Journal of Control, Automation and Electrical Systems (2022) 33:625–644
https://doi.org/10.1007/s40313-021-00858-y

Explainable Artificial Intelligence for COVID-19 Diagnosis Through Blood Test Variables

Lucas M. Thimoteo[1] · Marley M. Vellasco[1] · Jorge Amaral[2] · Karla Figueiredo[3] · Cátia Lie Yokoyama[4] · Erito Marques[2]

Received: 30 April 2021 / Revised: 1 October 2021 / Accepted: 20 October 2021 / Published online: 3 January 2022

Abstract

This work proposes an explainable artificial intelligence approach to help diagnose COVID-19 patients based on blood test and pathogen variables. Two glass-box models, logistic regression and explainable boosting machine, and two black-box models, random forest and support vector machine, were used to assess the disease diagnosis. Shapley additive explanations were used to explain predictions for the black-box models, while glass-box models feature importance brought insights into the most relevant features. All global explanations show the eosinophils and leukocytes, white blood cells are among the essential features to help diagnose the COVID-19. Moreover, the best model obtained an AUC of 0.87.

Keywords COVID-19 diagnosis · Machine learning · Explainability · Interpretability · Shapley additive explanations · Explainable boosting machine

1 Introduction

The Coronavirus Disease, better known as COVID-19, has gained humanity's attention since the first reports of its occurrences in Wuhan, Hubei province, China (Andersen et al. 2020) in late 2019. In March 2020, the World Health Organization (WHO) declared the COVID-19 to be a pandemic outbreak (WHO 2020). Until July 2021, the COVID-19 has affected over 190 million people in the world (WHO 2021).

Due to the fact of COVID-19 highly contagious nature, governments around the globe implemented policies for social distancing, quarantine, and eventually lockdowns. These events brought severe health and economic challenges to several countries, and Brazil was no exception. There is an urgent need to test the majority of the population to assist federal and local government's decision-making. Due to limited resources, COVID-19 tests are restricted to healthcare professionals and people with severe conditions, hence not reaching the vast majority of the population.

Since the 1990s, diagnostic medicine has been taking advantage of Machine Learning (Singh et al. 2019). Advances in technology provided faster, easier and reliable ways of applying machine learning in several fields of medicine, such as diagnostic of respiratory diseases (Amaral et al. 2012), diabetes (Zou et al. 2018) and cancer (Liu et al. 2017). However, in areas such as Finance, Government, and

✉ Lucas M. Thimoteo
lucasthim@yahoo.com

Marley M. Vellasco
marley@ele.puc-rio.br

Jorge Amaral
jamaral@uerj.br

Karla Figueiredo
karlafigueiredo@ime.uerj.br

Cátia Lie Yokoyama
catialie@gmail.com

Erito Marques
mederitomarques@gmail.com

[1] Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ, Brasil

[2] Programa de Pós-Graduação em Engenharia Eletrônica (PEL), Universidade do Estado do Rio de Janeiro, Rio de Janeiro, RJ, Brasil

[3] Programa de Pós-Graduação em Ciências Computacionais (CCOMP), Programa de Pós-Graduação em Telessaúde, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, RJ, Brasil

[4] Departamento de Biologia Geral, Universidade Estadual de Londrina, Londrina, PR, Brasil

Medicine, the need to explain and interpret the predictions of a model has become a pressing issue. Understanding the reasons behind a model prediction or understanding the model itself can guide a more reliable and trustworthy development of artificial intelligence in these areas (Tjoa and Guan 2019). Several interpretability methods have been proposed over the last few years. Some distinctive approaches are agnostic models. They can explain deep learning predictions regarding cancer diagnosis through images (de Sousa et al. 2019). Other methods use intrinsically interpretable models, such as a fuzzy model that can build a rule set to predict lung cancer with liquid biopsy variables (Potie et al. 2019).

This work is an extended version of the congress paper (Thimoteo 2020). It expands the study on how explainable artificial intelligence and interpretable machine learning can help diagnose COVID-19 through blood test and pathogen clinical variables. A larger dataset was collected from suspected cases from the Hospital Israelita Albert Einstein. Moreover, a new glass-box model is tested on this larger dataset, the explainable boosting machine. Therefore, we bring a total of two glass-box models: logistic regression and explainable boosting machine—and also two black-box models: random forests and support vector machines. The glass-box models can be interpreted globally and locally right after they are trained, and explanations are given for two black-box models using the Shapley additive explanations. The sections ahead are divided into COVID-19 Symptoms and Diagnosis, Methods, Experimental Assessment, Results, Interpretation of Results, and Conclusion.

2 COVID-19 Symptoms and Diagnosis

COVID-19 has varied clinical specifics, varying from asymptomatic to acute respiratory distress syndrome (ARDS). The most common clinical features of this disease are fever and cough (Huang et al. 2020). Loss of smell is also a strong indicator of COVID-19 infection. Other symptoms such as headache, rhinorrhea, sore throat, fatigue, chest pain and tightness were noted (Cascella et al. 2020). Atypical symptoms such as diarrhea, nausea, and even abdominal pain can also occur (Chakraborty et al. 2020). In a subset of patients, by the end of the first week the disease can progress to pneumonia, respiratory failure and death (Singhal 2020).

Rapid and accurate detection of COVID-19 is essential to control outbreaks in the community and in hospital. Current diagnostic tests for coronavirus include the reverse-transcription polymerase chain reaction (RT-PCR). In patients with confirmed COVID-19 diagnosis, the laboratory evaluation should be repeated to evaluate for viral clearance prior to being released from observation. However, the availability of testing will vary based on which country a person lives in with increasing availability occurring nearly daily (Zhai et al. 2020).

Serological testing for SARS-CoV-2 is widely available. This test is carried out using enzyme linked immunosorbent assay (ELISA), immunofluorescence (IFA) or, in case of limited lab capacity, rapid diagnostic tests (RDTs) (Krammer and Simon 2020). The RDT is becoming a crucial tool for the early diagnosis of SARS-CoV-2, particularly in situations where laboratory workloads are high and RT-PCR tests are in short supply (Porte et al. 2020). Despite the numerous antibody tests available to date, serologic diagnosis has limitation in both specificity and sensitivity. This diagnosis, however, can have an important role in broad-based surveillance of COVID-19 worldwide (Cascella et al. 2020).

Until now, there is no specific antiviral treatment that has been proven to be effective for COVID-19 (Pardi and Weissman 2020). The treatment is symptomatic, and oxygen therapy represents the first step for addressing respiratory impairment. Noninvasive (NIV) and invasive mechanical ventilation (IMV) may be required in cases of respiratory failure refractory to oxygen therapy (Huang et al. 2020) .

2.1 Blood Count Exam and Relationship with COVID-19

Monocytes and macrophages have been reported to play crucial roles in the immune pathogenesis upon virus infection (Maggi et al. 2020). Macrophage polarization during viral infections contributes to either antiviral responses or immune pathogenesis (Teijaro et al. 2014). Previous studies have also found increased monocyte-derived macrophages with an inflammatory phenotype, expressing serious gene important in cell adhesion and activation in severe COVID-19 patients (Liao et al. 2020; Merad and Martin 2020). Macrophages were also observed in the kidneys of patients with COVID-19, and acute kidney tubular damage was linked with marked accumulation of monocytes and macrophages (Mehta et al. 2020).

Also, it has been reported that the severity of pulmonary immune injury is correlated with extensive infiltration of neutrophils and macrophages in the lungs, followed by increased numbers of neutrophils and monocytes counts in the peripheral blood samples (Martinez et al. 2020b; Zhang et al. 2020). Common results among hospitalized patients with COVID-19 include lymphopenia, elevated inflammatory markers (e.g., erythrocyte sedimentation rate, C-reactive protein, ferritin, tumor necrosis factor-α, IL-1, IL-6), and abnormal coagulation parameters (e.g., prolonged prothrombin time, thrombocytopenia, elevated D-dimer, low fibrinogen) (Skevaki et al. 2020). Moreover, different acute phase biomarkers, like ferritin, and a hypercoagulability state, indicated by elevated D-dimer, have all been linked with illness severity and mortality.

Therefore, with the combination of few laboratory parameters, in particular D-dimer and lymphocyte with the detection and serial monitoring of inflammatory monocytes could be of great help in guiding the prognostication and treatment of patients with COVID-19 (Biamonte et al. 2021; Zhang et al. 2020).

3 Machine Learning for COVID-19 Diagnosis

Throughout the year 2020, several machine learning (ML)- and deep learning (DL)-based studies and solutions have been proposed to help the correct diagnosis of the COVID-19. DeCaprio et al. (2020) used machine learning algorithms to build a vulnerability index for COVID-19. Since datasets for COVID-19 were not readily available, the CV19 index was measured in terms of the short-term risk of developing serious complications from respiratory infections. Jiang et al. (2020) used several machine learning algorithms to predict, from the historical data of patients whose test was positive for coronavirus, identify attributes that may be indicative of the acute respiratory distress syndrome (ARDS). The results presented show that slightly elevated Alanine aminotransferase (ALT), myalgias (body pain), and elevated hemoglobin are the most predictive clinical factors of the patient's condition.

Several studies lie on the COVID-19 diagnosis through chest imaging, such as radiology or computer tomography (CT). Although accurate in correctly identifying positive cases, most of these tools do not address the issue of differentiating the COVID-19 from other pulmonary diseases (Ghaderzadeh and Asadi 2021). Moreover, studies show unrealistic optimal results, most likely consisting of data leakage problems and lack of clarity during experiment procedures, making their work not reproducible or market-ready. By analyzing Roberts et al. (2021) and Mohammad-Rahimi et al. (2021), it is clear that most of the approaches make use of imaging data and sophisticated deep learning and computer vision techniques. However, little to no importance is given model interpretability and prediction explanations to help doctors and field specialists.

