Hematological Analysis for Anemia Diagnosis

by

Naga Sai Dhanya Veerepalli, Akhila Kudupudi, Suparna Mannava, Srija Anasuri

Abstract

This work investigates the use of machine learning models for haematological parameter-based anaemia diagnosis and classification. We hope to shed light on the patient distribution, important diagnostic characteristics, and predictive capabilities of different models by utilising a dataset that includes a variety of anaemia types and associated disorders. The most common conditions are Normocytic hypochromic anaemia (279 patients) and Normocytic normochromic anaemia (269 patients), followed by Iron deficiency anaemia (189 patients) and Healthy individuals (336 patients), according to an analysis of the diagnose distribution using Exploratory Data Analysis (EDA). Additionally, rare diseases such as leukaemia with thrombocytopenia (11 individuals) and macrocytic anaemia (18 patients) were found.

Haemoglobin (HGB) level research revealed important patterns among diagnoses. The average HGB values were highest in healthy persons and significantly lower in those with iron deficiency anaemia and macrocytic anaemia. Outliers offered distinctive characteristics for diagnosis, such as high HGB levels in leukaemia with thrombocytopenia. To predict the forms of anemia, machine learning models such as Random Forest, Logistic Regression, Multilayer Perceptron, and J48 Decision Trees were used. The Random Forest model, which identified important predictors such haemoglobin (HGB), mean corpuscular haemoglobin concentration (MCHC), and mean corpuscular volume (MCV), had the greatest accuracy of 98% among these.

These results highlight how crucial it is to combine machine learning and statistical insights in order to improve diagnosis accuracy. The models helped classify common and uncommon anaemia types by demonstrating a strong capacity to recognise trends across haematological characteristics. In the future, class imbalance will be addressed, real-time data will be included, and explainable AI approaches will be used to improve model interpretability. With its strong framework for enhancing anaemia detection and treatment in clinical settings, this study adds to the expanding field of data-driven diagnostics.

I. Introduction

Anemia is a major global health issue, affecting millions of people across various populations and age groups. It is defined by a shortage of red blood cells or hemoglobin, both essential for oxygen transport in the body. This deficiency can result in symptoms like fatigue, weakness, and reduced cognitive and physical performance. Anemia's severity can range from mild forms, such as Iron Deficiency Anemia, to serious hematological disorders like leukemia, which can pose life-threatening risks. The diversity of anemia subtypes highlights the urgent need for precise and timely diagnoses to ensure effective treatment and management.

Anemia includes multiple subtypes, each with distinct causes, clinical features, and treatment strategies. For example, Iron Deficiency Anemia typically arises from insufficient iron intake or chronic blood loss, while Macrocytic Anemia is often associated with vitamin B12 or folate deficiencies. In contrast, leukemia and its related forms, such as Leukemia with Thrombocytopenia, involve malignant processes with abnormal blood cell production. This variety necessitates advanced diagnostic tools and methods to accurately differentiate between these conditions.

This study utilizes hematological data from Complete Blood Count (CBC) tests, one of the most frequently conducted diagnostic tests in clinical settings. CBC tests yield extensive information about various blood components, making them essential for diagnosing and monitoring hematological disorders. Key parameters analyzed in this project include Hemoglobin (HGB), Mean Corpuscular Volume (MCV), and Mean Corpuscular Hemoglobin Concentration (MCHC), which are critical indicators of blood health. Deviations from their normal ranges often signify specific types of anemia. Additional parameters, such as Platelet Count (PLT), White Blood Cell Count (WBC), and Red Blood Cell Count (RBC), enhance the dataset, providing a comprehensive view of the hematological profile.

To improve the diagnostic process, this study employs machine learning models, particularly focusing on Random Forest, which is a robust and interpretable classification algorithm. By examining patterns in the CBC data, these models aim to identify key features that distinguish different types of anemia. This data-driven approach enhances the accuracy of anemia classification and reveals underlying relationships and trends within the data.

By employing data-driven techniques, this study aims to enhance anemia diagnosis, moving beyond traditional methods to more accurate and tailored approaches. The insights gained can assist healthcare practitioners in making better-informed decisions, optimizing treatment plans, and ultimately improving patient outcomes.

II. Methods

a. Data source

The dataset used in this project consists of 1,277 records derived from Complete Blood Count (CBC) tests, which are routinely performed to assess blood health. These records include manually diagnosed anemia cases categorized into subtypes such as Iron Deficiency Anemia, Macrocytic Anemia, and Leukemia, among others. The dataset was sourced from healthcare institutions, ensuring accuracy and reliability, as each record was validated by healthcare professionals.

b. Variables

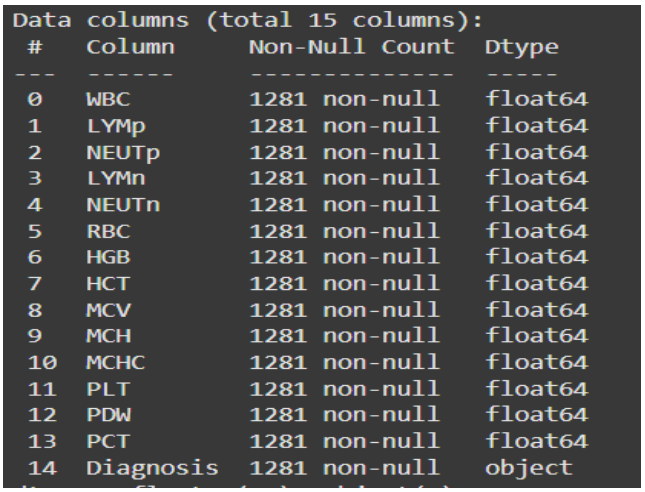
The hematological parameters analyzed include -

- HGB (Hemoglobin) - A critical measure of the blood's capacity to transport oxygen.
- MCV (Mean Corpuscular Volume) - Represents the average size of red blood cells, aiding in the classification of anemia types.
- MCHC (Mean Corpuscular Hemoglobin Concentration) - Indicates the concentration of hemoglobin within red blood cells, crucial for diagnosing certain anemias.
- PLT (Platelet Count) - Monitors blood clotting ability.
- WBC (White Blood Cell Count) - Assesses immune system health.
- RBC: Red blood cell count, responsible for oxygen transport.
- PDW: Platelet size variability.
- PCT: A marker for bacterial infections and sepsis risk.
- PlT: The number of platelets in the blood, involved in blood clotting.

The target variable represents the diagnosis, comprising both healthy and multiple anemia types.

Categories of the diagnosis -

- Healthy - Indicates no detectable health issues.
- Iron Deficiency Anemia - A condition marked by a lack of healthy red blood cells due to insufficient iron, impairing oxygen transport.
- Leukemia - A cancer affecting blood and bone marrow, resulting in the abnormal production of white blood cells.
- Leukemia with Thrombocytopenia - A condition involving leukemia accompanied by an abnormally low platelet count, increasing bleeding risk.
- Macrocytic Anemia - Characterized by red blood cells that are larger than normal, often linked to vitamin B12 or folate deficiency.
- Normocytic Hypochromic Anemia - Red blood cells are normal in size but have reduced hemoglobin content, causing pale cells and impaired oxygen transport.
- Normocytic Normochromic Anemia - A condition where red blood cells are normal in size and color but are present in insufficient numbers.
- Other Microcytic Anemia - A category of anemia where red blood cells are smaller than normal, excluding common causes like iron deficiency.
- Thrombocytopenia - Defined by a low platelet count, leading to a higher risk of bleeding and bruising.



```
Data columns (total 15 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   WBC        1281 non-null   float64
 1   LYMp       1281 non-null   float64
 2   NEUTp      1281 non-null   float64
 3   LYMn       1281 non-null   float64
 4   NEUTn      1281 non-null   float64
 5   RBC        1281 non-null   float64
 6   HGB        1281 non-null   float64
 7   HCT        1281 non-null   float64
 8   MCV        1281 non-null   float64
 9   MCH        1281 non-null   float64
 10  MCHC       1281 non-null   float64
 11  PLT        1281 non-null   float64
 12  PDW        1281 non-null   float64
 13  PCT        1281 non-null   float64
 14  Diagnosis  1281 non-null   object
```

Figure1. The columns in the dataset

The dataset was divided into training (869 records) and testing (359 records) subsets to evaluate model performance.

c. **Data preprocessing**

Data preprocessing is a crucial step in the data analysis and machine learning pipeline, involving various techniques to clean and prepare raw data for analysis. The primary goals of data preprocessing are to improve data quality, enhance model performance, and ensure that the data is suitable for the specific tasks at hand.

