

# Predicting Anemia Risk Using Lifestyle and Demographic Data Using Machine Learning Models

Mahir Ahmed<sup>1</sup>, Tasnim Jahan Rimvy<sup>2</sup>, Badhon Dalbot<sup>3</sup>, Sayed Hasan Sami<sup>4</sup>, and Sheikh Sadi<sup>5</sup>

<sup>1</sup>Department of Computer Science Engineering, United International University

<sup>2</sup>Department of Computer Science Engineering, United International University

<sup>3</sup>Department of Computer Science Engineering, United International University

<sup>4</sup>Department of Computer Science Engineering, United International University

<sup>5</sup>Department of Computer Science Engineering, United International University

---

## Abstract

Anemia remains a pervasive global public health concern, profoundly impacting socio-economic development, particularly in nations like Bangladesh, where prevalence rates are critically high across vulnerable groups (e.g., up to 64% in young children). Traditional diagnosis relies on invasive, costly, and often inaccessible laboratory testing, creating a significant diagnostic and intervention gap in resource-constrained community settings. While machine learning (ML) has shown promise, existing models are often limited by outdated data, a focus solely on children, or a reliance on clinical biomarkers, failing to incorporate easily obtainable lifestyle and behavioral risk factors. This study addresses these limitations by developing a high-accuracy, non-invasive anemia risk prediction model for the general population using readily available demographic, socio-economic, and lifestyle data from contemporary national health surveys. We employ rigorous feature selection techniques (including the Chi-square test and Recursive Feature Elimination) to identify the most potent risk factors. We then develop and evaluate advanced ensemble ML models, including Random Forest, Gradient Boosting, and XGBoost, to predict an individual's anemia status. This approach aims to demonstrate the efficacy of tree-based ensembles in capturing the complex, non-linear relationships within survey data. The resulting model is intended to serve as a robust, low-cost digital screening tool, maximizing the utility of non-invasive data for early risk detection and facilitating targeted public health interventions.

---

## 1. Introduction

Anemia, a global public health crisis, is fundamentally characterized by a deficiency in the number of red blood cells or a reduced concentration of hemoglobin, consequently impairing the blood's capacity to transport oxygen to the body's tissues. Affecting an estimated 1.8 billion people worldwide, anemia is not merely a health condition but a disorder with severe socio-economic consequences. Its impact ranges from fatigue and impaired cognitive and physical development in children to reduced work productivity and significantly increased risks of maternal and perinatal mortality in adults [?]. The condition arises from a complex interplay of factors, including nutritional deficiencies (iron, folate, vitamin B<sub>12</sub>), chronic diseases, and recurrent infections.

The challenge of anemia is particularly acute in developing nations, and Bangladesh exemplifies this crisis. According to recent data from the World Health Organization (WHO) and the Bangladesh Demographic and Health

Survey (BDHS), the country faces persistently high prevalence rates across all vulnerable groups. For instance, anemia affects an alarming 64% of children aged 6–23 months, 42% of children aged 24–59 months, and between 46% and 50% of pregnant women [?]. Furthermore, 33% to 41% of non-pregnant women are also anemic, with overall rates among children and adolescents hovering around 46.8%. While severe anemia is comparatively rare (affecting only 2–3%), the high prevalence of mild-to-moderate cases still places a profound burden on the national healthcare system and economy.

Despite this magnitude, a significant diagnostic gap persists. Traditional methods of diagnosis, relying on invasive blood sampling and laboratory analysis (e.g., Complete Blood Count), are often expensive, time-consuming, and inaccessible in rural and low-resource community settings. Consequently, a large proportion of at-risk individuals, particularly those outside major urban centers, remain undiagnosed until the condition becomes severe.

This need for accessible, low-cost screening has moti-

vated previous research to employ machine learning (ML) techniques. Existing studies have generally fallen into two categories: those achieving high accuracy using invasive clinical or hematological data (which require laboratory infrastructure) and those utilizing socio-demographic survey data (which are accessible but can be outdated, lack model interpretability, or omit critical behavioral and lifestyle indicators). The present study is designed to bridge this gap.

This study is designed to address the aforementioned gaps through a two-pronged approach. First, we aim to systematically identify the most significant predictors associated with anemia risk across the general population by employing advanced feature selection techniques on a comprehensive set of demographic, socio-economic, and critical lifestyle factors. Second, we will develop and rigorously evaluate advanced machine learning models, including Random Forest, Gradient Boosting, and XGBoost, to predict an individual's anemia status (anemic vs. non-anemic). This commitment to developing a high-performing and accurate model will maximize its utility for early risk detection and informed intervention design in community health programs, providing a low-cost, non-invasive screening solution.

## 2. Related Works

Anemia prediction and diagnosis have been widely explored using data-driven and machine learning (ML) techniques in recent years. The increasing availability of health-related datasets, coupled with the limitations of conventional diagnostic methods, has motivated researchers to develop automated systems for early detection and risk forecasting. A review of the existing literature reveals a growing trend toward integrating demographic, nutritional, and biochemical factors to model anemia risk with high predictive accuracy.

