SINTECHCOM: Science, Technology and Communication Journal
Vol. 5, No. 2, February 2025, pp. 81-90, DOI: 10.59190/stc.v5i2.281

Analysis of anemia disease in Pakistan using logistic regression

Agnes Lee Si Tian, Kang Yuan Chin, Nik Azlin Nik Aziz, Norhaidah Mohd Asrah*
Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia, Muar 84600, Malaysia

ABSTRACT

Anemia, a global health issue affecting over two billion individuals, is characterized by a deficiency in red blood cells or hemoglobin, impairing oxygen transport in the body. Early detection of anemia is critical, particularly in resource-constrained regions. This research aims to develop a robust anemia prediction model leveraging machine learning techniques and non-invasive data inputs, including red, green, and blue (RGB) pixel intensities and hemoglobin levels. The research focuses on three objectives which are to analyze the relationship between predictor variables and anemia status using a correlation heatmap, to assess the contribution of RGB pixel intensities and hemoglobin levels in predicting anemia using feature importance analysis, and to identify significant predictors through recursive feature elimination. The model, developed using logistic regression, achieved an exceptional accuracy of 99.33% and an AUC score of 1.00. The hemoglobin level emerged as the most significant predictor, showing a strong negative correlation of -0.84 with anemia status. This approach not only enhances understanding of anemia's determinants but also provides actionable insights for healthcare professionals to devise targeted therapies and public health measures. Addressing these risk factors is vital to improving health outcomes, particularly for vulnerable populations at higher risk of anemia.

ARTICLE INFO

Article history:

Received Jan 16, 2025
Revised Feb 17, 2025
Accepted Feb 24, 2025

Keywords:

Anemia
Hemoglobin
Logistic Regression
Machine Learning
Public Health

* **Corresponding Author**
E-mail address: norhaida@uthm.edu.my

1. INTRODUCTION

Anemia is a condition characterized by a deficiency in red blood cells or hemoglobin, which reduces the body's capacity to carry oxygen efficiently. This leads to symptoms such as fatigue, dizziness, shortness of breath, and pale skin [1, 2]. There are several types of anemia, including iron-deficiency anemia, vitamin deficiency anemia, and aplastic anemia, with iron-deficiency anemia being the most prevalent [3]. The primary causes of anemia include nutritional deficiencies, chronic illnesses, genetic disorders, and infections [3, 4]. Vulnerable groups, such as children, menstruating adolescents, pregnant women, and the elderly, are at a higher risk of developing anemia [4, 5].

If left untreated, anemia can have severe consequences, including developmental delays in children, complications during pregnancy such as low birth weight and preterm delivery, organ damage due to oxygen deprivation, reduced cognitive and physical performance, and even an increased risk of mortality in severe cases [2, 3]. Early detection is crucial to mitigate these adverse effects, especially in low and middle income regions where anemia prevalence is high due to limited healthcare resources and widespread dietary deficiencies [4].

Recent advancements have introduced innovative methods for anemia detection and prediction, particularly through image-based and computational techniques. [6] explored the use of image analysis and processing algorithms to detect hemoglobin deficiencies, showcasing the potential of non-invasive diagnostic tools in clinical applications. Another research utilized RGB pixel analysis of fingertip video images to distinguish between low and high hemoglobin levels in sickle cell

patients. This research highlighted the effectiveness of image-based approaches in improving diagnostic accuracy and efficiency. Furthermore, it demonstrated the utility of logistic regression in identifying key anemia predictors, such as nutritional status and chronic illnesses, emphasizing its straightforward interpretability for healthcare practitioners [7].

Machine learning techniques, including logistic regression, have also been explored for anemia prediction. Logistic regression, in particular, has proven effective due to its ability to provide interpretable insights into key predictors, such as nutritional status and chronic illnesses. Studies have shown that logistic regression outperforms other algorithms like Naïve Bayes, Decision Tree, and XGBoost, achieving high accuracy and reliability, with one research reporting 95% accuracy and an AUC score of 0.99 on a dataset of 1,000 pathology records [8].

