Strathmore UNIVERSITY

SU+ @ Strathmore University Library

Electronic Theses and Dissertations

2025

Classification of anemia types using supervised machine learning techniques.

Onwong'a, Cindy Kerubo
School of Computing and Engineering Sciences
Strathmore University

Recommended Citation

Onwong'a, C. K. (2025). *Classification of anaemia types using supervised machine learning techniques* [Strathmore University]. https://hdl.handle.net/11071/16451

Follow this and additional works at: https://hdl.handle.net/11071/16451

This work is availed for free and open access by Strathmore University Library.
It has been accepted for digital distribution by an authorized administrator of SU+ @Strathmore University.
For more information, please contact library@strathmore.edu

Classification of Anemia 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 Scoences for examination with my approval as the university supervisor.

Signature: _____ Date:___17th April 2024_____

Name: Dr. Vincent Omwenga

Abstract

Anemia occurs in all ages, is pathophysiologically diverse and multifactorial, and is the most predominant blood disease among human beings. Given its complex aetiology, classification of anemia types using traditional methods have limitations and drawbacks. Traditional methods of classification rely heavily on laboratory testing and analysis of blood samples which need specialized equipment and trained personnel and present potential for errors in sample collection and processing. These traditional methods are also expensive and time consuming. Machine learning based algorithms offer more accuracy and efficiency in the classification of anemia. This study aimed at building a classification model for classifying nutritional types of anemia using supervised machine learning techniques. UML diagrams were designed to visualize the model. The dataset that was utilized in this study was retrieved from Kaggle, an open-source dataset repository. It contained complete blood count test results for patients with proven cases of nutritional anemia. The data was checked for missing and null values. The features were all used in the model development because all the variables are different, and they contribute to the classification of anemia. The data was split into training and testing sets in an 80 – 20 ratio. The testing and training datasets were used to train and test the classification models developed. Given the imbalanced nature of the dataset, the SMOTE resampling technique was implemented to generate synthetic samples for the minority classes. Various parameters were tuned to improve the performance of the various classifier models. They include but are not limited to max depth, n estimators, gamma, var smoothing, learning rate, batch size, epoch, and optimizer. 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. Other classifiers that were developed and evaluated were Random Forest and Decision Trees. 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 precision was 0.9897, the recall was 0.9885 and the F1 score was 0.9888.

Keywords: Anemia Classification, Nutritional anemia classification, Supervised learning, Neural networks, Multiclass Classification, SMOTE, XG Boost, Random Forest, Naive Bayes, Decision Trees, machine learning

Table of Contents

Declarationi

Abstractii

List of Figuresvi

List of Equationsvii

Abbreviations/Acronymsviii

Chapter 1: Introduction1

1.1. Background1

1.2. Problem Statement3

1.3. Objectives4

1.3.1. General Objective4

1.3.2. Specific Objectives4

1.4. Research Questions4

1.5. Justification5

1.6. Scope and Limitation5

Chapter 2: Literature Review6

2.1. Introduction6

2.2. Discriminant Analysis6

2.2.1. Discriminant Analysis and Machine Learning7

2.1.2. Discriminant Analysis Based on Statistical Models8

2.2. Classification Systems for Anemia9

2.2.1. Anemia Classification in Research9

2.2.2. Classification of Anemia in Medicine9

2.2.3. Anemia Classification as a Public Health Concern10

2.3. Technologies Used in the Classification of Anemia11

2.3.1. Data Mining Techniques11

2.3.2. Decision Trees12

VT OMNES VNVM SINT

2.3.3. Fuzzy Logic12
2.3.4. k-Nearest Neighbor13
2.3.5. Artificial Neural Networks14
2.3.6. Neuro-Fuzzy Network14
2.4. Related Works15
2.4.1. Artificial Learning Methods Classifier for Anemia Types15
2.4.2. Deep Learning and Genetic Algorithms for Nutritional Anemia Classification16
2.4.3. Application of Artificial Intelligence in Diagnosis and Classification of Anemia16
2.5. Conceptual Framework17
Chapter 3: Research Methodology19
3.1. Introduction19
3.2. Agile Dynamic System Development Methodology19
3.2.1. Planning and Requirement Analysis Phase19
3.2.2. Designing Phase19
3.2.3. Building/Development Phase20
3.2.4. Testing Phase20
3.3. Research Design21
3.4. Study Area21
3.5. Data Collection21
3.6. Data Analysis and Classifier Modelling22
3.7. Research Quality23
3.8. Ethical Considerations23
Chapter 4: System Analysis, Design and Architecture25
4.1. Introduction25
4.2. System Analysis25
4.2.1. Requirements Gathering25
4.2.2. Functional Requirements25
4.2.3. Non-Functional Requirements26
4.3. System Architecture26