Zoabi et al. (2021) presented a simple and super effective ML approach to diagnosing the COVID-19 using simple features based on a patient questionnaire. Among the utilized features were confirmed contact with an infected person, symptoms such and sore throat and fever, and also age. Results show an AUC measure of 0.9 in test sets, which is considered a high standard result in the medical area. Shapley additive explanations identified the most critical variables such as cough, fever and contact with an infected person. This approach has two significant limitations: the first one is not able to identify asymptomatic patients, and the second one is that it might be susceptible to bias toward false negatives. Since, as mentioned by the authors, underestimation of symptoms from negative tested patients might happen, this tool can certainly help field agents on early trials to better guide patient ward allocation and manage COVID-19 tests under scarcity conditions.

In Gangloff et al. (2021), authors show an exciting strategy to combine RT-PCR, CT images and clinical variables to increase the overall performance on identifying the COVID-19. The primary motivation for this work was that CT images alone are not reasonable indications, and the RT-PCR test still fails on some false-negative cases. For that, they propose a framework that processed clinical variables such as blood test variables, symptoms and blood pressure, along with CT and RT-PCR tests. Simple feature importance was made over logistic regression feature weights, and authors state that their findings are in accordance with medical studies. Results show an AUC measure of 0.778 (CI 95% 0.682–0.873) when using only clinical variables, 0.852 (CI 95% 0.764–0.940) when combining CT with clinical variables, and 0.930 (0.867–0.992) for the combination of RT-PCR and clinical variables. Although authors achieved high performance, using CT and RT-PCR to correctly diagnose patients might not be feasible under many circumstances. It is worth mention that they proposed that the trained model with clinical variables could be used to triage COVID-19 patients and ease the burden in the healthcare system.

Other papers focus on different aspects of the patient, such as predicting the mortality rate to redirect patients to intensive care units beforehand (de Moraes Batista et al. 2020) and also monitoring the respiratory functions and body temperature in order to identify abnormalities. This last work is only based on an automation setup and a set of rules. It shows the potential of incorporating ML-based predictions to identify early anomalies to increase patients' chances of survival in emergency calls.

Now coming to an epidemic-centered scope, we see works that employ several computational models in order to predict a short-term spreading rate on a local or global scale. In Jha et al. (2020), the authors show a Bayesian learning model for short-term prediction on infected and deceased people. They used data provided by the state government of Texas and divided the population into five fractions: susceptible, exposed, infected, recovered and deceased (SEIRD). The authors argue that the proposed model shows satisfactory results for predicting the number of deceased but might not be adequate for predicting infected cases due to the rapid acceleration in the spreading rate.

Also, Viguerie et al. (2021) propose a similar SEIRD model for the region of Lombardy (Italy), stating that their validation is adequate to predict spreading rate, taking into account different scenarios for restriction or relaxation of lockdown rules. Martinez et al. (2020a) proposes a simpler Holt's model to predict the short-term number of infected subjects in some states of Brazil. Their model falls short of predicting the increasing infection in the state of Rio de Janeiro. Moreover, other ways of helping in public health and healthcare policies are possible, such as the work that is shown in Zohdi (2020) which predicts infection zones by modeling and simulating coughing dynamics. This prediction might be particularly interesting for public spaces and areas nearby health facilities such as clinics and hospitals.

In conclusion, the number of studies proposed so far in such a short time after the beginning of the pandemic is impressive. However, all of them should be further studied and also applied to more diverse populations. The next section presents the methods used in our work.

4 Methods

Four models were chosen to assess the COVID-19 diagnosis problem. The first two are glass-box models; that is, the models are transparent and intelligible: the logistic regression and the explainable boosting machine. The other two are black-box models; that is, they are neither clear nor transparent about how they make predictions. These are: random forest (RF) and support vector machine (SVM).

To address the interpretability and explainability of the black-box models, Shapley additive explanations (SHAP) method will be used. For the logistic regression and EBM, feature weights can be simply visualized and evaluated.

4.1 Logistic Regression

The logistic regression (LR) is one of the most used models for classification tasks over the last decades (Fan et al. 2019). One major advantage of this model is its intrinsically interpretable nature, since it projects the input features into linear a regression problem and thus can be seen as a glass-box model. The decision of the model is given by:

$$p(t) = \frac{1}{1 + e^{-t}} \tag{1}$$

$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdot + \beta_n x_n \tag{2}$$

If for a given feature x_1 there is a β_1 weight, for another feature x_2 there is a weight β_2 and β_1 is ten times greater than β_2, it is reasonable to say that the feature x_1 is considerably more important to the decision of the model. That is, x_1 will have ten times more impact than x_2. Therefore, it is possible to make sense of a global feature importance by analyzing the values of the β feature weights. One limitation of this model is that due to its linear nature, local explanations cannot be directly assessed, that is because the model adjusts the feature importances as being the same for the entire features distribution.

4.2 Explainable Boosting Machine

The explainable boosting machine (EBM) proposed in Nori et al. (2019) is a glass-box-type model that leverages state-of-the-art performing algorithms, such as boosting and bagging, while maintaining intrinsic interpretability. The main idea of this model is to build a generalized additive model with interactions (GA^2M) by executing a round-robin training procedure on one feature at a time, using a meager learning rate. Its general is given in the form of:

$$g(E[y]) = \beta_0 + \Sigma f_j(x_j) + \Sigma f_{ij}(x_i, x_j) \tag{3}$$

where g is the link function and f_j and f_{ij} are called smooth functions. This approach brings two major improvements when compared to the original GAM (Hastie and Tibshirani 1986).

The first one is its potential to capture feature interaction importance: the compound effect of two features that become relevant to the model. In real-world scenarios, feature interactions are not only present, but many times they are a constraint of the problem, i.e., a change made in one feature impacts the value of other features. This constraint is often not considered in many explainability algorithms since most of them rely on local linear approximations. The ability to capture the interactions between two features contributes to enhancing the model's accuracy while keeping it explainable. Two-dimension interactions can be explored as heatmaps $f_{ij}(x_i, x_j)$; as a result, a model that presents only one- and two-dimensional components is still explainable.

The second improvement is FAST, which is an extremely efficient tool for measuring and ranking the strength of all pairs of variables' interactions. EBM is a fast implementation of the Lou et al. (2013).

Due to the additive nature of this model, it is easy to evaluate the contribution of each feature or pairwise interaction. Therefore, it is possible to achieve both global interpretation and local explanations with this method. The local explanation can be assessed by computing the result of each smooth function, given the values of the features. It is also possible to achieve global interpretation by taking the mean of the smooth functions, providing a result similar to the logistic regression feature weights.

4.3 Random Forest

The random forest classifier is random forest (RF) that is an ensemble strategy that assembles and compounds several base decision trees (Breiman 2001). It can be used both for regression and for classification problems. In the former, RF outputs the class that is the majority of the class' output by individual trees, while in the later, it presents the mean of the individual tree's results. It employs the bootstrap aggregation (bagging) that helps to alleviate the variance by calculating the average of many decision trees, which present low bias, and at the same time, they are still able to capture complex interaction structures in data. Breiman (1996) also observed that in an ensemble of decision trees, the trees were deeply correlated, as long as, in the tree-growing procedure, the algorithm could select any of the available features. Hence, the random selection of features decreases the correlation between the trees. Notably, in the process of building an individual tree on a bootstrapped dataset, before each split, a subset of $m \leq p$ of the p input features is designated at random as candidates for inferring the best split of the training set.

Random forest is fast to train and execute and achieves state-of-the-art performance. It can handle high-dimension input vectors and offers an internal assessment of the generalization error as forest-building advances. In addition, another critical virtue is the capacity to build feature importance plots. At each split in each tree, the algorithm records the improvement in the split criterion as an importance score associated to the splitting variable. These importance scores are compiled over all the trees in the forest separately for each input and can also be employed as a feature selection method by choosing the features with higher scores in the importance plots.

4.4 Support Vector Machine

The basic principles from which support vector machines (SVM) were conceived were established by statistical learning theory (Vapnik 2000) . Considering a binary linearly separable classification problem, SVM provides a decision boundary that is hyperplane with optimal geometric margin from the classes, which, in turn, presents the highest generalization capacity. This conception can be extended to a nonlinear separable problem by applying an artifice is called a “kernel trick.” There is a wide variety of kernel functions, in order to explore different linear and nonlinear relationships, for example, polynomial, Gaussian and hyperbolic. This scheme transforms the data into a new high-dimensional space, where one expects the classes to be effortlessly separable. Albeit the decision surface is a hyperplane in the high-dimensional space, when it is examined in the primary feature space, it is no longer a hyperplane, indicating that SVM can also be employed when data that are not linearly separable. In order to address the issue of patterns that are not so easily separable, the SVM can have a soft-margin implemented. That is, as the sample moves across the decision boundary, a loss function assigns an uncertainty value to its prediction (Chang and Lin 2011).

4.5 Shapley Additive Explanations

In order to address the interpretability of models, Lundberg and Lee (2017) proposed SHapley Additive exPlanations or SHAP, which generalizes the feature importance and combine earlier methods with game theory and local explanations to provide a mathematically rigorous, special and reliable additive feature attribution. SHAP explains the output of a model as a sum of the effects that each feature on the final conditional expectation. Averaging over all possible feature orderings and game-theoretic proofs are utilized to guarantee a consistent solution.