A comprehensive research on the determinants of anemia among women of reproductive age across Sub-Saharan African countries used multilevel mixed-effects modeling and ordered logistic regression analysis. The findings identified critical factors influencing anemia prevalence, highlighting socioeconomic, nutritional, and healthcare-related determinants, and emphasizing the importance of targeted interventions to mitigate anemia in vulnerable populations [9].

In the healthcare industry, logistic regression remains especially relevant due to its capability to provide interpretable coefficients that quantify the relationship between predictors and outcomes. This empowers healthcare providers to identify and prioritize key risk factors for anemia while understanding their contributions to the disease. Logistic regression is a statistical modeling technique commonly used for binary classification problems, where the outcome variable has two possible categories, such as "Yes" or "No." It estimates the probability of a specific event occurring by modeling the relationship between one or more predictor variables and the binary outcome using the logistic function [10].

This research aims to validate a robust anemia prediction model leveraging machine learning based on RGB pixel intensities and hemoglobin levels. The primary focus is on exploring non-invasive diagnostic techniques by leveraging RGB pixel intensities from images as predictors for anemia detection. The first objective of this research is to analyze the relationship between predictor variables and the response variable using a correlation heatmap. The second objective focuses on assessing the contribution of RGB pixel intensities and hemoglobin levels in predicting anemia status through feature importance analysis, including coefficients, odds ratios, and Recursive Feature Elimination (RFE). Finally, the third objective seeks to identify the most significant predictors for accurate anemia detection using recursive feature elimination.

2. MATERIALS AND METHOD

In conducting research on anemia prediction, it is important to have a comprehensive dataset that contains important variables influencing the condition. This section outlines the data sources and dataset used in this research, detailing the characteristics of the data and its relevance to the research objectives.

2.1. Data Sources and Dataset

The dataset used in this research is a secondary dataset obtained from Kaggle website, contributed by Humair M. (https://www.kaggle.com/datasets/humairmunir/anaemia-prediction-dataset/data) and contains 500 entries with six variables: Sex, % Red Pixel, % Green Pixel, % Blue Pixel, Hemoglobin Level (Hb), and Anemia Status. The predictor variables used in this research are x_1 (Percentage of Red Pixel), x_2 (Percentage of Green Pixel), x_3 (Percentage of Green Pixel), x_4 (Hemoglobin Level l) and x_5 (Sex), while the response variable is y (Anemia Status). This research should be able to analyze the contribution of RGB pixel intensities and hemoglobin levels in predicting anemia status. Table 1 shows the data description of the anemia's dataset [11].

2.2. Data Preprocessing

Data preprocessing is a crucial step in the machine learning pipeline, as it involves preparing raw data for analysis and modeling. Raw data collected from various sources often contains inconsistencies, missing values, noise, and irrelevant information that can affect the performance of machine learning algorithms. By applying data preprocessing techniques, data can be transformed into

a clean, structured, and meaningful format that is suitable for accurate and efficient analysis. This process not only ensures the quality and reliability of the data but it also enhances the overall performance of the machine learning models by enabling better feature extraction and improved algorithm efficiency [12].

Table 1. Data description of absence or presence of anemia.

Variable	Description	Type of variable
Y	Absence or presence of anemia: 0 represent absence 1 represent present	Qualitative
x_1	Percentage of Red Pixel	Quantitative
x_2	Percentage of Green Pixel	Quantitative
x_3	Percentage of Blue Pixel	Quantitative
x_4	Hemoglobin Level	Quantitative
x_5	Sex: 0 represent female 1 represent male	Qualitative

2.2.1. Data Cleaning

Data cleaning involves checking for missing values, as missing data can skew results and reduce model reliability. Missing data was handled using Simple Imputer from the scikit-learn library, replacing missing values with the mean of each column. Next, categorical variables were encoded numerically, with "Sex" mapped to 0 (Male) and 1 (Female), and "Anemia Status" to 0 (No) and 1 (Yes). Subsequently, all numeric features were standardized using Standard Scaler from scikit-learn which adjusts data to have a mean of 0 and a standard deviation of 1, ensuring all variables contribute equally to model performance [13]. The formula for Standard Scaler is shown in Equation (1):

$$Standard\ Scaler = \frac{X_i - \bar{X}}{std} \tag{1}$$

where, X_i is the original value of the feature for a particular instance in the dataset, $\bar{X}$ is the mean value of the feature across all instances in the dataset and std is the standard deviation of the feature, which measures how spread out the values are from the mean.

