

PREDICTING ANEMIA FOR PREGNANT WOMEN USING DIFFERENT MACHINE LEARNING ALGORITHMS

Chrakhan Hassan Karim¹, Nzar Abdulqader Ali², and Soran Husen Mohamad³

¹ College of Administration and Economic, Statistics and Informatic department,
University of Sulaimani, Email:chrakhan.hassan@univsul.edu.iq.

² College of Administration and Economic, Statistics and Informatic department,
University of Sulaimani, Email:nzar.ali@univsul.edu.iq.

³ College of Administration and Economic, Statistics and Informatic department,
University of Sulaimani, Email:soran.abdulrahman@univsul.edu.iq.

<https://doi.org/10.30572/2018/KJE/170328>

ABSTRACT

Anemia appears as a common medical condition which affects pregnant women internationally since it produces serious health issues for mothers and their developing babies. The main cause of anemia comes from red blood cell and hemoglobin deficiencies specifically related to iron deficiency while vitamin B12 deficiency and chronic diseases contribute to its progression. Tangible methods for early diagnosis matter to deliver proper management of pregnancy-related issues yet they prove hard to achieve in areas with limited resources. The study tackles the requirement for advanced anemia in pregnancy predictive methods by implementing machine learning algorithms to strengthen diagnostic performance. A team compiled the dataset at sulaimaniya maternity hospital which contained socio-demographic information and medical histories and blood test results from pregnant females. The collected dataset presented data on age as well as occupation and gestational age together with dietary habits and chronic diseases and blood parameters measured as hemoglobin and iron levels. The study optimizes model performance through the use of multiple feature selection methods including PCA, SelectKBest, Chi-Square and RFE. The study trained and evaluated eight machine learning models which included Random Forest (RF), Gradient Boosting (GB), AdaBoost, Extra Trees, Support Vector Machines (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN) and Decision Trees (DT) and Naive Bayes (NB). Each model received evaluation through assessments of accuracy, precision and recall and F1-score measures. The integration of feature selection led to substantially better model performance because RF, Extra Trees, SVM and LR



demonstrated near-perfect results. The performance of KNN remained significantly lower than other models because it demonstrated weak abilities in selecting features. Among the evaluated algorithms SelectKBest and Chi-Square along with RFE demonstrated superior performance than PCA for obtaining 100% accuracy levels. This research shows that machine learning systems have potential in predicting anemia in pregnant women thus promoting improved diagnostic strategies in underprivileged healthcare settings. This research utilizes maternity hospital information to demonstrate the advantages of data-based approaches in maternal healthcare and presents machine learning as a tool for better healthcare results.

KEYWORDS

Anemia prediction, machine learning, feature selection, pregnant women.

1. INTRODUCTION

Anemia stands as one of the most common health conditions that affects women globally including those who are pregnant because it comes from red blood cell or hemoglobin deficiencies in blood. The body becomes unable to supply sufficient oxygen to tissues thus creating severe health risks for both the mother and her unborn child ([Cappellini et al. 2022](#)). The global healthcare community considers anemia in pregnancy an important health concern because it generates multiple complications which range from preterm birth to fetal and maternal death. WHO data shows anemia affects 42% of pregnant women worldwide and iron deficiency stands as the primary source of this maternal condition ([Wu et al. 2024](#)). Additionally, to iron deficiency, vitamin B12 deficiency along with infections such as malaria and chronic diseases including tuberculosis and chronic kidney disease account for elevated anemia cases among pregnant women ([Ifeanyi, E.O., & Uzoma 2024](#)).

The natural process of pregnancy creates higher food requirements that increases a woman's risk for essential vitamin and mineral deficiencies particularly affecting her iron stores and folic acid levels. Pregnant women who experienced multiple pregnancies or frequent abortions or who have pregnancies that occur in short intervals face higher chances of developing anemia ([Verulava and Gogua 2024](#)). Insufficient dietary consumption of both iron-rich foods and fruits and vegetables together with proteins leads to nutritional deficiencies which develops anemia. The risk of anemia among pregnant women depends on their socio-economic standing which includes family finances along with both educational attainment and healthcare service capabilities and nutritional food choices ([Gibore et al. 2021](#)).

Anemia exists throughout the population but healthcare providers frequently miss this condition due to restricted access to prenatal care services and laboratory diagnostic tools in resource-limited areas. Both maternal and fetal health benefit from the early diagnosis followed by prompt medical supervision to prevent anemia-related issues ([Obeagu and Obeagu 2024](#)). The current diagnostic practices depend heavily on blood tests which include evaluation of hemoglobin concentration and red blood cell count together with serum ferritin levels and multiple other tests. These diagnostic techniques prove difficult to obtain in developing nations and select geographies because of limited accessibility. People require new predictive methods to identify anemia that use advanced tools which combine different health data sources for accurate results ([Alsafi et al. 2024](#)).

Machine learning (ML) has proven itself as a valuable healthcare instrument which demonstrates promising capabilities for disease diagnosis and prediction such as anemia within recent years ([Shhatha and Alsaif 2025](#)). Machine learning algorithms analyze extensive datasets

consisting of patient characteristics like demographics alongside medical records and life habit information and blood test results to discover concealed patterns that lead to dependable predictions (Furizal, Ma'arif, and Rifaldi 2023). Healthcare professionals gain improved diagnostic and customized therapeutic capabilities through ML models because these systems process extensive patient data more effectively than conventional statistical approaches (Javaid et al. 2022).

The analysis implements machine learning algorithms to forecast anemia in pregnant women while processing maternity hospital-based dataset information. The information set contains different variables which include personal information about patient age alongside education level and pregnancy history and the presence of chronic diseases and dietary patterns and blood analysis data (Alesawi and Alawadi 2025). The research variables enable analysts to measure what affects anemia as well as pinpoint its primary risk factors. Multiple feature selection methods such as Principal Component Analysis (PCA), SelectKBest, Chi-Square, and Recursive Feature Elimination (RFE) are used because they enhance the prediction accuracy. The feature selection methods reduce data complexity by finding significant attributes which enhance model effectiveness through performance improvement (Shneen et al. 2025).

Various machine learning algorithms such as Random Forest, Gradient Boosting, AdaBoost, Extra Trees, Support Vector Machines (SVM), Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees (DT) and Naive Bayes were evaluated to determine their performance in this study. Main performance indicators of these prediction models include accuracy and precision along with recall and F1-score which deliver a thorough understanding of anemia prediction capabilities. Through the investigation of machine learning applications in maternal healthcare this research aims to enhance knowledge about predictive healthcare while proving the potential of ML to advance maternal health outcomes.

2. RELATED WORKS

The research by Meitei & Jiran et al. investigated child anemia prediction in North-Eastern Indian states through machine learning methodology (2022). The dataset included 21,000 children and the essential predicting variables consisted of maternal education status alongside social class indicators as well as health-related information. A LASSO regression achieved best results using penalized regression methods with an accuracy exceeding 64% during analysis. Due to its adequate predictive ability and high accuracy levels LASSO regression proves suitable for predicting anemia among children between 6 to 59 months of age (Meitei et al. 2022).

