Classification of Anaemia Types Using Supervised Machine Learning Techniques

By

Cindy Kerubo Onwong'a

120762

A Research Thesis Submitted to the School of Computing and Engineering Sciences in partial fulfilment of the requirements for the award of the Degree of Master on Science degree in Information Technology.

Master of Science in Information Technology

Strathmore University

June 2024

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: Cindy Kerubo Onwong'a

Adm. No: 120762

Signature: _____ _____ Date: ____17th April 2024_____

Approval

This thesis has been submitted to the School of Computing and Engineering Sciences for examination with my approval as the university supervisor.

Signature: _____ Date:___17th April 2024_____

Name: Dr. Vincent Omwenga

Abstract

Anaemia reduction is one of the World Health assembly goals for 2025. Given the complex aetiology of anaemia, classification of nutritional anaemia using traditional methods has limitations and drawbacks. Traditional methods of classification rely heavily on analysis of complete blood count tests which need specialists and trained personnel, and present potential for errors in analysis. These traditional methods are also expensive and time consuming given the wait time between testing and getting the results. Machine learning based algorithms offer more accuracy and efficiency in the classification of anaemia given their ability to learn data and identify patterns. This study aimed at building a classification model for classifying nutritional types of anaemia using supervised machine learning techniques. The dataset that was utilized in this study was retrieved from Kaggle, an open-source dataset repository and used in accordance with the Open Database license. The dataset contained complete blood count test results for patients with proven cases of nutritional anaemia. The data was pre-processed and explored in preparation for model building. The features were all used in the model development because all the variables are different, and they contribute to the classification of anaemia. The models that were built are Naïve Bayes, random forest, XG Boost, decision trees, and multilayer perceptron. These models were tested using the testing set and their performances compared to find the better performing one. Hyperparameter tuning was done on some of the poorer performing models to try and improve their performance. The best performing model was the XG Boost classifier which achieved an accuracy of 98.85%. The poorest performing model was the Gaussian Naive Bayes model with an accuracy score of 0.7872. The SVM model was very computationally heavy and could not build. For deeper analysis of the model, metrics like recall, precision and F1 scores were measured. The XG boost model was then loaded to an interface for functionality testing. The tool was able to classify nutritional classes of anaemia based on complete blood count data entered by a user. This tool could potentially be plugged into hospitals and clinics to aid in the early detection, diagnosis and treatment by reducing the wait time between getting tested and getting results. This can be considered one of many steps towards anaemia reduction.

Keywords: Anaemia Classification, Nutritional Anaemia classification, Supervised learning, SMOTE, XG Boost, Random Forest, Naive Bayes, Decision Trees, machine learning, multilayer perceptron**.**

Table of Contents

List of Figures

List of Tables

List of Equations

Abbreviations/Acronyms

ANN	-	artificial neural network
ANOVA	-	analysis of variance
AUC	-	area under the curve
CBC	-	Complete Blood Count
CNN	-	convolutional neural network
HCT	-	haematocrit
HGB	-	haemoglobin
HIV	-	Human Immunodeficiency Virus
HSWC	-	hyper-segmented white cell
IDA	-	Iron deficiency anaemia
kNN	-	k-Nearest Neighbours
LDA	-	linear discriminating analysis
MCHC	-	mean corpuscular haemoglobin concentration
MCV	-	Mean Corpuscular Volume
NHANES	-	National Health and Nutrition Examination Survey
RAD	-	Rapid Application Development
RBC	-	red blood cell
RDW	-	Red blood cell width
ROC	-	Receiver Operating Characteristic
SDG	-	Sustainable Development Goals
SVM	-	support vector machine
TIBC	-	total iron-binding capacity

WBC - white blood cells

WHA - World Health Assembly

Glossary

Haemoglobin - A protein that is crucial for oxygen transportation to the body and body tissues (World Health Organization, 2017).

Nutritional anaemia - The body's inability to synthesize enough haemoglobin and erythrocytes because of a deficiency of certain nutrients (Balarajan et al., 2011)

Chapter 1: Introduction

1.1. Background

Anaemia is among the most predominant blood diseases among human beings. It occurs in all ages and is pathophysiologically diverse and multifactorial. In as much as it may be a unique disease, it is frequently associated with other pathologic conditions, meaning almost all specialists in the medical field, deal with it in different etiologies and degrees. There is a very slim chance that a physician can in their entire pathology career not be confronted by an anaemic patient. In some specific conditions, the absence or presence of anaemia refines the prognosis like in the elderly population. The World Health Organization defines anaemia as a medical condition where red blood cells become fewer or the concentration of haemoglobin (Hb) in blood falls below the standard level thus becoming inadequate in fulfilling the body's physiological need or demand (World Health Organization, 2017, 2011). Haemoglobin is a protein that is crucial for oxygen transportation to the body and body tissues. With abnormal red blood cells or a decreased number of them, there is a decreased capacity of oxygen being transported to body tissues. Anaemia can also be characterized by a reduction of the concentration of erythrocyte mass, or concentration of blood haemoglobin and haematocrit (Karagül Yıldız et al., 2021). Normal values of haemoglobin and haematocrit may differ in classification based on parameters like age and gender, with values below these determined thresholds showing anaemia presence. Anaemia serves as a marker for compromised health and inadequate nourishment.

Anaemia is typically caused by three main mechanisms: ineffective erythropoiesis which is characterized by the reduced production of red blood cells, haemolysis which involves red blood cell destruction, and blood loss (World Health Organization, 2017). The popular contributors of anaemia are genetic haemoglobin disorders, nutritional deficiencies, and diseases (World Health Organization, 2020, 2017). The top three factors leading to anaemia are inadequate iron consumption, insufficient vitamin A intake, and having the beta-thalassaemia trait respectively (Gardner & Kassebaum, 2020). Iron deficiency anaemia constitutes roughly half of all instances of anaemia in pregnant women as well as non-pregnant women, as well as around 42% of cases in children under five worldwide (Safiri et al., 2021). The contribution of iron deficiency as a

cause varies based on factors like age, gender, as well as geographical location of the population under study, and the prevalence of other factors contributing to anaemia in the region. Whilst iron deficiency results in lowered production of red blood cells and haemoglobin that results in decreased haemoglobin concentration and haematocrit, which is used to identify anaemia, there exist alternative factors that can lead to anaemia that are not related to iron. Other factors that can lead to anaemia are nutritional deficiencies like folate, vitamins A and B12, inflammation whether acute or chronic, communicable diseases like tuberculosis, malaria, HIV, and other parasitic infections, and haemoglobinopathies and other disorders that affect the synthesis of haemoglobin, survival, or production of red blood cells, that maybe be either acquired or inherited.

Anaemia may be categorized based on its underlying cause like nutritional deficiency or haemolytic anaemia. Additionally, different types of anaemia are distinguishable by red blood cell shape, size and colour. For instance, in the case of microcytic anaemias like iron deficiency anaemia where red blood cells become smaller than usual, and the concentration of haemoglobin is each red blood cell is decreased. Iron deficiency anaemia is also characterized as hypochromic since the red blood cells have a lower-than-normal colour intensity (World Health Organization, 2017). Folate or vitamin B12 deficiencies are typical of megaloblastic anaemia, which is a condition where the presence of red blood cells that are larger than normal is characteristic. When the body is unable to synthesize enough haemoglobin and erythrocytes because of a deficiency of certain nutrients, it is referred to as "nutritional anaemias" (Balarajan et al., 2011; World Health Organization, 2017). Inadequate intake or deficiencies in specific vitamins and minerals like vitamin A, riboflavin (B2), pyridoxine (B6), cobalamin (B12), C, D and E, folate, and copper may cause anaemia because of the vital role they play in haemoglobin as well as erythrocyte production (World Health Organization, 2017).

Anaemia can be caused by various infectious diseases through several mechanisms such as hindering the processes of nutrient intake and processing by the body, ineffective erythropoiesis or increased loss of nutrients. The inflammatory response in acute and chronic infections also leads to "Anaemia of chronic inflammation" or "Anaemia of chronic disease." In this case, pro-inflammatory cytokines can cause changes in iron metabolism, leading to the accumulation of

iron as ferritin in stores, while reducing red blood cell production and lifespan. This can result in a class of anaemia, called normocytic anaemia, characterized by low red blood cell count (Wiciński et al., 2020). In low- and middle-income countries, prevalent infections like malaria, tuberculosis, HIV, and parasitic infections are among anaemia causes. After iron deficiency anaemia, anaemia of chronic inflammation or chronic disease, is the second most prevalent form (Wiciński et al., 2020).

Anaemia can also be caused by structural changes or decreased production of globin chains in haemoglobin. Inherited haemoglobin disorders are estimated to result in more than 300, 000 births globally every year, with around 80% of which are occurring in low-income and middle-income countries (Hess et al., 2019). The most common genetic haemoglobin disorder is sickle cell disease, which is linked to chronic haemolytic anaemia and primarily found in countries located in sub-Saharan Africa. Next in frequency are β- and α-thalassaemia and are more prevalent in Southeast Asia (Hess et al., 2019). Approximately 5% of the world's population is estimated to carry a significant haemoglobin variant, with a higher prevalence of 18% in Africa and 7% in Asia (Hess et al., 2019).

To detect anaemia, a complete blood count test is performed to examine haemoglobin levels. Iron deficiency anaemia which is the most common form of anaemia in both genders, mostly results from gastrointestinal bleeding, inadequate dietary iron intake, poor iron absorption in the intestines, as well as reduced erythropoiesis due to iron deficiency. Without early diagnosis, it can cause symptoms such as fatigue, weakness, shortness of breath, among others (Gardner & Kassebaum, 2020). B12 deficiency anaemia, also known as pernicious anaemia, is prevalent in underdeveloped countries due to insufficient intake of animal protein, and its main cause is poor absorption of vitamin B12 in the small intestine. This anaemia develops insidiously and presents symptoms like dizziness and neurological issues and can be associated with pancreatic diseases and certain cancers (Gardner & Kassebaum, 2020). Folate deficiency anaemia is caused by a lack of folic acid, leading to reduced production of blood cells. If undetected, it can result in gingivitis, diarrhoea, nervous system disorders, forgetfulness and depression. Haemoglobin anaemia occurs when there are not enough healthy red blood cells to carry adequate oxygen to organs and tissues, usually caused by impaired production or increased destruction of red blood

cells (Gardner & Kassebaum, 2020). Anaemia is also a contributing factor in the development of leukaemia and solid organ cancers.

Anaemia, particularly nutritional anaemia, remains a significant global health issue with serious health consequences if not diagnosed and managed early. Advancements in technology like machine learning, show promise in enhancing the detection and classification of nutritional anaemia, given their ability to automatically learn from data and identify patterns. Machine learning models can perform analysis of complete blood count tests more accurately and swiftly in order to predict anaemia, which can revolutionize anaemia diagnosis, treatment and management, and ultimately reducing the burden of anaemia worldwide.

1.2. Problem Statement

Anaemia is a critical global health issue that impacts various population groups such as young children, adolescent girls, females in their childbearing age, expectant women, and older adults in nations with limited financial resources (Kassebaum, 2016; World Health Organization, 2020). In 2019, over 30% of the female population residing in Africa and Asia experience a significant impact from anaemia while in Northern America and Europe only 24.6% of the female population is affected. The regional disparities are high, with Africa being almost thrice as prevalent as Northern America and Europe (FAO, 2021). Anaemia continues to be highly prevalent worldwide, especially in low-income settings where it is assumed that a significant percentage of young children and females of childbearing age likely to have anaemia (FAO, 2021; Gardner & Kassebaum, 2020). Reduction of anaemia is included as one of the WHA Global Nutrition Targets for 2015 and is also part of the Sustainable Development Goals (SDGs) (FAO, 2021). Although some progress has been achieved regarding anaemia reduction, global progress is off track towards the achievement of WHA 2025 target to reduce by 50% anaemia in females of childbearing age (FAO, 2021).

Traditional methods of anaemia classification can be limited. These methods are time-consuming because of the protracted process and wait time between getting tested and getting the results. They also tend to be expensive, especially for individuals living below the poverty level without

access to free healthcare. They are also prone to inaccuracies, given their subjectivity to other interpretations (Kovačević et al., 2022).

1.3. Objectives

1.3.1. General Objective

This study aimed at classifying nutritional anaemia types by developing a classification model using supervised machine learning techniques.

1.3.2. Specific Objectives

i. To investigate standards used in categorizing anaemia.

ii. To review machine learning techniques for classifying anaemia.

iii. To develop a model for classifying anaemia types using machine learning techniques.

iv. To test and validate the model.