2.2.2. Exploratory Data Analysis

Exploratory data analysis (EDA) is the process of examining and summarizing datasets to uncover patterns, relationships, and anomalies using visual and statistical techniques. It serves as an initial step in the data analysis process, helping to understand the data's structure, test assumptions, and guide further analysis. EDA ensures data is clean and suitable for modelling, providing valuable insights that influence the choice of methods and models for deeper investigation [14].

Next, correlation heatmap is one of the analyses under EDA. In this research the correlation between features and Anemia Status was analyzed using Python's correlation heatmap. Correlation matrices are an essential tool of exploratory data analysis, it identifies highly correlated predictors, providing a clearer understanding of the relationships between variables [15]. Numerous numerical variables are included in a correlation plot, each of which is represented by a column. Each pair of variables' connection is shown by the rows. The relationship's strength is shown by the values in the cells where a positive relationship is indicated by a positive value, while a negative relationship is indicated by a negative value. Finding possible links between variables and determining their strength can be achieved using correlation heatmaps [16]. The correlation coefficient is calculated using the Pearson correlation formula as in Equation (2):

$$r = \frac{\Sigma(x_1 - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \tag{2}$$

where, r is the Pearson correlation coefficient, x_i and y_i are the individual sample points and $\bar{x}$ and $\bar{y}$ are the means of x and y samples.

2.2.3. Data Splitting

Data splitting is a fundamental technique in machine learning and data science used to partition a dataset into training set and testing set. This division ensures that the model is trained on one portion of the data and evaluated on another, preventing overfitting and promoting the development of a robust and generalizable model. The training set is used to fit the model, while the testing set is reserved for assessing its predictive performance on unseen data [17, 18]. In this research, the dataset was split into training (70%) and testing (30%). In this analysis data splitting was carried out using train_test_split from scikit-learn. This ensures that the model is trained on one portion of the data and evaluated on another.

2.3. Model Building

Model building aims at finding more realistic ways to describe the stochastic behavior observed in data, converting raw data into actionable insights or predictions. This process involves critical steps such as correlation analysis and feature selection to develop a robust predictive model. Python's scikit-learn library was used in these steps. Scikit-learn provides a comprehensive suite of tools for data preprocessing, model training, and evaluation, making it an ideal choice for building predictive models. It can efficiently implement various machine learning algorithms and utilize techniques such as cross-validation to ensure that the model is both accurate and reliable [19].

2.3.1. Logistic Regression

Logistic regression is a statistical method used to model the relationship between a dependent variable, which is categorical in nature, and one or more independent variables. Unlike linear regression, which predicts continuous outcomes, logistic regression predicts probabilities that lies between 0 and 1, making it particularly useful for binary classification problems [20]. The logistic regression model is based on the logistic function which maps any real-valued number into a value between 0 and 1. The general form of the logistic regression equation is expressed as in Equation (3):

$$logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (3)$$

where, p is the probability of the event occurring, β_0 is the intercept, β_1, β_2, …, β_k are the coefficients for the independent variables $X_1, X_2, \ldots, X_k$

In this research, the logistic regression model was used to predict the probability of an individual being anemic. The model was built using the Logistic Regression class from the scikit-learn library with the liblinear solver. The regression equation for this research can be expressed as in Equation (4):

$$logit(p) = + \beta_1 \cdot \text{Hb} + \beta_2 \cdot RGB_1 + \beta_3 \cdot RGB_2 + \beta_4 \cdot RGB_3 + \beta_5 \cdot Sex \quad (4)$$

where, p is the probability of being anemic. Hb is the hemoglobin level, RGB_1, RGB_2 and RGB_3 are color intensity values, and Sex is a binary variable indicating male or female.

2.3.2. Odd ratios

Feature importance was analyzed using the coefficients of the logistic regression model. In logistic regression, the absolute value of a coefficient indicates the relative contribution of the corresponding predictor to the model. Positive coefficients indicate an increase in the likelihood of the target event with an increase in the predictor, while negative coefficients indicate a decrease.

