'. The KR file has one record for every child of interviewed women born in the last five years (0–59 months).

Study Sample

The study population included U5C in Gujarat. Cases (6–59 months) with complete data on anemia status were included for analysis. In NFHS-5, anemia testing was performed for children aged 6–59 months after obtaining informed consent from a parent or guardian. Blood samples were collected via finger and heel pricks using a microcuvette and then analyzed on-site with a portable, battery-operated HemoCue Hb 201 + analyzer. This point-of-care testing provided immediate hemoglobin concentration values. Children with hemoglobin levels below 11.0 g/dL were classified as anemic, as per WHO guidelines, with altitude adjustments applied. Those with severe anemia (Hb < 7.0 g/dL) were referred to a healthcare facility for further assessment and treatment.[16] The inclusion criteria were children aged 6–59 months from Gujarat in NFHS-5 with available hemoglobin values. On the other hand, the exclusion criteria included children < 6 months and records with missing or implausible hemoglobin values. After removing 1,810 observations with missing data about anemia status, a total of 8,058 U5C (6–59 months) were included in the study.

Sample Size Consideration

The NFHS-5 adopted a stratified, two-stage sampling design to ensure national and state-level representativeness. In the first stage, primary sampling units (PSUs)-villages in rural areas and census enumeration blocks in urban areas were selected using probability proportional to size sampling. This method gives larger units a higher probability of selection, thereby ensuring proportional representation. The probability of choosing a PSU is calculated using the following formula:

$$P_i = \frac{M_i}{\sum_{j=1}^{N} M_j}$$

where P_i denotes the probability of selecting PSU I, and M_i is the population size of PSU i. Moreover, $\sum_{j=1}^{N} M_j$ represents the sum of the population sizes of all PSUs in the stratum.

In the second stage, 28 households were systematically selected from each PSU. The NFHS-5 KR (children's recode) file included data on children aged 0–59 months, collected from all eligible women who were interviewed. At the national level, the KR file covered approximately 232,920 children, and in Gujarat, it included 8,058 children aged 6–59 months after removing observations with missing anemia status and incomplete covariate data.[16]

Outcome and Independent Variables

The existing literature on factors affecting anemia in children in India was comprehensively reviewed to identify potential independent features. Based on the literature review and data availability, independent features were selected from a wide range of categories, including child factors (age, gender, size at birth, breastfeeding status, and bed net usage) and health and nutritional indicators (recent episodes of fever, cough, diarrhea, and treatment received, vitamin A supplementation). Other categories were maternal characteristics (age, educational level, and marital status), socioeconomic and household attributes (wealth index, parental employment, and household size), and reproductive and delivery history (birth order, total number of children, number of living children, type of birth, place of delivery, delivery by caesarean, and ANC). The remaining categories encompassed environmental factors (source of drinking water, type of cooking fuel, type of residence, and type of toilet facility) and cultural background (religion and caste).[15,17-21] Furthermore, anemia status was the outcome variable, which was categorized as anemic or non-anemic based on hemoglobin concentration (<11.0 g/dL), according to WHO guidelines. Independent variables were selected based on prior literature and data availability and were grouped into child, maternal, household, reproductive, environmental, and cultural domains. Among child factors, age was grouped as 6–23 months and 24–59 months. In addition, gender was classified as male or female, and size at birth was categorized as small, average, or large. Moreover, breastfeeding status and bed-net usage were grouped into currently breastfeeding or not breastfeeding and using or not using, respectively. Additionally, health and nutritional indicators included whether the child had a fever, cough, or diarrhea in the past two weeks (yes/no), whether treatment for fever or diarrhea was received (yes/no), and whether the child received vitamin A supplementation (yes/no). For maternal factors, mothers' age was grouped as 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, and 45–49 years, and maternal education was categorized as no education, primary, secondary, or higher education. Further, socioeconomic and household attributes contained the wealth index, divided into the poorest, poorer, middle, richer, and richest quintiles. Within reproductive and delivery history, birth order was grouped as 1, 2, 3, or 4 and above. Furthermore, the type of birth was classified as single or multiple, and the place of delivery was categorized as institutional, non-institutional, or other. Likewise, delivery by caesarean section and ANC visits were determined as "yes" or "no". Similarly, environmental factors included the source of drinking water (protected, unprotected, or other), household cooking fuel (electricity, liquified petroleum gas, biogas, and the like), type of residence (urban or rural), and type of toilet facility (no facility, improved sanitation, or unimproved sanitation). Finally, cultural factors consisted of religion (Hindu, Muslim, Christian, or other) and caste (scheduled caste, scheduled tribe, other backward class, or other).

In this study, the wealth index variable was derived directly from the NFHS-5 dataset, which constructs the index using the principal component analysis of household assets, housing characteristics, and access to essential services (e.g., water and sanitation facilities). Households are ranked based on asset scores and divided into the poorest, poorer, middle, richer, and richest quintiles. For the analysis, all five quintiles were retained as separate categories.[16]

Variable Encoding

As mentioned earlier, the outcome variable was anemia status (anemic/not anemic). All categorical independent variables were appropriately encoded before model development. In addition, nominal variables were transformed into dummy variables. Ordinal variables, such as wealth index and maternal education, were retained as categorical factors with ordered levels based on a conceptual or literature-supported hierarchy. Eventually, continuous variables were used in their original form unless normalization was necessary.

Statistical Analysis

All the statistical analyses were performed using R, version 4.3.2. The obtained data were imported with the haven package and processed by the dplyr and Hmisc packages. Stepwise forward and backward selection were conducted using the MASS and stats packages, while correlation filtering was performed by Cramér's V and Kendall's Tau from the DescTools package. Model training and evaluation used caret, with receiver operating characteristic/area under the curve (ROC/AUC) from pROC and F1-score from MLmetrics. Finally, LASSO logistic regression was fitted using glmnet, with λ selected via 10-fold cross-validation. Moreover, descriptive statistics were conducted on the entire dataset in order to characterize the study population. Qualitative variables were presented as frequencies with percentages, and continuous variables were presented as means and standard deviations (SD). Anemic and non-anemic group comparisons (Table 1) were performed using Pearson's chi-square test. Additionally, Fisher's exact test and Pearson's chi-square test were utilized for 2×2 tables with small expected counts and larger contingency tables, respectively. Two-sided *P*-values were reported, and the variance inflation factor (VIF) was employed to assess multicollinearity among variables. The data were randomly divided into a training set and a test set in an 80:20 ratio.