1.4. Research Questions

i. What standards are referred to when categorizing anaemia?

ii. What machine learning techniques have been used in the classification of anaemia?

iii. How can classifying anaemia types be improved using technology?

iv. How will the model's performance be evaluated?

1.5. Justification

Approximately a third of the global population is affected by anaemia, making it the most prevalent blood disorder worldwide. As a symptom of serious diseases and a disease, anaemia can significantly impact the quality of life of those affected. In settings with lower income and resources like low-income and middle-income counties, it is crucial to implement interventions to lower the prevalence of anaemia, particularly among those affected by it. Traditional classification techniques can be expensive, time-consuming, and sometimes inaccurate given the

complexity of anaemia etiologies. Machine learning algorithms can automatically learn from data and identify patterns, which are crucial in classifying anaemia. As an anaemia reduction effort, the classification model developed in this study assists in the classification of nutritional anaemia, and hence early detection, diagnosis and treatment, which can improve health outcomes and reduce the burden of anaemia worldwide.

1.6. Scope

The study was limited to classifying nutritional types of anaemia, specifically folate anaemia, B12 anaemia, iron deficiency, and hemoglobin anaemia. The classification models developed utilized decision tree algorithm, random forest algorithm, naïve bayes algorithm, multilayer perceptron algorithm and XG boost algorithm. The dataset used in the study did not include pregnant women, children, and patients with cancer. The classifier was utilized to enable the classification and prediction of the input data into the correct classes.

Chapter 2: Literature Review

2.1. Introduction

The literature review section intended to examine the relevant literature in relation to anaemia classification. It also looked at the universal standards applied when classifying anaemia, as well as existing techniques, employed in the classification of anaemia. It also explored various machine learning classification techniques which formed the basis of the study.

2.2. Theoretical Review

Classification, is the process of allocating objects to already defined categories based on the characteristics of the objects (Bouveyron et al., 2019). The categories are already known, usually derived form a dataset of previously classified objects, and the goal is to develop algorithms that can accurately assign objects to the appropriate category, based on their features, which is referred to as a supervised problem, since it relies on the input of experts to indicate the categories being classified. The process of classification involves comparing the features of new objects to those of the objects in the predefined categories and assigning them to the most similar category (Bouveyron et al., 2019). For centuries, humans like biologists, botanists, or doctors have been carrying out the task of classification, by learning how to categorize or classify new examinations to particular species or ailments (Bouveyron et al., 2019). Prior to the twentieth century, this process was performed without the use of automated algorithms.

Fisher (1936), introduced the earliest statistical method used for classification in 1936, called the Fisher's linear discriminant analysis or simply linear discriminant analysis. It involved determining the optimal linearly weighted integration of multiple attributes to discriminate among multiple populations. He applied this method to a dataset on irises derived from Edgar Anderson's work. In a later publication, Fisher demonstrated the relationship connecting the discriminant analysis method and other existing statistical methods including Mahalanobis generalized distance, Hotelling's T-squared distribution, and analysis of variance (ANOVA) (Bouveyron et al., 2019; R. A. Fisher, 1936, 1937; Anderson, 1936; Hotelling, 1931; Mahalanobis, 1930). Discriminant analysis found numerous applications in various disciplines including handwriting recognition, spam detection, fraud and fault detection, computer vision

and medical diagnosis. Fisher's linear discriminant analysis contributed effective resolution for various practices, but some practices necessitated the invention of specialized approaches (Bouveyron et al., 2019).

One of the fundamental classification methods is logistic regression, which was introduced by Cox (1958). Logistic regression extended the linear regression model to categorical dependent variables, making performing binary classification attainable. Logistic regression has been widely adopted in various fields including economics, political science, medicine, and marketing and continues to gain popularity in various industries such as banks to predict mortgage defaults and marketing firms to predict click-through rates (Bouveyron et al., 2019).

The perceptron by Roseblatt (1958) was another crucial early classification method, initially designed for image recognition by emulating the decision-making behaviour of neurons. Although the initial results were promising, the perceptron was limited in its ability to identify multiple categories without multiple additional layers. Nonetheless, it is among the earliest artificial neural networks, and with advancements in computing power in recent years, convolutional neural networks that utilize modified versions of multilayer perceptron have transformed the field of classification, exhibiting remarkable results in specific cases (Lecun et al., 1998; Rosenblatt, 1958).

Prior to the advent of deep learning and convolutional neural networks, support vector machines were instrumental in advancing classification performance in the late 1990s. Although the first version of the Support Vector Machine (SVM) was introduced in 1963, it was first implemented in 1992, courtesy of the "kernel trick" by Boser et al. (1992), thus making a major impact on classification performances (Boser et al., 1992; Cortes & Vapnik, 1995). SVMs are a group of classifiers that rely on the kernel selection, that through a nonlinear projection, reconstructs the original dataset to a high dimensional space, where it is linearly separable with a hyperplane (Bouveyron et al., 2019). SVMs were popular for their ability to handle various types of data due to the concept of kernel.

As researchers faced new data characteristics such as high dimensionality, limited sample sizes, partially labelled data, and non-normal distributions, they proposed various extensions to FDA.

For example, to elude the curse of dimensionality, McLachlan acknowledged the vitality of selecting variables. Alternative methods using constrained Gaussian models were proposed by Banfield & Raftery, and Bensmail & Celeux in 1993 and 1996 respectively (Bouveyron et al., 2019; Banfield & Raftery, 1993; Bensmail & Celeux, 1996). Celeux & Mkhadri (1992) introduced a regularized discriminant analysis method for high-dimensional discrete data whereas Hastie & Tibshirani (1996) explored the categorization of non-normal data by making use of mixtures of Gaussionas (Bouveyron et al., 2019; Celeux & Mkhadri, 1992; Hastie & Tibshirani, 1996; McLachlan, 1976).

2.2. Anaemia Classification Systems

Classification of anaemia is an essential task that helps in understanding the disease and developing effective treatment strategies. Classification systems for anaemia aim to provide a common language and framework for clinicians and researchers to communicate and compare data, facilitate diagnosis and management, and enhance understanding of the pathology and epidemiology of anaemia. The most widely used classification system for anaemia is the World Health Organization (WHO) classification which is based upon the etiology and morphology of red blood cells (World Health Organization, 2011), the Centre for Disease and Prevention (CDC), the International Classification of Diseases (ICD) classification system, the haematological system, as well as the CLSI classification which accounts for both the red blood cell indices and laboratory findings (Wayne, 2008). Other classification systems have also been proposed, like the New York Heart Association (NYHA) classification for anaemia in heart failure, and the Severity of Anaemia Score (SAS) for predicting the outcome of critically ill patients with anaemia (Walsh et al., 2010; J. D. Fisher, 1972).

According to Cascio & DeLoughery (2017), anaemia can be classified using two different classification systems which are based on red cell size and underlying mechanism of anaemia. The initial system relies on Wintrobe's findings that variations in the size of red blood cells can help in distinguishing probable causes of anaemia, and is divided into microcytic, normocytic, and macrocytic anaemias. Microcytic anaemias, identifiable by a less than normal mean corpuscular volume (<80fL), reflect defects in haemoglobin synthesis caused by a lack of iron, thalassemia, or sideroblastic anaemias. Macrocytic anaemias (MCV > 100fL) are caused by

membrane defects or synthesis defects in the synthesis of DNA as witnessed in the cases of megaloblastic anaemia and chemotherapy-induced anaemia. Macrocytic anaemia can also be caused by the presence of a reticulocytosis, which results in an elevated MCV in the context of haemolysis. However, it can be difficult to distinguish potential etiologies for anaemia drawing on the size of red cells by dint of classification, since red cells can demonstrate normal size at the onset or in cases where multiple processes occur concurrently.

The second classification system is based on the underlying mechanism of anaemia, which is whether there is a rise in red cell loss or fall in red cell. The initial divergence point is determined by the reticulocyte count, which determines whether red blood cell production rose or fell. If red cell production rises, the primary consideration would be haemolysis and blood loss. On the other hand, if there is a fall in red cell production, it is important to consider primary agents of impaired production of bone marrow like deficiencies due to insufficient nutrient intake, marrow failure, lack of growth factors, or myelophthisic processes caused by cancer or infection (Cascio & DeLoughery, 2017). These classification systems have been widely used and adopted in the medical field.

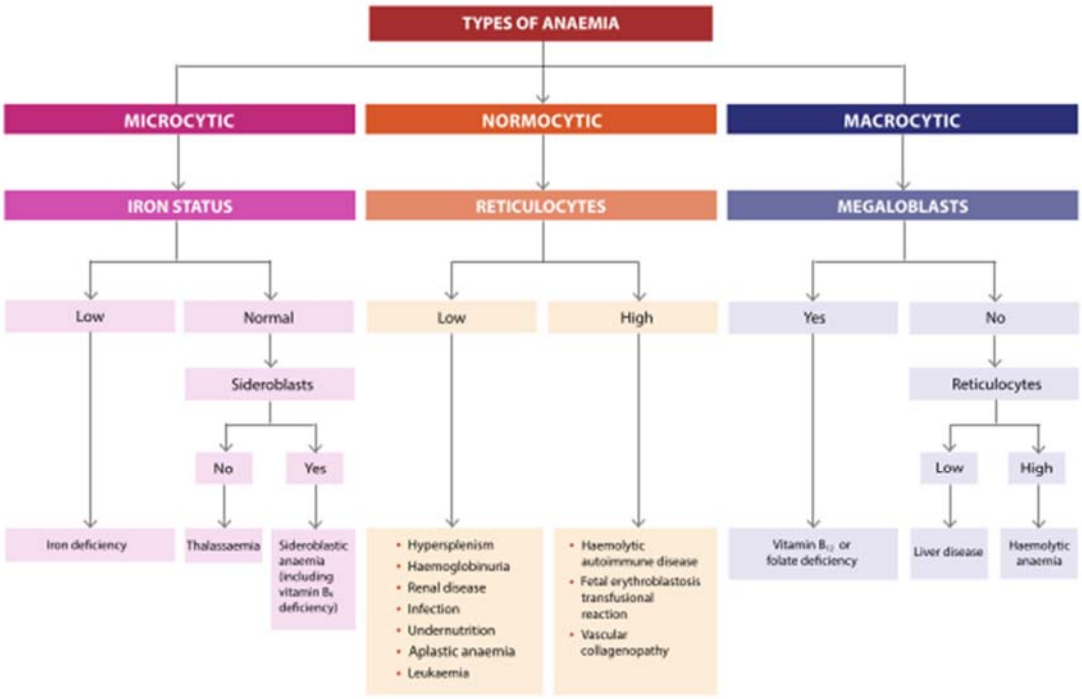


Figure 2. 1: Types of anaemia (World Health Organization, 2017)

2.2.1. Anaemia Classification as a Public Health Concern

Categorization of anaemia as a population issue typically involves categorizing according to its prevalence and severity in a specific population. This includes determining the prevalence of anaemia or the proportion of the population that is affected, and the severity of anaemia or the degree of reduction in haemoglobin levels, as well as identifying the specific causes of anaemia in that population. This classification requires a comprehensive understanding of the prevalence, severity, and underlying cause in a particular population. Chaparro & Suchdev (2019) in their study reviewed literature on anaemia including its definitions, classifications, epidemiology, causes, and public health significance. They relied on three sources: the WHO, a global analysis of the burden of anaemia and analysis from the BRINDA project. They also consulted gray literature, including documents from international organizations (Chaparro & Suchdev, 2019).

2.3. Techniques Used in the Classification of Anaemia

Various approaches leveraging machine learning have been explored in anaemia classification. These technologies use algorithms to learn from data and to identify classes to categorize anaemia. In this study, we will focus on data mining techniques, decision trees, fuzzy logic, k-Nearest Neighbour, artificial neural networks, and neuro-fuzzy network.

2.3.1. Data Mining Techniques

Sanap et al. (2011) in a study to analyse the conduct of different categorization techniques on an anaemia dataset, started by collecting a dataset of anaemia patients and healthy individuals and then pre-processed the data by imputing missing values and normalizing the attribute values in preparation for analysis. The dataset contained 424 instances and 10 attributes related to anaemia, like level of haemoglobin, mean corpuscular haemoglobin, red blood cell count, mean corpuscular volume, etc. Sanap et al. (2011) used feature selection techniques to identify the most relevant attributes for classification. Sanap et al. (2011) then applied various approaches of mining data, like support vector machines, naive bayes, decision trees and k-Nearest neighbours, to classify the anaemia dataset. The results were evaluated and compared using metrics like specificity, sensitivity, accuracy and area under the curve (AUC). These results showed that naive bayes and decision trees had the best performance with an accuracy of 99% and 98%, respectively. Support vector machines and k-Nearest Neighbours also performed well, with an accuracy of 97% and 96%, respectively. Sanap et al. (2011) also conducted an analysis of the important attributes for classification and realized that mean corpuscular haemoglobin, mean corpuscular volume, and haemoglobin level were the most important attributes for classifying anaemia (Sanap et al., 2011).