The coefficients were further transformed into odds ratios, allowing for an intuitive interpretation of the impact of each feature. The logistic regression model was trained using the fit method, and its performance was evaluated on the test dataset. Predictions for the test data were generated using the predict method, and model performance was assessed using accuracy and the odds

ratio. The odds ratio represents the change in odds for a one-unit increase in the predictor variable, holding all other variables constant [20]. The odds ratio is calculated using formula in Equation (5):

$$OR = e^{\beta} \tag{5}$$

where, β is the coefficient of the predictor.

2.3.3. Recursive Feature Elimination

Recursive feature elimination (RFE) is a wrapper-based feature selection technique that iteratively identifies the most predictive subset of features by leveraging a machine learning algorithm. In each iteration, RFE evaluates feature importance such as coefficients in logistic regression, removes the least significant feature, and retrains the model until the desired number of features is achieved. This process accounts for feature interactions and model-specific dependencies, reducing dimensionality, mitigating multicollinearity, and improving model interpretability and efficiency [21]. Incorporating cross-validation into the RFE process ensures that the selected features generalize well to unseen data, enhancing robustness and reducing overfitting, as noted in recent intelligent systems research [8, 22]. RFE's integration of feature selection and model training makes it highly adaptable and precise, making it suitable for applications requiring high accuracy, such as medical diagnostics and automated decision-making systems. By retaining only the most predictive features, RFE optimizes model performance and efficiency, ensuring improved reliability in real-world scenarios.

2.4. Model Validation and Evaluation

Model validation and evaluation are integral steps in assessing a model's performance and ensuring its reliability for real-world applications. Validation helps identify whether the model is underfitting or overfitting the data, while evaluation metrics provide a quantitative measure of its effectiveness. This section explores the methods and metrics employed to validate and assess the predictive accuracy of the model [23].

2.4.1. Performance Metrics

Model validation assesses a model's ability to generalize unseen data, ensuring robustness and reliability. Effective validation prevents overfitting and underfitting while enhancing predictive performance. Key metrics used for classification tasks include accuracy, precision, recall, and F1-score. These metrics provide insights into a model's diagnostic accuracy and areas for improvement [23]. Key metrics derived from the confusion matrix are defined as in equation 6 until 13:

$$True\ Positive\ Rate = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{6}$$

$$False\ Positive\ Rate = \frac{False\ Positive}{True\ Negative + False\ Positive} \tag{7}$$

$$True\ Negative\ Rate = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{8}$$

$$False\ Negative\ Rate = \frac{False\ Negative}{True\ Positive + False\ Negative} \tag{9}$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Predictions} \tag{10}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{11}$$

$$(Sensitivity)\ Recall = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{12}$$

$$F-1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{13}$$

These metrics were implemented in the research using Python's scikit-learn library, while confusion matrices were visualized using matplotlib. By integrating RFE as a feature selection method with model validation techniques, the research highlighted how efficient feature reduction and robust validation improve predictive accuracy and reduce computational overhead, particularly in high-dimensional datasets like those used in anemia disease prediction [8, 22].

2.4.2. Receiver Operating Characteristic

To further evaluate the model's discrimination capability, a receiver operating characteristic (ROC) curve was generated. The ROC curve is a graphical representation that illustrates the diagnostic ability of a binary classifier system by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) across various threshold settings. The area under the curve (AUC) quantifies the overall ability of the model to discriminate between positive and negative classes, with a value closer to 1 indicating better performance [24, 25].

In Python, the ROC curve is implemented using the roc_curve and auc functions from the sklearn.metrics module. First, probabilities for the positive class are predicted using the predict_proba method of the trained logistic regression model, and the probabilities for the positive class are extracted. These probabilities are used to compute the true positive rate (TPR), false positive rate (FPR), and thresholds using the roc_curve function. The auc function calculates the area under the ROC curve (AUC), providing a numerical value to quantify the model's discriminatory power.