This method main advantage is its ability to explain any predictor by assigning features of samples a score (SHAP value), based on their participation in the prediction task. The explanation (Molnar 2019) for a prediction is given by:

$$g(x') = \phi_0 + \sum_{n=1}^{M} \phi_n x'_n \tag{4}$$

where g is the explanation model, $z' \in \{0, 1\}^M$ is the simplified feature vector, M is the size of the simplified vector, and $\phi_n \in R$ is the weight (SHAP value) for each feature n of each sample.

The SHAP values estimation steps are the following:

1. Randomly choose some features in samples and replace for a random value
2. Get the prediction for each modified sample
3. Compute the SHAP values with the SHAP Kernel

The estimation described before is based on random sampling, therefore the variance of the estimation will decrease as the number of samples increases. Besides, it makes an assumption of feature independence. If, for example, features x_1, x_2 are strongly correlated in our training data, substituting x_1 by random values ignores this dependency and produces predictions based on x_1, x_2 instances that are unlikely to appear in the training set, rendering SHAP value estimation less accurate.

A Kernel SHAP Explainer fits a linear model for every prediction of a given dataset. This Kernel aims to optimize the following loss equation:

$$L(f, g, \pi_x) = \sum [f(h_x(z')) - g(z')]^2 \pi_x(z') \tag{5}$$

The π_x term is a compliant weighting strategy that assigns higher values to predictions that depend on lesser features or at almost all of them (Molnar 2019).

The SHAP Tree Explainer can only be used with tree-based methods (Lundberg 2020). It uses the tree structure instead of random sampling to simulate missing features by simply ignoring decision paths that depend on the missing features. As a result, the TreeExplainer output is deterministic and does not change depending on the context dataset. Instead of iterating over each possible feature combination (or a subset thereof), all combinations are pushed through the tree at the same time, using a more complex algorithm to keep track of each combination's result—reducing complexity from $O(TL2^M)$ for all possible combinations to the polynomial $O(TLD^2)$, where M is the number of features, T is number of trees, L is maximum number of leaves, and D is maximum tree depth).

This is a faster way of calculating SHAP values, since no linear models is fitted. The SHAP values are calculated by the change in the conditional expectation of all features given a subset of features. In other words, this calculation ignores the nodes of features that are not present in the sample subset and calculate change on the conditional expectation, given this subset sample.

5 Experiments

The main steps of the experiments consisted of data acquisition, data pre-processing, model tuning, model evaluation, and interpretation of models and their predictions.

All the experiments ran in a free tier Google Colab environment, which consisted of an Intel Xeon 2.34Ghz quad-core processor, 25Gb of RAM, and Linux Ubuntu 18.04 operational system. The algorithms were implemented in Python 3.7 using PIP packages. Implementation details can be found here.

5.1 Data Acquisition

The experiment was executed with data provided by the COVID-19 Data Sharing/BR, which is an initiative of the São Paulo Research Foundation (FAPESP) in collaboration with the University of São Paulo. The dataset consisted of suspected COVID-19 patients admitted to the Hospital Israelita Albert Einstein from February 2020 to June 2020. Additionally, we confirmed that it is possible to have two samples from the same patient but from different days. It happens because a patient can be admitted more than once in the same unit, and also, a doctor can request a COVID-19 for the patient more than once while they are still under treatment.

It is important to note that data present itself with some limitations by its definition since it was built gathering only suspected cases of COVID-19. In most cases, people have had some symptoms or at least have come into direct contact with other infected people. Therefore, it is unlikely that asymptomatic cases can be found in this data source. Moreover, the Hospital Israelita Albert Einstein is a private hospital and located in a wealthy area of the city from São Paulo, which makes the access of lower-income families to be restricted. Raw data are available here.

5.2 Data Pre-processing

The raw dataset contains over 50 thousand patients containing several types of exams, such as blood tests, urine tests, influenza pathogen tests, other coronavirus pathogen tests, and several types of COVID-19 tests.

In order to make a fair comparison with our previous results (Thimoteo et al. 2020), we selected only features that were similar to the previous final dataset. These features are: age, sex, complete blood count variables (which is a type of blood test), other coronavirus pathogen diagnosis and other influenza pathogen diagnosis. When considering only blood test features, it was possible to obtain approximately 5800 samples with no missing data. After aggregating the pathogen features, this number dropped to approximately 1500. These two datasets are roughly 15 times and 4 times larger than obtained in (Thimoteo et al. 2020) since the previous dataset consisted of nearly 360 patients. Also, in contrast to the previous dataset, now we had access to the original value of the features, which makes our current datasets more reliable.

Literature mentions (Schmidt et al. 2015; Agor et al. 2019) that several trials containing missing values might be due to the physician coming to conclusion that certain exams are not necessary, because the patient shows no apparent anomalies. Despite having the reference values for all the features (values that are considered normal), we opted to avoid data imputation and exclude samples with missing values. To support this decision, we state that: First, considering that data were obtained through an early stage of the COVID-19 pandemic in Brazil, healthcare professionals did not have much knowledge about what factors could really matter to detect and monitor during the infection (Michelen et al. 2020). Second, the state of São Paulo faced several periods of high admission rates (BRAZIL 2021), both regular ward and ICUs. Therefore, missing values of features might have been caused by the lack of resources, since hospitals were crowded. Finally, all patients in the dataset are people that had potential symptoms of the COVID-19; therefore, it is not possible to infer a mean value or normal values for these patients, because a control group of healthy people are not present in this cohort.

As for the target class, we used only the RT-PCR COVID-19 test, which is still currently considered the golden standard for the COVID-19 diagnosis. That way, positive cases are considered 1 and negative cases 0. However, this test still has its limitations, especially for false negatives, and when applied outside its effective window of 3-4 days of infection, it might result in false negatives. The total number of positive cases for our current dataset is roughly 30%, which still makes this problem an imbalanced classification, but no longer an anomaly detection.

The final datasets were randomly split into 70% training and 30% holdout (test) in order to test model's generalization capacity.

The training set was linearly scaled with a maximum value of 1 and a minimum of 0. The exception was for the age feature, which was only divided by 100. It is essential to mention that the **test sets were scaled based on the parameters of the training sets** to avoid data leakage and optimistic unreal results.

Table 1 shows the age and blood test features distribution. Regarding the pathogen diagnosis features, there was a distribution of approximately 1–10% cases among them. Also, both datasets have approximately 50% female patients.

5.3 Model Training and Validation

In order to ensure model's capacity of generalization, we optimized their hyperparameter values by using fivefold stratified cross-validation. This cross-validation consisted of executing five iterations, separating the training data in five folds at random but keeping positive/negative class ratio. For each iteration, one fold is used for testing, and the other four are used for training. Hence, we can increase the model's generalization by testing several hyperparameters values with different sets.

The search for the best combination of the hyperparameters was performed using the Optuna framework, a hyperparameter optimization tool (Akiba et al. 2019). Optuna is a platform-agnostic API that utilizes searching, and pruning strategies based on the Tree-structured Parzen Estimator (Bergstra et al. 2011).

While we tuned the models for the F2 measure in our previous work, we now opted to tune them for the AUC measure. That is because we want to find models that can maintain a good balance between a high true positive rate and a low false positive rate.

6 Results

Table 2 summarizes the results for the blood test hold out dataset and the blood test with pathogen hold out dataset. For practicality, results from our previous dataset are shown to make a direct comparison. Metrics are presented along with their respective 99% confidence interval calculated as the Wilson score interval (Wallis 2013).

The overall results on this work are slightly lower than our previous work; however, we had access to more patients, making the obtained results more reliable. Besides, previous results are probably unrealistic because the original dataset had already been normalized, and we had no access to original feature values. That led to data leakage because the normalization was made before the train/test split. Unfortunately, by that time, that was the only COVID-19 dataset publicly available in Brazil. Moreover, the accuracy measure should not be directly compared since, in the previous work, the patients cohort contained only 10% positive cases, while the current cohort contains 30% of positive patients.

When considering the AUC measure, the random forest and the explainable boosting machine models were the top-performing models for the current datasets.

The dataset containing only the blood test features had poorer performance on test sets. It is clear that the pathogen variables play an essential role in the performance, probably because it is less likely for a person to be diagnosed with more than one pathogen at a time. That is, is unlikely that a person has COVID-19 and other diseases such as Influenza or other types of coronavirus. The closest work found in the literature (Gangloff et al. 2021) was briefly discussed in the machine learning for COVID-19 Diagnosis section. The authors found that using only clinical-biological, they achieved an AUC measure of 0.778 (CI 95% ±0.7) and suggested that this model could be used as a first triage. Our approach was evaluated in a more extensive dataset (500 against 107), and we did not perform any missing imputation, which indicates that our procedure did not alter any feature distribution. Besides, our approach shows better results and a narrower confidence interval. When considering the blood test and the pathogens, we obtained an AUC measure of AUC = 0.874 (CI 95% ± 0.04) for RF, and they obtained 0.778 (CI 95% ±0.7). It is also worth mentioning that the explainable boosting model achieved an AUC = 0.873 with the benefit of explaining. We believe that the results could be further improved with the addition of the clinical signs such as the proportion of cough, hyperthermia, myalgia, asthenia, diarrhea and confusion.