The first step of the project is to load the anemia dataset into a Pandas DataFrame. This dataset contained 1281 entries across 15 columns, which provided a structured view of various blood parameters relevant to anemia classification. Next, the crucial step is to check for any missing values in the dataset, as these could significantly impact the analysis and model performance. Fortunately, the check revealed that there were no missing values in any of the columns, indicating a complete dataset. This completeness is essential for ensuring reliable analysis and modeling. Following the assessment of missing values, the next step is to identify and handle outliers. Outliers can distort the results of machine learning models, so the Z-score method was employed to detect them. It was determined that certain entries deviated significantly from the mean.

As a result, the dataset was reduced from 1281 to 1277 entries by removing these outlier records, enhancing the quality of the data. The next step involved checking for duplicate entries within the dataset. Duplicates can lead to overfitting and biased results, so it was important to address them. The analysis revealed 49 duplicate rows, which were subsequently removed. This action further refined the dataset, leaving 1228 unique entries that better represent the population under study.

After cleaning the dataset of outliers and duplicates, the filtered data is saved for future use in modeling. This step ensured that the integrity of the preprocessing efforts was preserved, allowing for easy retrieval of the refined dataset. With a clean dataset in hand, feature scaling was performed to standardize the range of the independent variables. This process is vital as it ensures that no single feature dominates the model due to its scale, leading to improved model performance and faster convergence. Finally, the dataset is split into training and testing sets, which is crucial for evaluating the model's performance on unseen data. The division was made so that 70% of the data would be used for training and 30% for testing, establishing a robust framework for model validation.

In summary, the preprocessing steps transformed the initial dataset of 1281 entries into a well-prepared set of 1228 unique entries, free from outliers and duplicates, with scaled features ready for effective machine learning modeling. This thorough preprocessing ensures that the data is clean, consistent, and suitable for building predictive models, ultimately leading to better performance in classifying different types of anemia.

```
SELECT TOP (1000) [WBC]
      ,[LYMp]
      ,[NEUTp]
      ,[LYMn]
      ,[NEUTn]
      ,[RBC]
      ,[HGB]
      ,[HCT]
      ,[MCV]
      ,[MCH]
      ,[MCHC]
      ,[PLT]
      ,[PDW]
      ,[PCT]
      ,[Diagnosis]
  FROM [anemia].[dbo].[anemia_filtered];
-- Create Training and Test Datasets with Stratification
WITH StratifiedData AS (
    SELECT *,
           NTILE(10) OVER (PARTITION BY Diagnosis ORDER BY NEWID()) AS Bucket
    FROM [anemia].[dbo].[anemia_filtered]
)
-- Create Training Data
SELECT *
INTO TrainData
FROM StratifiedData
WHERE Bucket <= 7;  -- 70% for training

-- Create Test Data
WITH StratifiedData AS (
    SELECT *,
           NTILE(10) OVER (PARTITION BY Diagnosis ORDER BY NEWID()) AS Bucket
    FROM [anemia].[dbo].[anemia_filtered]
)
SELECT *
INTO TestData
FROM StratifiedData
WHERE Bucket > 7;   -- 30% for testing
```

Figure 2: Data stratification and split

d. Data Modeling

1. Model Training

RANDOM FOREST

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         852               98.0437 %
Incorrectly Classified Instances        17                1.9563 %
Kappa statistic                          0.9759
Mean absolute error                      0.0234
Root mean squared error                  0.0765
Relative absolute error                 12.9592 %
Root relative squared error             25.4793 %
Total Number of Instances              869
```

Figure 3: Random forest metrics

The Random Forest model did remarkably well in classifying various anaemia kinds, with an overall accuracy of 98.04% and 852 properly categorised occurrences out of 869. Beyond what would be predicted by chance, the model's predictions and the actual labels have almost perfect agreement, as indicated by the Kappa statistic of 0.9759. Additionally, the model's accuracy in predicting anemia types is demonstrated by its low error rates, which include a Mean Absolute Error of 0.0234.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.996	0.006	0.983	0.996	0.989	0.985	1.000	0.999	Healthy
	0.992	0.003	0.985	0.992	0.989	0.986	1.000	1.000	Iron deficiency anemia
	0.969	0.001	0.969	0.969	0.969	0.968	1.000	1.000	Leukemia
	0.875	0.003	0.700	0.875	0.778	0.780	0.999	0.935	Leukemia with thrombocytopenia
	0.692	0.000	1.000	0.692	0.818	0.830	0.998	0.939	Macrocytic anemia
	0.989	0.001	0.995	0.989	0.992	0.990	1.000	0.999	Normocytic hypochromic anemia
	0.989	0.006	0.978	0.989	0.983	0.979	1.000	0.998	Normocytic normochromic anemia
	0.923	0.000	1.000	0.923	0.960	0.959	0.999	0.985	Other microcytic anemia
	0.961	0.002	0.961	0.961	0.961	0.958	1.000	0.996	Thrombocytopenia
Weighted Avg.	0.980	0.004	0.981	0.980	0.980	0.978	1.000	0.997	

Figure 4: Class metrics

With high precision, recall, and F1-scores for the majority of categories, the model demonstrated consistent performance across all anemia classes. While disorders like "Macrocytic anaemia" and "Leukaemia with thrombocytopenia," which are more difficult to distinguish, showed somewhat worse recall, classes like "Healthy" and "Iron deficiency anaemia" performed almost flawlessly. The overall measures, which include an average weighted F1-score of 0.980, demonstrate balanced and strong performance across all categories in spite of these small misclassifications.

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	<-- classified as
226	1	0	0	0	0	0	0	0	a = Healthy
1	129	0	0	0	0	0	0	0	b = Iron deficiency anemia
1	0	31	0	0	0	0	0	0	c = Leukemia
0	0	1	7	0	0	0	0	0	d = Leukemia with thrombocytopenia
0	0	0	0	9	1	3	0	0	e = Macrocytic anemia
1	0	0	0	0	187	1	0	0	f = Normocytic hypochromic anemia
0	0	0	2	0	0	178	0	0	g = Normocytic normochromic anemia
0	1	0	0	0	0	0	36	2	h = Other microcytic anemia
1	0	0	1	0	0	0	0	49	i = Thrombocytopenia

Figure 5: Confusion matrix

The confusion matrix confirms the model's effectiveness, with most predictions falling along the diagonal, representing correct classifications. Misclassifications were minimal and typically occurred between similar conditions, such as "Leukemia" and "Leukemia with thrombocytopenia," reflecting the inherent challenges in distinguishing closely related anemia types.

LOGISTIC REGRESSION

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         720               82.8539 %
Incorrectly Classified Instances       149               17.1461 %
Kappa statistic                          0.7891
Mean absolute error                      0.0617
Root mean squared error                  0.1822
Relative absolute error                 34.1747 %
Root relative squared error             60.6623 %
Total Number of Instances              869
```

Figure 6: Logistic regression metrics

The Logistic Regression model achieved an accuracy of 82.85%, correctly classifying 720 out of 869 instances. While this performance is reasonable, it is lower compared to models like Random Forest. The Kappa statistic of 0.7891 indicates substantial agreement between predictions and actual labels but also highlights some limitations in capturing the complexity of the dataset. The error metrics, such as Mean Absolute Error (0.0617) and Root Mean Squared Error (0.1822), further reflect the challenges in accurately predicting all classes, especially those with subtle differences.

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.903    0.036    0.899      0.903    0.901      0.866    0.971     0.933     Healthy
                 0.877    0.015    0.912      0.877    0.894      0.876    0.964     0.920     Iron deficiency anemia
                 0.625    0.011    0.690      0.625    0.656      0.644    0.898     0.579     Leukemia
                 0.625    0.009    0.385      0.625    0.476      0.484    0.988     0.325     Leukemia with thrombocytopenia
                 0.538    0.007    0.538      0.538    0.538      0.531    0.989     0.501     Macrocytic anemia
                 0.778    0.049    0.817      0.778    0.797      0.742    0.917     0.837     Normocytic hypochromic anemia
                 0.883    0.042    0.846      0.883    0.864      0.828    0.956     0.871     Normocytic normochromic anemia
                 0.692    0.019    0.628      0.692    0.659      0.643    0.921     0.480     Other microcytic anemia
                 0.706    0.017    0.720      0.706    0.713      0.695    0.911     0.626     Thrombocytopenia
Weighted Avg.    0.829    0.033    0.832      0.829    0.829      0.796    0.947     0.834
```

Figure 7: Class metrics

For classes like "Healthy" and "Iron deficiency anaemia," Logistic Regression works well, with high precision and recall values above 0.89, according to class-level metrics. More complicated classes like "Leukaemia" and "Macrocytic anaemia," where precision and recall fall to 0.69 and 0.54 respectively, are difficult for the model to handle. This variability suggests that for rarer or more complicated anaemia types, the complex interactions or overlaps between features cannot be well captured by Logistic Regression, a linear model. Lower PRC areas indicate difficulties in striking a balance between accuracy and recall, while ROC areas above 0.9 for the majority of classes indicate the model can still distinguish between classes quite well.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
 205   1   1   1   0  10   5   3   1 |   a = Healthy
   0 114   0   2   0   9   0   4   1 |   b = Iron deficiency anemia
   6   1  20   0   0   2   2   1   0 |   c = Leukemia
   0   0   1   5   0   1   0   0   1 |   d = Leukemia with thrombocytopenia
   0   0   0   0   7   2   4   0   0 |   e = Macrocytic anemia
   5   4   6   1   4 147  13   5   4 |   f = Normocytic hypochromic anemia
   7   0   0   1   0   5 159   2   6 |   g = Normocytic normochromic anemia
   1   4   0   0   0   1   5  27   1 |   h = Other microcytic anemia
   4   1   1   3   2   3   0   1  36 |   i = Thrombocytopenia
```

Figure 8: Confusion matrix

The majority of categories fall along the diagonal, indicating accurate predictions, according to the confusion matrix. Notable misclassifications do exist, nevertheless, in overlapping groupings like "Leukaemia" and "Leukaemia with thrombocytopenia."