2.3.2. Decision Trees

Setsirichok et al. (2012) proposed leveraging computational techniques based on machine learning to classify haemoglobin typing data and complete blood count data to screen for thalassemia, a genetic blood disorder. Setsirichok et al. (2012) evaluated the performance of three classification algorithms –multilayer perceptron, naive bayes classifier, as well as C4.5

decision trees– in a dataset consisting of 422 haemoglobin typing data as well as complete blood count data samples. The dataset was split into training and testing data after which the classifiers were trained on the training data and tested using the testing data. Results showed that the C4.5 decision tree algorithm performed better than the other two algorithms with an accuracy of 98.2%. The naive Bayes classifier and multilayer perceptron achieved accuracies of 96.7% and 94.3%, respectively. The authors also noted that the C4.5 decision tree had better sensitivity and specificity than the other two algorithms (Setsirichok et al., 2012).

2.3.3. Fuzzy Logic

Shaik & Subashini (2017) proposed a fuzzy logic-based method to identify types of anaemia using LabVIEW software, to overcome the limitations of traditional methods that depended on laboratory equipment and required skilled professionals to operate. Shaik & Subashini (2017) used a membership function to define the anaemia grades and a rule base to determine the final diagnosis. The membership functions were used to map the input variables to fuzzy sets that represented the degree of membership of the input variable to a specific category. The input parameters used were haemoglobin (HGB), haematocrit (HCT), white blood cells (WBC), mean corpuscular haemoglobin concentration (MCHC), mean corpuscular volume (MCV), mean corpuscular volume (MCV), serum iron, total iron binding capacity (TIBC), and hyper-segmented white cell (HSWC) laboratory tests. The output parameters were six types of anaemia: iron deficiency anaemias, myelophitistic anaemia, chronic anaemia, megaloblastic anaemia and aplastic anaemia (AA). The rule base was used to relate the fuzzy sets of input variables to the fuzzy set of output variables. Shaik & Subashini (2017) compared the proposed method to two other methods: one that used traditional decision tree algorithm and another that used a neural network. The proposed method achieved higher accuracy than the two other methods. Shaik & Subashini (2017) also showed that the proposed method could diagnose anaemia in real-time, which was an advantage over traditional methods that require considerable time and resources (Shaik & Subashini, 2017). However, the study has limitations like the small sample size and the lack of validation in different populations or settings.

2.3.4. k-Nearest Neighbor

Kovačević et al. (2022) successfully used k-Nearest Neighbour (kNN) algorithm in categorization of anaemia based on haemoglobin concentrations and MCV. They employed a dataset obtained from NHANES 2003-2004, that contained 5098 records of individuals with or without anaemia. The kNN algorithm was applied in the classification of cases into anaemic and non-anaemic as a consideration of haemoglobin and MCV values. Kovačević et al. (2022) first applied kNN algorithm with k=3 and measured its performance using metrics like sensitivity, specificity, and accuracy, and found that the kNN algorithm achieved an overall accuracy of 95.6% in classifying anaemic and non-anaemic cases. Kovačević et al. (2022) conducted further experiments with different values of k and found that the performance improved with larger values of k and concluded that kNN algorithm is a promising tool for the diagnosis and classification of anaemia.

2.3.5. Artificial Neural Networks

Jamei & Talarposhti (2016) developed a method for distinguishing between the iron deficiency anaemia and β-Thalassemia trait employing an artificial neural network (ANN) algorithm with a pattern-based input selection approach by incorporating (ANN) with a pattern-based input selection (PBIS) approach, by incorporating the decision-making capability of artificial neural networks with a human expert. Jamei & Talarposhti (2016) collected blood samples from patients diagnosed with IDA or β-TT, and analysed the samples for various blood parameters, that were used as input features for the ANN which was trained using PBIS approach to identify the most significant variables for discrimination. The most notable features for discrimination were found to be haemoglobin, red blood cell distribution width, and mean corpuscular volume. The model's performance was assessed using specificity, sensitivity, and accuracy metrics, then compared to other models like support vector machines, k-nearest neighbour, and decision trees. Results indicated that ANN-PBIS model had better performance than the other models, regarding accuracy, and had a specificity of 95.83% and a sensitivity of 95.75% (Jamei & Talarposhti, 2016).

2.3.6. Neuro-Fuzzy Network

Allahverdi et al. (2011) presented an approach to determine anaemia in children based on a neuro-fuzzy network of the Takagi-Sugeno type, after noting that anaemia was among the frequently encountered nutritional issues worldwide, especially amongst developing countries where it affects children under-5. The development of accurate and reliable methods is crucial to prevent and treat this disease. Allahverdi et al. (2011) proposed a method that combined the strengths of both neural networks and fuzzy-logic to develop a hybrid Takagi-Sugeno type neuro-fuzzy network. The approach involved three steps: data preprocessing, rule generation, and model construction. In data preprocessing, the input and output variables were normalized to avoid any bias. In rule generation, the authors used the algorithm for fuzzy C-means clustering to obtain a set of fuzzy rules built upon the input-output relationships in the training data. Allahverdi et al. (2011) finally used the neuro-fuzzy network of the Takagi-Sugeno type to construct the final model. Allahverdi et al. (2011) tested the model on a dataset consisting of 200 children with and without anaemia, and reported a specificity of 98.5%, a sensitivity of 94.5% and an accuracy of 96.5%. These findings indicate that the model had a better performance than other techniques like artificial neural networks and support vector machines (Allahverdi et al., 2011). Although the use of a hybrid approach combining the strengths of both fuzzy logic and neural networks showed great promise, the dataset used in the study was small thus limiting validation of the proposed approach.

2.4. Related Works

2.4.1. Artificial Learning Methods Classifier for Anaemia Types

Karagul Yildiz et al. (2021), presented a study which aims to classify various categories of anaemia by leveraging artificial learning techniques. Karagül Yıldız et al. (2021) pointed out the importance of early detection of anaemia as well as proper management and treatment, and highlighted the difficulty in differentiating between several types of anaemia based on traditional laboratory methods. Karagül Yıldız et al. (2021) used a dataset consisting of 420 patients, with four distinct categories of anaemia: aplastic anaemia, haemolytic anaemia, iron deficiency anaemia and vitamin B12 deficiency. The dataset contained 12 features including age, gender, haemoglobin level, mean corpuscular volume as well as red blood cell count. Karagül Yıldız et al. (2021) used four different machine learning algorithms to classify anaemia types: artificial

neural networks, support vector machines, decision trees, and k-nearest neighbour. The performance of each algorithm was evaluated using various metrics like sensitivity, specificity, area under the curve of the receiver operating characteristic curve, and accuracy. The results showed that all four algorithms performed well in classifying anaemia types. The SVM and ANN algorithms achieved the highest accuracy rates of 91.7% and 91.2%, respectively. The DT and kNN algorithms had slightly lower accuracy rates but still achieved satisfactory performance. Karagül Yıldız et al. (2021) also conducted an analysis to identify the incredibly important attributes of anaemia classification, finding that mean corpuscular volume, red blood cell count, and haemoglobin concentration were the most notable features (Karagül Yıldız et al., 2021).

The study is a demonstration of the potential of artificial learning methods in accurately classifying several types of anaemia with the results suggesting that techniques of machine learning are usable as a complementary tool to traditional laboratory methods for anaemia diagnosis and may improve the accuracy of diagnosis and selection of appropriate treatment. However, additional studies are recommended for validation of the results in larger and more diverse datasets.

2.4.2. Deep Learning and Genetic Algorithms for Nutritional Anaemia Classification

Kilicarslan et al. (2021) proposed a hybrid model that combined deep learning and genetic algorithms in the classification of nutritional anaemias, since nutritional anaemia is a widespread problem in developing countries, and early detection can help prevent complication. The proposed hybrid approach used genetic algorithms for the optimization of hyperparameters for the deep learning model, thus improving accuracy of the classification. The convolutional neural network approach was used in the deep learning model since it is a popular choice for image classification tasks. Karagül Yıldız et al. (2021) used a dataset consisting of images of blood smears, which were analysed to identify the occurrence of anaemia. The model's performance was assessed using metrics like F-score, specificity, sensitivity, and accuracy. Results showed that the hybrid approach outperformed the individual deep learning model and conventional approaches leveraging machine learning like random forests and support vector machines and achieved an accuracy score of 94.35%.

According to the results, a hybrid model can be an effective tool for early detection and classification of diverse types of nutritional anaemia including folate deficiency, B12 deficiency and iron deficiency anaemia (Kilicarslan et al., 2021). However, the study did not perform an evaluation and comparison of the presented model with other advanced categorization models, which could have provided additional insights into the proposed model's performance.

2.4.3. Application of Artificial Intelligence in Diagnosis and Classification of Anaemia

Kovačević et al. (2022) presented a research investigation into the utilization of machine learning methods in the detection as well as the classification of anaemia. Kovačević et al. (2022) presents a comprehensive examination of traditional diagnosis and classification of anaemia and compared the performances of these techniques with traditional methods. Kovačević et al. (2022) detailed the dataset used in the study, including the selection criteria and the various features extracted from the blood samples, and then performed a performance comparison of computational techniques based on machine learning like support vector machines, k-nearest neighbour, and decision trees, with traditional classification methods. The results of the study demonstrate that machine learning techniques perform better than traditional classification in terms of accuracy, sensitivity, and specificity. Support vector machine yielded the highest level of accuracy of 97.25% as well as the highest F1 score of 0.97, an indication of a high degree of precision and recall (Kovačević et al., 2022). The study however only used a single dataset thus limiting validation as it is not representative of the wider population and did not consider the cost-effectiveness of implementing machine techniques in clinical settings.

2.5. Research Gap

When reviewing studies related to anaemia, it is observed that there is a notable gap in research focused on nutritional anaemia using digitized 1D datasets. Most existing studies rely on small datasets, limiting their generalizability and robustness of their findings. Utilizing small datasets, exacerbates the challenge of achieving reliable classification results. Moreover, many existing approaches do not fully leverage advanced data analytics and machine learning techniques that could enhance accuracy and efficiency of anaemia classification. The scarcity of extensive, high-quality datasets and the underperformance of classification models highlights the need for

comprehensive research that employs large scale, digitized dataset and advanced analytical techniques to develop a more accurate and scalable classification system for nutritional anaemia. Addressing this gap can lead to improved diagnostic tools and better patient outcomes by providing more precise and individualized classification of nutritional anaemia.

