TÉCNICO LISBOA

Classification of microcytic anaemias using machine learning methods

Beatriz Neves Leitão

Thesis to obtain the Master of Science Degree in

Biotechnology

Supervisor(s): Prof. Susana de Almeida Mendes Vinga Martins
Dr. Maria Paula Duarte Faustino Gonçalves

Chairperson: Prof. Miguel Nobre Parreira Cacho Teixeira
Supervisor: Prof. Susana de Almeida Mendes Vinga Martins
Member of the Committee: Dr. Sara Guilherme Oliveira da Silva

October 2021

All beings that in my life touched my heart, I dedicate this thesis

Preface

The work presented in this thesis was performed at INESC-ID, “Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento em Lisboa”, of Instituto superior Técnico (Lisbon, Portugal) and at ”Grupo de I&D em Hemoglobinopatias, Metabolismo do Ferro e Patologias Associadas” of Human Genetics Department of National Institute of Health Doctor Ricardo Jorge (INSA) (Lisbon, Portugal), during the period February-October 2021, under the supervision of Prof. Susana Vinga at IST and co-supervision of Dra. Paula Faustino at INSA.

Acknowledgments

First of all, I want to thank my internal supervisor Professor Susana Vinga, for giving me the opportunity to work on a subject that gets outside of my academic background. Thank you very much for the trust and all the support throughout this thesis.

I would like to thank Doctor Paula Faustino from National Institute of Health Doctor Ricardo Jorge (INSA), for making this project possible, and all the guidance. To thank Pedro Lopes for the collaboration and technical support when carrying the laboratory techniques at INSA. I also want to thank Bárbara Faleiro and Daniela Santos for the molecular characterization of β and α-thalassemia carriers, to Marta Barreto and Irina Kislaya for providing samples and data obtained within the scope of the INSEF 2015 project, and to João Batista for the help with the M3GP algorithm.

In addition, I want to thank my parents and sister who are always there for me no matter what, could not be more grateful for all the love. And to my grandparents who followed me closely along my university journey, thank you for all the lunches and company.

I also want to express my thanks to my boyfriend Pedro, the best friend I could ever have, for encouraging me throughout this thesis and for always making my days better.

Lastly, I want to appreciate my friends and colleagues, for all the good times, as they have allowed me to rest my mind outside of my research.

Resumo

A prevalência de anemia na população mundial é de 24.8%. A discriminação adequada entre anemias microcíticas é fundamental para fornecer o tratamento adequado e providenciar aconselhamento genético.

Uma vez que os métodos mais fidedignos para diagnosticar talassemias e anemia ferropénica (AF), algumas das anemias microcíticas mais comuns, são caros e demorados, vários índices foram desenvolvidos ao longo dos anos. Contudo, esses índices revelaram não ser 100% fiáveis.

Nesta tese foram utilizados dados hematológicos de uma amostra da população portuguesa constituída por 390 indivíduos e respetivo diagnóstico para treinar e testar diferentes algoritmos de aprendizagem automática. O propósito foi desenvolver um classificador binário, especificamente adaptado à população portuguesa, a fim de discriminar entre portadores de β-talassemia e doentes com AF. Para além disso, foi desenvolvido um classificador multi-classe capaz de distinguir entre portadores de β-talassemia, portadores de α-talassemia, doentes com AF e indivíduos saudáveis. De forma a não comprometer o objetivo principal, a obtenção dum diagnóstico rápido e acessível, os classificadores desenvolvidos foram baseados apenas em informações obtidas através de um hemograma, um dos exames laboratoriais mais comuns em medicina.

Embora não tenha sido possível ultrapassar o desempenho com os classificadores binários criados do índice mais fiável para a população portuguesa, RDWI (índice de distribuição de largura dos glóbulos vermelhos), que apresentou uma exatidão mediana de 95.4%, foi possível igualar esta exatidão com o algoritmo florestas aleatórias. Este algoritmo apresentou um ótimo desempenho tanto na classificação binária, como na classificação multi-classe, onde obteve resultados promissores revelando uma exatidão mediana de 93.0%.

Palavras-chave: anemia microcítica, talassemia, anemia ferropénica, classificação, aprendizagem automática.

Abstract

The prevalence of anaemia in the world population is 24.8%. Proper discrimination between microcytic anaemias is essential to provide the right treatment and genetic counselling.

As the most reliable methods to diagnose thalassemias and IDA (iron deficiency anaemia), some of the most common microcytic anaemias are expensive and time-consuming, many indexes have been developed through the years. These indexes, however, have not been revealed to be 100% accurate.

In this thesis, haematological data from a sample of the Portuguese population constituted by 390 individuals and their diagnosis was used to train and test different machine learning algorithms. The objective was to develop a binary classifier, specifically adapted to the Portuguese population, to discriminate β-thalassemia carriers from IDA patients. Beyond that, a multi-class classifier capable of distinguishing between β-thalassemia carriers, α-thalassemia carriers, IDA patients, and healthy subjects was also developed. In order not to compromise the main objective, to obtain a quick and accessible diagnosis, the classifiers developed were only based on information obtained through a complete blood count test, one of the most common laboratory tests in medicine.

Although it was not possible to surpass the performance with the binary classifiers created of the most reliable index for the Portuguese population, RDWI (red cell distribution width index), which presented a median accuracy of 95.4%, it was possible to match it with the random forest algorithm. This algorithm showed an excellent performance in the binary and in the multi-class classification, where it achieved promising results, revelling a median accuracy of 93.0%.

Keywords: microcytic aneamia, thalassemia, iron deficiency anaemia, classification, machine learning.

Contents

Preface v
Acknowledgments vii
Resumo ix
Abstract xi
List of Tables xv
List of Figures xvii
Glossary xix

1 Introduction **1**
1.1 Motivation 1
1.2 Objectives and Contributions 2
1.3 Thesis Outline 3

2 Theoretical Background **5**
2.1 Microcytic Anaemia 5
2.1.1 Diagnosis 8
2.2 Machine Learning Algorithms 10
2.2.1 Hyperparameter Optimization 16
2.3 Evolutionary Algorithms 17
2.4 Evaluation Metrics 19
2.5 Outlier Detection 21
2.6 State of the Art of Anaemia Classifiers 23
2.6.1 Anaemia Indexes 23
2.6.2 Machine Learning in Anaemia Classification 24

3 Implementation **27**
3.1 Dataset Description 27
3.2 Molecular Diagnosis 28
3.2.1 β-thalassemia 28
3.2.2 α-thalassemia 29
3.3 Indexes Evaluation 30
3.4 Machine Learning Classification 30

3.4.1 Features Selection . . . 30
3.4.2 Hyperparameter Optimization . . . 31
3.5 Genetic Programming Classification . . . 32
3.6 Outliers Detection . . . 32

4 Results **33**
4.1 Molecular Diagnosis . . . 33
4.2 Data Description . . . 35
4.3 Indexes Evaluation . . . 38
4.4 Machine Learning Classification . . . 39
4.4.1 Models Improvement and Evaluation . . . 39
4.5 Genetic Programming Classification . . . 44
4.6 Oultiers Detection . . . 45

5 Conclusions **51**
5.1 Achievements . . . 51
5.2 Future Work . . . 53

Bibliography **55**

List of Tables

2.1 Confusion matrix 19
2.2 Discriminant indexes, formula and cut-off values 24
2.3 Anaemia classifiers constructed with machine learning algorithms 25

3.1 Data description 27
3.2 Machine learning hyperparameters tested and their range of values 31

4.1 β-thalassemia diagnosis result 34
4.2 Dataset features' properties 36
4.3 β-thalassemia p-values of the T-test 37
4.4 Indexes performance with 30 random splits of the data 38
4.5 Indexes performance with all the data 38
4.6 Median accuracy of the machine learning binary classifiers 40
4.7 Median accuracy of the machine learning multi-class classifiers 42
4.8 Median accuracy per class of the best machine learning multi-class classifiers 43
4.9 Median accuracy of the M3GP classifiers 44
4.10 Most frequently misclassified instances in binary classification 47
4.11 Most frequently misclassified instances in multi-class classification 47
4.12 Mean feature values per class and some instances' values 48

List of Figures

2.1 Prevalance of anaemia in the world (preschool-age children) . . . 5
2.2 Microcytic disorders . . . 6
2.3 Microcytic anaemia evaluation . . . 9
2.4 Sigmoid function . . . 11
2.5 Artificial neural network example . . . 13
2.6 Support vector machine example . . . 14
2.7 Data set example and corresponding decision tree . . . 15
2.8 K-nearest neighbors example . . . 16
2.9 Genetic algorithm scheme . . . 18
2.10 Genetic programming scheme . . . 19
2.11 ROC curve . . . 20
2.12 Monte Carlo cross-validation scheme . . . 21
2.13 Illustration of the elements involved in the computation of $s(i)$ . . . 23

3.1 Illustration of the human β-globin multigene loci residing on chromosome 11 (11p15.4). . 28
3.2 Gap-PCR analysis for diagnosis of $-\alpha^{3.7}$ kb deletion (α-thalassemia) . . . 29

4.1 Example of two Sanger Sequencing results obtained . . . 33
4.2 Photo of an agarose gel electrophoresis of DNA fragments obtained by gap-PCR . . . 34
4.3 Boxplots with all the normalized features per class . . . 35
4.4 β-thalassemia normalized features . . . 37
4.5 Accuracy of the binary logistic regression model in the genetic algorithm, over the number of generations and the population size . . . 40
4.6 Machine learning binary models accuracy with 30 data random splits and with all data . . 41
4.7 Machine learning multi-class models accuracy with 30 data random splits and with all data 43
4.8 Graphical representation of the data separation in feature space of the M3GP classifiers . 44
4.9 Silhouette analysis and respective class mean . . . 45
4.10 Cook's distance outlier detection . . . 46
4.11 Venn diagrams of the models most misclassified instances and the outliers detected . . . 48

Glossary

ACD	Anaemia of Chronic Disease
ANN	Artificial Neural Network
AUC	Area Under the ROC Curve
CBC	Complete Blood Count
DT	Decision Tree
EA	Evolutionary Algorithm
EC	Evolutionary Computation
ELISA	Enzyme-linked Immunosorbent Assays
EP	Evolutionary Programming
ES	Evolution Strategies
FS	F-Score
Fe	Iron
Ft	Ferritin
GA	Genetic Algorithm
GP	Genetic Programming
G&K	Green and King
Hb	Hemoglobin
Hct	Hematocrit
IDA	Iron Deficiency Anaemia
IUPAC	International Union of Pure and Applied Chemistry
KNN	k-nearest neighbors
LR	Logistic Regression
M3GP	Multidimensional Multi-class Genetic Programming with Multidimensional Populations
MCHC	Mean Corpuscular Hemoglobin Concentration
MCH	Mean Corpuscular Hemoglobin
MCV	Mean Corpuscular Volume
NB	Naive Bayes
PCR	Polymerase Chain Reaction

PLT	Platelet
RBC	Red blood cell
RDW	Red Blood Cell Distribution Width
RF	Random Forest
ROC	Receiver Operating Characteristics
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machines
TIBC	Total Iron Binding Capacity
TS	Transferrin Saturation
Tf	Transferrin
WBC	White Blood Cell
acc	Accuracy
dNTP	2'-deoxynucleotide triphosphate
ddNTP	2', 3'-dideoxynucleotide triphosphate
err	Error Rate
fn	False Negative
fp rate	False Positive Rate
fp	False Positive
gap-PCR	gap-Polymerase Chain Reaction
p	Precision
r	Recall
sn	Sensitivity
sp	Specificity
tn	True Negative
tp rate	True Positive Rate
tp	True Positive

Chapter 1

Introduction

1.1 Motivation

According to the World Health Organization (WHO [1]) anaemias are highly prevalent diseases worldwide. Aneamias can be divided in three major classes: macrocytic, normocytic, and microcytic. The microcytic anaemias are a group of anaemias characterized by a low mean corpuscular volume. Some of the most common are the thalassemias and the IDA (iron deficiency anaemia) [2]. Hence, it becomes relevant to develop fast, trustful, and cost-effective methods that can access the correct diagnosis of this diseases, in order to provide the right treatment.

To reduce the cost of the diagnosis, several indexes have been developed to distinguish between these different types of anaemias, as it would be especially advantageous in countries with less financial resources and more limited healthcare systems. These indexes however, are not totally accurate and have presented different performances when tested in populations from different countries [3–5], suggesting that they are not making a good generalization across all populations. Therefore, it is important to continue the study of this field of research in order to develop even better indexes able to make an accurate diagnosis for specific target populations.

To develop new classifiers this thesis makes use of machine learning. Machine learning is a subset of artificial intelligence, its algorithms are very useful tools for pattern recognition. Resorting to statistics and optimization, these algorithms are able to learn from the data provided and construct models which are capable of making predictions [6].

In respect to anaemia, multiple studies have used machine learning algorithms to create classifiers able to distinguish between anaemic and non-anaemic subjects [7–10] and even between IDA patients and thalassemia carriers [11]. These classifiers were developed using hematological data from individuals whom the diagnosis was known. This collection of data was used to train the algorithms, so that afterwards it was possible to make predictions about new subjects, which has led to accurate results and, therefore, to good generalization capacity of the estimated models. Nevertheless, even though few studies tried to distinguish between β-thalassemia carriers and IDA patients, like the known indexes do, none used data from the Portuguese population.

1.2 Objectives and Contributions

This project intended to develop a new binary microcytic anaemia classifier to discriminate between β-thalassemia carriers and IDA patients with a better performance on the Portuguese population than the existing indexes and go a step further, with the creation of a multi-class classifier capable of discriminating between β-thalassemia carriers, α-thalassemia carriers, IDA patients, and healthy subjects (control group). In order not to compromise the main objective of creating these classifiers, which is to obtain a quick and accessible diagnosis, the classifiers developed in this thesis are only based on hematological data from the Portuguese population obtained through a CBC (complete blood count) test, one of the most common laboratory tests in medicine.

Under this context, it was necessary to evaluate the performance of the indexes that already exist with the hematological data and the subject's diagnosis confirmed by molecular analysis that were made available. In addition, in order to exemplify and gain expertise, the molecular diagnosis of β-thalassemia and α-thalassemia was also performed, from the DNA sample to the final diagnosis.

After that, several machine learning algorithms were trained and tested with the hematological data and the molecular diagnosis of multiple individuals. To improve the accuracy of the models generated by the algorithms, a genetic algorithm was employed to select new features created from existing ones, and a hyperparameter optimization was done so that the learning process could be as successful as possible.

Besides the machine learning algorithms, with the purpose of obtaining a classifier with the best performance possible, other artificial intelligence technique, genetic programming, was employed to generate binary and multi-class classifiers.

That being so, with this thesis it was possible evaluate several indexes and conclude that the index that has a better performance on the Portuguese population is the RDWI, this index achieved a median accuracy of 95.4%. Furthermore, multiple machine learning algorithms were tested for the first time with Portuguese hematological data and even though was not possible to surpass the performance of the RDWI index with the created binary classifiers, it was possible to match it with the random forest algorithm. Among all the algorithms the random forest presented the best performance not only in the binary classification but also in the multi-class classification, where it achieved promising results revelling a median accuracy of 93.0%.

In addition, it was possible to develop a semi-automatic model able to identify instances that present features different from what would be expected according to the attributed disease and, therefore, may require a second analysis.

This thesis involved both a laboratory and a computational part. The laboratory methods were performed at "Grupo de I&D em Hemoglobinopatias, Metabolismo do Ferro e Patologias Associadas" of Human Genetics Department of National Institute of Health Doctor Ricardo Jorge (INSA), and the code (Python) written in this thesis is all open source and is freely available on my GitHub repository [12].

1.3 Thesis Outline

Including this introductory chapter, Chapter 1, this thesis is divided into 5 chapters.

In Chapter 2, the theoretical background of this thesis is explained in detail. It starts by characterizing the different microcytic anaemia diseases, followed by their genetic basis and clinical manifestations, plus an explanation of the current process of diagnosis.

Subsequently, the machine learning artificial intelligence technique is explained, as well as some of its best-known algorithms, and how to optimize the learning process. After that, another artificial intelligence technique, evolutionary algorithms, is explained along with some of its types, genetic algorithm and genetic programming. Following the artificial intelligence techniques are their evaluation metrics, used to evaluate the performance of the models generated.

Later on, is the state of the art of the anaemia classifiers introducing some of the best-known indexes that discriminate between β-thalassemia carriers and IDA patients, and examples of the application of machine learning algorithms in the classification of anaemia.

Chapter 3, explains all the techniques that were implemented in this thesis, along with a description of the data used. Right after, comes the results chapter, Chapter 4, containing all the results obtained throughout this thesis.

Finally, the last chapter, Chapter 5, contains the conclusions that were possible to draw from the results obtained, as well as future perspectives to continue the work developed here.

Chapter 2

Theoretical Background

2.1 Microcytic Anaemia

Anaemia is a condition characterized by a decrease of red blood cell mass and low level of hemoglobin, having as principal consequence a diminished oxygen carrying capacity of the blood [13]. The main symptoms are therefore related with lack of oxygen, such as fatigue and shortness of breath [14].

According to the World Health Organization [1] anaemia was estimated to affect 24.8% of the global population, having a bigger prevalence in low income countries, as depicted in Figure 2.1.

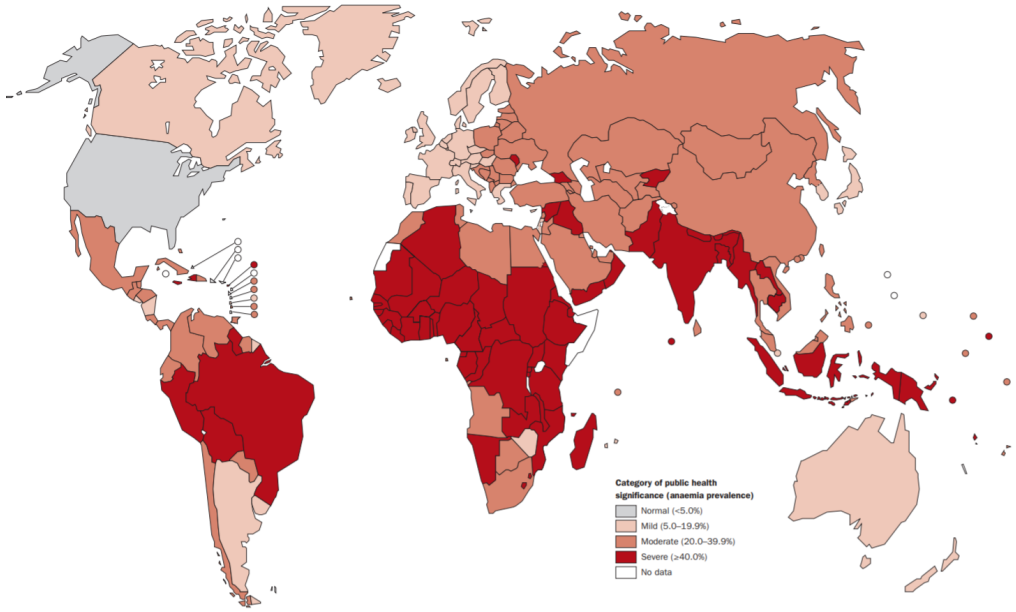


Figure 2.1: Prevalance of anaemia in the world (preschool-age children), adapted from WHO [1].

Regarding Portugal, two different studies assessed the prevalence of anaemia obtaining very different results. The first study indicated an anaemia prevalence of 19.9% [15], detecting that subjects with ages between 18 and 34 years old and older adults ($\geq$65) had the highest prevalence of anaemia. The most recent study only included subjects with ages between 25 and 74 years old and revelled a prevalence of only 5.8% [16]. Since the two studies do not cover the same age groups, the preva-

lence of anaemia in Portugal is debatable. On one hand, the most recent study does not include some of the most at-risk ages, such as young adults with ages between 18 and 24 years old, on the other hand 19.9% seems to be a very high value for a European country, leaving some uncertainty as to the prevalence of anaemia in Portugal.