7 Interpretation and Explanations

Explainable artificial intelligence (XAI) is a new topic of study that focuses on machine learning interpretability and aspires to create a more transparent AI. The major goal is to develop a set of interpretable models and methodologies that result in more understandable models while maintaining excellent prediction performance (Adadi and Berrada 2018). Regrettably, there is no universally accepted definition of explainable. Some researchers use the terms interpretability and explainability interchangeably, while others distinguish between the two. Authors Doshi-Velez and Kim (2017) define interpret as "to explain or present in language that humans can understand." Authors in Samek et al. (2019) define interpretation as the translation of abstract concepts into a domain humans can understand, whereas explanation is the collection of the features of the interpretable domain

Table 1 Blood test feature distribution

Feature name	Mean	std	min	max	Reference
Blood test + Pathogen dataset (approximately 1500 patients)					
Age	48.87	18.36	0.0	89.0	0 years
PLATELETS	238.57	81.64	12.0	1027.0	$\times 10^3$/uL
MPV	10.21	0.93	7.6	14.7	6.5–15.0 fL
RED CELLS	4.59	0.68	1.2	7.03	3.90–5.00 $\times 10^6$/uL
BASOPHILS (%)	0.43	0.3	0.0	3.2	%
BASOPHILS	31.48	22.88	0.0	286.0	0–100 μL
EOSINOPHILS (%)	2.02	2.4	0.0	28.9	%
LYMPHOCYTES (%)	25.19	11.57	0.5	83.3	%
LYMPHOCYTES	1792.04	945.67	60.0	14370.0	900–2900 μL
MONOCYTES (%)	8.75	3.44	0.4	41.6	%
MONOCYTES	638.21	300.72	10.0	2996.0	300–900 μL
CHCM	34.07	1.12	27.9	37.5	31.0–36.0 g/dL
HCM	29.55	1.94	18.3	43.0	26.0–34.0 pg
VCM	86.74	5.06	60.8	123.2	82.0–98.0 fL
RDW	13.26	1.57	10.7	26.3	11.5–16.5 %
LEUKOCYTES (%)	7.66	3.1	0.53	33.82	%
LEUKOCYTES	7669.34	3115.6	530.0	33820.0	3500–10500 μL
Blood test only dataset (approximately 5800 patients)					
Age	48.68	18.6	0.0	89.0	0 years
PLATELETS	225.18	78.97	21.0	1027.0	$\times 10^3$/uL
MPV	10.2	0.9	7.8	13.9	6.5–15.0 fL
RED CELLS	4.64	0.63	2.34	6.78	3.90–5.00 $\times 10^6$/uL
BASOPHILS (%)	0.4	0.29	0.0	2.29	%
BASOPHILS	29.27	23.01	0.0	225.0	0–100 μL
EOSINOPHILS (%)	1.82	2.32	0.0	21.8	%
LYMPHOCYTES (%)	23.75	11.87	0.5	83.3	%
LYMPHOCYTES	1706.48	1101.95	60.0	14370.0	900–2900 μL
MONOCYTES (%)	9.02	3.62	0.6	33.9	%
MONOCYTES	651.83	319.92	39.0	2954.0	300–900 μL
CHCM	34.11	1.08	28.5	37.5	31.0–36.0 g/dL
HCM	29.6	1.95	18.6	39.4	26.0–34.0 pg
VCM	86.8	5.07	63.8	118.4	82.0–98.0 fL
RDW	13.24	1.38	11.0	24.1	11.5–16.5 %
LEUKOCYTES (%)	7.61	3.16	0.76	33.82	%
LEUKOCYTES	7614.73	3190.62	760.0	33820.0	3500–10500 μL

that have led to the production of a choice in a specific example. The notion of explanation in this work is aligned with Samek et al. (2019).

As stated previously, it is possible to analyze a model relation between input and output in two ways: looking at local explanations and global interpretation/explanations (Molnar 2019). Here, we use these two terms interchangeably . While global interpretations enlighten scientists about what features may be more representative to the model given a data sample, local explanations show what led the model to the current output.

In our scope, global interpretations can help healthcare researchers to question data sample quality and also direct the search for more relevant features. Single explanations of the COVID-19 diagnosis can assist healthcare professionals that are working directly with potentially infected patients in their decision-making.

7.1 Global Interpretation

For the glass-box models, global interpretation can be qualitatively evaluated by analyzing its feature weights. For these

Table 2 Results of experiments on test sets

Dataset	Model	AUC	F2	Accuracy
Blood test + Pathogen	LR	0.860 ± 0.042	0.771 ± 0.051	0.773 ± 0.051
	EBM	0.873 ± 0.041	**0.779 ± 0.051**	**0.813 ± 0.048**
	SVM	0.856 ± 0.043	0.778 ± 0.051	0.779 ± 0.051
	RF	**0.874 ± 0.041**	0.773 ± 0.051	0.809 ± 0.048
Blood test only	LR	0.803 ± 0.024	0.689 ± 0.028	0.771 ± 0.026
	EBM	0.830 ± 0.023	0.725 ± 0.027	0.787 ± 0.025
	SVM	0.816 ± 0.024	0.732 ± 0.027	0.737 ± 0.027
	RF	**0.838 ± 0.023**	**0.762 ± 0.026**	**0.789 ± 0.025**
Previous	LR	**0.899 ± 0.074**	**0.844 ± 0.090**	0.872 ± 0.083
	SVM	0.866 ± 0.084	0.813 ± 0.096	0.853 ± 0.088
	RF	0.896 ± 0.075	0.834 ± 0.092	**0.908 ± 0.071**

Values in bold show the best metric result for each dataset

models, we can look at these weighting coefficients as the importance that a model gives to a feature. Figure 1 shows the feature weights for the logistic regression. Weights with positive values indicate features that contribute to the positive diagnosis, and features with negative values contribute to the negative diagnosis.

Figure 2 shows the feature weights for the explainable boosting machine. However, since this is a nonlinear model, it is impossible to relate the positive and negative contributions of features. Since the black-box models offer no intelligible interpretation, we apply the Kernel Explainer and the Tree Explainer of the SHAP algorithm to the SVM and random forest, respectively. The SHAP values were calculated with the test set samples. Figures 3 and 4 show the SHAP Summary plot, which provide a sense of global interpretation. The blue dots indicate low feature values, while red dots indicate high feature values. Features that appear with dots spread across the horizontal axis are important, while features containing dots closer to the zero lines are non-important features. Also, it is important to mention that the higher an absolute SHAP value is, the higher the contribution to model output.

Similar to our previous findings, we see that in pretty much all plots mentioned so far, the models recognize the white blood cell features (leukocytes, lymphocytes, eosinophils, and basophils) as being the most important. These cells are responsible for maintaining our immunological system.

Also, the monocytes still appear as an important feature, and in the LR plots, it has a positive contribution for the COVID-19 diagnosis. This is also in accord with our previously reported study and also with the following findings (Meidaninikjeh et al. 2021; Skevaki et al. 2020; Mehta et al. 2020). Age is a moderate feature in all scenarios, while sex has no importance for the LR model and lesser importance for the EBM.

It is worth mentioning that pathogens play an essential role in the discrimination of COVID-19. High values of other pathogens in the SHAP summary, such as H1N1 and INFLUENZA B, are vital to dismiss that the patient has COVID-19.

Interestingly, the EBM presents all feature interaction pairs and being less critical than the age factor. Despite that, the model still captures some importance for the compound effect of monocytes and white blood cells.

We see that low levels of white blood cells (indicated by blue dots) significantly impact models output for all SHAP summaries. Moreover, the monocytes behavior found in the logistic regression is also seen in the SHAP summaries. That is, with high levels of monocytes (indicated by red dots), we see a positive contribution to the models output. For the age variable, we see that for younger patients, their age has an important negative impact on the output. However, the age is not so relevant for older patients, given that SHAP values are closer to zero. The SHAP summary for the SVM with the blood test variables shows the most distinct interpretation results. It is important to mention that this model had a lower performance when compared to other models on both current and previous results.

7.2 Local Explanations

As mentioned earlier, local explanations can assist healthcare professionals in their decision-making about possibly infected patients. However, even for research purposes, they play an essential role. For example, it is possible to compare the explanation of a true-positive patient with a false-positive and understand what might confuse the models.

Since both datasets contain over 1000 samples, it is impossible to show all outcomes for all models due to space limitations. Furthermore, the logistic regression is a linear model and intrinsically interpretable; all local explanations would be the same, since the model's feature importance does not change with feature values. Therefore, we will show examples where all models correctly predicted the