3. RESULTS AND DISCUSSIONS

Each result is contextualized to highlight its significance and how it contributes to the overarching aim of predicting anemia status using logistic regression. The results and discussions section presents the findings of the analysis, providing insights into the model's performance and its implications. This section also discusses the relationships and patterns identified during exploratory data analysis and their relevance to the research objectives. Each result is contextualized to highlight its significance and how it contributes to the overarching aim of predicting anemia status using logistic regression.

3.1. Correlation Heatmap

The insights on the relationships between variables were explored using a correlation heatmap. This visual representation highlights the strength and direction of linear relationships among features.

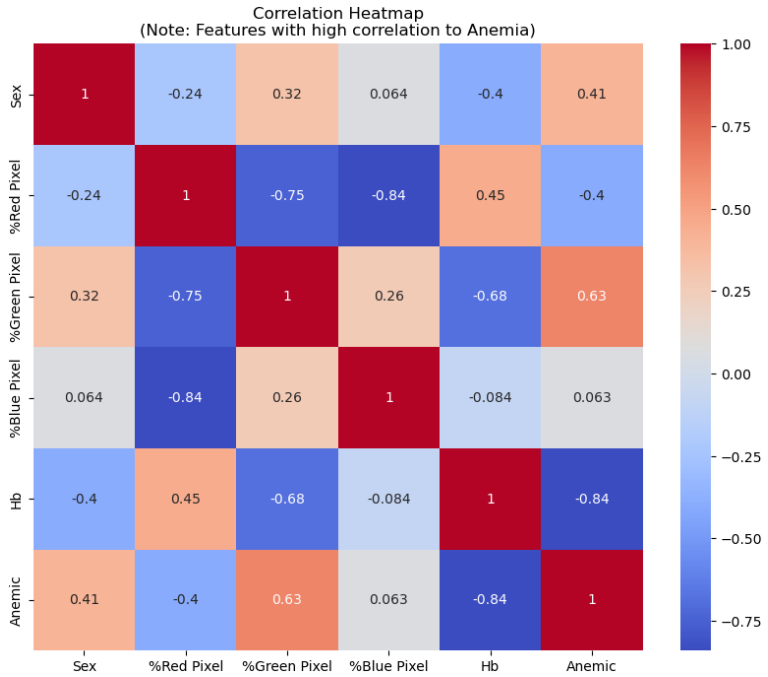

Figure 1. Correlation heatmap.

Based on Figure 1, there is a strong negative correlation (-0.84), indicating that as hemoglobin levels increase, the likelihood of anemia decreases significantly. On the other hand, moderate positive correlation (0.41), showing that females are more likely to be anemic compared to males. The percentage of red pixel has moderate negative correlation with anemia (-0.4). Meanwhile, percentage of green pixel also has moderate positive correlation with anemia (0.63). Besides that, percentage of blue pixel has minimal correlation (0.063), suggesting it has little to no linear relationship with anemia. In addition, there is a strong negative correlation between percentage of red pixel and percentage of green pixel (-0.75), suggesting multicollinearity. The percentage of red pixel is also positively correlated with hemoglobin level (0.45) while percentage of green pixel shows a negative correlation with hemoglobin level (-0.68). Sex and hemoglobin level have a moderate negative correlation (-0.4), indicating that females tend to have lower hemoglobin levels. Therefore, the key takeaways here are, hemoglobin is the most influential predictor, as evidenced by its strong correlation with anemia, and sex is moderately correlated with anemia, whereby reinforcing its importance as a predictor.

3.2. Odd Ratios

The impact of each predictor on the likelihood of anemia was assessed using odds ratios. These provide a quantitative measure of association between variables and outcomes.

Table 2. Odd ratios.

Variable	Odd Ratio
Percentage of red pixel	0.812874
Percentage of green pixel	2.590927
Percentage of blue pixel	0.718941
Hemoglobin level	0.004554
Sex	1.945302

Based on Table 2, a 1% increase in the red pixel percentage decreases the odds of anemia by approximately 18.7% (since $1-0.812874=0.1871261$). Meanwhile, a 1% increase in the green pixel percentage increases the odds of anemia by about 159.1% (since $2.590927-1=1.590927$). Furthermore, a 1% increase in the blue pixel percentage decreases the odds of anemia by about 28.1%. Moreover, a 1-unit increase in hemoglobin level decreases the odds of anemia by over 99.5%. This suggests that hemoglobin level is a highly significant predictor of anemia status. Lastly, females (coded as 1) have nearly double the odds (94.5% higher odds) of being anemic compared to males (coded as 0), holding all other variables constant.