As the name itself suggests, in the microcytic anaemias the red blood cells are smaller than the usual. This occurs because of the decreased production of a major constituent of the red blood cells, hemoglobin [2].

Inside the microcytic anaemias there are multiple variants, some of the most common are the thalassemias and IDA (iron deficiency anaemia) [2], as illustrated in Figure 2.2.

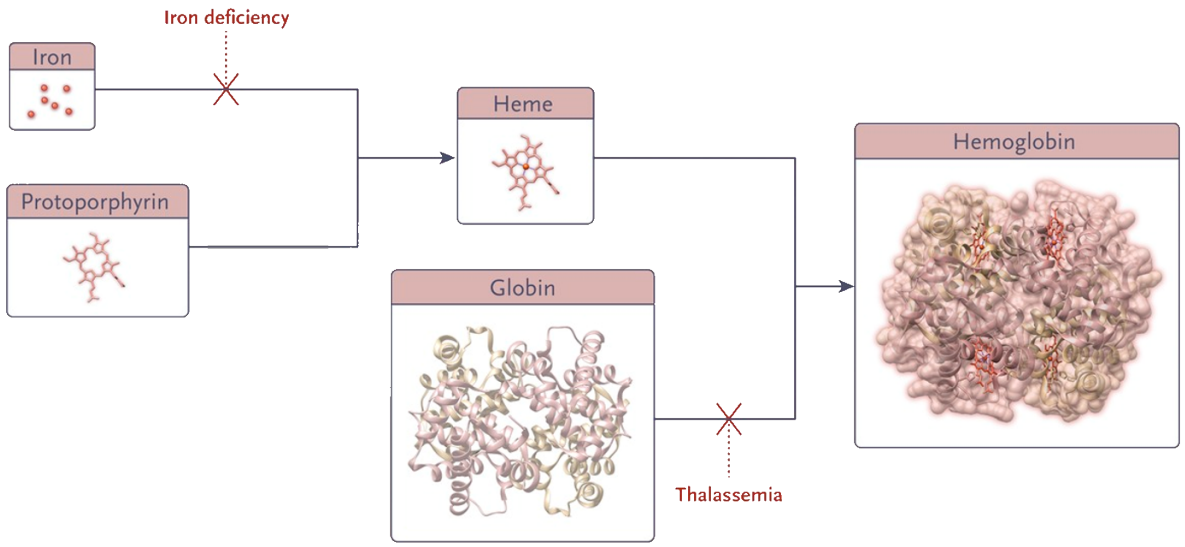


Figure 2.2: Microcytic disorders. In the case of IDA, the lack of hemoglobin arises from the shortage of iron; in thalassemias it results from the defective hemoglobin production, adapted from DeLoughery [2].

IDA is the most common cause of anaemia. In this condition, anaemia is associated with low levels of iron. Iron is needed for the hemoglobin synthesis in erythropoiesis and essential for the oxygen transport by the red blood cells [17].

Thalassemias are hemoglobinopathies autosomal recessive disorders [18]. The thalassemias, β-thalassemia and α-thalassemia are associated with a defect in the synthesis of two globin chains, β-globin and α-globin, respectively. These proteins form, in equal quantities, hemoglobin A, which represents about 97.0% of the total red blood cell hemoglobin of a human adult [19]. Most of the remaining hemoglobin present in the red blood cells is hemoglobin A2, representing about 2.0% of the total hemoglobin. Hemoglobin A2 differs from the hemoglobin A, being composed of two α-globin chains and two δ-globin chains [14].

Epidemiology

In the most developed countries, IDA arises mainly from eating habits (such as vegetarian diet or no read meat intake), and pathologic conditions (chronic blood loss or malabsorption). In the developing countries it typically results from insufficient dietary intake and/or loss of blood [17]. In that sense it is crucial to find the cause, since there are no natural mechanisms, other than menstruation, for ridding the body of iron [2].

While the β-thalassemia is more concentrated in the Mediterranean and Asia, the α-thalassemia is more frequent in Africa, Oceania, and India. Other hemoglobinopathies have their highest prevalence in Africa, Saudi Arabia, India, and south-east Asia [20]. These mentioned regions are associated with high prevalence of endemic malaria, which can be explained by the fact that hemoglobinopathies trait are thought to provide protection against this infectious disease. In malaria regions natural selection can even be responsible for maintaining higher frequencies of the genes causing those pathologies. However, the distribution of hemoglobinopathies does not always coincide with the presence of malaria, as the case of the Pacific region, where there is no malaria, albeit those exceptions can be explained by genetic drift [20]. Hence, the presence of β-thalassemia in Northern Europe, America, Caribbean, and Australia can be explained by population migration and intermarriage between different ethnic groups [21].

Despite the epidemical and clinical studies that propose protection against malaria in the presence of hemoglobinopathies trait, the mechanism underlying it is still unknown. The only consensus is that there may exists an enhanced phagocytosis of the red blood cells infected with malaria [22].

More recently a study with the Italian population has also suggested a potential association between SARSCoV-2 immunity and β-thalassaemic heterozygote population [23].

Genetic Basis and Clinical Manifestations

In the case of β-thalassemia the severity of the disease is directly related to the pathogenicity of the β-globin gene mutation and the degree of excess of the α-globin chain, which will precipitate in the red blood cell precursors, leading to ineffective erythropoiesis [24]. Given that there is one copy of the β-globin chain gene in each chromosome 11, subjects can either be heterozygous or homozygous [2]. Besides, these mutations can either partially or completely eliminate the synthesis of the β-globin chain, being classified as β^+ and β^0, respectively [25]. As a result, there are three main forms of β-thalassemia [2, 24]:

- thalassemia major, a transfusion-dependent anaemia, subjects are homozygotes or compound heterozygotes for β^0 or β^+ genes;

- thalassemia intermedia, subjects are mostly homozygotes or compound heterozygote (β^+/β^+ or β^0/β^+), still have residual β-chain synthesis, causing mild to moderate microcytic hypochromic (in which the red blood cells are paler than normal) anaemia;

- thalassemia minor, the β-thalassemia carrier state, where subjects are mostly heterozygotes (β^{+}/β or β^{0}/β), leading to microcytosis and hypochromia, and mild or no anaemia.

Regarding α-thalassemia, there are two equal genes encoding the α-globin chain in each chromosome 16 and, for that reason, there are four types of α-thalassemias, two carrier states (trait 1 and 2) and two clinically relevant forms (hemoglobin H and Bart diseases) [2, 26–28]:

- trait 1, defect in only one α-globin gene ($-\alpha/\alpha\alpha$), cause no or very mild microcytic hypochromic anaemia;
- trait 2, defect in two α-globin genes in one allele ($--/\alpha\alpha$) or in one per allele ($-\alpha/-\alpha$), causing mild or severe microcytic hypochromic anaemia;
- hemoglobin H disease, deletion or mutation in three α-globin genes ($--/-\alpha$), which may lead to moderately severe microcytic hypochromic anaemia;
- hemoglobin Barts, complete absence of α-globin production ($--/--$), causing severe intrauterine anaemia resulting in hydrops fetalis which is lethal in utero or soon after birth.

The distinction between IDA and thalassemias is fundamental in order to prevent iron therapy in individuals with thalassemia trait, which could lead to iron overload, and also to provide genetic counselling to thalassemia carriers and their families, and to evaluate the need for prenatal diagnosis of thalassemia [29].

2.1.1 Diagnosis

Anaemia diagnosis is characterized by a hemoglobin concentration $<$ 13 g/dL for men and $<$ 12 g/dL for non-pregnant women [1]. Once diagnosed, it is classified into categories based on the mean corpuscular volume (MCV) of the red blood cells as: microcytic (MCV, $<$ 80 fL), normocytic (MCV, 80 $-$ 100 fL), or macrocytic (MCV, $>$ 100 fL) [19].

To get information about the MCV a complete blood count (CBC) test, one of the most common laboratory tests in medicine, is required. This test, apart from the MCV, provides information regarding [30, 31]:

- Hemoglobin (Hb);
- Red blood cell (RBC) count and red blood cell indexes, which give information on the physical features of the red blood cells: red blood cell distribution width (RDW), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC) and the MCV;
- White blood cell (WBC) count;
- Platelet (PLT) count;
- Hematocrit (Hct), which reflects percentage of the total blood volume that consists of packed red blood cells.

After the microcytic anaemia diagnosis, there are three major possibilities: IDA, thalassemia, and anaemia of chronic disease (ACD), see Figure 2.3.

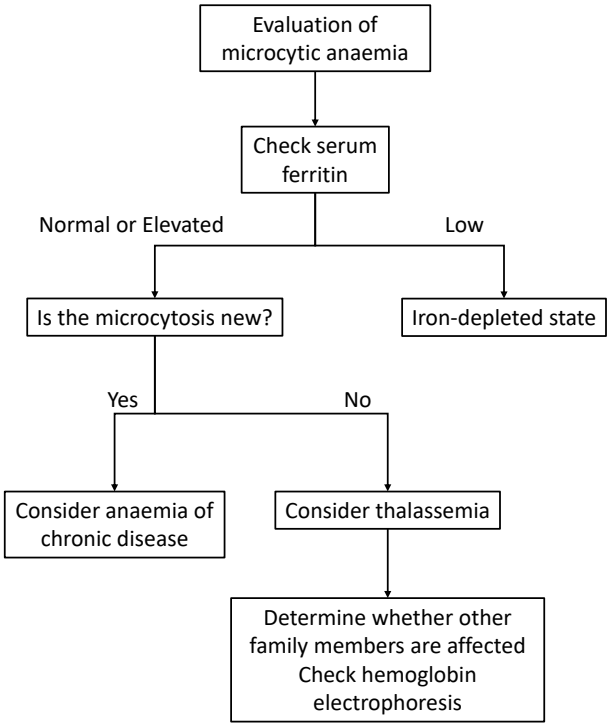


Figure 2.3: Microcytic anaemia evaluation, adapted from Tefferi [32].

First of all, it is important to evaluate the possibility of IDA since it is the most common cause. The most accurate test for the diagnosis of IDA is bone marrow biopsy, however this approach is costly and invasive [2]. Therefore, the diagnosis is usually based on low serum ferritin (Ft), low serum iron levels (Fe), low transferrin saturation (TS), high transferrin (Tf), and total iron binding capacity (TIBC), being the most sensitive for the IDA diagnostic the serum ferritin [33].

In the absence of inflammation, the concentration of serum ferritin, the ferritin secreted into the plasma, is an indicator of the size of the total body iron store, therefore a low serum ferritin level is a sign of depleted iron stores. Serum ferritin is normally accessed through enzyme-linked immunosorbent assays (ELISA) or enzyme immunoassays after venous blood collection [34].

When the serum ferritin level is normal, it is important to understand if the microcytosis is recent in the patient and in that case nonthalassemic conditions associated with microcytosis other than IDA such as ACD should be considered. On the contrary, if it had been earlier identified, it insinuates a congenital disorder, which is a strong indicator of a thalassemia diagnosis [19].

Once there is a strong suspicion of a thalassemic diagnosis the question is whether it is an α-thalassemia or a β-thalassemia. In both cases, a microcytic and/or hypochromic anaemia (in which the red blood cells are paler than normal) is observed. However, in presence of β-thalassemia trait there is an increased level of hemoglobin A2 ($>$3.5%).

As previously mentioned hemoglobin A is composed by two α-globin chains and two β-globin chains. However, hemoglobin A2 is different, being composed of two α-globin chains and two δ-globin chains.

That said, in the case of β-thalassemia the results of the hemoglobin evaluation (by electrophoresis or by high-performance liquid chromatography) are irregular, due to an increased in the hemoglobin A2 production as an adjust mechanism to compensate the reduce production of the β-globin chain. By a process of elimination, if the result of the analysis is normal, it is an indicator of α-thalassemia trait [14].

Nevertheless, DNA testing, as example by gap-polymerase chain reaction (gap-PCR) analysis, is mandatory to diagnose α-thalassemia and to determine the exact mutation that led to the α-thalassemic trait [18].

2.2 Machine Learning Algorithms

Nowadays large amounts of data are generated in the modern healthcare systems. This data can be analysed through machine learning algorithms to identify patterns with the aim of disease prevention as well as personalized disease diagnosis and treatment. As a result, the combination of machine learning and healthcare data can enhance the efficiency and quality of medical care, while reducing its costs [35]. One example of the application of these algorithms is the iLet, a bionic pancreas that manages the blood sugar levels in patients with type 1 diabetes mellitus, reducing the costs and burden of diabetes care. In comparison with an insulin pump, the bionic pancreas improves the management of the blood sugar, reducing the frequency of the hypoglycemic episodes [36].

Machine learning comes from the assumption that there is a process that explains the data observed, that there are patterns in the data, and even though we cannot completely identify them, it is possible to make a good and applicable approximation of it [6].

Supposing that the future will not be that different from the past, we can collect samples from which we know the outcome, constituting our training set, and make predictions about new samples by learning from the previous data.

One example of a training set can be a collection of CBC tests from a group of subjects, in this case the features could be Hb, RBC, MCV, and RDW values for each of them. A training set X is then constituted by N rows corresponding to independent and identically distributed instances, our subjects CBC tests, and d columns which are the features of those instances. In some cases, it is even possible to have corresponding outputs of those instances, which in this example would be the respective diagnosis of each of the subjects, healthy, IDA, or thalassemia carrier, Y:

$$X = \begin{bmatrix} X_1^1 & X_2^1 & \dots & X_d^1 \\ X_1^2 & X_2^2 & \dots & X_d^2 \\ \vdots & & & \\ X_1^N & X_2^N & \dots & X_d^N \end{bmatrix}, \mathrm{Y} = \begin{bmatrix} Y^1 \\ Y^2 \\ \vdots \\ Y^N \end{bmatrix} \tag{2.1}$$

That said, we expect that the computer (machine) can automatically extract a model: instructions to transform the input into an output.

In order to build these algorithms, machine learning uses the theory of statistics and optimization, to make inferences from the training set. For this, is necessary to define some parameters in our model

and using the training set the computer program will optimize them, in other words, it will learn from the data [6].

The model learning approach can be supervised or unsupervised (to get knowledge from the data).

In supervised learning we have a training set with variables from our data and their corresponding outputs. The most common types of supervised learning are regression, when the outputs may have any numerical value within a range, and classification, in which we may want to classify subjects into some discrete categories based on their characteristics, like a CBC test result.

An unsupervised model is when we have a training set with variables, an input, but no corresponding output. In this case the objective of pattern recognition can be, for example, to create groups of similar inputs, clustering our data, or for instance to transform data with high-dimensional space into more easy-to-visualize dimensions (two or three) [37].

From the wide range of classification algorithms, this work will only focus on some of the most important machine learning classification techniques, briefly described below.

Logistic Regression

The logistic regression approach for classification problems is an extension of the linear regression model, the main difference being that the outcome of logistic regression is a binary variable [38].

In linear regression models the relationship between output and input is given by a linear function of the parameters:

$$Z = w_o + w_1 X_1 + \cdots + w_p X_p. \tag{2.2}$$

Due to the fact that in logistic regression models it is more desirable to have values between 0 and 1, as we wish to predict discrete class labels, the logistic regression model forces the output to assume these values using an activation function, which can be the sigmoid function, that maps the whole real axis into a finite interval [37], see Figure 2.4.

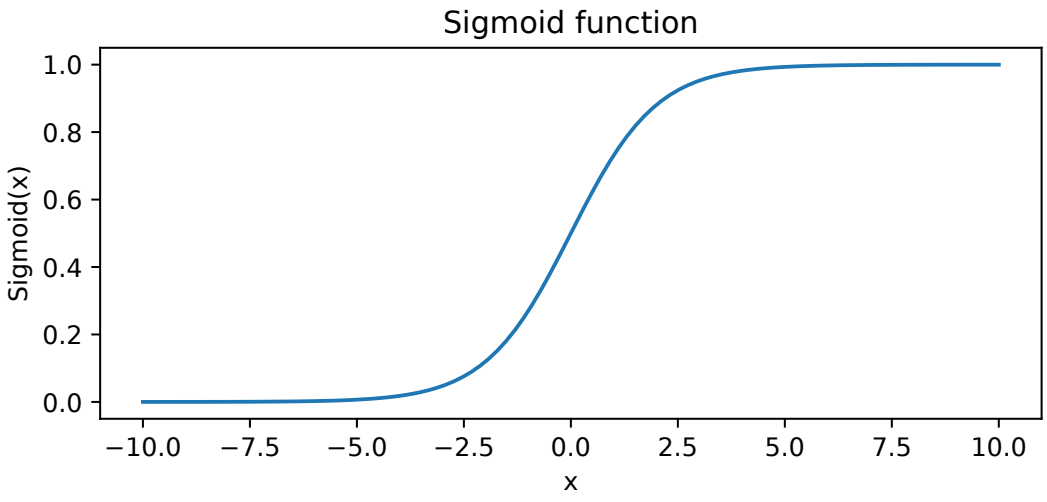


Figure 2.4: Sigmoid function.

Therefore, considering the case of two-classes classification (C_1 and C_2), and $Y \in (0, 1)$, $Y = 1$ represents C_1 and $Y = 0$ represents C_2. We can represent the probability of C_1 as the $P(Y = 1 \mid X)$ [37]:

$$P(Y = 1 \mid X) = \frac{1}{1 + \exp(-(w_o + w_1 X_1 + \cdots + w_p X_p))}. \tag{2.3}$$

As a result we can use 0.5 as a threshold and assume that if the $P(Y = 1 \mid X) = 0.7$, for example, our input belongs to C_1, bearing in mind that it has a 70% probability of belonging to C_1 and 30% probability of belonging to C_2.

Logistic regression can be generalize to multi-class problems, i.e. more than two possible discrete outcomes ($k > 2$), using the multinomial logistic regression algorithm. This algorithm is a simple extension of the logistic regression comprising k discriminant functions, each evaluating the probability of an instance X belonging to a certain class $P(C_k \mid X)$. Then the instance is simply assigned to the class C_k if $P(C_k \mid X) > P(C_j \mid X)$ for all $j \neq k$ [37].

Naive Bayes

The naive Bayes algorithm arises from the "naive" assumption that the features of the instances are independent given the class (conditional independence), along with the application of the Bayes' theorem:

$$P(Y \mid X_1, \ldots, X_p) = \frac{P(Y)P(X_1, \ldots, X_p \mid Y)}{P(X_1, \ldots, X_p)}. \tag{2.4}$$

Ignoring the probability of an instance ($P(X_1, \ldots, X_p)$), as it is a common factor for all classes, and therefore it does not influence the result. The probability of class Y given an instance's features ($P(Y \mid X_1, \ldots, X_p)$) is calculated by multiplying the probability of a class (the prior probability, $P(Y)$) by the probability of the instances features given that class ($P(X_1, \ldots, X_p \mid Y)$). Since we assumed that the features are independent given the class, it is simply calculated by multiplying the probability of each feature given the class [39]:

$$P(X_1, \ldots, X_p \mid Y) = P(X_1 \mid Y) \times P(X_2 \mid Y) \times, \ldots, P(X_p \mid Y) = \prod_{j=1}^{p} P(X_j \mid Y). \tag{2.5}$$

Therefore the probability of class given the instance' features can be calculated as:

$$P(Y \mid X_1, \ldots, X_p) \propto P(Y) \prod_{j=1}^{p} P(X_j \mid Y). \tag{2.6}$$

Afterwards the Maximum A Posterior (MAP) classification rule is applied: for a certain instance we search among all the classes, for the class where the $P(Y \mid X_1, \ldots, X_p)$ is the highest and assign the instance to that class.

In case of a continuous data set it is commonly assumed that the values within each class have a normal (Guassian) distribution. This distribution can be represented by its mean (μ) and standard

deviation (σ), allowing the probability of the instance feature given that class to be calculated as [40]:

$$P(X_j \mid Y) = g(X_j; \mu_Y; \sigma_Y) = \frac{1}{\sqrt{2\pi\sigma_Y}} e^{-\frac{(X_j - \mu_Y)^2}{2\sigma_Y^2}}. \tag{2.7}$$

Even though the conditional independence assumption rarely holds true, the naive Bayes algorithm has shown a good performance even when compared to other classifiers in real-world situations, being one of the most efficient and effective algorithms. Although it is a good classifier it shows poor performance in the output probability estimation [41].