2.6. Conceptual Framework

This study proposes the development of a model for classifying nutritional anaemia types using machine learning techniques. Figure 2.2 below represents the conceptual framework for the classifier model:

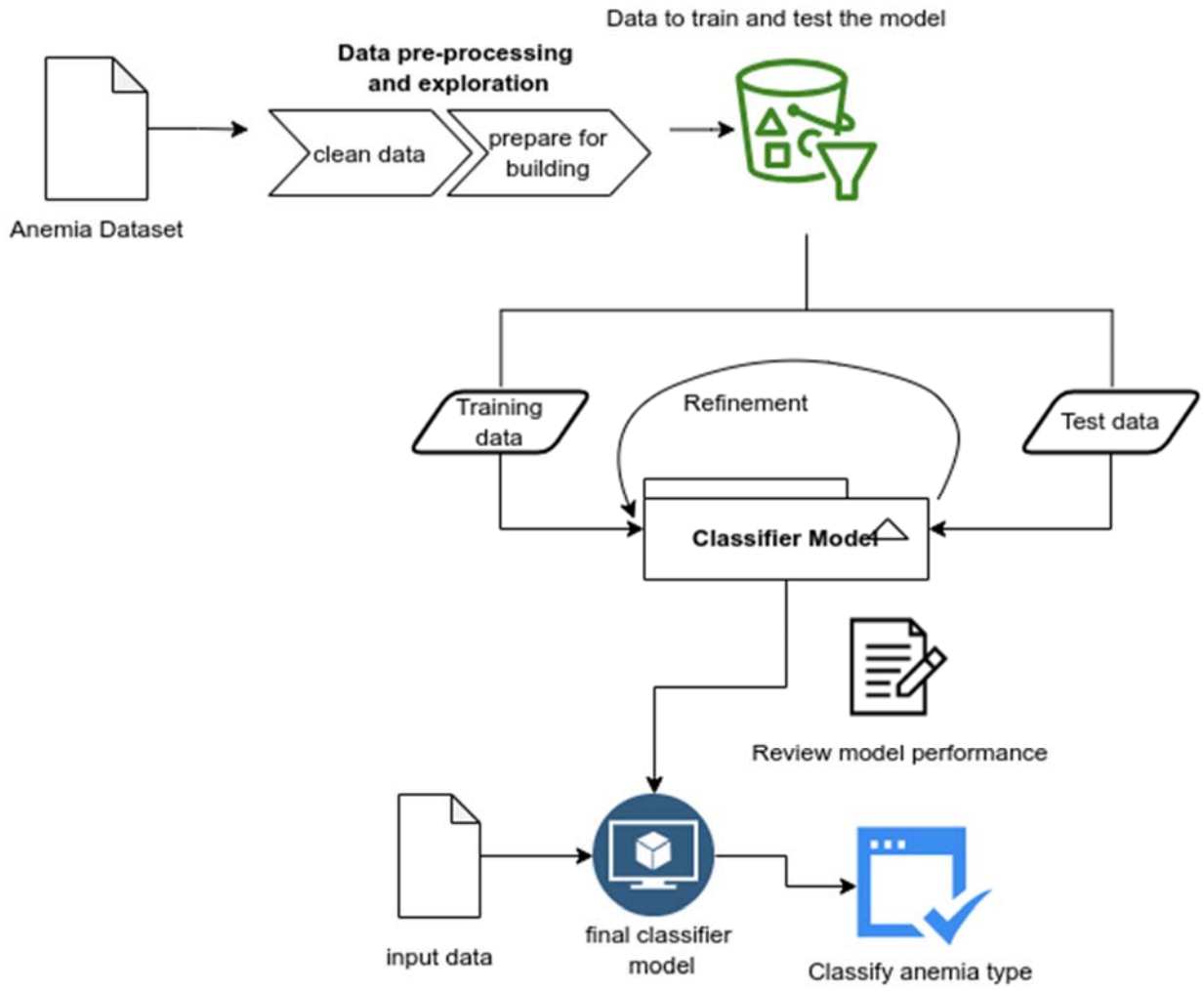


Figure 2. 2: Conceptual Framework for the Classifier

Chapter 3: Research Methodology

3.1. Introduction

Research methodology pertains to the structured exploration of the research procedure, from the initial planning phase to the final reporting of findings. It encompasses the principles and philosophy that underpin research methods. Research methods refer to the specific techniques and procedures that a researcher uses to achieve their research objectives (Thomas, 2021). This chapter detailed the different techniques and approaches that were utilized while conducting the research. The research design that was found most suited for this study was explored, as well as procedures for data collection and analysis. The system development methodology was equally examined. This research was directed by the stipulated objectives for fulfilment by the research.

3.2. Research Design

One of the most important steps in providing direction to the research problem is establishing a research design. This design encompasses the overall plan that outlines various aspects like the type of study, data collection methods, experimental designs, and statistical techniques for analysing data samples (Bairagi & Munot, 2019). In this study, experimental research design was employed. The data was experimented on while trying to make it suitable for developing the models. The models were tested by varying various parameters to find the best parameters for optimum model performance. The features or parameters were tested against each other to analyse and understand the relationship and correlation between them. The data was randomly assigned into training, and testing sets. Various algorithms were fit to the data to check for the one with the best performance. Various performance metrics were compared to find the most optimum metrics and the better performing model.

3.3. Data Collection

The data that was employed in this study, was data collected over a 5-year interval from 2013 to 2018 by Tokat Gaziosmanpaşa University, Faculty of Medicine, Turkey. The data was retrieved from Kaggle, an open-source repository on https://www.kaggle.com/datasets/serhathoca/anemia-disease/download?datasetVersionNumber=2. The dataset was retrieved in excel document

format and converted to comma separated or csv for easier access and manipulation in the notebook. The data contained the complete blood count test results of 15300 patients, of which 10,379 were female and 4921 were male. Of all the patients, 1019 have anaemia based on the haemoglobin values, 4182 have iron deficiency, 199 have B12 deficiency, 153 have folate deficiency and 9747 have no anaemia. The dataset excluded pregnant women, children, and patients with cancer. The dataset was employed to improve the concept of nutritional anaemia type classification, by using the anonymized participants provided in the dataset to fit various machine learning model to improve the understanding and research in the field. The dataset consisted of 24 features that are gender, haemoglobin, red cell distribution width, mean corpuscular volume, platelets, mean corpuscular haemolglobin concentration, red blood cells, mean corpuscular haemoglobin, mean platelet volume, hematocrit, neutrophils, basophils, eozinofiller, ferritin, folate, lenfosit, monositler, neutrophils, B12, serum iron, white blood cells, platelet, total serum iron, and transferrin saturation (SDTSD) obtained by the formula: SDTSD = (SD/TSD) * 100

3.4. Agile Dynamic System Development Methodology

Agile Dynamic System Development Method (DSDM) was employed in this study. DSDM is an agile approach to system development that uses the rapid application development approach (RAD) (Alsaqqa et al., 2020). It is an iterative and incremental approach that focuses on delivering functional and tested systems in a timely manner.

In the context of the study, the agile DSDM approach involved these stages:

3.4.1. Planning and Requirement Analysis Phase

In this phase, relevant secondary data was collected. The data source for this secondary data was Kaggle which is an open-source dataset repository. The dataset was downloaded from https://www.kaggle.com/datasets/serhathoca/anemia-disease/download?datasetVersionNumber=2. Although the repository is open source, an account was needed to access the dataset. The dataset was downloadable in excel format (.xlsx). The dataset was converted into comma separated or csv before it was applied in the development of

the classifier models. The model was built using Tensorflow and Keras among other Python libraries.

3.4.2. Designing Phase

In this phase, the designs of the classifier model were developed. The architecture of the system was also developed. Unified Modelling Language (UML) diagrams that describe the flow of information and the different interactions between the various components were also designed. The various UML diagrams were drawn using diagrams.net, which was previously draw.io, a free online diagram software.

3.4.3. Building/Development Phase

In the building or development phase, classification models were built. Designs from the previous design phase were converted into a working model. The classifier models were developed using Python programming language, using TensorFlow and Keras as well as other open-source Python libraries. The data acquired in the requirements analysis was pre-processed by checking for missing values and outliers and found that there were none. The parameters were plotted against each other to check for correlation and no parameters or variables were discarded. The data was investigated and discovered to be imbalanced. Synthetic Minority Over-sampling Technique (SMOTE) was used to balance the dataset by generating synthetic samples for the minority classes. The data was then split into training and testing sets in the ratio of 80-20 with 80% being for training and 20% for testing the models. The data was then normalized, to make it better suited for analysis since there were features that were significantly larger than others. The models were trained to fit the training data and was tested using the testing data. For some the models depending on performance, the parameters were tuned to improve the model's performance, for instance the number of layers was increased for the multilayer perceptron as well as the input nodes. Other parameters like batch size and number of epochs were also varied to try and improve its performance.

3.4.4. Testing Phase

In the testing phase, the classifier models were tested, and their performances were monitored, to check model behaviour when exposed to new data. The testing data was fed into the models and the predictions compared with the actual values. The models and their performance were measured. Performance metrics that were used to check the performance of the models are accuracy, precision, recall and F1 scores. The model was then loaded to the system, which was then tested to ensure functionality in the classification of nutritional anaemia in a user setting.

3.5. Data Analysis

The analysis of data in this study involved cleaning the data by checking for missing and null values, comparing the correlation between the various parameters to find the insignificant or redundant parameters, reducing the noise in the data, resampling the dataset to generate synthetic samples for the minority classes, and normalizing the data to make it better suited for using in building the model. The data was used as was retrieved from the source, except for the preprocessing as indicated. The manipulation was of course guided by the terms stipulated in the license agreement.

3.6. Model Development

The classifier models were developed through these steps:

i. Data pre-processing – The collected data was cleaned, resampled, and normalized to make it more suitable for model building.

ii. Data splitting – The data was split into testing, and training data in 80-20 interval: 80% for training and 20% for testing.

iii. Model training – The models were trained to fit the data using various machine learning algorithms.

iv. Model testing – The generated models were tested against the testing data and were measured for their performance.

3.7. Research Quality

Validity and reliability were utilized in evaluation of the study's quality. The model's performance was evaluated to check the accuracy of the model. The performance metrics that were measured include precision, recall, accuracy, and F1-score. True Positive (TP) is the number of samples identified as belonging to a specific class, and they actually belong to that class, whereas True Negative (TN) is the number of samples identified as not belonging to a certain class, and they actually do not belong to that class. False Positive is the number of samples identified as belonging to a certain class, yet they belong to a different class, and False negative is the number of samples identified as not belonging to a certain class, yet they belong to that class. These performance metrics were computed as below:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Equation 3. 1: Accuracy Metrics (Karagül Yıldız et al., 2021)

$$\text{Recall} = \frac{TP}{TP+FN}$$

Equation 3. 2: Recall Metrics (Karagül Yıldız et al., 2021)

$$\text{Precision} = \frac{TP}{TP+FP}$$

Equation 3. 3: Precision Metrics (Karagül Yıldız et al., 2021)

$$\text{F1-Score} = \frac{2*P*R}{P+R}$$

Equation 3. 4: F1 Score (Karagül Yıldız et al., 2021)

where P is precision and R is Recall. The definition of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) was critical in better analysing the models' performance.

3.8. Ethical Considerations

The study utilized secondary data retrieved from Kaggle open-source datasets in accordance with the Open Data Commons Open Database License that can be retrieved from http://opendatacommons.org/licenses/odbl/1.0/. The data was used in accordance with the terms stipulated in the license. The data was solely employed for research purposes. All cited authors were acknowledged and given credit for their work and to avoid plagiarism.

Chapter 4: System Analysis, Design and Architecture

4.1. Introduction

In this section the overall architecture and design of the classifier was incorporated with various requirements. UML diagrams were used to: describe the overall architecture of the system and to give detailed descriptions of the various model components. This was modelled by use of class diagrams, use case diagrams, system sequence diagrams and entity relationship diagrams.

4.2. System Analysis

System analysis and design is a systematic approach to developing high-quality information systems (Tilley & Rosenblatt, 2017). System analysis involves understanding the components and processes by analysing data, identifying system requirements and creating UML models and diagrams. The requirements gathered for this study are categorized into; functional, and non-functional requirements.

4.2.1. Functional Requirements

Functional requirements describe the behaviour of the system. The functional requirements for the classification tool are:

i. The classifier should register a new user.
ii. The classifier should accept the complete blood count data inputs entered manually.
iii. The classifier should classify the input data as belonging to one of the five classes.
iv. The classifier should inform the user what class of anaemia the input data belongs in.

4.2.2. Non-Functional Requirements

Non-functional requirements for the model include:

i. Performance – The classifier provides a class prediction within minutes.

ii. Maintainability – The classifier is easy to maintain and support.

iii. Security – The classifier is secure from alterations by unauthorized personnel.

iv. Reliability – The classifier is set with certain parameters that enable it to give consistent results when provided with the same input data.

4.3. System Architecture

Various machine learning models were developed, and their performances compared to find the best performing one for use in testing the system functionality. The dataset used in the study was prepared for model development, and then used to train the models. After training, the various models were tested, and their performances checked and compared. The best performing model was loaded into the system to test its functionality. The user enters the complete blood count test results and the class of nutritional anaemia of the data is predicted.