In the training set, stepwise, backward, forward, correlation, and LASSO feature selection methods were applied to identify key variables for anemia in U5C. In the stepwise selection procedures, logistic regression models were fitted, and variables were iteratively added or removed based on model fit statistics. In addition, variables were introduced one by one in forward selection. In backward elimination, all variables were initially included and removed sequentially. Further, both approaches were combined in stepwise selection, allowing variables to enter or leave the model at each step. In these procedures, the *alpha-to-enter* parameter was set at 0.05, implying that a variable was included in the model only if its *P*-value was less than or equal to 0.05, indicating sufficient statistical significance for entry. Conversely, the *alpha-to-remove* parameter was set at 0.10, which defined the threshold for variable removal. Any variable with a *P*-value greater than or equal to 0.10 was excluded from the model. This iterative process continued until a stable set of variables satisfying both inclusion and exclusion thresholds was obtained. In the correlation-based method, pairwise associations between variables and the outcome were evaluated for feature selection. For nominal variables, associations were quantified using Cramér's V. Furthermore, Kendall's tau was used for ordinal variables. Variables with moderate-to-high correlation strength were selected for model inclusion. A binary logistic regression model was constructed on this training set for each selected feature set (except for LASSO) in order to investigate the correlation between selected variables and anemia status. LASSO logistic regression performs both parameter estimation and variable selection simultaneously using the L1 penalty term in order to reduce the coefficients of some redundant variables to zero. The results are presented as odds ratios (ORs) with 95% confidence intervals (CIs). The final model was developed using the variables retained after feature selection. The LASSO identified a broader array of health-related variables than other methods, owing to its regularization mechanism, which suppresses redundant variables while retaining clinically meaningful ones. In the LASSO, the tuning parameter λ was estimated via 10-fold cross-validation, selecting the value that minimized the cross-validated deviance. Moreover, the algorithm iterated until convergence, with a maximum of 10,000 iterations per model fit (as the glmnet default).

All feature selection methods were applied using identical logistic regression model structures. The outcome variable "anemia status" was consistently coded as binary (0 = not anemic and 1 = anemic) across all models. A uniform classification threshold of 0.5 was used, and the same 80:20 data split was maintained to ensure comparability across model performance evaluations.

On the test set, performance metrics derived from the confusion matrix were calculated to evaluate the predictive performance of models constructed with different feature selection methods. The confusion matrix was generated to compute performance measurement criteria, such as accuracy, F1-score, precision, and recall, with TP, FP, FN, and TN as true positive, false positive, false negative, and true negative, respectively. TP represents the number of children who have anemia and are correctly identified by the model as having anemia. In addition, FP demonstrates the number of children who do not have anemia but are incorrectly classified as having anemia by the model. Further, FN is the number of children who have anemia but are incorrectly classified as not having anemia, and TN denotes the number of children who do not have anemia and are correctly identified by the model as not having anemia. The area under the ROC curve was computed using the trapezoidal rule from ROC analysis. The ROC curve plots sensitivity (true positive rate) against 1 – specificity (false positive rate) at all possible classification thresholds and provides a comprehensive evaluation of the model's discriminative ability. Eventually, deviance was calculated to indicate model fit, with lower deviance indicating better fit.

Results

Table 1 presents the descriptive statistics of the study population. Among the 8,058 U5C included in the analysis, 80.54% were identified as anemic based on WHO-defined hemoglobin thresholds. Among the 4,142 male children, 81.4% were anemic, while among the 3,916 females, the prevalence was slightly lower at 79.6%. The mean age of children in the anemic group was 31.64 months (SD = 15.36), whereas it was higher among non-anemic children at 36.87 (SD = 15.96) months. Moreover, the age distribution ranged from 6 months to 59 months in both cohorts. Significant group differences between anemic and non-anemic groups ($P < 0.05$) were observed in terms of child's age, gender, breastfeeding status, bed net usage, diarrhea, treatment for fever and diarrhea, vitamin A supplementation, maternal age, maternal education, wealth index, birth order, caesarean delivery, drinking water source, household fuel type, type of residence, toilet facility, religion, and caste (Table 1). However, no significant differences were found regarding birth weight, history of fever or cough, kind of birth, place of delivery, or ANC visits.

Multicollinearity diagnostics revealed acceptable VIF values, with all variables falling below the conventional threshold of 5, except for the variable "Received treatment for fever or cough" (VIF = 5.011), which was only marginally elevated and did not compromise the overall model stability.[22]

Multiple feature selection methods, including stepwise, backward, and forward logistic regression, correlation-based filtering, and LASSO, were employed to identify influential variables for anemia.

The feature selection performance of each method was evaluated on the testing set, and 95% CIs were obtained for each method. Table 2 lists the performance metrics for each feature selection method, with 95% CIs for the testing and training sets. Stepwise, backward, and forward selection methods yielded moderate accuracy (0.612 [0.598–0.636], 0.615 [0.602–0.631], and 0.621 [0.614–0.630]) and modest AUCs (0.651 [0.639–0.663], 0.655 [0.642–0.668], and 0.663 [0.642–0.674]), but recall was consistently low (0.133 [0.128–0.148], 0.145 [0.137–0.153], and 0.124 [0.118–0.131]), indicating limited ability to identify actual anemia cases. Likewise, F1-scores remained low (0.202 [0.197–0.216], 0.215 [0.201–0.229], and 0.194 [0.179–0.212]), reflecting the jointly

Table 1. Descriptive Statistics of Sociodemographic, Behavioral, and Health-Related Characteristics of the Study Population

