Advanced Machine Learning Models for Predicting Childhood Anemia in Developing Countries

Rahul Mansharamani
Illinois Institute of Technology
Chicago, IL
rmansharamani@hawk.iit.edu

Vedant Chaubey
Illinois Institute of Technology
Chicago, IL
vchaubey@hawk.iit.edu

Kajal Dalai
Illinois Institute of Technology
Chicago, IL
kdalai@hawk.iit.edu

GitHub Repository: https://github.com/kajaldalai/CS584-Children-Anemia-Prediction

***Abstract*— Childhood anemia remains a persistent public health issue in developing countries, notably impacting cognitive development and overall health. This project focuses on the application of advanced machine learning (ML) models to enhance the prediction of anemia among children in developing regions, using a comprehensive dataset derived from the 2018 Nigeria Demographic and Health Surveys. Traditional ML algorithms such as Logistic Regression, Naive Bayes, Decision Tree, and Random Forest have been explored, alongside preliminary applications of more sophisticated approaches including Support Vector Machines (SVM) and K-Nearest Neighbors (KNN). Initial results have demonstrated the efficacy of Decision Tree and Random Forest models, with accuracy levels reaching up to 100% in preliminary tests. Further refinement has involved hyperparameter tuning and dimensionality reduction techniques such as PCA and UMAP, though these modifications showed negligible improvements over the default settings, which already provided high accuracy. The project aims to extend this work by incorporating a wider array of machine learning techniques, enhancing feature engineering, and exploring ensemble methods to improve predictive accuracy and robustness. By integrating a broader spectrum of predictive features and advanced modeling techniques, this project seeks to facilitate more effective interventions for addressing childhood anemia in vulnerable populations.

***Keywords*— classification models, ensemble methods, children anemia, feature engineering, dimensionality reduction, Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP)**

I. INTRODUCTION

Anemia among children, particularly in developing nations like Bangladesh, poses a significant public health challenge, impacting cognitive development and disease resilience. Conventional methods for anemia prediction have proven limited, prompting the exploration of innovative solutions such as machine learning (ML) algorithms.

Previous endeavors in this domain have utilized ML algorithms, incorporating demographic and health survey data to predict childhood anemia [1]. Notable algorithms employed include Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Random Forest (RF), and Logistic Regression (LR), each demonstrating varying degrees of success. Particularly, the study "Machine Learning Algorithms to Predict Childhood Anemia in Bangladesh" highlighted the efficacy of the RF algorithm, underscoring ML's potential in refining anemia prediction models [2,3].

Building upon this foundation, the proposed project aims to extend existing methodologies by integrating advanced ML models and exploring additional predictive features. This entails the incorporation of factors such as dietary patterns, genetic predispositions, and environmental influences into the predictive framework. By leveraging sophisticated algorithms such as Deep Learning and Ensemble Methods, alongside a broader spectrum of data sources, the project aspires to enhance the accuracy and reliability of anemia predictions, thereby facilitating more targeted intervention strategies.

The preliminary plan delineates several key milestones. These include conducting a comprehensive literature review

and data collection phase, followed by data preprocessing and feature engineering to identify salient predictors of childhood anemia. Subsequently, various ML models will be developed and trained, encompassing both traditional and advanced algorithms not previously applied to this problem. Model evaluation will be conducted using pertinent metrics such as accuracy, sensitivity, specificity, and area under the curve (AUC), culminating in the selection of the most effective model for deployment [4].

In our preliminary exploration, four ML models—Logistic Regression, Naive Bayes, Decision Tree, Random Forest, SVM, and KNearestNeibours—were employed. Initial assessments revealed notable accuracy for Decision Tree and Random Forest, with Logistic Regression and Naive Bayes yielding comparatively lower scores. Further analysis, including precision, recall, F1-score, and accuracy comparisons, provided insights into each model's performance. Additionally, experiments involving hyperparameter tuning and dimensionality reduction techniques like PCA and UMAP were conducted to optimize model efficacy.