UT OMNES VNVM SINT

4.4. System Design27

4.4.1. Context Diagram27

4.4.2. Data Flow Diagram28

4.4.3. Entity Relationship Diagram29

Chapter 5: System Implementation and Testing31

5.1. Introduction31

5.2. Development Environment31

5.3. Hardware Resources31

5.4. Software Resources32

5.5. Data Pre-processing33

5.6. Model Training and Testing33

5.7. Model Evaluation and Selection36

5.7.1. Hyper-parameter Tuning39

Chapter 6: Discussions41

6.1. Introduction41

6.2. Investigation of Classification Systems for Anaemia41

6.3. Technologies Used in Anaemia Classification42

6.4. The Classifier Model42

6.5. Validating the Classifier Model43

Chapter 7: Conclusions, Recommendations and Future Work45

7.1. Conclusion45

7.2. Limitations of the Study45

7.3. Recommendations45

7.4. Future Work46

References47

Appendices53

List of Figures

Figure 2.1: Types of anemia 11

Figure 2.2: Conceptual framework for the classifier model 18

Figure 4.1: System Architecture 26

Figure 4.2: Context Diagram 27

Figure 4.3: Data Flow Diagram 28

Figure 4.4: Entity Relationship Diagram 29

Figure 5.1: Correlation Matrix 33

Figure 5.2: Data Reshaping and Normalization 34

Figure 5.3: Decision Forest 34

Figure 5.4: Random Forest 35

Figure 5.5: Naive Bayes 35

Figure 5.6: XG Boost 35

Figure 5.7: Multilayer Perceptron 36

Figure 5.8: Decision Tree evaluation 36

Figure 5.9: Random Forest evaluation 36

Figure 5.10: Naive Bayes evaluation 36

Figure 5.11: Multilayer Perceptron evaluation 36

Figure 5.12: XG Boost evaluation 36

Figure 5.13: Hyperparameter Tuning 36

Figure 6.1: Performance metrics 41

OMNES VNVM SINT

List of Equations

Equation 3.1: Accuracy Metrics.....23

Equation 3.2: Recall Metrics.....23

Equation 3.3: Precision Metrics.....23

Equation 3.4: F1-score Metrics.....23

Equation 6.1: Accuracy.....40

Equation 6.2: Recall.....40

Equation 6.3: Precision.....40

Equation 6.4: F1 - Score.....40

VT OMNES VNVM SINT

Abbreviations/Acronyms

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

IDA - Iron deficiency anemia

RDW - Red blood cell width

Chapter 1: Introduction

1.1. Background

Anaemia 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 anemia presence. Anaemia serves as a marker for compromised health and inadequate nourishment.

Anaemia is typically caused by three main mechanisms: ineffective erythropoieses 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 anemia are inadequate iron consumption, insufficient vitamin A intake, and having the beta-thalassaemia trait respectively (Gardner & Kassebaum, 2020). Iron deficiency anemia constitutes roughly half of all instances of anemia 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

of 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 anemia 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 anemia are nutritional deficiencies like folate, and 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 anemia. Additionally, different types of anaemia are distinguishable by red blood cell shape, size and colour. For instance, in the case of microcytic anemias like iron deficiency anemia where red blood cells become smaller than usual, and the concentration of haemoglobin is each red blood cell is decreased. Iron deficiency anemia 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 anemia, which is a condition where the presence of red blood cells that are larger than normal is characteristic. When they 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 erythropoiesus or increased loss of nutrients. The inflammatory response in acute and chronic infections also leads to “anemia of chronic inflammation” or “anemia 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 anemia, called normocytic anemia, 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 anemia, anemia 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 that 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 anemia and primarily found in countries located in sub-Saharan Africa. Next in frequency are β- and α-thalasaemmia, 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).

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 anemia 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). Anemia 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 anemia (FAO, 2021; Gardner & Kassebaum, 2020). Reduction of anemia 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 anemia reduction, global progress is

off track towards the achievement of WHA 2025 target to reduce by 50% anemia in females of childbearing age (FAO, 2021).

Traditional methods of anemia classification like blood tests, physical examinations and patient history 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 anemia types by developing a classification model using supervised machine learning techniques.

1.3.2. Specific Objectives

i. To investigate standards utilized in anemia classification.

ii. To review existing approaches and methodologies for classifying anemia.

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

iv. To test and validate the model.

1.4. Research Questions

i. How is classification of anemia done?

ii. What approaches and methodologies are utilized in the classification of anemia?

iii. How can anemia types be classified?

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, anemia 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 anemia, particularly among those affected by it. Traditional classification techniques can be expensive, time-consuming, and sometimes inaccurate given the complexity of anemia etiologies. In an effort to combat the problem of anemia, the classification model proposed in this study will assist in the classification of nutritional anemia. Machine learning algorithms can automatically learn from data and identify patterns, which are crucial in classifying anemia. Traditional methods of classifying anemia rely on predefined rules and assumptions and may not be able to capture the complexity of anemia.

1.6. Scope and Limitation

The study was limited to classifying nutritional types of anemia, specifically folate anaemia, B12 anaemia, iron deficiency, and hemoglobin anemia. The dataset used in the study was free and available for download from an open-source dataset repository, meaning that no experiments were carried out hence no experimental data collected. The classifier model was based on neural networks, that enable the classification and prediction of the data into the correct classes.

VT OMNES VNVM SINT

Chapter 2: Literature Review

2.1. Introduction

The literature review section intended to examine the relevant literature in relation to anemia classification. It also looked at existing approaches, methodologies, and technologies, employed in the classification of anemia: both by researchers and in medical perspectives. It also explored various machine learning classification techniques which formed the basis of the study.

2.2. Discriminant Analysis

Discriminant analysis, also known as 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 involves determining the optimal linearly weighted integration of multiple attributes to discriminate among multiple populations (R. A. Fisher, 1936). Fisher (1936) applied this method to a dataset on irises derived from Edgar Andern's work (Anderson, 1936). 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) (R. A. Fisher, 1937; Hotelling, 1931;

Mahalanobis, 1930). Fisher (1936) recognized the usage of a comparable method in craniometry for assessing variations in mandible quantifications between sexes (R. A. Fisher, 1936).

2.2.1. Discriminant Analysis and Machine Learning

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) (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 (Rosenblatt, 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 perceptrons have transformed the field of classification, exhibiting remarkable results in specific cases (Lecun et al., 1998).

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.