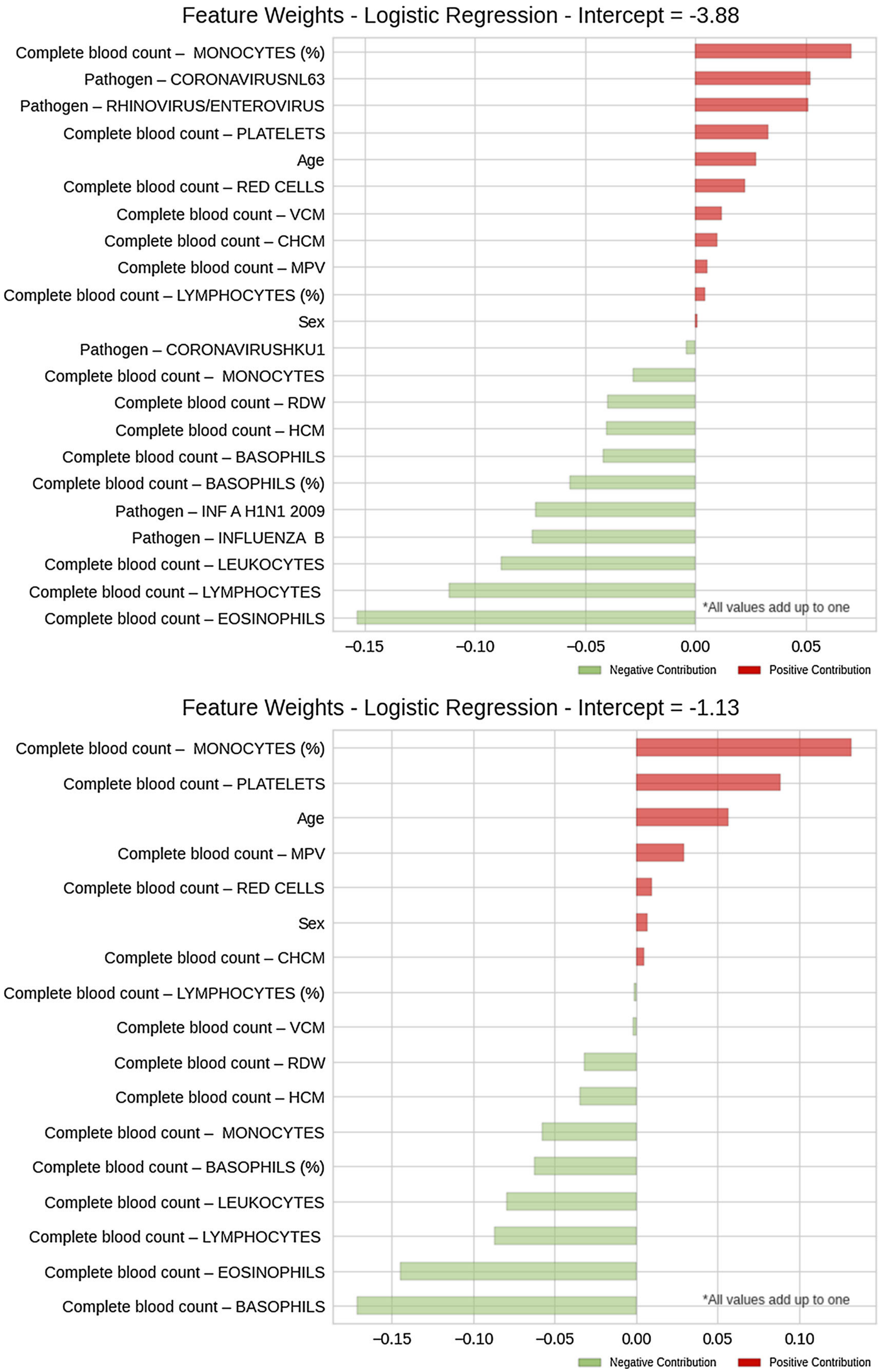


Fig. 1 Global interpretation for logistic regression. Top: dataset with blood tests and pathogens; bottom: dataset with only blood tests

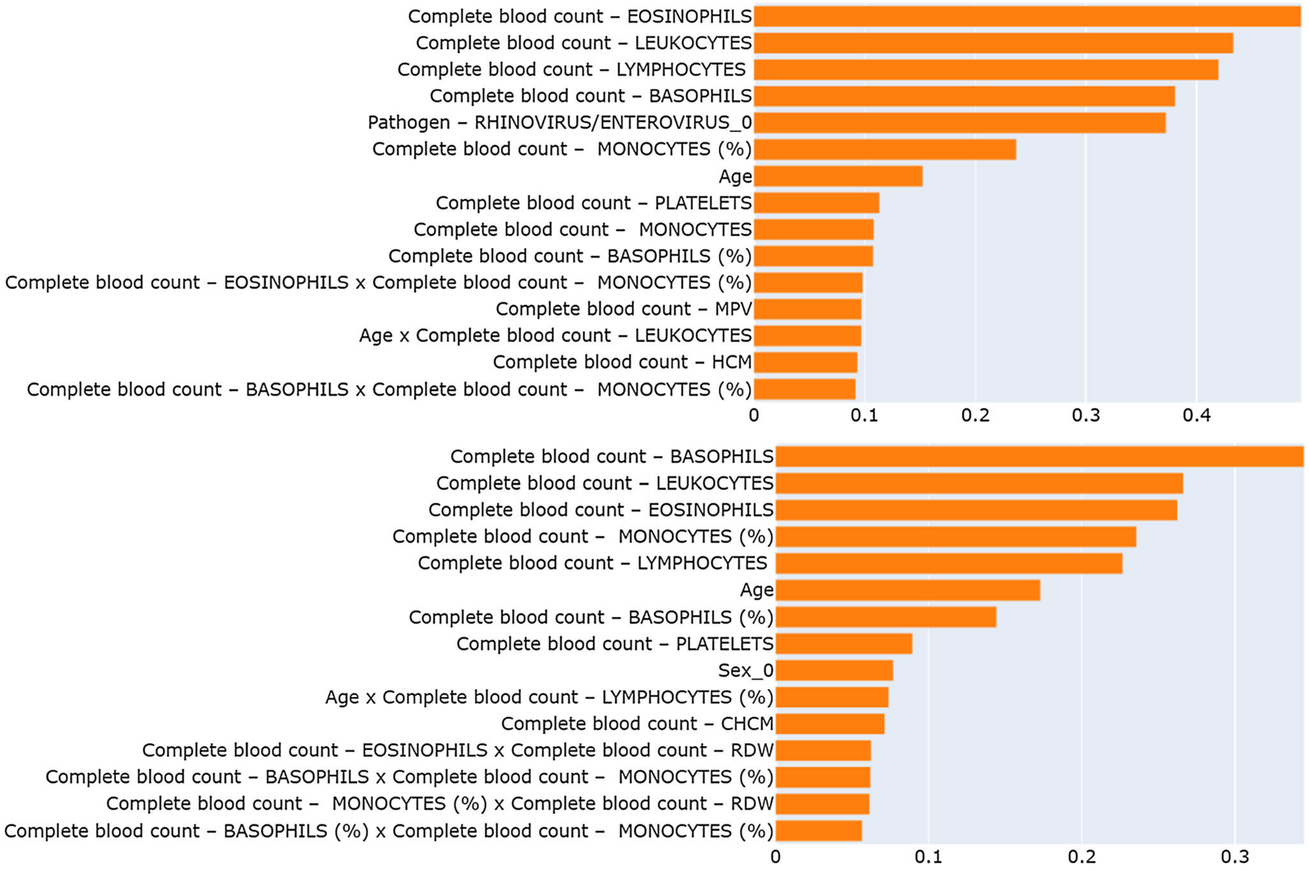


Fig. 2 Global interpretation for explainable boosting machine. Top: dataset with blood tests and pathogens; Bottom: dataset with only blood tests

outcome, some models failed, while others succeeded, all models failed.

Thus, we will show the reader how one can interpret local explanations for the COVID-19 diagnosis in a myriad of situations. In other words, we want to show that although the overall results are really good, the models still make mistakes and get confused with the features, and local explanations can help with identify that. If the reader wants to check more examples, we encourage seeing our GitHub, provided at the beginning of the Experiments section.

Figure 5 shows explanations for a given patient that led to a true-positive diagnosis and Fig. 6 for a true-negative patient. Recalling that SVM and RF explanations are obtained through SHAP values, while EBM explanations result from its smooth functions for each feature.

The pink arrows in the SHAP values waterfall plots indicate values contributing to the positive diagnosis, while the blue arrows contribute to the negative diagnosis.

For the true positive, low levels of white blood cells and elevated values of red cells greatly impact the positive outcome for this patient, for all models.

For the true negative, the white blood cells are also responsible for the majority impact on models output. Interestingly, the SVM was the only model that took in consideration the monocytes and sex of the patient as relevant features.

Figure 7 shows a negative patient where the EBM correctly identified and RF and SVM classified as positive. Even though EBM got the classification right, we notice that several variables have positive contributions, similarly to the SVM and RF models. Also, we see that the model output for SVM and RF are nearly 0.5. Although not reported in the figure, the model output for the EBM is 0.48. Therefore, this patient lies in a low confidence zone for all models.

Figure 8 shows a positive patient that all the models failed to diagnose, for the hemogram dataset. We notice that all models predicted a score near to 0 probability of a positive case and most of the features actually contributed to a negative diagnosis in all cases. This implies that there might be a group of patients that do not suffer from clear alterations in these blood test features, indicating that more information about the patient is needed.

Finally, for all presented cases, all models considered the white blood cells as the most relevant, according to the global interpretation plots. However, besides the white blood cell variables, each model considered different aspects to obtain their outcomes. It suggests that it is possible to characterize

Fig. 3 Global interpretation for SVM. Top: dataset with blood tests and pathogens; bottom: dataset with only blood tests