Variables		Total	Non-Anemic	Anemic	*P*-Value
		n (%)	n (%)	n (%)	
Child factors					
Age of child	6-23 months	2,659 (33)	378 (14.2)	2,281 (85.8)	<0.0001
	24-59 months	5,399 (67)	1,190 (22.0)	4,209 (78.0)	
Gender of the child	Male	4,142 (51.4)	770 (18.6)	3,372 (81.4)	0.043
	Female	3,916 (48.6)	798 (20.4)	3,118 (79.6)	
Size of the child*	Large	2,055 (25.7)	401 (19.5)	1,654 (80.5)	0.858
	Average	5,006 (62.5)	966 (19.3)	4,040 (80.7)	
	Small	947 (11.8)	190 (20.1)	757 (79.9)	
Breastfeeding status	Not breastfeeding	5,092 (63.2)	1,108 (21.8)	3,984 (78.2)	<0.0001
	Currently breastfeeding	2,966 (36.8)	460 (15.5)	2,506 (84.5)	
Bed net usage	Not using	1,042 (13.1)	176 (16.9)	866 (83.1)	0.029
	Using	6,923 (86.9)	1,368 (19.8)	5,555 (80.2)	
Health and nutrition					
Fever	No	7,096 (88.1)	1,403 (19.8)	5,693 (80.2)	0.054
	Yes	962 (11.9)	165 (17.2)	797 (82.8)	
Cough	No	7,272 (90.2)	1,428 (19.6)	5,844 (80.4)	0.220
	Yes	786 (9.8)	140 (17.8)	646 (82.2)	
Diarrhea	No	7,344 (91.1)	1,469 (20.0)	5,875 (80.0)	<0.0001
	Yes	714 (8.9)	99 (13.9)	615 (86.1)	
Received treatment for fever	Not received	7,269 (90.2)	1,437 (19.8)	5,832 (80.2)	0.033
	Received	789 (9.8)	131 (16.6)	658 (83.4)	
Received treatment for diarrhea	Not received	7,508 (93.2)	1,491 (19.9)	6,017 (80.1)	0.001
	Received	550 (6.8)	77 (14.0)	473 (86.0)	
Vitamin A supplementation	No	4,435 (55.0)	1,067 (24.1)	3,368 (75.9)	<0.0001
	Yes	3,623 (45.0)	501 (13.8)	3,122 (86.2)	
Maternal factors					
Age of mother	15-19	97 (1.2)	10 (10.3)	87 (89.7)	<0.0001
	20-24	2,126 (26.4)	356 (16.7)	1,770 (83.3)	
	25-29	3,363 (41.7)	649 (19.3)	2,714 (80.7)	
	30-34	1,721 (21.4)	364 (21.2)	1,357 (78.8)	
	35-39	608 (7.5)	150 (24.7)	458 (75.3)	
	40-44	111 (1.4)	27 (24.3)	84 (75.7)	
	45-49	32 (0.4)	12 (37.5)	20 (62.5)	
Mothers' education	No education	1,737 (21.6)	312 (18.0)	1,425 (82.0)	0.001
	Primary	1,152 (14.3)	192 (16.7)	960 (83.3)	
	Secondary	4,441 (55.1)	891 (20.1)	3,550 (79.9)	
	Higher	728 (9.0)	173 (23.8)	555 (76.2)	
Socioeconomic and household					
Wealth index	Poorest	1,465 (18.2)	184 (12.6)	1,281 (87.4)	<0.0001
	Poorer	1,791 (22.2)	295 (16.5)	1,496 (83.5)	
	Middle	1,745 (21.7)	335 (19.2)	1,410 (80.8)	
	Richer	1,679 (20.8)	404 (24.1)	1,275 (75.9)	
	Richest	1,378 (17.1)	350 (25.4)	1,028 (74.6)	
Reproductive and delivery history					
Birth order	1	3,208 (39.8)	663 (20.7)	2,545 (79.3)	0.007
	2	2,697 (33.5)	538 (19.9)	2,159 (80.1)	
	3	1,266 (15.7)	224 (17.7)	1,042 (82.3)	
	4+	887 (11.0)	143 (16.1)	744 (83.9)	
Type of birth	Single birth	7,937 (98.5)	1,542 (19.4)	6,395 (80.6)	0.570
	Multiple birth	121 (1.5)	26 (21.5)	95 (78.5)	

Table 3. Continued.

Variables		Total n (%)	Non-Anemic n (%)	Anemic n (%)	*P*-Value
Place of delivery	Institutional	7,470 (92.7)	1,459 (19.5)	6,011 (80.5)	0.445
	Non-institutional	579 (7.2)	106 (18.3)	473 (81.7)	
	Other	9 (0.1)	3 (33.3)	6 (66.7)	
Delivery by caesarean	No	6,581 (81.7)	1,247 (18.9)	5,334 (81.1)	0.015
	Yes	1,477 (18.3)	321 (21.7)	1,156 (78.3)	
ANC visits	No	8,011 (99.4)	1,554 (19.4)	6,457 (80.6)	0.073
	Yes	47 (0.6)	14 (29.8)	33 (70.2)	
Environmental factors					
Drinking water source	Protected	3,328 (41.3)	563 (16.9)	2,765 (83.1)	<0.0001
	Unprotected	4,450 (55.2)	956 (21.5)	3,494 (78.5)	
	Other	280 (3.5)	49 (17.5)	231 (82.5)	
Household cooking fuel	Electricity	39 (0.5)	6 (15.4)	33 (84.6)	<0.0001
	LPG	3,930 (48.8)	888 (22.6)	3,042 (77.4)	
	Biogas	20 (0.2)	3 (15.0)	17 (85.0)	
	Other	4,069 (50.5)	671 (16.5)	3,398 (83.5)	
Type of residence	Urban	2,206 (27.4)	524 (23.8)	1,682 (76.2)	<0.0001
	Rural	5,852 (72.6)	1,044 (17.8)	4,808 (82.2)	
Type of toilet facility	No facility	2,290 (28.4)	343 (15.0)	1,947 (85.0)	<0.0001
	Improved sanitation	5,453 (67.7)	1,168 (21.4)	4,285 (78.6)	
	Unimproved sanitation	315 (3.9)	57 (18.1)	258 (81.9)	
Cultural factors					
Religion	Hindu	7,060 (87.6)	1,304 (18.5)	5,756 (81.5)	<0.0001
	Muslim	924 (11.5)	241 (26.1)	683 (73.9)	
	Christian	47 (0.6)	16 (34.0)	31 (66.0)	
	Other	27 (0.3)	7 (25.9)	20 (74.1)	
Caste	Scheduled Caste	894 (11.1)	195 (21.8)	699 (78.2)	<0.0001
	Scheduled Tribe	2,074 (25.7)	258 (12.4)	1,816 (87.6)	
	OBC	3,728 (46.3)	772 (20.7)	2,956 (79.3)	
	Other	1,362 (16.9)	343 (25.2)	1,019 (74.8)	

Note. *Size of the child refers to the mother's perception of the baby's size at birth as recorded in the National Family Health Survey-5. ANC: Antenatal care; LPG: Liquefied petroleum gas; OBC: Other backward class.
Source. [16].

Table 2. Performance Metrics With 95% Confidence Intervals for Feature Selection Methods on Testing and Training Sets

	Accuracy	Recall	Precision	F1-Score	AUC	Deviance
Testing Set Results						
Stepwise	0.612 (0.598–0.636)	0.133 (0.128–0.148)	0.421 (0.407–0.432)	0.202 (0.197–0.216)	0.651 (0.639–0.663)	11221.4 (10976.4-1466.4)
Backward	0.615 (0.602–0.631)	0.145 (0.137–0.153)	0.419 (0.402–0.436)	0.215 (0.201–0.229)	0.655 (0.642–0.668)	11183.8 (10948.6–11419.0)
Forward	0.621 (0.614–0.630)	0.124 (0.118–0.131)	0.448 (0.426–0.472)	0.194 (0.179–0.212)	0.663 (0.642–0.674)	11197.0 (10955.9–11438.1)
Correlation	0.895 (0.883–0.907)	0.857 (0.843–0.871)	0.650 (0.630–0.681)	0.739 (0.725–0.760)	0.685 (0.667–0.703)	10803.0 (10616.8–10989.2)
LASSO	0.945 (0.939–0.953)	0.915 (0.907–0.926)	0.783 (0.772–0.804)	0.843 (0.836–0.860)	0.747 (0.738–0.764)	7610.4 (7449.7–7771.1)
Training Set Results						
Stepwise	0.623 (0.585–0.636)	0.138 (0.121–0.152)	0.434 (0.413–0.448)	0.209 (0.198–0.222)	0.654 (0.647–0.668)	11189.1 (10942.5–11424.9)
Backward	0.626 (0.600–0.638)	0.151 (0.125–0.160)	0.438 (0.421–0.442)	0.218 (0.207–0.231)	0.660 (0.647–0.671)	11145.2 (10913.3–11352.1)
Forward	0.630 (0.609–0.643)	0.135 (0.112–0.145)	0.453 (0.421–0.465)	0.202 (0.195–0.218)	0.668 (0.656–0.679)	11106.5 (10798.3-11335.0)
Correlation	0.900 (0.884–0.909)	0.861 (0.848–0.879)	0.657 (0.624–0.684)	0.742 (0.726–0.759)	0.689 (0.680–0.711)	10785.6 (10612.7–10963.4)
LASSO	0.948 (0.932–0.953)	0.914 (0.900–0.921)	0.789 (0.778–0.799)	0.848 (0.824–0.851)	0.747 (0.728–0.769)	7542.2 (7328.0–7803.6)