The research of Dejene Belayneh [Endalamaw et al. \(2022\)](#) analyzed anemia prediction levels for Ethiopian pregnant women by applying homogeneous ensemble machine learning methods. The research included 11,174 cases with a total of 34 elements including occupation types and water supply types as well as pregnancy length information. Research revealed that the combination of Catboost with class decomposition produced outstanding accuracy results which reached 97.6%. Catboost demonstrates outstanding capabilities as an anemia prediction tool for pregnant women which surpasses Random Forest and XGBoost techniques ([Dejene, Abuhay, and Bogale 2022](#)).

The research by Maasthi Manasvi [Jagadeesh et al. \(2023\)](#) created an anemia prevention and malnutrition elimination support system through predictive decision tools. A total of 1,799 subjects participated in this study which employed age, gender and multiple blood cell count variables as predictors. The research examined K-Nearest Neighbor (kNN) and Naïve Bayes (NBC), Random Forest (RF) alongside Decision Tree (DT) as testing algorithms. The Random Forest model delivered 95.49% accuracy in predicting anemia across both male and female populations which established it as an optimal model for this purpose ([Maasthi et al. 2023](#)).

The researchers behind Kassaw Abdulaziz Kebede et al., 2023 employed machine learning approaches to determine the factors that lead to anemia among children under five in Ethiopia. The study adopted data from 9,501 children while using maternal wealth index and health insurance coverage and cooking fuel type as its main predictors. The Boruta algorithm used its methodology to find the most important variables for anemia occurrence. The research data confirms machine learning methods show usefulness for understanding what factors affect anemia in Ethiopian young children ([Kebede Kassaw et al. 2023](#)).

The researchers at Pan Yunhui et al., 2023 developed a machine learning system to forecast iron deficiency anemia among Chinese premenopausal women after sleeve gastrectomy at the twelve-month mark. The clinical data contained 407 patients while preoperative ferritin levels together with age and hemoglobin levels represented the main predictors. A Linear Support Vector Machine (SVM) was used for the study while achieving an AUC score of 0.858 in one analysis and 0.799 in another which demonstrated high predictive accuracy. The calibration curve established that the prediction model worked effectively to forecast anemia levels in this particular patient demographic ([Pan et al. 2023](#)).

According to Shweta & Pande, 2023, the research analyzed multiple ensemble learning together with boosting methods for anemia prediction. A total of 16,721 dataset instances contained the six predicting elements which were hemoglobin and MCV levels. The experts ran ANN, LR, SVM and ensemble algorithms which contained AdaBoost and XGBoost methods in their

analytical workflow. XGBoost reached 100% prediction accuracy for anemia in specific groups as it executed its process in 0.2404 seconds (Shweta and Pande 2023).

The researchers at Zemariam Alemu Birara et al. (2024) conducted an analysis to forecast anemia prevalence in Ethiopian young girls through supervised machine learning algorithms. The researchers analyzed 5,642 weighted samples through which they evaluated predictors from family size and wealth index and education. The performance analysis revealed Random Forest as the best algorithm out of SVM, Random Forest and LR when it came to predicting anemia based on ANN. The random forest algorithm proved its suitability as a prediction tool for anemia in youth girls because it reached 72% accuracy and 82% AUC value (Zemariam et al. 2024).

Researchers at Göl Mehmet et al. (2024) evaluated the application of machine learning to forecast anemia due to malnutrition in elderly patients. This research analyzed anemia among 438 participants while evaluating factors that included body weight together with alcohol consumption and muscle composition variables and clinical testing results. The Random Forest (RF) proved to be the optimal algorithm which delivered 85.39% accuracy results. The J48 algorithm demonstrated superior performance when analyzing geriatric patient anemia through the addition of hemogram data to achieve 97.77% accuracy (Göl et al. 2024).

The analysis of childhood anemia in Ethiopia during 2016 was conducted through the examination of the Ethiopian Demographic Health Survey by Tesfaye Solomon Hailemariam et al. in 2024. A total of 8,482 children were part of the dataset which used stunting along with maternal education and family income as predictors. Logistic Regression (LR) proved itself as the most suitable model among the tested methods which included Random Forest (RF) and Decision Trees (DT). This method demonstrated a 66% accuracy level together with a 69% AUC rating thus establishing itself as a dependable tool to predict anemia in children between 6-59 months old (Tefaye, Seboka, and Sisay 2024).

3. BACKGROUND THEORY

3.1. Socio-Demographic and Health-Related Parameters of Pregnant Women

Various health-related and socio-demographic indicators appear in Table 1 to help identify pregnancy-related anemia factors. The dataset contains categorical and numerical variables that produce an all-inclusive snapshot of patient characteristics and health situation. The age variable divides into three age groups due to its vital impact on pregnancy results (20-30, 30-40, >40). The anemia risk during pregnancy depends on occupation and educational level and family income since these factors determine healthcare and nutritional resource access. The development of anemia during pregnancy depends heavily on three health-related elements

including gestational age and the total number of pregnancies along with any past abortions (Shulman 2014).

Table 1 Questionnaire about socio-demographic and other parameters related to pregnant women

Variable	Category
Age	20-30, 30-40, >40
Occupation	Housewife, Work, Employed
Level of Education	Illiterate, Primary school, High school, Higher education
Gestational Age	First trimester, Second trimester, Third trimester
No. of Pregnancy (Gravidity)	1 to 2, 3 to 4, 5 or more
Abortion	Yes (number of abortions), No
Birth Interval	<2 years, >2 years, First birth
Iron Supplementation on Current Pregnancy	Yes (types), No
Presence of Chronic Disease	Yes (type of chronic diseases), No
Body Mass Index (BMI) in kg/m ²	Low, Normal, High
Family Income	Low, Intermediate, High
Type of Food/Style of Food (weekly)	Fish, Meat, Fruit, etc.
Daily Activity	Exercise (Yes, No)
Residency	Urban, Rural
Blood Pressure	Lower than normal, Normal, Higher than normal
WBC (White Blood Cells)	Low (<3.8), Normal (3.8–11), High (>11)
MCH (Mean Corpuscular Hemoglobin)	Low (<25), Normal (25–34), High (>34)
MCHC (Mean Corpuscular Hemoglobin Concentration)	Low (<30), Normal (30–36), High (>36)
RBC (Red Blood Cells)	Low (<3.8), Normal (3.8–5.6), High (>5.6)
MCV (Mean Corpuscular Volume)	Low (<80), Normal (80–95), High (>95)
HCT (Hematocrit)	Low (<35.5), Normal (35.5–48), High (>48)
RDW% (Red Cell Distribution Width)	Low (<11.2), Normal (11.2–15), High (>15)
PLT (Platelet)	Low (<150), Normal (150–400), High (>400)
S. Iron (Serum Iron)	Low (<37), Normal (37–145), High (>145)
S. Ferritin	Low (<13), Normal (13–150), High (>150)
Folic Acid	Low (<3.1), Normal (3.1–17.5), High (>17.5)
Vitamin B12 (V.B12)	Low (<200), Normal (200–835), High (>835)

Pregnant women face substantial impact on their nutritional intake and pregnancy health because of their lifestyle choices which incorporate their physical activity habits and eating patterns particularly regarding fish consumption and meat intake and fruit consumption frequency. The evaluation of blood pressure alongside body mass index (BMI) and chronic disease identification helps determine anemia risk because it shows how well a person resists anemia-related conditions (Morales-Suárez-Varela et al. 2023). White blood cell count (WBC) together with hemoglobin levels (MCH, MCHC, RBC) and hematocrit (HCT) and red cell distribution width (RDW%) and serum measurements of iron, ferritin, folic acid, and vitamin B12 supply vital information about blood health status and deficiencies which enables anemia diagnosis. These factors work together to offer a complete view of the complex factors affecting anemia in pregnant women to establish a strong basis for forecasting and treating this condition (Farhan and Jabbar 2014).