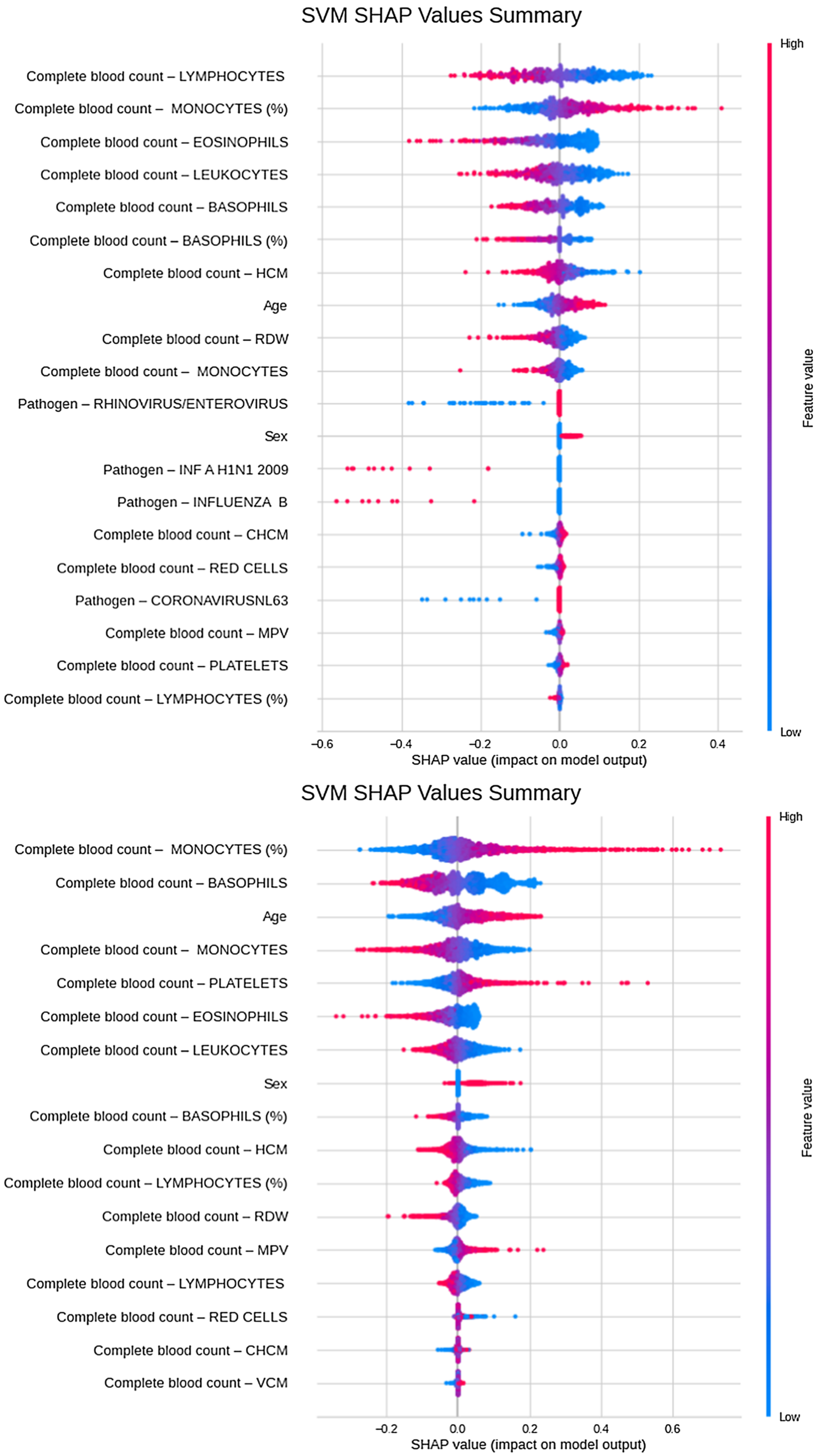
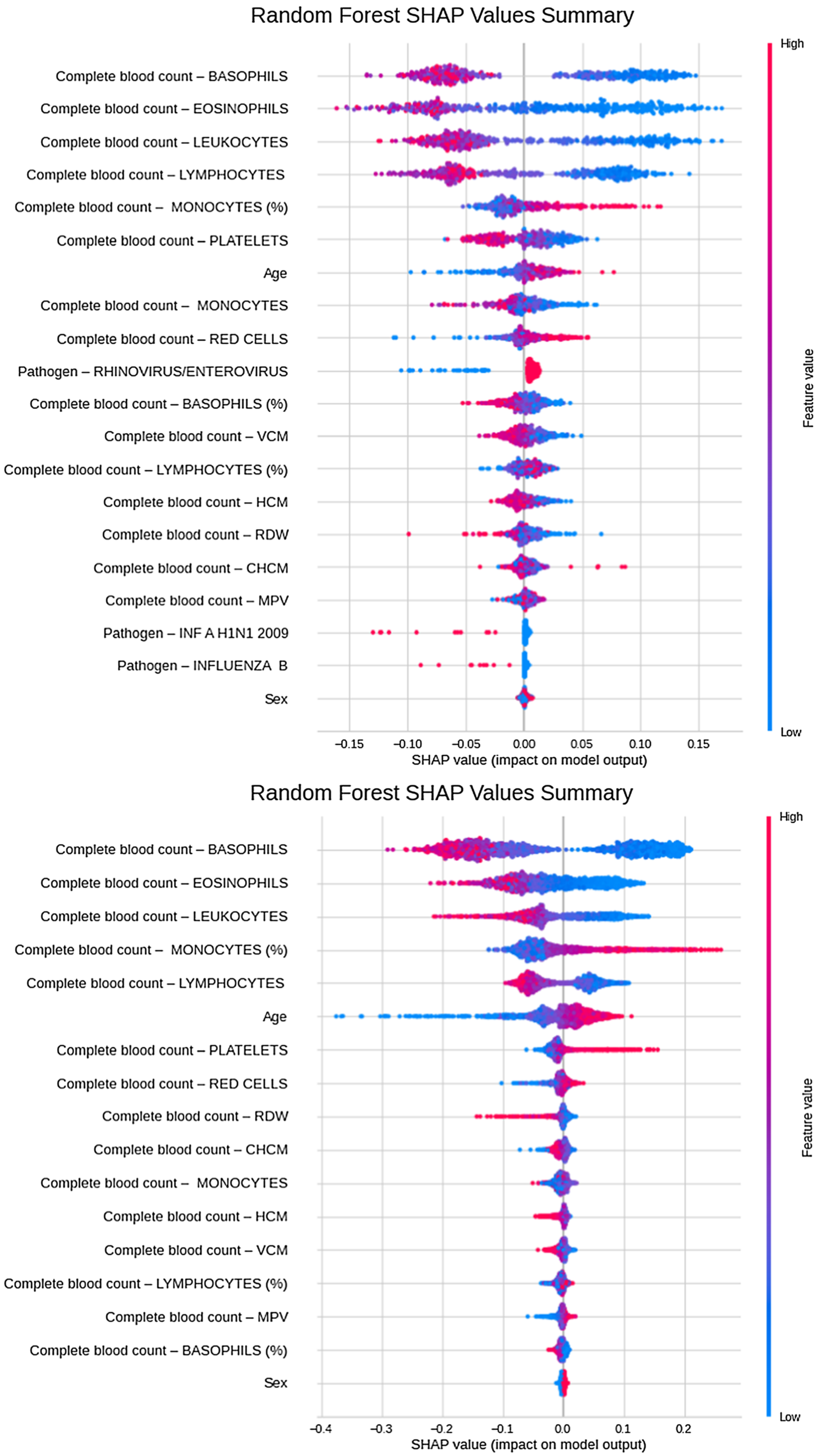


Fig. 4 Global interpretation for random forest. Top: dataset with blood tests and pathogens; bottom: dataset with only blood tests



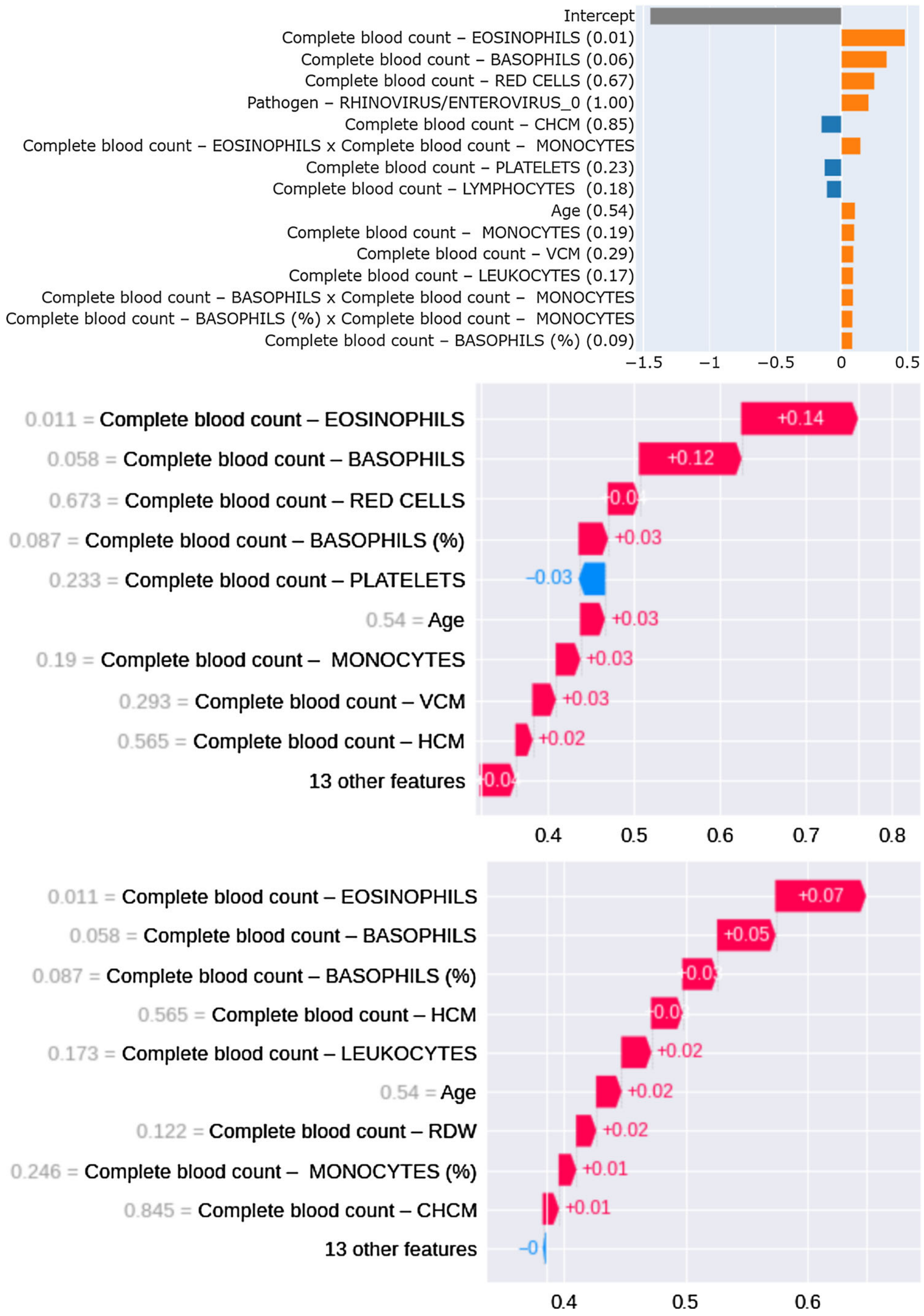


Fig. 5 Local explanations for a true positive patient on hemogram and pathogen dataset. Top: EBM; middle: RF; bottom: SVM

the same diagnosis based on a set of different features; that is, it is possible to characterize the same event but from different perspectives.