Note. AUC: Area under the curve; LASSO: The least absolute shrinkage and selection operator.

limited sensitivity and precision. Nonetheless, deviance values were comparatively higher (11221.4 [10976.4–11466.4], 11183.8 [10948.6–11419.0], and 11197.0 [10955.9–11438.1]), consistent with the weaker overall fit. Additionally, the correlation-based method substantially improved feature selection performance, with a recall of 0.857 (0.843–0.871) and a more balanced precision of 0.650 (0.626–0.674), yielding a markedly higher F1-score of 0.739 (0.720–0.758). Similarly, accuracy increased to 0.895 (0.883–0.907). Although the AUC remained moderate at 0.685 (0.667–0.703), the deviance decreased to 10803.0 (10616.8–10989.2), indicating a better model fit than stepwise methods.

It should be noted that LASSO consistently outperformed all feature selection methods across all metrics, achieving the highest accuracy, recall, precision, F1-score, and AUC at 0.945 (0.939–0.953), 0.915 (0.907–0.926), 0.783 (0.772–0.804), 0.843 (0.836–0.860), and 0.747 (0.738–0.764), respectively, but the lowest deviance at 7610.4 (7449.7–7771.1). These consistent improvements indicate that the variables retained by LASSO constitute a robust and informative subset of features, thereby enhancing both discriminative ability and overall model fit. Overall, these findings affirm the effectiveness of LASSO in identifying both direct and indirect risk factors while maintaining reduced model complexity [23]. Performance metrics on the training set demonstrated a similar pattern, with slightly higher values across all methods. Notably, LASSO maintained its superior performance across both the training and test datasets, representing successful variable selection and stability.

Table 3 provides the variables retained by each method, along with ORs and 95% CIs. Wealth index, source of drinking water, and type of cooking fuel consistently appeared across all methods, emphasizing their robustness. However, bed net use, diarrhea treatment, maternal education, and breastfeeding status were retained inconsistently, underscoring the methodological sensitivity of feature-selection procedures. This pattern highlights the necessity of a multi-method approach to reliably capture diverse variables based on their underlying statistical associations and collinearity structures. Remarkably, LASSO also identified several indirect variables, such as vitamin A supplementation and ANC, which are often underrepresented in traditional selection methods, demonstrating its sensitivity in uncovering complex variable relationships. Each method identified a somewhat distinct subset of predictors; however, the comparative analysis revealed that the LASSO method achieved superior performance in identifying key predictors of childhood anemia compared to the other intended methods.

LASSO-derived logistic regression results identified several key variables associated with anemia. Maternal education emerged as a strong protective factor, with children of mothers who had higher education levels exhibiting significantly lower odds of anemia (OR = 0.571; 95% CI: 0.451–0.722). In addition, improved drinking water access reduced the risk (OR = 0.761; CI: 0.584–0.992), while reliance on unprotected sources increased anemia risk. Moreover, improved sanitation was related to lower anemia risk (OR = 0.547; CI: 0.424–0.706), underscoring the role of hygiene in micronutrient status.

Cooking fuel type had a considerable effect. The use of biogas significantly reduced the risk of anemia (OR = 0.044; CI: 0.031–0.063), likely through reduced indoor air pollution. Conversely, liquified petroleum gas and natural gas users had slightly elevated risks (OR = 1.196 and OR 1.205), suggesting that even relatively clean fuels may have residual adverse effects or reflect underlying confounding factors, such as socioeconomic status.

Socioeconomic status was measured via the wealth index, which was inversely associated with anemia. Compared to the poorest group, children from middle-income households had 60% reduced odds (OR = 0.400; CI: 0.366–0.437), while the richer (OR = 0.836; CI: 0.765–0.913) and richest (OR = 0.859; CI: 0.786–0.938) groups also demonstrated statistically significant protective associations.

The use of bednets was linked with lower odds of anemia (OR = 0.736; CI: 0.674-0.804), reflecting its role in malaria prevention. Similarly, breastfeeding was protective (OR = 0.671; CI: 0.614–0.733), possibly due to enhanced iron bioavailability and immune defense in breastfed infants.

Unexpectedly, children who received treatment for diarrhea had an elevated risk of anemia (OR = 1.309; CI: 1.186–1.443), potentially indicating reverse causality or confounding by more severe underlying conditions. Likewise, vitamin A supplementation showed a strong protective association (OR = 0.739; CI: 0.618–0.833), likely through improved iron metabolism and immune function. Finally, ANC visits also emerged as a protective maternal factor (OR = 0.753; CI: 0.689–0.825), highlighting the role of maternal health in fetal and postnatal iron status.

These findings emphasize that a constellation of modifiable factors (nutritional, environmental, maternal, and socioeconomic) will generally contribute to anemia in early childhood. Accordingly, interventions targeting these domains may offer substantial gains in anemia prevention.

Discussion

This study evaluated the significant risk factors of anemia among U5C in Gujarat and assessed the predictive performance of logistic regression models using various feature selection methods. The findings revealed that toilet facility type, wealth index, cooking fuel type, drinking water source, mother's education, bed net usage, breastfeeding status, treatment for diarrhea and fever, vitamin A supplementation, and antenatal visits were the most critical variables identified by the LASSO method. The wealth index demonstrated an inverse relationship with anemia risk. Children from wealthier households had lower odds of anemia, which conforms

Table 3. Odds Ratios and 95% Confidence Intervals for Variables Retained by Each Feature Selection Method