2.1.2. Discriminant Analysis Based on Statistical Models

Statistical methods have kept up with various breakthroughs in supervised classification, with some of the older methods being considered benchmark methods due to their simplicity and effectiveness. By assuming that the classes have a normal distribution with a shared covariance matrix, the Fisher discriminant analysis (R. A. Fisher, 1936) for instance can be reinterpreted as a technique that is based on a statistical model. The method constructs a model which based on Fisher's discriminant function, can produce a classification rule. This approach is called linear discriminating analysis (LDA), and it allows for the calculation of an optimal threshold to differentiate classes within Fisher's discriminant subspace (Fukunaga, 1999). Welch (1939), who developed the initial version of a classification rule for the case of two classes with normal distributions, and Wald (1939, 1949), who established the theory of decision functions (Wald, 1939, 1949; Welch, 1939) are some of the preliminary research works on class-conditional distribution in discriminant analysis. Miles (1945) tackled the issue of minimizing classification error in scenarios with multiple classes (Mises, 1945). Rao (1948, 1952, 1954) further expanded on this work by considering class rule estimation from samples (Rao, 1948, 1952, 1954).

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 (1976) acknowledged the vitality of selecting variables (McLachlan, 1976). Alternative methods using constrained Gaussian models were proposed by Banfield & Raftery (1993) and Bensmail & Celeux (1996) (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 (Celeux & Mkhadri, 1992; Hastie & Tibshirani, 1996).

2.2. Classification Systems for Anemia

Classification of anemia is an essential task that helps in understanding the disease and developing effective treatment strategies. Classification systems for anemia aim to provide a common language and framework for clinicians and researchers to communicate and compare data, facilitate diagnosis and management, and enhance our understanding of the pathology and epidemiology of anemia. The most widely used classification systems for anemia are the World Health Organization (WHO) classification which is based upon the etiology and morphology of red blood cells (World Health Organization, 2011), the Centers 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 (J. D. Fisher, 1972) for anemia in heart failure, and the Severity of Anemia Score (SAS) for predicting the outcome of critically ill patients with anemia (Walsh et al., 2010).

2.2.1. Anemia Classification in Research

Anaemia can be classified in research using various techniques like laboratory tests and statistical models. Laboratory tests involve measuring the levels of blood components like haemoglobin, haematocrit, mean corpuscular volume, or red blood cell count, and comparing them to already established reference ranges. Based on the laboratory results, anaemia may be categorized into different classes like anemia of chronic disease, vitamin-deficiency anaemia and iron deficiency anaemia. Statistical models can also be used for the classification of anemia in research. These models may involve algorithms of machine learning like artificial neural networks, random forests and support vector machines which use distinctive features, like demographic and clinical data, to predict the likelihood of an individual having a specific type of anemia. These models can help to identify the key risk factors and develop targeted interventions to prevent or treat anemia.

2.2.2. Classification of Anemia in Medicine

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 anemia. 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 anemias. 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 anameia 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 anemia 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 anemia, 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.

2.2.3. Anemia 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 anemia or the proportion of the population that is affected, and the severity of anemia or the degree of reduction in haemoglobin levels, as well as identifying the specific causes of anemia 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 anemia 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).

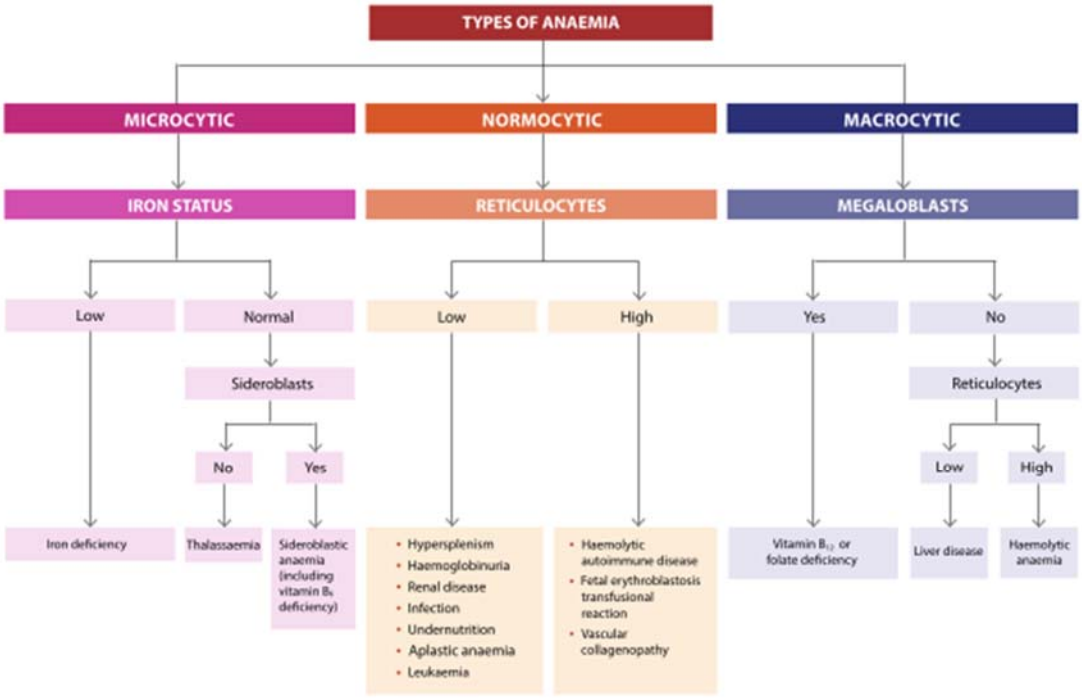


Figure 2. 1: Types of Anaemia (World Health Organization, 2017)

2.3. Technologies Used in the Classification of Anemia

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 Neighbor, 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 anemia dataset, started by collecting a dataset of anemia 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 anemia, 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 anemia 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 Neighbors 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 anemia (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 training data 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 anemia 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 anemia 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 anemia: 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 anemia 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 Neighbor (kNN) algorithm in categorization of anemia based on haemoglobin concentrations and MCV. They employed a dataset obtained from NHANES 2003-2004, that contained 5098 records of individuals with or without anemia. The kNN algorithm was applied in the classification of cases into anaemic and non-anemiac 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 anemic and non-anemic 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 anemia.