7.3 Limitations

It is worth noting that the data used to train the machine learning models were gathered in the same geographic area (São Paulo), which may restrict their ability to generalize to other regions with different profiles.

Moreover, since all our models considered the white blood cells variables as the most important ones, there might be a bias toward infected patients with good levels of white cells. Not only that, but specialists might criticize that looking only at a complete blood count test might not be enough to identify the COVID-19 among other similar diseases. Therefore, further studies should be carried to assess the predictive power of these variables on potentially false negative patients and also comparing the COVID-19 with other diseases.

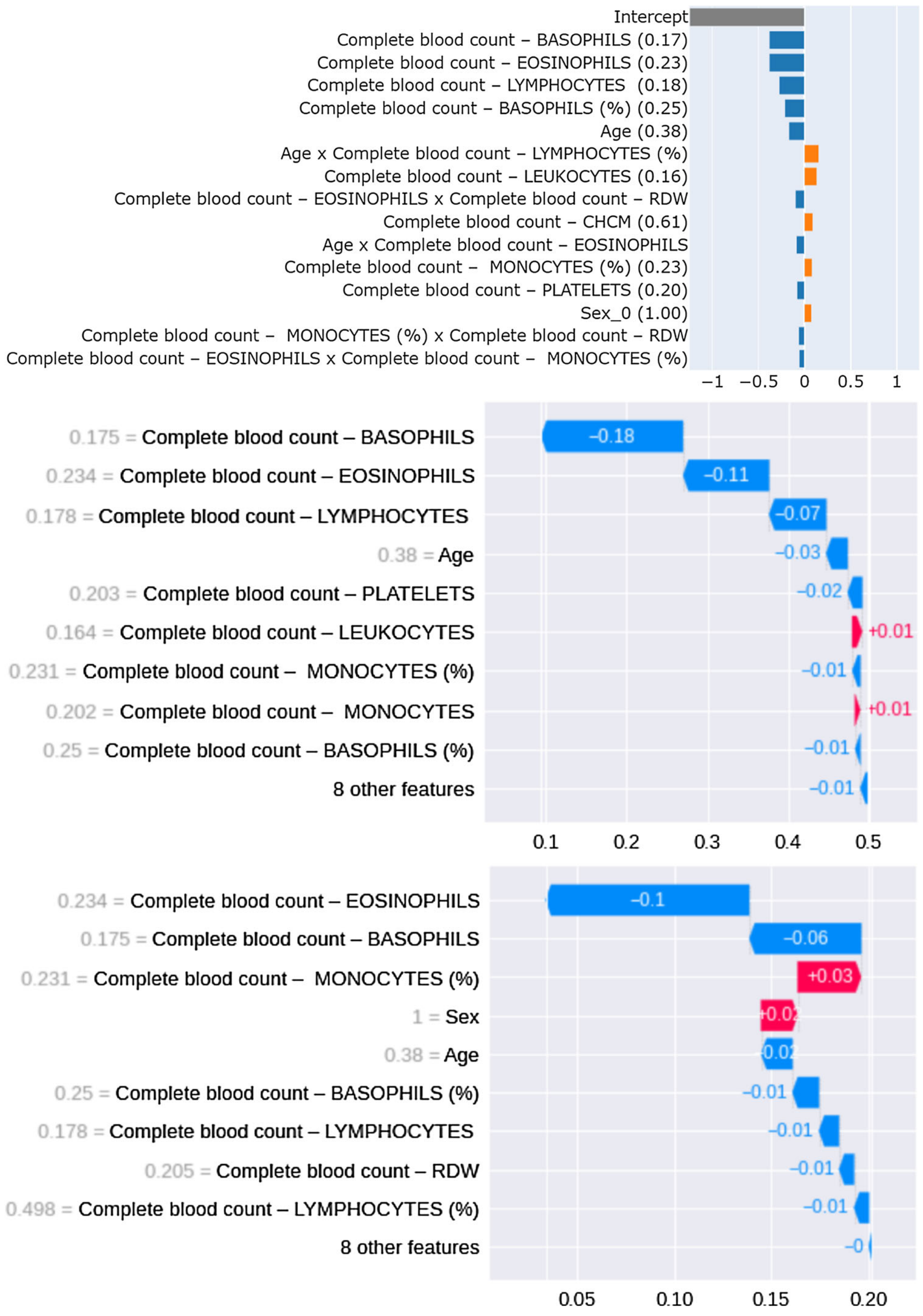


Fig. 6 Local explanations for a true negative patient on hemogram. Top: EBM; middle: RF; bottom: SVM

8 Conclusion

An explainable artificial intelligence system to help the diagnostic of COVID-19 patients, based on the blood test and pathogen variables, was presented in this paper. Despite the limitations presented in this study, the results obtained showed to be entirely satisfactory. The algorithms with the best results were EBM and RF. They achieved an AUC of 0.873 and 0.874, respectively. An AUC of at least 0.80 is widely accepted as adequate for diagnostic use, which indicates moderate diagnostic accuracy (0.70–0.90), and it is very close to the high accuracy range (AUC $\geq$ 0.9).

As mentioned earlier, it is important to notice that the concept of explainable artificial intelligence is still under discussion in the community and has no universal definition up to the point that this work is being presented. The major goal of XAI is to develop a set of methodologies that provide more understandable models while maintaining their predictive power. The analysis with the explainable artificial intelligence tools has shown that white blood cell fea-

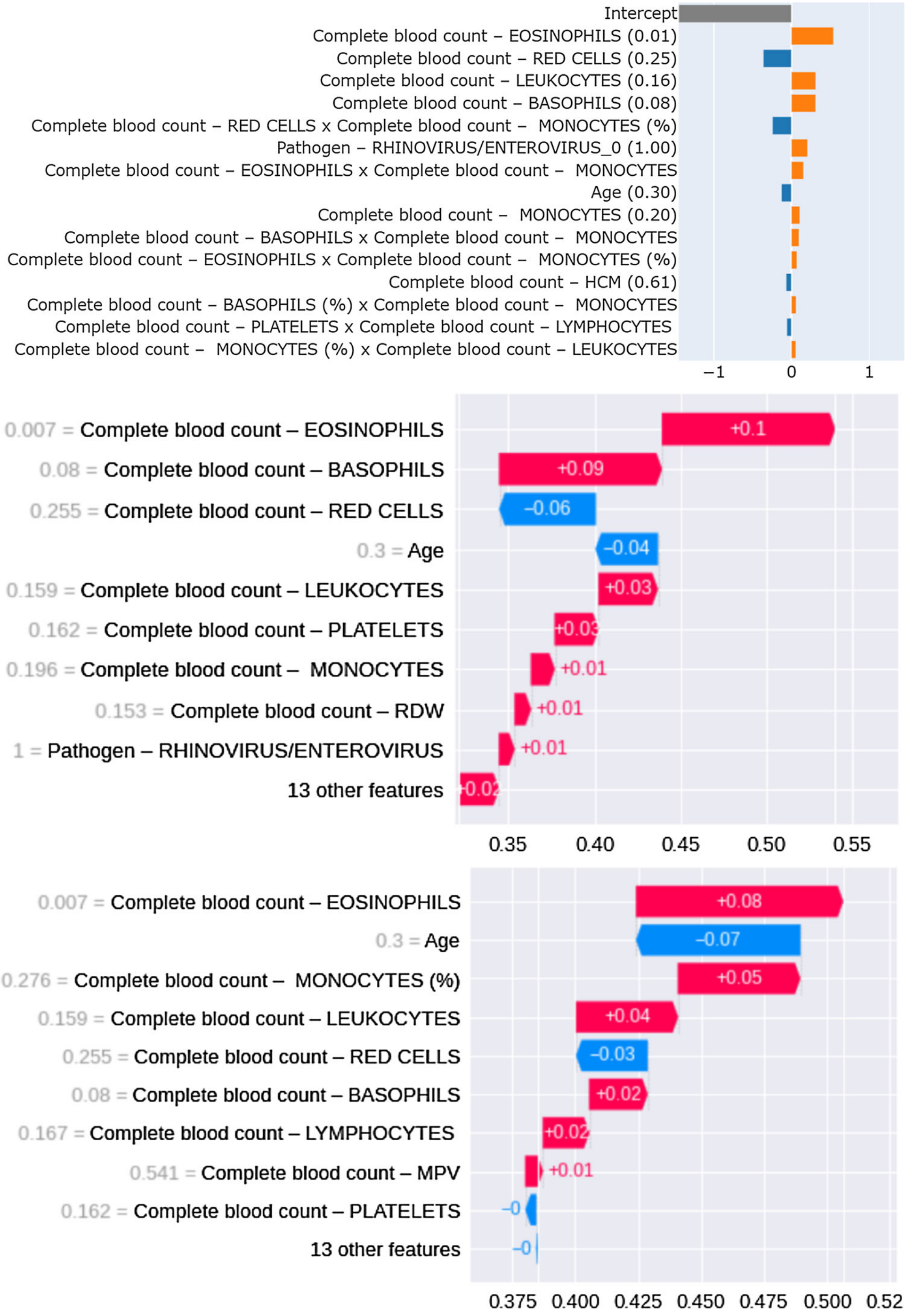


Fig. 7 Local explanations for mixed diagnosis negative patient on hemogram and pathogen dataset. Top: EBM; middle: RF; bottom: SVM

tures (leukocytes, lymphocytes, eosinophils and basophils) and the presence of other pathogens (such as H1N1 and INFLUENZA B) play an essential role in the COVID-19 detection. The SHAP summaries indicate if the contribution is favorable or not.