3.3. Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) was applied to determine the most important features for predicting anemia status. This method iteratively removes less significant features to retain the most impactful ones.

The Recursive Feature Elimination (RFE) process identified Sex, %Red Pixel, %Green Pixel, %Blue Pixel, and Hb (Hemoglobin level) as the most important features for predicting anemia status. This selection means that these variables contribute the most to the model's predictive performance, while other features were deemed less influential and excluded from the final model.

3.4. Confusion Matrix

The confusion matrix was used to evaluate the classification model's performance by summarizing predictions against actual outcomes. Based on Figure 2, the model correctly predicted 82 individuals as "Not Anemic" who were actually not anemic (True Negative). However, the model incorrectly predicted 1 individual as "Anemic" who was actually not anemic (False Positive). The model did not make any false predictions of "Not Anemic" for individuals who were actually anemic (False Negative). The model correctly predicted 67 individuals as "Anemic" who were actually anemic (True Positive)

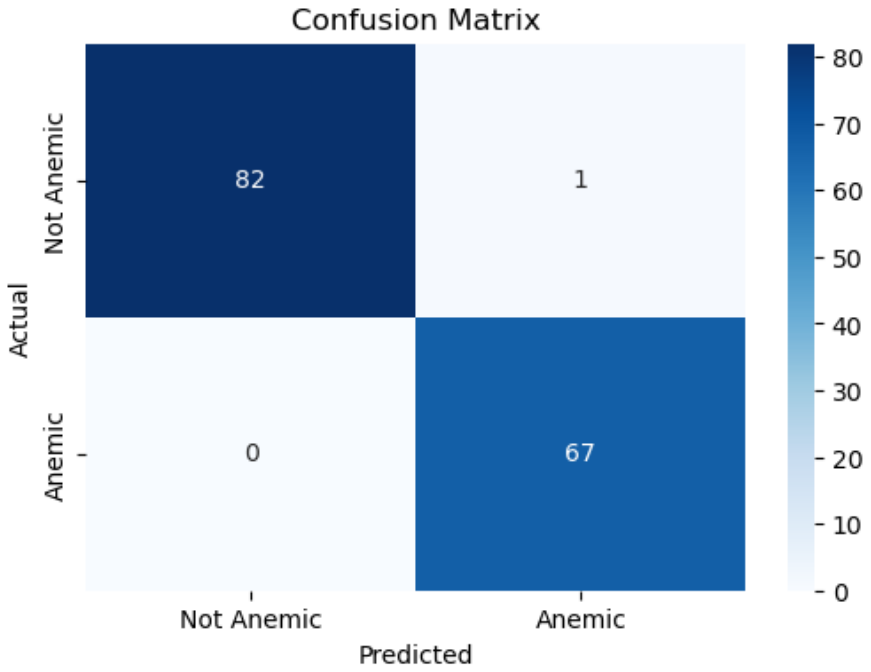

Figure 2. Confusion matrix.

3.5. Performance Metrics

Evaluation of the model's classification accuracy and effectiveness was performed using performance metrics.

Table 3. Classification report.

	Precision	Recall	F1-score	Support
0.0	1.00	0.99	0.99	83
1.0	0.99	1.00	0.99	67
Accuracy			0.99	150
Macro avg	0.99	0.99	0.99	150
Weighted avg	0.99	0.99	0.99	150

Based on Table 3, the classification report indicates exceptional performance of the model across all metrics. For Class 0 (Non - Anemic), the precision of 1.00 signifies that every prediction for this class is accurate, with no false positives. The recall of 0.99 reflects that nearly all actual instances of Class 0 are correctly identified, with minimal false negatives. The F1-score is 0.99, confirming strong performance. Similarly, for Class 1 (Anemic), the model achieves a precision of 0.99, indicating very few false positives, and a recall of 1.00, signifying all instances of anemia are correctly identified. The F1-score for Class 1 is also 0.99, highlighting the model's ability to balance precision and recall effectively for both classes. The overall accuracy of the model is 0.99, meaning it correctly predicts 99% of the 150 total samples, demonstrating excellent general performance.

