Journal Pre-proof

Explaining Machine Learning based Diagnosis of COVID-19 from Routine Blood Tests with Decision Trees and Criteria Graphs

Marcos Antonio Alves, Giulia Zanon de Castro, Bruno Alberto Soares Oliveira, Leonardo Augusto Ferreira, Jaime Arturo Ramírez, Rodrigo Silva, Frederico Gadelha Guimarães

PII: S0010-4825(21)00129-3

DOI: https://doi.org/10.1016/j.compbiomed.2021.104335

Reference: CBM 104335

To appear in: *Computers in Biology and Medicine*

Received Date: 18 December 2020

Revised Date: 25 February 2021

Accepted Date: 10 March 2021

Please cite this article as: M.A. Alves, G. Zanon de Castro, B.A. Soares Oliveira, L.A. Ferreira, J.A. Ramírez, R. Silva, F.G. Guimarães, Explaining Machine Learning based Diagnosis of COVID-19 from Routine Blood Tests with Decision Trees and Criteria Graphs, *Computers in Biology and Medicine*, https://doi.org/10.1016/j.compbiomed.2021.104335.

Explaining Machine Learning based Diagnosis of COVID-19 from Routine Blood Tests with Decision Trees and Criteria Graphs

Marcos Antonio Alves[a], Giulia Zanon de Castro[a], Bruno Alberto Soares Oliveira[a], Leonardo Augusto Ferreira[a], Jaime Arturo Ramírez[b], Rodrigo Silva[c], Frederico Gadelha Guimarães[b,*]

[a] *Graduate Program in Electrical Engineering, Universidade Federal de Minas Gerais*
Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil
[b] *Department of Electrical Engineering, Federal University of Minas Gerais, Brazil*
email: {jramirez, fredericoguimaraes}@ufmg.br
[c] *Department of Computer Science, Federal University of Ouro Preto, Brazil*
email: rodrigo.silva@ufop.edu.br

Abstract

The sudden outbreak of coronavirus disease 2019 (COVID-19) revealed the need for fast and reliable automatic tools to help health teams. This paper aims to present understandable solutions based on Machine Learning (ML) techniques to deal with COVID-19 screening in routine blood tests. We tested different ML classifiers in a public dataset from the Hospital Albert Einstein, São Paulo, Brazil. After cleaning and pre-processing the data has 608 patients, of which 84 are positive for COVID-19 confirmed by RT-PCR. To understand the model decisions, we introduce (i) a local Decision Tree Explainer (DTX) for local explanation and (ii) a Criteria Graph to aggregate these explanations and portrait a global picture of the results. Random Forest (RF) classifier achieved the best results (accuracy 0.88, F1–score 0.76, sensitivity 0.66, specificity 0.91, and AUROC 0.86). By using DTX and Criteria Graph for cases confirmed by the RF, it was possible to find some patterns among the individuals able to aid the clinicians to understand the interconnection among the blood parameters either globally or on a case-by-case basis. The results are in accordance with the literature and the proposed methodology may be embedded in an electronic health record system.

Keywords: COVID–19, Criteria Graph, Decision Tree, Explainable Artificial Intelligence, Machine Learning

*Corresponding author
Email address: fredericoguimaraes@ufmg.br (Frederico Gadelha Guimarães)

1. Introduction

COVID-19, the disease associated with the SARS-CoV-2 virus, was declared a pandemic by the World Health Organization (WHO) on March 11th 2020 [1]. This pandemic has impacted all aspects of life, politics, education, economy, social, environment and climate and set off a warning about how governments, civil society and health systems can deal with an unknown disease. Although many scientific advances have been made and an intense vaccination program is being carried out in several countries, the severe situation is not effectively controlled yet.

An accurate and reliable diagnosis is crucial in providing timely medical aid to suspected or infected individuals and helps the government agencies to prevent its spread and save people's lives. The standard test for COVID-19 is the Reverse Transcriptase Polymerase Chain Reaction, known as RT-PCR, reviewed in [2]. However, it has limitations in terms of resources and specimen collection [3], it is time-consuming [3, 4, 5, 6], it has high specificity and low sensitivity[1] [3, 7, 8], high misclassification in the early symptomatic phase [6] and, also, it is unavailable in many countries and societies making the real extent of the spread still unknown [8, 9].

In addition to the RT-PCR, AI-based approaches may be used to assist in the screening of patients suspected of being contaminated by SARS-CoV-2, supporting the medical decision. In the field of Machine Learning (ML), a branch of Artificial Intelligence (AI) that studies methods that allow computers to learn tasks by examples, many researches studied the diagnosis of COVID-19 either through the analysis of medical images or routine blood tests, as in [6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19].

Routine blood tests play an important role in the diagnosis of COVID-19 and other respiratory diseases. Parameters such as white blood cells (WBC), C-reactive protein (CRP), neutrophils (NEU), lymphocytes (LYM), monocytes (MONO), eosinophils (EOS), basophils (BAY), aspartate and alanine aminotransferase (AST and ALT, respectively), lactate dehydrogenase (LDH) and others have shown high correlations in patients diagnosed with COVID-19 [6, 8, 9, 10, 14, 18, 19, 20, 21, 22, 23].

These hematological features have been used for identifying patterns through ML approaches to verify whether the patient is infected or not. Meng et al. [4] used different indicators of whole blood count, coagulation test, and biochemical examination

[1] In medical diagnosis, specificity refers to the potential of a test to correctly identify those without the disease (true negative rate), whereas sensitivity refers to the potential of a test to correctly detect those with the disease (true positive rate).

to build a Multivariate Logistic Regression (MLR) that was embedded in a COVID-19 diagnosis aid system. Kukar et al. [11] provided a model called "Smart Blood Analytics (SBA)" based on routine blood tests for patients with various bacterial and viral infections and COVID-19 patients. Wu et al. [13] extracted 11 blood indices through Random Forest (RF) algorithm to build an online assistant discrimination tool. Batista et al.[9] used Artificial Neural Networks (ANN), RF, Gradient Boosting Tree (GBT), Logistic Regression (LR) and Support Vector Machines (SVM) to predict the risk of positive COVID-19 using as predictors only results from emergency care admission exams. Brinati et al. [8] developed two classification models using hematological values from Italian patients. RF and Three-Way RF (TWRF) models showed the best results. A Decision Tree was used for explanation. Barbosa et al. [14, 15] proposed the Heg.IA as a support system for the diagnosis of COVID-19. RF is used as the classifier.

Although these models bring promising results in COVID-19 diagnosis, their transparency and trust can be questionable. A model can be defined as explainable if a human can understand its decisions [24]. Any fully automated method without the possibility for human verification would be potentially dangerous in a practical setting, in particular, in the medical field. Explainable ML, or Explainable AI (xAI), typically refers to *post hoc* analysis and techniques used to understand a pre-trained model or its predictions. The ability of a system to explain its decisions is a central paradigm in symbolic or logic-based machine learning [25]. A model-agnostic explainer [25] can interpret a black-box model prediction without assumptions on the underlying black-box model. They are usually employed after the training step (*post-hoc* explainability), see for instance LIME [26] and SHAP [27], providing an understandable output by showing graphically the results and highlighting the features that most contributed to the black-box model decision.

In this work, we search for an accurate ML model for COVID-19 screening based on hematological data and propose the use of a decision tree explainer to improve the interpretability of the best model. We argue that a decision tree more closely resembles the decision-making process of a human healthcare worker and because of that it may be more useful in a real-world environment. We also introduce a criteria graph to aggregate explanations allowing for a generalization of the decision process and a deeper understanding of the interaction of factors leading to a diagnosis.

The main contributions and findings are listed below:

- A literature review of ML methods applied to COVID-19 screening in routine blood tests;
- Reasonable results from different ML techniques (including an ensemble) to

support the diagnosis of COVID-19 using usual blood exams;

- A decision tree-based methodology for the explanation of the model which can be given to the health teams;
- A methodology for aggregating the individual explanations in a graph that shows the relative importance of each attribute and their interactions;
- Further evidence that simple blood tests might help identifying false positive/negative RT-PCR tests.

The remainder of the paper is organized as follows: Section 2 reviews the application of AI for diagnosing COVID-19. Section 3 discusses the Decision-Tree based Explainer (DTX) used for local interpretation. Section 4 presents the proposed Criteria Graph that can be used for global model interpretation. Section 5 explains the ML process, such as models and dataset used, evaluation process and explainability. Section 6 presents the results and discussion. Section 7 provides future directions and conclusions.

2. AI-based approaches in the COVID-19 pandemic

Since the announcement of the pandemic, the scientific community has been working hard to investigate SARS-CoV-2 dynamics. As a result, the volume of papers about COVID-19 has increased exponentially [5]. Reviews were carried out to organize, summarize, and merge the amount of information available in such a short time. For instance, Mohamadou, Halidou and Kapen [28] revised 61 studies dealing with mathematical modelling, AI and datasets related to COVID-19. They reported that most models are either based on Susceptible-Exposed-Infected-Removed (SEIR) as in [29] or SIR model. Toledo et al. [17] provided a historic review of the virus, its epidemiology and pathophysiology, emphasizing the laboratory diagnosis, particularly in hematological changes found during the disease. Wynants et al. [30] provided a systematic review and critic appraisal of current models for COVID-19 for the prognosis of patients and for identifying people at increased risk of becoming infected or being admitted to hospital with the disease. Kermali et al. [10] revised 34 papers discussing biomarkers and their clinical implications. Zheng et al. [31] provided a meta-analysis of the risk factors of critical/mortal cases and non-critical COVID-19 patients, with 13 studies including 3027 patients, in which critical patient conditions and parameters were highlighted.

Regarding AI and ML-based works, Yan et al. [21] applied an Extreme Gradient Boosting Machine (XGBoost) algorithm to predict risk mortality, in which a

single-tree was used to build an explanation for the model. Tian et al. [32] investigated the predictors of mortality in hospitalized patients in a total of 14 studies documenting the outcome of 4659 patients. Comorbidities such as hypertension, coronary heart disease, and diabetes were associated with a significantly higher risk of death amongst infected patients. Clinical manifestation laboratory examinations that could imply the progression of COVID-19 were presented. Shi et al. [33] analyzed AI techniques in imaging data acquisition, segmentation, and diagnosis. These images, either X-ray or CT images, can improve the work efficiency of the specialists by an accurate delineation of infections. Also in the AI context, Bullock et al. [5] revised datasets, tools, resources to confront many aspects of the COVID-19 crisis at different scales including molecular, clinical, and societal applications. In the clinical aspect, medical images, outcomes prediction and noninvasive measurements were discussed. Although these works have made valuable contributions to dealing with the pandemic, the decision made by the automatic learning model on the samples is still unclear.

In the revised literature, important hematological features were highlighted such as CRP [21], LDH [8, 21, 23, 34], AST, ALT, NEU [8], LYM and WBC [8, 9, 11, 34], EOS [8, 9] and others, see also [4, 6, 13, 14, 16, 20]. These features are detailed in Table 1 with a short description of each hematological parameter, the reference value for males and female and the percentage of missing rates presented in the dataset used. In the literature, they were commonly estimated either through statistics as in [6, 16, 20] or a ML model or metric, such as RF in [8], Least Absolute Shrinkage and Selection Operator (LASSO) in [4], Multi-tree XGBoost in [21] or an evolutionary strategy as in [14].

The state-of-the-art algorithms have been the most used, such as the Support Vector Machine (SVM) in [9, 16], XGBoost in [11, 21] and RF in [8, 13]. For the sake of simplicity, in Table 2 we summarize the works that used ML techniques to classify patients suspected of being infected with SARS-CoV-2 using hematological parameters. There is a short description of the papers, methods used (the best one is in bold), features analyzed, and the results for each performance metric.

A series of recently published papers have reported the epidemiological and clinical characteristics of patients with COVID-19 disease, however there is no standard for data collection. Many public datasets available have different features and a large number of missing values, making it difficult to aggregate this data into a single ML model.

Although many papers have presented ML-based support approaches to deal with COVID-19 screening in routine blood tests, only Brinati et al. [8] and Yan et al. [21] have raised the necessity of some sort of transparency in the model's decisions. The

Abb.	Feature	Description	Reference Value		Miss. %	Ref
			Female	Male		
HCT	Hematocrit	The amount of whole blood that is made up of red blood cells	36–46 %	41–53 %	0.82	[11, 35]
HGB	Hemoglobin	It is the oxygen-carrying component of red blood cells	12–16 g/dL	13.5–17.5 g/dL	0.82	[11, 35]
PLT	Platelets	A tiny, disc-shaped piece of cell that helps form blood clots to slow or stop bleeding and to help wounds heal	150–400 $\times 10^9/L$	150–400 $\times 10^9/L$	0.98	[6, 10, 35]
RBC	Red blood Cells	The blood cell that carries oxygen	3.5–5.5 $\times 10^{12}/L$	4.3–5.9 $\times 10^{12}/L$	0.98	[35]
LYM	Lymphocytes	A type of white blood cells	0.5–4.0 $\times 10^9/L$	0.5–4.0 $\times 10^9/L$	0.98	[10, 21, 36]
MCH	Mean corpuscular hemoglobin	It corresponds to the average hemoglobin weight in a population of erythrocytes	25.4–34.6 pg/cell	25.4–34.6 pg/cell	0.98	[11, 35]
MCHC	MCH concentration	Mean of the internal hemoglobin concentration in a population of erythrocytes	31–36 % Hb/cell	31–36 % Hb/cell	0.98	[11, 35]
WBC	Leukocytes	White Blood Cells that help the body fight infections and other diseases.	4500–11000 $/mm^3$	4500–11000 $/mm^3$	0.98	[11, 12, 34, 35]
BAY	Basophils	Type of white blood cell (leukocyte) with coarse, bluish-black granules of uniform size within the cytoplasm	0.0–0.1 $\times 10^9/L$	0.0–0.1 $\times 10^9/L$	0.98	[13, 36]
EOS	Eosinophils	Normal type of white blood cell that has coarse granules within its cytoplasm	0.1–0.5 $\times 10^9/L$	0.1–0.5 $\times 10^9/L$	0.98	[11, 37, 36]
LDH	Lactate dehydrogenase	Enzyme of the anaerobic metabolic pathway, that catalyzes the conversion of lactate to pyruvate, important in energy production	140–280 U/L	140–280 U/L	0.98	[23, 38]
MCV	Mean corpuscular volume	Average volume of an erythrocyte population	80–100 μm^3	80–100 μm^3	0.98	[35]
RWD	Red blood cell distribution width	A measurement of the range in the volume and size of red blood cells	< 15 %	< 15 %	0.98	[39]
MONO	Monocytes	A type of immune cell that has a single nucleus and fights off bacteria, viruses and fungi	0.3–0.8 $\times 10^9/L$	0.3–0.8 $\times 10^9/L$	1.15	[39]
MPV	Mean platelet volume	Average size of platelets	7.2–11.7 fL	7.2–11.7 fL	1.48	[40]
NEU	Neutrophils	A type of immune cell that is one of the first cell types to travel to the site of an infection and help by ingesting microorganisms and releasing enzymes that kill them	1.8–7.7 $\times 10^9/L$	1.8–7.7 $\times 10^9/L$	15.62	[16, 10, 20, 34, 39]
CRP	C-reactive protein	Plasma protein produced by the liver and induced by various inflammatory mediators such as interleukin-6	<10 mg/L	<10 mg/L	16.77	[6, 10, 8, 20, 21, 34]
CREAT	Creatinine	A chemical waste molecule generated from muscle metabolism.	44–97 μmol/L	53–106 μmol/L	30.26	[11, 13, 13, 41]
UREA	Urea	A nitrogen-containing substance normally cleared from the blood by the kidney into the urine.	2.5–7.1 mmol/L	2.5–7.1 mmol/L	34.70	[11, 20]
K+	Potassium	A metallic element that is important in body functions such as regulation of blood pressure	3.5–5.5 mEq/L	3.5–5.5 mEq/L	38.98	[42]
Na	Sodium	A mineral needed by the body to keep body fluids in balance	135–145 mmol/L	135–145 mmol/L	39.14	[34]
AST	Aspartate transaminase	An enzyme found in the liver, heart, and other tissues. A high level of AST released into the blood may be a sign of liver or heart damage, cancer, or other diseases	0–35 U/L	0–35 U/L	62.82	[6, 20, 34]
ALT	Alanine transaminase	An enzyme that is normally present in liver and heart cells and it is released into blood when the liver or heart is damaged	<41.0 U/L	<31.0 U/L	62.99	[6, 20]

Table 1: Description of the features used, abbreviation (Abb.) often used/adopted, reference values for male and female, missing rates (Miss. %) and some related references that reported the feature's relationship with COVID-19.

Ref	Description	Dataset	Methods	Features	Inter.	Metric results
[9]	Predict the risk of positive cases using as predictors only results from emergency care admission exams	235 patients from Hospital Israelita Albert Einstein in São Paulo, Brazil.	NN, RF, GB-Trees, LR, **SVM**	15 blood parameters	No	AUC 0.85, SE 0.68, SP 0.85, PPV 0.74, NPV 0.77
[4]	ML-based diagnosis model and a COVID-19 diagnosis aid application	620 patients from West China Hospital	**MLR**	Age, gender and more 35 indicators	No	AUC 0.87, PPV 0.86, NPV 0.85
[8]	ML models using hematochemical values from routine blood exams	279 patients from San Rafaele Hospital in Milan, Italy	DT, ET, KNN, LR, NB, **RF**, SVM, **TWRF**	Several	DT	For RF: ACC 0.82, AUC 0.84, SE 0.92, SP 0.65, PPV 0.83. For TWRF: ACC 0.86, SE 0.95, SP 0.75, PPV 0.86
[11]	Smart Blood Analytics (SBA) predictive model on patients with various bacterial and viral infections, and COVID-19 patients	5333 patients from Department of Infectious Diseases, University Medical Centre Ljubljana, Slovenia.	RF, DNN, **XGBoost**	35 blood parameters	No	AUC 0.97, SE 0.82, SP 0.98
[13]	RF model and an online assistant tool.	253 samples from 169 suspected patients collected from multiple sources.	**RF**	49 clinical available blood test data.	No	ACC 0.96, AUC 0.96, SE 0.95, SP 0.97, MCC 0.96, Related AUC 1.00
[14]	Heg.IA: An intelligent system to support the diagnosis of Covid-19 based on blood tests	5644 patients provided by Hospital Israelita Albert Einstein (São Paulo, Brazil). 559 had positive diagnosis.	MLP, SVM, RT, RF, **BN** and NB	24 blood tests	No	ACC 0.95, PR 0.94, SE 0.97, SP 0.94, Kappa index 0.90
[21]	Predict the mortality risk and explain the model.	2779 validated or suspected COVID-19 patients from Tongji Hospital in Wuhan, China.	**XGBoost**	Several	Single Tree XGB	F1 0.93, PR 0.95, SE 0.92
[16]	Detect the COVID-19 severely ill patients from those with only mild symptoms.	137 clinically confirmed cases from the Tongji Hospital Affiliated to Huazhong University of Science and Technology.	LR, **SVM**, RF, KNN, AdaBoost	100 features (8 clinical, 76 blood, and 16 urine)	No	ACC 0.79, SE 0.76, SP 0.70
[22]	Predict mortality risk	70 survivors from SMS Medical College, Jaipur (Rajasthan, India).	**LR**	Several	No	ACC 0.70, AUC 0.95, SE 0.90, SP 0.89
[34]	Identify patients at risk for deterioration during their hospital stay	6995 patients were evaluated at Sheba Medical Center, China	RF, NN, CRT	Several	No	ACC 0.79, AUC 0.79, SE 0.68, SP 0.81. All of them with Apache II
[18, 19]	Prediction of the diagnosis based on blood count results and age	1157 patients made available by the repository COVID-19 Data Sharing/BR	**XGBoost**	Several	No	ACC 0.80, F1 0.70, AUC 0.81, SE 0.76, PPV 0.65, NPV 0.88

Papers that applied ML models for prediction of COVID-19, datasets and models used (the best model reported is the bold one), features analyzed, interpretability ter.), m
models for prediction of COVID-19, datasets and models used (the best model reported is the bold one), features
metric results in each paper. The methods are BN: Bayesian Networks, CRT: Classification and Regression Tree,
Decision Trees, ET: Extremely Randomized Trees, GBT: Gradient Boosting Trees, KNN: k-Nearest Neighbors,
ilayer Perceptron, MLR: Multivariate Logistic Regression, NB: Naive Bayes, NN: Neural Networks, RF: Randon
ne, TWRF: Three-Way RF, XGBoost: Extreme Gradient Boosting Machine.

former presents a Decision Tree as an interpretable model but in doing so accuracy is getting sacrificed. In the latter, the authors used the XGBoost algorithm to obtain the relative importance of the features and built a Single-Tree XGBoost on the three most important (LDH, LYM and high-sensitivity CRP). Again, this is an approach that trades accuracy by interpretability.

In this paper, we evaluate different ML methods, including ensembles, for COVID-19 diagnosis from routine blood tests. Besides, our methods include cleaning and pre-processing steps, imbalance class treatment, the creation of ensemble models, and an interpretability module. The proposed methodology can be generalized to other contexts as a pipeline for the ML workflow. Local interpretability is provided by using a Decision Tree-based explainer (DTX) (Section 3) and global interpretability is obtained with the criteria graph (Section 4) proposed herein. The DTX presents an explanation for the high-accuracy black-box model. Therefore, the quality of the predictions does not have to be sacrificed. On the other hand, this means that the explanations are individual. Thus, to get an insight into the models global behaviour, the Criteria Graph compresses the information of all the explanations and presents it in a single image.

3. Decision–Tree based Explainer

The *post hoc* explanation approaches aim to explain the predictions of a particular pre-trained ML model. These explanations can be of two types:

- **Instance explanation:** aims to explain predictions of the black-box model for individual instances. It provides local scope for interpretability.
- **Model explanation:** it is usually the result of aggregating instance explanations over many training or testing instances. This approach provides global level interpretability, generalizing local explanations. The aggregation of many instances enables the identification of the impact of features in the classification and knowledge extraction from the ML model.

The interpreter applied in this work is known as Decision Tree-based Explainer (DTX). DTX can be defined as a model-agnostic, *post hoc*, perturbation-based, feature selector explainer. This approach generates a readable tree structure that provides classification rules, which reflect the local behaviour of the complex ML model around the instance to be explained. The explainer can understand the black-box

model according to:

$$\sum_{i=0}^{|\eta|} g(f_{dtx}(x_i) - f_{bbm}(x_i)) \ \ \forall \ \ x_i \in \eta \tag{1}$$

where $f_{dtx}(x_i)$ is the DT prediction, $f_{bbm}(x_i)$ is the black-box model prediction, η is a noise set created around the instance to be explained, $|\eta|$ is the number of samples around the instance to be explained, $g(.)$ measures the distance between the black-box prediction and DTX prediction, for instance, in classification problems we can use accuracy.

The set η is created with artificial samples generated around the instance that we want to explain. This set is used to fit the explainer and to measure the accuracy of the explainer concerning the black-box model. Equation (1) implies local fidelity of the explainer to the predictions provided by the black-box model. The correctness of the prediction is orthogonal to the correctness of the explanation, but enforcing local fidelity to better models (in terms of higher accuracy) might enable better explanations.

Figure 1 illustrates how the DTX presents an understandable visual output. The left side shows the noise set η around the sample ($\mathbf{x}$) that is going to be explained. It also shows the decision boundaries defined by the explainer. The right side shows the tree structure generated by DTX for a local explanation. Also, DTX works as a feature selector, since the features presented in the tree are the most important for the method around the neighbourhood of $\mathbf{x}$.

In the example in Figure 1, the explanation provided for why sample $\mathbf{x}$ is classified as class 1 (positive class), is given by the path in the tree that lead to this outcome: $x_2 \geq 0.074$ and $x_1 \leq -0.04$.

4. Criteria Graph for pattern identification in explanations

From the previous section, one can see that the decision tree explainer returns a rule of the type:

$$\textbf{if } criterion_1 \textbf{ and } ... \textbf{ and } criterion_n \textbf{ then } \text{class} = X \tag{2}$$

where a criterion is defined as $attribute \diamond value$ and $\diamond$ is one of $\leq$, $\geq$, $<$ or $>$ operators.

This kind of rule is easy to understand and provides valuable information to the health worker. Nevertheless, each patient will have its own local explanation and it might be useful to understand relationships between criteria over the whole population.

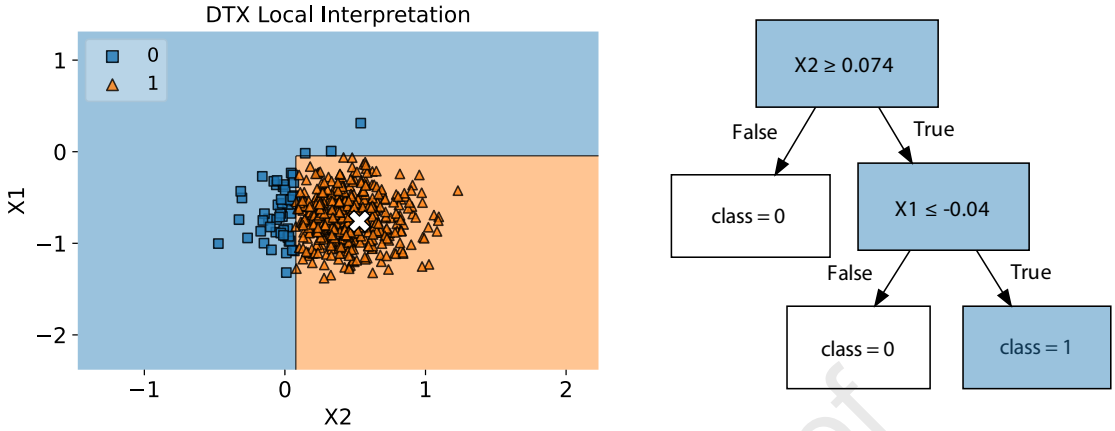


Figure 1: In the left side, there is a noise set η generated by DTX around the instance to be explained, $\mathbf{x}$. The decision boundary is based on the DTX output. In the right side there is a tree structure representing the rules responsible for explaining the black-box prediction.

To provide this information, in this work, we also propose a global interpretability method named Criteria Graph, which works as follows:

Given a set of rules, $\mathcal{R} = \{R_1, R_2, ..., R_m\}$, where each rule, R_i, is the explanation for the i^{th} patient's diagnosis, and m is the number of patients. First, for each attribute, we discretize the values of each criterion. Being the mean value of that attribute, μ, and the standard deviation, σ, if a *value* is in the interval $[\mu - 0.5\sigma, \mu + 0.5\sigma]$ it gets the label *medium*. If $value < \mu - 0.5\sigma$ it gets the label *low* and if $value > \mu + 0.5\sigma$ it gets the value *high*.

After discretization, each criterion becomes a node in the graph. The size of the node is proportional to the number of patients for which that criterion was used in the diagnosis. If two criteria appear in the same rule, a link is created between them and the width of the link is proportional to the number of patients for which the two criteria are used in the diagnosis.

Figure 2 shows the result of this procedure applied to the set below. Notice that the color of each node provides an extra visual cue related to the value of the criterion. Red for *low*, Blue for *high* and Yellow for *medium*.

$$\mathcal{R} = \begin{bmatrix} (\textbf{if } a_1 > low \textbf{ and } a_2 < medium \textbf{ then } class = 1) \\ (\textbf{if } a_1 > low \textbf{ and } a_3 < high \textbf{ then } class = 1) \\ (\textbf{if } a_1 > low \textbf{ and } a_2 < medium \textbf{ then } class = 1) \end{bmatrix}$$

The Criteria Graph is a model explanation obtained by aggregating instance explanations, as provided by DTX, over many instances, to identify patterns in

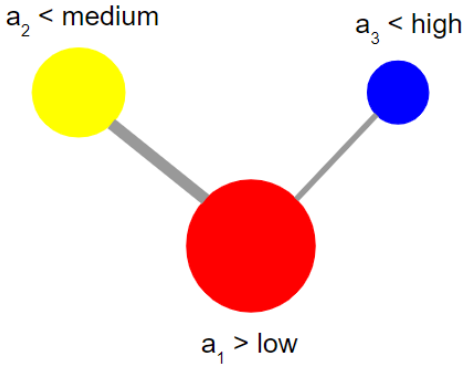


Figure 2: Criteria Graph

explanations.

5. Methods

In this paper, we focus on COVID-19 binary classification using a public dataset detailed in subsection 5.1. The ML procedures for generating classifiers with evolving explanations consist, basically, of two main steps: (i) evaluation of different artificial learning models, and (ii) comparison among SHAP, LIME and DTX for local interpretation of the output and criteria graph for global interpretation. Figure 3 provides an overview of the entire process.

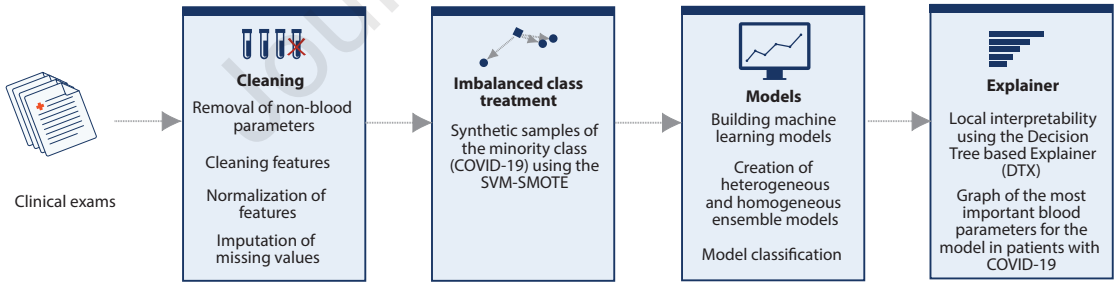


Figure 3: Diagram of the proposed method of generating ensemble classifiers with local explainability.

5.1. Dataset

The dataset contains anonymous data from patients seen at the Hospital Israelita Albert Einstein, São Paulo, Brazil, and who had samples collected to perform the

SARS-CoV-2 by RT-PCR and additional laboratory tests during the visit. The dataset is publicly available in [43] for collaborative research and it is often updated. The raw version we used contained 5644 samples and 111 features, standard normalized (z-score), related to the medical tests, such as blood, urine and others.

5.2. Pre-Processing

To select the most representative parameters in the dataset we first define a threshold of 95% for removing features with several missing values greater than it. Non-blood features were also discarded, such as urine tests and other contagious infectious diseases. These diseases include respiratory infections, such as influenza A and B; parainfluenza 1, 2, 3 and 4; enterovirus infections and others. We remove these features since the dependence of the diagnosis on a variety of other infectious diseases for COVID-19 prediction is not a practical situation in the emergency context. Furthermore, a false negative result of one of these diseases would generate a spread of the error.

However, the diagnostic results for the others infectious diseases could be used to train a multiple output classifier, which may assist the health professional in the process of diagnosing simultaneous diseases. But this is not the focus of this work.

The set of final features were detailed in Table 1. After the cleaning process, we found a total of 608 observations, being 84 positive and 524 negative COVID-19 confirmed cases through RT-PCR being, thus, an imbalanced data problem. The distribution for each class is approximately 1:6 ratio. Since many null values remained, it was necessary an imputation technique to deal with. The "Iterative Imputer" technique from Scikit-learn package [44] showed the best performance in experimental tests compared with mean or median.

5.3. Evaluation of Predictive Models

In this paper, we use as a baseline the state-of-the-art of Logistic Regression [45], XGBoost [46] and Random Forest [47], since these algorithms have shown good results in problems with imbalanced data, as in [8, 13, 11, 21]. We also tested the SVM and MLP methods.

We train and evaluate these models through a nested cross-validation procedure [48]. As illustrated in Figure 4, first, in each iteration, the dataset is stratified between two subsets: training + validation and test set. In the inner loop, training + validation are divided into k folds and the model being trained in $k-1$ partitions. The other fold, which does not participate in the training, is used for model validation and for selecting the best set of hyperparameters through the Grid Search algorithm. At the end of an iteration, the model is evaluated in the test set. In the outer loop,

this process is repeated in other different training + validation and test set folds, mutually exclusive. The nested cross-validation method, in this way, allows a more reliable evaluation of the model generalization.

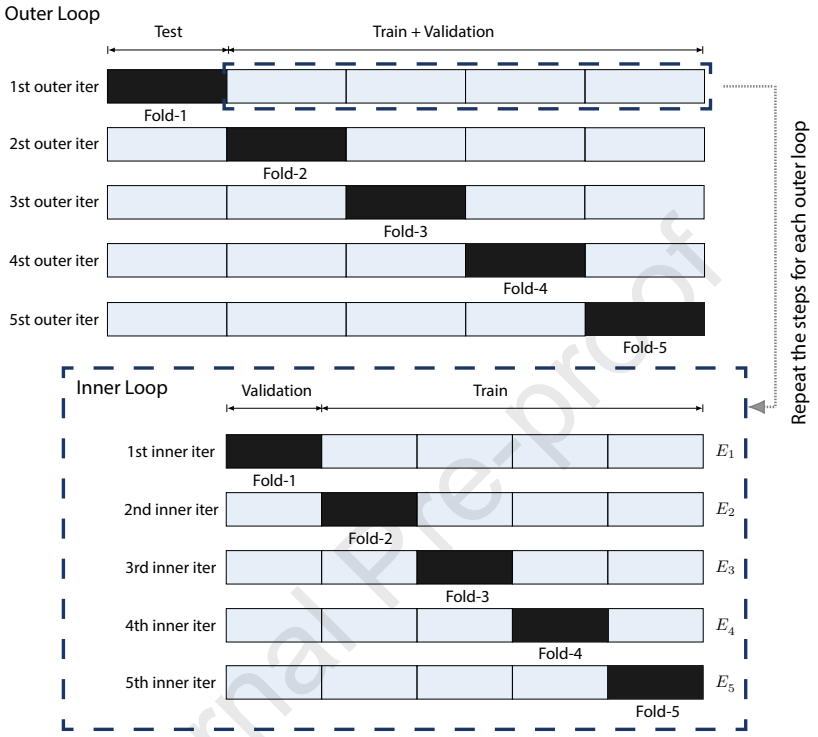


Figure 4: The nested cross validation method

For the evaluation of the models, we chose the known f1–score [49] to measure the best set of hyperparameters. Since 524 patient observations had no detection of the SARS-CoV-2 (86% of the dataset), the evaluation of accuracy does not provide a representative measure. F1–score, in its turn, provides a measure of the discrimination capacity of the models.

For the RF algorithm, we vary the number of estimators in the set of $\{10, 20, 30, 45, 50, 55, \cdots, 90, 95, 100\}$ trees, while we change the maximum depth of the tree in the set of $\{2, 4, 8, 16, 32, 64\}$. For the XGBoost, the same set of hyperparameters was applied, adding a learning rate of $\{0.1, 0.05, 0.01\}$. For the SVM, we vary the cost hyperparameter in the set of $\{0.001, 0.01, 0.05, 0.1, 0.5, 1, 10\}$ and linear and rbf kernels. In the MLP algorithm, we test hidden layers of size $(64, 32, 16)$, $(32, 32, 16)$, $(64, 32, 32)$ and $(64, 64, 32, 32)$, with constant or adaptive learning rate. We define the hyperpa-

rameter alpha in the set of $\{0.01, 0.05, 0.005, 0.001, 0.0001\}$.

We train each algorithm using the SVM–Synthetic Minority Over-sampling Technique (SVM–SMOTE) [50]. Through this technique, minority class data are synthetically over-sampled, presenting for the training subset the same proportion of instances for the positive and the negative class. Resampling by this technique is performed by creating a synthetic sample between the k neighbors closest to the instance, as shown in Figure 5. For this task, we select a number of $k = 5$ neighbors.

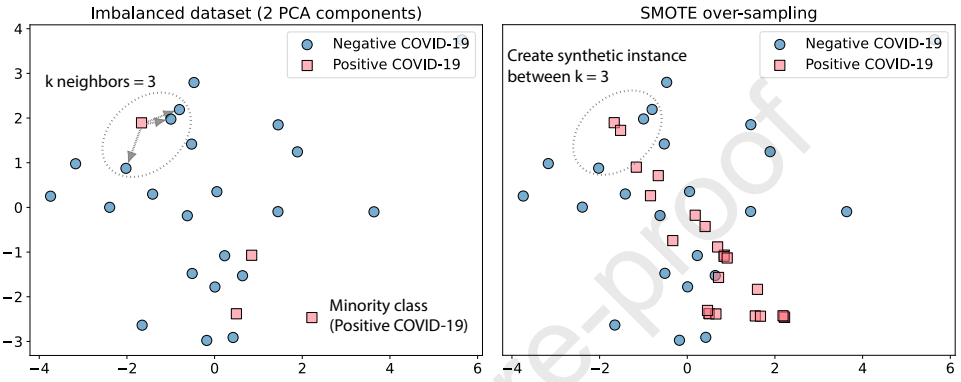


Figure 5: Example of synthetic sample generated by SMOTE

Through the nested cross-validation method, we generate five final models for each algorithm, which correspond to the number of external partitions. Thus, we choose the best of the five models generated for each method and retrain it in 10 iterations using the selected hyperparameters to measure their ability to generalize. For each iteration, we split the data in 80% for training and the rest for the test set. Considering the imbalanced data, we applied the SMOTE again, but only for the training data, for each of the interactions, synthetically super-sampling the minority class data.

5.4. Ensemble

To compose the ensemble, we combine the best nested cross-validation models of RF, LR, XGBoost, SVM, and MLP. The label was predicted based on the majority voting decision. For weighting the votes, the model that obtained the best performance received a weight equal 2 and the worst one a weight equal 0.

After generating the ensemble, we evaluated the combined models in each test subset of the 10 iterations, using the following evaluation metrics: accuracy, f1–score, sensitivity and specificity. In the end, the average and standard deviation values are calculated for each of the metrics, obtaining the result that represents the model's generalization.

5.5. Explainability

We propose a methodology to provide a local explanation of the black box model using a single decision tree. In this step, we performed the following experiments:

1. Select a test instance for local explainability;
2. Generate new samples around the instance (noise set η);
3. Using the RF, classify the noise set and also the test instance;
4. The classification results are assigned as labels for these new samples;
5. With these labels and data, a DT is trained;
6. Then, the DT is used to provide a local explanation of the black-box model by taking the path in the tree that leads to the classification.

For global explanation, the local explanations obtained with DTX are aggregated over many instances to build the Criteria Graph (see section 4).

6. Results and Discussion

Table 3 shows the results for the classification of COVID-19 using the metrics accuracy, f1–score, sensitivity, specificity and area under the ROC curve (AUROC). We also summarize the classification results in the normalized confusion matrix per class (positive or negative) for each algorithm in Table 4.

Table 3: Results of the classification of COVID-19

Model/Score	Accuracy	F1–score	Sensitivity	Specificity	AUROC
LR	0.82 ± 0.03	0.71 ± 0.05	**0.73 ± 0.13**	0.84 ± 0.02	0.85 ± 0.05
RF	**0.88 ± 0.02**	**0.76 ± 0.03**	0.66 ± 0.10	0.91 ± 0.02	0.86 ± 0.05
XGBoost	0.87 ± 0.02	0.73 ± 0.03	0.60 ± 0.10	0.91 ± 0.02	0.85 ± 0.04
SVM	0.84 ± 0.02	0.70 ± 0.05	0.56 ± 0.14	0.89 ± 0.02	0.85 ± 0.05
MLP	0.85 ± 0.02	0.68 ± 0.06	0.42 ± 0.13	**0.92 ± 0.02**	0.81 ± 0.04
Ensemble	**0.88 ± 0.02**	**0.76 ± 0.03**	0.67 ± 0.10	0.91 ± 0.02	**0.87 ± 0.05**

Fig. 6 shows the average of the ROC curve obtained for each one of the algorithms evaluated. This curve is computed by varying the decision threshold, obtaining true positive and false positive rates for each of them. The closer the area is to 1, the greater the discrimination capacity of the model in the diagnostic test.

Using the f1–score for comparison, the best models obtained were the RF, with maximum tree depth equal to 8 and 45 estimators, and the heterogeneous ensemble. In both models we obtain an f1–score of 76%. Thus, prioritizing simplicity, we chose the RF model to apply our proposed Criteria Graph for the global explanations and

Table 4: Normalized confusion matrices for the ML methods tested. For each actual class, the sum of the corresponding row is 1.00

		Predicted	
		Negative	Positive
Actual	Negative	0.84	0.16
	Positive	0.26	0.74

(a) LR

		Predicted	
		Negative	Positive
Actual	Negative	0.91	0.09
	Positive	0.34	0.66

(b) RF

		Predicted	
		Negative	Positive
Actual	Negative	0.91	0.09
	Positive	0.40	0.60

(c) XGBoost

		Predicted	
		Negative	Positive
Actual	Negative	0.89	0.11
	Positive	0.44	0.56

(d) SVM

		Predicted	
		Negative	Positive
Actual	Negative	0.92	0.08
	Positive	0.58	0.42

(e) MLP

		Predicted	
		Negative	Positive
Actual	Negative	0.90	0.10
	Positive	0.35	0.65

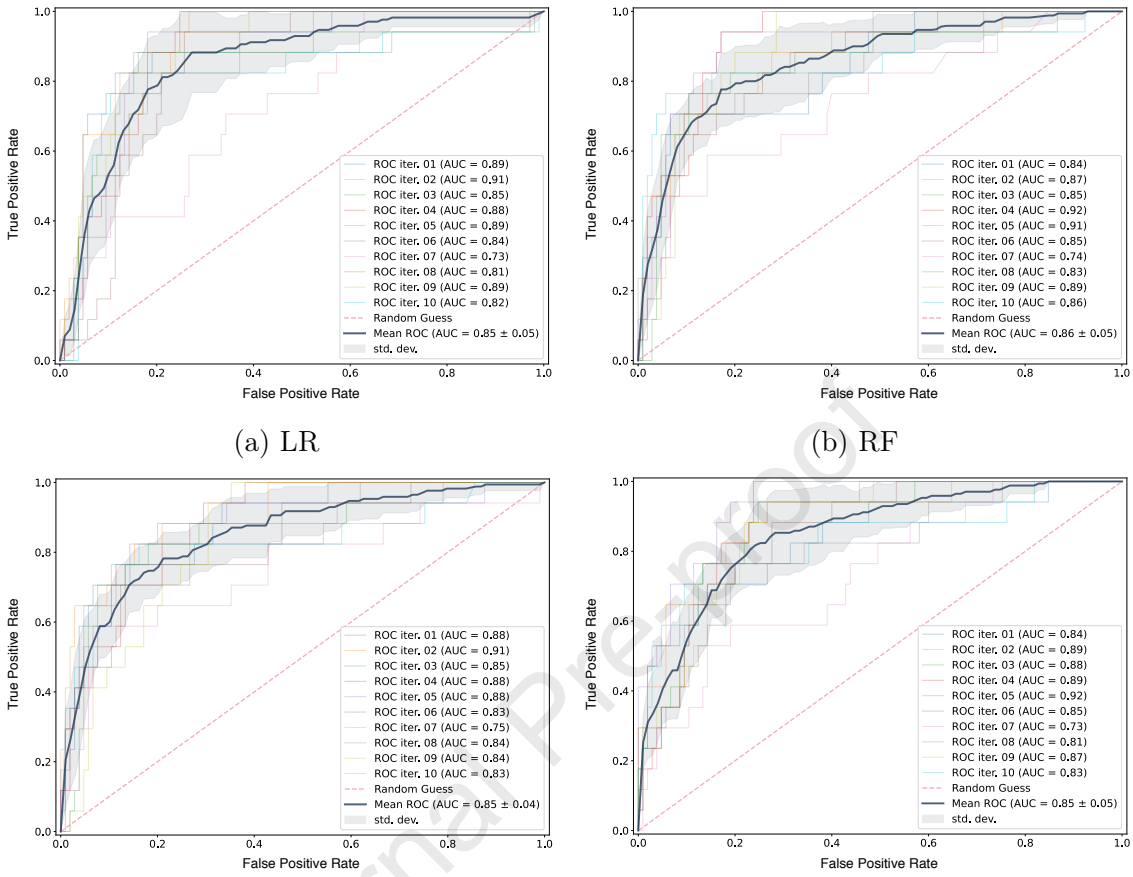
(f) Ensemble

the DTX for the local explanations. For the RF model, in 9 of the 10 iterations, the area under the curve ROC was ≥ 0.83 and the final average was equal to 0.86.

Fig. 7a shows the importance of the blood features for the model decision using the global SHAP values, which reflects the positive or negative contributions of each feature to the model output. A positive SHAP value represents a positive contribution to the target variable, while a negative SHAP value represents a negative contribution. These importances are classified in a descending way, suggesting that the main features that contributed to the target variable are the WBC, PLT and the EOS.

In addition to this information, the coloring of the points on the chart is related to the normalized values of the blood parameters of the patient, such as the number of WBC. The closer to blue, the lower the value of the characteristic and the closer to pink, the higher its value. Thus, a low value of the number of WBC, as well as the number of PLT, seen in blue, tends to positively impact the positive COVID–19 output. To corroborate this result, Fig. 8 shows the kernel density estimate for each of these two variables, for visualizing the distribution of observations of SARS–CoV–2 exam result across the dataset. For WBC and PLT values there is a central tendency around normalized values lowest of these characteristics. This is consistent with the literature, that suggest that the platelet count may reflect the pathological changes of patients with COVID–19 [51]. This tendency is also observed for EOS and the eosinopenia, characterized low EOS levels, appear to be related to disease severity [52]. In the case of CRP, higher values of this marker tend to positively impact the positive COVID–19 output.

Fig. 7b and 7c show examples of local explanations for two different patients with



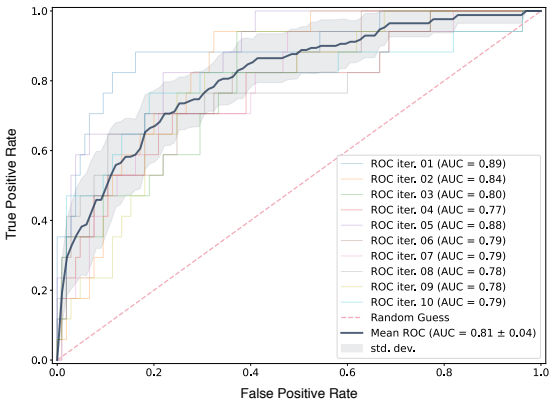
(a) LR (b) RF

(c) XGBoost (d) SVM

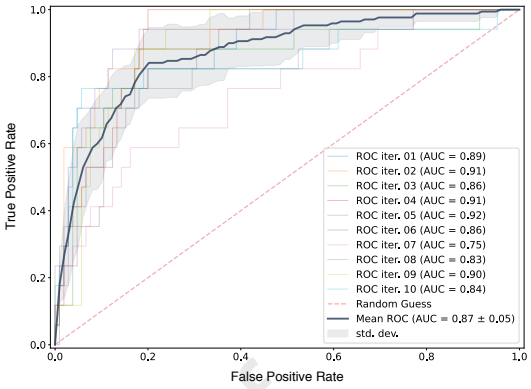
COVID-19, using the Local Interpretable Model–agnostic Explanations (LIME). This algorithm works by generating new samples around the instance to be explained and obtaining the prediction of the local noise using the original model. Then, based on the proximity to the given instance, the sample is weighted and a linear regression is constructed using these new samples and the considered instance. Through this method, the learned linear model is valid on a local scale.

The bars pointing to the right in Fig. 7b and 7c display the features that have a positive correlation with the output, while the bars on the left show the features that have a negative correlation. Thus, for the two patients, low WBC values (with WBC $\leq$ -0.64) and low EOS values (with EOS $\leq$ -0.67) have a positive correlation with the positive COVID–19, according to LIME explanations.

We also show the values of the first three most important blood parameters

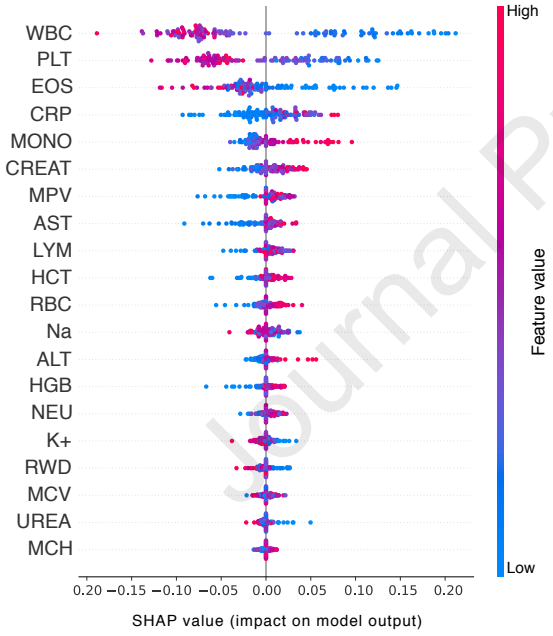


(e) MLP

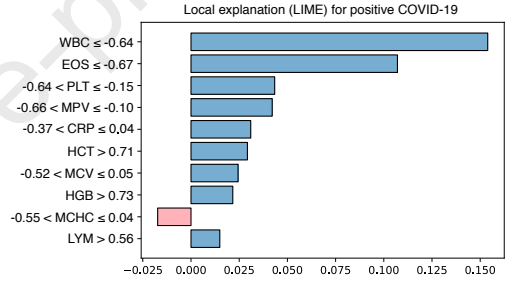


(f) Ensemble

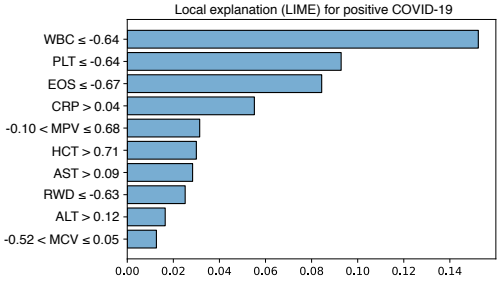
Figure 6: AUROC for each algorithm



(a) Global SHAP

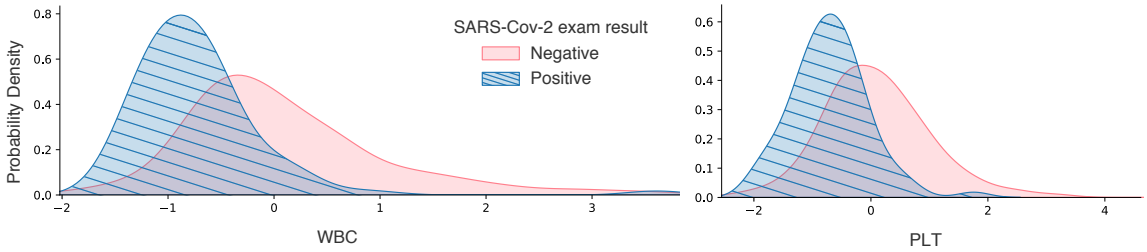


(b) Local explanation (LIME) for a positive COVID–19 individual



(c) Local explanation (LIME) for another positive COVID–19 individual.

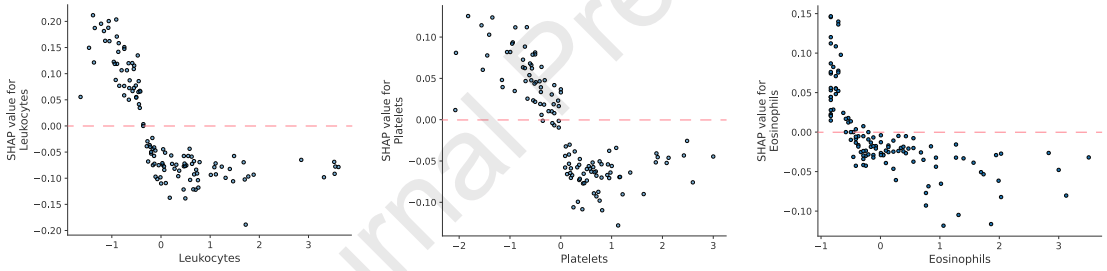
Figure 7: Explanations provided by SHAP and LIME



(a) WBC (b) PLT

Figure 8: Kernel density estimation of WBC and PLT

presented in Fig. 7a as a function of the corresponding SHAP value (Fig. 9), which represents the marginal effect that these features have on the predicted result of the model. Values of the normalized number of WBC, PLT and EOS above the highlighted lines, tend to contribute to increasing the probability of the positive class.



(a) WBC (b) PLT (c) EOS

Figure 9: Marginal effect of blood features on the target variable

6.1. Decision–Tree based Explainer and Criteria Graph

Table 5 shows the rules for the decision tree-based explainer for 12 positive COVID–19 patients which reflect the models behaviour. Since the explanations are local and built with high fidelity to the high accuracy model, differently from [8] and [21] one does not have to compromise accuracy. Also, the decisions trees allow us to represent non-linear behaviour which is an advantage against LIME.

It can be seen that the model uses different criteria to "diagnose" each patient. This indicates that the COVID-19 affects a number of parameters in the blood and that the variation of these parameters is individual dependent.

ID	Decision Tree Explanation
1	EOS $\leq$ -0.51 **and** PLT $\leq$ 0.16 **and** CRP $>$ -1.74 **and** EOS $\leq$ -0.61 **and** MPV $>$ -0.82 **and** NEU $\leq$ -0.42 **and** MCHC $\leq$ 1.90
2	CRP $>$ -0.43 **and** EOS $>$ 0.63 **and** AST $\leq$ -0.41 **and** UREA $>$ -0.91 **and** MCV $>$ 0.12 **and** CREAT $>$ -0.88 **and** K+ $>$ -0.52
3	EOS $>$ 0.54 **and** WBC $>$ -0.97 **and** MCV $\leq$ -0.13 **and** ALT $\leq$ 2.13 **and** CRP $>$ -0.51 **and** PLT $>$ -2.98 **and** HGB $>$ 0.96 **and** Sodium $>$ 0.12
4	AST $>$ -0.43 **and** CRP $>$ -0.46 **and** PLT $\leq$ 0.26 **and** WBC $\leq$ -0.44 **and** LYM $>$ -1.29 **and** EOS $\leq$ 0.76 **and** CREAT $>$ -0.75 **and** PLT $\leq$ 0.06 **and** AST $>$ -0.37 **and** PLT $>$ -3.57
5	EOS $>$ -0.59 **and** CRP $>$ -0.53 **and** PLT $\leq$ -0.33 **and** CREAT $>$ -0.30 **and** AST $>$ -0.34 **and** EOS $\leq$ 0.39 **and** MONO $>$ -0.49 **and** WBC $\leq$ -0.58 **and** LYM $\leq$ 1.11 **and** HCT $>$ -0.50
6	CRP $>$ -0.50 **and** MPV $>$ -0.99 **and** EOS $\leq$ 0.82 **and** PLT $\leq$ 0.21 **and** LYM $>$ -1.17 **and** CREAT $>$ -0.64 **and** EOS $\leq$ 0.37 **and** WBC $\leq$ -0.40 **and** HGB $\leq$ 0.44 **and** PLT $>$ -4.22
7	CRP $>$ -0.52 **and** PLT $\leq$ 0.08 **and** EOS $\leq$ -0.07 **and** HGB $>$ -0.83 **and** CREAT $>$ -0.87 **and** EOS $\leq$ -0.67 **and** RBC $>$ -1.02 **and** MONO $>$ -0.16
8	HGB $>$ -0.83 **and** LYM $>$ -1.13 **and** CRP $>$ -0.47 **and** CREAT $>$ -0.48 **and** HCT $>$ -1.09 **and** EOS $\leq$ 0.77 **and** AST $>$ 0.23 **and** MCV $>$ -6.31 **and** WBC $\leq$ -0.88 **and** PLT $\leq$ -0.05
9	EOS $\leq$ -0.59 **and** PLT $\leq$ -0.08 **and** MPV $>$ -1.00 **and** HCT $>$ 0.48 **and** UREA $\leq$ 2.35 **and** WBC $\leq$ -1.04 **and** MPV $>$ -0.97 **and** MCHC $>$ -1.08
10	EOS $\leq$ -0.55 **and** PLT $\leq$ 0.13 **and** MPV $>$ -1.02 **and** WBC $\leq$ 0.09 **and** PLT $\leq$ -0.11 **and** ALT $>$ -1.13 **and** WBC $\leq$ -0.29 **and** MONO $>$ -0.28
11	EOS $>$ -0.62 **and** AST $>$ -0.46 **and** EOS $\leq$ 0.52 **and** WBC $\leq$ -0.47 **and** CREAT $>$ -0.46 **and** CRP $>$ -0.68 **and** PLT $\leq$ -0.04 **and** MONO $>$ -0.03 **and** MCH $>$ -1.77 **and** AST $\leq$ 1.04
12	PLT $\leq$ 0.10 **and** MPV $>$ -1.04 **and** EOS $\leq$ -0.54 **and** MPV $>$ -1.01 **and** WBC $\leq$ -0.58 **and** LYM $>$ -1.48 **and** MCH $>$ -1.48 **and** HGB $>$ 1.07 **and** ALT $>$ -0.54 **and** MCHC $>$ -0.22

Table 5: Explanations for the COVID-19 inference of the 12 COVID-19 positive patients in the test set.

Looking at the set of rules it is hard to identify patterns that can be important in the search for a more universal and robust diagnosis methodology. For these reasons, the criteria graph (Section 4) was built for the explanations described in Table 5 and can be visualized in Figure 10. Different from LIME [26] and SHAP [27], the Criteria Graph not only shows the importance of the features (the area of the nodes which they represent) but also how the features are inter-connected.

The five largest nodes in the criteria graph correspond to PLT, MPV, EOS, CRP, and AST. Meanwhile, the 5 most important attributes according to their SHAP values (see Fig. 7a) were WBC, PLT, EOS, CRP and MONO. Notice that there

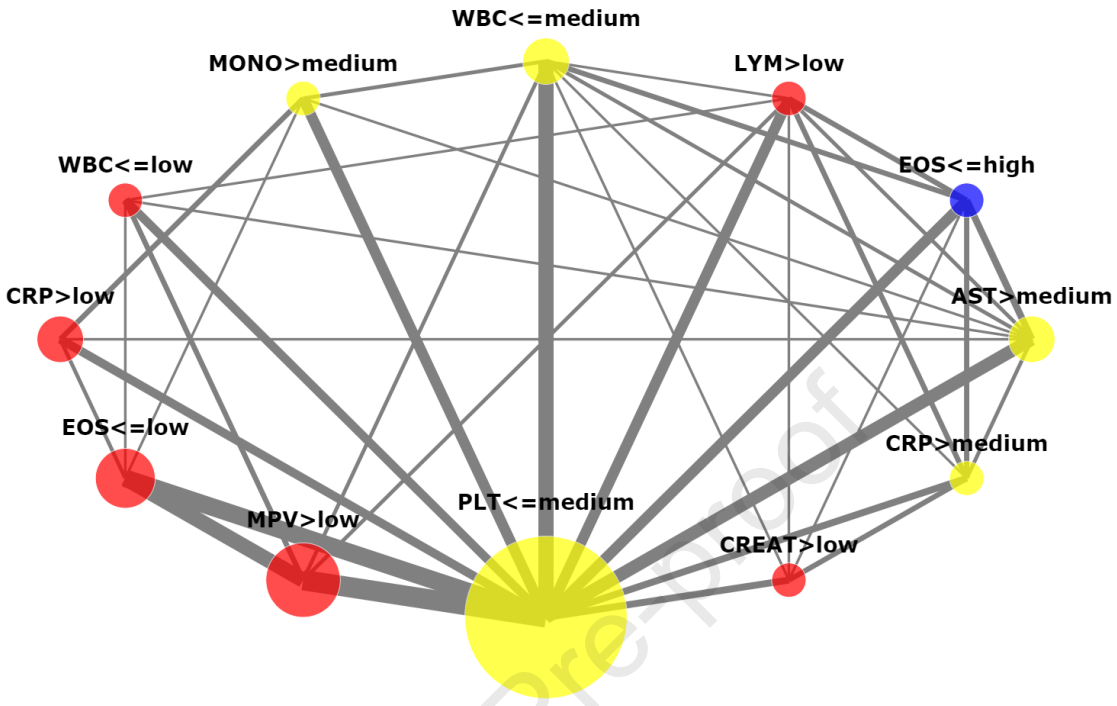


Figure 10: Criteria Graph for the decision tree explanations. Only factors and interactions that appeared in more than one third of the patients are depicted.

is a lot of overlap between the two rankings. Although WBC does not figure in the top five attributes, it has two nodes in the graph. That means that the WBC was important for the inference but its threshold value was not very clear. Thus, it seems to make sense that as a whole the attribute loses strength. The graph also shows a strong relationship between the criteria $PLT \leq Medium$, $MPV > low$ and $EOS \leq low$ pointing to a route towards a more reliable diagnosis procedure.

Increasing the number of patients used to produce the graph may increase the strength of the identified patterns. Nevertheless, the Criteria Graph provides information that other explanation methods lack and that this information may be extremely useful for the application expert. For instance, neither SHAP [27] nor LIME [26] present information about features interactions.

In Figures 7b and 7c it can be seen that LIME presents information about the thresholds used in the classification. However, as it happens with the DTX, the information is only local (individual dependent). The criteria graph addresses this drawback by aggregating the results of all the explanations.

SHAP can inform the user about possible feature thresholds with the marginal effects plot as shown in Figure 9. Such approach can be cumbersome if the number of features is high. In this context, the criteria graph is able to more clearly show the robustness of the thresholds by compressing the information about each feature in few nodes which are all displayed in a single plot. Thus, the amount artefacts presented to the user is reduced which tends to reduce the analysis time.

6.2. Practical Application

As aforementioned stated, the RF and heterogeneous ensemble models achieved the best results. Looking for the simplest model (often called parsimony), we follow with the RF as the preferred one plus the Criteria Graph for global explanations and DTX for local ones. Utilizing a web application, the healthcare professional may be able to input the patient's blood test results (similar, for instance, with that available in [13]). The system may be able, for instance, (i) to provide for the decision-maker both the results (infected or not), (ii) shows the rules to facilitate her/his valuable interpretation regarding local and global explanations, (iii) to be pre-configured to streamline the medical work and provide faster and more reliable diagnostics and (iv) offer intelligent prescription, which can be filled automatically in the correct standards of the medical prescription. The implementation must be focused on reusing the code, since once new strains of the virus are appearing, adaptations in the code/system may be required to make it useful in the future.

There are many advantages of using electronic medical records, such as security and availability of patient information, standardization/integration of data, and automation of procedures, to name a few. We know that SARS-CoV-2 is highly transmissible and rapid tests are already in place to diagnose the disease. Therefore, we emphasize that the proposed solution has the objective of supporting the decision making of clinicians, providing more information for helping them. Moreover, a considerable differential of the proposed methodology is the presentation of explanations of the model, making such information comprehensible to the health professional, being able to assist her/him in the final result of the diagnosis.

7. Conclusion

Recent research suggests that some parameters assessed in routine blood tests are indicative of COVID-19. It is well known that machine learning techniques excel in finding correlations in all sorts of data. Thus, it seems natural to try these techniques for the problem of COVID-19 screening through routine blood test data. However, there is significant barrier to the application of such methods in the real world due

to their lack of transparency, meaning that human specialists may find it difficult to trust the ML decisions.

In this context, in this work, we search for an accurate machine learning model for COVID-19 screening based on hematological data and propose two methods to improve the interpretability of the ML decisions, a Decision Tree Explainer and a Criteria Graph. The decision Tree Explainer is used to provide an individual explanation for each classified sample in terms of *If ... then* rules. The Criteria Graph is used to aggregate the set of rules produced by the decision tree to provide a global picture of the criteria that guided the model decisions and show the interactions among these criteria.

From the tested ML techniques, the best results were obtained with a RF which is an opaque model. It presented an accuracy of 0.88 $\pm$ 0.02, F1–score of 0.76 $\pm$ 0.03, Sensitivity of 0.66 $\pm$ 0.10, and Specificity 0.91 $\pm$ 0.02. The Decision Tree was then used to produce explanations for the classification of twelve confirmed COVID-19 cases and finally, the Criteria Graph was used to aggregate the explanations and portrait a global picture of the model results. The obtained Criteria Graph was in accordance with the well know techniques for interpretability SHAP and LIME indicating its adequacy and the adequacy of the Decision Tree Explainers. In addition, it could be seen that the Criteria Graph presents valuable information, such as the interaction among different criteria and the robustness of a criteria with respect to its threshold value, which is not provided by other techniques.

Given the urgency of the pandemic and the need to generate immediate results, much of the research has been published in repositories such as arXiv or medRxiv. Some methodologies discussed in the literature review are not clear enough to be reproducible or the model decision is not comprehensible. Lastly, we made comparisons between our proposed work and others from the literature that have not been peer-reviewed and published yet in the scientific literature. However, their data confirm our finding that ML models using routine blood parameters are useful in the diagnosis of COVID-19.

7.1. Future work

We employed hematological data from the Hospital Israelita Albert Einstein in São Paulo, Brazil, which is available as public data. However, this data is arguably not large and it is normalized (using z-normalization). Since we do not have access to the values used to normalize the data, the original values of the features are not accessible. Applying the proposed methods with larger data is an important step in our future work.

Still, the solution we offer brings good results, it is reproducible and the model

explainable. Additionally, we intend to integrate it with other fronts, such as chest X-rays and CT scans. In this way, ML models may serve as a way to support the diagnosis of the disease, regardless of the stage of contagion, and can help in the validation of RT-PCR.

Declaration of interests

The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

M.A. Alves, G.Z. de Castro and B.A.S. Oliveira declare that this work has been supported by the Brazilian agency CAPES (Coordination of Improvement of Higher Education Personnel).

J.A. Ramírez, R. Silva and F.G. Guimarães declare that this work has been supported by the Brazilian agencies CAPES, CNPq (National Council for Scientific and Technological Development), and FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais).

[1] World Health Organization, Coronavirus disease (covid-19) pandemic, 2020. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

[2] K. Zimmermann, J. W. Mannhalter, Technical aspects of quantitative competitive pcr, BioTechniques 21 (1996) 268–279. doi:10.2144/96212rv01.

[3] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, L. Xia, Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases, Radiology (2020) 200642. doi:10.1148/radiol.2020200642.

[4] Z. Meng, M. Wang, H. Song, S. Guo, Y. Zhou, W. Li, Y. Zhou, M. Li, X. Song, Y. Zhou, et al., Development and utilization of an intelligent application for aiding covid-19 diagnosis, medRxiv (2020).

[5] J. Bullock, K. H. Pham, C. S. N. Lam, M. Luengo-Oroz, et al., Mapping the landscape of artificial intelligence applications against covid-19, arXiv preprint arXiv:2003.11336 (2020).

[6] D. Ferrari, A. Motta, M. Strollo, G. Banfi, M. Locatelli, Routine blood tests as a potential diagnostic tool for covid-19, Clinical Chemistry and Laboratory Medicine (CCLM) 1 (2020). doi:10.1515/cclm-2020-0398.

[7] J. P. Kanne, B. P. Little, J. H. Chung, B. M. Elicker, L. H. Ketai, Essentials for radiologists on covid-19: an update—radiology scientific expert panel, Radiology (2020). doi:10.1148/radiol.2020200527.

[8] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, F. Cabitza, Detection of covid-19 infection from routine blood exams with machine learning: a feasibility study, Journal of medical systems 44 (2020) 1–12. doi:10.1007/s10916-020-01597-4.

[9] A. F. d. M. Batista, J. L. Miraglia, T. H. R. Donato, A. D. P. Chiavegatto Filho, Covid-19 diagnosis prediction in emergency care patients: a machine learning approach, medRxiv (2020).

[10] M. Kermali, R. K. Khalsa, K. Pillai, Z. Ismail, A. Harky, The role of biomarkers in diagnosis of covid-19–a systematic review, Life Sciences (2020) 117788. doi:10.1016/j.lfs.2020.117788.

[11] M. Kukar, G. Gunčar, T. Vovko, S. Podnar, P. Černelč, M. Brvar, M. Zalaznik, M. Notar, S. Moškon, M. Notar, Covid-19 diagnosis by routine blood tests using machine learning, arXiv preprint arXiv:2006.03476 (2020).

[12] H.-Y. Zheng, M. Zhang, C.-X. Yang, N. Zhang, X.-C. Wang, X.-P. Yang, X.-Q. Dong, Y.-T. Zheng, Elevated exhaustion levels and reduced functional diversity of t cells in peripheral blood may predict severe progression in covid-19 patients, Cellular & molecular immunology 17 (2020) 541–543.

[13] J. Wu, P. Zhang, L. Zhang, W. Meng, J. Li, C. Tong, Y. Li, J. Cai, Z. Yang, J. Zhu, et al., Rapid and accurate identification of covid-19 infection through machine learning based on clinical available blood test results, medRxiv (2020). doi:10.1101/2020.04.02.20051136.

[14] V. A. d. F. Barbosa, J. C. Gomes, M. A. Santana, J. E. Almeida Albuquerque, R. G. Souza, R. E. Souza, W. P. Santos, Heg. ia: An intelligent system to support diagnosis of covid-19 based on blood tests, medRxiv (2020).

[15] V. A. d. F. Barbosa, J. C. Gomes, M. A. Santana, C. L. Lima, et al., Covid-19 rapid test by combining a random forest based web system and blood tests, medRxiv (2020).

[16] N. Zhang, R. Zhang, H. Yao, H. Xu, M. Duan, T. Xie, J. Pan, J. Huang, Y. Zhang, X. Xu, et al., Severity detection for the coronavirus disease 2019 (covid-19) patients using a machine learning model based on the blood and urine tests, SSRN: https://ssrn.com/abstract=3564426 (2020). doi:http://dx.doi.org/10.2139/ssrn.3564426.

[17] S. L. d. O. Toledo, L. S. Nogueira, M. d. G. Carvalho, D. R. A. Rios, M. d. B. Pinheiro, Covid-19: Review and hematologic impact, Clinica Chimica Acta 510 (2020) 170–176. doi:10.1016/j.cca.2020.07.016.

[18] E. C. Silveira, Prediction of covid-19 from hemogram results and age using machine learning, Frontiers in Health Informatics 9 (2020) 39. doi:10.30699/fhi.v9i1.234.

[19] E. C. Silveira, Prediction of covid-19 from hemogram results and age using machine learning, Iranian Journal of Medical Informatics 9 (2020).

[20] R. Mardani, A. A. Vasmehjani, F. Zali, A. Gholami, S. D. M. Nasab, H. Kaghazian, M. Kaviani, N. Ahmadi, Laboratory parameters in detection of covid-19 patients with positive rt-pcr: a diagnostic accuracy study, Archives of Academic Emergency Medicine 8 (2020).

[21] L. Yan, H.-T. Zhang, Y. Xiao, M. Wang, C. Sun, J. Liang, et al., Prediction of criticality in patients with severe covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in wuhan, MedRxiv (2020). doi:10.1101/2020.02.27.20028027.

[22] S. Bhandari, A. S. Shaktawat, A. Tak, B. Patel, J. Shukla, S. Singhal, K. Gupta, J. Gupta, S. Kakkar, A. Dube, et al., Logistic regression analysis to predict mortality risk in covid-19 patients from routine hematologic parameters, Ibnosina Journal of Medicine and Biomedical Sciences 12 (2020) 123.

[23] B. M. Henry, G. Aggarwal, J. Wong, S. Benoit, J. Vikse, M. Plebani, G. Lippi, Lactate dehydrogenase levels predict coronavirus disease 2019 (covid-19) severity and mortality: A pooled analysis, The American Journal of Emergency Medicine (2020). doi:10.1016/j.ajem.2020.05.073.

[24] L. A. Ferreira, F. G. Guimarães, R. Silva, Applying genetic programming to improve interpretability in machine learning models, in: 2020 IEEE Congress on Evolutionary Computation (IEEE CEC 2020), IEEE, 2020. doi:10.1109/CEC48606.2020.9185620.

[25] C. Molnar, Interpretable Machine Learning, 2019. URL: https://christophm.github.io/interpretable-ml-book/.

[26] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. doi:10.1145/2939672.2939778.

[27] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774.

[28] Y. Mohamadou, A. Halidou, P. T. Kapen, A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of covid-19, Applied Intelligence 50 (2020) 3913–3925. doi:10.1007/s10489-020-01770-9.

[29] P. C. Silva, P. V. Batista, H. S. Lima, M. A. Alves, F. G. Guimarães, R. C. Silva, Covid-abs: An agent-based model of covid-19 epidemic to simulate health and economic effects of social distancing interventions, Chaos, Solitons & Fractals 139 (2020) 110088. doi:10.1016/j.chaos.2020.110088.

[30] L. Wynants, B. Van Calster, M. M. Bonten, G. S. Collins, T. P. Debray, M. De Vos, M. C. Haller, G. Heinze, K. G. Moons, R. D. Riley, et al., Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal, BMJ 369 (2020). doi:10.1136/bmj.m1328.

[31] Z. Zheng, F. Peng, B. Xu, J. Zhao, H. Liu, J. Peng, Q. Li, C. Jiang, Y. Zhou, S. Liu, et al., Risk factors of critical & mortal covid-19 cases: A systematic literature review and meta-analysis, Journal of Infection (2020). doi:10.1016/j.jinf.2020.04.021.

[32] W. Tian, W. Jiang, J. Yao, C. J. Nicholson, R. H. Li, H. H. Sigurslid, L. Wooster, J. I. Rotter, X. Guo, R. Malhotra, Predictors of mortality in hospitalized covid-19 patients: A systematic review and meta-analysis, Journal of Medical Virology 92 (2020) 1875–1883. doi:10.1002/jmv.26050.

[33] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, D. Shen, Review of artificial intelligence techniques in imaging data acquisition, segmentation

and diagnosis for covid-19, IEEE Reviews in Biomedical Engineering (2020). doi:10.1109/RBME.2020.2987975.

[34] D. Assaf, Y. Gutman, Y. Neuman, G. Segal, S. Amit, S. Gefen-Halevi, N. Shilo, A. Epstein, R. Mor-Cohen, A. Biber, et al., Utilization of machine-learning models to accurately predict the risk for critical covid-19, Internal and emergency medicine (2020) 1–9. doi:10.1007/s11739-020-02475-0.

[35] L. Dean, L. Dean, Blood groups and red cell antigens, volume 2, NCBI Bethesda, Md, USA, 2005.

[36] NHSFoundation, Full blood count (fbc) reference ranges, 2020. URL: https://www.yorkhospitals.nhs.uk/seecmsfile/?id=2396.

[37] M. Xiuli Ding, M. Geqing Xia, M. Zhi Geng, Z. Wang, L. Wang, A simple laboratory parameter facilitates early identification of covid-19 patients, medRxiv (2020). doi:https://doi.org/10.1101/2020.02.13.20022830.

[38] A. Farhana, S. L. Lappin, Biochemistry, lactate dehydrogenase (ldh), in: StatPearls [Internet], StatPearls Publishing, 2020.

[39] M. A. Lichtman, K. Kaushansky, J. T. Prchal, M. M. Levi, L. J. Burns, J. Armitage, Williams manual of hematology, McGraw Hill Professional, 2017.

[40] H. Demirin, H. Ozhan, T. Ucgun, A. Celer, S. Bulur, H. Cil, C. Gunes, H. A. Yildirim, Normal range of mean platelet volume in healthy subjects: Insight from a large epidemiologic study, Thrombosis research 128 (2011) 358–360. doi:10.1016/j.thromres.2011.05.007.

[41] K. D. Pagana, T. J. Pagana, Mosby's Diagnostic and Laboratory Test Reference-E-Book, Elsevier Health Sciences, 2012.

[42] A. Rastegar, Serum potassium, in: Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition, Butterworths, 1990.

[43] Hospital Israelita Albert Einstein, Diagnosis of covid-19 and its clinical spectrum - ai and data science supporting clinical decisions (from 28th mar to 3st apr), https://www.kaggle.com/einsteindata4u/covid19, 2020. Online; accessed 10 June 2020.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.

[45] D. W. Hosmer Jr, S. Lemeshow, R. X. Sturdivant, Applied logistic regression, volume 398, John Wiley & Sons, 2013.

[46] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794. doi:10.1145/2939672.2939785.

[47] T. K. Ho, Random decision forests, in: Proceedings of 3rd international conference on document analysis and recognition, volume 1, IEEE, 1995, pp. 278–282. doi:10.1109/ICDAR.1995.598994.

[48] G. C. Cawley, N. L. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, Journal of Machine Learning Research 11 (2010) 2079–2107.

[49] L. A. Jeni, J. F. Cohn, F. De La Torre, Facing imbalanced data–recommendations for the use of performance metrics, in: 2013 Humaine association conference on affective computing and intelligent interaction, IEEE, 2013, pp. 245–251. doi:10.1109/ACII.2013.47.

[50] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357. doi:10.1613/jair.953.

[51] X. Zhao, K. Wang, P. Zuo, Y. Liu, M. Zhang, S. Xie, H. Zhang, X. Chen, C. Liu, Early decrease in blood platelet count is associated with poor prognosis in covid-19 patients—indications for predictive, preventive, and personalized medical approach, The EPMA Journal (2020) 1. doi:10.1007/s13167-020-00208-z.

[52] Y. Sun, Y. Dong, L. Wang, H. Xie, B. Li, C. Chang, F.-s. Wang, Characteristics and prognostic factors of disease severity in patients with covid-19: The beijing experience, Journal of autoimmunity 112 (2020) 102473. doi:10.1016/j.jaut.2020.102473.

Highlights of the paper:

- A literature review of ML methods applied to COVID-19 screening in routine blood tests
- Results from different ML techniques – including an ensemble – to support the diagnosis of COVID-19 using usual blood exams
- A decision tree–based methodology for the explanation of the model which can be given to the health teams
- A methodology for aggregating the individual explanations in a graph that shows the relative importance of each attribute and their interactions
- Further evidence that simple blood tests might help identifying false positive/negative RT-PCR tests

Conflicts of Interest Statement

Manuscript title: Explaining Machine Learning based Diagnosis of COVID-19 from Routine Blood Tests with Decision Trees and Criteria Graphs

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.