Artificial Neural Network

Artificial neural networks (ANN) were designed with the aim of finding a mathematical representation of the way a human brain processes and analyses information. In the brain a neuron is a cell that communicates with other cells through synapses, in a simplified form we can say that neurons work as computational units that receive an input in the dendrites and propagate it until the axons.

On an ANN, given an input with a set of features and an output, the algorithm learns a non-linear function. A network learns by processing examples forming a probability-weighted association between the input and the output, see Figure 2.5).

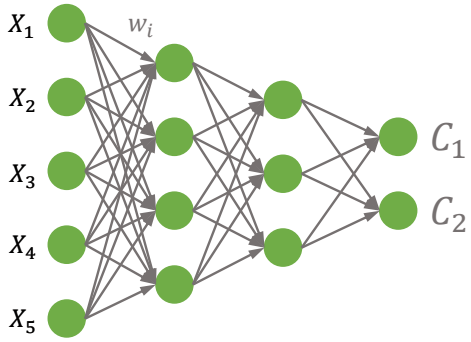


Figure 2.5: Artificial neural network example.

In an ANN the connections between the neurons are associated with a weight (w_i), by which the inputs are multiplied and their sum is sent as an input to the hidden layer. The value obtained is passed to an activation function and then the signal is propagated to the neuron of the next layer. In the last layer (the output layer) in the case of a classification problem, there is an activation function, and therefore the neurons indicate the probability of the given input belonging to each class. In regression problems there is no activation function in the output layer, so the output is a set of continuous values. For the learning task back-propagation is commonly used. This procedure repeatedly adjusts the weights of the neurons connections in order to minimize the loss function, which measures the difference between the prediction obtained and the true output [42].

One of the main differences between ANN and logistic regression is that linking the input to the output there can be a series of non-linear hidden functions, layers with several neurons, that perform different transformations on their inputs for either classification or regression problems.

Support Vector Machine

Support vector machine (SVM) models are used to solve problems of classification, regression, and outliers detection. The idea of the SVM algorithm is to find an optimal hyperplane, defined as the decision function, with maximal margin between the support vectors, the training instances of the two classes that are closer to the hyperplane. That said, hyperplanes are decision boundaries that allow us to classify new instances. Depending on the side of the hyperplane that the new instances are assigned to, they are attributed to different classes [43].

In the simplest case where we have two classes (C_1 and C_0) with only two features (X_1 and X_2) and the instances are linearly separable, we start by defining two hyperplanes, each hyperplane delimits a class, allowing for the distances between the hyperplanes to be the largest possible, in this example:

$$w_1 \times X_1 + w_2 \times X_2 + w_o = 1 \text{ for } C_0 \tag{2.8}$$

$$w_1 \times X_1 + w_2 \times X_2 + w_o = -1 \text{ for } C_1 \tag{2.9}$$

These two hyperplanes now define a region called margin and the decision function with the maximal margin between the classes that we wish to define is simply a third hyperplane right in the middle of the other two:

$$w_1 \times X_1 + w_2 \times X_2 + w_o = 0 \tag{2.10}$$

In Figure 2.6 is a plot that illustrates how these hyperplanes separate the data. As a result, SVM works as a decision machine so does not provide the posterior probabilities [37].

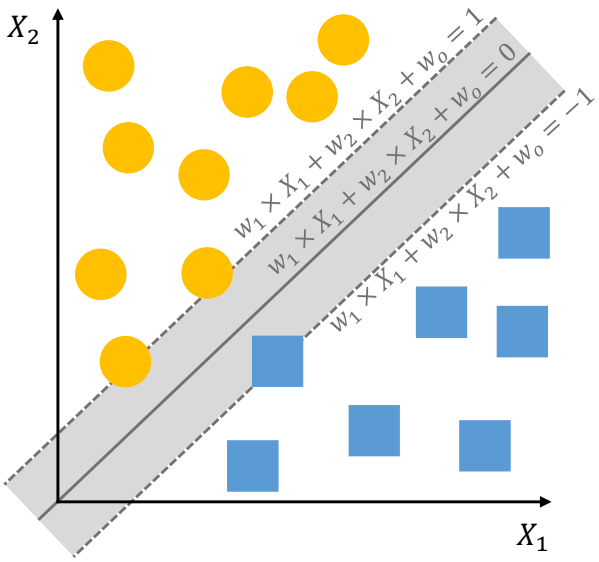


Figure 2.6: Support vector machine example, adapted from Cortes and Vapnik [43].

Decision Tree

Decision trees are a hierarchical data structure, which can be used for classification or regression. Decision trees are models that predict an output by learning decision rules from the training data. After learning the decision rules, the predictive model is able to go from observations about a new instance (the branches of the tree) to and output of that instance (the leaves of the tree) by applying those decision rules [44].

One of the main advantages of decision trees is the fact that it can be simply understood and visualized, at least until 3 dimensions, as in Figure 2.7, facilitating its interpretation.

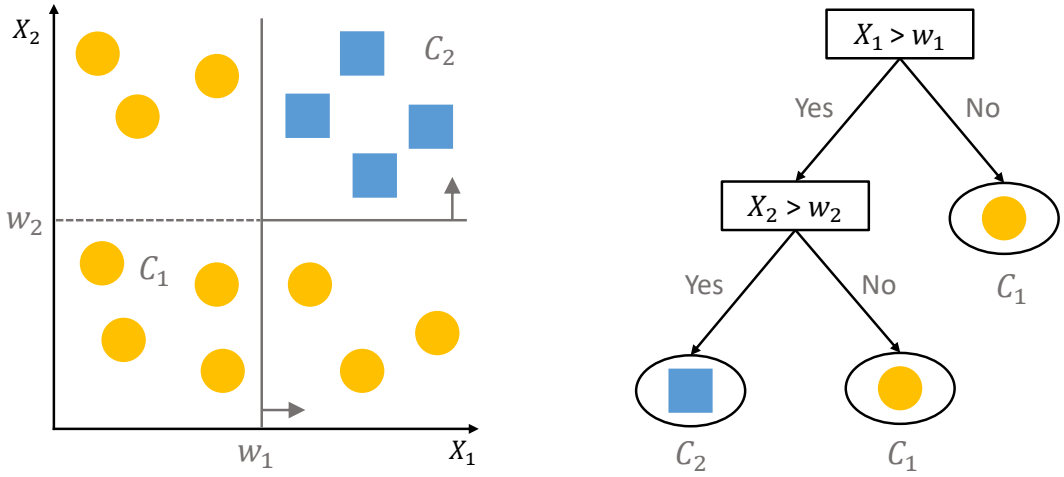


Figure 2.7: Data set example and corresponding decision tree. The rectangle nodes are decision rules, while the ovals, the leaves of the tree, are the output. Adapted from Alpaydin [6].

The most common decision tree algorithms are the ID3 (Iterative Dichotomiser 3), the C4.5 (the successor to ID3), C5.0 and the CART (Classification and Regression Trees). These algorithms seek to find for each node of the tree the feature and threshold that will yield the largest information gain for the classification of an input (in the case of the CART, classification or regression) [45].

Random Forest

Random forest classifier is an ensemble method that results from a combination of decision trees.

In random forest, each tree in the set is constructed from a sample of the training data taken with replacement, known as bootstrap, which reduces variance and helps to avoid overfitting of the model. Other source of randomness to decrease the variance is in the construction of the tree nodes, where the best split can either be found from all input features or a random subgroup of them.

That being so, the objective is to construct a number of estimators independently and then, the overall output is the classification among all the trees having the most votes, improving the generalizability and robustness of the prediction [46].

K-Nearest Neighbors

K-nearest neighbors is a nonparametric classifier/estimator used for classification or regression. In nonparametric classification it is simply assumed that similar observations share the same class [6].

In that sense, in K-nearest neighbors classification, a new instance is classified based on a predefined number of training samples (K) closest in distance to the new instance, that way the new instance is ranked in the most common class among the K training samples [47]. If $K = 1$, the new instance is simply assigned to the same class as the nearest training sample. In Figure 2.8 there is a graphical example of a classification with different values of K. There are several ways to calculate this distance, being the most common the Euclidean distance (straight line distance).

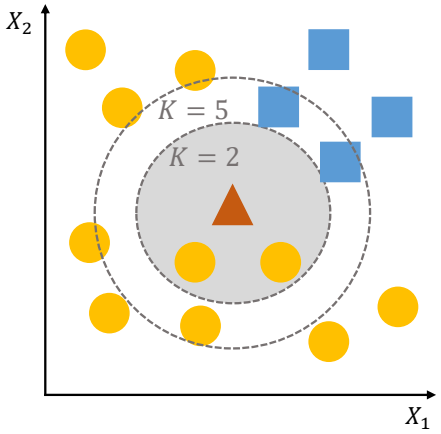


Figure 2.8: K-nearest neighbors example.

2.2.1 Hyperparameter Optimization

As previously stated in machine learning the training set is used to estimate the parameters of our model, the learning process. The hyperparameters are the parameters whose values are used to control the learning process.

Different machine learning models have different hyperparameters and the problem of selecting good values for those hyperparameters is known as hyperparameter optimization [48]. Some of those hyperparameters are for example, in the case of random forest, the maximum depth of the trees and the number of trees in the forest; in logistic regression, the algorithm used to solve the optimization problem of the model and the maximum number of iterations of the solver; in the K-nearest neighbors, the K number of neighbors; while in the ANN, the number and size of hidden layers and the activation function of those layers.

Being Λ list of hyperparameters candidates, A_λ a learning algorithm with a set of chosen hyperparameters, $X^{(train)}$ the training set, $X^{(test)}$ the test set and L a loss function (which measures the error between the predicted value and the actual value), the hyperparameter (λ) optimization problem, in practice, can be addressed as the following equation [48]:

$$\lambda^{(*)} \approx \underset{\lambda \in \Lambda}{\operatorname{argmin}} \underset{\chi \in X^{(test)}}{\operatorname{mean}} L(\chi; A_\lambda(X^{(train)})). \tag{2.11}$$

The most commonly use strategy to solve the hyperparameters optimization problem is a combination of manual search and grid search. Manual search is simply the manual selection and testing of different hyperparameter values, without any automation in the selection of those. In grid search it is necessary to manually choose a set of values for each hyperparameter, but afterwards an exhaustive search through a set of trials with every possible combination of those values is done, which implies that the number of trials grows exponentially with the number of hyperparameters, hence this solution ends up suffering from the curse of dimensionality. Although these solutions are the most widely used, the best approach has proven to be the random search, which replaces the exhaustive enumeration of all possible combinations by randomly selecting them, making it a much more efficient solution because not all hyperparameters are equally important to tune [48].

2.3 Evolutionary Algorithms

Evolutionary algorithms (EA) are a subset of evolutionary computations (EC) that performs optimization or learning tasks. EA are well known by their efficiency in developing good approximate solutions to difficult problems [49].

Inspired by Darwinian natural evolution, EA simulates evolution resorting to mechanisms such as reproduction, mutation, recombination, and selection to generate solutions for complex real-world problems. Starting with an initial population of random individuals (solutions to a specific problem) a fitness function is applied to evaluate the candidates and the best candidates are then chosen to seed the next generation. The next generation will be created by applying recombination and/or mutation to two or more parents (the solutions with the best fitness), leading to the creation of an offspring. Afterwards this new offspring suffers a selective pressure, that is, using the same fitness function the offspring is evaluated, having to compete not only with their siblings but also with the previous generation for a place in the next generation. These steps are then iterated until an individual achieves a desired fitness score or the number of generations reaches a limit previously established [50]. An example of an EA pseudo code is depicted in Algorithm 1.

Algorithm 1 Evolutionary algorithm pseudocode, adapted from Eiben et al. [50].

```
INITIALISE population with random candidate solutions
EVALUATE each candidate
while TERMINATION CONDITION is not satisfied do
  SELECT parents
  RECOMBINE pairs of parents
  MUTATE individuals
  EVALUATE new candidates
  SELECT individuals for the next generation
end while
```

While maintaining these general outline EA are further divided in genetic algorithm (GA), genetic programming (GP), evolutionary programming (EP) and evolution strategies (ES) [50].

Genetic algorithm

The most known EA is the genetic algorithm (GA). In genetic algorithms the candidate solutions are often referred as chromosomes. Chromosomes are a way of coding the features rather than working with the features themselves, in that sense, a chromosome is a string with finite length of alphabets (genes) of certain cardinality [51]. As an example, we can think about the features obtained in a CBC test. In order to make a classifier it is possible to use several different combinations of the features, however, testing exhaustively all those combinations would be very demanding. Instead, we can use a string of 0s or 1s, where 0 means that a feature is not used to construct the classifier and 1 the opposite. Now each chromosome represents a set of features chosen to build a classifier and each gene a feature of that classifier. Afterwards by using the steps represented in Figure 2.9, an optimal solution can be achieved.

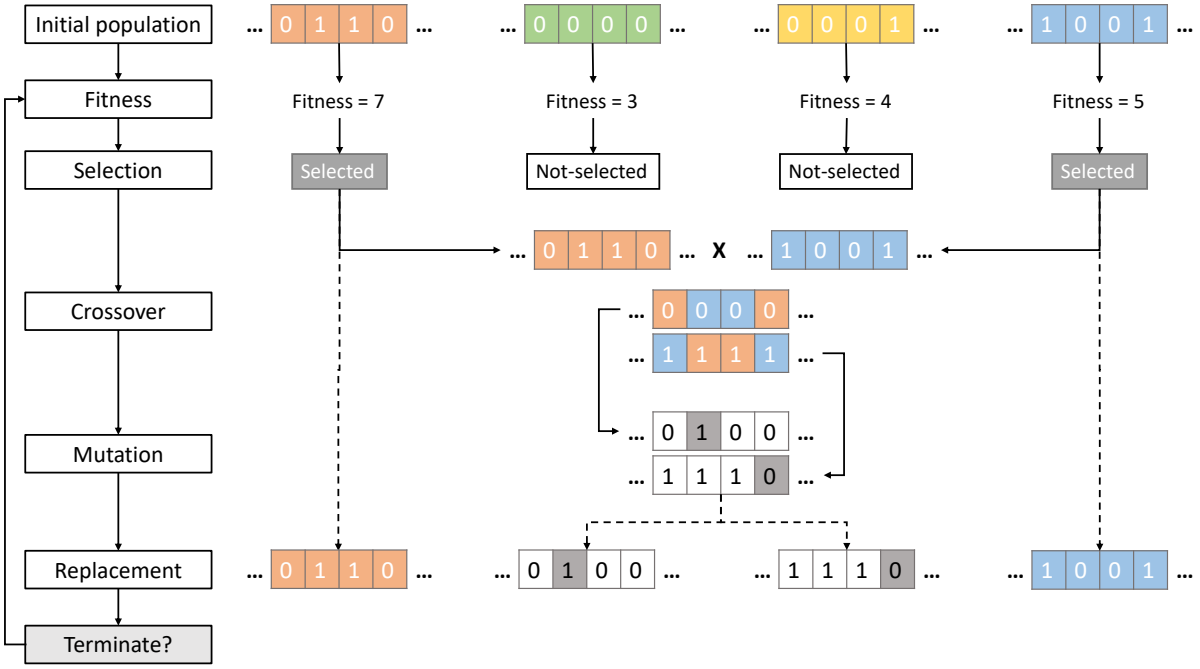


Figure 2.9: Genetic algorithm scheme, adapted from [52]

Genetic programming

Genetic programming (GP) was first introduced by Koza [53] in 1992 as an evolutionary algorithm. In GP it is not required to specify the structure of the solution in advance and, therefore, GP allows to solve the problem without been explicitly told how to do it. In this case the evolution is applied to a population of random programs to solve a specific problem [54]. This programs are traditionally represented as tree structures, as illustrated in Figure 2.10. In this structures the functions, arithmetic operations, are the internal nodes of the tree and the terminals (the leaves) the features and constants in the program (combining features with operators instead of just selecting them), allowing mathematical expressions/ programs to be easily evolved and evaluated.

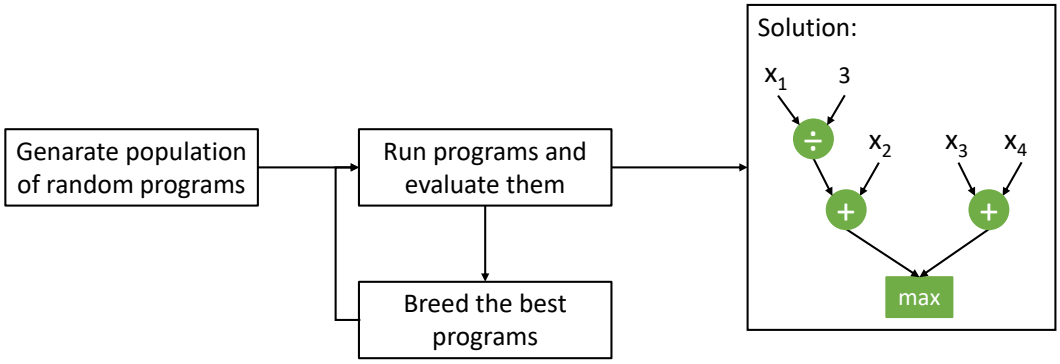


Figure 2.10: Genetic programming scheme, adapted from Poli et al. [54].

2.4 Evaluation Metrics

To build the classifiers, and to further evaluate which is the optimal among them, it is necessary to measure the effectiveness of the classifiers through evaluation metrics.

In case we are dealing with a binary classification problem, C_1 if the patient has a certain disease and C_0 otherwise, a confusion matrix, illustrated in Table 2.1, can be built in order evaluate our classifiers [55].

Table 2.1: Confusion matrix.

	Actual Positive Class	Actual Negative Class
Predicted Positive Class	True positive (*tp*)	False positive (*fp*)
Predicted Negative Class	False negative (*fn*)	True negative (*tn*)

In a confusion matrix the *tp* and *tn* represent correctly classified instances while *fp* and *fn* the misclassified. Afterwards, using these scoring parameters several evaluation metrics with different focus can be generated to further characterize our models.

Accuracy (acc),

$$\text{Accuracy} = \frac{tp + tn}{tp + fn + fp + tn}, \tag{2.12}$$

which measures the fraction of correct predictions over all the instances, is one of the most applied.

Error Rate (err),

$$\text{Error Rate} = \frac{fp + fn}{tp + fn + fp + tn}, \tag{2.13}$$

on the contrary, indicates the fraction of incorrect predictions over all the instances and therefore is the same as $1 -$ Accuracy.

Sensitivity (sn), **Recall** (r) or **true positive rate** (*tp rate*),

$$\text{Sensitivity} = \frac{tp}{tp + fn}, \tag{2.14}$$

measures the ratio of positive predictions that are correctly identified over all the actual positives.

Specificity (sp),

$$\text{Specificity} = \frac{tn}{tn + fp}, \tag{2.15}$$

in contrast, indicates the ratio of negative predictions that are correctly identified over all the actual negative.

The ratio of negative predictions that are incorrectly identified are called the **false positive rate** (*fp rate*) and can be calculated as $1 -$ Specificity.

Precision (p),

$$\text{Precision} = \frac{tp}{tp + fp}, \tag{2.16}$$

refers to the portion of positive predictions that were correctly evaluated over all the positive predictions.