Method	Variable	Category	Odds Ratio	95% Confidence Interval for Odds Ratio
LASSO	Educational level of the mother	No Education (Ref.)	—	—
		Primary	0.841	(0.623, 1.135)
		Secondary	0.879	(0.670, 1.154)
		Higher	0.571*	(0.451, 0.722)
	Source of drinking water	Protected (Ref.)	—	—
		Unprotected	1.313*	(1.008, 1.711)
		Other	1.016	(0.747, 1.380)
	Type of toilet facility	No facility (Ref.)	—	—
		Improved Sanitation	0.547*	(0.424, 0.706)
		Unimproved sanitation	0.714	(0.522, 1.020)
	Type of cooking fuel	Electricity (Ref.)	—	—
		LPG	1.196*	(1.003, 1.427)
		Biogas	0.044*	(0.031, 0.063)
		Other	0.950	(0.901, 1.002)
	Wealth index	Poorest (Ref.)	—	—
		Poorer	0.244	(0.093, 1.002)
		Middle	0.400*	(0.366, 0.437)
		Richer	0.836*	(0.765, 0.913)
		Richest	0.859*	(0.786, 0.938)
	Bed net usage	No (Ref.)	—	—
		Yes	0.736*	(0.674, 0.804)
	Breastfeeding status	Not breastfeeding (Ref.)	—	—
		Currently breastfeeding	0.671*	(0.614, 0.733)
	Received treatment for diarrhea	Not received (Ref.)	—	—
		Received	1.309*	(1.186, 1.443)
	Had a fever	No (Ref.)	—	—
		Yes	1.354	(1.277, 1.436)
	Vitamin A supplementation	No (Ref.)	—	—
		Yes	0.739*	(0.618, 0.883)
	Antenatal visits	No (Ref.)	—	—
		Yes	0.753*	(0.689, 0.822)
Stepwise	Wealth index	Poorest (Ref.)	—	—
		Poorer	0.921	(0.823, 1.063)
		Middle	0.863*	(0.773, 0.963)
		Richer	0.884*	(0.681, 0.985)
		Richest	0.902	(0.737, 1.021)
	Source of drinking water	Protected (Ref.)	—	—
		Unprotected	1.273*	(1.034, 1.475)
		Other	1.054	(0.840, 1.277)
	Caste	OBC (Ref.)	—	—
		SC	1.221*	(1.040, 1.425)
		ST	0.940	(0.833, 1.078)
		Other	1.084	(0.951, 1.232)
	Religion	Hindu (Ref.)	—	—
		Muslim	1.241*	(1.064, 1.587)
		Christian	1.185	(0.949, 1.373)
		Other	1.108	(0.910, 1.321)

Table 3. Continued.

Method	Variable	Category	Odds Ratio	95% Confidence Interval for Odds Ratio
Stepwise	Type of cooking fuel	Electricity (Ref.)	—	—
		LPG	0.932	(0.840, 1.063)
		Biogas	0.583*	(0.414, 0.893)
		Other	1.031	(0.945, 1.149)
	Bed net usage	No (Ref.)	—	—
		Yes	1.283*	(1.076, 1.402)
	The mother's level of education	No education (Ref.)	—	—
		Primary	0.866*	(0.770, 0.976)
		Secondary	0.773*	(0.654, 0.898)
		Higher	0.962	(0.837, 1.131)
	Birth order	1 (Ref.)	—	—
		2	1.165*	(1.013, 1.338)
		3	1.094	(0.951, 1.255)
		4+	1.030	(0.923, 1.166)
	Had a cough	No (Ref.)	—	—
		Yes	1.137*	(1.011, 1.197)
	Marital status of mother	Other (Ref.)	—	—
		Married	1.196*	(1.020, 1.456)
Backward	Age of mother	25–29 (Ref.)	—	—
		15–19	0.920	(0.710, 1.189)
		20–24	0.954	(0.864, 1.073)
		30–34	1.185*	(1.014, 1.321)
		35–39	1.083	(0.943, 1.241)
		40–44	1.039	(0.960, 1.145)
		45-49	1.054	(0.941, 1.203)
	Source of drinking water	Protected (Ref.)	—	—
		Unprotected	1.331*	(1.110, 1.626)
		Other	1.065	(0.872, 1.295)
	Type of toilet facility	No facility (Ref.)	—	—
		Improved	0.843*	(0.744, 0.988)
		Unimproved	0.928	(0.795, 1.094)
	Religion	Hindu (Ref.)	—	—
		Muslim	1.125*	(1.017, 1.306)
		Christian	1.161	(0.982, 1.393)
	Type of cooking fuel	Electricity (Ref.)	—	—
		LPG	0.918	(0.793, 1.048)
		Other	1.084	(0.938, 1.142)
		Biogas	0.552*	(0.404, 0.781)
	Wealth index	Poorest (Ref.)	—	—
		Poorer	0.950	(0.854, 1.083)
		Middle	0.894*	(0.804, 0.993)
		Richer	0.958	(0.867, 1.062)
	Children ever born	1 (Ref.)	—	—
		2	0.851	(0.774, 1.002)
		3	1.121	(0.966, 1.287)
		4+	1.169*	(1.014, 1.302)

Table 3. Continued.

Method	Variable	Category	Odds Ratio	95% Confidence Interval for Odds Ratio
Backward	Birth order	1 (Ref.)	—	—
		2	1.125*	(1.022, 1.497)
		3	1.023*	(1.009, 1.124)
		4+	1.105	(0.988, 1.220)
	Breastfeeding	No (Ref.)	—	—
		Currently	0.775*	(0.673, 0.894)
	Had a fever	No (Ref.)	—	—
		Yes	1.159*	(1.031, 1.314)
	Caste	OBC (Ref.)	—	—
		SC	1.113*	(1.004, 1.273)
		ST	0.945	(0.838, 1.074)
Forward	Wealth index	Poorest (Ref.)	—	—
		Poorer	0.943	(0.837, 1.060)
		Middle	0.884*	(0.690, 0.988)
		Richer	0.921	(0.838, 1.037)
		Richest	0.863*	(0.652, 0.969)
	Religion	Hindu (Ref.)		
		Muslim	1.194*	(1.052, 1.547)
		Christian	1.152	(0.976, 1.389)
	Had a fever (Yes vs. No)	No (Ref.)	—	—
		Yes	1.230*	(1.083, 1.445)
	Type of cooking fuel	Electricity (Ref.)	—	—
		LPG	1.143*	(1.035, 1.267)
		Biogas	0.608*	(0.442, 0.811)
		Other	0.902	(0.741, 1.056)
	Source of drinking water	Protected (Ref.)	—	—
		Unprotected	1.251*	(1.072, 1.512)
		Other	1.044	(0.874, 1.281)
	Age of mother	15-19	0.620*	(0.401, 0.950)
		20-24	0.965	0.854, 1.087)
		25-29 (Ref.)	----	—
		30-34	0.834*	(0.724, 0.973)
		35-39	0.921	(0.748, 1.158)
		40-44	0.861	(0.566, 1.301)
		45-49	0.887	(0.544, 1.334)
	Type of toilet facility	No facility (Ref.)	—	—
		Improved	0.851*	(0.673, 0.958)
		Unimproved	0.889	(0.685, 1.004)
	Bed net usage	No (Ref.)	—	—
		Yes	0.886*	(0.684, 0.942)
Correlation-based	Caste	OBC (Ref.)	—	—
		SC	1.193*	(1.015, 1.220)
		ST	0.941	(0.831, 1.075)
	Religion	Hindu (Ref.)	—	—
		Muslim	1.163*	(1.034, 1.383)
		Christian	1.225*	(1.094, 1.405)
	Type of residence	Urban (Ref.)	—	—
		Rural	0.780*	(0.542, 0.856)

Table 3. Continued.