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 anemia 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 algotihm for fuzzy C-menas 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 anemia, 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 Anemia 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 anemia 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, hemolytic anaemia, iron deficiecncy 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 anemia 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 anemia 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 anemia 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 anemia with the results suggesting that techniques of machine learning are usable as a complementary tool to traditional laboratory methods for anemia 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 Anemia Classification

Kilicarslan et al. (2021) proposed a hybrid model that combined deep learning and genetic algorithms in the classification of nutritional anemias, since nutritional anemia 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 Anemia

Kovačević et al. (2022) presented a research investigation into the utilization of machine learning methods in the detection as well as the classification of anemia. Kovačević et al. (2022) presents a comprehensive examination of traditional diagnosis and classification of anemia, 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. Conceptual Framework

This study proposes the development of a model for classifying nutritional anemia types using a hybrid model leveraging both neural network as well as logistic regression techniques. Figure 2-2 represents the conceptual framework for the proposed classifier model:

VT OMNES VNVM SINT

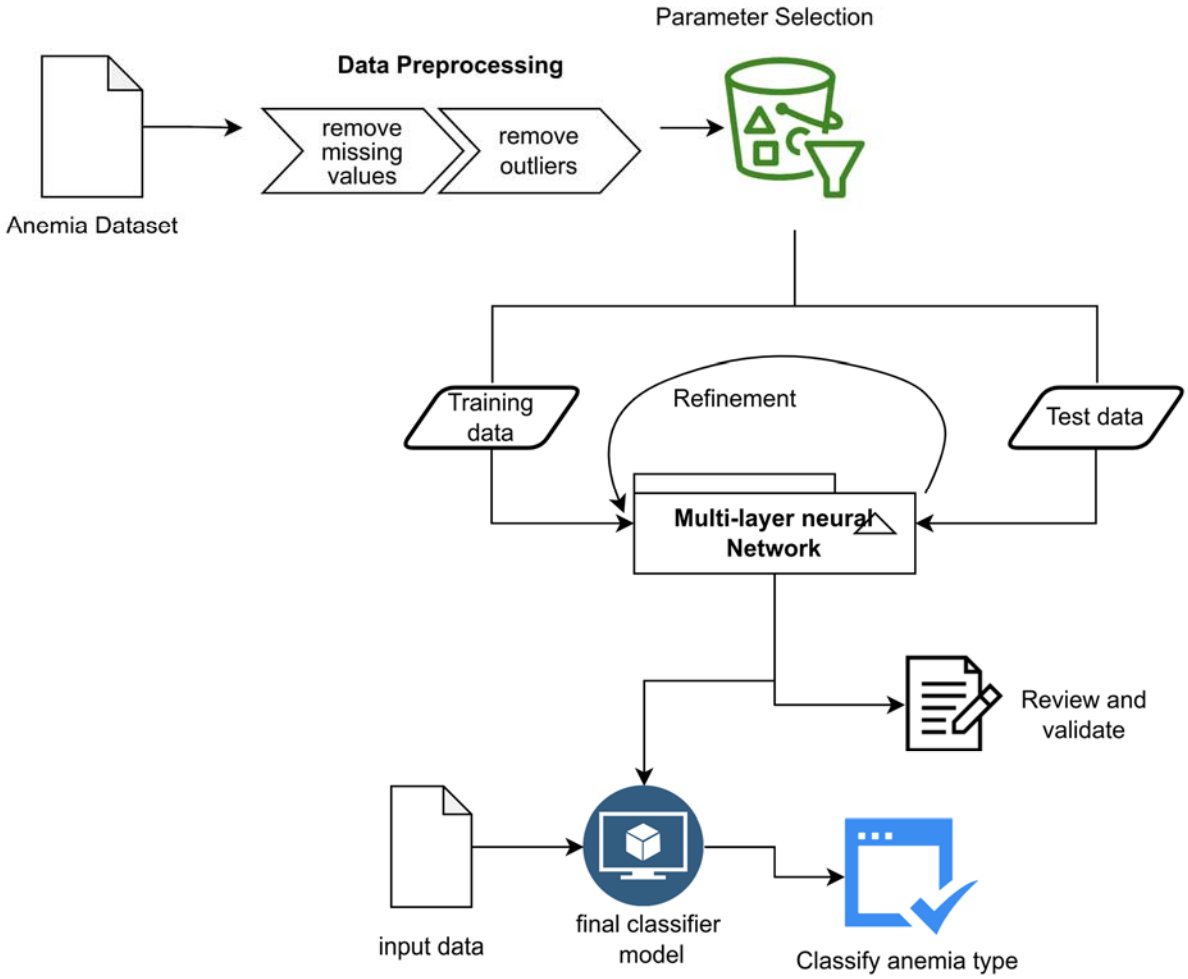


Figure 2. 2: Conceptual Framework for the Classifier

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. 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 our study, the agile DSDM approach involved these stages:

3.2.1. 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 on Anemia Disease Dataset. 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 for classifying nutritional types of anemia, specifically iron deficiency, folate deficiency, B12 deficiency. The model was built using Tensorflow and Keras among other Python libraries.

3.2.2. 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.