For this dataset, logistic regression performs rather well overall. Because it cannot model non-linear interactions, it suffers with complicated anaemia types but does well for simpler or linearly separable classes. More sophisticated models like Random Forest or Gradient Boosting are better suited to identifying the complex patterns in this dataset, even while its interpretability and efficiency make it a good choice for simple analysis.

MULTILAYER PERCEPTRON

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         656               75.4891 %
Incorrectly Classified Instances       213               24.5109 %
Kappa statistic                          0.6958
Mean absolute error                      0.0751
Root mean squared error                  0.198
Relative absolute error                 41.5449 %
Root relative squared error             65.9144 %
Total Number of Instances              869
```

Figure 9: Multilayer perceptron metrics

By properly categorising 656 out of 869 instances, the Multilayer Perceptron (MLP) model assessed using 10-fold cross-validation obtained an accuracy of 75.49%. Compared to the 5-fold examination, this is a moderate improvement, suggesting that the higher folds improved generalisation. While the Root Mean Squared Error (0.198) and Mean Absolute Error (0.0751) show respectable performance with occasional misclassifications, the Kappa statistic of 0.6958 indicates strong agreement between the model's predictions and the actual labels.

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.899    0.034    0.903      0.899    0.901      0.866    0.982     0.961     Healthy
               0.731    0.054    0.704      0.731    0.717      0.666    0.951     0.786     Iron deficiency anemia
               0.719    0.011    0.719      0.719    0.719      0.708    0.988     0.720     Leukemia
               0.375    0.002    0.600      0.375    0.462      0.471    0.792     0.428     Leukemia with thrombocytopenia
               0.077    0.000    1.000      0.077    0.143      0.275    0.765     0.175     Macrocytic anemia
               0.630    0.110    0.613      0.630    0.621      0.514    0.883     0.656     Normocytic hypochromic anemia
               0.856    0.073    0.755      0.856    0.802      0.749    0.953     0.871     Normocytic normochromic anemia
               0.333    0.011    0.591      0.333    0.426      0.425    0.859     0.447     Other microcytic anemia
               0.863    0.007    0.880      0.863    0.871      0.863    0.967     0.810     Thrombocytopenia
Weighted Avg.  0.755    0.057    0.756      0.755    0.747      0.697    0.939     0.792
```

Figure 10: Class metrics

With a high precision of 0.903 and recall of 0.899, the model did well for "Healthy" instances at the class level. With F1-scores above 0.7, classes such as "Iron deficiency anaemia" and "Normocytic normochromic anaemia" also exhibit balanced metrics. Rarer and more complicated classes, such "Macrocytic anaemia," whose precision and recall values were as low as 0.143, were difficult for the model to handle. The model's moderate ability to balance precision and recall across all anaemia kinds is indicated by the overall weighted average F1-score of 0.747.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
 204   0   2   0   0  11  10   0   0 |   a = Healthy
   0  95   0   0   0  34   0   1   0 |   b = Iron deficiency anemia
   3   3  23   0   0   2   1   0   0 |   c = Leukemia
   0   0   3   3   0   0   0   0   2 |   d = Leukemia with thrombocytopenia
   0   1   0   0   1   5   6   0   0 |   e = Macrocytic anemia
   9  31   1   0   0 119  22   6   1 |   f = Normocytic hypochromic anemia
   4   1   2   2   0  13 154   2   2 |   g = Normocytic normochromic anemia
   1   3   1   0   0   9  11  13   1 |   h = Other microcytic anemia
   5   1   0   0   0   1   0   0  44 |   i = Thrombocytopenia
```

Figure 11: Confusion matrix

Although misclassifications happen in closely related classes, the confusion matrix verifies that the majority of classifications line up along the diagonal, indicating accurate predictions. As an example, "Leukaemia" was frequently mislabeled as "Leukaemia with

thrombocytopenia," underscoring the difficulty in differentiating between two overlapping conditions. The MLP model performs quite well overall, especially for more prevalent forms of anemia. Its inability to handle more uncommon or overlapping cases, however, points to the necessity of further feature engineering or model optimisation. The MLP is nevertheless a useful technique for managing fairly complex datasets like this one and comprehending feature interactions in spite of its drawbacks.

J48 DECISION TREES

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         857               98.6191 %
Incorrectly Classified Instances        12                1.3809 %
Kappa statistic                          0.983
Mean absolute error                      0.0039
Root mean squared error                  0.0542
Relative absolute error                  2.1314 %
Root relative squared error             18.0562 %
Total Number of Instances              869
```

Figure 12: J48 Decision tree metrics

The J48 decision tree classifier accurately classified 857 out of 869 cases, achieving an excellent accuracy of 98.62%. This illustrates how well the model can manage the categorisation of anaemia types. The trustworthiness of this method is further highlighted by the nearly perfect agreement between the model's predictions and the genuine labels, as indicated by the Kappa statistic of 0.983. The model's accuracy is confirmed by the extraordinarily low error measurements, such as Mean Absolute Error (0.0039) and Root Mean Squared Error (0.0542).

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.003    0.991      1.000    0.996      0.994    0.997     0.982     Healthy
               0.992    0.001    0.992      0.992    0.992      0.991    0.994     0.977     Iron deficiency anemia
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Leukemia
               0.750    0.002    0.750      0.750    0.750      0.748    0.874     0.721     Leukemia with thrombocytopenia
               0.923    0.002    0.857      0.923    0.889      0.888    0.960     0.792     Macrocytic anemia
               0.989    0.000    1.000      0.989    0.995      0.993    0.993     0.992     Normocytic hypochromic anemia
               0.983    0.004    0.983      0.983    0.983      0.979    0.995     0.979     Normocytic normochromic anemia
               0.923    0.000    1.000      0.923    0.960      0.959    0.985     0.966     Other microcytic anemia
               1.000    0.002    0.962      1.000    0.981      0.980    1.000     0.993     Thrombocytopenia
Weighted Avg.  0.986    0.002    0.986      0.986    0.986      0.984    0.993     0.978
```

Figure 13: Class metrics

For the majority of classes, the class-level performance metrics show excellent precision, recall, and F1-scores. For example, with F1-scores of 0.996 and 0.992, respectively, the "Healthy" and "Iron deficiency anemia" classes obtained nearly flawless precision and recall. With F1-scores of 0.750 and 0.857, even the more difficult classes, such as "Leukaemia with thrombocytopenia" and "Macrocytic anaemia," have acceptable metrics. The model's overall consistency and robustness across all anemia types is highlighted by the weighted average F1-score of 0.986.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
 227   0   0   0   0   0   0   0   0 |   a = Healthy
   1 129   0   0   0   0   0   0   0 |   b = Iron deficiency anemia
   0   0  32   0   0   0   0   0   0 |   c = Leukemia
   0   0   0   6   0   0   2   0   0 |   d = Leukemia with thrombocytopenia
   0   0   0   0  12   0   1   0   0 |   e = Macrocytic anemia
   1   0   0   0   1 187   0   0   0 |   f = Normocytic hypochromic anemia
   0   0   0   2   1   0 177   0   0 |   g = Normocytic normochromic anemia
   0   1   0   0   0   0   0  36   2 |   h = Other microcytic anemia
   0   0   0   0   0   0   0   0  51 |   i = Thrombocytopenia
```

Figure 14: Confusion matrix

The confusion matrix shows that accurate classifications are indicated by the majority of predictions aligning along the diagonal. Misclassifications are rare; only a small number of closely related anaemia types have been misclassified. For instance, there is no confusion between "leukaemia" and "macrocytic anaemia" with other groups. All things considered, the J48 decision tree performs exceptionally well in identifying the different forms of anemia with great accuracy and precision. It is a strong model for classifying medical data since it can efficiently handle both common and uncommon classes. The scatter plot's potential to give medical professionals useful insights is confirmed by the few misclassifications and distinct clusters.