3.6. Receiver Operating Characteristics (ROC) Curve

The model's ability to distinguish between classes was evaluated using the ROC curve. This graph plots the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. Based on Figure 3, the ROC curve reaching the top-left corner (TPR = 1 and FPR = 0) reflects perfect classification. The AUC of 1.00 confirms that the model is excellent, as it achieves maximum separability between the two classes. Hence, there is no overlap in predicted probabilities for the two classes and this model provides extremely precise threshold selection for classification. The model’s overall accuracy is 99.33%.

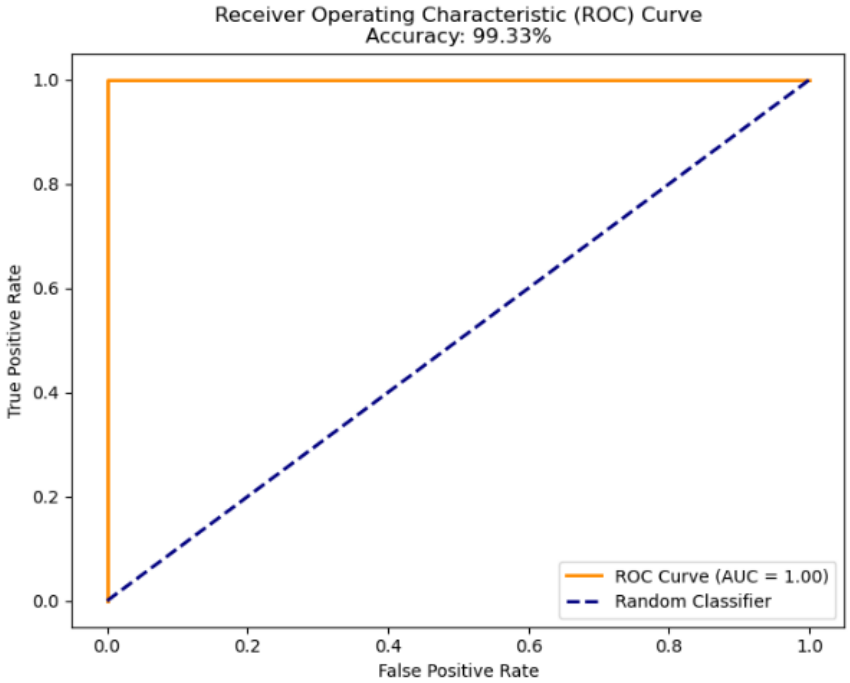

Figure 3. Receiver operating characteristics (ROC) curve.

4. CONCLUSION

In conclusion, the correlation heatmap provided a clear analysis of the relationships between predictor variables and anemia status, highlighting strong correlations such as hemoglobin levels' negative correlation with anemia and sex's moderate positive correlation with anemia. Feature importance analysis, including coefficients, odds ratios, and Recursive Feature Elimination (RFE), demonstrated the contributions of RGB pixel intensities and hemoglobin levels to anemia prediction. Hemoglobin levels emerged as the most significant predictor, while RGB pixel intensities especially red and green along with sex were also influential. Recursive Feature Elimination further validated these findings by identifying sex, hemoglobin levels, and RGB pixel intensities as the key predictors, enabling highly accurate anemia detection. These results, coupled with the model's exceptional performance metrics achieving an accuracy of 99.33% and an AUC score of 1.00, confirm that the research objectives were met, underscoring the model's robustness and practical applicability. The high performance of the model demonstrates its potential for non-invasive anemia detection paving the way for cost-effective, accessible diagnostic tools that could benefit healthcare providers and patients.