F-Score (FS),

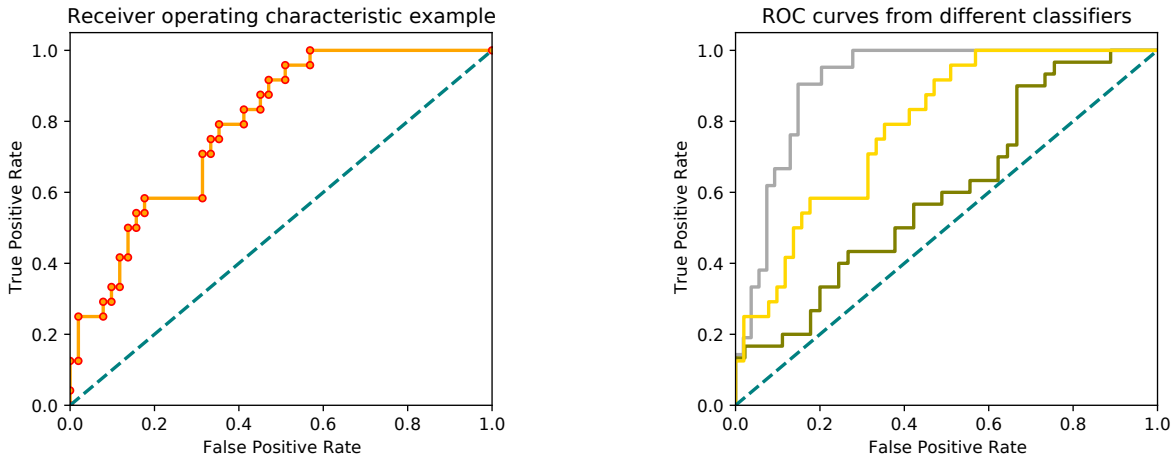
$$\text{F-Score} = \frac{2 \times p \times r}{p + r}, \tag{2.17}$$

is the harmonic mean between the precision and recall.

The area under the ROC curve (AUC)

A receiver operating characteristic (ROC) curve is a graphical plot that can be used to optimize and compare models as it illustrates the sensitivity and specificity of the classifier using different discriminant thresholds - the value that must be exceeded for an observation to be classified in a certain class. This is relevant in classifiers created with algorithms like the logistic regression, where we assume that if the $P(Y \mid X)$ crosses a previously established discriminant threshold, that instance (Y) belongs to a certain a class.

The ROC curve is simply constructed by plotting the *tp rate* (Sensitivity) on the Y axis and the *fp rate* $(1 -$ Specificity) on the X axis, as shown in Figure 2.11.



(a) ROC curve example.

(b) ROC curves from different classifiers.

Figure 2.11: ROC curve, adapted from Alpaydin [6].

By selecting different thresholds in our model, we can plot different points in the graph which will represent the trade-off between the benefits (true positives) and the costs (false positives). The perfect

threshold would be the one representing the point $(0, 1)$ in the graph as it would mean a 0 *fp rate* and an 1 *tp rate*. Hence, we can consider that the best operating point of the ROC curve, i.e., best threshold, corresponds to the closest point to the top left corner.

Afterwards to compare different classifiers we must have in attention the area under the curve (AUC), which summarizes the performance of our classifier into a single scalar value in the interval $[0, 1]$.

A random classifier would generate the diagonal line $y = x$, whose AUC value would be 0.5, that said no reasonable classifier should have a AUC value below 0.5.

Monte Carlo cross-validation

We can simulate how well a model will generalize to an independent data set by splitting the data that we have in a training set ($X^{(train)}$), to fit the model, and in a test set ($X^{(test)}$) to test the generalization ability, this method is known as the holdout method [56]. However, one disadvantage of this solution is assuming that the instances are not equally difficult to classify, the model's accuracy estimation can have large variation depending on the data split done. Some data splitting will end up having an easier/harder test set to classify than others, which will have an impact on the estimated model's accuracy and therefore can introduce a bias. To work around this problem the repeated random sub-sampling validation, also know as the Monte Carlo cross-validation, can be used instead. The Monte Carlo cross-validation is basically a repetition of the holdout method k times, and the estimated accuracy is derived by averaging the k iterations [56], see Figure 2.12.

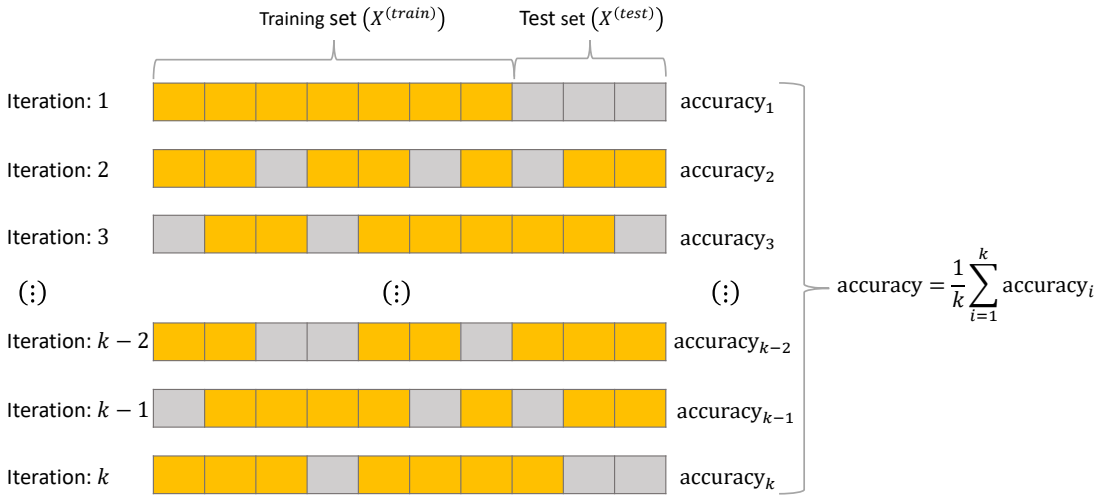


Figure 2.12: Monte Carlo cross-validation scheme, adapted from Patro [57].

2.5 Outlier Detection

An outlier, also referred to as an abnormality, a discordant, a deviant, or an anomaly in the data, can be defined as "(...) an observation which deviates so much from the other instances as to arouse suspicions that it was generated by a different mechanism." [58]. Therefore, it consists of instances that occur when the process that generates the data behaves in an unusual way.

As stated earlier, in machine learning we seek to find patterns in data. Thus, if an instance does not follow the general rule, i.e. it differs significantly from others, it can be considered an outlier which, if not detected, can compromise the model. For this reason, over the years, several methods have been developed to perform outlier detection.

Cook's Distance

Cook's distance, named in honour of R. Dennis Cook, is a measure of an instances' influence on a linear regression model, used to indicate which instances need to be checked for validity [59].

The Cook's distance of instance i (D_i) (for $i = 1, \ldots, n$) is defined as the sum of the changes in the regression model when instance i is deleted from it, measuring the effect on the predictions when that instance is removed:

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{y}_{\,j} - \widehat{y}_{\,j(i)} \right)^2}{ps^2}, \tag{2.18}$$

where $\widehat{y}_{\,j}$ is the fitted response value obtained with the full sample, $\widehat{y}_{\,j(i)}$ is the fitted response value obtained when excluding i from the sample, p is the number of features fitted in the model, and

$$s^2 = \frac{1}{n} \sum_{k=1}^{n} \left(y_{\,k} - \widehat{y}_{\,k} \right)^2, \tag{2.19}$$

is the mean squared error of the regression model, where $y_{\,k}$ is the actual response value of instance k and $\widehat{y}_{\,k}$ is the fitted response value of that instance.

Generally, an instance is considered an outlier when $D(i) > 4/n$, being n the number of instances [60].

Silhouette analysis

The silhouette method is used to find an optimal number of classes, as well as the consistency within classes, providing a succinct graphical representation of how well each instance is considered to be classified [61].

In a silhouette analysis, each class is represented by several silhouette coefficients that shows how similar each instance is to its own class compared to other classes. Then by combining the silhouettes of the classes in a single graph, an appreciation of the relative quality of the classes can be done. In order to assess class validity, the mean silhouette is calculated and can be used to select an optimal number of classes [61].

The silhouette coefficient of an instance (i) can be calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \tag{2.20}$$

Using as an example Figure 2.13.

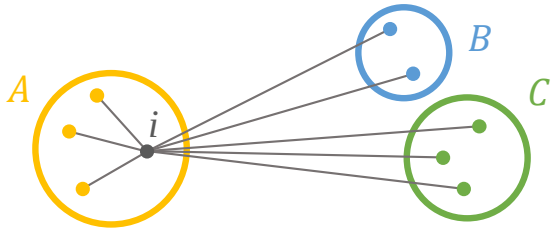


Figure 2.13: Illustration of the elements involved in the computation of $s(i)$, where the instance i belongs to class A. Adapted from Rousseeuw [61].

Where instance i belongs class A, $a(i)$ is defined as:

$$a(i) = \text{average dissimilarity of } i \text{ to all other instances of } A. \tag{2.21}$$

For all classes $Z \neq A$ (in this example classes B and C), calculate the average length of all lines going from i to Z:

$$d(i, Z) = \text{average dissimilarity of } i \text{ to all instances of } Z. \tag{2.22}$$

Afterwards select the smallest of those numbers, effectively the average distance of i to all the instances in the closest class, and denote it by:

$$b(i) = \min_{Z \neq A} d(i, Z), \tag{2.23}$$

From $s(i)$ equation we can easily see that $-1 \leq s(i) \leq 1$. When $s(i)$ has a high value it means that the within dissimilarity $a(i)$ is smaller than the smallest dissimilarity of $d(i, Z)$ and therefore the instance i is well clustered, being the worst situation when $s(i)$ is close to -1, suggesting that the instance has been misclassified [61].

2.6 State of the Art of Anaemia Classifiers

2.6.1 Anaemia Indexes

Due to the clinical relevance of a proper distinction between microcytic anaemias, several indexes, based on the blood cell parameters obtained in the CBC test, were constructed over the years. The intention was to propose a suitable discrimination between IDA and β-thalassemia trait, while avoiding a time-consuming and expensive method. In 1973 England et al. [62] published a discriminant function, able to differentiate IDA from β-thalassemia trait. In the exact same year, two more new formulas were published. Mentzer [63], built a simpler formula, as well as Shine and Lal [64]. Since then several indexes have been developed, as summarized in Table 2.2, representing important indexes to discriminate between both conditions. These indexes can be especially advantageous in countries with less financial resources and more limited health systems.

Table 2.2: Discriminant indexes, formula and cut-off values.

Index	Formula	Cut-off	Reference
England and Fraser	$\text{MCV} - \text{RBC} - (5 \times \text{Hb}) - 3.4$	0	[62]
Mentzer	MCV/RBC	13	[63]
Shine and Lal	$\text{MCV}^2 \times \text{MCH}$	1.53	[64]
Ricerca	RDW/RBC	4.4	[65]
Green and King (G&K)	$(\text{MCV}^2 \times \text{RDW})/(100 \times \text{Hb})$	65	[66]
RDWI	$(\text{MCV} \times \text{RDW})/\text{RBC}$	220	[67]
Sirdah	$\text{MCV} - \text{RBC} - (3 \times \text{Hb})$	27	[5]
Ehsani	$\text{MCV} - (10 \times \text{RBC})$	15	[68]
Telmissani–MCHD	MCH/MCV	0.34	[69]
Telmissani–MDHL	$(\text{MCH} \times \text{RBC})/\text{MCV}$	1.75	[69]
Index26	Combination of multiple indexes	16	[4]
CRUISE	$\text{MCHC} + 0.603 \times \text{RBC} + 0.523 \times \text{RDW}$	42.63	[4]
Matos and Carvalho	$1.91 \times \text{RBC} + 0.44 \times \text{MCHC}$	23.85	[70]
Bessman	RDW	14	[71]
Srivastava	MCH/RBC	3.8	[72]

MCV is the mean corpuscular volume of the red blood cells, RBC the red blood cell count, Hb the hemoglobin concentration, MCH the mean corpuscular hemoglobin, RDW the red blood cell distribution width, and MCHC mean corpuscular hemoglobin concentration.

All the indexes shown in Table 2.2, are based on the MCV, RBC, Hb, MCH, RDW or MCHC, indicating that the values that depend on the physical features of the red blood cells obtained on a CBC test are the ones that give more information when differentiating between these diseases.

In the Portuguese population it was found that the index that has a better performance on the female sex is the RDWI [73]. However, during the research for this thesis, was not found any study that evaluated the performance of these indexes in individuals of both sexes in the Portuguese population. Yet, multiple studies have tested these indexes in other populations. For example, it was concluded that G&K and RDWI indexes, with an accuracy of 88.4% and 92.0% respectively, provided the highest reliabilities in differentiating β-thalassemia trait from IDA in the Brazilian population [3]. In the Iranian population the discriminating formula with the better performance was found to be Index26 with an accuracy of 84.7% [4]. In another study, conducted with the Palestinian population the best indexes were Sirdah, G&K and the RDWI, all with an AUC of 0.91 [5]. Accordingly, it is plausible to say that some discriminating formulas are better adjusted to a specific population than others. Even though some show a better performance, none of them is 100% reliable, revealing the need to continue this search to develop a more efficient index.

2.6.2 Machine Learning in Anaemia Classification

As previously mentioned the application of machine learning algorithms is not a novelty in disease diagnosis and not even in the classification of anaemia, as in the last decade previous works have already demonstrated the usefulness of different machine learning algorithms, see Table 2.3.

Table 2.3: Anaemia classifiers constructed with machine learning algorithms.

Method	Description	Sample size	Acc (%)	Reference
Naive Bayes	anaemic/ non-anaemic	2151	85	[7]
Naive Bayes	anaemic/ non-anaemic	200	96	[8]
Random forest	anaemic/ non-anaemic	200	95	[8]
C4.5 decision tree	anaemic/ non-anaemic	200	95	[8]
C4.5 decision tree	anaemic/ non-anaemic	514	98	[9]
Support vector machine	anaemic/ non-anaemic	514	87	[9]
Support vector machine	β-thalassemia/ non-anaemic	20	99	[10]
K-nearest neighbors	β-thalassemia/ non-anaemic	20	99	[10]
Artificial neural network	β-thalassemia/ non-anaemic	20	99	[10]
Artificial neural network	IDA/ β-thalassemia	268	93	[11]

These studies referred in Table 2.3 have presented very exciting results regarding the performance of their models, presenting great accuracies either in the discrimination between anaemic and non-anaemic individuals, or between microcytic aneamia individuals. Nevertheless, it is important to keep in mind that not all of these studies used the same methodology to assess the accuracy of their models.

Regarding the information used to build these models, in the vast majority of the studies mentioned in Table 2.3 the models are only based on information obtained through a CBC test, only the study conducted by Purwar et al. [10] used CBC test data fused with blood film features, which probably explains why its accuracy is so high. However, none of the studies analysed used data from the Portuguese population which may hamper the correct classification for these patients. Besides this problem, few studies used machine learning in the classification of microcytic anaemia. Hence, it becomes relevant to test these algorithms in the Portuguese population as it may allow us to construct new indexes, specifically adjusted to this population, able to differentiate IDA from β-thalassemia trait, and even go a little further and build a multi-class classifier able to discriminate between various microcytic anaemias and consequently provide a more efficient, accurate and cost-effective diagnosis.

Chapter 3

Implementation

The implementation of this dissertation is divided into two parts, a laboratory part and a computational part, reflecting the two components of this work.

3.1 Dataset Description

The DNA samples used in the molecular diagnosis were made available by the ”Grupo de I&D em Hemoglobinopatias, Metabolismo do Ferro e Patologias Associadas” of INSA.

The data used to test the indexes mentioned in section 2.6.1 and to train and test the predictive models had been previously obtained in Exame Nacional de Saúde com Exame Físico, providing samples and data obtained within the scope of the INSEF 2015 project [74] carried out in 2015 by Department of Epidemiology of INSA, and by Bárbara Faleiro [73] and Daniela Santos during their masters dissertations research. This data was acquired in the Portuguese population and in order to address one of the main objectives of this thesis, which is to reduce the cost of the diagnosis, is solely composed by information obtained with a complete blood count test [hemoglobin (Hb), red blood cell distribution width (RDW), mean corpuscular hemoglobin (MCH), mean corpuscular volume (MCV)], the subjects' sex, and his confirmed diagnosis. The dataset used to train and test these predictive models is composed by a total of 390 instances, Table 3.1. While all those instances were used in the multi-class models, only 196 (β-thalassemia carriers plus IDA patients) were used in the binary models. In order to be able to depict the different groups of numerical data a descriptive data analysis was done resorting to boxplots, and the calculation of p-values, mean, median, etc.

Table 3.1: Data description.

	β-thalassemia	α-thalassemia	IDA	Control	**Total**
Female	68	32	54	97	251
Male	64	20	10	45	139
Total	132	52	64	142	**390**

3.2 Molecular Diagnosis

In this thesis in order to exemplify, the molecular diagnosis of β-thalassemia and α-thalassemia was carried out with samples from five individuals with β-thalassemia and three with α-thalassemia. In case of the β-thalassemia, the diagnosis was made through the analysis of mutations in the β-globin gene (exon 1, 2 and 3) by Sanger sequencing, while gap-PCR analysis was used for the identification of $-\alpha^{3.7}$ Kb deletion, a common deletion underlying α-thalassemia, being a predominant mutation in African, Mediterranean and Asian subjects [75]. For both methods DNA quantity and quality were assessed using a NanoDrop One (Thermo Fisher Scientific, USA) spectrophotometer, the concentration of DNA (ng/µL) was estimated by measuring the absorbance at 260 nm and the quality assessed by measuring the absorbance at 260/280 nm and a 260/230 nm, to determine the presence of proteins, salts, and residual phenol.

3.2.1 β-thalassemia

The human β-globin gene cluster in chromosome 11 consists of five functional genes being one of them the *HBB* gene, which encodes the β-globin chain, [76], as exemplified in Figure 3.1.

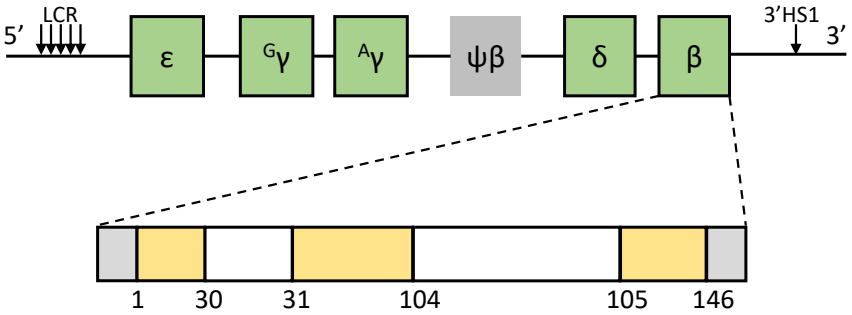


Figure 3.1: Illustration of the human β-globin multigene loci residing on chromosome 11 (11p15.4). Each locus consists of functional genes (arranged according to their order of developmental expression) and pseudogene (depicted in grey). The β-globin gene consists of 3 exons (yellow boxes) and 2 introns (white boxes), flanked by untranslated regions (UTRs, grey boxes), the codon numbers are indicated underneath. Adapted from Patrinos et al. [76]

In order to find mutations in the *HBB* gene to diagnose β-thalassemia, after the DNA quantity and quality assessment, the *HBB* gene was amplified through polymerase chain reaction (PCR). This technique is performed in a thermocycler, in this particular case in a Biometra® thermocycler, that in repeated cycles heats and cools the reaction tubes, with all the necessary reagents:

- DNA sample, with target DNA;
- Primers, two short DNA sequences designed to bind to the start and end of the DNA target;
- *Taq* DNA polymerase, which catalyses the DNA synthesis;
- Free 2'-deoxynucleotide triphosphates (dNTPs), for DNA synthesis;
- Buffer, to maintain the pH of the solution relatively stable.

At the end of each cycle the amount of DNA target doubles, as the newly synthesized DNA segments will serve as DNA templates in later cycles, through repetitive cycles the concentration of the DNA target will increase exponentially [77].

After the PCR, the amplification of the DNA target was confirmed by agarose gel electrophoresis and a purification of PCR products was performed with illustra^{TM} ExoProStar^{TM} 1-Step (GE Healthcare, USA).