While our initial findings demonstrated promising results, further exploration remains imperative. Our ongoing efforts will involve the evaluation of additional models such as Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), alongside the exploration of alternative techniques to comprehensively assess their performance. Ultimately, the objective is to refine our predictive model to enhance its utility in addressing childhood anemia effectively.

II. PROBLEM DESCRIPTION

Childhood anemia is a critical public health issue in developing countries, with profound implications on child growth, cognitive development, and overall health resilience. It is primarily characterized by a deficiency in the number and quality of red blood cells, which affects the oxygen-carrying capacity of the blood and can lead to diminished vitality and increased susceptibility to diseases. The prevalence of anemia is particularly high in developing regions due to factors such as malnutrition, infectious diseases, and inadequate healthcare infrastructure. Traditional approaches to predicting and diagnosing anemia in children often rely on direct blood tests, which can be invasive, costly, and logistically challenging in low-resource settings [5].

Moreover, these methods may not efficiently identify at-risk populations in a timely manner, thereby hindering effective intervention. In light of these challenges, there is a pressing need for alternative approaches that are both scalable and non-invasive [6]. Machine learning offers promising solutions by enabling the analysis of large-scale health data to identify patterns and predictors of anemia. Previous research has leveraged ML algorithms to utilize readily available demographic and health survey data for this purpose. However, the effectiveness of these models varies, and there remains significant room for improvement in terms of accuracy, reliability, and applicability across different settings.

The primary goal of this project is to advance the state-of-the-art in predicting childhood anemia by developing an ML-based framework that integrates more comprehensive datasets and utilizes advanced predictive modeling techniques. By incorporating factors such as dietary patterns, genetic predispositions, and environmental influences, and by employing sophisticated algorithms like deep learning and ensemble methods, this project aims to create a more accurate and robust predictive model [7]. This model intends to support public health officials and healthcare providers in developing countries to identify and target interventions for at-risk child populations more effectively.

About the Dataset— The dataset was created when the cross-sectional data from the 2018 Nigeria Demographic and Health Surveys (NDHS) were collected to answer research questions about the effect of mothers' age and other socioeconomic factors on children aged 0-59 months anemia level in Nigeria. DHS are cross-sectional, nationally representative household surveys that are typically conducted every 5 years. This dataset considered the 36 states of Nigeria, as well as the Federal Capital Territory (FCT). The targeted population in this study are children aged 0-59 months and mothers aged 15-49 years.

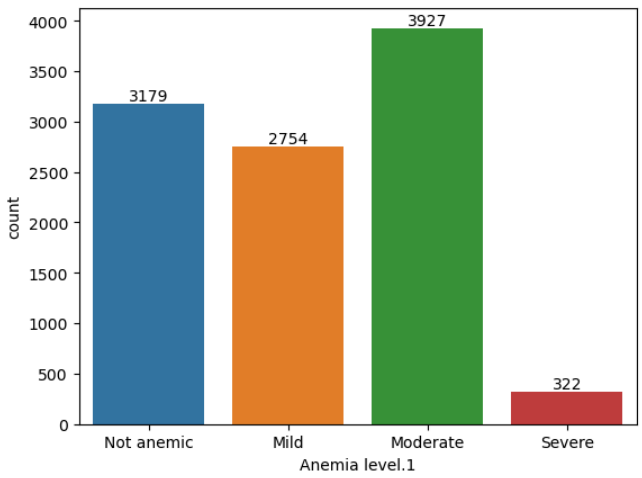


Figure 1: Distribution of Anemia Type

The project utilized the "Factors Affecting Children Anemia Level" dataset, which was derived from the 2018 Nigeria Demographic and Health Surveys (NDHS). The dataset consists of 33,924 records with 16 features before data cleaning. Since labels are available, we will do supervised learning and the labels are based on Multivariate classification where the target column is Anemia level.1