To further enhance the reliability and applicability of the anemia prediction model, it is recommended to standardize imaging conditions to mitigate variability caused by environmental factors such as lighting and camera quality. This would ensure more consistent and accurate measurements of RGB pixel intensities across different settings and populations. Additionally, expanding the dataset to include a more diverse range of participants is crucial. Incorporating variables such as age, body mass index (BMI), socioeconomic status, dietary habits, and medical history can also provide a more comprehensive understanding of the factors influencing anemia. This approach would not only improve the model's generalizability across different ethnic groups and demographics but also enhance its predictive power and real-world utility.

ACKNOWLEDGEMENTS

The authors would like to thank the Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia for its support.

REFERENCES

[1] Lights, V., & Seladi-Schulman, J. (2023). What is anemia?. *Healthline*.

[2] Mayo Clinic. (2023). *Anemia: Symptoms and causes*. Retrieved from https://www.mayoclinic.org/diseases-conditions/anemia/symptoms-causes/syc-20351360.
[3] HealthDirect. (2023). *Anemia: Causes, symptoms, and treatment*. Retrieved from https://www.healthdirect.gov.au/anaemia.
[4] World Health Organization. (2023). *Anaemia: Overview and statistics*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/anaemia.
[5] Deivita, Y., Syafruddin, S., Nilawati, U. A., Aminuddin, A., Burhanuddin, B., & Zahir, Z. (2021). Overview of Anemia; risk factors and solution offering. *Gaceta Sanitaria*, **35**, 235–241.
[6] Nalini, M., Sriharipriyan, P., Matheswaran, P., Vikash, K. V., & Sudharsan, M. (2023). Anemia detection through image analysis and image processing. *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, 1–4.
[7] Hasan, M. K., Sakib, N., Love, R. R., & Ahamed, S. I. (2017, October). RGB pixel analysis of fingertip video image captured from sickle cell patient with low and high level of hemoglobin. *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, 499–505.
[8] Rahman, M. M., Mojumdar, M. U., Shifa, H. A., Chakraborty, N. R., Stenin, N. P., & Hasan, M. A. (2024, February). Anemia disease prediction using machine learning techniques and performance analysis. *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, 1276–1282.
[9] Mare, K. U., Aychiluhm, S. B., Sabo, K. G., Tadesse, A. W., Kase, B. F., Ebrahim, O. A., Tebeje, T. M., Mulaw, G. F., & Seifu, B. L. (2023). Determinants of anemia level among reproductive-age women in 29 Sub-Saharan African countries: A multilevel mixed-effects modelling with ordered logistic regression analysis. *Plos one*, **18**(11), e0294992.
[10] Bobbitt, Z. (2020). Introduction to logistic regression. *Statology*.
[11] Munir, H. (2024). Anaemia prediction dataset. *Kaggle*.
[12] Singh, A. (2014). Data Preprocessing in Machine Learning: Steps, Techniques. *Applied Roots*.
[13] Liu, Q. (2022). Scale and standardize data with Scikit-learn. *GitHub*.
[14] Mahadevan, M. (2022). Step-by-Step Exploratory Data Analysis (EDA) using Python. *Analytics Vidhya*.
[15] Szabo, B. (2020). How to Create a Seaborn Correlation Heatmap in Python?. *Medium*.
[16] Kumar, A. (2022). Correlation concepts, matrix & heatmap using Seaborn. *Analytics Yogi*.
[17] Gillis, A. S. (2024). Data splitting. *TechTarget*.
[18] Muraina, I. (2022). Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts. *7th international Mardin Artuklu Scientific Research Conference*, 496–504.
[19] Geeks for Geeks. (2023). Model building for data analytics. Retrieved from https://www.geeksforgeeks.org/model-building-for-data-analytics.
[20] Srimaneekarn, N., Hayter, A., Liu, W., & Tantipoj, C. (2022). Binary response analysis using logistic regression in dentistry. *International Journal of Dentistry*, **2022**(1), 5358602.
[21] Kohavi, R. & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, **97**(1-2), 273–324.
[22] Jain, D. & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, **19**(3), 179–189.
[23] GeeksforGeeks. (2024). What is model validation and why is it important? Retrieved from https://www.geeksforgeeks.org/what-is-model-validation-and-why-is-it-important.
[24] Chugh, V. (2024). AUC and the ROC curve in machine learning. *Datacamp*.
[25] Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, **10**(4), 565–577.