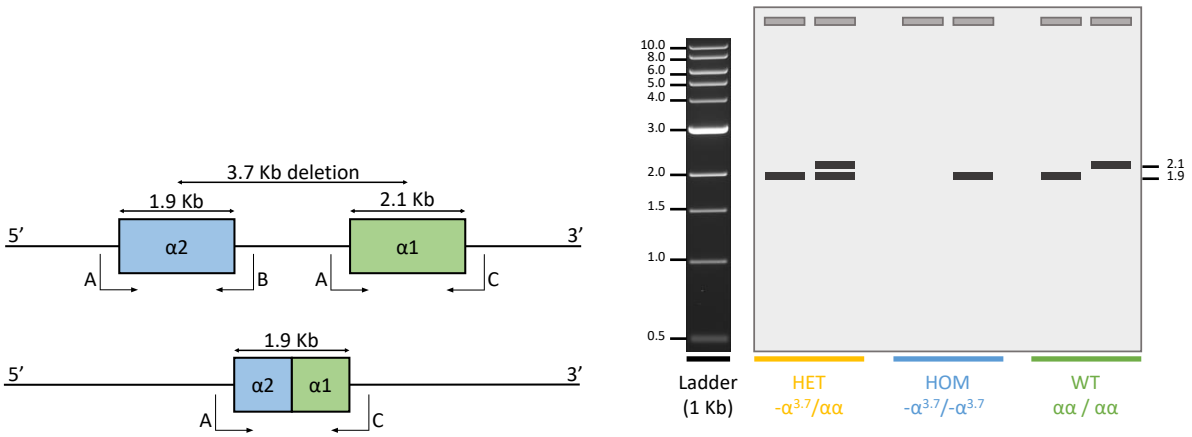
Later, to search for possible mutations causing the β-thalassemia, the *HBB* amplified gene was sequenced by Automated Sanger Sequencing, a well-known technique developed by Sanger et al. [78] that allows DNA sequencing based on the selective incorporation of fluorescent chain-terminating 2', 3'-dideoxynucleotide triphosphates (ddNTPs) by DNA polymerase during in vitro DNA replication.

The results were then analysed using FinchTV 1.4.0 [79] and compared with the RefSeq sequence of the *HBB* gene (Transcript: HBB-201 ENST00000335295.4) accessed in Ensembl (release 104) [80].

3.2.2 α-thalassemia

The human α-globin gene cluster (HBAC) in chromosome 16 consists of four functional genes including *HBA2* and *HBA1*, which are the two equal genes encoding the α-globin chain previously mentioned.

To diagnose α-thalassemia the gap-PCR technique was performed to search for the common $-\alpha^{3.7}$ Kb deletion, Figure 3.2(a). This PCR technique is a fast and specific method of detecting previously characterized recurrent deletions, it uses specific primers that only amplify a gene sequence if a deletion joins the flanking gene sequences together [81]. The result of this PCR is then analysed by agarose gel electrophoresis, as exemplified in Figure 3.2(b).



(a) Schematic representation of the breakpoints in genes *HBA2* (α2) and *HBA1* (α1)

(b) Interpretation of gap-PCR results. Where HET is a heterozygous result, HOM a homozygous result and WT a wild-type result.

Figure 3.2: Gap-PCR analysis for diagnosis of $-\alpha^{3.7}$ kb deletion (α-thalassemia), adapted from Faleiro [73].

3.3 Indexes Evaluation

The accuracy of the discriminant formulas mentioned in section 2.6.1 was evaluated with the data from β-thalassemia carrier individuals and IDA patients.

The indexes' formulas were computed in the Python 3.8 programming language [82] and their accuracies were calculated resorting to the libraries `sklearn` [83], `pandas` [84] and `numpy` [85].

These indexes accuracy was calculated in two different ways, using all the data from β-thalassemia carriers and IDA patients, and through the median accuracy of 30 random splits, each with 30% of the individuals data, so that it would be more fairly compared with the machine learning models which were evaluated using the Monte Carlo cross-validation technique mentioned in section 2.4.

3.4 Machine Learning Classification

Regarding the machine learning classification two different types of machine learning classifiers were created: 1) binary to distinguish between β-thalassemia carriers and IDA patients (like the indexes) and 2) multi-class to distinguish between the β-thalassemia carriers, α-thalassemia carriers, IDA patients, and control subjects.

All the machine learning algorithms explained in section 2.2 were used to create the predictive models. Their implementation was computed in the Python 3.8 programming language [82], and the theoretical principles of the algorithms as well as the hyperparameter optimization were implemented with the library `sklearn` [83], with the help of `pandas` [84], `numpy` [85] and `Matplotlib` [86].

The machine learning models were evaluated using the Monte Carlo cross-validation technique resorting to the median of the accuracy during 30 random splits, where 70% of the data was allocated to the training set ($X^{(train)}$) and the remaining 30% of the data to the test set ($X^{(test)}$). In addition, for comparison, the models were also evaluated using all data as training and testing.

3.4.1 Features Selection

To construct the models, multiple different features resulting from the multiplication and division of the original features were generated. To select the best features among them a genetic algorithm adapted from Codes [87] was used. The recombination between the pairs of parents was only single-point crossover, and the genetic algorithm parameter mutation probability was kept at 50%. To keep the number of features used small, for each feature used in the model the fitness function, in this case the accuracy, was penalized with a drop of 0.5%. The generation number and population size were explored in only one machine learning model in order to find an appropriate number of generations and population size.

3.4.2 Hyperparameter Optimization

To solve the hyperparameter optimization problem, the random search solution, explained in section 2.2.1, was applied with successive halving [88]. Successive halving works like a tournament between a set of random combinations of hyperparameters candidates, that is, using a small amount of data all the set of random combinations of the hyperparameters are evaluated. Then the best candidates will continue to be evaluated with different data while the worst are discarded. This process is then repeated until the best candidate is obtained. To implement this solution a distribution over possible parameter values has to be established first. Table 3.2 displays, for each machine learning algorithm used, the hyperparameters tested and their respective range of values.

Table 3.2: Machine learning hyperparameters tested and their range of values.

Model	Hyperparameters	Range of values
Logistic regression	C	[0.1, 0.2, ..., 1.5]
	max_iter	[100, 200, ..., 500]
	solver	['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
	fit_intercept	[True, False]
Support vector machine	C	[0.1, 0.2, ..., 1.5]
	kernel	['linear', 'poly', 'rbf', 'sigmoid']
	degree	[1, 2, ..., 6]
	gamma	['scale', 'auto']
	shrinking	[True, False]
	max_iter	[-1, 100, 200, ..., 500]
	decision_function_shape	['ovo', 'ovr']
Artificial neural network	solver	['lbfgs', 'sgd', 'adam']
	activation	['identity', 'logistic', 'tanh', 'relu']
	hidden_layer_sizes	[(100,), (300, 5), (400, 6), (250, 6), (500, 6), (300,2)]
	max_iter	[100, 200, ..., 1000]
Decision tree	criterion	['gini', 'entropy']
	splitter	['best', 'random']
	max_features	['auto', 'sqrt', 'log2']
Random forest	n_estimators	[50, 100, ..., 300]
	max_features	['auto', 'sqrt', 'log2']
	min_samples_split	[2, 3, ..., 10]
	bootstrap	[True, False]
	max_depth	[50, 51, ..., 100]
	criterion	['gini', 'entropy']
K-nearest neighbors	weights	['uniform', 'distance']
	n_neighbors	[1, 2, ..., 10]
	leaf_size	[5, 10, ..., 50]
	algorithm	['auto', 'ball_tree', 'kd_tree', 'brute']
Naive Bayes	var_smoothing	[1e-11, 1e-10, 1e-9, 1e-8, 1e-7]

The explanation of the meaning of each hyperparameter in this table is available at the `sklearn` website [89].

3.5 Genetic Programming Classification

Genetic programming, other artificial intelligence technique, was also used to distinguish between β-thalassemia carriers and IDA patients (binary classification) and between the β-thalassemia carriers, α-thalassemia carriers, IDA patients, and control subjects (multi-class classification).

The genetic programming algorithm used was M3GP (multidimensional multi-class genetic programming with multidimensional populations) [90], previously implemented in Python by Batista [91].

The M3GP algorithm works by evolving a population of models that maps all the *p*-dimensional features of the training set into new *d*-dimensional features and subsequently calculating the covariance matrix and class centroid, for each of the training data classes. Later, using the Mahalanobis distance, the test data will be classified according to the class whose centroid is closer, in the end the model obtained will be the one with the best fitness (higher accuracy). The biggest advantage of the M3GP algorithm is that there is no need to specify the number of dimensions (*d*), since the algorithm evolves a population of individuals whose dimensions can change during evolution, as there are genetic operators that can add or remove dimensions. This way it is able to progressively search for the optimal dimensions that maximize the classification accuracy.

Just like the machine learning models, the accuracy of M3GP models obtained in the end of the algorithm was calculated using Monte Carlo cross-validation technique resorting to the median of the accuracy during 30 random splits, where 70% of the data was allocated to the training set and the remaining 30% of the data to the test set.

3.6 Outliers Detection

The objective of the outliers detection section was to develop a semi-automatic model able to identify instances that present features different from what would be expected according to the attributed disease, that are atypical, and that in the worst case scenario may actually require a second analysis. The diagnosis of the subjects that make up the data used has already been confirmed by molecular diagnosis. The study of the presence of outliers is not aimed at discarding data, because in this case we would be simplifying the classification task and the performance results would be biased.

So to evaluate the possible presence of outliers in the data set used to train and test all the models, the techniques Cook's distance and silhouette analysis, explained in section 2.5, were applied in Python using the libraries `Yellowbrick` [92] and `sklearn` [83] respectively. Besides these techniques, the instances that were more often misclassified using the best binary and multi-class models were sought and compared with the outliers found to see if there was any overlap.

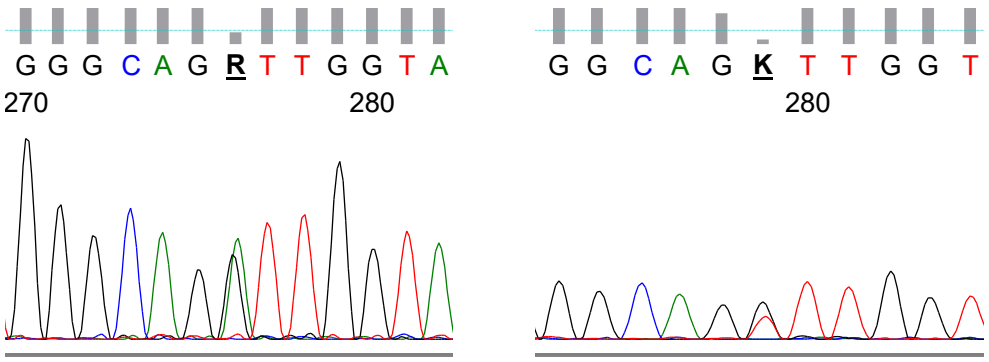
Chapter 4

Results

The following chapter presents the results obtained throughout this thesis. First, the results of the molecular diagnosis performed at the National Institute of Health Doctor Ricardo Jorge (INSA), followed by the performance results of the existing indexes, introduced in the section 2.6.1 and the classifiers created in this thesis. At the end are the results of the outlier detection performed.

4.1 Molecular Diagnosis

As previously described, this thesis also involved laboratory experimental tasks. Although all the datasets used were retrieved and analysed by the "Grupo de I&D em Hemoglobinopatias, Metabolismo do Ferro e Patologias Associadas" of INSA, additionally molecular diagnosis experiments were conducted. The goal was to perform the molecular diagnosis of β-thalassemia and α-thalassemia to gain expertise in the steps required for the analysis, from the DNA sample to the final diagnosis. Two examples of the results obtained after the gene sequencing of the *HBB* gene for the β-thalassemia diagnosis are represented in Figure 4.1.



(a) Case number 77287.

(b) Case number 145457.

Figure 4.1: Example of two Sanger Sequencing results obtained.

Both cases present single nucleotide polymorphisms (SNPs) in heterozygosity at the same DNA positions. In the case number 77287 there is a R in the Sanger sequencing result, in the International

Union of Pure and Applied Chemistry (IUPAC) nucleotide code it means that in that specific position the subject has different alleles for certain gene (a heterozygosity). In one allele he has an adenine in that position and in the other a guanine. A healthy individual only presents guanine, therefore he has guanine $>$ adenine substitution located at the first nucleotide of intron 1, which is a known pathogenic mutation. Consequently, this individual is a β-thalassemia carrier. The case number 145457 is similar but, instead of having a R in Sanger sequencing result he has K, which stands for guanine and thymine, so the subject instead of having only guanine in both alleles, has a guanine $>$ thymine substitution also located at the first nucleotide of intron 1, which is also a known pathogenic mutation. Thus, this individual is also a β-thalassemia carrier. Mutations at this specific position severely affect gene splicing and the synthesis of the β-globin chain (β^0). As both individuals are heterezygotes, the other allele has no mutation (β), they both have thalassemia minor (β/β^0).

In Table 4.1 are summarized all the different mutations found in the β-thalassemia cases. These results exemplify that there are several different mutations that can lead to β-thalassemia, having in turn different impacts on the gene's phenotype as some mutations lead to the partial or complete elimination of the β-globin chain synthesis (β^+ and β^0, respectively).

Table 4.1: β-thalassemia diagnosis result.

Case number	Mutation		
	HGVS name	Common name	Allele Phenotype
77287	HBB:c.92+1 G $>$ A	IVS I-1 (G $>$ A)	β^0
145457	HBB:c.92+1 G $>$ T	IVS I-1 (G $>$ T)	β^0
040997	HBB:c.93-21 G $>$ A	IVS I-110 (G $>$ A)	β^+
961401	HBB:c.92+6 T $>$ C	IVS I-6 (T $>$ C)	β^+
950756	HBB:c.118 C $>$ T	CD 39 (CAG $>$ TAG)	β^0

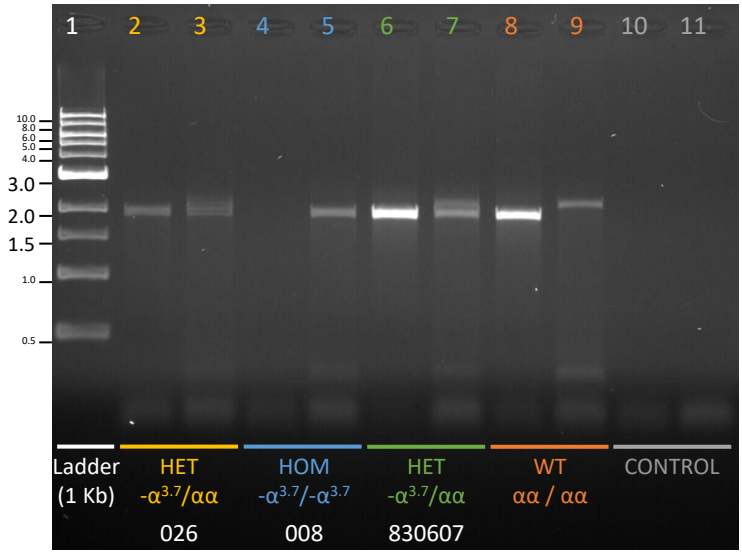


Figure 4.2: Photo of an agarose gel electrophoresis of DNA fragments obtained by gap-PCR. The analysis of fragments obtained in lanes 2&3, 4&5, and 6&7 allow an α-thalassemia diagnosis, where HET is a heterozygous diagnosis and HOM a homozygous diagnosis. In lanes 8&9 is a normal control, where WT is a wild-type diagnosis, and in lanes 10&11 is the negative control.

With regards to the α-thalassemia diagnosis, the results and interpretation of gap-PCR are presented in Figure 4.2. These results reveal different severities of the disease as some individuals present deletions in homozygosity and others in heterozygosity.

As a consequence of the different mutations that can lead to β-thalassemia and α-thalassemia it is expected that subjects with the same thalassemia diagnosis present slightly dissimilar CBC test results, which will complicate the diagnosis as the pattern recognition task is more demanding.

4.2 Data Description

A descriptive analysis of the data was performed using boxplots with all the normalized features per class and by calculating the mean, median, minimum and maximum of each feature for each class, see Figure 4.3 and Table 4.2, respectively.

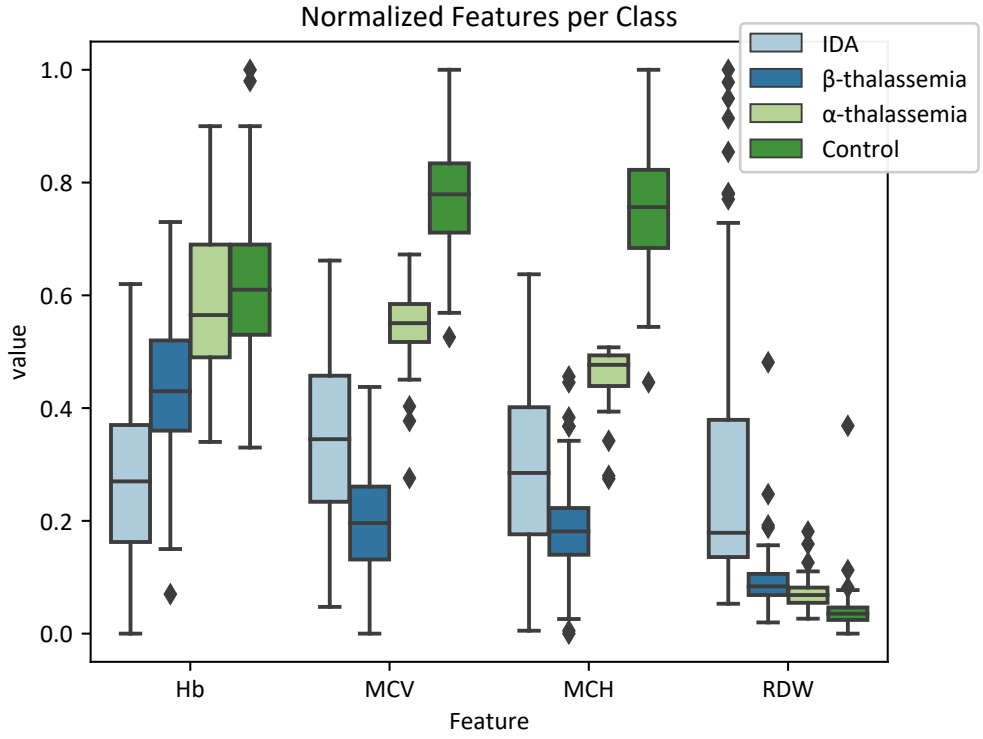


Figure 4.3: Boxplots with all the normalized features per class. Where Hb is the hemoglobin concentration, MCV the mean corpuscular volume of the red blood cells, MCH the mean corpuscular hemoglobin, and RDW the red blood cell distribution width.

Both in Figure 4.3 and Table 4.2 it is possible to observe that there are several differences in the median of some features between the different classes.

The hemoglobin (Hb) is lower in individuals with microcytic anaemia than in the control group (healthy subjects), which is expected. The microcytic anaemias are associated with a decreased production of hemoglobin. In the case of IDA, the lack of hemoglobin arises from the shortage of iron, in thalassemias it

results from the defective hemoglobin production [2]. However, in the case of α-thalassemia, hemoglobin is higher than in other microcytic anaemias, probably because most of the α-thalassemia data used are from individuals with heterozygous mutations, suggesting that these individuals must have the α-thalassemia trait 1, or at worst 2, and therefore still have some α-globin synthesis.

The decreased production of hemoglobin also affects the MCV (mean corpuscular volume) value, because hemoglobin is a major constituent of the red bold cells, which explains why the individuals with microcytic anaemia have a lower MCV than the control group.

In respect to the MCH (mean corpuscular hemoglobin), it is much lower in individuals with microcytic anaemia, which is expected because it is the average amount of hemoglobin in each red blood cell, so it is also dependent on the production of hemoglobin.

The RDW (red blood cell distribution width), represents the coefficient of variation of the red blood cell volume distribution. In the β-thalassemia and α-thalassemia carriers practically all red blood cells are microcytic due to mutations in the globin chain genes and therefore the RDW is low. However, the IDA subjects have the highest RDW. This has been explained in previous studies by the administration of iron therapy in patients with IDA, which results in a rise in the RDW few days after the initiation of the iron therapy and during the next moth [93], which also probably explains why IDA patients present the highest variation in all the features.

Table 4.2: Dataset features' properties.

Class		Features			
		Hb	MCV	MCH	RDW
β-thalassemia	mean	11.8 ± 1.2	65.0 ± 4.3	20.8 ± 1.5	15.0 ± 2.1
	median	11.8	64.8	20.6	14.7
	minimum	8.2	55.7	17.1	11.8
	maximum	14.8	76.0	25.9	32.7
α-thalassemia	mean	13.5 ± 1.4	81.1 ± 3.2	25.9 ± 1.0	14.1 ± 1.3
	median	13.1	81.2	26.3	14.0
	minimum	10.9	68.5	22.4	12.1
	maximum	16.5	86.9	26.9	19.1
IDA	mean	10.1 ± 1.4	71.5 ± 6.8	22.7 ± 2.9	24.8 ± 12.3
	median	10.2	71.7	22.6	19.0
	minimum	7.5	57.9	17.2	13.3
	maximum	13.7	86.4	29.4	56.2
Control	mean	13.7 ± 1.2	91.4 ± 4.5	31.6 ± 1.9	12.7 ± 1.5
	median	13.6	91.8	31.7	12.5
	minimum	10.8	80.1	25.7	10.9
	maximum	17.5	102.1	36.4	27.6

Hb is the hemoglobin concentration, MCV the mean corpuscular volume of the red blood cells, MCH the mean corpuscular hemoglobin, and RDW the red blood cell distribution width.

As the β-thalassemia carriers have different allele phenotypes according to the mutation that they have, boxplots that separates the β^0 mutation carriers from the β^+ was also constructed (Figure 4.4), as well as a table with the p-values, obtained in a T-test, of the different features of the β-thalassemia carriers (Table 4.3).

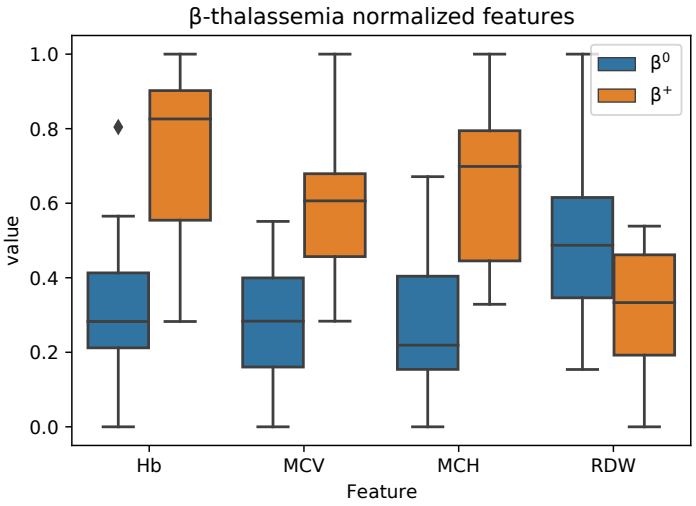


Figure 4.4: β-thalassemia normalized features.

Table 4.3: β-thalassemia p-values of the T-test.

Feature	P-value		
	Both sexes	Female	Male
Hb	0.002	0.068	0.002
MCV	0.001	0.091	0.012
MCH	0.001	0.046	0.020
RDW	0.059	0.513	0.109

The null hypothesis is that the mean values of the features do not differ between the β^0 and the β^+ mutation groups.

According to both Figure 4.4 and Table 4.3, there is a significant difference in almost all features of the β-thalassemia carriers depending on the mutation they have, the only feature who's p-value is not statistically significant is the RDW. However, when we look to the p-values of each sex the values are very different, while in the females only the MCH is statistically significant, in the males there is only one that is not, the RDW. However, it should be noted that among all cases of β-thalassemia we only have information regarding the mutation of 21 individuals from which 11 are female and 10 male. Therefore to think that the type of mutation that a β-thalassemia carrier has is only significant in the male sex can be hasty and therefore it would be advisable to have more data in order to draw this conclusion.

4.3 Indexes Evaluation

As introduced in the section 2.6.1 there are already different discriminant formulas based on the CBC test results to distinguish β-thalassemia carriers from IDA patients. With the aim of understanding which formula to distinguish β-thalassemia carriers from IDA presents the higher accuracy in the Portuguese population the reviewed indexes were tested. These indexes accuracy was calculated in two different ways, using all the data from β-thalassemia carriers and IDA patients and through the median accuracy of 30 random splits, each with 30% of the individuals data, so that it would be more fairly compared with the machine learning models which were evaluated using the Monte Carlo cross-validation technique mentioned in section 2.4. The resulting values are represented in Table 4.4 and 4.5, in descending order.

Table 4.4: Indexes performance with 30 random splits of the data.

Index	Median accuracy %
RDWI	95.4
Green and King (G&K)	92.3
Ehsani	84.6
Ricerca	83.1
Sirdah	79.2
England and Fraser	76.9
Srivastava	75.4
Telmissani–MDHL	75.4
Shine and Lal	73.8
Mentzer	66.2
Matos and Carvalho	66.2
CRUISE	66.2
Telmissani–MCHD	63.1
Bessman	47.7

The indexes performance was calculated through the median accuracy of 30 random splits, each with 30% of the individuals data, so that it would be more fairly compared with the machine learning models which were evaluated using the Monte Carlo cross-validation technique.

Table 4.5: Indexes performance with all the data.

Index	Accuracy %
RDWI	94.9
Green and King (G&K)	91.8
Ricerca	83.7
Ehsani	83.2
Sirdah	78.6
England and Fraser	76.0
Srivastava	74.5
Telmissani–MDHL	74.5
Shine and Lal	74.0
Mentzer	67.3
Matos and Carvalho	67.3
CRUISE	67.3
Telmissani–MCHD	63.3
Bessman	48.0

These tables show clearly that the RDWI and G&K are the most reliable indexes in the studied population, just like what was observed in the Brazilian [3] and Palestinian [5] populations. Both indexes have in common the inclusion in their formulas of the MCV, RDW and Hb (since, RBC $=$ Hb$/$MCH $\times$ 10). This suggests that these three features may be very relevant in the discrimination between β-thalassemia carriers and IDA patients in the Portuguese population.

4.4 Machine Learning Classification

This section presents the results of the machine learning classification. It begins with the application of a feature selection algorithm to select which features to use in the models, followed by a hyperparameter optimization of these same models. At the end, the performance results of all the models are compared.

4.4.1 Models Improvement and Evaluation

Two different types of machine learning classifiers were created: 1) binary to distinguish between β-thalassemia carriers and IDA patients (like the indexes) and 2) multi-class to distinguish between the β-thalassemia carriers, α-thalassemia carriers, IDA patients, and control subjects.

First, all binary and multi-class models were trained using the features: sex, Hb, MCV, MCH and RDW. Then, in an attempt to optimize these models, several different features resulting from the multiplication and division of the mentioned ones were generated and a genetic algorithm was used to select those that could achieve greater accuracy with each model.

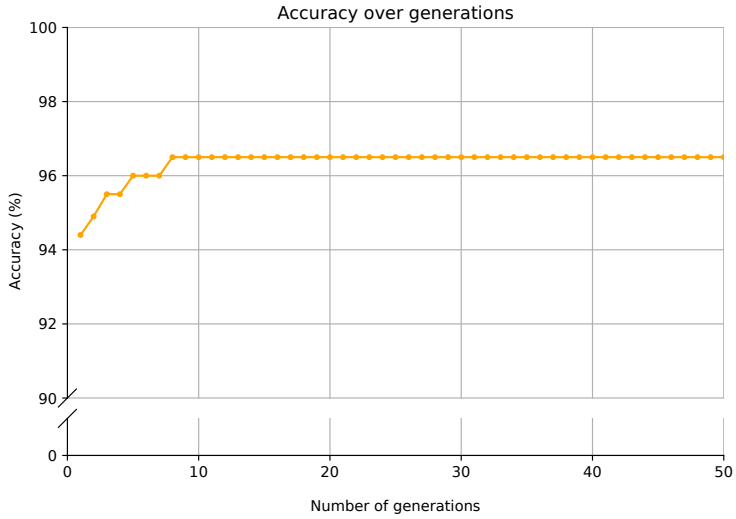
Figure 4.5(a) shows how the number of generations in the genetic algorithm influenced the final accuracy of a binary logistic regression model. It seems that after 10 generations the accuracy no longer improves. For this reason, the genetic algorithm used with all models to select the features was set to 20 generations, as it gives a good margin for improvement. The same rationale was used to select the initial population size of the genetic algorithm, Figure 4.5(b). The initial population size chosen was 50, since a huge initial population does not seem to translate into greater final accuracy, and so 50 was considered to be a good enough size to run the evolution.

After selecting the features that could possibly achieve greater accuracy in each model, the random search solution technique with successive halving was used to solve the hyperparameter optimization problem and thus allow the models to further improve their accuracy. In addition, the random search solution was also used in the models with the initial features (sex, Hb, MCV, MCH and RDW), as it could also improve their accuracy.

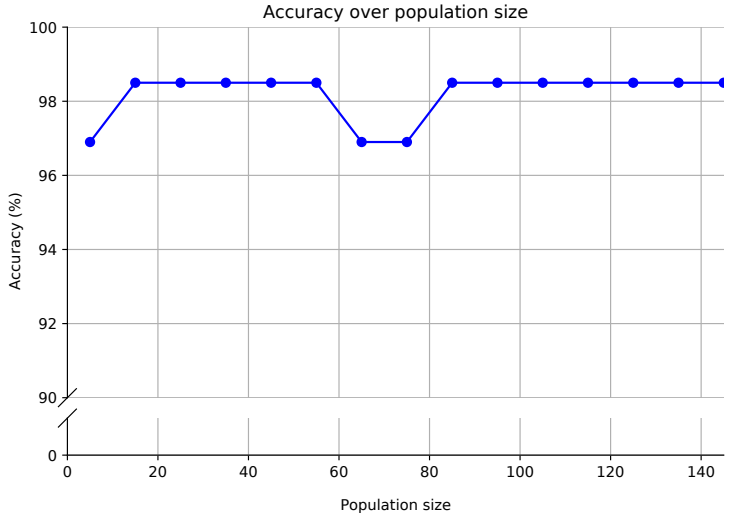
The accuracy of the binary models is summarized in Table 4.6. This accuracy was calculated using the median accuracy over 30 random divisions of the data, where 70% of the data were allocated to the training set and the remaining 30% to the test.

In the binary classifiers, the highest accuracy achieved was 95.4%. This performance was reached with two different algorithms, random forest and artificial neural network. Both algorithms were using the features created and later selected by the genetic algorithm. These new features contributed significantly to the improvement of the accuracy of most classifiers, almost all the best classifiers of each algorithm are the ones using the new features.

There can be several reasons why this optimization through feature creation and selection did not work in all models. The genetic algorithm used to select the features evaluated the set of features chosen with only on split of the data, which may have introduced a bias, since some splits of the data can result in a test set easier to classify and therefore that group of features is wrongly associated with a high



(a) Accuracy over number of generations.



(b) Accuracy over population size.

Figure 4.5: Accuracy of the binary logistic regression model in the genetic algorithm, over the number of generations and the population size.

Table 4.6: Median accuracy of the machine learning binary classifiers

Model	Median accuracy (%)			
	Without hyperparameter optimization		With hyperparameter optimization	
	Initial features	Selected features	Initial features	Selected features
Random forest (RF)	92.3	**95.4**	90.8	93.8
Artificial neural network (ANN)	93.8	93.8	94.6	**95.4**
Logistic regression (LR)	93.8	**94.6**	93.1	93.8
Decision tree (DT)	90.8	93.1	90.8	**93.8**
Support vector machine (SVM)	80.0	75.4	**93.8**	93.8
Naive Bayes (NB)	90.8	**92.3**	90.8	88.5

accuracy. In Table 4.6 the models were evaluated resorting to a median accuracy of 30 different splits of the data and therefore that bias is mitigated. That said some of the features selected were not the best, as the original features actually presented a higher accuracy. This also leads us to think that maybe if these initial features were forced to be in the initial population of the genetic algorithm, they could have been selected. However this genetic algorithm initial population was seeded completely randomly, which may also justify why this optimization did not work in all models. Even if the initial features were in the initial population or had appear through mutations the genetic algorithm may have discarded them due to the penalization on the number of features.

The hyperparameter optimization, despite having improved the performance of some algorithms, did not have as notable an impact as the new features, suggesting that the default hyperparameters were already adequate in most classifiers. Besides this, this optimization just like the feature selection was done resorting to only one split of the data which may also have introduced the bias previously explained.

It is curious to note that accuracy value obtained, 95.4%, was also the highest value achieved by the existing indexes, which suggests that there is some difficulty in surpassing this value.

To assess the variation in accuracy across the different divisions of the data, boxplots were constructed with the best binary models, Figure 4.6.

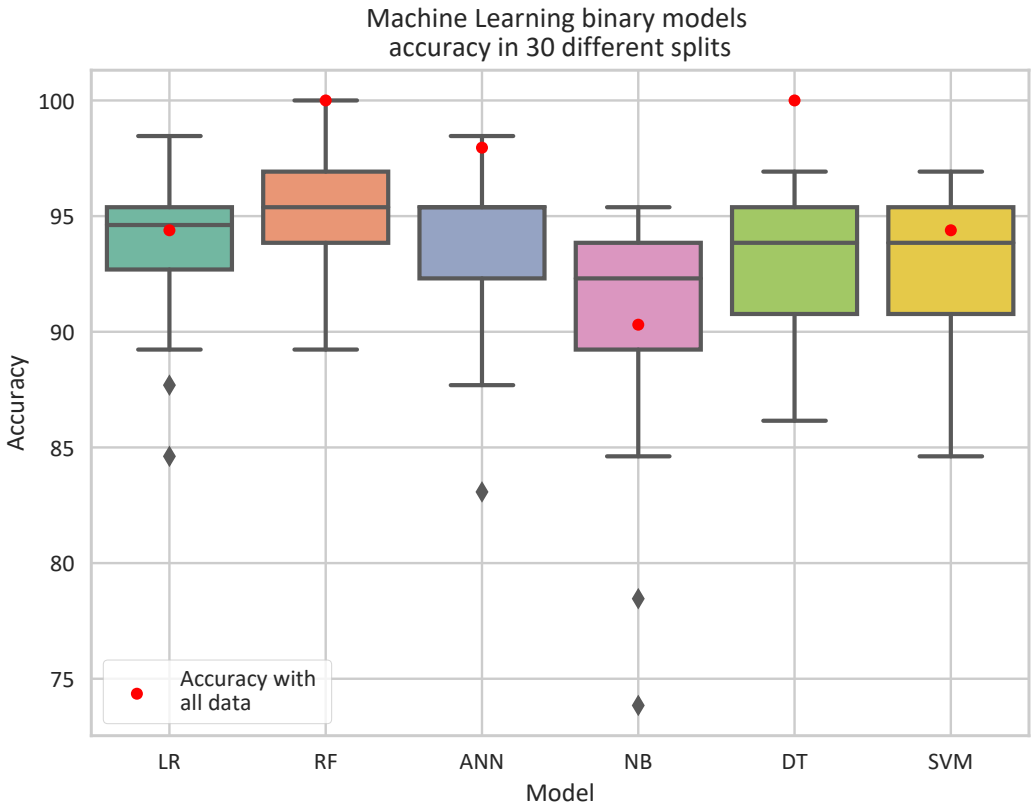


Figure 4.6: Machine learning binary models accuracy with 30 data random splits and the accuracy when training and testing the models with all data.

The random forest algorithm not only had the highest median accuracy, but also showed a very little performance variation in classifying microcytic anaemia across different data splits. In contrast, the naive Bayes algorithm obtained the lowest median accuracy and a much larger variation.

This boxplot also shows the accuracy when training and testing the algorithms with all data. Even though in the random forest and in the decision tree it led to an accuracy of 100%, it does not represent the true accuracy of a classifier built with these algorithms, as the accuracy rarely reached this value, in the case of the random forest, and never, with the decision tree, when the models were trained and tested with different data. This 100% accuracy when training and testing with all data, however, indicates a greater tendency to overfit the data than the remaining algorithms. The logistic regression and support vector machine, for example, had very similar accuracy when using all the data and when splitting the data and therefore appear to be more proof against overfitting. In the case of the decision tree algorithm this overfitting can be explained by the fact that the maximum depth of the tree was not defined and therefore, the nodes of the tree were expanded until all leaves were pure or until all leaves contain less than the minimum number of samples required to split an internal node, which by default was established as two.

In Table 4.7 it is summarized the accuracy of the microcytic anaemia multi-class models.

Table 4.7: Median accuracy of the machine learning multi-class classifiers.

Model	Median accuracy			
	Without hyperparameter optimization		With hyperparameter optimization	
	Initial features	Selected features	Initial features	Selected features
Random forest (RF)	92.6	92.2	**93.0**	91.5
k-nearest neighbors (KNN)	**92.6**	92.2	91.5	91.5
Artificial neural network (ANN)	89.1	89.9	**92.2**	89.9
Naive Bayes (NB)	89.1	**91.5**	89.1	91.5
Decision tree (DT)	89.9	89.9	**90.7**	90.7

The highest accuracy achieved with multi-class classifiers was 93.0%, with the random forest algorithm. This accuracy is not as good as that reached in the binary classifiers but considering that this classification is between four classes it is a more difficult task. In multi-class classification, the best classifiers for the different algorithms were almost never the ones that used the selected features, although they improved the accuracy of some classifiers, in multi-class classification the hyperparameters optimization was more significant in improving its accuracy. As explained before, the feature selection may not have worked due to the bias in the test set, the lack of the initial features on the population that seeded the genetic algorithm or due to the penalization on the number of features.

The Table 4.8 presents the accuracy of the best microcytic anaemia multi-class classifiers, with the median accuracy overall the classes and the median accuracy per class. The control class had the highest accuracy regardless of the algorithm used to create the model, on the other hand, patients with IDA almost always had the worst accuracy, which means that with the data used, this class is the most difficult to be identified.

Table 4.8: Median accuracy per class of the best machine learning multi-class classifiers.

Model	Median accuracy (%)				
	IDA	α-thalassemia	β-thalassemia	Control	Overall %
Random forest (RF)	93.4	96.9	95.3	98.4	93.0
k-nearest neighbors (KNN)	93.8	96.9	95.3	99.2	92.6
Artificial neural network (ANN)	95.3	96.1	95.3	96.9	92.2
Naive Bayes (NB)	93.4	96.9	94.6	99.2	91.5
Decision tree (DT)	91.5	96.1	95.0	98.1	90.7

As for the binary classifiers, boxplots with the accuracy across the different divisions of the data were also constructed with the best multi-class models, Figure 4.7.

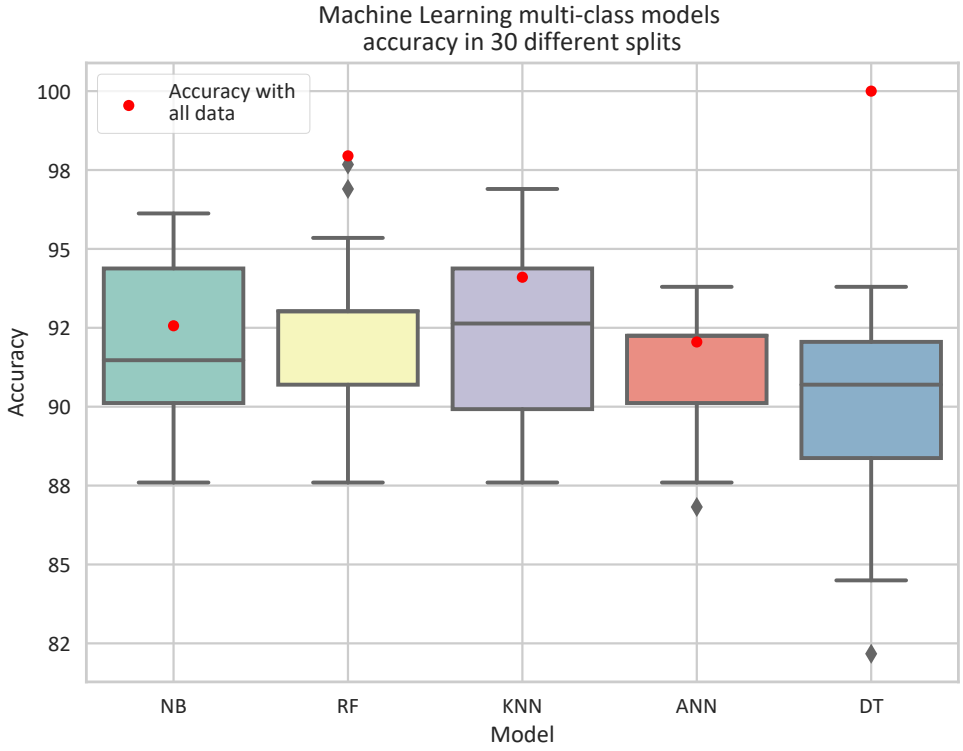


Figure 4.7: Machine learning multi-class models accuracy with 30 data random splits and the accuracy when training and testing the models with all data.