Method	Variable	Category	Odds Ratio	95% Confidence Interval for Odds Ratio
Correlation-based	The education level of the mother	No education (Ref.)	—	—
		Primary	0.943	(0.820, 1.193)
		Secondary	0.884*	(0.672, 0.978)
		Higher	0.821*	(0.614, 0.940)
	Wealth index	Poorest (Ref.)	—	—
		Poorer	0.930	(0.808, 1.159)
		Middle	0.805*	(0.717, 0.938)
		Richer	0.764*	(0.681, 0.904)
		Richest	0.714*	(0.614, 0.832)
	Type of cooking fuel	Electricity (Ref.)	—	—
		LPG	0.895*	(0.775, 0.954)
		Biogas	0.574*	(0.413, 0.819)
		Other	0.902	(0.853, 1.147)
	Type of toilet facility	No facility (Ref.)	—	—
		Improved	0.834*	(0.714, 0.945)
		Unimproved	0.924	(0.784, 1.073)
	Age of mother	15-19	0.820*	(0.715, 0.994)
		20-24	0.996	(0.881, 1.100)
		25-29 (Ref.)	----	—
		30-34	0.934	(0.786, 1.059)
		35-39	0.901	(0.857, 1.099)
		40-44	0.957	(0.808, 1.287)
		45-49	0.821	(0.619, 1.274)
	Source of drinking water	Protected (Ref.)	—	—
		Unprotected	1.281*	(1.093, 1.475)
		Other	0.984	(0.901, 1.257)

Note. *$P<0.05$; Ref.: Reference category. LASSO: The least absolute shrinkage and selection operator; LPG: Liquefied petroleum gas; OBC: Other Backward Class; SC: Scheduled Castes; ST: Scheduled Tribes.

to the results of other studies, linking economic status to better nutrition, improved living conditions, and greater healthcare access.[10,24-26] In families with low socioeconomic status, cow's milk is generally the main component of complementary nutrition [27]. The consumption of cow's milk is associated with iron deficiency anemia, especially in early childhood. Cow's milk, which contains minerals such as calcium and phosphorus that can prevent intestinal iron absorption, has a low iron content and can cause microscopic intestinal bleeding, leading to iron loss; it should, therefore, be limited in infants' diets [28].

On the other hand, fever may indicate ongoing infections, and long-term infections can adversely affect iron metabolism, causing the depletion of iron stores. The heightened inflammatory response during infection elevates hepcidin levels, thereby reducing intestinal iron absorption and sequestering iron. Over time, this process can contribute to the development of iron-deficiency anemia [29]. Chronic diarrhea disrupts intestinal mucosal integrity, thereby negatively affecting nutrient absorption. By reducing the absorption of essential micronutrients such as iron, this condition can ultimately lead to iron-deficiency anemia [25]. In cases of prolonged diarrhea, the intestinal epithelium may be damaged, and inflammation can increase hepcidin levels, further restricting iron absorption [30]. Hence, effective treatment and prevention of diarrhea can be considered a crucial protective factor in reducing anemia. Furthermore, toilet facility type has been associated with anemia, particularly in low-income settings where poor sanitation may contribute to the spread of gastrointestinal infections [31]. Moreover, consistent with the findings of the current study, the results of another study confirmed a significant association between the use of treated bed nets and lower anemia prevalence among pregnant and non-pregnant women in Ghana, reinforcing the protective role of malaria prevention in reducing the anemia burden [32].

According to some studies, the effects of cooking fuel may stem from exposure to indoor air pollution or socioeconomic factors associated with fuel choice [33-34]. Biogas, being cleaner and less harmful, can potentially mitigate respiratory and inflammatory conditions that can exacerbate anemia. Additionally, recent research revealed that the use of polluting cooking fuels (e.g., wood and charcoal) was significantly associated with higher odds of childhood anemia across 29 sub-Saharan African countries. Their findings reinforce the importance of transitioning to cleaner fuels (e.g., biogas) to reduce anemia risk by minimizing household air pollution and related respiratory infections.[35]

Improving maternal education has been shown to directly affect childcare practices, dietary diversity, and healthcare utilization.[5,10,24,27] These findings align with those of research by Black et al, highlighting that educated mothers are more likely to adopt better childcare practices, ensure adequate nutrition, and seek timely healthcare, all contributing to reduced anemia risk.[27] Education fosters awareness about the importance of iron-rich foods and healthcare services, which are essential in combating anemia. In addition, increasing uptake of ANC services ensures that pregnant women receive iron-folic acid supplementation, nutritional counselling, and treatment for potential infections, all of which mitigate the risk of anemia in both mothers and their children.[36-37] Furthermore, the results revealed that while breastfeeding plays a protective role in reducing anemia risk, sustained nutritional interventions are necessary for long-term anemia prevention. These findings conform to the results of previous studies, indicating that breastfeeding provides essential micronutrients and immune protection, thereby lowering the risk of infections that contribute to anemia.[38] Further, prior research emphasized that exclusive breastfeeding in the early months of life can help maintain adequate iron stores. However, iron-rich complementary foods should be introduced after six months to sustain optimal hemoglobin levels.[39] Similarly, promoting vitamin A supplementation can strengthen immunity and improve iron metabolism, thereby enhancing hemoglobin levels.[9] Among the variables identified by the LASSO method, many are modifiable and have the potential to significantly influence anemia risk in U5C. These modifiable factors underscore the need for integrated public health and social interventions that address maternal education, environmental health, and nutritional support to effectively combat anemia.

Several studies have directly compared alternative feature selection methods. Hastie et al evaluated best subset selection, forward stepwise, and LASSO and concluded that LASSO provided more efficient and stable feature selection in large datasets.[40] Similarly, Zhou et al compared LASSO and stepwise regression in psychological research, reporting that LASSO achieved higher predictive accuracy and model stability.[41] As in our study, LASSO generally outperformed stepwise and correlation-based methods; however, this issue does not guarantee the best results across all datasets and application domains, as data characteristics should always guide method choice. For example, Haury et al found that simple filtering methods could outperform embedded techniques, such as LASSO, in terms of accuracy and stability.[42]

In contrast, several studies using the NFHS-5 dataset to identify childhood anemia risk factors have not applied or compared feature selection methods, relying instead on direct logistic regression modelling.[11,13-15] In this context, our study addressed this gap by systematically comparing stepwise, forward, backward, correlation-based filtering, and LASSO methods on the NFHS-5 Gujarat dataset, thereby empirically demonstrating the superior feature selection performance of LASSO in childhood anemia data.

Theoretically, LASSO's strength lies in its L1 regularization, which performs coefficient shrinkage and variable selection simultaneously. More precisely, this penalization drives less informative predictors toward zero, thereby reducing model complexity, a common problem in large-scale datasets like NFHS-5.[43] Unlike stepwise or correlation-based methods, which depend on discrete inclusion thresholds and may vary across samples, LASSO's continuous penalization yields more stable, reproducible results.[44] Its ability to balance bias and variance through cross-validated penalty selection improves predictive generalizability. Consequently, LASSO often outperforms traditional selection methods, which produce parsimonious yet robust models that identify the most relevant predictors while minimizing overfitting. This study, thus, provided a methodological benchmark for feature selection in childhood anemia research.

Strengths and Limitations

This study provided a comprehensive assessment of childhood anemia risk factors in Gujarat using NFHS-5 data, with a robust comparison of multiple feature selection techniques. By incorporating clinically relevant diagnostic metrics (sensitivity, specificity, precision, F1-score, and AUC), it presented a more reliable evaluation of model performance beyond accuracy. To the best of our knowledge, this work is among the first to compare several feature selection methods (LASSO, stepwise, backward, forward, and correlation-based) for identifying key predictors of anemia among U5C, providing a useful methodological reference for future studies.

As with most studies, this research had some limitations. The high anemia prevalence (80.54%) created a class imbalance, potentially affecting model interpretability, calibration, and generalizability. Moreover, the reliance on traditional logistic regression may have limited the ability to capture nonlinear or complex predictor interactions. Therefore, future work should apply advanced machine learning and resampling strategies (e.g., synthetic minority over-sampling technique, cost-sensitive learning, and ensemble methods) to improve accuracy and robustness.

Finally, as the model was developed using data from a single state with high anemia prevalence, external validation in other regions is needed to assess generalizability and broader applicability.

Conclusion

This study explored the key risk factors affecting anemia status among U5C in Gujarat using NFHS-5 data. Among the feature selection methods used and compared in this study, LASSO demonstrated superior performance in identifying risk factors for anemia. Among the variables retrieved from the LASSO method, maternal education, hygiene conditions (source of drinking water and type of toilet facility), access to nutrition and health services (cooking fuel, use of bed nets, breastfeeding status, vitamin A supplements, and antenatal visits), and disease management (treatment received for diarrhea) were modifiable. Improvements in these areas may reduce the risk of anemia. In this context, policymakers and health institutions should develop comprehensive strategies to address these factors, improve child health, and prevent anemia.

Acknowledgments
We would like to express our sincere gratitude to Dr. Nermin Ankay, a specialist in the Department of Pediatrics at the Ministry of Health Emergency Hospital in Northern Cyprus, for her valuable insights into this study.

Authors' Contribution
Conceptualization: Nuriye Sancar.
Data curation: Anjana Vasthava.
Formal analysis: Nuriye Sancar, Anjana Vasthava.
Investigation: Nuriye Sancar, Anjana Vasthava.
Methodology: Nuriye Sancar, Anjana Vasthava.
Project administration: Nuriye Sancar, Anjana Vasthava.
Resources: Nuriye Sancar, Anjana Vasthava.
Software: Nuriye Sancar, Anjana Vasthava.
Supervision: Nuriye Sancar.
Validation: Nuriye Sancar, Anjana Vasthava.
Visualization: Nuriye Sancar, Anjana Vasthava.
Writing–original draft: Nuriye Sancar, Anjana Vasthava.
Writing–review & editing: Nuriye Sancar, Anjana Vasthava.

Competing Interests
The authors declare that there is no conflict of interests.

Ethical Approval
This study utilized secondary data from the NFHS-5 for Gujarat, obtained with authorization from the Department of Homeland Security Program (dated October 21, 2024).

Funding
This study received no financial support.

References

1. Sankar D, Oviya IR. Multidisciplinary approaches to study anaemia with special mention on aplastic anaemia (Review). Int J Mol Med. 2024;54(5):95. doi: 10.3892/ijmm.2024.5419
2. Grantham-McGregor S, Ani C. A review of studies on the effect of iron deficiency on cognitive development in children. J Nutr. 2001;131(2):649S-68S. doi: 10.1093/jn/131.2.649S
3. Lozoff B, Corapci F, Burden MJ, Kaciroti N, Angulo-Barroso R, Sazawal S, et al. Preschool-aged children with iron deficiency anemia show altered affect and behavior. J Nutr. 2007;137(3):683-9. doi: 10.1093/jn/137.3.683
4. World Health Organization (WHO). Global Anemia Reduction Efforts: Progress and Challenges. WHO; 2021. Available from: https://www.who.int/health-topics/anaemia#tab=tab_3. Accessed February 13, 2025.
5. Balarajan Y, Ramakrishnan U, Ozaltin E, Shankar AH, Subramanian SV. Anaemia in low-income and middle-income countries. Lancet. 2011;378(9809):2123-35. doi: 10.1016/s0140-6736(10)62304-5
6. International Institute for Population Sciences (IIPS). National Family Health Survey (MCH & Family Planning), India 1992-93: Summary Report. Bombay: IIPS; 1995. Available from: https://dhsprogram.com/pubs/pdf/SR162/SR162.pdf. Accessed February 13, 2025.
7. Ministry of Health & Family Welfare (MoHFW), Government of India. Anemia Mukt Bharat: Intensified National Iron Plus Initiative (I-NIPI): Operational Guidelines for Programme Managers. New Delhi: MoHFW; 2018.
8. Stevens GA, Finucane MM, De-Regil LM, Paciorek CJ, Flaxman SR, Branca F, et al. 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 Glob Health. 2013;1(1):e16-25. doi: 10.1016/s2214-109x(13)70001-9
9. Taneja DK, Rai SK, Yadav K. Evaluation of promotion of iron-rich foods for the prevention of nutritional anemia in India. Indian J Public Health. 2020;64(3):236-41. doi: 10.4103/ijph.IJPH_65_20
10. Rahman MM, Abe SK, Rahman MS, Kanda M, Narita S, Bilano V, et al. Maternal anemia and risk of adverse birth and health outcomes in low- and middle-income countries: systematic review and meta-analysis. Am J Clin Nutr. 2016;103(2):495-504. doi: 10.3945/ajcn.115.107896
11. Pasricha SR, Black J, Muthayya S, Shet A, Bhat V, Nagaraj S, et al. Determinants of anemia among young children in rural India. Pediatrics. 2010;126(1):e140-9. doi: 10.1542/peds.2009-3108
12. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med. 2019;112:103375. doi: 10.1016/j.compbiomed.2019.103375
13. Dutta M, Bhise M, Prashad L, Chaurasia H, Debnath P. Prevalence and risk factors of anemia among children 6-59 months in India: a multilevel analysis. Clin Epidemiol Glob Health. 2020;8(3):868-78. doi: 10.1016/j.cegh.2020.02.015
14. Preethi V, Hemalatha V, Arlappa N, Thomas MB, Jaleel A. Trends and predictors of severe and moderate anaemia among children aged 6-59 months in India: an analysis of three rounds of National Family Health Survey (NFHS) data. BMC Public Health. 2024;24(1):2824. doi: 10.1186/s12889-024-20328-9
15. Islam MA, Afroja S, Khan MS, Alauddin S, Nahar MT, Talukder A. Prevalence and triggering factors of childhood anemia: an application of ordinal logistic regression model. Int J Clin Pract. 2022;2022:2212624. doi: 10.1155/2022/2212624
16. International Institute for Population Sciences (IIPS), ICF. National Family Health Survey (NFHS-5), 2019-21: India. Mumbai: IIPS; 2021. Available from: https://dhsprogram.com/pubs/pdf/FR375/FR375.pdf
17. Vart P, Jaglan A, Shafique K. Caste-based social inequalities and childhood anemia in India: results from the National Family Health Survey (NFHS) 2005-2006. BMC Public Health. 2015;15:537. doi: 10.1186/s12889-015-1881-4
18. Puranik A, Shreya N. Anemia among children under 5 years in India: a district-level spatial analysis using NFHS-5 data. Asia Pac J Public Health. 2022;34(2-3):206-12. doi: 10.1177/10105395211058288
19. Kundu RN, Chatterjee A, Bharati S, Pal M, Hossain MG, Bharati P. Zonal distribution and determinants of stunting, wasting, underweight, and thinness among the Indian under-five children: findings from NFHS-5. In: Pal M, Hossain MG, Mahumud RA, Bharati P, eds. Public Health and Nutrition: The South Asian and Southeast Asian Landscape. Singapore: Springer; 2024. p. 63-84. doi: 10.1007/978-981-97-7890-4_6
20. Gurudatha S, Majhi R. Robust machine learning methods for prediction of childhood anemia–a case of the empowered action group states of India. In: 2024 2nd International Conference on Recent Advances in Information Technology

for Sustainable Development (ICRAIS). Manipal: IEEE; 2024. p. 188-93. doi: 10.1109/icrais62903.2024.10811745

21. Baranwal A, Baranwal A, Roy N. Association of household environment and prevalence of anemia among children under-5 in India. Front Public Health. 2014;2:196. doi: 10.3389/fpubh.2014.00196
22. O'brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant. 2007;41(5):673-90. doi: 10.1007/s11135-006-9018-6
23. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996;58(1):267-88. doi: 10.1111/j.2517-6161.1996.tb02080.x
24. Kunjumon A, Nagarajan S, Thodukayil FS, Panneer S. Factors associated with under-five mortality in scheduled tribes in India: an analysis of National Family Health Survey-5 (2019-2021). Indian J Med Res. 2024;160(1):31-9. doi: 10.25259/ijmr_1869_23
25. World Health Organization (WHO). Global Nutrition Report: Actions and Accountability to Advance Nutrition and Sustainable Development. WHO; 2015. Available from: https://globalnutritionreport.org/reports/2015-global-nutrition-report/.
26. Black RE, Allen LH, Bhutta ZA, Caulfield LE, de Onis M, Ezzati M, et al. Maternal and child undernutrition: global and regional exposures and health consequences. Lancet. 2008;371(9608):243-60. doi: 10.1016/s0140-6736(07)61690-0
27. Black RE, Victora CG, Walker SP, Bhutta ZA, Christian P, de Onis M, et al. Maternal and child undernutrition and overweight in low-income and middle-income countries. Lancet. 2013;382(9890):427-51. doi: 10.1016/s0140-6736(13)60937-x
28. Ziegler EE. Consumption of cow's milk as a cause of iron deficiency in infants and toddlers. Nutr Rev. 2011;69 Suppl 1:S37-42. doi: 10.1111/j.1753-4887.2011.00431.x
29. Weiss G, Goodnough LT. Anemia of chronic disease. N Engl J Med. 2005;352(10):1011-23. doi: 10.1056/NEJMra041809
30. Guerrant RL, Oriá RB, Moore SR, Oriá MO, Lima AA. Malnutrition as an enteric infectious disease with long-term effects on child development. Nutr Rev. 2008;66(9):487-505. doi: 10.1111/j.1753-4887.2008.00082.x
31. Prüss-Ustün A, Bartram J, Clasen T, Colford JM, Jr., Cumming O, Curtis V, et al. Burden of disease from inadequate water, sanitation and hygiene in low- and middle-income settings: a retrospective analysis of data from 145 countries. Trop Med Int Health. 2014;19(8):894-905. doi: 10.1111/tmi.12329
32. Ansah RK, Tackie S, Twum RA, Tawiah K, Boadi RK, Addo DA, et al. The relationship between anaemia and the use of treated bed nets among pregnant and non-pregnant women in Ghana. PLoS One. 2024;19(5):e0300431. doi: 10.1371/journal.pone.0300431
33. He Y, Liu X, Zheng Y, Zhai Z, Wu X, Kang N, et al. Lower socioeconomic status strengthens the effect of cooking fuel use on anemia risk and anemia-related parameters: findings from the Henan Rural Cohort. Sci Total Environ. 2022;831:154958. doi: 10.1016/j.scitotenv.2022.154958
34. Balakrishnan K, Cohen A, Smith KR. Addressing the burden of disease attributable to air pollution in India: the need to integrate across household and ambient air pollution exposures. Environ Health Perspect. 2014;122(1):A6-7. doi: 10.1289/ehp.1307822
35. Amadu I, Seidu AA, Afitiri AR, Ahinkorah BO, Yaya S. Household cooking fuel type and childhood anaemia in sub-Saharan Africa: analysis of cross-sectional surveys of 123, 186 children from 29 countries. BMJ Open. 2021;11(7):e048724. doi: 10.1136/bmjopen-2021-048724
36. World Health Organization (WHO). The Global Prevalence of Anaemia in 2019. 2021. Available from: https://www.who.int/data/gho/data/themes/topics/anaemia_in_women_and_children#:~:text=In%202019%2C%20global%20anaemia%20prevalence%20was%2029.9%25%20%2895%25,over%20half%20a%20billion%20women%20aged%2015-49%20years.
37. Kapil U, Kapil R, Gupta A. Prevention and control of anemia amongst children and adolescents: theory and practice in India. Indian J Pediatr. 2019;86(6):523-31. doi: 10.1007/s12098-019-02932-5
38. Meinzen-Derr JK, Guerrero ML, Altaye M, Ortega-Gallegos H, Ruiz-Palacios GM, Morrow AL. Risk of infant anemia is associated with exclusive breast-feeding and maternal anemia in a Mexican cohort. J Nutr. 2006;136(2):452-8. doi: 10.1093/jn/136.2.452
39. Edmond KM, Zandoh C, Quigley MA, Amenga-Etego S, Owusu-Agyei S, Kirkwood BR. Delayed breastfeeding initiation increases risk of neonatal mortality. Pediatrics. 2006;117(3):e380-6. doi: 10.1542/peds.2005-1496
40. Hastie T, Tibshirani R, Tibshirani RJ. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. ArXiv [Preprint]. July 27, 2017. Available from: https://arxiv.org/abs/1707.08692.
41. Zhou DJ, Chahal R, Gotlib IH, Liu S. Comparison of lasso and stepwise regression in psychological data. Methodology (Gott). 2024;20(2):121-43. doi: 10.5964/meth.11523
42. Haury AC, Gestraud P, Vert JP. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS One. 2011;6(12):e28210. doi: 10.1371/journal.pone.0028210
43. Friedrich S, Groll A, Ickstadt K, Kneib T, Pauly M, Rahnenführer J, et al. Regularization approaches in clinical biostatistics: a review of methods and their applications. Stat Methods Med Res. 2023;32(2):425-40. doi: 10.1177/09622802221133557
44. Sztepanacz JL, Houle D. Regularized regression can improve estimates of multivariate selection in the face of multicollinearity and limited data. Evol Lett. 2024;8(3):361-73. doi: 10.1093/evlett/qrad064