3.2. Features Selection Algorithms

Principal Component Analysis (PCA): The dimensionality reduction method Principal Component Analysis (PCA) reduces feature dimensions in datasets by maintaining maximum possible information. PCA converts the initial features into principal components which exhibit no correlation between each other. The first components maintain the maximum amount of original data variance when the components are arranged in descending order. Standardization of the dataset follows as a first step before moving forward to covariance matrix calculation. The eigen analysis of the covariance matrix reveals its principal components together with their corresponding eigenvalues. The ranking of components occurs according to their captured variance where the most significant components become part of the reduced feature selection. PCA follows an equation composed of the covariance matrix Σ eigenvalue decomposition (Salem and Hussein 2019).

$$\Sigma v = \lambda v \quad (1)$$

Where:

Σ : is the covariance matrix of the data,

v : is the eigenvector (principal component),

λ : is the eigenvalue (variance explained by the component).

SelectKBest: Using SelectKBest users can find the most important k features through statistical testing. The scoring function assesses relevance in each feature until the method generates feature ranks. The user determines the value of k along with the method identifying the most suitable k features that explain the outcome variable. The selection of statistical tests includes chi-squared tests and ANOVA F-test and mutual information among others for assessing feature relevance. The technique functions in classification and regression tasks. A chi-squared test applies to classification tasks as SelectKBest implements it for identifying relevant features in this context (Abid-Althaqafi, Alsalamah, and Ismail 2024). The ultimate formula of SelectKBest relies on the selection of the chosen scoring function. For example, for the chi-squared test:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

Where:

O_i : is the observed frequency,

E_i : is the expected frequency.

Chi-Square Test: A statistical assessment named the Chi-Square test examines two categorical variables for independence. This test serves feature selection needs by verifying if features remain uncorrelated from the target variable when applied to categorical variables. The Chi-

Square test compares the observed frequency of occurrences in each category to the expected frequency under the null hypothesis of independence ([Rana and Singhal 2015](#)).

Recursive Feature Elimination (RFE): The recursive feature elimination methodology functions as a feature selection tool through model performance-driven elimination of unimportant features. The method begins with all features during which it constructs a model (e.g. linear regression, decision tree). RFE sorts features by their relevance after which it removes the least crucial elements. The procedure iterates until it reaches the specified number of selected features. The iterative RFE process demands that the model trains repeatedly while the number of features steadily decreases. The model obtains assessments based on performance metrics which include accuracy and mean squared error ([Priyatno and Widiyaningtyas 2024](#)).

3.3. Machine Learning Models

Random Forest: functions as an ensemble learning approach by grouping multiple decision trees for better accuracy achievement. The algorithm constructs numerous decision trees by selecting random data subsets with random features before performing an average computation that lowers overfitting effects and boosts model reliability ([Salman, Kalakech, and Steiti 2024](#)).

Gradient Boosting: Each new tree in Gradient Boosting models builds sequentially to fix errors that occurred in preceding trees. Gradient descent minimization of loss functions through its algorithm enables high effectiveness in complex problems but requires appropriate tuning to prevent overfitting ([Natekin and Knoll 2013](#)).

AdaBoost: The combination of weak classifiers through AdaBoost results in a strong classifier by using decision trees. The weighting system grants additional importance to misclassified examples so that new models must tackle more challenging predicaments. Through this approach the performance gains become noticeable before the number of trees gets excessive ([Hornýák and Iantovics 2023](#)).

Extra Trees: The ensemble method called Extra Trees introduces additional randomness to Random Forests through its Extremely Randomized Trees approach. The technique constructs numerous decision trees which apply random attribute splits and threshold divisions so training occurs quickly and generalized performance can potentially improve ([Bui, Nguyen, and Soukhanouvong 2022](#)).

Support Vector Machine (SVM): operates as a robust classifier by defining the optimal hyperplane which divides different classes in spaces of high dimensionality. The method works to establish maximum space between different classes to produce better generalization results. SVM excels in high-dimensional settings yet its computational requirements may become substantial ([Ghaddar and Naoum-Sawaya 2017](#)).

Logistic Regression: The linear statistical method called Logistic Regression serves for two-class predictive modeling. A logistic function enables this model to calculate the class probability. This model represents a basic approach which delivers effective results when the target has linear approximation with its features (Skiena 2017).

K-Nearest Neighbors (KNN): represents a non-parametric instance-based learning algorithm. In KNN classification a new point receives its label from the most common class type found among its nearest neighboring points across the dimensional space. The ease of KNN implementation comes with expensive computational cost during the prediction stage (Taunk et al. 2019).

Decision Tree: functions as a model which sorts data through feature value-based splits. The model consists of decision points at nodes which transition into class labels found at leaf nodes. KNN offers easy interpretability yet its performance suffers from overfitting so practitioners normally implement Random Forest with pruning to enhance predictability (Zhou et al. 2021).

Naive Bayes: implements a probabilistic classifier which bases its methods on Bayes' theorem under the condition that features maintain independence from each other. Naive Bayes proves effective when classifying text while working well in contexts where features demonstrate independence properties to deliver quick scalable predictions (Jiang et al. 2013).

4. METHODOLOGY

The method requires multiple successive procedures for machine learning dataset analysis and processing. A Data Frame contains the dataset after it gets loaded into the system. During data preprocessing the non-numeric variables become numeric results and missing data points get replaced by computing the mean value from each respective column. After cleaning the data, it experiences standardization-based feature scaling to normalize all feature values so each characteristic brings equal contribution to the model. The following step includes applying feature selection methods PCA, KBest and Chi-Square and Recursive Feature Elimination (RFE) to choose the most significant features. The SMOTE technique applies Synthetic Minority Oversampling Technique to create artificial minority class samples to achieve class balance in the selected features dataset.

The dataset splits into training and testing subsets where 70% serves training purposes and 30% remains for testing. A total of eight machine learning algorithms gets trained including Random Forest, Gradient Boosting, AdaBoost, Extra Trees, SVM, Logistic Regression, KNN, Decision Tree and Naive Bayes. Evaluation of trained models includes an assessment through confusion matrices and classification reports. After training the models they are saved for future applications and the results are printed while providing an effectiveness summary. The

optimization process through this methodology allows machines to evaluate their prediction accuracy for the target variable. Fig.1 describe model methodology.