Within the multi-class models, random forest is not only the one with the highest median accuracy, but it also continues to show a very low variation in accuracy between the different divisions of the data, as what was observed with the binary classifiers, which means that it is the most stable method. However, random forest and decision tree are again the classifiers most likely to suffer from overfitting, their accuracy reaches significantly higher values when using all the data, never reaching these values when the data is split in the training and testing set.

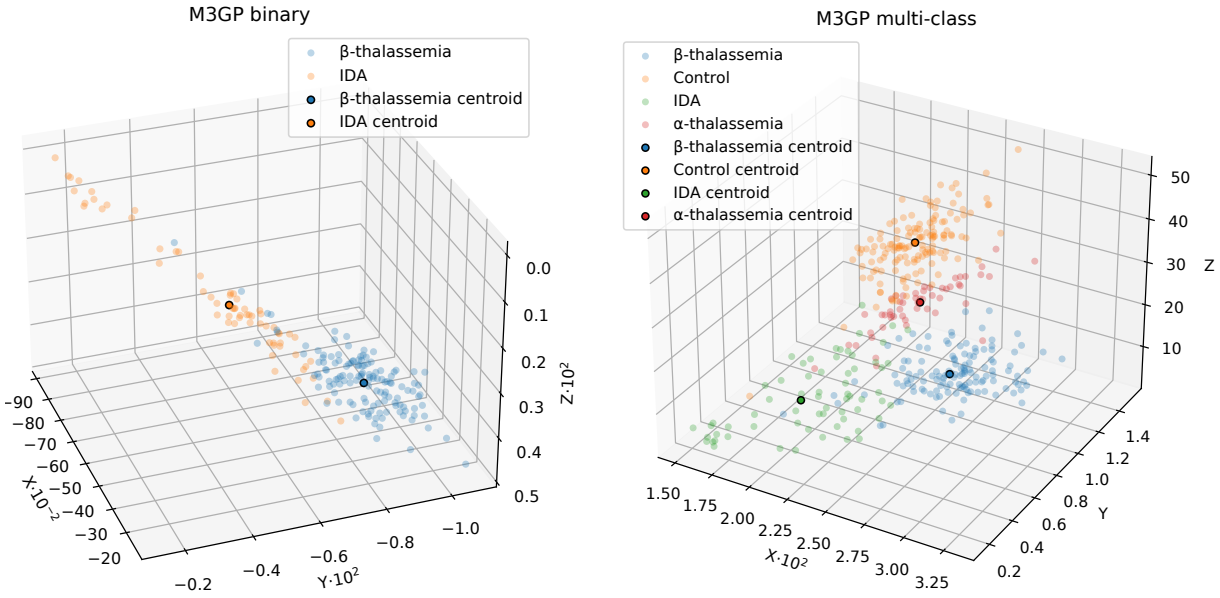
Overall, we can conclude that the algorithm that reaches the best accuracy with little variation, both in the binary classification and in the multi-class classification of microcytic anaemia, is the random forest.

Yet, as this algorithm tends to overfit the data, measures must be taken to avoid overfitting the model like for example, the cross-validation technique. This technique was used to select hyperparameters that do not cause overfitting and thus allow the correct classification of new instances by the classifier obtained with the trained algorithm.

4.5 Genetic Programming Classification

Just like the machine learning algorithms, the genetic programming algorithm M3GP was used to create binary and multi-class classifiers.

In Figure 4.8 are two graphical representations of the data separation in the feature space of classifiers obtained with the M3GP algorithm. In Figure 4.8(a) a binary classifier with 3 dimensions, that is, using 3 new features, and in Figure 4.8(b) a multi-class classifier, which to simplify the graphical representation was forced to have only 3 dimensions.



(a) Data separation of a M3GP binary classifier.

(b) Data separation of a M3GP multi-class classifier.

Figure 4.8: Graphical representation of the data separation in feature space of the M3GP classifiers.

As we can see, there seems to be a good separation between all the classes, however, when evaluating the accuracy of the classifiers obtained with the M3GP algorithm in the same way as the machine learning classifiers, Table 4.9 (binary and multi-class), the accuracy was not better than that of these classifiers.

Table 4.9: Median accuracy of the M3GP classifiers.

Classifier	Median accuracy %				
	IDA	β-thalassemia	α-thalassemia	Control	Overall
Binary	93.8	93.8	-	-	93.8
Multi-class	95.0	95.3	96.9	98.4	92.2

4.6 Oultiers Detection

Considering that the best median accuracy achieved with both binary classifiers and indexes was not able to surpass the value of 95.4% and the best multi-class median accuracy achieved is 93.0%, the suspicion of the presence of outliers arose. For this reason, the consistency within classes as well as the presence of outliers were evaluated.

To assess consistency within classes, a silhouette analysis was performed, Figure 4.9.

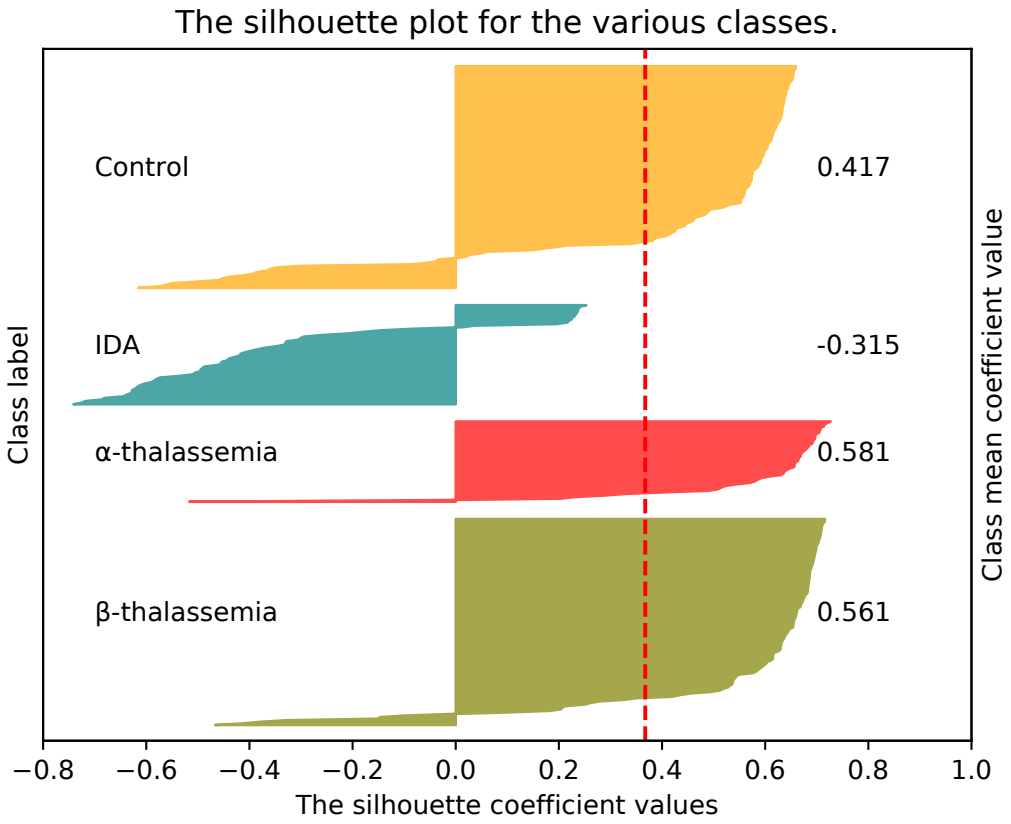


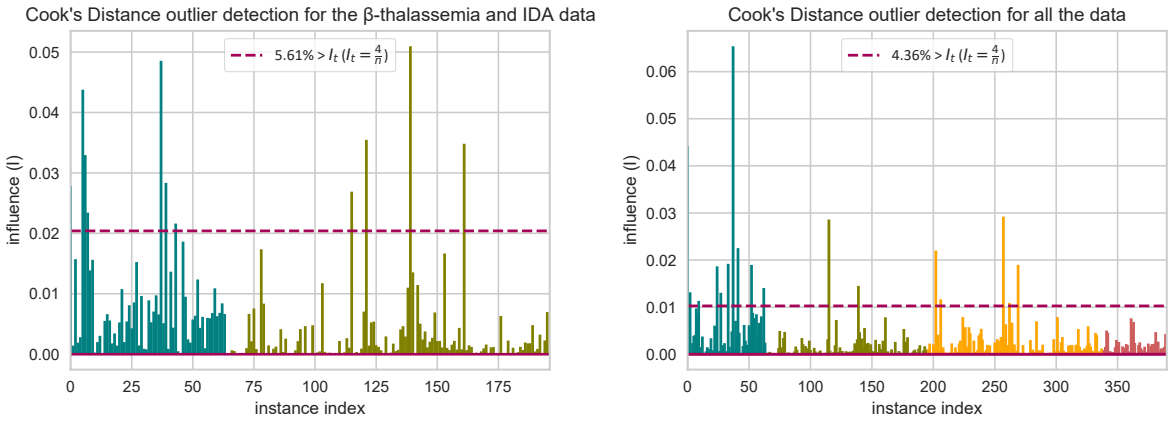
Figure 4.9: Silhouette analysis and respective class mean.

As reported by the silhouette plot, α-thalassemia is the class with the highest mean silhouette coefficient. However, it is important to emphasize that it is the class with the lowest number of instances, which will make it difficult to train the models and probably for this reason it is not the class that registered the best accuracy in any of the models. The control class had the second lowest average silhouette coefficient, yet registered the best accuracy in the models, which demonstrates the importance of the number of instances in the performance of the algorithms since it is the class that has more instances.

With negative mean silhouette coefficient, IDA is the class with the lowest consistency, probably because no information about the iron status was used and also, due to the administration of iron therapy in some patients with IDA. However, this low consistency did not compromise the differential diagnosis between IDA and β-thalassemia in the binary classifiers, most likely because the β-thalassemia class has a very high class consistency, yet, it had a greater impact on the performance of multi-class models.

Owing to the fact that the average silhouette coefficient is 0.37, we can say that the consistency of the classes is, in general, good.

To evaluate the presence of outliers the Cook's distance was calculated for all the instances used in the binary and multi-class models, Figures 4.10(a) and 4.10(b), respectively. According to the Cook's distance, the percentage of outliers found was 5.61% in the data used in binary models and 4.36% within all the data. This may explain why the accuracy could not surpass the value of 95.4% in the binary models and 93.0% in the multi-class. In both cases the class where most outliers were found was IDA, which is consistent with the results obtained with the silhouette analysis. In opposition, the control class does not have any outliers which certainly helped this class to be the class that always presented the best accuracy.



(a) Cook's distance outlier detection for the β-thalassemia and IDA data.

(b) Cook's distance outlier detection for all data.

Figure 4.10: Cook's distance outlier detection. IDA (blue), β-thalassemia (green), Control (yellow) and α-thalassemia (red).

Afterwards in order to understand if these outliers found by the Cook's distance were the instances that were being misclassified in the models, the models were trained and tested with all the instances, that is, without cross-validation, and the instances that were most frequently misclassified with the best models of each algorithm were identified. In addition, the classes in which these instances were placed were also identified, both with the created models and with the already existing indexes.

In Table 4.10 are the most frequently misclassified instances in the binary models. The instances that are most often misclassified are not only from IDA patients, these instances belong to both IDA patients and β-thalassemia carriers. However, this does not mean that both classes are equally hard to identify because the β-thalassemia carriers in the data used are more than double of the IDA patients. For this reason, if both classes were equally difficult to classify, β-thalassemia cases should be more than double in this table. Since there are less than those with IDA, it can be concluded that the IDA class is much more difficult to be identified by the created models. It is also interesting to note that, in the case of the most frequently misclassified instances of β-thalassemia carriers, the models created are very divided on the class they assign to these individuals, while the existing indices vote mainly on the IDA class. A possible justification is that due to the fact that IDA is more common than β-thalassemia, these indexes consider right from the start that the individual is more likely to have IDA.

Table 4.10: Most frequently misclassified instances in binary classification.

Instance	Best machine learning models			Best indexes		Diagnosis
	Times misclassified	IDA %	β-thalassemia %	IDA %	β-thalassemia %	
29	4	33	67	75	25	IDA
34	3	50	50	75	25	IDA
36	3	50	50	25	75	IDA
50	3	50	50	50	50	IDA
51	3	50	50	50	50	IDA
59	3	50	50	75	25	IDA
115	3	50	50	100	0	β-thalassemia
121	4	66	34	75	25	β-thalassemia
139	3	50	50	75	25	β-thalassemia
153	2	33	67	75	25	β-thalassemia
189	2	33	67	75	25	β-thalassemia

The machine learning models used were the best for each type and the best indexes were the best 4.

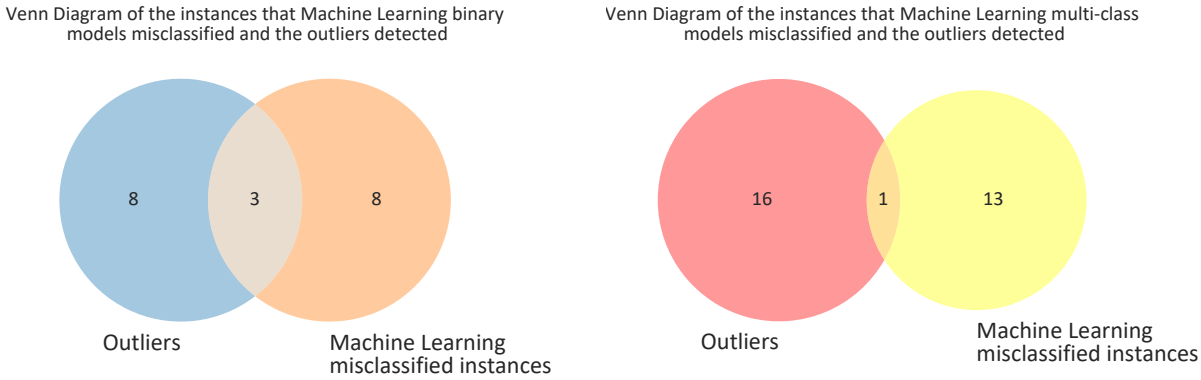
Regarding the most incorrectly classified instances in the multi-class models, Table 4.11, the class that is less frequently incorrectly classified is the control class, being the easiest class to classify, probably due to the fact that it is the class with the most instances and has a good consistency. On the other hand, IDA and α-thalassemia seem to be the most difficult classes to classify, which can be justified by the scarcity of instances of these classes, since both have less than half of the instances of β-thalassemia and control, and in the case of the IDA class also by the lack of information about the iron status of the body that can have led to the low consistency of the IDA class. Even so, one would expect fewer instances of the α-thalassemia class compared to IDA since the α-thalassemia class did not show any outliers and is the class with the highest consistency.

Table 4.11: Most frequently misclassified instances in multi-class classification.

Instance	Best machine learning models					Diagnosis
	Times misclassified	IDA %	β-thalassemia %	α-thalassemia %	Control %	
10	4	20	0	0	80	IDA
21	3	40	0	40	20	IDA
29	3	40	60	0	0	IDA
115	3	60	40	0	0	β-thalassemia
159	3	40	40	20	0	β-thalassemia
170	3	60	40	0	0	β-thalassemia
189	3	60	40	0	0	β-thalassemia
195	3	60	40	0	0	β-thalassemia
333	4	80	0	0	20	Control
338	3	40	20	40	0	α-thalassemia
353	4	20	60	20	0	α-thalassemia
357	4	80	0	20	0	α-thalassemia
363	4	0	80	20	0	α-thalassemia
364	3	60	0	40	0	α-thalassemia

The machine learning models used were the best for each type.

To understand if there was any overlap between the outliers found by Cook's distance and the most frequently misclassified instances Venn diagrams were made, Figure 4.11.



(a) Venn diagrams of the machine learning binary models most misclassified instances and the outliers detected in the β-thalassemia and IDA data.

(b) Venn diagrams of the machine learning multi-class models most misclassified instances and the outliers detected in all data.

Figure 4.11: Venn diagrams of the machine learning models most misclassified instances and the outliers detected.

In the binary classification three of the instances that were considered outliers, instances 115, 121 and 139, they are also among the instances that are most frequently misclassified by binary models. In the case of multi-class classification, only instance 115 was considered outlier, while it is among the most frequently misclassified instances by multi-class models. These three instances, 115, 121 and 139, are all from individuals with β-thalassemia trait.

To understand if there is in fact anything strange about these instances, in Table 4.12 are the features' values from these patients, the mean values of the features in the different classes and the mean values of a group of five individuals with both β-thalassemia trait and IDA. The data from the individuals with both β-thalassemia trait and IDA were also made available by Bárbara Faleiro [73] and Daniela Santos during their masters dissertations research, but due to the small number of instances (only five observations) with both conditions it was not included in the multi-class classification.

Table 4.12: Mean feature values per class and some instances' values.

		Features			
		Hb	MCV	MCH	RDW
Instances	115	12.4	69.9	22.9	**32.7**
	121	11.6	74.6	25.7	15.2
	139	8.2	68.1	20.6	15.9
Class mean	β-thalassemia	11.8 ± 1.2	65.0 ± 4.3	20.8 ± 1.5	15.0 ± 2.1
	IDA	10.1 ± 1.4	71.5 ± 6.8	22.7 ± 2.9	24.8 ± 12.3
	Control	13.7 ± 1.2	91.4 ± 4.5	31.6 ± 1.9	12.7 ± 1.5
	α-thalassemia	13.5 ± 1.4	81.1 ± 3.2	25.9 ± 1.0	14.1 ± 1.3
	β-thalassemia and IDA	12.3 ± 0.9	60.3 ± 1.5	19.9 ± 0.5	$\mathbf{34.6} \pm 0.8$

Regarding instances 121 and 139, some values are a little farther from the β-thalassemia average, such as the MCV in instance 121 and the Hb in instance 139. Although nothing extraordinary, due to the high sensitivity of the algorithms it seems to have been enough to lead them to misclassify these instances. In the case of instance 115, the situation is different, as the RDW is more than double the average obtained in the β-thalassemia class. A similar value was only obtained in individuals who had both β-thalassemia trait and IDA, which leads to the suspicion that this subject was misdiagnosed and that, in fact, instead of just having β-thalassemia trait, he has both β-thalassemia trait and IDA.

In general, due to the percentage of outliers detected through Cook's distance, we can admit that in fact some instances may have values a little more distant than what was expected for their class and therefore it was not possible to achieve greater accuracy. However, it is also important to bear in mind that the absence of data regarding iron status of the body seems to affect the performance of these classifiers in identifying patients with IDA and that effectively the more data there is to train the classifiers, the better their performance will be.