3.2.3. 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 the correlation was found to be weak and no parameters or variables were discarded. The data was investigated and discovered to be imbalanced. Synthethic 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 the normalized, to make it better suited for analysis since there are features that were significantly larger than others. The model was 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. 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.

3.2.4. 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 model are accuracy, precision, recall and F1 scores.

3.3. 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 model. The model was 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 corelation between them. The data was randomly assigned into training, testing and validation 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.

3.4. Study Area

The study area for this research was nutritional type of anemia. The data that was employed in this, was data retrieved from patients records of 15300 patients from Tokat Gaziosmanpaşa University, Faculty of Medicine, Turkey, collected over a 5-year interval. Of all the patients, 1019 have anemia based on the haemoglobin values, 4182 have iron deficiency, 199 have B12 deficiency, 153 have folate deficiency and 9747 have no anemia. The dataset excluded pregnant women, children, and patients with cancer. The dataset was employed to improve the concept of nutritional anemia 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.

3.5. Data Collection

The dataset deployed for this study was retrieved from kaggle, an open-source repository. The dataset was retrieved in excel document format and converted to comma separated or csv for

easier access and manipulation in the notebook. The dataset included complete blood count test results of 15,300 patients spanning 5 years; from 2013 to 2018. The features of the dataset included 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

3.6. Data Analysis and Classifier Modelling

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.

The classifier models will be 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

UT OMNES UNUM SINT

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 definition of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) was critical in analysing the model's performance better. The performance metrics that were measured included precision, recall, accuracy, and F-score. These 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

3.8. Ethical Considerations

The study did not require primary collection of data from the field. The study utilized secondary data retrieved from Kaggle open-source datasets. Since the data was open source, informed consent was not required. The data was solely employed for research purposes. All cited authors were acknowledged and given credit for their work and to avoid plagiarism.

VT OMNES
VNVM SINT

Chapter 4: System Analysis, Design and Architecture

4.1. Introduction

In this section the overall architecture and design of the classifier model 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 data flow 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. Requirements Gathering

Data used in this study was retrieved from Kaggle, an open-source dataset repository that provides datasets for improvement of research. Data required for classification of anemia is complete blood count. The dataset contained complete blood count test results from 15300 patients over a 5-year period between 2013 and 2018. Pregnant women, children and patients with cancer were excluded from the dataset. Noise in the dataset was removed. The data was resampled and normalized to better fit building the model. The data was split into training and testing sets. All 23 parameters were considered important.

4.2.2. Functional Requirements

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

i. The model accepts input data in CSV file format.
ii. The model fits to the data for optimum performance.

iii. The model classifies the data as belonging to one of the five classes.
iv. The model predicted the correct class of nutritional anemia with an accuracy of 98.88%.

4.2.3. Non-Functional Requirements

Non-functional requirements for the model include:

i. Performance – The model provides a class prediction within minutes.
ii. Maintainability – The model is easy to maintain and support.
iii. Security – The model is secure from alterations by unauthorized personnel.
iv. Reliability – The model is set with certain parameters that enable it to give consistent results when provided with the same input data.

4.3. System Architecture

The anemia dataset was pre-processed to remove noise, then used to train the model. The pre-processed data was separated into training testing and validation data. The training data was used to train the model and the test data to test and refine the model. With hyperparameter optimization and a number of iterations, the models were trained iteratively until they fit. The testing data was used as input data to test the models. The testing data was loaded into the classifiers and the class of nutritional anemia of the data was predicted.
The architecture of the system is as shown below:

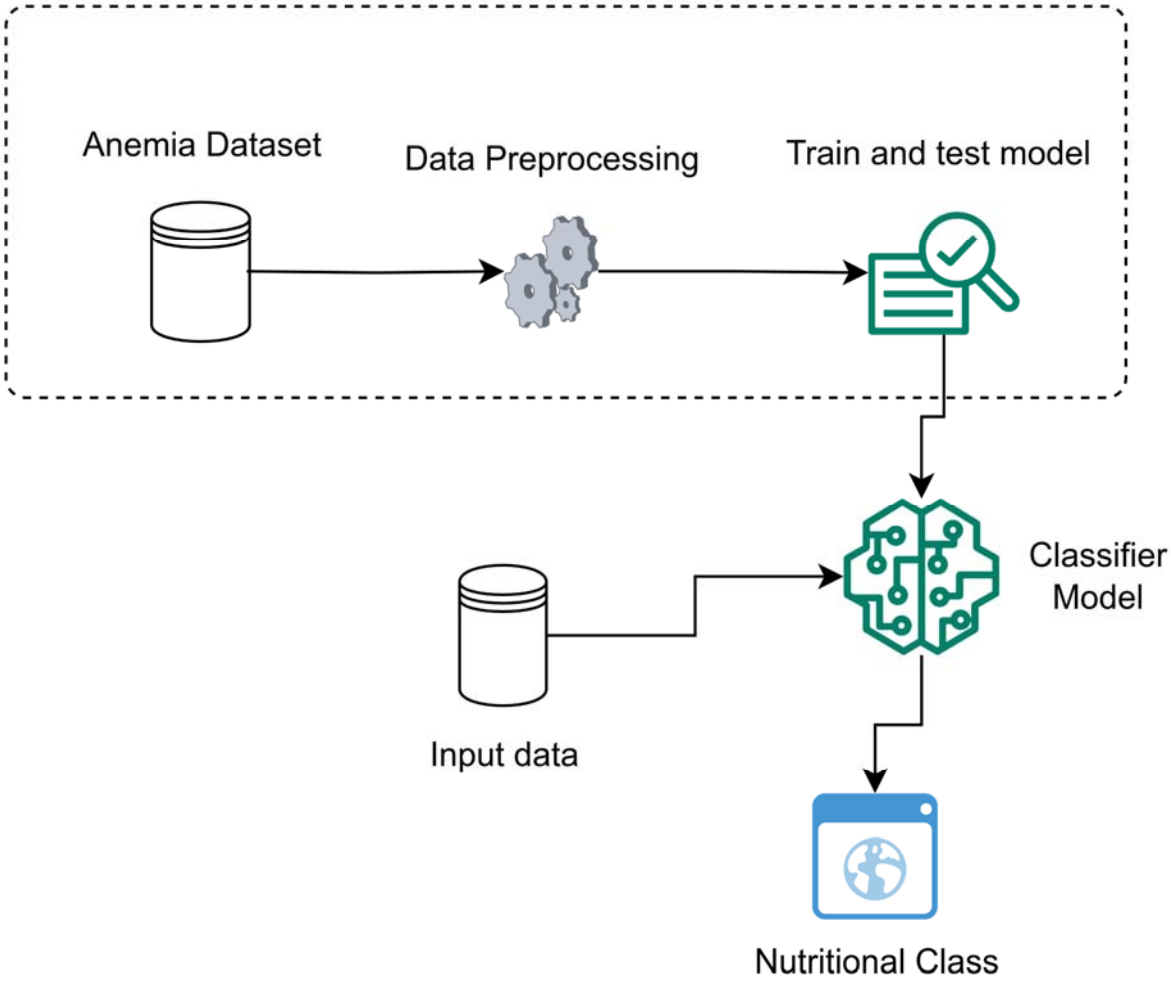


Figure 4. 1: System Architecture

4.4. System Design

System Design phase of system development was where the design of the human interface and the database was done. The structured system design methodology was utilized in this study.

4.4.1. Context Diagram

Context diagrams visually detailed 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 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 below:

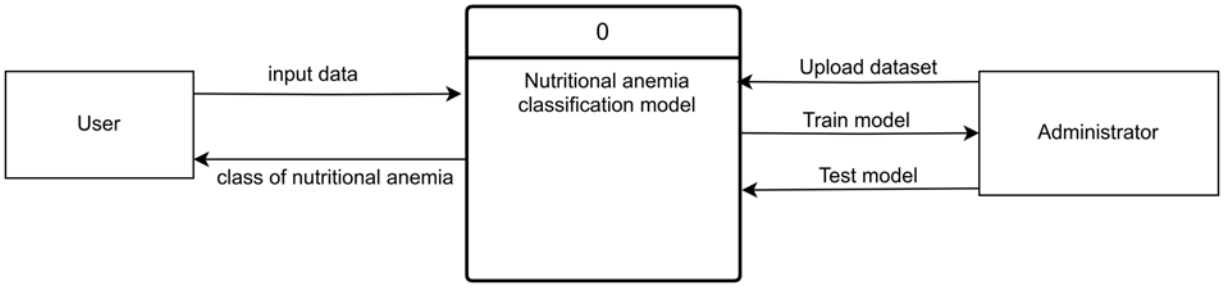


Figure 4. 2: Context Diagram

4.4.2. Data Flow Diagram

The Data Flow Diagram (DFD) is a visual representation of the flow of data within the model, outlining the processes that handle the data. DFDs typically have multiple levels, starting with level 0 that represents the entire model in a single process, followed by more detailed diagrams at level 1, level 2, etc,. With an increase in complexity and detail, the diagram level increased to give detailed representation of the model's data flow. The data flow diagram of our model is as below:

VT OMNES VNVM SINT

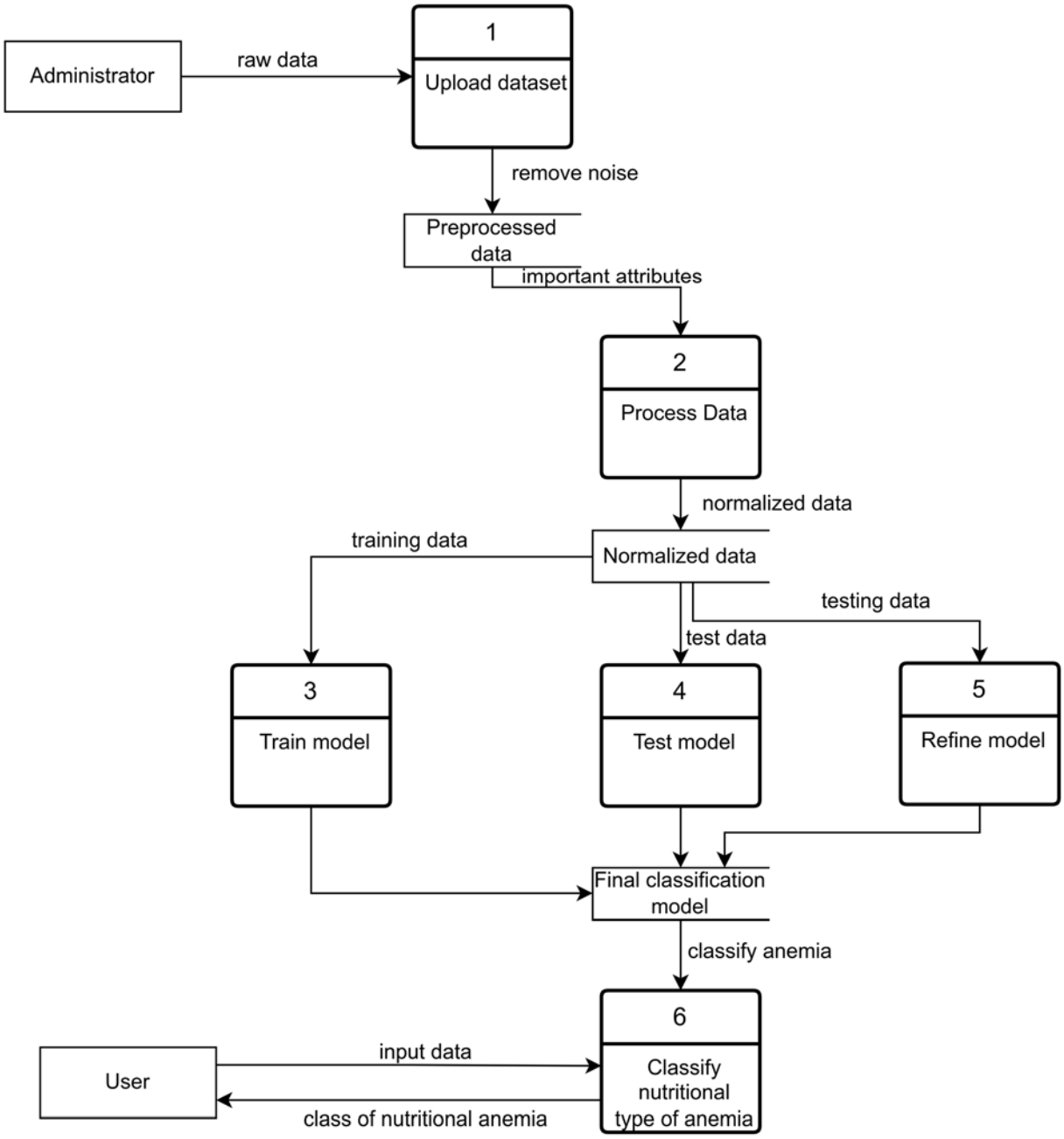


Figure 4. 3: Data Flow Diagram

4.4.3. 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 below:

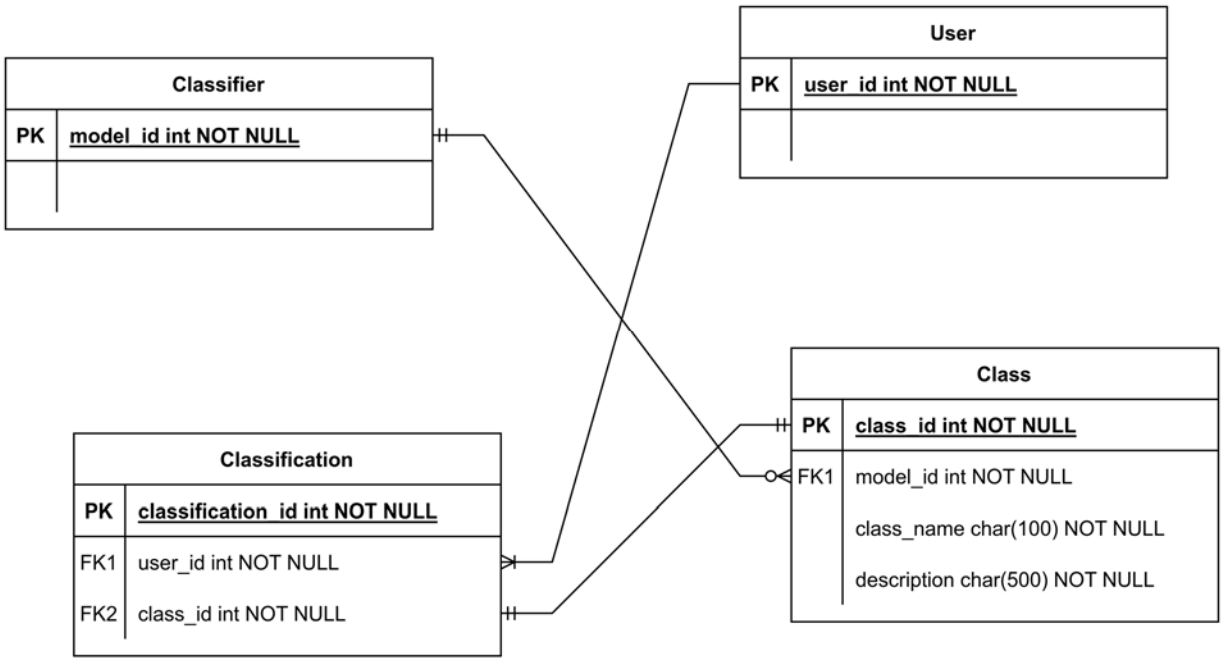


Figure 4. 4: Entity Relationship Diagram

VT OMNES VNVM SINT

Chapter 5: System Implementation and Testing

5.1. Introduction

The study was aimed at building a classification model for classifying types of nutritional anemia, specifically into no anemia, haemoglobin anemia, iron deficiency anaemia, folate anaemia, and B12 anaemia, by using supervised machine learning technologies. This chapter gives an overview of how the classification models to classify types of nutritional anemia using supervised learning techniques were built/developed, trained, and tested. Details of what environment was used to develop the model 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 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. A personal computer was also utilized in the development of the model.

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 model 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

5.5. Data Pre-processing

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 synthesised for the minority classes. The data was discovered to have some parameters with more values than others, which would affect the performance of the model. As a result, normalization was performed on the data.