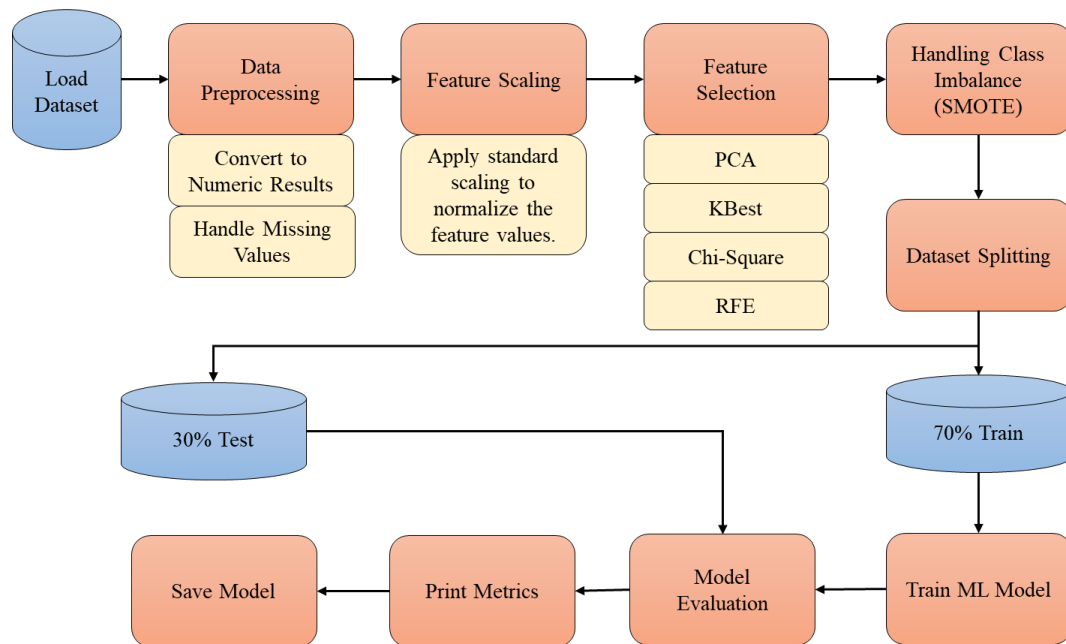


Fig. 1. Model Methodology.

4.1. Dataset Description

The dataset in Table 2 consists of 386 rows alongside 32 columns which detail different elements regarding anemia together with possible influence factors. The features in the dataset cover demographic information, medical history, lifestyle habits, and various blood test results. The database provides essential information required for anemia analysis and its variable associations including age groups as well as occupation type and educational level and pregnancy status and chronic diseases and nutritional patterns and blood test results. The data collection from sulaimaniya maternity hospital involved manual methods which delivered important insights about regional anemia and its origins together with its effects in the population.

The chosen variables within the data collection demonstrate a complete understanding of elements which affect anemia development. The demographic variable Age affects both health status and nutritional requirements. The level of education in the population demonstrates both knowledge about health matters and ability to obtain proper dietary nourishment. Abortion serves as a pregnancy-related factor that impacts both iron levels and health in women. The Fruit feature demonstrates how frequently people eat fruit while revealing nutritional patterns that affect vitamin consumption particularly vitamin C which helps the body absorb iron. The Other feature expands the analysis of food consumption by providing expanded details about nutritional deficits within an individual's diet.

Table 2. Dataset Features

Feature	Description
Age	Age of the individual (in years).
Occupation	Occupation of the individual (encoded as an integer).
Level of Education	Education level of the individual (encoded as an integer).
Gestational Age	The number of weeks the individual is pregnant (in weeks).
No. of Pregnancy (Gravidity)	The number of pregnancies the individual has had.
Abortion	Number of abortions the individual has experienced.
Birth Interval	The time gap between the individual's pregnancies (in years).
Presence of Chronic Disease	Whether the individual has any chronic diseases (encoded as an integer).
Body Mass Index (BMI) in kg/m ²	Body Mass Index (BMI) of the individual, calculated as weight in kg/height ² in meters.
Family Income	The annual family income (encoded as an integer).
Fish	Frequency of fish consumption (encoded as an integer).
Meat	Frequency of meat consumption (encoded as an integer).
Fruit	Frequency of fruit consumption (encoded as an integer).
Other	Frequency of other food consumption (encoded as an integer).
Daily Activity	Level of daily physical activity (encoded as an integer).
Blood Pressure	Blood pressure of the individual (encoded as an integer).
WBC (White Blood Cell Count)	White blood cell count in the blood (measured in thousands/ μ L).
HGB (Hemoglobin)	Hemoglobin level in the blood (measured in g/dL).
MCH (Mean Corpuscular Hemoglobin)	Average amount of hemoglobin in a red blood cell (measured in pg).
MCHC (Mean Corpuscular Hemoglobin Concentration)	Average concentration of hemoglobin in a red blood cell (measured in g/dL).
RBC (Red Blood Cell Count)	Number of red blood cells per microliter of blood (measured in millions/ μ L).
MCV (Mean Corpuscular Volume)	Average volume of red blood cells (measured in fL).
HCT (Hematocrit)	Percentage of blood volume occupied by red blood cells.
RDW% (Red Cell Distribution Width)	Variation in the size of red blood cells (measured in %).
PLT (Platelet Count)	Platelet count in the blood (measured in thousands/ μ L).
S. Iron (Serum Iron)	Iron level in the blood serum (measured in μ g/dL).
S. Ferritin	Ferritin level, indicating iron storage in the body (measured in ng/mL).
Folic Acid	Level of folic acid in the blood (measured in ng/mL).
V. B12	Vitamin B12 level in the blood (measured in pg/mL).
HGB2	Secondary measure of hemoglobin level in the blood (measured in g/dL).
Type	Type of anemia (encoded as an integer).
Anemia	Indicator of anemia status (1 = Anemic, 0 = Not Anemic).

The amount of physical activity recorded in Daily Activity affects anemia since it changes health status and energy availability. WBC (White Blood Cell Count) measures immune system function yet it might indicate potential infections or chronic diseases which could affect anemia development. The vital indicator of anemia is HGB (Hemoglobin) because its reduced levels always indicate the presence of this condition. Anemia severity and its different forms can be examined through blood tests which measure MCH (Mean Corpuscular Hemoglobin) along with MCV (Mean Corpuscular Volume). Doctors use HCT (Hematocrit) to determine the proportion of red blood cells that fill up blood volume because this measurement allows clear

anemia diagnosis. The anemia diagnostic marker RDW% determines red blood cell size variability so it might increase with specific forms of anemia.

The (Serum Ferritin) blood test reveals storage levels of body iron to help confirm iron-deficiency anemia diagnosis. The anemia diagnostic process benefits from the supplementary information provided by HGB2 which serves as an alternate method to assess hemoglobin levels. The Type feature classifies the anemia types which function as the main outcome variable for analytical purposes. The features collaborate to produce important information about anemia-related elements including dietary patterns and blood test results and patient medical background.

4.2. Dataset Preprocessing

The analysis data required multiple preprocessing stages before it could train machine learning models. The CSV file imported data which underwent numeric conversion while solving non-numeric characters by either conversion or substitution to numeric values. The research team separated "Animia" as the target variable from all other features. The preprocessing method resolved missing feature values by replacing them with the mean value from each individual column thus creating a null value-free dataset. A Min-Max Scaler performed the standardization of feature values as the next operation. The data transformation using this scaling method creates values within a 0 to 1 range that benefits specific algorithms particularly support vector machines and k-nearest neighbors because they require such normalized feature scales.