2. TEST DATASET RESULT

RANDOM FOREST

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         344               95.8217 %
Incorrectly Classified Instances        15                4.1783 %
Kappa statistic                          0.948
Mean absolute error                      0.0346
Root mean squared error                  0.1031
Relative absolute error                 19.2656 %
Root relative squared error             34.4503 %
Total Number of Instances              359
```

Figure 15: Random forest metrics

With stratified cross-validation to guarantee that the data is evenly distributed across classes, the findings are for the Random Forest model applied to the test set for anaemia classification. With only 15 cases incorrectly classified, the model accurately classifies 344 out of 359 occurrences, achieving 95.82% accuracy. The Kappa number of 0.948, which shows great agreement between predictions and actual data, reflects this high accuracy.

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.990    0.011    0.969      0.990    0.979      0.972    0.999     0.998     Healthy
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iron deficiency anemia
                 0.833    0.000    1.000      0.833    0.909      0.910    0.999     0.979     Leukemia
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.989     0.344     Leukemia with thrombocytopenia
                 0.000    0.000    ?          0.000    ?          ?        0.986     0.516     Macrocytic anemia
                 0.975    0.004    0.987      0.975    0.981      0.976    0.999     0.996     Normocytic hypochromic anemia
                 0.987    0.014    0.949      0.987    0.967      0.959    0.997     0.991     Normocytic normochromic anemia
                 0.867    0.003    0.929      0.867    0.897      0.893    0.999     0.983     Other microcytic anemia
                 0.952    0.015    0.800      0.952    0.870      0.864    0.999     0.978     Thrombocytopenia
Weighted Avg.    0.958    0.008    ?          0.958    ?          ?        0.999     0.985
```

Figure 16: Class metrics

With precision, recall, and F-measure values near or equal to 1, the Random Forest model excels for classes like "Healthy" and "Iron deficiency anaemia," according to the Detailed Accuracy By Class. Certain groups, such as "Leukaemia with thrombocytopenia" and "Macrocytic anaemia," provide difficulties in differentiation, with comparatively low precision and recall, indicating areas in need of improvement. The model's overall good performance is further demonstrated by the Weighted Average metrics, which indicate near-perfect separability with an F-measure of 0.958 and ROC Area of 0.999.

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 95  0  0  0  0  0  0  0  1 |  a = Healthy
  0 54  0  0  0  0  0  0  0 |  b = Iron deficiency anemia
  2  0 10  0  0  0  0  0  0 |  c = Leukemia
  0  0  0  0  0  0  0  0  3 |  d = Leukemia with thrombocytopenia
  0  0  0  0  0  1  2  0  0 |  e = Macrocytic anemia
  0  0  0  0  0 78  1  1  0 |  f = Normocytic hypochromic anemia
  0  0  0  1  0  0 74  0  0 |  g = Normocytic normochromic anemia
  0  0  0  0  0  0  1 13  1 |  h = Other microcytic anemia
  1  0  0  0  0  0  0  0 20 |  i = Thrombocytopenia
```

Figure 17: Confusion matrix

Additional information about certain areas of misclassification is provided by the Confusion Matrix. Examples of highly accurate classifications include "Healthy" and "Iron deficiency anaemia," yet there is considerable misunderstanding between closely related illnesses like "Normocytic hypochromic anaemia" and "Normocytic normochromic anaemia." This misunderstanding may result from minor distinctions or overlapping characteristics between these classifications. Although there may be room for improvement in feature distinction for closely related classes, the Random Forest model performs well overall and is ideally suited for the categorisation of anaemia.

LOGISTIC REGRESSION

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         291               81.0585 %
Incorrectly Classified Instances        68               18.9415 %
Kappa statistic                          0.766
Mean absolute error                      0.0477
Root mean squared error                  0.2021
Relative absolute error                 26.5592 %
Root relative squared error             67.5171 %
Total Number of Instances              359
```

Figure 18: Logistic regression metrics

The results of the anaemia classification test using the logistic regression model show an overall classification accuracy of 81.05%. Although there is potential for improvement, this shows that the model can distinguish between the nine forms of anemia with reasonable accuracy. The predicted and actual classifications show a moderate level of agreement, as indicated by the Kappa score of 0.766. The model's average and squared error magnitudes are shown by metrics such as the mean absolute error (0.0477) and root mean squared error (0.2021), while the relative error metrics reveal forecast variability.

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.865    0.049    0.865      0.865   0.865      0.815    0.975     0.901     Healthy
               0.870    0.016    0.904      0.870   0.887      0.867    0.977     0.922     Iron deficiency anemia
               0.583    0.014    0.583      0.583   0.583      0.569    0.939     0.512     Leukemia
               0.000    0.022    0.000      0.000   0.000      -0.014   0.922     0.065     Leukemia with thrombocytopenia
               0.333    0.000    1.000      0.333   0.500      0.576    0.890     0.403     Macrocytic anemia
               0.763    0.043    0.836      0.763   0.797      0.744    0.887     0.740     Normocytic hypochromic anemia
               0.867    0.042    0.844      0.867   0.855      0.817    0.933     0.797     Normocytic normochromic anemia
               0.800    0.017    0.667      0.800   0.727      0.718    0.987     0.712     Other microcytic anemia
               0.714    0.021    0.682      0.714   0.698      0.679    0.974     0.644     Thrombocytopenia
Weighted Avg.  0.811    0.037    0.825      0.811   0.816      0.778    0.945     0.799
```

Figure 19: Class metrics

There are conflicting findings when examining the performance at the class level. The model yields good precision, recall, and F1-scores for common anaemia kinds, including healthy and iron deficiency anaemia, indicating reliable predictions for these categories. However, recall and F1-scores are much lower for more complicated or uncommon categories, such as leukaemia with thrombocytopenia and macrocytic anaemia. This finding is corroborated by the confusion matrix, which reveals frequent misclassifications, especially for these difficult classes. For instance, leukaemia with thrombocytopenia shows a great deal of overlap with other classifications, and macrocytic anaemia is frequently mislabeled as normocytic hypochromic anaemia.

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 83  0  5  0  0  2  2  1  3 |  a = Healthy
  0 47  0  0  0  4  0  3  0 |  b = Iron deficiency anemia
  5  0  7  0  0  0  0  0  0 |  c = Leukemia
  0  0  0  0  0  0  0  0  3 |  d = Leukemia with thrombocytopenia
  0  0  0  0  1  0  2  0  0 |  e = Macrocytic anemia
  3  3  0  3  0 61  7  2  1 |  f = Normocytic hypochromic anemia
  2  0  0  2  0  6 65  0  0 |  g = Normocytic normochromic anemia
  0  2  0  0  0  0  1 12  0 |  h = Other microcytic anemia
  3  0  0  3  0  0  0  0 15 |  i = Thrombocytopenia
```

Figure 20: Confusion matrix

MULTILAYER PERCEPTRON

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         285               79.3872 %
Incorrectly Classified Instances        74               20.6128 %
Kappa statistic                          0.7429
Mean absolute error                      0.0589
Root mean squared error                  0.1901
Relative absolute error                 32.8091 %
Root relative squared error             63.5079 %
Total Number of Instances              359
```

Figure 21: Multilayer perceptron metrics

With an overall accuracy of 79.39%, the Multilayer Perceptron (MLP) model's results on the test dataset show that it can successfully classify anaemia kinds. 285 cases were correctly diagnosed and 74 cases were misclassified, according to the confusion matrix, which shows the distribution of correctly and erroneously classified instances among the nine anaemia groups. The model's resilience in managing the dataset is highlighted by the Kappa score of 0.7429, which shows a high degree of agreement between the anticipated and actual classifications.

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.885    0.053    0.859      0.885    0.872      0.824    0.972     0.941     Healthy
               0.852    0.033    0.821      0.852    0.836      0.807    0.978     0.879     Iron deficiency anemia
               0.417    0.020    0.417      0.417    0.417      0.396    0.936     0.413     Leukemia
               0.000    0.008    0.000      0.000    0.000      -0.008   0.570     0.012     Leukemia with thrombocytopenia
               0.000    0.000    ?          0.000    ?          ?        0.774     0.034     Macrocytic anemia
               0.800    0.075    0.753      0.800    0.776      0.709    0.909     0.786     Normocytic hypochromic anemia
               0.827    0.039    0.849      0.827    0.838      0.796    0.934     0.856     Normocytic normochromic anemia
               0.467    0.003    0.875      0.467    0.609      0.629    0.928     0.690     Other microcytic anemia
               0.762    0.021    0.696      0.762    0.727      0.710    0.975     0.696     Thrombocytopenia
Weighted Avg.  0.794    0.046    ?          0.794    ?          ?        0.943     0.822
```

Figure 22: Class metrics

The performance of the MLP model for each type of anemia is shown by the detailed accuracy by class. Different classes had different True Positive (TP) rates; the "Healthy" and "Iron deficiency anaemia" groups had comparatively high TP rates of 0.885 and 0.852, respectively. Some categories, such as "Leukaemia with thrombocytopenia," show lower TP rates, nevertheless, suggesting that the model has trouble finding cases in these groups. A possible difficulty in resolving class imbalance is demonstrated by the fact that precision and memory scores are greater for dominating classes but decrease for under-represented ones.

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 85  1  3  1  0  1  3  0  2 |  a = Healthy
  0 46  0  0  0  7  0  1  0 |  b = Iron deficiency anemia
  4  0  5  1  0  1  1  0  0 |  c = Leukemia
  0  0  1  0  0  1  0  0  1 |  d = Leukemia with thrombocytopenia
  0  0  0  0  0  1  2  0  0 |  e = Macrocytic anemia
  5  4  1  1  0 64  4  0  1 |  f = Normocytic hypochromic anemia
  0  0  2  0  0  8 62  0  3 |  g = Normocytic normochromic anemia
  1  4  0  0  0  2  1  7  0 |  h = Other microcytic anemia
  4  1  0  0  0  0  0  0 16 |  i = Thrombocytopenia
```

Figure 23: Confusion matrix

The model's decision boundaries and the grouping of cases across various anaemia categories are shown in the test data's scatter plot visualisation. Some classes' clusters show overlaps, which would have led to the misclassifications seen in the confusion matrix, while other classes' clusters are closely clustered, indicating clear decision-making. The MLP model performs admirably overall, especially for classes that are well-represented. Its poorer performance for fewer common classes, however, points to the possibility of improvement, maybe with the use of sophisticated loss functions or oversampling to lessen class imbalance. The findings offer a strong basis for comprehending the distribution of anaemia kinds and further refining categorisation techniques.

J48 DECISION TREES

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         353               98.3287 %
Incorrectly Classified Instances         6                1.6713 %
Kappa statistic                          0.9793
Mean absolute error                      0.0057
Root mean squared error                  0.0626
Relative absolute error                  3.1513 %
Root relative squared error             20.9274 %
Total Number of Instances              359
```

Figure 24: J48 decision tree metrics

With a high accuracy of 98.33%, the analysis of J48 decision trees for the test data shows impressive performance. Only six examples were incorrectly categorised by the model, which properly classified 353 out of 359 instances. A Kappa number of 0.9793 supports this result, showing a high degree of agreement between the actual and projected categories. The accuracy and dependability of this model are further illustrated by the mean absolute error (0.0057) and root mean squared error (0.0626).

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Healthy
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iron deficiency anemia
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Leukemia
                 1.000    0.003    0.750      1.000    0.857      0.865    0.997     0.750     Leukemia with thrombocytopenia
                 0.000    0.003    0.000      0.000    0.000      -0.005   0.777     0.073     Macrocytic anemia
                 1.000    0.004    0.988      1.000    0.994      0.992    0.997     0.972     Normocytic hypochromic anemia
                 0.973    0.007    0.973      0.973    0.973      0.966    0.986     0.973     Normocytic normochromic anemia
                 0.933    0.000    1.000      0.933    0.966      0.965    0.965     0.936     Other microcytic anemia
                 1.000    0.003    0.955      1.000    0.977      0.976    0.997     0.922     Thrombocytopenia
Weighted Avg.    0.983    0.002    0.979      0.983    0.981      0.979    0.993     0.971
```

Figure 25: Class metrics

The model obtained excellent scores for the majority of anaemia categories, including "Iron deficiency anaemia" and "Normocytic hypochromic anaemia," with precision and recall at 1.000 when examining the detailed accuracy by class. Although they still performed well, other categories such as "Leukaemia with thrombocytopenia" had somewhat worse recall and precision, which is still within an acceptable range. The model's strong ability to discriminate between various forms of anaemia is demonstrated by the high ROC and PRC area values across all classes.

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 96  0  0  0  0  0  0  0  0 |  a = Healthy
  0 54  0  0  0  0  0  0  0 |  b = Iron deficiency anemia
  0  0 12  0  0  0  0  0  0 |  c = Leukemia
  0  0  0  3  0  0  0  0  0 |  d = Leukemia with thrombocytopenia
  0  0  0  0  0  1  2  0  0 |  e = Macrocytic anemia
  0  0  0  0  0 80  0  0  0 |  f = Normocytic hypochromic anemia
  0  0  0  1  1  0 73  0  0 |  g = Normocytic normochromic anemia
  0  0  0  0  0  0  0 14  1 |  h = Other microcytic anemia
  0  0  0  0  0  0  0  0 21 |  i = Thrombocytopenia
```

Figure 26: Confusion matrix

The model's capacity for categorisation is revealed via the confusion matrix. Misclassifications of categories like "Normocytic normochromic anaemia," "Healthy," and "Iron

deficiency anaemia" have been hardly nonexistent. In categories such as "Macrocytic anaemia," where the model finds it difficult to accurately differentiate between some uncommon or complex anaemia types, a few inaccuracies are noted.

Overall, the J48 model has proven to be highly effective for classifying anemia types in this dataset.

e. Research Questions

1. What is the distribution of patients across different types of anemia and related conditions?

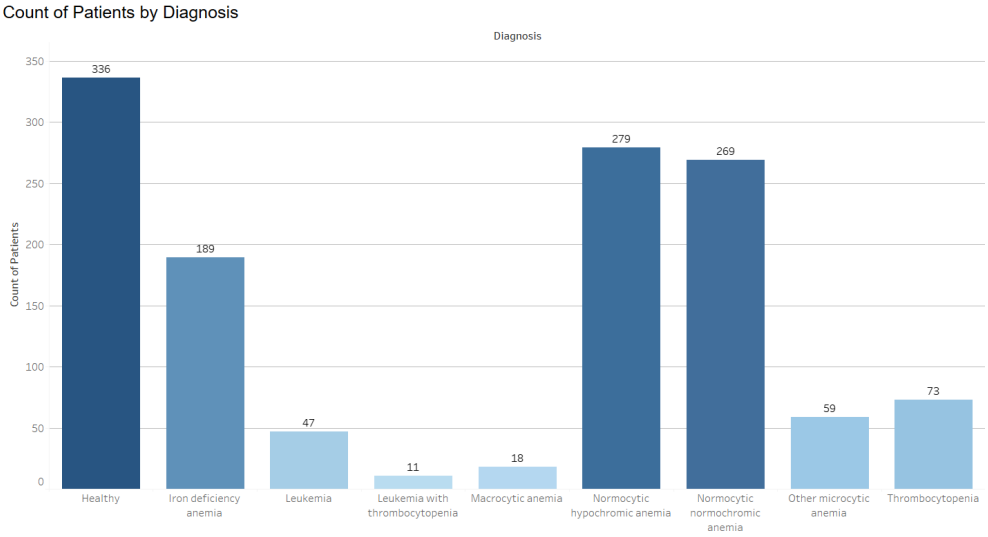


Figure 27: Bar chart of distribution of patients across different types of anemia and related conditions

This bar chart represents the distribution of patients in different diagnoses. The two most frequent diagnoses are Normocytic hypochromic anemia with 279 patients and Normocytic normochromic anemia with 269 patients, followed by Iron deficiency anemia with 189 patients and Healthy individuals with 336 patients. The rare conditions include Macrocytic anemia with 18 patients and Leukemia with thrombocytopenia with 11 patients.

2. How do hemoglobin (HGB) levels, including their variation and averages, differ across various diagnoses, and what patterns or anomalies can be observed to distinguish between healthy individuals and anemia types?

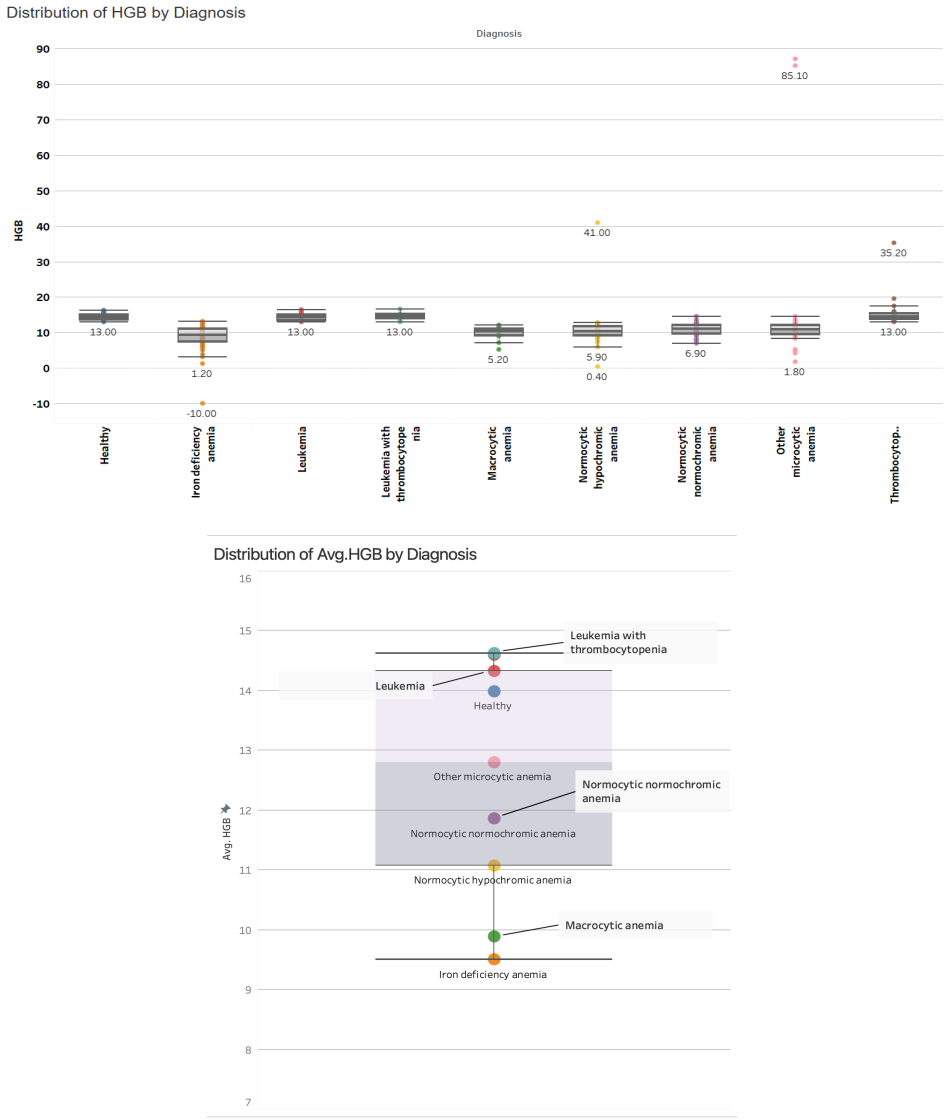


Figure 28: Box plots of HGB levels of different anemia classes

The analysis of HGB levels across diagnoses shows clear trends. Healthy diagnoses demonstrate the highest average values of HGB, while Iron deficiency anemia and Macrocytic anemia are significantly lower. Outliers also show extreme levels of HGB in conditions such as Leukemia with thrombocytopenia, which aids diagnosis and helps to separate one type of anemia from another.

3. What are the factors that contribute to predicting the different types of anemia ?

	precision	recall	f1-score	support
Healthy	0.98	1.00	0.99	93
Iron deficiency anemia	1.00	1.00	1.00	61
Leukemia	1.00	1.00	1.00	7
Leukemia with thrombocytopenia	0.60	1.00	0.75	3
Macrocytic anemia	0.00	0.00	0.00	3
Normocytic hypochromic anemia	0.97	0.99	0.98	88
Normocytic normochromic anemia	1.00	0.97	0.99	75
Other microcytic anemia	1.00	1.00	1.00	17
Thrombocytopenia	1.00	0.95	0.98	22
accuracy			0.98	369
macro avg	0.84	0.88	0.85	369
weighted avg	0.98	0.98	0.98	369

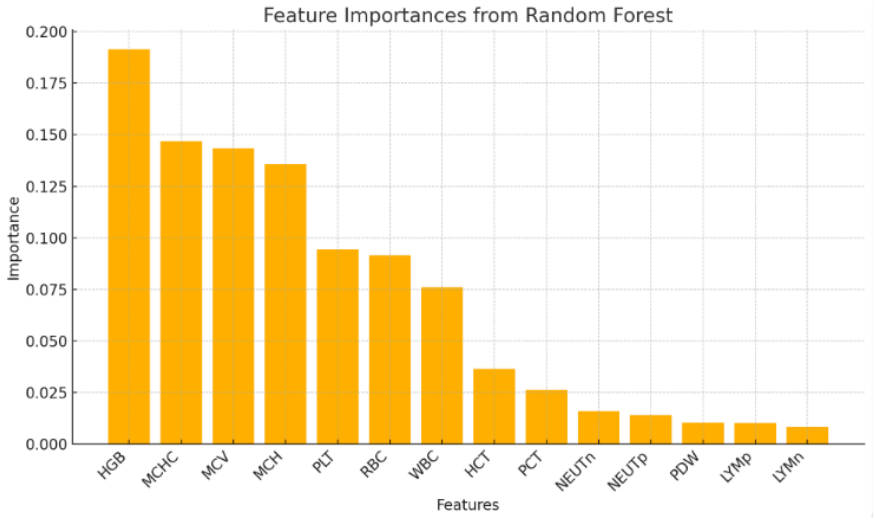


Figure 29: Feature importance from Random forest model

Consequently, the Random Forest model's most important predictors for anemia types are: Hemoglobin (HGB), Mean Corpuscular Hemoglobin Concentration (MCHC), and Mean Corpuscular Volume (MCV). On the overall accuracy amounting to 98%, the model showed substantial performance in the detection of both common and rare diagnoses. These features are really crucial in the understanding of blood health and provide relevant clinical information for the purpose of anemia classification.

f. Exploratory Data Analysis

Exploratory Data Analysis (EDA) provides valuable insights into the anemia dataset, helping to identify trends, patterns, and anomalies across different diagnoses. The visualizations

below explore relationships between key hematological parameters, including Hemoglobin (HGB), Platelet Count (PLT), White Blood Cell Count (WBC), and Red Blood Cell Count (RBC), to better understand the diagnostic features of various anemia types.

1. HGB vs. PLT Relationship

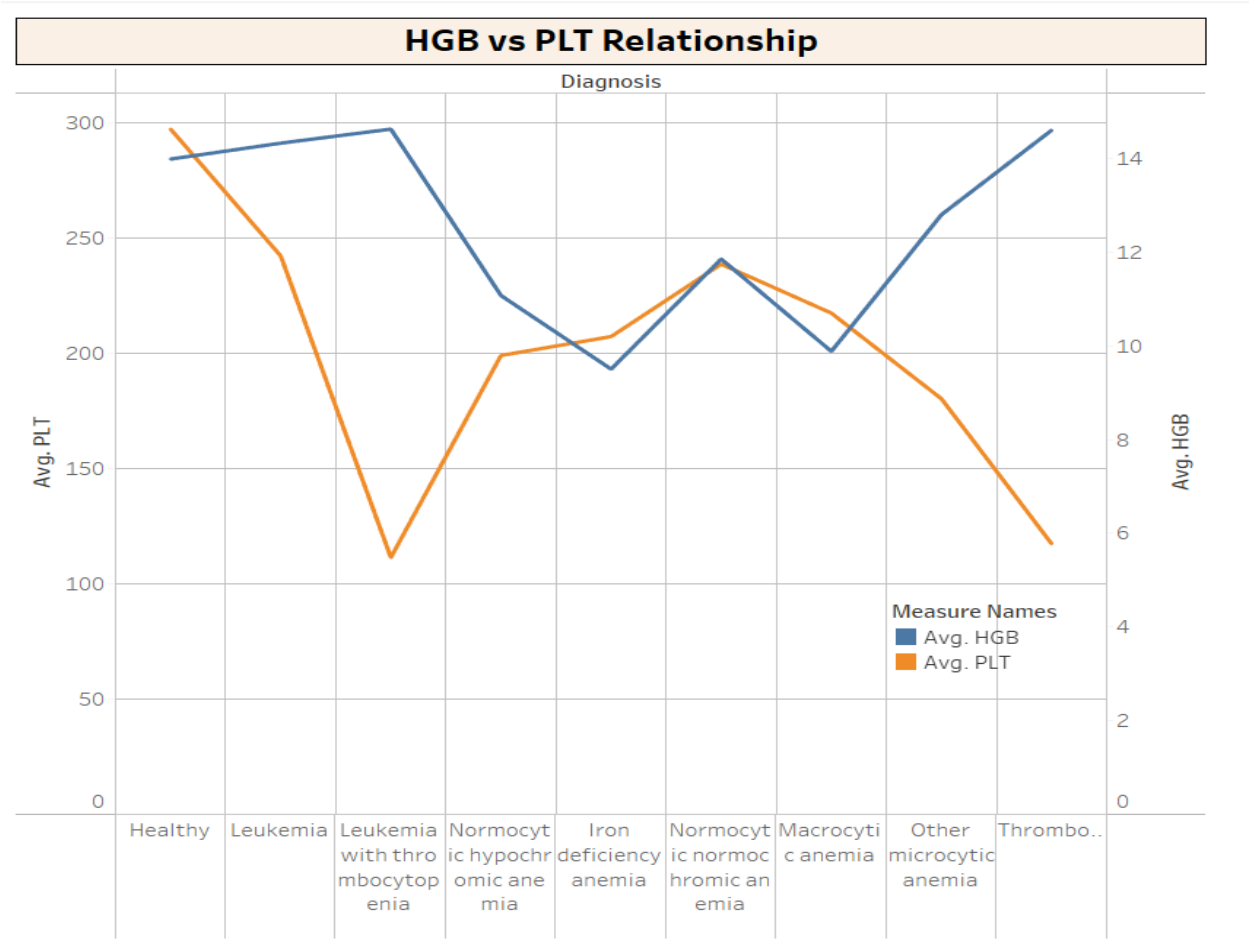


Figure 29: Avg PLT vs Avg HGB for different anemia classes

This plot illustrates the relationship between **Hemoglobin (HGB)** and **Platelet Count (PLT)** across different diagnoses. Healthy individuals have high HGB and moderately high PLT levels, while conditions like **Leukemia with Thrombocytopenia** exhibit both low HGB and

PLT levels. The distinct trends across diagnoses highlight the diagnostic value of these parameters, especially in distinguishing severe anemia types.

2. Avg WBC & RBC by Diagnosis

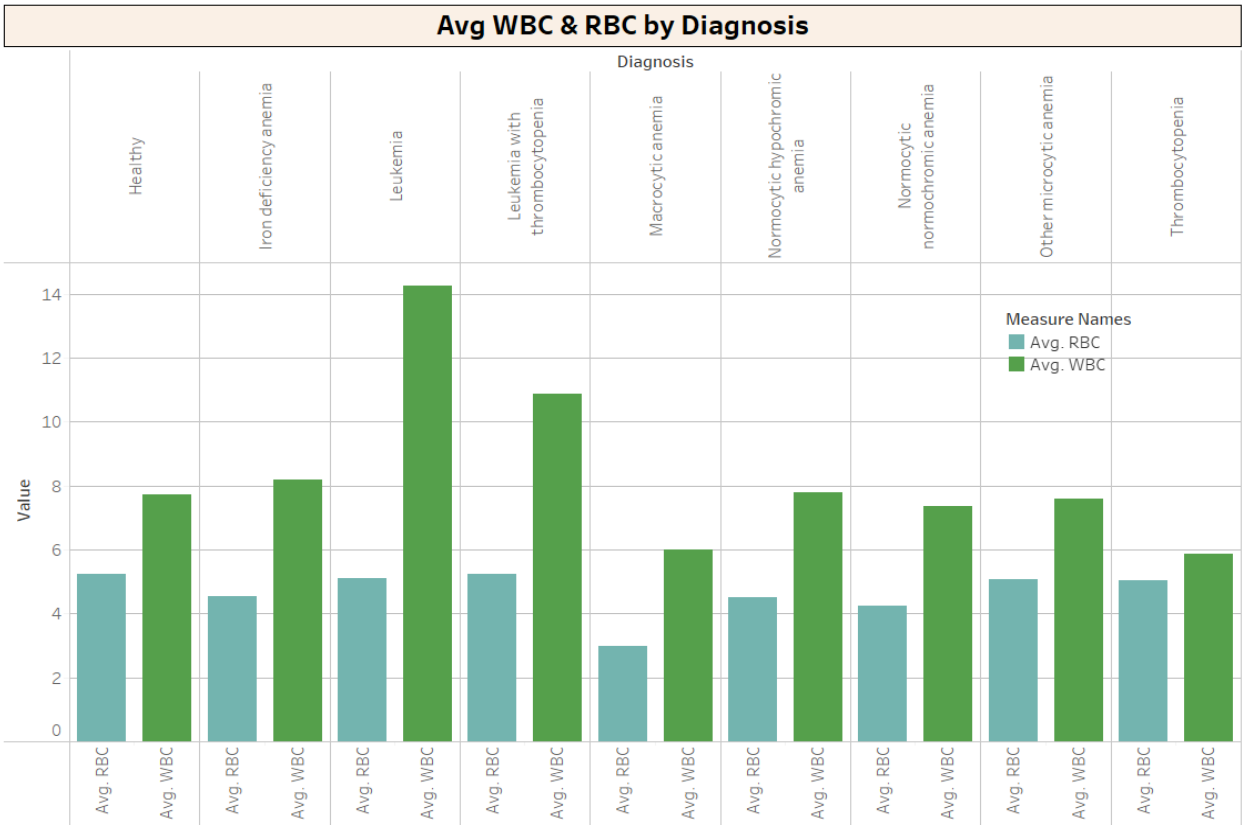


Figure 30: Average WBC and RBC for each class of anemia

This chart compares the average **White Blood Cell (WBC)** count and **Red Blood Cell (RBC)** count for each diagnosis. Leukemia exhibits the highest WBC counts, reflecting abnormal white blood cell production, while Macrocytic Anemia shows the lowest RBC counts due to large, dysfunctional red blood cells. Most anemia types show reduced RBC counts with normal WBC levels, emphasizing the importance of these parameters in diagnosis.

3. Diagnosis Contribution to Blood Stats

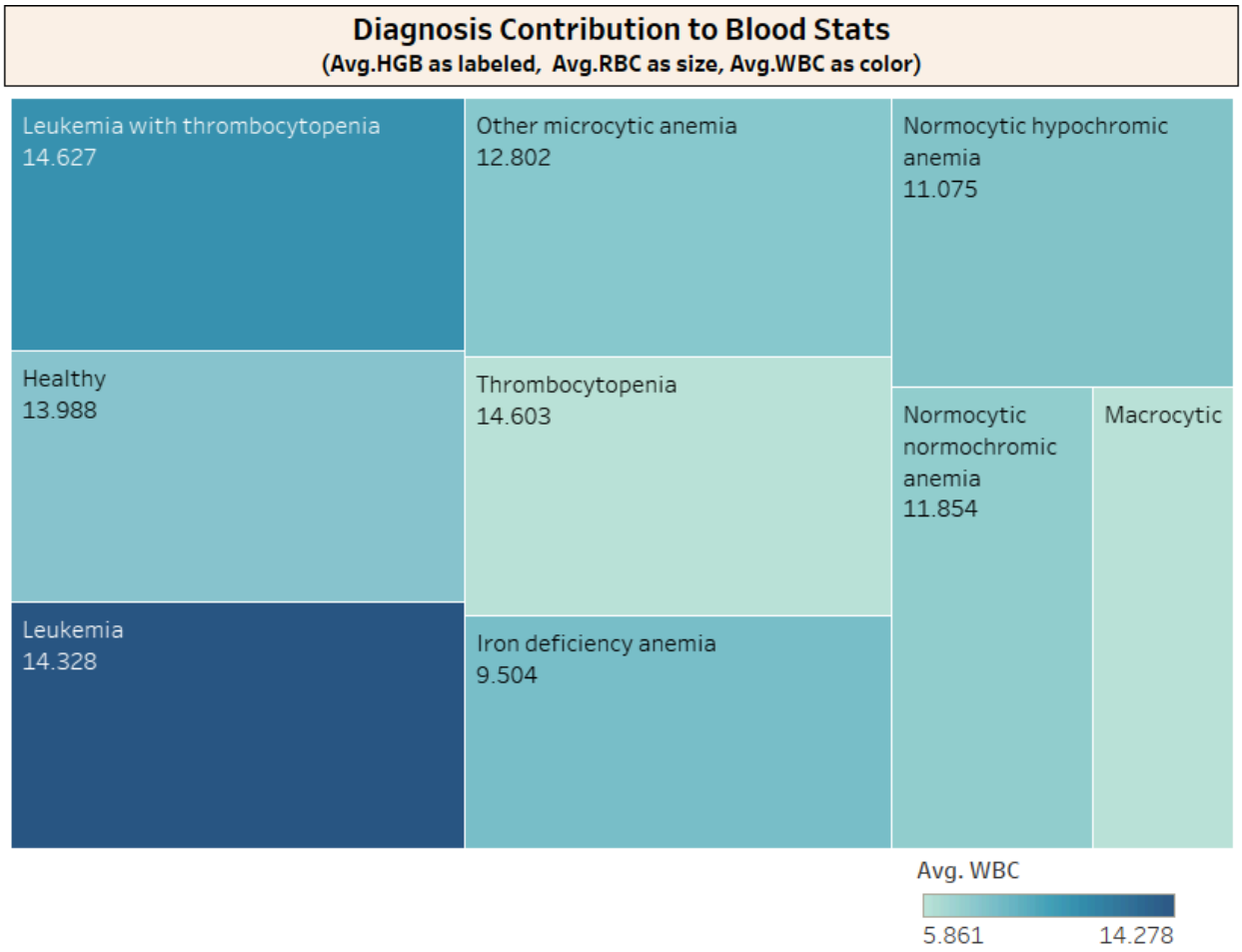


Figure 31: Tree map of average WBC for each anemia class

This treemap shows the contribution of average **Hemoglobin (HGB)** (as labels), **Red Blood Cell (RBC)** count (as size), and **White Blood Cell (WBC)** count (as color intensity) for each diagnosis. It emphasizes the variations in blood parameters among different conditions, such as:

- **Leukemia with Thrombocytopenia**: High WBC (darker color) and lower RBC (smaller size).
- **Iron Deficiency Anemia**: Low HGB (label) and RBC (size).
- **Healthy Individuals**: Balanced HGB, RBC, and WBC values.

This visualization highlights the interplay of key blood parameters across diagnoses, aiding in understanding diagnostic patterns.

4. HGB and MCHC by Diagnosis

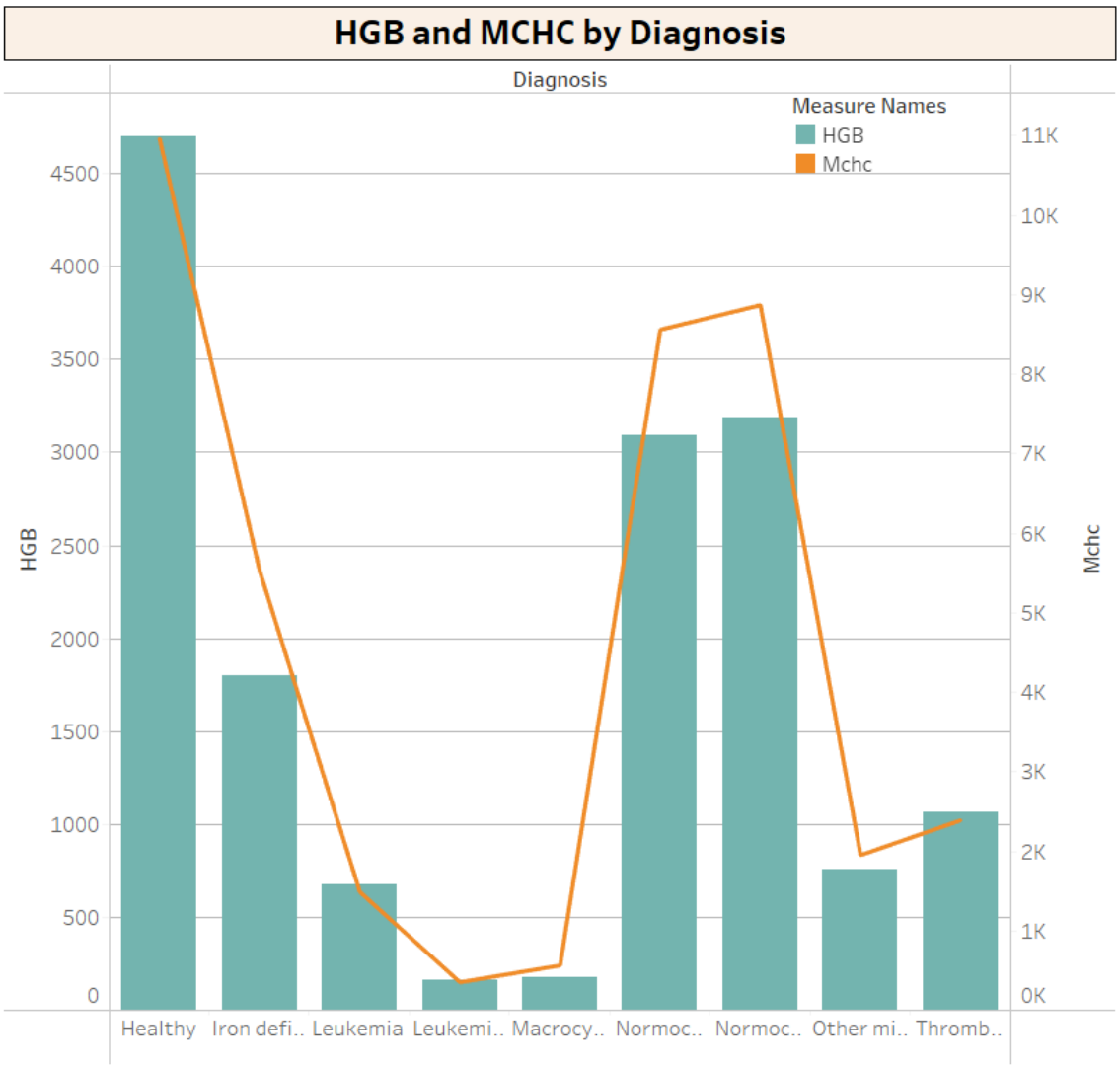


Figure 32: HGB vs MCHC for each anemia class

This chart displays the relationship between **Hemoglobin (HGB)** and **Mean Corpuscular Hemoglobin Concentration (MCHC)** across various diagnoses:

- **Healthy Individuals**: High HGB and moderately high MCHC values, indicative of normal blood health.

- **Iron Deficiency Anemia**: Significantly lower HGB and reduced MCHC, reflecting insufficient hemoglobin production.
- **Leukemia**: Shows drastically reduced HGB and MCHC levels, corresponding to abnormal blood cell production.
- **Macrocytic Anemia**: HGB levels are low, while MCHC values remain stable, highlighting the distinct characteristics of this condition.

This visualization emphasizes the interplay between HGB and MCHC, providing diagnostic insights into the severity and type of anemia. Conditions with overlapping or extreme values indicate complexity in differentiation.

III. CONCLUSION

Based on the given dataset, the models' outcomes—which include Random Forest, Logistic Regression, Multilayer Perceptron (MLP), and J48 Decision Trees—showcase a thorough method for categorising different forms of anaemia. With a classification accuracy of 98.33% on the test data, the J48 Decision Tree model outperformed the others. This performance shows how well it can handle categorical divides and spot important patterns in the data. Its capacity to accurately classify important categories such as "Healthy," "Iron deficiency anaemia," and "Normocytic normochromic anaemia" with few misclassifications is demonstrated by the confusion matrix.

Using its capacity to recognise non-linear patterns, the Multilayer Perceptron model also showed impressive performance. Its marginally worse accuracy when compared to the J48 model, however, raises the possibility that it is overfitting some classes or having trouble with unbalanced data. Despite being efficient and interpretable, logistic regression fared poorly for more complex anaemia types because of its intrinsic linearity, which hampered its ability to identify subtle, non-linear connections in the data. With comprehensive handling of class imbalances and strong feature importance analysis, Random Forest offered balanced performance across all classes.

Haematological indicators such haemoglobin (HGB), platelet count (PLT), red blood cell count (RBC), and mean corpuscular haemoglobin concentration (MCHC) are crucial for identifying and distinguishing between different types of anaemia, according to the analysis. The diagnostic use of these measures is demonstrated by distinct patterns, such as decreased HGB and MCHC in Iron Deficiency Anaemia or low HGB and PLT in Leukaemia with Thrombocytopenia. Blood parameters are balanced in healthy persons, offering a reference point for comparison. Relationships such as HGB vs. PLT and HGB vs. MCHC highlight the interplay between these qualities and highlight how difficult it is to differentiate between overlapping circumstances. These findings support the outcomes of Random Forest and other classification models and open the door for automated tools that improve human judgement and facilitate accurate diagnosis.

IV. FUTURE WORK

1. **HANDLING CLASS IMBALANCE:** The dataset contained fewer instances of several anaemia classes, such as "Leukaemia with thrombocytopenia," which resulted in decreased recall and precision. Future research can concentrate on improving balance by modifying class weights in the models or oversampling under-represented classes using strategies like SMOTE.
2. **FEATURE SELECTION**: Although the models did well, investigating more complex feature engineering strategies, including domain-specific transformations or interaction terms, could increase classification accuracy even more. Model complexity could also be reduced by incorporating feature importance insights to eliminate superfluous or irrelevant elements.
3. **EXPLAINABILITY AND INTERPRETABILITY**: Explainability is essential for therapeutic applications. Future research can concentrate on using explainable AI (XAI) methods, like SHAP or LIME, to offer insights into model predictions and help medical practitioners comprehend the rationale behind particular classifications.
4. **DATASET EXPANSION**: Adding more examples from a range of demographics and more medical characteristics to the dataset may enhance the models' generalisability and suitability for use in practical situations.

REFERENCES

[1] Patel, A., Shah, S. D., & Desai, M. (2021). Machine learning-based classification of anemia using blood test data. *International Journal of Medical Informatics*, 149, 104424. DOI: 10.1016/j.ijmedinf.2021.104424

[2] Arevalo, H. J., et al. (2019). Hematology data-driven approaches for predicting anemia subtypes. *Journal of Clinical Pathology*, 72(3), 188-194.

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). DOI: 10.1145/2939672.2939785

[4] Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20), 1920-1930. DOI: 10.1161/CIRCULATIONAHA.115.001593

[5] Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques" (4th Edition)*. Morgan Kaufmann. Available at: https://www.cs.waikato.ac.nz/ml/weka/