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.

```
#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

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:

```
#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

5.7. 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. 8: 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. 9: 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. 10: 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. 11: 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. 12: XG Boost evaluation

5.7.1. Hyper-parameter Tuning

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. Hyperparamer tuning was performed using the GridSearchCV package. Even with hyperparameter tuning, the accuracy was 82.91% which is relatively low compared to the other classifiers.

```python
#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))
```

```
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. 13: Hyperparameter Tuning

Chapter 6: Discussions

6.1. Introduction

The aim of the study was to develop a classification model for classifying nutritional types of anemia, 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. Investigation of Classification Systems for Anaemia

While reviewing the literature, a common classification pattern for anaemia types was identified. As a population, anameia classification was done based on the percentage severity of haemoglobin concentration in each population group based on age. This means that, if the percentage of individuals with anemia as per haemoglobin concentration levels in a specific age group is between a stipulated range, then that population group, as a collective, is classified as suffering from either mild, moderate or severe anaemia. In research, anaemia was either classified based on the underlying cause or based on the severity of the haemoglobin concentration. Haemoglobin concentrations varies by age and gender, among other factors. In the field of medicine, anemia can be classified based on the size of the red blood cells or based on the underlying mechanism of anemia. Some of the classes of anemia recognized in medicine are microcytic, normocytic, or macrocytic based on the underlying cause, as well as mild, moderate, or severe, based on the degree of haemoglobin concentration, depending on the age and gender of the patient. The ICD coding system is also utilized in the classification of anemia, although the severity is not defined explicitly. Other classification systems that were identified are the World health organization classification based on red blood cell etiology and morphology, the Clinical and laboratory standards institute classification, the New York heart association classification, and the severity of anaemia score.

6.3. Technologies Used in Anaemia Classification

With the advancement of technology in the present day, various techniques have been employed in the classification of anemia, be it differential diagnosis of two specific anemias, or the general classification of anemia. 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 anemia based of different factors or parameters. The most utilized parameters because of their ease of access are haemoglobin, hematocrit, mean corpuscular mean, volume are 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. These data has been implemented in the early detection and differential diagnosis of anemia, including studies like Jamie & Talarposhti (2016) (Jamei & Talarposhti, 2016).

There are a few studies that hold high significance in the classification of anemia. 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 anemia 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 anemia like iron deficiency, B12 and folate anemias, and patients without anemia. The proposed hybrid genetic algorithm and convolutional neural networks algorithm achieved an accuracy of 98.50% (Kilicarslan et al., 2021).

6.4. The Classifier Model

The study aimed at developing a machine learning model for classifying nutritional types of anemia, specifically iron deficiency anemia, B12 anemia, folate anemia, and haemoglobin anemia as well as no anemia. The model was designed by utilizing context diagrams, dataflow diagrams and entity relationship diagrams. The model was built using a multilayer neural

network and a dataset containing blood count data with 24 parameters. All the 24 parameters were used in the development of the model. The dataset was pre-processed, trained and tested and then evaluated using metrics like accuracy, precision, recall and F1 scores.

A decision tree classifier, a random forest classifier, a naive bayes classifier, an XG Boost classifier, and a multilayer perceptron neural network were all trained and tested and their performance tested. The XG Boost classifier performed the best over the others.

6.5. Validating the Classifier Model

The metrics utilized in evaluating the performance of the model are accuracy, sensitivity or recall, precision and F1-score since accuracy alone does not give a deep analysis of the model performance on new data. The formulas for each of these metrics is as shown below:

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

Equation 6. **1**: Accuracy (Karagül Yıldız et al., 2021)

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

Equation 6. **2**: Recall (Karagül Yıldız et al., 2021)

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

Equation 6. **3**: Precision (Karagül Yıldız et al., 2021)

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

Equation 6. **4**: F1 Score (Karagül Yıldız et al., 2021)

where P is precision and R is recall. 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 yet, yet they belong to that class.

The accuracy of the model was generated using sklearn.metrics. The specific metric as shown below:

```
#performance metrics
accuracy = accuracy_score(test_labels1, predict_results)
print('Accuracy: %f' % accuracy)

precision = precision_score(test_labels1, predict_results, average='weighted')
print('Precision: %f' % precision)

recall = recall_score(test_labels1, predict_results, average='weighted')
print('Recall: %f' % recall)

f1 = f1_score(test_labels1, predict_results, average='weighted')
print('F1 score: %f' % f1)
```

Figure 6. **1**: Performance metrics

The accuracy of the model was 0.9503, the precision was 0.9491, the recall was 0.9503 and the F1 score was 0.9457.

VT OMNES VNVM SINT

Chapter 7: Conclusions, Recommendations and Future Work

7.1. Conclusion

This study was aimed at developing a classification model for nutritional anemia, specifically iron deficiency anemia, B12 anemia and folate anemia, as well as haemoglobin anemia and no anemia, using machine learning algorithms. This was achieved by building and comparing the performance metrics of the various classifiers. The models utilized complete blood count data with 24 parameters. The dataset used in the study was pre-processed, and then split into training and testing sets. The training set was in the building and tuning of the model, while the testing set was fed into the models to evaluate the performance of the model. Some model parameters were tuned to improve the performance of these models. The XG Boost classifier 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.

7.2. Limitations of the Study

The limitations identified from this study are:

i. The dataset was not balanced as some classes had very few datapoints, as compared with others

ii. The study was only limited to classification of nutritional type of anemia

iii. The study did not include the most vulnerable populations, including women and children

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 be made more accessible with the help of policy makers and healthcare institutions.

iii. Using a more rodust dataset which includes examples belonging to more than one class.

7.4. Future Work

Current findings could be extended by:

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

ii. Acquiring more data, especially for some classes with datapoints.

iii. Including several types of anemia.

iv. 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 anemia. *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). Anemia 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 Anemia 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 anemia 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 Anemia. *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 Anemia 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 Anemia. *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 anemia 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 anemia 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). Anemia 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). Anemia 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

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

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu