Seminário Ricardo Jorge

Classification of microcytic anemias using machine learning methods

Beatriz Neves Leitão

Departamento de Genética Humana, INSA

&

Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento (INESC-ID), IST

3 de fevereiro de 2025

REPÚBLICA PORTUGUESA | SAÚDE

SNS SERVIÇO NACIONAL DE SAÚDE

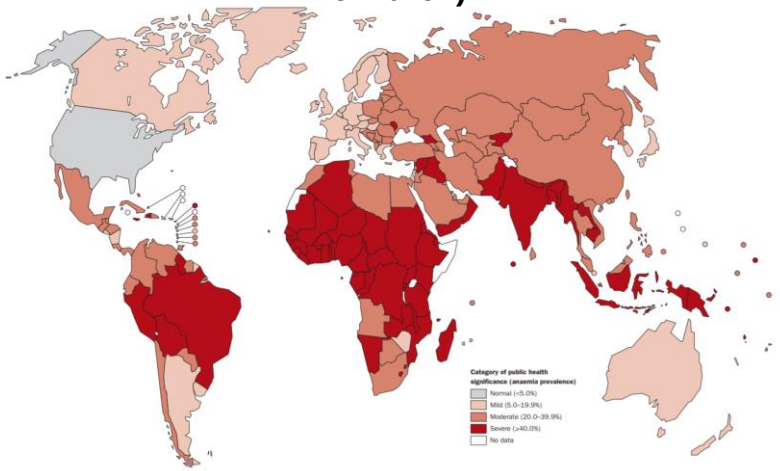
1899 · 2024 125_ANOS INSTITUTO RICARDO JORGE

Anemia

- Anemia is defined as a hemoglobin (Hb) concentration below a specified cut-off point: Hb level <13 g/dL in men, <12 g/dL in women, <11 g/dL in pregnant women.

World Health Organization (2008)

Prevalance of anemia in the world (preschool-age children)



From: World Health Organization (2008)

Main symptoms:

- Fatigue
- Shortness of breath
- Chest pain
- Headache
- Dizziness or lightheadedness
- Fast heartbeat

Estimated to affect 24.8% of the world population

World Health Organization (2008)

Microcytic anemia

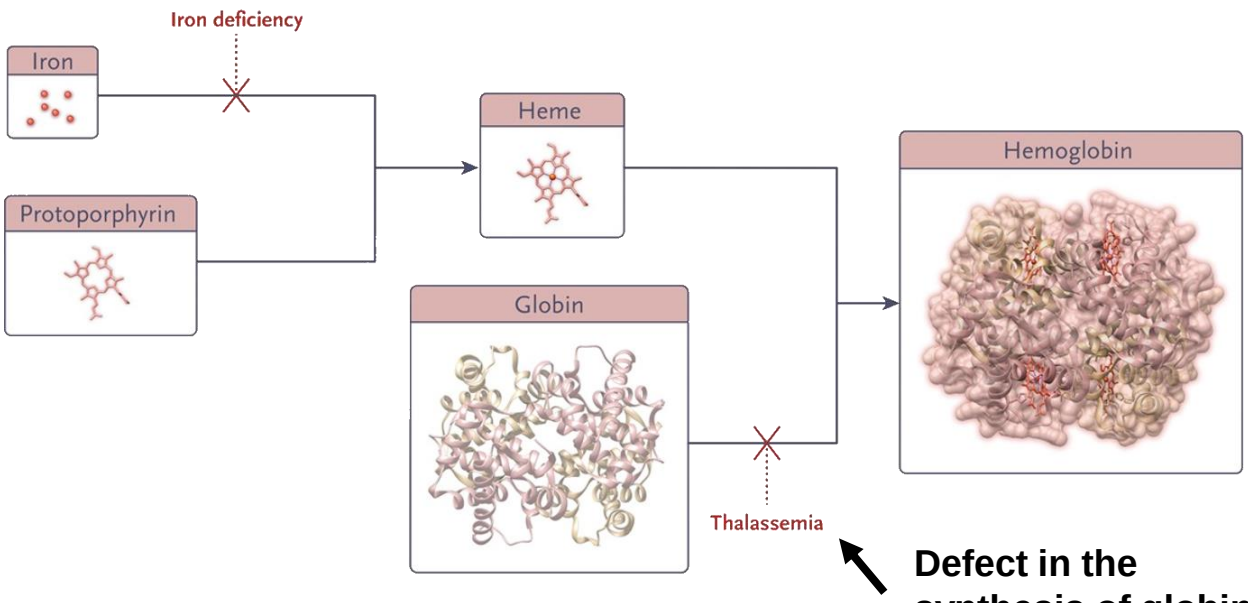
small cell

The most common causes are:

- **Iron deficiency anemia**
- **Thalassemia trait**

β-thalassemia **α-thalassemia**

Low levels of iron



Defect in the synthesis of globin chains due to a genetic disease (thalassaemia)

Adapted from: Thomas DeLoughery (2014)

Diagnosis

In case of Iron Deficiency Anemia (IDA):

- Blood examination to determine Red Blood Cell indices:
  microcytosis = low Mean Corpuscular Volume (MCV <80 fL)
  hypochromia = low Mean Corpuscular Haemoglobin (MCH <27 pg)
- Serum iron-biomarkers - A low serum ferritin level (<15 µg/L) is a sign of depleted iron stores.
- Additional iron index – low transferrin saturation, high Transferrin or Total Iron Binding Capacity.

In case of thalassemia trait:

Beta-thalassemia

- Blood examination to determine Red Blood Cell indices (microcytosis, hypochromia) + elevated HbA2 level (>3.5%)
- DNA testing is required to diagnose thalassemia mutation

Alpha-thalassemia

- Blood examination to determine Red Blood Cell indices (microcytosis, hypochromia)
- DNA testing is mandatory to diagnose α-thalassemia

Discriminant mathematical indices/formulas*

based on Red Blood Cells parameters

A Complete Blood Count (CBC, hemogram) test provides information about:

- hemoglobin (Hb)
- red blood cell (RBC) count
- red blood cell distribution width (RDW)
- mean corpuscular hemoglobin (MCH)
- mean corpuscular hemoglobin concentration (MCHC)
- mean corpuscular volume (MCV) of the red blood cells

* for distinguishing thalassaemia trait from IDA

Index	Formula	Accuracy (%)
England and Fraser	MCV-RBC-(5*Hb)-3.4	**77.11**
Mentzer	MCV/RBC	**84.78**
Shine and Lal	MCV^2 *MCH	**66.37**
Ricerca	RDW/RBC	**61.30**
Green and King	(MCV^2 *RDW)/(100*Hb)	**83.35**
RDWI	(MCV*RDW)/RBC	**81.04**
Sirdah	MCV-RBC-(3*Hb)	**83.68**
Ehsani	MCV-(10*RBC)	**85.89**
Telmissani-MCHD	MCH/MCV	**59.65**
Telmissani-MDHL	(MCH*RBC)/MCV	**68.36**
Index26	Combination of multiple indexes	**84.67**
CRUISE	MCHC+0.603*RBC+0.523*RDW	**75.08**
Matos and Carvalho	1.91*RBC+0.44*MCHC	**78.94**
Bessman	RDW	**36.38**
Srivastava	MCH/RBC	**77.29**

Adapted from: Jahangiri, M. et al. (2019)

Discriminant mathematical formulas

Discriminant formulas

Index	Formula	Accuracy (%)
England and Fraser	MCV-RBC-(5*Hb)-3.4	**77.11**
Mentzer	MCV/RBC	**84.78**
Shine and Lal	MCV^2 *MCH	**66.37**
Ricerca	RDW/RBC	**61.30**
Green and King	(MCV^2 *RDW)/(100*Hb)	**83.35**
RDWI	(MCV*RDW)/RBC	**81.04**
Sirdah	MCV-RBC-(3*Hb)	**83.68**
Ehsani	MCV-(10*RBC)	**85.89**
Telmissani-MCHD	MCH/MCV	**59.65**
Telmissani-MDHL	(MCH*RBC)/MCV	**68.36**
Index26	Combination of multiple indexes	**84.67**
CRUISE	MCHC+0.603*RBC+0.523*RDW	**75.08**
Matos and Carvalho	1.91*RBC+0.44*MCHC	**78.94**
Bessman	RDW	**36.38**
Srivastava	MCH/RBC	**77.29**

Adapted from: Jahangiri, M. et al. (2019)

Machine learning in anemia classification:

Method	Description	Sample size	Accuracy (%)
Naive Bayes	anemic/ non-anemic	2151	**85**
Naive Bayes	anemic/ non-anemic	200	**96**
Random forest	anemic/ non-anemic	200	**95**
C4.5 decision tree	anemic/ non-anemic	200	**95**
C4.5 decision tree	anemic/ non-anemic	514	**98**
Support vector machine	anemic/ non-anemic	514	**87**
Support vector machine	β-thalassemia/ non-anemic	20	**99**
K-nearest neighbors	β-thalassemia/ non-anemic	20	**99**
Artificial neural network	β-thalassemia/ non-anemic	20	**99**
Artificial neural network	IDA/ β-thalassemia	268	**93**

- None of the studies analyzed used data from Portugal
- Few studies used machine learning in the classification of microcytic anemia

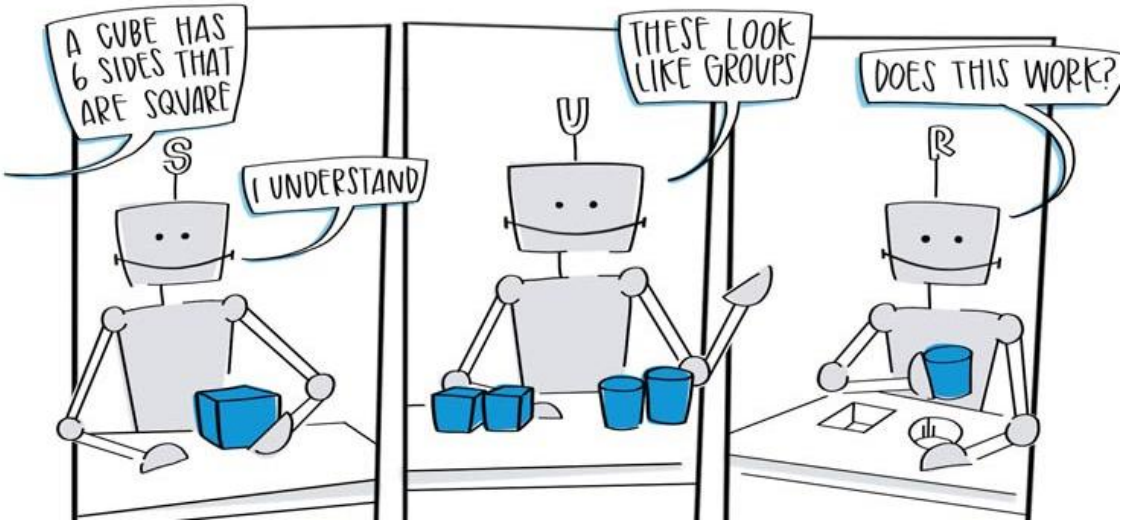
In face of a microcytic anemia revealed by the Complete Blood Count parameters (Hemogram) a correct diagnosis is crucial, which require proceeding for time-consuming, and cost-effective tests and require laboratorial equipment not available in all laboratories and hospitals.

Objectives

- Evaluation the performance of several published mathematical indices for distinguishing the ethiology of microcytic anemias in our population
- Develop a new binary classifier to discriminate between β-thalassemia carriers and iron deficiency anemia (IDA) patients
- Create a multi-class classifier capable of discriminating between β-thalassemia carriers, α-thalassemia carriers, IDA patients, and healthy subjects (control group)

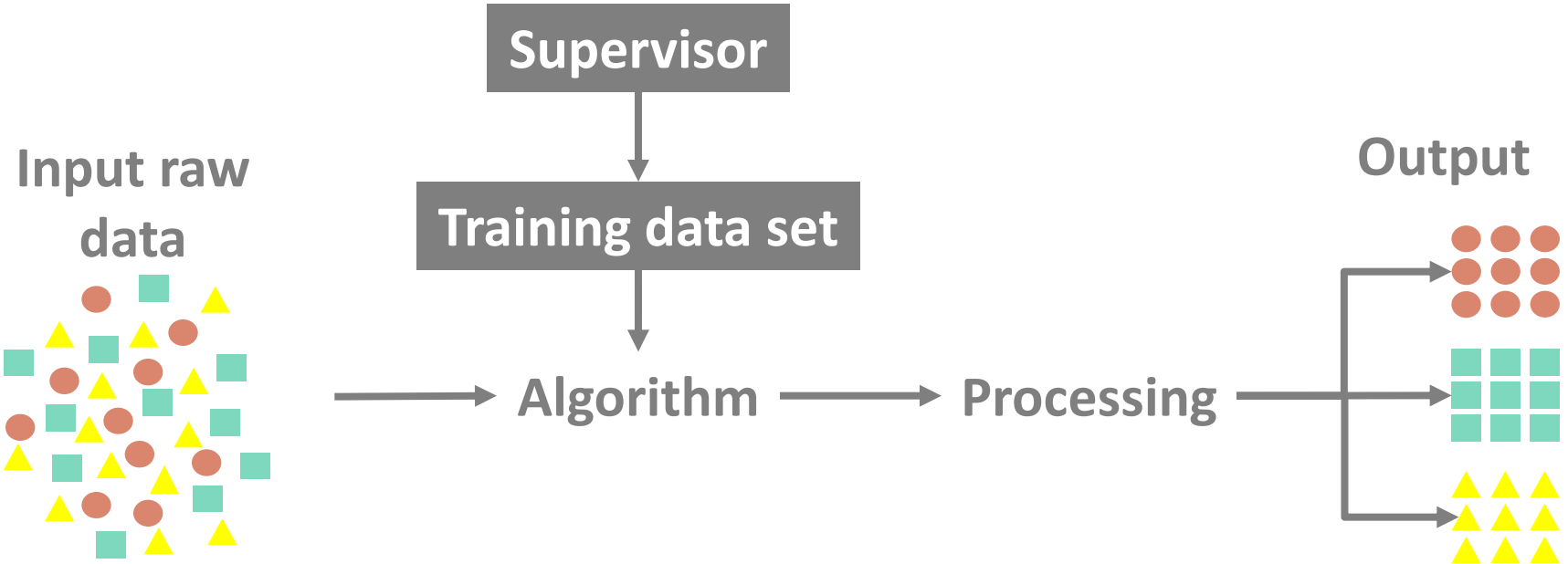
Machine Learning

Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed



Supervised **Unsupervised** **Reinforcement**

Supervised Learning

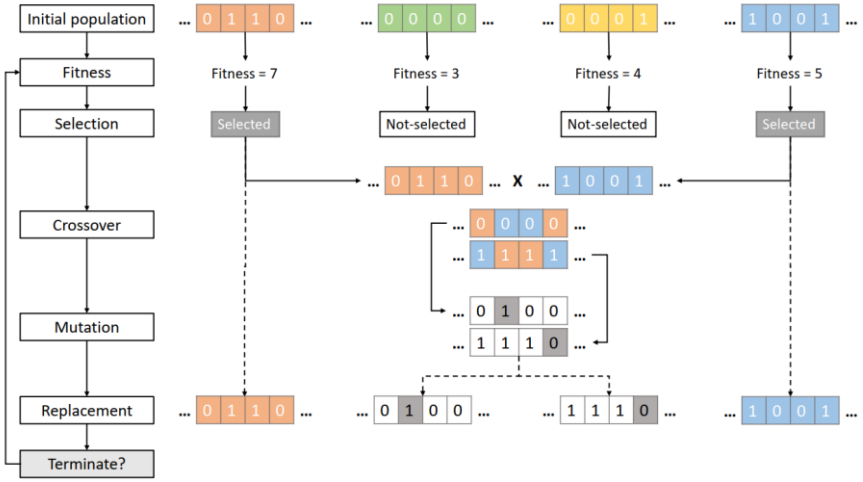


Adapted from: Ronald van Loon (2018)

Machine Learning

Evolutionary algorithms (EA) perform optimization or learning tasks. Inspired by Darwinian natural evolution, EA simulates evolution to generate solutions for complex real-world problems.

Genetic algorithm



Adapted from: S. Hamblin (2013)

Sample

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

- Complete CBC results
- Complete molecular diagnosis of β-thalassaemia and α-thalassaemia
- Incomplete serum ferritin measurements

DGH-UID biobank

Instituto_Nacional de Saúde Doutor Ricardo Jorge

&

INSEF
Inquérito Nacional de Saúde com Exame Físico 2013-2016

Martins R, *et al*. J Hum Genet 2004; 49:651-655.

Nunes B, *et al*. J Public Health (Oxf); 2019; 41:511-517.
Santos D, *et al*. Acta Med Port 2023; 36:467-474.

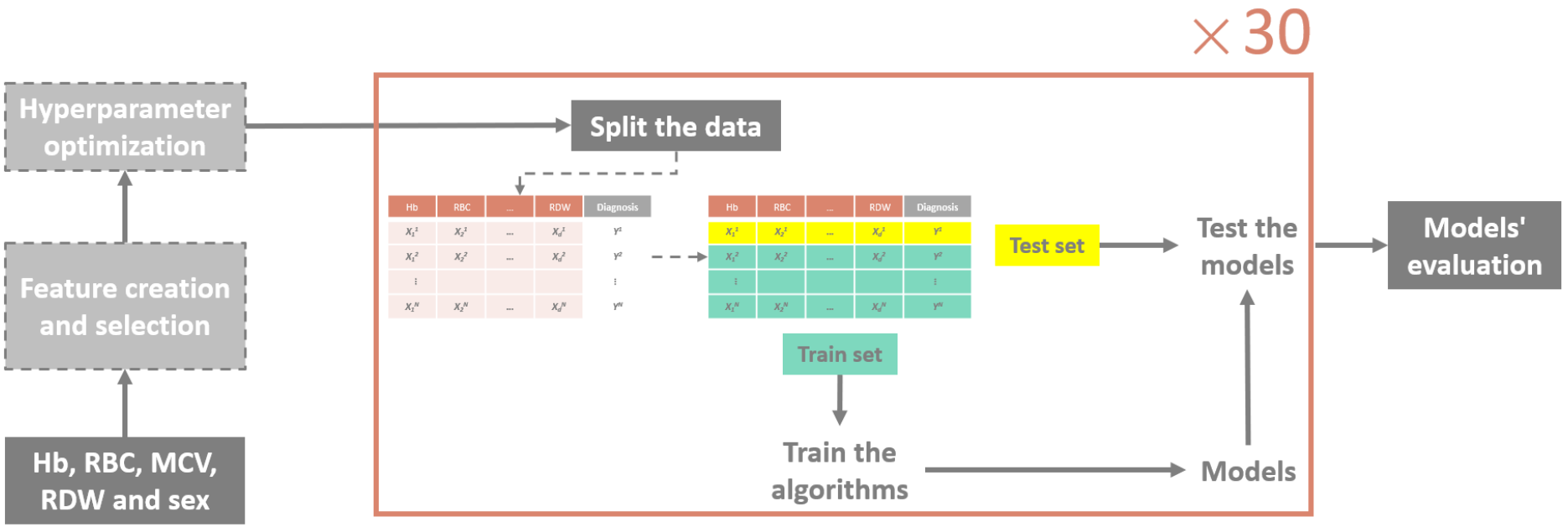
Indexes Evaluation

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

- RDWI and G&K are the most reliable indexes in Portuguese population, just like what was observed in the Brazilian and Palestinian populations

From: J. F. Matos et al. (2013) and M. Sirdah et al. (2008)

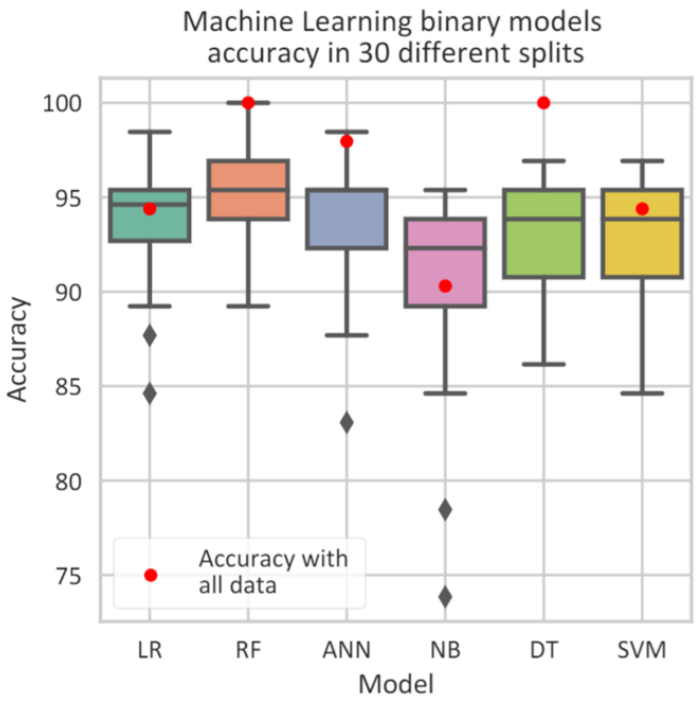
Machine learning classifiers: experimental design



python scikit learn pandas matplotlib seaborn NumPy

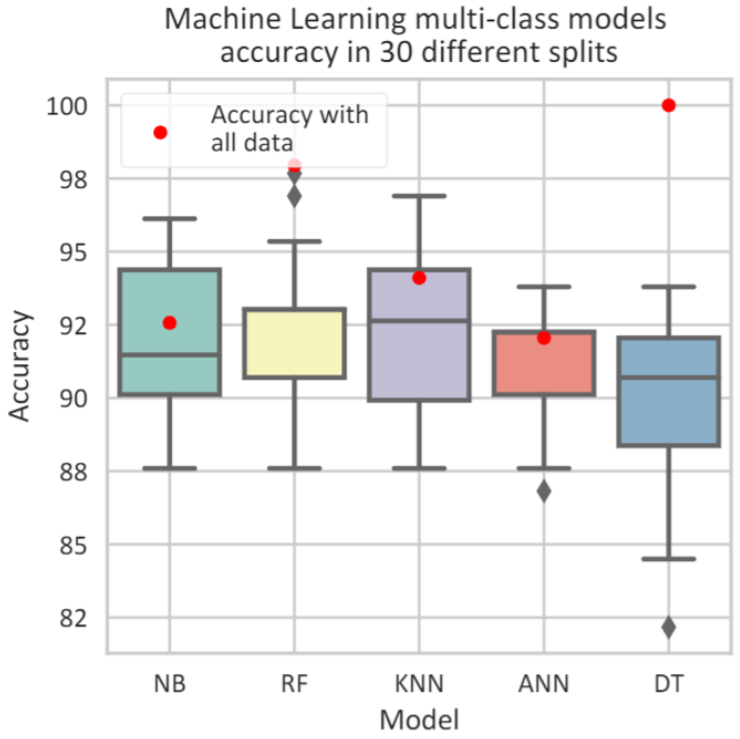
Machine learning classifiers: models' evaluation (binary)

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



Machine learning classifiers: models' evaluation (multi-class)

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



Machine learning classifiers: models' evaluation (multi-class)

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

Models' evaluation

Binary classification:

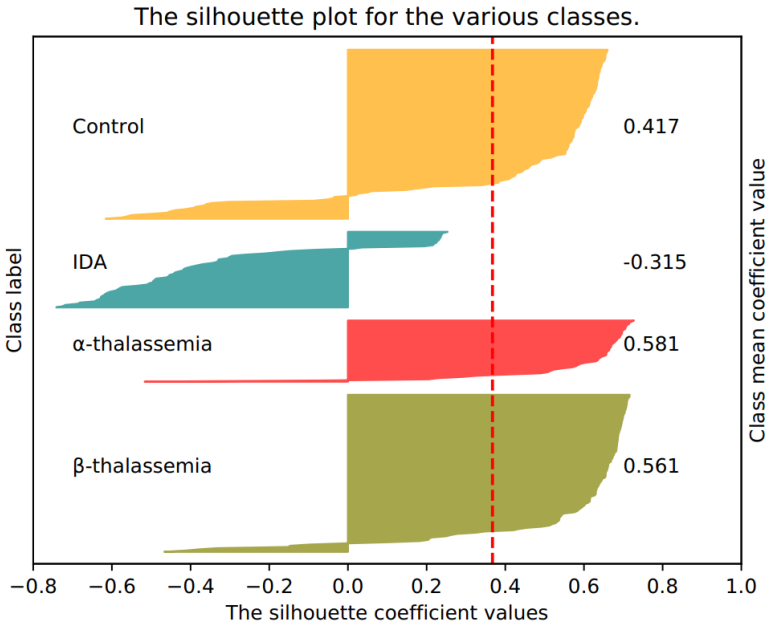
Index	Median accuracy %
RDWI	**95.4**

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**

Multi-class classification:

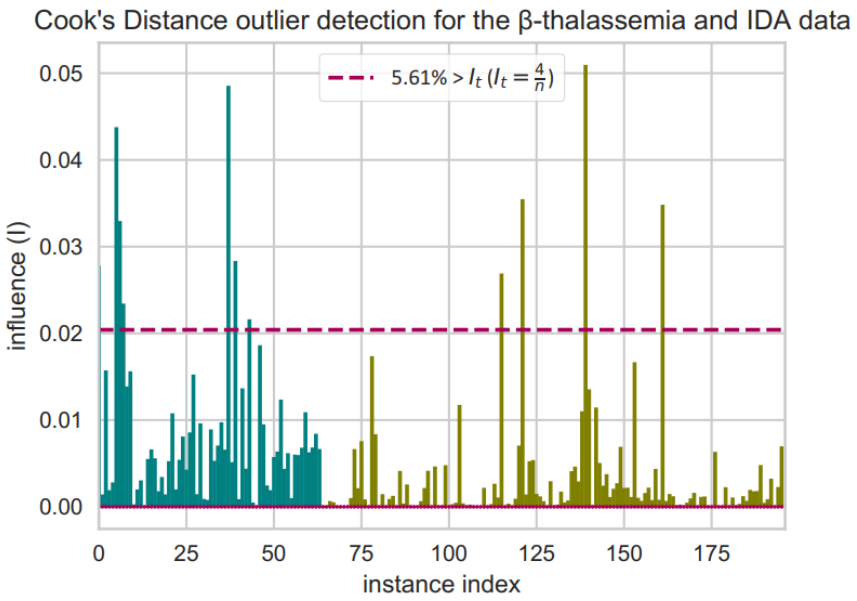
Model	Median accuracy (%)			
	Without hyperparameter optimization		With hyperparameter optimization	
	Initial features	Selected features	Initial features	Selected features
Random forest (RF)	92.6	92.2	**93.0**	91.5

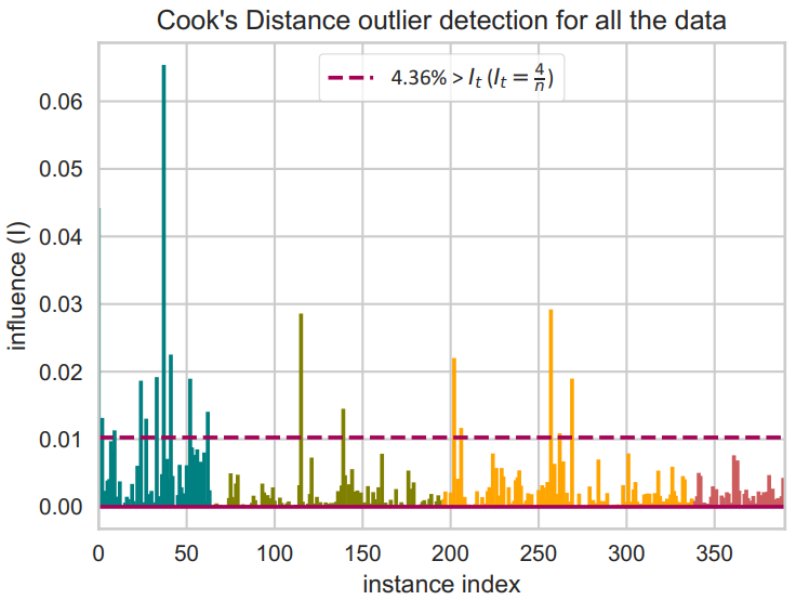
Class Consistency

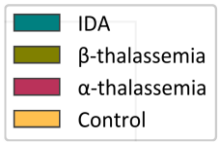


- The α-thalassemia class is the class with the highest mean silhouette coefficient
- The control class had the 2nd lowest average silhouette coefficient
- IDA is the class with the lowest consistency
- β-thalassemia class has a very high consistency
- The average silhouette coefficient is 0.37

Outliers Detection







- IDA is the class where most outliers were found
- The control class does not have any outliers

Outliers Detection

Most frequently misclassified instances:

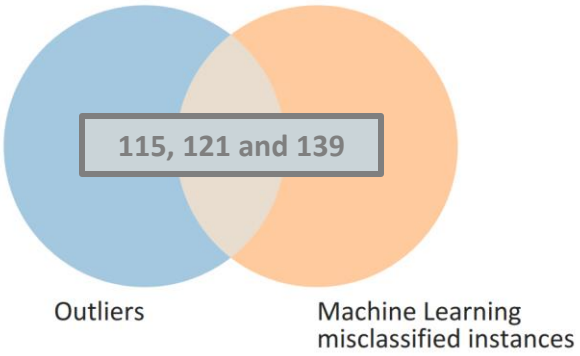
- Binary classification

Class	Nº of instances
IDA	6
β-thalassemia	5

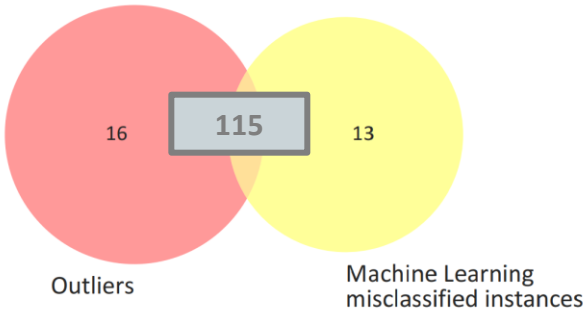
- Multi-class classification

Class	Nº of instances
IDA	3
β-talassemia	5
α-thalassemia	5
Control	1

Venn Diagram of the instances that Machine Learning binary models misclassified and the outliers detected



Venn Diagram of the instances that Machine Learning multi-class models misclassified and the outliers detected



- Instances, 115, 121 and 139, are all from individuals with β-thalassemia trait

Outliers Detection

		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	**34.6** ± 0.8

- Some values are a little farther from the β-thalassemia average
- In instance sample 115 the RDW is more than double the average obtained in the β-thalassemia class.

Conclusions

- The existing indexes are well adapted to the Portuguese population, especially the RDWI [(MCV*RDW)/RBC] which presented a median accuracy of 95.4%
- The random forest algorithm, among all the algorithms presented the best performance, both in the binary (95.4% median accuracy) and in the multi-class (93.0% median accuracy)
- 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

Leitão, B. N., Faustino, P., & Vinga, S. (2022). Comparative Evaluation of Classification Indexes and Outlier Detection of Microcytic Anaemias in a Portuguese Sample. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). (Vol. 13566 LNAI, pp. 219-231). https://doi.org/10.1007/978-3-031-16474-3_19

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

- Since the latest Hematology Analyzer equipment is already capable of using classifiers for diagnosis, it would be an advantage to be able to add the best multiclassifier developed to provide an early warning when a suspicious sample is detected.

Leitão, B. N., Faustino, P., & Vinga, S. (2022). Comparative Evaluation of Classification Indexes and Outlier Detection of Microcytic Anaemias in a Portuguese Sample. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). (Vol. 13566 LNAI, pp. 219-231). https://doi.org/10.1007/978-3-031-16474-3_19

Acknowledgments

Susana Vinga

- DGH

Paula Faustino
Bárbara Faleiro
Daniela Santos
Pedro Lopes

- DEP

Marta Barreto
Irina Kislaya
Carlos Matias Dias

Sara Silva

João Batista

Funding:

Partially supported through FCT (UIDB/50021/2020, PTDC/CCI-BIO/4180/2020, DSAIPA/DS/0026/2019) and EU Horizon 2020 (No. 951970).

INSEF, developed within the scope of the Pre-defined Project of the Programa Iniciativas em Saúde Pública, was promoted by INSA-DEP and benefited from financial support granted by Iceland, Liechtenstein and Norway, through the EEA Grants.

Thank you