The architecture of the system is as shown in Figure 4.1 below:

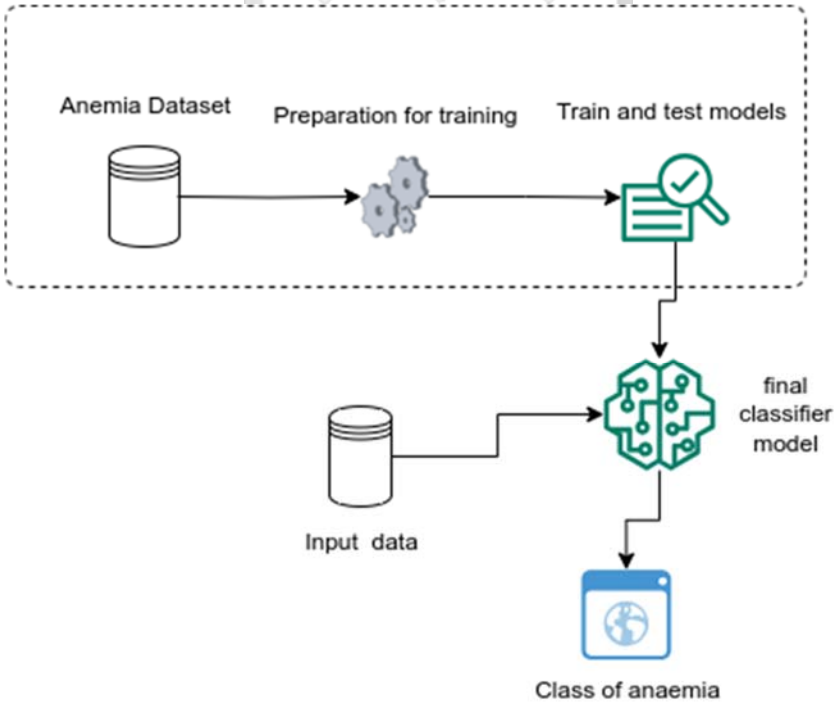


Figure 4. 1: System Architecture

4.4. System Design

4.4.1. Use Case Diagram

Use case diagram is a representation of the requirements of the system and the interactions between the system and its external entities. The main actors are the administrator and the user. The administrator builds the model that is used to make the classifications whereas the user enters the complete blood count test results through the interface to get a classification from the model. The use case diagram of our model is as detailed in Figure 4.2 below:

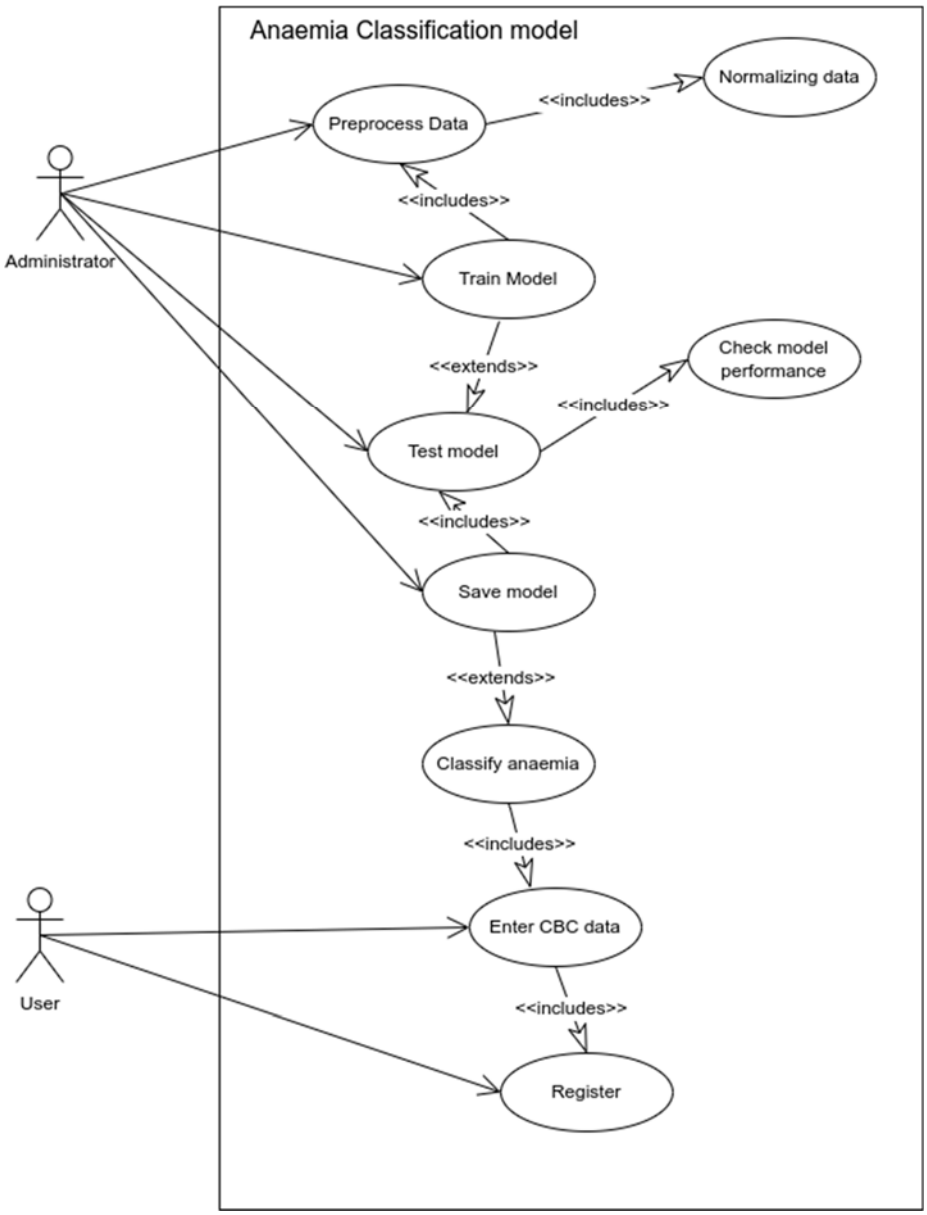


Figure 4. 2: Use Case Diagram

The actors and their roles in the system are defined as shown in Table 4.1 below :

Table 4. 1: Use Case Description

Actor	**Use Case**	**Description**
User	Register	The user must first register by providing the necessary information, in order to use the classifier.
	Enter CBC data	A registered user should input data, which is the complete blood count test results and submit it to the classifier through the interface, in order to get the classification.
Administrator	Preprocess data	The administrator should preprocess the data ahead of using it to train the models, by for example normalizing it.
	Train model	The administrator uses the preprocessed data to train the models.
	Test model	After training the model, the administrator tests its performance and compares to the other models to get the best performing model.
	Save model	The best performing model is saved and loaded to the classifier for use in classifying the class of anaemia, based on the input data entered by the user.

4.4.2. Context Diagram

Context diagrams visually detail the scope of the developed model and illustrate how the information flows between the model and external entities. In a context diagram, the whole system is typically depicted as a single process, also known as a level 0 data-flow diagram. The model consisted of two major user level: regular users and the administrator. Both user types interacted with the model by providing various data items and instructions and receiving different results. The context diagram for the model is presented in Figure 4.3 below:

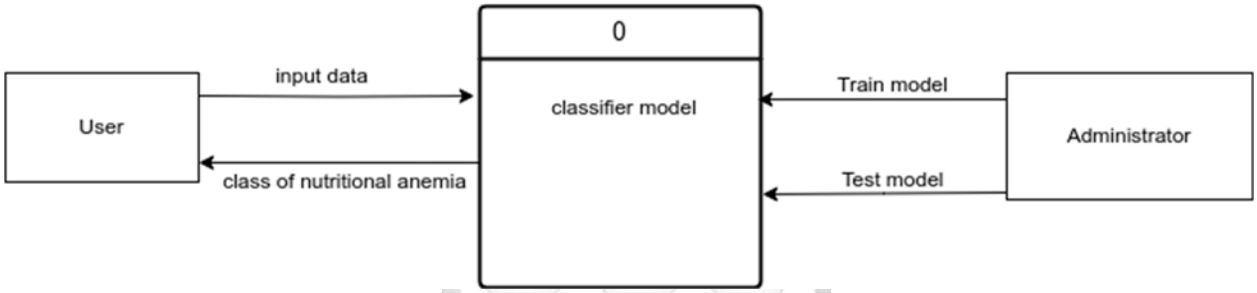


Figure 4. 3: Context Diagram

4.4.3. Sequence Diagram

Sequence Diagrams illustrated sequentially, how the system achieved its specific objectives. It shows the various interactions between the internal components. Before training the model, the administrator loads our dataset for preprocessing. Once the dataset is preprocessed, it is used to train the model. After training, the model is tested to check its performance. The performances of the various models are compared, and the best performing model is saved. After saving, the model is loaded into the interface. The user then logs into the system and inputs the results of the complete blood count test. The results are loaded into the model and the results are shown to the user. Figure 4.4 below presents the sequence diagram for the model:

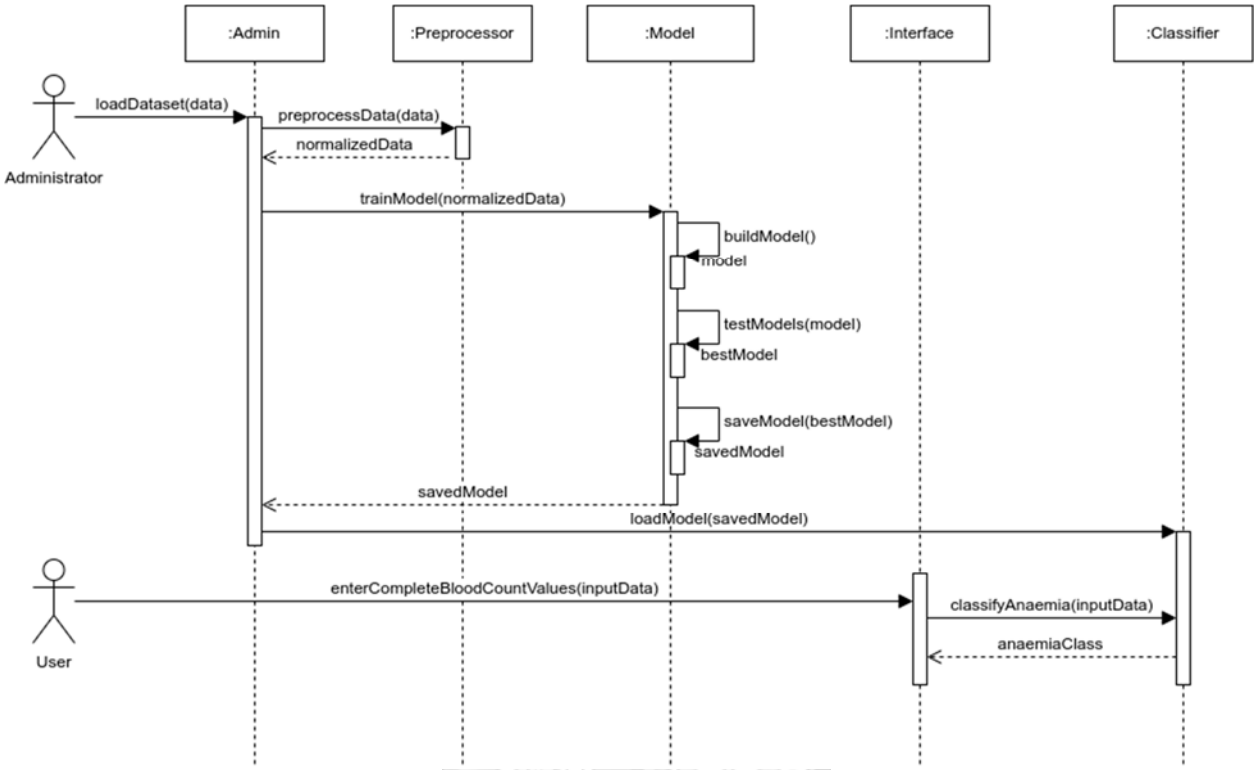


Figure 4. 4: Sequence Diagram

4.4.4. Entity Relationship Diagram

Entity relationship diagram is a graphical representation of entities, and their relationship with each other, representative of a database. It shows the data model and helps to visualize how data is organized and related in the system. The entity relationship diagram is a shown in Figure 4.5 below:

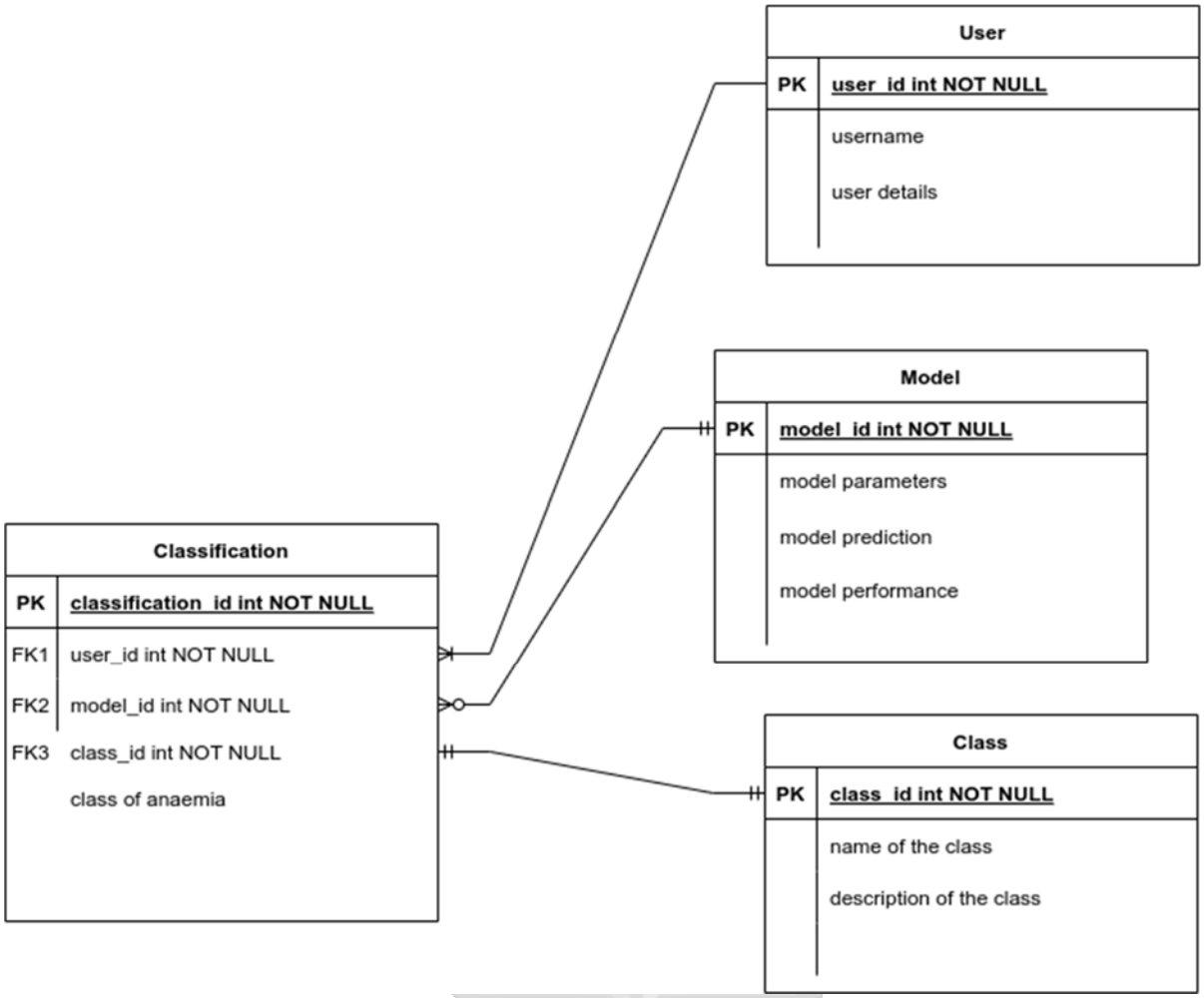


Figure 4. 5: Entity Relationship Diagram

Chapter 5: System Implementation and Testing

5.1. Introduction

The study was aimed at building a classification model for classifying types of nutritional anaemia, specifically into; no anaemia, haemoglobin anaemia, iron deficiency anaemia, folate anaemia, and B12 anaemia, by using supervised machine learning techniques. This chapter gives an overview of how the classification models to classify types of nutritional anaemia using supervised learning techniques were built/developed, trained, and tested. Details of what environment was used to develop the models are given, as well as the data pre-processed methods and techniques. The procedures followed in the development of the model are detailed.

5.2. Development Environment

The classification model was developed on Docker desktop, which is a desktop application that allows for the management of Docker containers, images, and other resources. It provides access to powerful Python libraries like Keras and Tensorflow, which were crucial in the development of the classifier models. Other libraries that were used in the development of the models are scikit-learn, pandas, numpy, xgboost, matplotlib, and seaborn. The interface was developed using the Flask python library as well as html and css. All these were done on a personal computer.

5.3. Hardware Resources

Developing the classifier model utilized these hardware resources:

Table 5. 1: Hardware Resources

Hardware	**Specifications**	
Personal Computer	RAM	24 GB
	CPU	Intel Core i5-8250U CPU @ 1.60GHz × 8
	SSD	512 GB

Hardware	**Specifications**	
Docker Desktop	RAM	5.62 GB
	Disk	67.32 GB

5.4. Software Resources

The software resources used in the development included Python, which was the primary programming language, as well as the libraries used in the development, as summarized below:

Table 5. 2: Software Resources

Software	**Library**	**Version**
Python 3.9.16	TensorFlow	2.14.0
	Keras	2.12.0
	Pandas	1.5.3
	NumPy	1.24.3
	Scikit-learn	1.4.2
	XGBoost	2.0.3
	Imbalanced-learn	0.12.2
	Matplotlib	3.7.1
	Seaborn	0.12.2

	Flask	3.0.3

5.5. Data Pre-processing and Exploration

The data was loaded into the notebook and checked for missing and null values. The correlation of the features with each other was determined by plotting the parameters against each other to show the correlation. The correlation was found to be low for most features, and as a result no parameters were removed or considered insignificant.

Figure 5. 1: Correlation Matrix

The dataset was resampled because of its imbalanced nature-samples were synthesized for the minority classes. The data was discovered to have some parameters with significantly larger values than others, which would affect the performance of the model. As a result, normalization was performed on the data.

```
#SMOTE resampling

smote = SMOTE(random_state=24)
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)
```

```
#Normalizing the data
stats = df.describe()

stats = stats.transpose()

#normalization function
def norm(x):
    return (x - stats['mean']) / stats['std']

df = norm(df)
df.head()
```

Figure 5. 2: Data Reshaping and Normalization

5.6. Model Training and Testing

The model train_test_split function of the scikit-learn library was utilised in splitting the data into training and testing data. 80% of the data was used in training the model with the remaining 20% being used for testing. Model testing was performed by comparing the predicted classes by the actual classes of the testing data.

The Decision Tree classifier was implemented as below:

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score, confusion_matrix


#Train  Decision Tree Classifier

dt_resamp = DecisionTreeClassifier(random_state=24)
dt_resamp.fit(X_train_resampled, y_train_resampled)

# Evaluate both classifiers
y_dt_pred_resamp = dt_resamp.predict(X_test)


print("\nEvaluation on resampled data:")
print(classification_report(y_test, y_dt_pred_resamp))
print(confusion_matrix(y_test, y_dt_pred_resamp))
accuracy = accuracy_score(y_test, y_dt_pred_resamp)
print('Accuracy: %f' % accuracy)
```

Figure 5. 3: Decision Forest

The Random Forest classifier was implemented as below:

```
from sklearn.ensemble import RandomForestClassifier

#Train Random Forest Classifier

rf_resamp = RandomForestClassifier(random_state=24)
rf_resamp.fit(X_train_resampled, y_train_resampled)

# Evaluate both classifiers
y_rf_pred_resamp = rf_resamp.predict(X_test)

print("\nEvaluation on resampled data:")
print(classification_report(y_test, y_rf_pred_resamp))
print(confusion_matrix(y_test, y_rf_pred_resamp))
accuracy = accuracy_score(y_test, y_rf_pred_resamp)
print('Accuracy: %f' % accuracy)
```

Figure 5. 4: Random Forest

The Naïve Bayes classifier was implemented as below:

```
#Naive Bayes Classifier
from sklearn.naive_bayes import GaussianNB

nb = GaussianNB()

nb_resamp = nb.fit(X_train_resampled, y_train_resampled)

ynb_pred_resamp = nb_resamp.predict(X_test)

print("\nEvaluation on resampled data:")
print(classification_report(y_test, ynb_pred_resamp))
print(confusion_matrix(y_test, ynb_pred_resamp))
accuracy = accuracy_score(y_test, ynb_pred_resamp)
print('Accuracy: %f' % accuracy)
```

Figure 5. 5: Naive Bayes

The XG Boost classifier was implemented as below:

```
#XG Boost Classifier
import xgboost as xgb

XGB = xgb.XGBClassifier(random_state=24)
xgb_resamp = XGB.fit(X_train_resampled, y_train_resampled)

y_xgb_pred_resamp = xgb_resamp.predict(X_test)

print("\nEvaluation on resampled data:")
print(classification_report(y_test, y_xgb_pred_resamp))
print(confusion_matrix(y_test, y_xgb_pred_resamp))
accuracy = accuracy_score(y_test, y_xgb_pred_resamp)
print('Accuracy: %f' % accuracy)
```

Figure 5. 6: XG Boost

The Multilayer Perceptron was implemented as below:

```python
#ANN(MLP) -- Resampled Data
from sklearn.preprocessing import StandardScaler
from tensorflow.keras.layers import Input


scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train_resampled)
X_test_scaled = scaler.transform(X_test)


mlp = Sequential([
    # Use an Input layer as the first layer, specifying the input shape
    Input(shape=(X_train_scaled.shape[1],)),

    Dense(32, input_shape=(X_train_scaled.shape[1])),
    Dense(32, activation='relu'),
    Dense(64, activation='relu'),
    Dense(128, activation='relu'),
    Dense(5, activation='softmax')
])

mlp.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])

history = mlp.fit(X_train_scaled, y_train_resampled, epochs=50, batch_size=32, validation_split=0.2)

test_loss, test_accuracy = mlp.evaluate(X_test_scaled, y_test)
print(f'Test Loss: {test_loss}, Test Accuracy: {test_accuracy}')

y_mlp_pred = mlp.predict(X_test_scaled)

y_mlp_pred_classes = np.argmax(y_mlp_pred, axis=1)

accuracy = accuracy_score(y_test, y_mlp_pred_classes)
print(f'Accuracy: {accuracy}')
print(confusion_matrix(y_test, y_mlp_pred_classes))
print(classification_report(y_test, y_mlp_pred_classes))
```

Figure 5. 7: Multilayer Perceptron

5.6.1. Hyper-parameter Tuning

Hyper-parameter tuning was performed on some of the models to improve their performance. Naive Bayes classifier performed the worst when loaded with the testing data. Hyper-parameter tuning was performed on it using the GridSearchCV package.

```
#Hyperparameter Tuning
from sklearn.model_selection import GridSearchCV


nb_param_grid = {
    'var_smoothing': [1e-9, 1e-8, 1e-7, 1e-6]
}

grid_search = GridSearchCV(nb, nb_param_grid, cv=5, scoring='accuracy')

grid_search.fit(X_train, y_train)

# Print the best hyperparameters and their score
print("Best var_smoothing:", grid_search.best_params_['var_smoothing'])
print("Best accuracy score:", grid_search.best_score_)

best_nb = grid_search.best_estimator_

y_pred_best = best_nb.predict(X_test)
print("Classification Report"),# accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred_best))
```

Figure 5. 8: Hyperparameter Tuning on Naïve Bayes

Even with hyperparameter tuning, the accuracy was 82.91% which is relatively low compared to the other classifiers.

```
Best var_smoothing: 1e-07
Best accuracy score: 0.8291666666666666
Classification Report
              precision    recall  f1-score   support

           0       0.89      0.91      0.90      1971
           1       0.50      0.29      0.37       200
           2       0.83      0.80      0.81       811
           3       0.37      0.46      0.41        39
           4       0.30      0.64      0.41        39

    accuracy                           0.83      3060
   macro avg       0.58      0.62      0.58      3060
weighted avg       0.83      0.83      0.83      3060
```

Figure 5. 9: Naive Bayes performance after hyperparameter tuning

5.7. Model Evaluation and Selection

The classifier models were compared using metrics like accuracy, precision, recall and F1 scores, as well as the confusion matrix. The Decision tree classifier achieved an accuracy of 97.15 %. The other metrics are as shown below:

```
Evaluation on resampled data:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98      1971
           1       0.86      0.99      0.92       200
           2       0.96      1.00      0.98       811
           3       0.93      0.95      0.94        39
           4       0.78      0.90      0.83        39

    accuracy                           0.97      3060
   macro avg       0.90      0.96      0.93      3060
weighted avg       0.97      0.97      0.97      3060

[[1892   31   35    3   10]
 [   0  199    1    0    0]
 [   1    0  810    0    0]
 [   1    1    0   37    0]
 [   4    0    0    0   35]]
Accuracy: 0.971569
```

Figure 5. 10: Decision Tree evaluation

The Random Forest classifier achieved an accuracy of 97.05 % with the other performance metrics as captured below:

```
Evaluation on resampled data:
              precision    recall  f1-score   support

           0       0.99      0.96      0.98      1971
           1       0.80      0.97      0.88       200
           2       0.99      1.00      0.99       811
           3       0.85      1.00      0.92        39
           4       0.70      0.97      0.82        39

    accuracy                           0.97      3060
   macro avg       0.87      0.98      0.92      3060
weighted avg       0.98      0.97      0.97      3060

[[1891   48    9    7   16]
 [   5  195    0    0    0]
 [   4    0  807    0    0]
 [   0    0    0   39    0]
 [   1    0    0    0   38]]
Accuracy: 0.970588
```

Figure 5. 11: Random Forest evaluation

The Naïve Bayes classifier achieved an accuracy of 78.72% with the other metrics as below:

```
Evaluation on resampled data:
              precision    recall  f1-score   support

           0       0.94      0.82      0.88      1971
           1       0.46      0.45      0.46       200
           2       0.76      0.79      0.78       811
           3       0.34      0.62      0.44        39
           4       0.15      0.90      0.25        39

    accuracy                           0.79      3060
   macro avg       0.53      0.71      0.56      3060
weighted avg       0.85      0.79      0.81      3060

[[1619   53  173   18  108]
 [  47   90   23   10   30]
 [  48   49  641   18   55]
 [   4    1    0   24   10]
 [   1    2    1    0   35]]
Accuracy: 0.787255
```

Figure 5. 12: Naive Bayes evaluation

The Multilayer Perceptron achieved an accuracy of 96.24% with the other metrics as below:

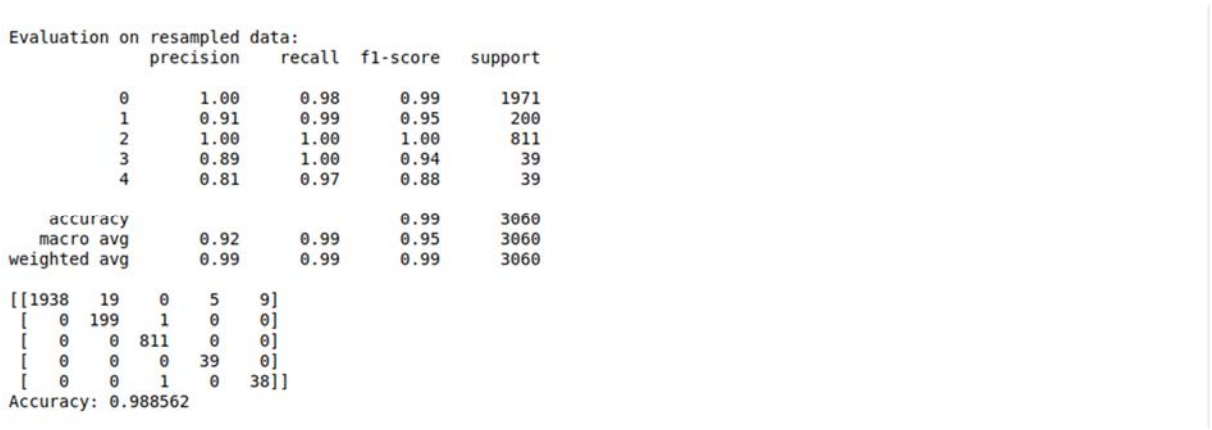
```
Accuracy: 0.9624183006535948
[[1965    0    6    0    0]
 [   5  158   21    4   12]
 [  11   15  779    5    1]
 [   1    9    0   27    2]
 [   3   12    8    0   16]]
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      1971
           1       0.81      0.79      0.80       200
           2       0.96      0.96      0.96       811
           3       0.75      0.69      0.72        39
           4       0.52      0.41      0.46        39

    accuracy                           0.96      3060
   macro avg       0.81      0.77      0.79      3060
weighted avg       0.96      0.96      0.96      3060
```

Figure 5. 13: Multilayer Perceptron evaluation

The XG Boost classifier achieved an accuracy of 98.88% with the other metrics as below:



```
Evaluation on resampled data:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99      1971
           1       0.91      0.99      0.95       200
           2       1.00      1.00      1.00       811
           3       0.89      1.00      0.94        39
           4       0.81      0.97      0.88        39

    accuracy                           0.99      3060
   macro avg       0.92      0.99      0.95      3060
weighted avg       0.99      0.99      0.99      3060

[[1938   19    0    5    9]
 [   0  199    1    0    0]
 [   0    0  811    0    0]
 [   0    0    0   39    0]
 [   0    0    1    0   38]]
Accuracy: 0.988562
```

Figure 5. 14: XG Boost evaluation

The XG Boost classifier was selected because of its performance compared to the other models.

5.8. System Testing

The selected model was loaded to the interface the user would interact with and then functional tests were conducted for registering, logging in and classification. Table 5.3 below details the functional tests conducted.

Table 5. 3: Functional Tests

No.	Component	Assumption	Test Case	Expected Results	Observed Results
1	Registration	User does not have an account	User provides their information	- Account information is saved - User is redirected to login	As expected
2	Sign/Log in	User has an account	User enters their username and password	User is redirected to landing page	As expected

3	Classification	User inputs complete blood count test data	User submits the input data and gets results	The class of anaemia is provided	As expected

Chapter 6: Discussion

6.1. Introduction

The aim of the study was to develop a classification model for classifying nutritional types of anaemia, specifically iron deficiency anaemia, folate anaemia, B12 anaemia, haemoglobin anaemia as well as no anaemia. This chapter discusses and reviews the realization of the research objectives as well as the findings of the study.

6.2. Results

With the advancement of technology in the present day, various techniques have been employed in the classification of anaemia, be it differential diagnosis of two specific anaemias, or the general classification of anaemia. These technologies and algorithms include data mining, decision trees, fuzzy logic, k-nearest neighbours, artificial neural networks, neuro-fuzzy networks, naive bayes, support vector machines, among others. These technologies and algorithms have been applied to various complete blood count datasets to build applications and models for detecting and categorizing anaemia based of different factors or parameters. The most utilized parameters because of their ease of access are haemoglobin, hematocrit, mean corpuscular volume, mean corpuscular haemoglobin concentration, red blood cell count, red blood cell width, platelet count among other. Some other studies/datasets provide more information/parameters like serum and ferritin concentration, specifically when detecting or monitoring iron stores. This data has been implemented in the early detection and differential diagnosis of anaemia, including studies like Jamie & Talarposhti (2016) (Jamei & Talarposhti, 2016).

There are a few studies that hold high significance in the classification of anaemia. One of the studies is by Karagul Yildiz et al. (2021) who utilized support vector machine, decision tree and k-nearest neighbour algorithms to classify 12 types of anaemia based on 1663 blood samples of patients. Support vector machine achieved the highest accuracy of 85.6% (Karagül Yıldız et al., 2021). Another study is by Kilicarlslan et al. (2021) who proposed two hybrid models utilizing genetic algorithm, and stacked autoencoder and convolutional neural network for prediction of

HGB anaemia, nutritional anaemia like iron deficiency, B12 and folate anaemias, and patients without anaemia. The proposed hybrid genetic algorithm and convolutional neural networks algorithm achieved an accuracy of 98.50% (Kilicarslan et al., 2021).

The dataset used in this study contained blood count data with 24 parameters which is higher than most previous studies. All the 24 parameters were used in the development of the model. A decision tree classifier, a random forest classifier, a naive bayes classifier, an XG Boost classifier, and a multilayer perceptron were all built, and their performances compared using metrics like accuracy, precision, recall and F1 scores. The XG Boost classifier performed the best in comparison with the others. The various models that were built were evaluated to check their performance. The metrics utilized in evaluating the performance are accuracy, sensitivity or recall, precision, F1-score, and confusion matrix since accuracy alone does not give a deep analysis of the model performances on new data. The Decision tree classifier when exposed to the test data performed well with an accuracy of 0.9715 and only a few samples from some of the classes being misclassified. The Random Forest classifier, also performed well with an accuracy of 0.9705, and even fewer samples being predicted as belonging in the incorrect classes. The Naïve Bayes performed poorly compared to the other models and managed an accuracy of 0.7872. Many classes were classified incorrectly, especially compared with the performance of the other models. The multilayer perceptron classifier performed relatively well, achieving an accuracy of 0.9624, and relatively many samples from each class being classified incorrectly. The XG Boost classifier gave the best performance attained an accuracy of 0.9885. Most of the classes were categorized correctly, with only a few being classified incorrectly. This was especially true for the minority classes thus making it the best performing classifier in comparison with the other models.

Chapter 7: Conclusions, Recommendations and Future Work

7.1. Conclusion

This study was aimed at developing a classification model for nutritional anaemia, specifically iron deficiency anaemia, B12 anaemia and folate anaemia, as well as haemoglobin anaemia and no anaemia, using machine learning algorithms. This was achieved by building and comparing the performance metrics of the various models to find the best performing one. The models utilized complete blood count data with 24 parameters: gender, haemoglobin, red cell distribution width, mean corpuscular volume, platelets, mean corpuscular haemoglobin concentration, red blood cells, mean corpuscular haemoglobin, mean platelet volume, hematocrit, neutrophils, basophils, eozinofiller, ferritin, folate, lenfosit, monositler, neutrophils, B12, serum iron, white blood cells, platelet, total serum iron, and transferrin saturation. The dataset used in the study was pre-processed, and then split into training and testing sets. The training set was utilized in the building of the model, while the testing set was fed into the models to evaluate their performances. Some model parameters were tuned to improve the performance of these models. Overall, the XG Boost model performed the best with an accuracy of 98.85%, with other metrics like precision, recall and F1-scores being 0.9897, 0.9885 and 0.9888, respectively. Using the XG boost model, the classifier tool was developed and tested with complete blood count values. If well implemented, this tool could potentially be plugged into hospitals for use by all patients, to aid in the early detection, diagnosis and treatment of nutritional anaemia. In clinics, it would help reduce the wait time between getting tested and getting results for complete blood count tests.

7.2. Limitations of the Study

One limitation of the dataset was that it was not balanced as some classes had very few datapoints as compared with others. Although synthetic datapoints were generated, it would be more authentic to use a balanced dataset. Another limitation is the study did not include the most vulnerable populations, including women and children. The dataset did not include patients with cancer, pregnant women and children. Despite these limitations, the study can be seen as one among many steps towards anaemia reduction as targeted by the World Health Assembly.

7.3. Recommendations

These recommendations were drawn from the study:

i. Training of the model be done with a more balanced dataset that does not include synthesized values for the minority classes.

ii. De-identified patient data from local settings be made more accessible with the help of policy makers and healthcare institutions.

iii. Using a more robust dataset which includes expectant women and children, and cancer patients.

7.4. Future Work

Current findings could be extended by:

i. Including vulnerable groups like pregnant women, young children and cancer patients in future studies.

ii. Acquiring more data, especially for the classes of nutritional anaemia with few datapoints.

iii. Including treatment or management information and recommendations.

References

Allahverdi, N., Tunali, A., Işik, H., & Kahramanli, H. (2011). A Takagi–Sugeno type neuro-fuzzy network for determining child anaemia. *Expert Systems with Applications*, *38*(6), 7415–7418. https://doi.org/10.1016/j.eswa.2010.12.083

Alsaqqa, S., Sawalha, S., & Abdel-Nabi, H. (2020). Agile Software Development: Methodologies and Trends. *International Journal of Interactive Mobile Technologies (IJIM)*, *14*(11), 246. https://doi.org/10.3991/ijim.v14i11.13269

Anderson, E. (1936). The Species Problem in Iris. *Annals of the Missouri Botanical Garden*, *23*(3), 457–509. https://doi.org/10.2307/2394164

Balarajan, Y., Ramakrishnan, U., Özaltin, E., Shankar, A. H., & Subramanian, S. V. (2011). Anaemia in low-income and middle-income countries. *The Lancet*, *378*(9809), 2123–2135. https://doi.org/10.1016/S0140-6736(10)62304-5

Banfield, J. D., & Raftery, A. E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, *49*(3), 803–821. https://doi.org/10.2307/2532201

Bensmail, H., & Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, *91*(436), 1743–1748.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152. https://doi.org/10.1145/130385.130401

Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108644181

Celeux, G., & Mkhadri, A. (1992). Discrete regularized discriminant analysis. *Statistics and Computing*, *2*, 143–151.

Chaparro, C. M., & Suchdev, P. S. (2019). Anaemia epidemiology, pathophysiology, and etiology in low- and middle-income countries. *Annals of the New York Academy of Sciences*, *1450*(1), 15–31. https://doi.org/10.1111/nyas.14092

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

FAO, I. (2021). *The State of Food Security and Nutrition in the World 2021: Transforming food systems for food security, improved nutrition and affordable healthy diets for all*. FAO. https://doi.org/10.4060/cb4474en

Fisher, J. D. (1972). New York Heart Association Classification. *Archives of Internal Medicine*, *129*(5), 836. https://doi.org/10.1001/archinte.1972.00320050160023

Fisher, R. A. (1936). Design of Experiments. *British Medical Journal*, *1*(3923), 554.

Fisher, R. A. (1937). The Wave of Advance of Advantageous Genes. *Annals of Eugenics*, *7*(4), 355–369. https://doi.org/10.1111/j.1469-1809.1937.tb02153.x

Fukunaga, K. (1999). Statistical pattern recognition. In *Handbook of Pattern Recognition and Computer Vision* (pp. 33–60). WORLD SCIENTIFIC. https://doi.org/10.1142/9789812384737_0002

Gardner, W., & Kassebaum, N. (2020). Global, Regional, and National Prevalence of Anaemia and Its Causes in 204 Countries and Territories, 1990–2019. *Current Developments in Nutrition*, *4*(Supplement_2), 830–830. https://doi.org/10.1093/cdn/nzaa053_035

Hastie, T., & Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 155–176.

Hess, S. Y., Wessells, K. R., Hinnouho, G.-M., Barffour, M. A., Sanchaisuriya, K., Arnold, C. D., Brown, K. H., Larson, C. P., Fucharoen, S., & Kounnavong, S. (2019). Iron status and

inherited haemoglobin disorders modify the effects of micronutrient powders on linear growth and morbidity among young Lao children in a double-blind randomised trial. *British Journal of Nutrition*, *122*(8), 895–909. https://doi.org/10.1017/S0007114519001715

Hotelling, H. (1931). The Economics of Exhaustible Resources. *Journal of Political Economy*, *39*(2), 137–175. https://doi.org/10.1086/254195

Jamei, M. K., & Talarposhti, K. M. (2016). *Discrimination between Iron Deficiency Anaemia (IDA) and β-Thalassemia Trait (β-TT) Based on Pattern- Based Input Selection Artificial Neural Network (PBIS- ANN)*. *7*(4).

Karagül Yıldız, T., Yurtay, N., & Öneç, B. (2021). Classifying anaemia types using artificial learning methods. *Engineering Science and Technology, an International Journal*, *24*(1), 50–70. https://doi.org/10.1016/j.jestch.2020.12.003

Kassebaum, N. J. (2016). The Global Burden of Anaemia. *Hematology/Oncology Clinics of North America*, *30*(2), 247–308. https://doi.org/10.1016/j.hoc.2015.11.002

Kilicarslan, S., Celik, M., & Sahin, Ş. (2021). Hybrid models based on genetic algorithm and deep learning algorithms for nutritional anaemia disease classification. *Biomedical Signal Processing and Control*, *63*, 102231. https://doi.org/10.1016/j.bspc.2020.102231

Kovačević, A., Lakota, A., Kuka, L., Bečić, E., Smajović, A., & Pokvić, L. G. (2022). Application of Artificial Intelligence in Diagnosis and Classification of Anaemia. *2022 11th Mediterranean Conference on Embedded Computing (MECO)*, 1–4. https://doi.org/10.1109/MECO55406.2022.9797180

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Mahalanobis, P. C. (1930). *On test and measures of group divergence: Theoretical formulae*. http://localhost:8080/xmlui/handle/10263/1639

McLachlan, G. J. (1976). A criterion for selecting variables for the linear discriminant function. *Biometrics*, 529–534.

Mises, R. v. (1945). On the classification of observation data into distinct groups. *The Annals of Mathematical Statistics*, *16*(1), 68–73.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*(2), 159–203.

Rao, C. R. (1952). *Advanced statistical methods in biometric research.*

Rao, C. R. (1954). A general theory of discrimination when the information about alternative population distributions is based on samples. *The Annals of Mathematical Statistics*, *25*(4), 651–670.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408. https://doi.org/10.1037/h0042519

Safiri, S., Kolahi, A.-A., Noori, M., Nejadghaderi, S. A., Karamzad, N., Bragazzi, N. L., Sullman, M. J. M., Abdollahi, M., Collins, G. S., Kaufman, J. S., & Grieger, J. A. (2021). Burden of anaemia and its underlying causes in 204 countries and territories, 1990–2019: Results from the Global Burden of Disease Study 2019. *Journal of Hematology & Oncology*, *14*(1), 185. https://doi.org/10.1186/s13045-021-01202-2

Sanap, S. A., Nagori, M., & Kshirsagar, V. (2011). Classification of anaemia using data mining techniques. *Swarm, Evolutionary, and Memetic Computing: Second International Conference, SEMCCO 2011, Visakhapatnam, Andhra Pradesh, India, December 19-21, 2011, Proceedings, Part II 2*, 113–121.

Setsirichok, D., Piroonratana, T., Wongseree, W., Usavanarong, T., Paulkhaolarn, N., Kanjanakorn, C., Sirikong, M., Limwongse, C., & Chaiyaratana, N. (2012). Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening. *Biomedical*

Signal Processing and Control, *7*(2), 202–212. https://doi.org/10.1016/j.bspc.2011.03.007

Shaik, M. F., & Subashini, M. M. (2017). Anaemia diagnosis by fuzzy logic using LabVIEW. *2017 International Conference on Intelligent Computing and Control (I2C2)*, 1–5.

Thomas, C. G. (2021). *Research Methodology and Scientific Writing*. Springer International Publishing. https://doi.org/10.1007/978-3-030-64865-7

Tilley, S. R., & Rosenblatt, H. J. (2017). *Systems analysis and design* (Eleventh edition). Cengage Learning.

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, *10*(4), 299–326.

Wald, A. (1949). Statistical decision functions. *The Annals of Mathematical Statistics*, 165–205.

Walsh, T. S., Wyncoll, D. L., & Stanworth, S. J. (2010). Managing anaemia in critically ill adults. *BMJ*, *341*, c4408. https://doi.org/10.1136/bmj.c4408

Wayne, P. A. (2008). CLSI Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory-Approved Guideline. *CLSI Document EP28-A3C. Third Edition. Available Online: Http://Shop. Clsi. Org/Site/Sample_pdf/EP28A3C_sample. Pdf (Accessed on 19 October 2010)*.

Welch, B. L. (1939). Note on discriminant functions. *Biometrika*, *31*(1/2), 218–220.

Wiciński, M., Liczner, G., Cadelski, K., Kołnierzak, T., Nowaczewska, M., & Malinowski, B. (2020). Anaemia of Chronic Diseases: Wider Diagnostics—Better Treatment? *Nutrients*, *12*(6), 6. https://doi.org/10.3390/nu12061784

World Health Organization. (2011). *Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity*. World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/85839/WHO_NMH_NHD_MNM_11.1_eng.pdf

World Health Organization. (2017). *Nutritional anaemias: Tools for effective prevention and control*. World Health Organization. https://apps.who.int/iris/handle/10665/259425

World Health Organization. (2020). *Global anaemia reduction efforts among women of reproductive age: Impact, achievement of targets and the way forward for optimizing efforts*. World Health Organization. https://apps.who.int/iris/handle/10665/336559

Appendices

Appendix A: Ethical Clearance Release Letter

Strathmore UNIVERSITY

16th May 2023

Cindy Kerubo Onwong'a

120762

cindy.onwonga@strathmore.edu

Dear Cindy,

RE: Classification of Anemia Types Using Supervised Machine Learning Techniques

This is to inform you that the Office of Graduate Studies on 15th May 2023 received your acknowledgement of breach in ethical processes given that you have already collected data and written the Thesis prior to obtaining Ethical clearance. The ethics approval process is ONLY done before any collection of primary or secondary data.

This is a letter for you to proceed with the next steps of your academic requirements.

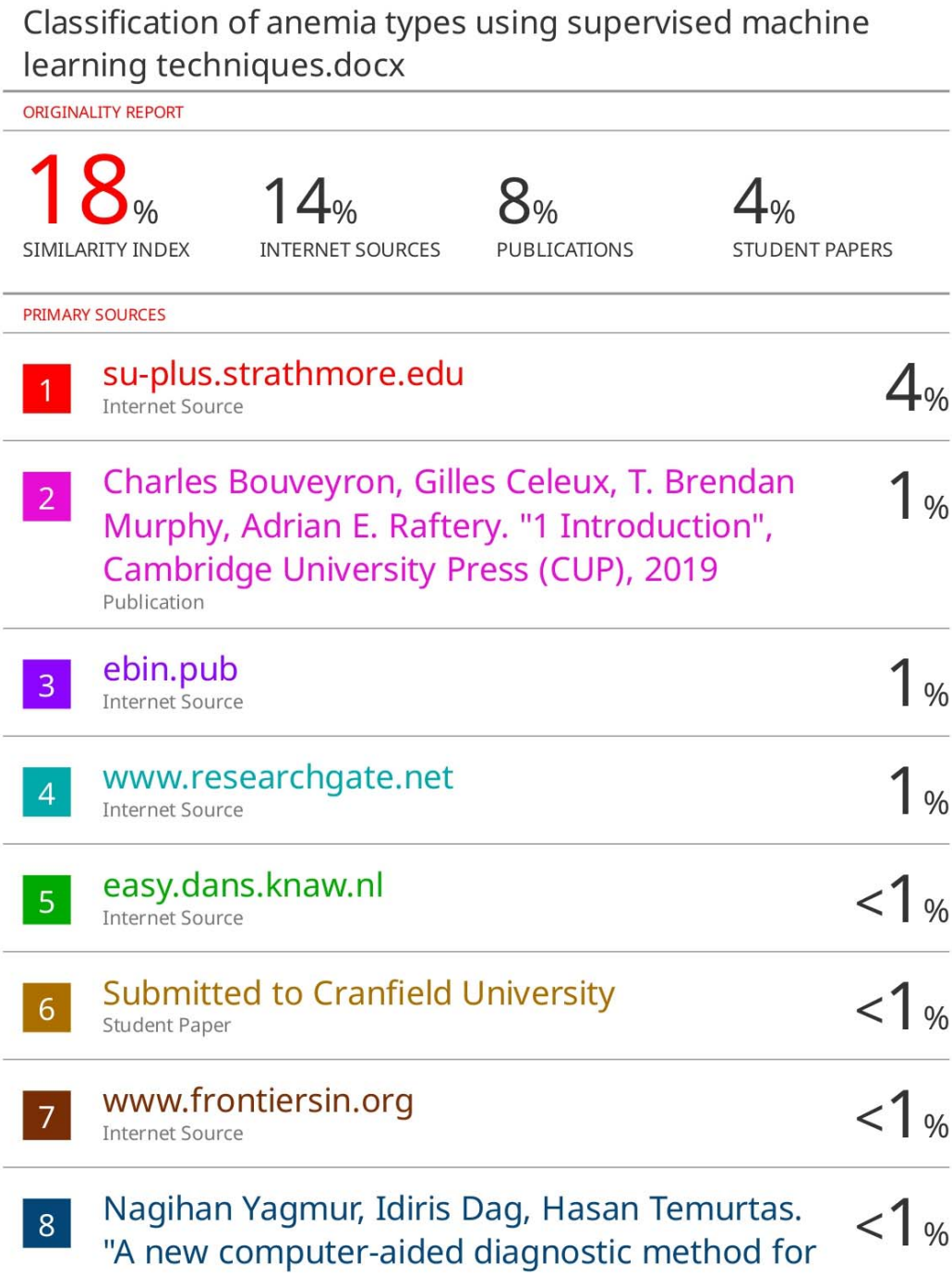
Please be advised, that in future, all research proposals should be submitted to the SU-ISERC through the RHInnO Ethics platform: https://strathmoreuniversity.rhinno.net/login

Disclaimer: This is not in any way an ethical approval letter.

Yours sincerely,

Dr. Bernard Shibwabo
Director of Graduate Studies

Appendix B: Similarity Report



Classification of anemia types using supervised machine learning techniques.docx

ORIGINALITY REPORT

18%	14%	8%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	su-plus.strathmore.edu Internet Source	4%
2	Charles Bouveyron, Gilles Celeux, T. Brendan Murphy, Adrian E. Raftery. "1 Introduction", Cambridge University Press (CUP), 2019 Publication	1%
3	ebin.pub Internet Source	1%
4	www.researchgate.net Internet Source	1%
5	easy.dans.knaw.nl Internet Source	<1%
6	Submitted to Cranfield University Student Paper	<1%
7	www.frontiersin.org Internet Source	<1%
8	Nagihan Yagmur, Idiris Dag, Hasan Temurtas. "A new computer-aided diagnostic method for	<1%