Chapter 5

Conclusions

5.1 Achievements

Anaemia is a disease that affects millions of people across the world and Portugal is not an exception. Regarding the microcytic anaemias, its differential diagnosis is important to provide the right treatment. This way iron overload can be avoided, and genetic counselling can be provided when appropriate. However, the most reliable methods to diagnose thalassemias and IDA are expensive and time-consuming. Therefore, indexes able to discriminate between β-thalassemia carriers and IDA patients have been created, in order to make the diagnosis faster and more accessible. These indexes however, have not revealed to be 100% accurate. The results obtained in this thesis indicate that the RDWI and G&K are the most reliable indexes for the Portuguese population, with a median accuracy of 95.4% and 92.3%, respectively.

Through molecular diagnosis it was possible to verify that multiple mutations of different severity can lead to β-thalassemia or α-thalassemia, which gives them variability in the CBC tests, making the screening by indexes more difficult. In order to pursuit for a higher accuracy, the principal objective of this thesis was to test different machine learning algorithms with data from the Portuguese population, as they could allow the creation of a new classifier specifically suited for this population. Beyond that, multi-class classification was also explored, to enable that, with just one multi-class classifier, it is possible to distinguish between β-thalassemia carriers, α-thalassemia carriers, IDA patients and healthy subjects, based on a CBC test alone.

Among all the algorithms tested in this thesis it was possible to conclude that the algorithm that reaches the best accuracy with little variation, both in the binary and in the multi-class classification of microcytic anaemia, is the random forest. This algorithm was able to achieve 95.4% accuracy in the binary classification, to distinguish between β-thalassemia carriers and IDA patients like the indexes, and 93.0% in the multi-class classification. However, as this algorithm tends to overfit the data, measures must be taken to avoid overfitting the model, which, in this case, was cross-validation.

In an attempt to make the classifiers even more accurate, the M3GP genetic programming algorithm was used to create binary and multi-class classifiers. However, its accuracy was below those obtained

with the models created by the random forest algorithm, having reached only a median accuracy of 93.8% in the binary classification and 92.2% in the multi-class classification.

Regardless of the algorithm used to create the model, in the multi-class classification the control class was found to be the class with the highest median accuracy. On the other hand, patients with IDA almost always had the worst median accuracy, which means that with the data used this class is the most difficult to be identified, probably due to the fact that information regarding the iron state of the body is not being incorporated in the classifiers. The low amount of data of patients with IDA also seems to have impaired the performance of their classification, with the same being observed in the classification of α-thalassemia, these classes have less than half the amount of data of the others and are the ones that present the smallest accuracy.

Considering that the best median accuracy achieved with both binary classifiers and indexes was not able to surpass the value of 95.4% and the best multi-class median accuracy achieved is 93.0%, the suspicion of the presence of outliers arose, and for this reason, the consistency within classes was evaluated, as well as the presence of outliers. Owing to the fact that the average silhouette coefficient is 0.37, we can say that the consistency of the classes is good in general. However, with negative mean silhouette coefficient, IDA is the class with the lowest consistency, reinforcing the need to use data regarding the body's iron status to identify this disease. This low consistency did not compromise the differential diagnosis between IDA and β-thalassemia trait in the binary classifiers, most likely because the β-thalassemia class has a very high class consistency, however, it had a greater impact on the performance of multi-class models.

To evaluate the presence of outliers, the Cook's distance was calculated for all the instances used in the binary and multi-class models. The percentage of outliers found was 5.61% in the data used in binary models (data from the classes: β-thalassemia and IDA) and 4.36% within all the data (data from all the four classes). Due to the percentage of outliers detected through the Cook's distance, we can admit that in fact some instances may have values a little more distant than what was expected for their class and therefore it was not possible to achieve greater accuracy.

When comparing the outliers found by Cook's distance with the most frequently misclassified instances, there was an overlap of three instances (115, 121 and 139) in the data used in the binary classification and one instance (115) in the multi-class classification. All these instances belong to the β-thalassemia class. Regarding instances 121 and 139, some values were found to be little farther from the β-thalassemia class average and due to the high sensitivity of the algorithms it seems to have been enough to lead them to misclassify these instances. In the case of instance 115, the situation is different, as the RDW is more than double the average obtained in the β-thalassemia class. This leads to the suspicion that this subject was misdiagnosed and that, in fact, instead of just having β-thalassemia trait, he has β-thalassemia trait and IDA, because its values are more similar to those obtained in subjects who have both β-thalassemia trait and IDA.

In conclusion, the existing indexes are well adapted to the Portuguese population, especially the RDWI which presented a median accuracy of 95.4% and even though was not possible to surpass its performance with the created binary classifiers, it was possible to match it with the random forest

algorithm, which among all the algorithms presented the best performance, both in the binary and in the multi-class classification. In addition, it was possible to develop a semi-automatic model able to identify instances that present features different from what would be expected according to the attributed disease and, therefore, may require a second analysis.

5.2 Future Work

As a future work, we should seek to obtain more data from patients with microcytic anaemia, especially from the classes that are less represented in the data used in the scope of this thesis, as it could allow to obtain classifiers for microcytic anaemias with a higher accuracy.

However, the availability of more data will probably not be enough to reach 100% accuracy considering that information about the iron status of the body seems to be important for the identification of IDA. So despite the acquisition of more data, since for now it is not possible to acquire information about the serum ferritin through the CBC test, it must be analysed whether the cost and time required to obtain information regarding the iron status of the body compensates for the increase in accuracy, as the main objective of creating these classifiers is to make the diagnosis faster and more accessible.

Furthermore, since the latest hematology analyzers are already able to use classifiers for diagnosis, it would be an advantage to be able to add the best multi-class classifier developed in this thesis, in order to give an early warming and prevent health complications.

Bibliography

[1] WHO. Worldwide prevalence of anaemia 1993-2005: Who global database on anaemia. *World Health Organization*, 2008.

[2] T. G. DeLoughery. Microcytic anemia. *New England Journal of Medicine*, 371(14):1324–1331, 2014.

[3] J. F. Matos, L. M. S. Dusse, R. V. B. Stubbert, M. R. Ferreira, W. Coura-Vital, A. P. S. M. Fernandes, J. R. de Faria, K. B. G. Borges, and M. d. G. Carvalho. Comparison of discriminative indices for iron deficiency anemia and β thalassemia trait in a Brazilian population. *Hematology*, 18(3):169–174, 2013.

[4] M. Jahangiri, F. Rahim, and A. S. Malehi. Diagnostic performance of hematological discrimination indices to discriminate between βeta thalassemia trait and iron deficiency anemia and using cluster analysis: Introducing two new indices tested in iranian population. *Scientific Reports*, 9(1):1–13, 2019.

[5] M. Sirdah, I. Tarazi, E. Al Najjar, and R. Al Haddad. Evaluation of the diagnostic reliability of different rbc indices and formulas in the differentiation of the β-thalassaemia minor from iron deficiency in palestinian population. *International Journal of Laboratory Hematology*, 30(4):324–330, 2008.

[6] E. Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2^{nd} edition, 2010.

[7] B. A. Patel and A. Parikh. Impact analysis of the complete blood count parameter using naive bayes. *2020 International Conference on Inventive Computation Technologies (ICICT)*, pages 7–12, 2020.

[8] M. Jaiswal, A. Srivastava, and T. J. Siddiqui. Machine learning algorithms for anemia disease prediction. *Recent Trends in Communication, Computing, and Electronics*, pages 463–469, 2019.

[9] S. A. Sanap, M. Nagori, and V. Kshirsagar. Classification of anemia using data mining techniques. *International Conference on Swarm, Evolutionary, and Memetic Computing*, pages 113–121, 2011.

[10] S. Purwar, R. K. Tripathi, R. Ranjan, and R. Saxena. Detection of microcytic hypochromia using cbc and blood film features extracted from convolution neural network by different classifiers. *Multimedia Tools and Applications*, 79(7):4573–4595, 2020.

[11] L. Kabootarizadeh, A. Jamshidnezhad, and Z. Koohmareh. Differential diagnosis of iron-deficiency anemia from β-thalassemia trait using an intelligent model in comparison with discriminant indexes. *Acta Informatica Medica*, 27(2):78, 2019.

[12] B. Leitão. ml-anaemia. `https://github.com/BeatrizNL/ml-anaemia`, 2021.

[13] K. Kaushansky, M. Lichtman, J. Prchal, M. Levi, O. Press, L. Burns, and M. Caligiuri. *Williams Hematology*. McGraw-Hill Education, 9^{th} edition, 2015.

[14] M. J. Cascio and T. G. DeLoughery. Anemia: evaluation and diagnostic tests. *Medical Clinics*, 101 (2):263–284, 2017.

[15] C. Fonseca, F. Marques, A. Robalo Nunes, A. Belo, D. Brilhante, and J. Cortez. Prevalence of anaemia and iron deficiency in Portugal: the EMPIRE study. *Internal medicine journal*, 46(4): 470–478, 2016.

[16] C. Samões, I. Kislaya, M. Sousa-Uva, V. Gaio, P. Faustino, B. Nunes, C. Matias-Dias, and M. Barreto. Prevalence of anemia in the portuguese adult population: results from the first national health examination survey (insef 2015). *Journal of Public Health*, pages 1–8, 2020.

[17] C. Camaschella. Iron-deficiency anemia. *New England journal of medicine*, 372(19):1832–1843, 2015.

[18] J. Old. Screening and genetic diagnosis of haemoglobin disorders. *Blood reviews*, 17(1):43–53, 2003.

[19] A. Tefferi. Anemia in adults: a contemporary approach to diagnosis. *Mayo Clinic Proceedings*, 78 (10):1274–1280, 2003.

[20] J. Flint, R. M. Harding, A. J. Boyce, and J. B. Clegg. The population genetics of the haemoglobinopathies. *Baillière's clinical haematology*, 11(1):1–51, 1998.

[21] A. Cao and R. Galanello. Beta-thalassemia. *Genetics in medicine*, 12(2):61–76, 2010.

[22] D. J. Roberts and T. N. Williams. Haemoglobinopathies and resistance to malaria. *Redox Report*, 8(5):304–310, 2003.

[23] E. Lansiaux, P. P. Pébaÿ, J.-L. Picard, and J. Son-Forget. Covid-19: beta-thalassemia subjects immunised? *Medical Hypotheses*, 142:109827, 2020.

[24] R. Origa. β-thalassemia. *Genetics in Medicine*, 19(6):609–619, 2017.

[25] S. Chaudhary, D. Dhawan, P. G. Bagali, P. S. Chaudhary, A. Chaudhary, S. Singh, and S. Vudathala. Compound heterozygous β+ β0 mutation of hbb gene leading to β-thalassemia major in a gujarati family—a case study. *Molecular Genetics and Metabolism Reports*, 7:51–53, 2016.

[26] D. R. Higgs. The molecular basis of α-thalassemia. *Cold Spring Harbor perspectives in medicine*, 3(1):a011718, 2013.

[27] J. Ferrão, M. Silva, L. Gonçalves, S. Gomes, P. Loureiro, A. Coelho, A. Miranda, F. Seuanes, A. B. Reis, F. Pina, et al. Widening the spectrum of deletions and molecular mechanisms underlying alpha-thalassemia. *Annals of Hematology*, 96(11):1921–1929, 2017.

[28] R. Galanello and A. Cao. Alpha-thalassemia. *Genetics in medicine*, 13(2):83–88, 2011.

[29] E. Urrechaga, J. Hoffmann, S. Izquierdo, and J. Escanero. Differential diagnosis of microcytic anemia: the role of microcytic and hypochromic erythrocytes. *International journal of laboratory hematology*, 37(3):334–340, 2015.

[30] A. Tefferi, C. A. Hanson, and D. J. Inwards. How to interpret and pursue an abnormal complete blood cell count in adults. *Mayo Clinic Proceedings*, 80(7):923–936, 2005.

[31] P. R. Sarma. *Clinical Methods: The History, Physical, and Laboratory Examinations*. Butterworths, 3^{rd} edition, 1990.

[32] A. Tefferi. Practical algorithms in anemia diagnosis. *Mayo Clinic Proceedings*, 79(7):955–956, 2004.

[33] G. H. Guyatt, A. D. Oxman, M. Ali, A. Willan, W. McIlroy, and C. Patterson. Laboratory diagnosis of iron-deficiency anemia. *Journal of general internal medicine*, 7(2):145–153, 1992.

[34] WHO. Serum ferritin concentrations for the assessment of iron status and iron deficiency in populations. *World Health Organization*, 2011.

[35] P. Chowriappa, S. Dua, and Y. Todorov. Introduction to machine learning in healthcare informatics. *Machine Learning in Healthcare Informatics*, pages 1–23, 2014.

[36] S. J. Russell, F. H. El-Khatib, M. Sinha, K. L. Magyar, K. McKeon, L. G. Goergen, C. Balliro, M. A. Hillard, D. M. Nathan, and E. R. Damiano. Outpatient glycemic control with a bionic pancreas in type 1 diabetes. *New England Journal of Medicine*, 371(4):313–325, 2014.

[37] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 1^{st} edition, 2006.

[38] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2^{nd} edition, 2013.

[39] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2):103–130, 1997.

[40] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 339–345, 1995.

[41] H. Zhang. The optimality of naïve bayes. *FLAIRS2004 conference*, pages 562–567, 2004.

[42] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[43] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[44] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and regression trees. *International Group*, 432:151–166, 1984.

[45] Scikit-learn. 1.10.6. Tree algorithms: ID3, C4.5, C5.0 and CART. `https://scikit-learn.org/stable/modules/tree.html`, last accessed on 11/10/2021.

[46] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[47] E. Fix and J. Hodges Jr. Discriminatory analysis-nonparametric discrimination: Consistency properties. *USAF School of Aviation Medicine*, pages 1–21, 1951.

[48] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2):281–305, 2012.

[49] P. A. Vikhar. Evolutionary algorithms: A critical review and its future prospects. *International conference on global trends in signal processing, information computing and communication*, pages 261–265, 2016.

[50] A. E. Eiben, J. E. Smith, et al. *Introduction to evolutionary computing*. Springer, 2^{nd} edition, 2015.

[51] K. Sastry, D. Goldberg, and G. Kendall. Genetic algorithms. *Search methodologies*, pages 97–125, 2005.

[52] S. Hamblin. On the practical usage of genetic algorithms in ecology and evolution. *Methods in Ecology and Evolution*, 4(2):184–194, 2013.

[53] J. R. Koza. *Genetic programming: on the programming of computers by means of natural selection*. MIT press, 2^{nd} edition, 1992.

[54] R. Poli, W. B. Langdon, and N. F. McPhee. *A field guide to genetic programming*. Lulu Enterprises, UK Ltd, 1^{st} edition, 2008.

[55] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[56] R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Appears in the International Joint Conference on Artificial Intelligence (IJCAI)*, 14(2):1137–1145, 1995.

[57] R. Patro. Cross-Validation: K Fold vs Monte Carlo. `https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b`, last accessed on 15/10/2021.

[58] D. M. Hawkins. *Identification of outliers*. Springer, 1^{st} edition, 1980.

[59] R. D. Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

[60] T. Van der Meer, M. Te Grotenhuis, and B. Pelzer. Influential cases in multilevel modeling: A methodological comment. *American Sociological Review*, 75(1):173–178, 2010.

[61] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[62] J. England, B. Bain, and P. Fraser. Differentiation of iron deficiency from thalassaemia trait. *The Lancet*, 301(7818):1514, 1973.

[63] W. Mentzer. Differentiation of iron deficiency from thalassaemia trait. *The Lancet*, 301(7808):882, 1973.

[64] I. Shine and S. Lal. A strategy to detect β-thalassaemia minor. *The Lancet*, 309(8013):692–694, 1977.

[65] B. Ricerca, S. Storti, G. d'Onofrio, S. Mancini, M. Vittori, S. Campisi, G. Mango, and B. Bizzi. Differentiation of iron deficiency from thalassaemia trait: a new approach. *Haematologica*, 72(5): 409–413, 1987.

[66] R. Green and R. King. A new red cell discriminant incorporating volume dispersion for differentiating iron deficiency anemia from thalassemia minor. *Blood cells*, 15(3):481–91, 1989.

[67] S. Jayabose, J. Giamelli, O. LevondogluTugal, C. Sandoval, F. Ozkaynak, and P. Visintainer. # 262 differentiating iron deficiency anemia from thalassemia minor by using an RDW-based index. *Journal of pediatric hematology/oncology*, 21(4):314, 1999.

[68] M. Ehsani, E. Shahgholi, M. Rahiminejad, F. Seighali, A. Rashidi, et al. A new index for discrimination between iron deficiency anemia and beta-thalassemia minor: results in 284 patients. *Pakistan Journal of Biological Sciences*, 12(5):473–475, 2009.

[69] O. A. Telmissani, S. Khalil, and G. T. Roberts. Mean density of hemoglobin per liter of blood: a new hematologic parameter with an inherent discriminant function. *Laboratory Hematology*, 5:149–152, 1999.

[70] J. F. Matos, L. Dusse, K. B. Borges, R. L. de Castro, W. Coura-Vital, and M. d. G. Carvalho. A new index to discriminate between iron deficiency anemia and thalassemia trait. *Revista brasileira de hematologia e hemoterapia*, 38(3):214–219, 2016.

[71] J. D. Bessman and D. I. Feinstein. Quantitative anisocytosis as a discriminant between iron deficiency and thalassemia minor. *Blood*, 53:288–293, 1979.

[72] P. Srivastava and J. Bevington. Iron deficiency and/or thalassaemia trait. *The Lancet*, 301(7807): 832, 1973.

[73] B. D. Faleiro. Hereditary anemia - characterization of the genetic basis and subjacent mechanisms. *Tese de mestrado em Biologia Humana e Ambiente, Universidade de Lisboa, Faculdade de Ciências*, 2020.

[74] A. J. Santos, A. P. Gil, I. Kislaya, L. Antunes, M. Barreto, S. Namorado, V. Gaio, B. Nunes, and C. M. Dias. 1º inquérito nacional de saúde com exame físico insef 2015. *Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA, IP)*, pages 1–80, 2016.

[75] J. Old, C. L. Harteveld, J. Traeger-Synodinos, M. Petrou, M. Angastiniotis, and R. Galanello. *Prevention of Thalassaemias and Other Haemoglobin Disorders: Volume 2: Laboratory Protocols*. Thalassaemia International Federation, 2^{nd} edition, 2012.

[76] G. P. Patrinos, P. Kollia, and M. N. Papadakis. Molecular diagnosis of inherited disorders: lessons from hemoglobinopathies. *Human mutation*, 26(5):399–412, 2005.

[77] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. 51:263–273, 1986.

[78] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977.

[79] Geospiza Research Team. FinchTV 1.4.0.

[80] K. L. Howe, P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, et al. Ensembl 2021. *Nucleic acids research*, 49(D1):D884–D891, 2021.

[81] C. Dodé, R. Krishnamoorthy, J. Lamb, and J. Rochette. Rapid analysis of $-\alpha^{3.7}$ thalassaemia and $\alpha\alpha\alpha^{\text{anti }3.7}$ triplication by enzymatic amplification analysis. *British journal of haematology*, 83(1): 105–111, 1993.

[82] Python Software Foundation. Python 3.8.

[83] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[84] Wes McKinney. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[85] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825): 357–362, 2020.

[86] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007.

[87] K. Codes. Genetic-algorithms. `https://github.com/kiecodes/genetic-algorithms`, 2020.

[88] K. Jamieson and A. Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. *Artificial Intelligence and Statistics*, pages 240–248, 2016.

[89] Scikit-learn: Machine learning in Python. `https://scikit-learn.org`, last accessed on 02/09/2021.

[90] L. Muñoz, S. Silva, and L. Trujillo. M3GP–Multiclass Classification with GP. *European Conference on Genetic Programming*, pages 78–91, 2015.

[91] J. Batista. Python-M3GP. `https://github.com/jespb/Python-M3GP`, 2019.

[92] B. Bengfort and R. Bilbro. Yellowbrick: Visualizing the scikit-learn model selection process. *Journal of Open Source Software*, 4(35):1075, 2019.

[93] D. Aslan, F. Gümrük, A. Gürgey, and C. Altay. Importance of RDW value in differential diagnosis of hypochrome anemias. *American journal of hematology*, 69(1):31–33, 2002.