### 2.1 Evaluating the Factors and Forecasting Childhood Anemia Through Machine Learning Algorithms

Salma and Majahar Ali [?] conducted a comprehensive machine learning-based study to predict childhood anemia in Bangladesh using nationally representative data from the Bangladesh Demographic and Health Survey (BDHS 2011). The study focused on children aged 6–59 months and aimed to identify socio-demographic, nutritional, and household-level factors associated with anemia prevalence in a low-resource setting. By relying on survey-based variables rather than laboratory measurements, the study emphasized the feasibility of non-invasive, population-level anemia screening.

The most significant predictors included parental education, child age, breastfeeding status, maternal age, maternal education, sanitation facilities, drinking water source,

and the number of children under five years of age in the household. These features reflect socio-economic status, caregiving practices, and environmental conditions known to influence nutritional health.

The study evaluated seven machine learning models: K-Nearest Neighbor (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Bagging, Gradient Boosting, and XGBoost. Model performance was assessed using ten-fold cross-validation and multiple evaluation metrics, including accuracy, sensitivity, specificity, precision, F1-score, Cohen's Kappa, and area under the ROC curve (AUC). Among the tested algorithms, Gradient Boosting achieved the best performance, reporting an accuracy of 87.46%, precision of 95.35%, specificity of 96.56%, sensitivity of 85.31%, and an AUC of 0.9099. The results demonstrated the effectiveness of ensemble-based models in capturing complex, non-linear relationships among socio-demographic risk factors.

Despite its strong methodological framework, the study presents several limitations. The analysis was restricted to children, excluding adolescents and adults who also face substantial anemia risk. The feature set primarily consisted of demographic and socio-economic variables, with limited inclusion of lifestyle and behavioral factors such as dietary diversity, physical activity, and daily habits. Furthermore, the study did not explicitly address class imbalance handling or model interpretability, which are critical considerations for deploying machine learning models in public health decision-making. These limitations highlight opportunities for extending anemia prediction research to broader populations using updated datasets and richer lifestyle-oriented features.

### 2.2 Detection of Iron Deficiency Anemia Using Palmar Images and Machine Learning

In a different methodological direction, Khawaga et al. [?] explored the feasibility of detecting iron deficiency anemia using medical images and machine learning techniques. The study utilized palmar images collected from hospitals in Ghana, focusing on children under five years of age. A total of 527 palm images were augmented to 2,635 samples to address data scarcity and improve model robustness. The core hypothesis of the study was that palmar pallor, a common clinical sign of anemia, could be quantitatively analyzed using computer vision and color-based features.

The authors extracted features from the region of interest (ROI) of palm images using the CIELAB color space, specifically the L\*, a\*, and b\* channels, which effectively capture variations in skin color and brightness. These image-based features were combined with basic demographic attributes such as age and gender, as well as hemoglobin (Hb) levels for validation. Several machine learning models were evaluated, including Convolutional Neural Networks (CNN), k-Nearest Neighbors (k-NN),

Table 1: Comparison of Related Studies on Anemia Prediction Using Machine Learning

Year	Authors	Target Group	Dataset / Country	Key Limitations
2023	Khawaga et al.	Children under five	Hospital palm images, Ghana	Small sample size, single-country data, image-dependent approach
2025	Salma & Majahar Ali	Children (6–59 months)	BDHS 2011, Bangladesh	Children-only focus, lack of lifestyle variables, less accuracy
2025	Moyo et al.	Women (15–49 years)	ZDHS 2015, Zimbabwe	Cross-sectional design, missing dietary and infection data
2025	Asare et al.	Children (6–59 months)	GDHS 2022, Ghana	Limited ML models, self-reported data, child-only scope
2025	Mwakyusa et al.	Children (6–59 months)	TDHS 2022, Tanzania	Model complexity, context-specific results, no biomarkers

Naïve Bayes, Support Vector Machines (SVM), and Decision Trees.

Among the tested models, Naïve Bayes achieved the highest classification accuracy of 99.96%, outperforming more complex deep learning approaches. The results demonstrated that palmar image analysis could serve as a highly accurate, non-invasive method for anemia detection in controlled clinical environments. The study highlighted the potential of low-cost imaging tools for rapid screening, especially in pediatric populations.

However, the study has notable limitations. The dataset was relatively small and geographically confined to a single country, limiting generalizability. The target population was restricted to young children, excluding adolescents and adults. Furthermore, the reliance on image-based clinical indicators may reduce applicability in non-clinical or community survey settings where controlled image acquisition is not feasible. These constraints limit the scalability of the approach for large-scale population screening.

### 2.3 Socioeconomic and Demographic Factors Associated with Anemia Among Women of Reproductive Age

Moyo et al. [?] investigated the socioeconomic and demographic determinants of anemia among women of reproductive age (15–49 years) in Zimbabwe using supervised machine learning techniques. The study analyzed data from the 2015 Zimbabwe Demographic and Health Survey (ZDHS), focusing on identifying key predictors of anemia using both traditional statistical methods and machine learning models.

The authors employed chi-square tests and multivariate logistic regression for initial association analysis, followed by Elastic Net regularization to identify important predic-

tors. Machine learning models including Random Forest, k-Nearest Neighbors, and Decision Trees were then trained to predict anemia status. To address class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, improving model sensitivity toward minority classes.

The reported anemia prevalence among women was 24.1%. Random Forest achieved the best predictive performance with an accuracy of approximately 74%. Significant predictors included age, marital status, education level, household wealth, occupation, modern contraceptive use, BMI, and province of residence. The study demonstrated that combining socio-demographic variables with ensemble learning methods can effectively model anemia risk in adult populations.

Despite these contributions, the study relied on cross-sectional data, limiting causal inference. Additionally, important lifestyle and dietary factors such as iron intake, physical activity, and infection history were not included. The dataset was also relatively dated, which may not reflect current socio-economic conditions. These limitations suggest the need for more comprehensive and updated datasets incorporating lifestyle behaviors.

### 2.4 Early Childhood Anemia in Ghana Using Machine Learning Techniques

Asare et al. [?] examined early childhood anemia in Ghana using data from the Ghana Demographic and Health Survey (GDHS 2022). The study targeted children aged 6–59 months and aimed to identify key predictors of anemia while comparing the performance of multiple machine learning models.

The feature set encompassed societal factors (region, residence, wealth index), parental characteristics (education, occupation, maternal iron intake), and child-specific

variables (age, sex, birth order, stunting status, dietary intake, and postnatal checkups). Logistic Regression, Decision Tree, k-Nearest Neighbors, and Random Forest models were evaluated.

The study reported an anemia prevalence of approximately 49% among Ghanaian children. Random Forest emerged as the best-performing model, achieving an accuracy of 94.7%. The findings highlighted the importance of parental education, socio-economic status, maternal nutrition, and postnatal healthcare in determining childhood anemia risk.

However, the study was limited by its cross-sectional design and reliance on self-reported survey data, which may introduce recall bias. Additionally, only a limited set of machine learning models was explored, and the analysis remained restricted to children, limiting broader population applicability.

## 2.5 Hybrid Machine Learning Model for the Prediction of Anemia

Mwakyusa et al. [?] proposed a hybrid machine learning framework for predicting childhood anemia using data from the Tanzania Demographic and Health Survey (TDHS 2022). The study focused on children aged 6–59 months and incorporated socio-demographic, maternal and child health, household, and environmental variables.

The authors developed a stacked ensemble model, combining Random Forest and Artificial Neural Networks as base learners, with XGBoost serving as the meta-learner. To optimize classification performance, the decision threshold was tuned using Youden’s J-index. The hybrid model achieved an accuracy of approximately 87%, with sensitivity and specificity values of 0.861 and 0.880, respectively, at an optimized threshold of 0.40. The ensemble approach consistently outperformed individual classifiers.

Despite its strong predictive performance, the proposed model introduced increased computational complexity, which may limit its deployment in routine public health systems. Furthermore, the results were context-specific to Tanzania and relied exclusively on survey-based variables without incorporating laboratory biomarkers. These factors may constrain the generalizability and interpretability of the model in broader settings.

## 3. Data and Methodology

### 3.1 Dataset Description

The study utilizes secondary data extracted from the Demographic and Health Survey (DHS), a nationally representative, population-based survey designed to capture demographic, socio-economic, lifestyle, environmental,

and health-related information. For this research, a curated subset of approximately 50 variables was selected from over 1,800 available features to focus on factors most relevant to predicting anemia risk in the general population. The selected variables represent four major domains: **demographic characteristics, socio-economic status, lifestyle and behavioral factors, and environmental and health access indicators**.

**Demographic variables** include age, sex, marital status, place of residence (urban or rural), region, household size, and relationship to the household head. These features capture population structure and living arrangements, which are known to influence nutritional status and health outcomes.

**Socio-economic indicators** such as education level, years of schooling, household wealth index, electricity access, housing materials, ownership of assets (radio, television, mobile phone), and banking access reflect economic capacity and living standards, which indirectly affect dietary quality, healthcare access, and disease vulnerability.

**Lifestyle and behavioral characteristics** were incorporated to account for daily practices that may contribute to anemia risk. These include smoking behavior, household exposure to tobacco smoke, type of cooking fuel and cooking environment, mosquito net usage, household pesticide spraying, and salt iodization. These variables capture exposure to indoor air pollution, parasitic infection risk, and micronutrient availability, all of which are linked to hemoglobin levels and anemia prevalence.

**Environmental and health-related variables** include body mass index (BMI), pregnancy status, drinking water source, toilet facility type, water treatment practices, handwashing facilities, and availability of soap. These indicators reflect nutritional status, sanitation, and hygiene conditions, which play a critical role in iron absorption, infection prevention, and overall health. Hemoglobin concentration and anemia status are included as outcome-related variables, with anemia categorized according to established clinical thresholds.

Overall, the dataset provides a comprehensive representation of individual- and household-level conditions influencing anemia risk. By integrating demographic, lifestyle, and environmental factors, the dataset supports the development of machine learning models capable of identifying complex, non-linear relationships associated with anemia in the general population, enabling early risk detection and informed public health interventions.