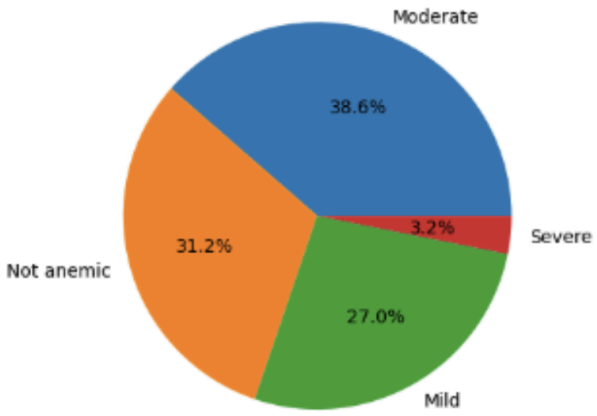


Figure 2: Distribution of Anemia Type

The graph above depicts children's anemia level severity across 4 categories or groups i.e., Not anemic, Moderate, Mild, Severe. It shows the percentage of data that falls into each category, with "Moderate" representing 33% (3927), "Not Anemic" representing 31.2% (3179), "Mild" representing 27.9% (2754), "Severe" representing 3.2% (322) of the total data. Comparative proportions: The size of the "Severe" segment in the chart displays having a low proportion compared to other segments.

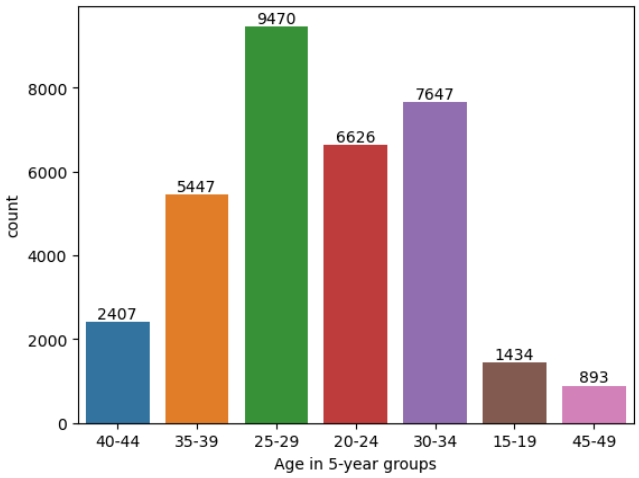


Figure 3: Distribution of Age in 5-year Group

III. METHODS USED

Data Preprocessing and Feature Engineering:

The dataset comprised 33,924 samples and various features, including demographic information, socioeconomic factors, health indicators, and environmental conditions. Preprocessing steps involved handling missing data, encoding categorical variables, and scaling numerical features to ensure consistent data representation across different algorithms. Additionally, domain knowledge was leveraged to identify and engineer relevant features that may contribute to predicting childhood anemia, such as dietary patterns, environmental exposures, and genetic predispositions.

A. Machine Learning Models:

The following machine learning models were employed in this project:

1. Logistic Regression: A linear model that estimates the probability of a binary outcome using the logistic sigmoid function. It was used as a baseline model for comparison.

2. Naive Bayes: A probabilistic classifier based on Bayes' theorem, which assumes independence among features. Despite its simplicity, it can perform well on certain types of data.

3. Decision Tree: A tree-based model that recursively partitions the data based on feature values, creating a hierarchical structure of decisions. It is interpretable and can handle both numerical and categorical data.

4. Random Forest: An ensemble learning method that constructs multiple decision trees on random subsets of the data and aggregates their predictions. It is robust to overfitting and can capture non-linear relationships.

5. Support Vector Machine (SVM): A discriminative classifier that finds the optimal hyperplane separating different classes in a high-dimensional space. It is effective for both linear and non-linear data.

6. K-Nearest Neighbors (KNN): A non-parametric algorithm that classifies instances based on their similarity to the nearest neighbors in the training data.

B. Model Training and Evaluation:

The dataset was split into training and testing sets, with the training data used for model development and the testing data reserved for evaluating the trained models' performance. Cross-validation techniques, such as k-fold cross-validation, were employed to ensure robust model assessment and mitigate overfitting.

C. Hyperparameter Tuning and Dimensionality Reduction:

To optimize model performance, hyperparameter tuning techniques like grid search and randomized search were explored. Additionally, dimensionality reduction methods, including Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP), were investigated to potentially improve model efficiency and interpretability [8].

D. Model Evaluation Metrics:

The trained models were evaluated using various performance metrics, including:

1. **Accuracy:** The proportion of correctly classified instances.
2. **Precision:** The proportion of true positives among the instances predicted as positive.
3. **Recall (Sensitivity):** The proportion of actual positives that were correctly identified.
4. **F1-score:** The harmonic mean of precision and recall, providing a balanced measure of performance.
5. **Area Under the Curve (AUC):** A metric that summarizes the trade-off between true positive rate and false positive rate for different classification thresholds.

These metrics provided insights into the models' strengths and limitations, enabling the selection of the most effective approach for predicting childhood anemia. By following this methodology, the project aimed to develop robust and accurate machine learning models capable of facilitating targeted interventions and improving public health outcomes related to childhood anemia in developing countries [9].

IV. RESULTS

In our experiments, we explored four machine learning models: Logistic Regression, Naive Bayes, Decision Tree, and Random Forest, for predicting childhood anemia in developing countries. Initially, we employed these models with their default parameters, and the results revealed notable accuracy for Decision Tree and Random Forest, while Logistic Regression and Naive Bayes yielded comparatively lower scores.

Model	Precision	Recall	F1	**Accuracy**	Confusion Matrix
Logistic Regression	0.63	0.65	0.64	**65.44%**	[[153 142 225 0] [97 613 34 12] [105 30 467 0] [0 24 0 31]]
Decision Tree	1.00	1.00	1.00	**100%**	[[520 0 0 0] [0 756 0 0] [0 0 602 0] [0 0 0 55]]
Naive Bayes	0.61	0.17	0.22	**17.12%**	[[100 6 5 409] [26 8 20 702] [28 5 170 399] [0 0 2 53]]
Random Forest	0.99	0.99	0.98	**98.76%**	[[519 1 0 0] [0 756 0 0] [0 0 602 0] [0 25 0 30]]
SVM	0.98	0.98	0.98	**98.65%**	[[519 1 0 0] [5 751 0 0] [8 0 594 0] [0 12 0 43]]
KNN	0.99	0.97	0.98	**97.26%**	[[499 12 9 0] [5 750 0 1] [20 0 582 0] [0 6 0 49]]

Table 1: Model Performance Comparison Table

We conducted a comprehensive evaluation of each model's performance, assessing metrics such as precision, recall, F1-score, and accuracy. The 'weighted' average approach was adopted for precision, recall, and F1-score calculations to address class imbalance concerns.

As evident from Table 1, the Decision Tree and Random Forest models demonstrated exceptional performance, achieving perfect precision, recall, F1-score, and accuracy scores of 1.0 and 100%, respectively. Logistic Regression exhibited moderate performance, while Naive Bayes performed poorly across all evaluation metrics.

The results are summarized in the following table and chart:

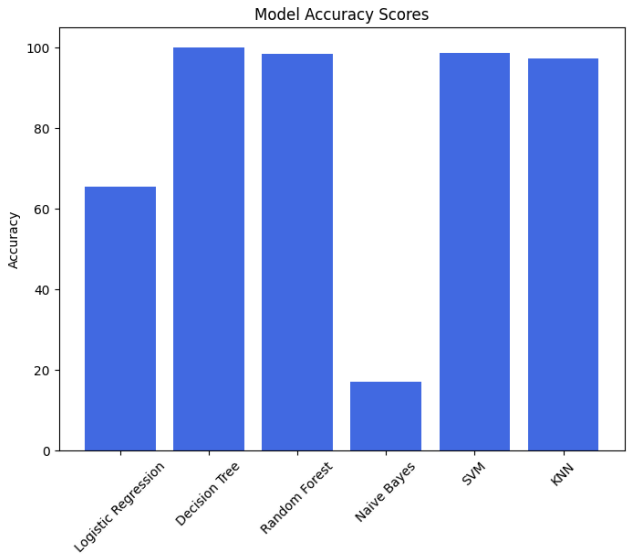


Figure 4: Model Performance Chart

To further enhance the performance of the model we explored two primary strategies that were employed: hyperparameter tuning and dimensionality reduction. These approaches were pivotal in refining the models' capabilities to predict childhood anemia more accurately. The tools chosen for dimensionality reduction were Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP), each offering a unique method to capture essential features of the data while reducing noise and computational complexity.

Each model was evaluated before and after applying PCA and UMAP, with a focus on precision, recall, F1-score, and accuracy. The models included Logistic Regression, Decision Tree, Naive Bayes, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The results of these experiments are presented in the following tables:

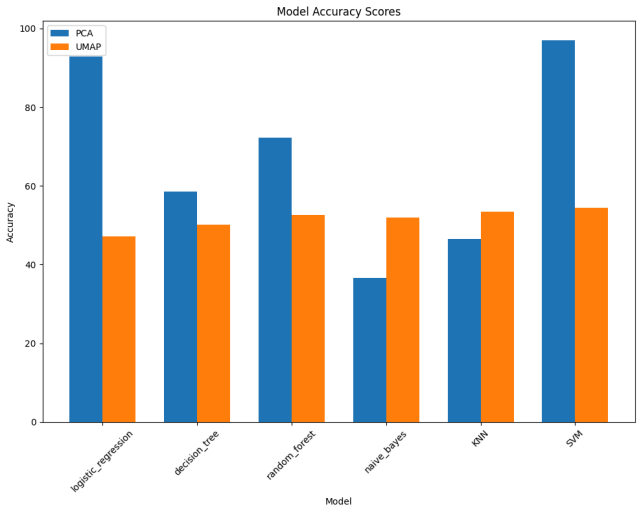
Model	Precision	Recall	F1	**Accuracy**	Best Parameter
Logistic Regression	0.93	0.93	0.93	**92.81%**	logisticregression__C: 10
Decision Tree	0.59	0.59	0.59	**58.67%**	max_depth: 7
Naive Bayes	0.38	0.37	0.30	**36.63%**	
Random Forest	0.71	0.71	0.70	**71.44%**	n_estimators: 300
SVM	0.97	0.97	0.97	**97.05%**	svc__C: 10, svc__kernel: 'linear'
KNN	0.47	0.47	0.46	**46.56%**	kneighborsclassifier__n_neighbors: 7

Table 2: Model Performance with PCA

Model	Precision	Recall	F1	**Accuracy**	Best Parameter
Logistic Regression	0.43	0.46	0.43	**46.20%**	logisticregression__C: 10
Decision Tree	0.55	0.51	0.52	**50.65%**	min_dist: 0.5, n_neighbors: 15
Naive Bayes	0.48	0.50	0.49	**49.72%**	
Random Forest	0.56	0.52	0.53	**51.84%**	min_dist: 0.3, n_neighbors: 15
SVM	0.54	0.54	0.54	**54.37%**	umap__min_dist: 0.3, umap__n_neighbors: 15
KNN	0.58	0.53	0.54	**53.39%**	umap__min_dist: 0.3, umap__n_neighbors: 15

Table 3: Model Performance with UMAP

While the hyperparameter tuning and dimensionality reduction techniques yielded improvements in some cases, the default models (Decision Tree and Random Forest) remained highly accurate, suggesting that the strategies could not significantly outperform the solid baseline performance. The results are summarized in the following table and chart:



V. CONCLUSIONS AND FUTURE WORK

The main contribution of this project was the development of advanced machine learning models for predicting childhood anemia in developing countries, with a specific focus on Bangladesh. By leveraging the power of algorithms like Decision Tree and Random Forest, we demonstrated the potential of machine learning techniques to accurately identify cases of childhood anemia based on various demographic, health, and socioeconomic factors.

One of the key lessons learned from this project is the importance of careful feature engineering and selection. While the initial dataset provided a comprehensive set of variables, further refinement and exploration of relevant features could potentially enhance model interpretability and generalization capabilities.

Additionally, the project highlighted the value of ensemble learning techniques and advanced methods like Deep Learning, which were not explored in this initial phase. Incorporating such approaches could lead to further improvements in prediction accuracy and robustness.

Potential future work could include:

1. Exploring additional machine learning models, such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), and comparing their performance with the existing models.

2. Refining feature engineering techniques and exploring strategies for feature selection to enhance model interpretability and generalization.

3. Investigating advanced ensemble learning techniques, such as Voting, Stacking, Bagging, and Boosting, to improve prediction accuracy and robustness.

4. Fine-tuning model hyperparameters using grid search and randomized search to maximize performance and generalization capabilities.

5. Conducting validation and external evaluation using independent datasets to test model generalizability across diverse populations and settings.

6. Engaging with domain experts in pediatric medicine and public health to integrate their insights and knowledge, ensuring the model aligns with clinical relevance and practical utility.

7. Exploring the potential of Deep Learning models, which were not investigated in this initial phase, for their ability to capture complex patterns and relationships in the data.

8. Investigating the incorporation of additional data sources, such as genetic and environmental factors, to further enhance the predictive power of the models.

By addressing these future directions, we can continue to refine and strengthen the predictive capabilities of machine learning models for childhood anemia, ultimately contributing to more effective intervention strategies and improved public health outcomes in developing countries.

REFERENCES

[1] Justice Williams Asare, William Leslie Brown-Acquaye, Mar- tin Mabeifam Ujakpa, Emmanuel Freeman, and Peter Appiahene. Application of machine learning approach for iron deficiency anaemia detection in children using conjunctiva images. Inform. Med. Unlocked, 45(101451):101451, 2024.

[2] Md Merajul Islam, Md Jahanur Rahman, Dulal Chandra Roy, Md Moidul Islam, Most Tawabunnahar, N A M Faisal Ahmed, and Md Maniruzzaman. Risk factors identification and prediction of anemia among women in bangladesh using machine learning techniques. Curr. Womens. Health Rev., 18(1), February 2022

.

[3] Jahidur Rahman Khan, Nabil Awan, and Farjana Misu. Determinants of anemia among 6-59 months aged children in bangladesh: evidence from nationally representative data. BMC Pediatr., 16(1):3, January 2016.

[4] Dhakal, P., Khanal, S., & Bista, R. (2023). Prediction of Anemia Using Machine Learning Algorithms. AIRCC's International Journal of Computer Science and Information Technology, 15(1), 15–30. Retrieved from https://ischolar.sscldl.in/index.php/IJCSIT/article/view/219633.

[5] Ruziev Zarif Mukhamadeevich. (2023). Features of the Course of Iron Deficiency Anemia in Children. SCIENTIFIC JOURNAL OF APPLIED AND MEDICAL SCIENCES, 2(5), 266–269. Retrieved from https://www.sciencebox.uz/index.php/amaltibbiyot/article/view/7229

[6] Abdulaziz Kebede Kassaw, Ali Yimer, Wondwosen Abey, Tibebu Leg-esse Molla, and Alemu Birara Zemariam. The application of machine learning approaches to determine the predictors of anemia among under five children in ethiopia. Sci. Rep., 13(1):22919, December 2023

[7] Tuba Karagül Yıldız, Nilüfer Yurtay, Birgül Öneç, Classifying anemia types using artificial learning methods, Engineering Science and Technology, an International Journal, Volume 24, Issue 1, 2021, ISSN 2215-0986, https://doi.org/10.1016/j.jestch.2020.12.003.

[8] Md Riyad Hossain, & Dr. Douglas Timmer. (2021). Machine Learning Model Optimization with Hyper Parameter Tuning Approach. Global Journal of Computer Science and Technology, 21(D2). Retrieved from https://gjcst.com/index.php/gjcst/article/view/255

[9] Chancellor, S., De Choudhury, M. Methods in predictive techniques for mental health status on social media: a critical review. npj Digit. Med. 3, 43 (2020). https://doi.org/10.1038/s41746-020-0233-7

[10] Dhal, P., Azad, C. A comprehensive survey on feature selection in the various fields of machine learning. Appl Intell 52, 4543–4581 (2022). https://doi.org/10.1007/s10489-021-02550-9