The preprocessing completed Recursive Feature Elimination (RFE) in conjunction with Logistic Regression for selecting features. The RFE method identifies and retains the 15 features expected to contribute most power to prediction. The chosen features underwent SMOTE (Synthetic Minority Over-sampling Technique) resampling treatment in order to balance the distribution of classes within the dataset. SMOTE creates new minority class examples to stop the model from showing bias towards the predominant class. The preprocessed information was divided into training and testing sections where 70% of the data served training purposes and 30% served testing purposes. The data partition allows performance assessment of the model on new information. [Table 3](#) summarized the dataset preprocessing operations.

Table 3 Dataset Preprocessing Operations.

Step	Description
Handling Non-Numeric Data	Converted non-numeric data to numeric values or replaced them.
Missing Value Imputation	Filled missing values in features with their respective column means.
Data Resampling (SMOTE)	Applied SMOTE to generate synthetic samples for the minority class, balancing the dataset.
Data Splitting	Split the data into 70% training and 30% testing sets to evaluate model performance.

4.3. Features Selection Methods Parameters

Table 4 outlines four different feature selection techniques with their essential parameters. PCA selects the initial 15 principal components from the dataset which represent the maximum variability in the data. SelectKBest with ANOVA F-value detects the most discriminating 15 features through an F-test analysis to differentiate class categories. The Chi-Square test in SelectKBest determines the 15 most important features through a statistical independence analysis of target-dependent categorical variables. Recursive Feature Elimination (RFE) together with Logistic Regression finds the 15 most valuable features through an iterative process which optimally removes lower importance features. The selection methods match distinct requirements based on the characteristics of data alongside the chosen modeling techniques.

Table 4 Features Selection Methods Parameters

Feature Selection Method	Model/Method	Key Parameter(s)	Description
Principal Component Analysis (PCA)	PCA	n_components=15	Reduces dimensionality by projecting the data onto the first 15 principal components, capturing the most variance in the dataset.
SelectKBest (ANOVA F-value)	f_classif	k=15	Selects the top 15 features based on the ANOVA F-value, which measures the variance between the feature groups and is useful for classification tasks.
SelectKBest (Chi-Square)	chi2	k=15	Selects the top 15 features based on the chi-square test, which measures the statistical independence between categorical features and the target variable.
Recursive Feature Elimination (RFE)	LogisticRegression	n_features_to_select=15	Recursively removes features by fitting a model and selecting the 15 most important features based on model performance. This method helps improve model accuracy by selecting only the most relevant features.

4.4. Machine Learning Models Training Parameters

The parameter configurations employed for different machine learning models appear in Table 5. The performance of each model depends on various designated parameters which control its operating behavior. The number of trees and maximum depth together with n_estimators and max_depth control the performance of Random Forest and Extra Trees models while SVM and Logistic Regression models require C, kernel, and solver parameters for optimization adjustments. The parameters enable the achievement of optimal results because they adapt to the dataset combined with the target task. The table lists standard setups yet

organizations can modify these values when facing different datasets or tasks.

Table 5 Machine Learning Models Training Parameters

Model	Parameter	Typical Value(s)
Random Forest	n_estimators	100
	max_depth	10
	min_samples_split	2
	max_features	integer value
Gradient Boosting	n_estimators	100
	learning_rate	0.01
	max_depth	5
	subsample	0.5
AdaBoost	n_estimators	50, 100, 200, etc.
	learning_rate	0.01
Extra Trees	n_estimators	100
	max_depth	10
	min_samples_split	5
	max_features	integer value
SVM	C	10
	kernel	linear
	gamma	0.01
	degree	3
Logistic Regression	C	0.01
	max_iter	100
	solver	Lbfgs
K-Nearest Neighbors	n_neighbors	5
	weights	distance
	algorithm	Auto
	max_depth	10
Decision Tree	min_samples_split	5
	min_samples_leaf	2
Naive Bayes	criterion	entropy
	var_smoothing	1e-9

5. RESULTS AND DISCUSSION

The research results evaluate machine learning model performance when using feature selection methods with different techniques. Model performance improvement requires feature selection as a fundamental step to eliminate features that are useless or repetitive because this technique reduces both overfitting and improves generalization. The experimental models encompass RF along with GB and AdaBoost as well as Extra Trees and SVM and then LR followed by KNN and Decision Trees and Naive Bayes. A performance evaluation of each model occurred through the implementation of Principal Component Analysis (PCA), SelectKBest, Chi-Square and Recursive Feature Elimination (RFE) as popular feature selection approaches.

5.1. PCA Feature Selection

PCA application for feature selection shows the majority of our models deliver excellent results as shown in [Table 6](#). All four models including Random Forest (RF), Extra Trees, Support Vector Machine (SVM) and Logistic Regression (LR) achieve remarkable 98% accuracy

alongside 98% precision and recall and F1-score values which prove their exceptional performance in this task. The dimensionality reduction achieved through PCA offers the perfect conditions for these models to operate on the dataset. The accuracy results of Gradient Boosting (GB) and AdaBoost exhibit performance degradation compared to other models while achieving 95% accuracy in conjunction with 95% precision and recall and 95% F1-score. These models demonstrate slightly worse efficiency than the previously mentioned models yet they still deliver satisfactory performance results. K-Nearest Neighbors (KNN) alongside Decision Trees (DT) demonstrate clear degradation in their performance assessment. The accuracy rate of KNN stands at 94% while its precision and recall perform worse than other models. Decision Trees attained the lowest accuracy rate of 86% following feature selection which indicates the least amount of improvement from this process. The decision tree nature possibly explains this finding because trees tend to overfit data and have reduced generalization especially after feature reduction. PCA demonstrates effective performance improvement for most models yet it shows low impact on KNN and DT.

Table 6. Models Results with PCA Feature Selection

Model	Accuracy	Precision	Recall	F1-Score
RF	98.0%	98.0%	98.0%	98.0%
GB	95.0%	96.0%	95.0%	95.0%
AdaBoost	97.0%	97.0%	97.0%	97.0%
Extra Tree	98.0%	98.0%	98.0%	98.0%
SVM	98.0%	98.0%	98.0%	98.0%
LR	98.0%	98.0%	98.0%	98.0%
KNN	94.0%	94.0%	94.0%	94.0%
DT	86.0%	87.0%	87.0%	86.0%
NB	92.0%	93.0%	92.0%	92.0%

5.2. SelectKBest Feature Selection

The models demonstrate exceptional performance after utilizing SelectKBest selection procedures as shown in Table 7. Random Forest (RF) together with Gradient Boosting (GB), AdaBoost, Extra Trees, SVM, Logistic Regression (LR), Decision Trees (DT) and Naive Bayes (NB) deliver 100% accuracy and achieve perfect scores for precision, recall and F1-score of each classification test. The selected features by SelectKBest enable the models to function optimally without overfitting because these features contain the most important information. The algorithms achieve their maximum predictive strength because of this approach. K-Nearest Neighbors (KNN) lowers its performance score although it maintains a 98% accuracy level. KNN reaches an accuracy level of 98% but its precision and recall and F1-score scores remain at 98% as well. The performance of KNN becomes better with SelectKBest but does not reach the equal level of excellence demonstrated by other models. KNN's distance-based approach

probably receives minimal benefits from feature selection since it performs differently than both tree-based and linear classification models. The application of SelectKBest improves most models to achieve perfect accuracy scores yet the KNN model still shows weaker performance.

Table 7. Models Results with SelectKBest Feature Selection

Model	Accuracy	Precision	Recall	F1-Score
RF	100.0%	100.0%	100.0%	100.0%
GB	100.0%	100.0%	100.0%	100.0%
AdaBoost	100.0%	100.0%	100.0%	100.0%
Extra Tree	100.0%	100.0%	100.0%	100.0%
SVM	100.0%	100.0%	100.0%	100.0%
LR	100.0%	100.0%	100.0%	100.0%
KNN	98.0%	98.0%	98.0%	98.0%
DT	100.0%	100.0%	100.0%	100.0%
NB	100.0%	100.0%	100.0%	100.0%

5.3. Chi-Square Feature Selection

The Chi-Square feature selection approach produces excellent results as shown in [Table 8](#) when assessing most of the tested models. All seven algorithms including Random Forest (RF), Gradient Boosting (GB) and AdaBoost, Extra Trees, SVM, Logistic Regression (LR) and Naive Bayes (NB) present perfect accuracy alongside ideal precision, recall and F1-scores. The algorithms demonstrate optimal performance because Chi-Square feature selection proves very effective at selecting important features. The Chi-Square method demonstrates strong reliability in improving model performance because the selected features maintain high consistency across models after feature selection. The K-Nearest Neighbors algorithm stands as the only model that does not benefit from Chi-Square selection. The accuracy of KNN remains at 95% but its precision and recall scores along with F1-score fall to 96% and 95%. The systematic relationship between features and class labels appears to bring minimal advantage to KNN under Chi-Square selection because KNN depends more heavily on different classification methodologies. Because KNN achieves lower performance with Chi-Square feature selection while other models perform better it demonstrates this approach may not provide KNN with maximum possible benefits.

Table 8. Models Results with Chi-Square Feature Selection

Model	Accuracy	Precision	Recall	F1-Score
RF	100.0%	100.0%	100.0%	100.0%
GB	100.0%	100.0%	100.0%	100.0%
AdaBoost	100.0%	100.0%	100.0%	100.0%
Extra Tree	100.0%	100.0%	100.0%	100.0%
SVM	100.0%	100.0%	100.0%	100.0%
LR	100.0%	100.0%	100.0%	100.0%
KNN	95.0%	96.0%	95.0%	95.0%
DT	100.0%	100.0%	100.0%	100.0%
NB	100.0%	100.0%	100.0%	100.0%

5.4. Recursive Feature Elimination (RFE) Feature Selection

The application of Recursive Feature Elimination (RFE) for feature selection produced identical performance outcomes in Table 9 as other methods because models reached 100% accuracy and attained full precision and recall and F1 scores. All performance of Random Forest (RF), Gradient Boosting (GB), AdaBoost, Extra Trees, SVM, Logistic Regression (LR), Decision Trees (DT), and Naive Bayes (NB) gets enhanced through Recursive Feature Elimination because it removes unneeded features while keeping only the most essential ones thus resulting in peak performance. RFE demonstrates exceptional effectiveness for these algorithms using this particular dataset because all metrics reach a perfect score. The K-Nearest Neighbors (KNN) algorithm demonstrates decreased performance by reaching 95% accuracy when precision stands at 96% and recall stands at 95% alongside an F1-score of 95%. The constant lower performance of KNN contrasts with every selection method tested which indicates KNN might not function optimally for this dataset or its structural aspects differ from the chosen techniques. The performance of most algorithms increases significantly because of RFE implementation yet KNN shows inferior results when compared to other models.

Table 9. Models Results with RFE Feature Selection

Model	Accuracy	Precision	Recall	F1-Score
RF	100.0%	100.0%	100.0%	100.0%
GB	100.0%	100.0%	100.0%	100.0%
AdaBoost	100.0%	100.0%	100.0%	100.0%
Extra Tree	100.0%	100.0%	100.0%	100.0%
SVM	100.0%	100.0%	100.0%	100.0%
LR	100.0%	100.0%	100.0%	100.0%
KNN	95.0%	96.0%	95.0%	95.0%
DT	100.0%	100.0%	100.0%	100.0%
NB	100.0%	100.0%	100.0%	100.0%

5.5. Discussion

Different feature selection approaches such as PCA, SelectKBest, Chi-Square and Recursive Feature Elimination (RFE) produced multiple findings that evaluate each model's effectiveness alongside the importance of feature selection methods. Every model test with Random Forest (RF), Extra Trees, SVM, and Logistic Regression (LR) demonstrated excellent results which achieved 100% accuracy and complete precision, recall, and F1-scores with minimal exceptions. The models show excellent suitability for the given dataset because they received major performance benefits from the applied feature selection approaches. Optimal performance resulted when ensemble methods (RF and Extra Trees) joined linear classifiers (SVM and LR) along with feature selection which demonstrated these models' effectiveness at handling irrelevant or redundant features.

The K-Nearest Neighbors (KNN) algorithm performed at a lower level consistently throughout every feature selection protocol. All feature selection techniques used with KNN resulted in precision, recall, F1-score, and accuracy levels between 94 and 98 percent. KNN does not seem to derive as much advantageous effect from feature selection methods as alternative models do. The distance-based model KNN depends strongly on data point distance calculation which results in reduced effectiveness when working with selected subset features. The performance decrease in KNN can be explained by its vulnerability to unimportant or redundant features that standard feature selection methods fail to remedy properly relative to tree-based models or linear classifiers.

Both Gradient Boosting (GB) and AdaBoost demonstrated solid outcomes although they delivered slightly lower results than the most effective approaches. The accuracy results of these models indicated between 95% and 97% accuracy based on different choices of feature selection methods. The ensemble characteristics of these models provide the benefit of reduced overfitting by implementing feature selection techniques. The optimal performance achieved by RF and Extra Trees surpassed the results obtained from Gradient Boosting and AdaBoost because boosting methods generally exhibit higher sensitivity to data noise when compared to bagging algorithms.

The performance of Decision Trees (DT) and Naive Bayes (NB) showed differences between different experimental conditions. The Restricted Boltzmann Machine model achieved the least successful performance in comparison to other models particularly with PCA feature selection which resulted in 86% accuracy but demonstrated improved results using alternative feature selection methods up to 100% accuracy. Naive Bayes (NB) demonstrated consistent perfect accuracy throughout all feature selection methods which proved its resistance to the feature selection processes. Naive Bayes demonstrates strong performance despite having limited features since it applies probability modeling to deliver accurate results as long as selected features retain strong class predictive strength.

Results indicate that SelectKBest, Chi-Square and RFE produce better results than PCA since they achieve 100% accuracy in most model tests. PCA functions as a dimensionality reduction technique yet it fails to conserve important features which stands in contrast to SelectKBest and RFE because they specifically target relevant features for the target variable. Chi-Square demonstrates effectiveness for RF together with LR and SVM since the models perform better with statistical correlations discovered through feature selection.

5.6. Related Works Comparison

[Table 10](#) details the similarities and differences between existing studies and the current

research based on dataset sizes and essential predictors together with employed algorithms and evaluation criteria along with major findings. Research studies utilize machine learning algorithms to forecast anemia conditions along with other associated medical problems across various patient groups including children and pregnant females as well as elderly adults. Meitei & Jiran et al. (2022) analyzed child anemia in North-Eastern India through LASSO regression with their dataset containing 21,000 children to reach an accuracy of more than 64%. The research by Dejene Belayneh Endalamaw et al. (2022) reached a remarkable 97.6% accuracy in predicting pregnant woman anemia through Catboost while surpassing Random Forest and XGBoost. XGBoost reached 100% accuracy for anemia prediction according to Shweta & Pande (2023) and required only 0.2404 seconds for execution time. The study used Random Forest and Gradient Boosting algorithms together with PCA and SelectKBest feature selection methods to reach 100% accuracy testing results. Machine learning models demonstrate adaptable and effective performance for healthcare predictions while handling different population types and medical scenarios.

Table 10. Related Works Comparison

Study	Dataset	Predictors	Algorithms	Findings
Meitei & Jiran et al., 2022	21,000 children	Maternal education, social status, health factors	Penalized Regression (LASSO, Ridge, Elastic Net)	LASSO regression achieved 64% accuracy, suitable for predicting anemia in children aged 6-59 months.
Dejene Belayneh Endalamaw et al., 2022	11,174 pregnant women	Occupation, water source, pregnancy duration, health-related factors	Catboost, Random Forest, XGBoost	Catboost with class decomposition achieved 97.6% accuracy, superior to other algorithms for predicting anemia in pregnant women.
Maasthi Manasvi Jagadeesh et al., 2023	1,799 individuals	Age, gender, red blood cells, blood cell count variables	KNN, Naïve Bayes, Random Forest, Decision Tree	Random Forest achieved 95.49% accuracy, suitable for predicting anemia in both men and women.
Kassaw Abdulaziz Kebede et al., 2023	9,501 children under 5	Maternal wealth index, health insurance, cooking fuel, family size	Boruta Algorithm	Machine learning methods identified key predictors for anemia in children under five in Ethiopia.
Pan Yunhui et al., 2023	407 Chinese women (post-sleeve gastrectomy)	Preoperative ferritin, age, hemoglobin levels	Linear SVM	SVM achieved an AUC of 0.858, suitable for predicting anemia in Chinese premenopausal women post-sleeve gastrectomy.
Shweta & Pande, 2023	16,721 instances	Hemoglobin, MCV, MCH, gender	ANN, Logistic Regression, SVM,	XGBoost achieved 100% accuracy in predicting anemia, ideal for specific demographic prediction.

Study	Dataset	Predictors	Algorithms	Findings
			XGBoost, AdaBoost	
Zemariam Alemu Birara et al., 2024	5,642 young girls in Ethiopia	Family size, wealth index, education status	Random Forest, SVM, Logistic Regression	Random Forest achieved 72% accuracy and 82% AUC, effective for predicting anemia in youth girls.
Göl Mehmet et al., 2024	438 elderly patients	Weight, alcohol use, muscle mass, strength, health tests	Random Forest, J48, Naïve Bayes, Decision Tree	Random Forest achieved 85.39% accuracy, J48 performed best with hemogram data, achieving 97.77% accuracy for geriatric anemia prediction.
Tesfaye Solomon Hailemariam et al., 2024	8,482 children in Ethiopia (2016 DHS)	Stunting, maternal education, family income	Logistic Regression, Random Forest, DT, KNN	Logistic regression achieved 66% accuracy and 69% AUC, suitable for predicting anemia in children aged 6-59 months.
Present Study	Custom dataset (current research)	Age, gender, RBC, hemoglobin, blood counts, and various health factors	Random Forest, Gradient Boosting, SVM, AdaBoost	The present study's models, particularly Random Forest and Gradient Boosting, achieved 100% accuracy with feature selection methods like PCA and SelectKBest.

6. CONCLUSIONS

Research findings establish that machine learning algorithms together with appropriate feature selection enhance the identification of pregnancy-related anemia in women. Random Forest alongside Extra Trees and SVM and Logistic Regression prove the best at predicting anemia in pregnant women through the implementation of feature selection methods SelectKBest and Chi-Square and RFE. The results show K-Nearest Neighbors experienced minimal enhancement from these methods because it lacks strong effectiveness in this application space. The collected data from sulaimaniya maternity hospital serves as a solid base for maternal health research by showing which factors lead to anemia in the area. The research reveals that machine learning shows promise to assist clinical practitioners in making timely accurate medical diagnoses which in turn leads to enhanced outcomes for both maternal and fetal patients.

7. REFERENCES

Abid-Althaqafi, Nahid R., Hessah A. Alsalamah and Walaa N. Ismail. (2024) "The Impact of the Weighted Features on the Accuracy of X-Platform's User Credibility Detection Using Supervised Machine Learning." IEEE Access 12: 8471-84.

- Alesawi, Kawthar Rasoul and Aymen Hasan Alawadi. (2025) "Enhancing Ddos Attack Classification Through Sdn and Machine Learning: A Feature Ranking Analysis." *Kufa Journal of Engineering* 16(2): 344-66.
- Alsafi, Walaa M. et al. (2024) "Diagnostic Accuracy of Hemoglobin Levels and Red Blood Cell Parameters for Iron Deficiency in Adolescents: A Cross-Sectional Study." *Global Pediatric Health* 11.
- Bui, Xuan-Nam, Hoang Nguyen and Phonepaserth Soukhanouvong. (2022) "Extra Trees Ensemble: A Machine Learning Model for Predicting Blast-Induced Ground Vibration Based on the Bagging and Sibling of Random Forest Algorithm." In, 643-52.
- Cappellini, Maria Domenica, Valeria Santini, Cecilia Braxs and Aryeh Shander. (2022) "Iron Metabolism and Iron Deficiency Anemia in Women." *Fertility and Sterility* 118(4): 607-14. <https://doi.org/10.1016/j.fertnstert.2022.08.014>.
- Dejene, Belayneh Endalamaw, Tesfamariam M. Abuhay and Dawit Shibabaw Bogale. (2022) "Predicting the Level of Anemia among Ethiopian Pregnant Women Using Homogeneous Ensemble Machine Learning Algorithm." *BMC Medical Informatics and Decision Making* 22(1): 1-11. <https://doi.org/10.1186/s12911-022-01992-6>.
- Farhan, Israa Dawed and Mayada Nazar Jabbar. (2014) "Evaluation of Some Blood Constituents in Dependence on Complete Blood Count of Patients in Baquba City, Diyala Province, Iraq." *Penambahan Natrium Benzoat Dan Kalium Sorbat (Antiinversi) Dan Kecepatan Pengadukan Sebagai Upaya Penghambatan Reaksi Inversi Pada Nira Tebu* (400).
- Furizal, Alfian Ma'arif and Dianda Rifaldi. (2023) "Application of Machine Learning in Healthcare and Medicine: A Review." *Journal of Robotics and Control (JRC)* 4(5): 621-31.
- Ghaddar, Bissan and Joe Naoum-Sawaya. (2017) "High Dimensional Data Classification and Feature Selection Using Support Vector Machines." *European Journal of Operational Research* 265.
- Gibore, Nyasiro S, Agatha F Ngowi, Mariam J Munyogwa and Mwanaisha M Ali. (2021) "Maternal and Pediatric Nutrition Dietary Habits Associated with Anemia in Pregnant Women Attending Antenatal Care Services." *Maternal and Pediatric Nutrition* 5(10): 1-8. <https://academic.oup.com/cdn/>.
- Göl, Mehmet et al. (2024) "Predicting Malnutrition-Based Anemia in Geriatric Patients Using Machine Learning Methods." *Journal of Evaluation in Clinical Practice* (July 2024).
- Hornyák, Olivér and László Barna Iantovics. (2023) "AdaBoost Algorithm Could Lead to Weak Results for Data with Certain Characteristics." *Mathematics* 11(8).

- Ifeanyi, E.O. and Uzoma, G.U. (2024) "Malaria, Anemia, and Iron Deficiency in Pregnancy: An Integrated Review." *Elite Journal of Scientific Research and Review* 2(61-76) https://www.researchgate.net/publication/382454638_Malaria_Anemia_and_Iron_Deficiency_in_Pregnancy_An_Integrated_Review?enrichId=rgreq-f550c670488e808156c760ae44ff2b16-XXX&enrichSource=Y292ZXJQYWdlOzM4MjQ1NDYzODtBUzoxMTQzMtI4MTI2MTg2NzUwNkAxNzIxNjY3Mzg0ND.
- Javaid, Mohd et al. (2022) "Significance of Machine Learning in Healthcare: Features, Pillars and Applications." *International Journal of Intelligent Networks* 3(February): 58-73. <https://doi.org/10.1016/j.ijin.2022.05.002>.
- Jiang, Liangxiao, Zhihua Cai, Harry Zhang and Dianhong Wang. (2013) "Naive Bayes Text Classifiers: A Locally Weighted Learning Approach." *Journal of Experimental & Theoretical Artificial Intelligence* 25.
- Kebede Kassaw, Abdulaziz et al. (2023) "The Application of Machine Learning Approaches to Determine the Predictors of Anemia among under Five Children in Ethiopia." *Scientific Reports* 13(1): 1-10. <https://doi.org/10.1038/s41598-023-50128-x>.
- Maasthi, Manasvi Jagadeesh et al. (2023) "Decision-Making Support System for Predicting and Eliminating Malnutrition and Anemia." *The Open Bioinformatics Journal* 16(1): 1-11.
- Meitei, A. Jiran, Akanksha Saini, Bibhuti Bhusan Mohapatra and Kh Jitenkumar Singh. (2022) "Predicting Child Anaemia in the North-Eastern States of India: A Machine Learning Approach." *International Journal of System Assurance Engineering and Management* 13(6): 2949-62. <https://doi.org/10.1007/s13198-022-01765-4>.
- Morales-Suárez-Varela, M et al. (2023) "Fish Intake in Pregnant Women and Its Impact on Maternal-Fetal Health Status." *Medicina de Familia. SEMERGEN* 49(6): 101996. <https://www.sciencedirect.com/science/article/pii/S1138359323000758>.
- Natekin, Alexey and Alois Knoll. (2013) "Gradient Boosting Machines, a Tutorial." *Frontiers in Neurorobotics* 7(DEC).
- Obeagu, Getrude Uzoma and Emmanuel Ifeanyi Obeagu. (2024) "Iron Deficiency Anaemia in Pregnancy: A Threat to Maternal and Fetal Wellbeing." *Lifeline Nursing and Health Sciences*. (November).
- Pan, Yunhui et al. (2023) "Machine Learning Prediction of Iron Deficiency Anemia in Chinese Premenopausal Women 12 Months after Sleeve Gastrectomy." *Nutrients* 15(15).

- Priyatno, Arif Mudi and Triyanna Widiyaningtyas. (2024) "A Systematic Literature Review: Recursive Feature Elimination Algorithms." *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)* 9(2): 196-207.
- Rana, Rakesh and Richa Singhal. (2015) "Chi-Square Test and Its Application in Hypothesis Testing." *Journal of the Practice of Cardiovascular Sciences* 1(1): 69.
- Salem, Nema and Sahar Hussein. (2019) "Data Dimensional Reduction and Principal Components Analysis." *Procedia Computer Science* 163: 292-99.
<https://doi.org/10.1016/j.procs.2019.12.111>.
- Salman, Hasan Ahmed, Ali Kalakech and Amani Steiti. (2024) "Random Forest Algorithm Overview." *Babylonian Journal of Machine Learning* 2024: 69-79.
- Shhatha, Amer Mohamed and Omar Ibrahim Alsaif. (2025) "Enhancing Cybersecurity Through Malware Detection Based on Machine Learning Technique." *Kufa Journal of Engineering* 16(3): 82-100.
- Shneen, Salam Waley, Rajaa Khalaf Gaber, Rasha Saad Salih and Suaad Makki Jiaad. (2025) "Artificial Neural Network (Ann) Based Proportional Integral Derivative (Pid) for Arm Rehabilitation Device." *Kufa Journal of Engineering* 16(1): 80-103.
- Shulman, Stanford T. (2014) "Saving Lives: Children's and Adults'." *Pediatric annals* 43(5): 163-64.
- Shweta, Nalluri and Sagar Dhanraj Pande. (2023) "Prediction of Anemia Using Various Ensemble Learning and Boosting Techniques." *EAI Endorsed Transactions on Pervasive Health and Technology* 9(1): 1-9.
- Skiena, Steven S. (2017) "Linear and Logistic Regression BT - The Data Science Design Manual." In ed. Steven S Skiena. Cham: Springer International Publishing, 267-302.
https://doi.org/10.1007/978-3-319-55444-0_9.
- Taunk, Kashvi, Sanjukta De, Srishti Verma and Aleena Swetapadma. (2019) "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification." *2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019 (May)*: 1255-60.
- Tesfaye, Solomon Hailemariam, Binyam Tariku Seboka and Daniel Sisay. (2024) "Application of Machine Learning Methods for Predicting Childhood Anaemia: Analysis of Ethiopian Demographic Health Survey of 2016." *PLoS ONE* 19(4 April): 1-14.
<http://dx.doi.org/10.1371/journal.pone.0300172>.
- Verulava, Tengiz and Inga Gogua. (2024) "Pregnant Mothers' Knowledge, Attitude, and Practice Towards Prevention of Iron Deficiency Anemia in Georgia." *Al-Kindy College Medical Journal* 20(3): 222-31.

Wu, Nuer et al. (2024) "The Global Burden of Maternal Disorders Attributable to Iron Deficiency Related Sub-Disorders in 204 Countries and Territories: An Analysis for the Global Burden of Disease Study." *Frontiers in Public Health* 12(September): 1-13.

Zemariam, Alemu Birara et al. (2024) "Employing Supervised Machine Learning Algorithms for Classification and Prediction of Anemia among Youth Girls in Ethiopia." *Scientific Reports* 14(1): 1-17. <https://doi.org/10.1038/s41598-024-60027-4>.

Zhou, HongFang et al. (2021) "A Feature Selection Algorithm of Decision Tree Based on Feature Weight." *Expert Systems with Applications* 164: 113842.