This method is analogous to the aforementioned COVID-19 rapid tests because of how the models interpreted the variables and predicted the events. These tests look for changes in the immune system and other factors that may indicate the existence of the disease.

Our results indicated that the proposed system could indeed be used to detect COVID-19 patients using clinical and laboratory data that is commonly used in clinical practice. In terms of healthcare, our approach could represent a new method that can help by quickly screening patients with COVID-19. This can be specially advantageous to expand testing protocols to areas where there is not available tests

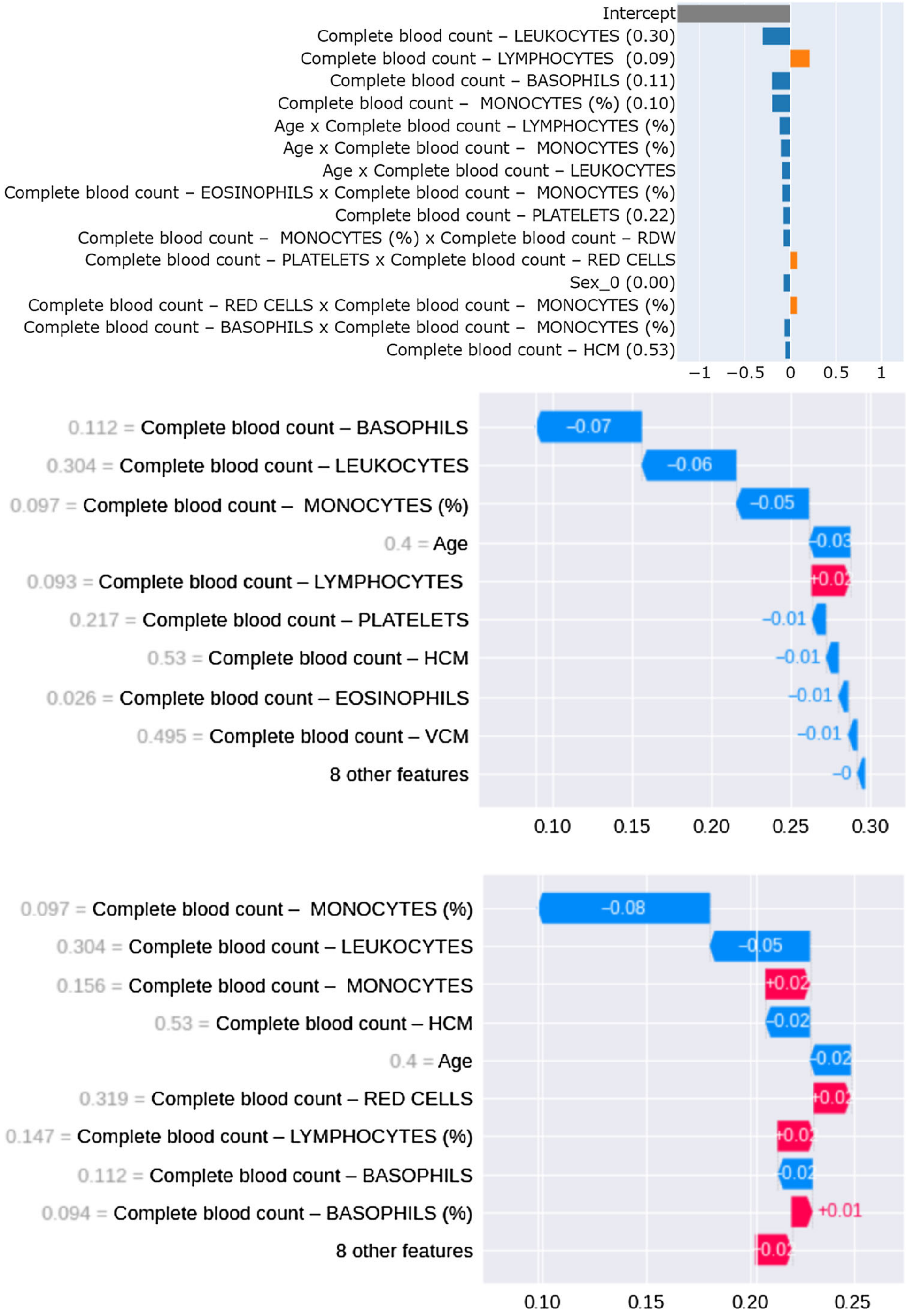


Fig. 8 Local explanations for failed diagnosis on hemogram dataset. Top: EBM; middle: RF; bottom: SVM

for the local population. Additionally, situations where laboratory workloads are heavy, and RT-PCR tests are scarce could also benefit from rapid blood sample tests. Therefore, it would help reduce the burden in the healthcare system and promote the optimal utilization of healthcare resources. Besides, we think that is possible to improve the results with the addition of the clinical signs such as the proportion of cough, hyperthermia, myalgia, asthenia, diarrhea and confusion.

We look forward to including more COVID-19 test types as well as data from different geographical and socioeconomic regions from Brazil. Due to the continental size of the country, the epidemic dynamics changes drastically between regions and our proposed approach might show different performance and also different explanations. Moreover, we would like to increase more blood test features in order to seek better model performance without depending on pathogen variables for future work. Although it was clear that all models performed well, there were some differences in feature importances. Therefore, it may be interesting to test an ensemble learning approach; to combine all models and explain predictions with SHAP or another agnostic explanation algorithm.

Lastly, since we have more knowledge about the COVID-19 disease and its behavior, we seek to impute missing data through domain knowledge , as well as machine learning-based methods such as decision trees (Lin and Tsai 2019) and investigate potential improvements in models performance.

Acknowledgements An early version of paper was presented at XXIII Congresso Brasileiro de Automática (CBA 2020). The authors would like to acknowledge the financial support provided by FAPERJ—Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro: E-26/210.320/2018 and CAPES - Finance Code 001.

Declarartions

Conflict of interest The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Agor, J., Özaltın, O. Y., Ivy, J. S., Capan, M., Arnold, R., & Romero, S. (2019). The value of missing information in severity of illness score development. *Journal of Biomedical Informatics, 97*, 103255. https://doi.org/10.1016/j.jbi.2019.103255

Akiba, T., Sano, S,. Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. arXiv:190710902 [cs, stat].

Amaral, J. L. M., Lopes, A. J., Jansen, J. M., Faria, A. C. D., & Melo, P. L. (2012). Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease. *Computer Methods and Programs in Biomedicine, 105*(3), 183–193. https://doi.org/10.1016/j.cmpb.2011.09.009

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nature Medicine, 26*(4), 450–452. https://doi.org/10.1038/s41591-020-0820-9

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, *24*.

Biamonte, F., Botta, C., Mazzitelli, M., Rotundo, S., Trecarichi, E. M., Foti, D., et al. (2021). Combined lymphocyte/monocyte count, D-dimer and iron status predict COVID-19 course and outcome in a long-term care facility. *Journal of Translational Medicine, 19*(1), 79. https://doi.org/10.1186/s12967-021-02744-2

BRAZIL MoH (2021) Covid-19 in Brazil. https://qsprod.saude.gov.br/extensions/covid-19_html/covid-19_html.html.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Di Napoli, R. (2020). Features, evaluation and treatment coronavirus (COVID-19). In StatPearls, StatPearls Publishing, Treasure Island (FL), http://www.ncbi.nlm.nih.gov/books/NBK554776/.

Chakraborty, C., Sharma, A. R., Sharma, G., Bhattacharya, M., & Lee, S. S. (2020). SARS-CoV-2 causing pneumonia-associated respiratory disorder (COVID-19): Diagnostic and proposed therapeutic options. *European Review for Medical and Pharmacological Sciences, 24*(7), 4016–4026.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 1–27. https://doi.org/10.1145/1961189.1961199

DeCaprio, D., Gartner, J., Burgess, T., Garcia, K., Kothari, S., Sayed, S., & McCall, C. J. (2020). Building a COVID-19 Vulnerability Index. arXiv:200307347 [cs, stat].

de Moraes Batista, A. F., Miraglia, J. L., Donato, T. H. R., & Filho, A. D. P. C. (2020). COVID-19 diagnosis prediction in emergency care patients: A machine learning approach. *Epidemiology*. https://doi.org/10.1101/2020.04.04.20052092

de Sousa, I. P., Vellasco, M. M. B. R., & da Silva, E. C. (2019). Local interpretable model-agnostic explanations for classification of lymph node metastases. *Sensors (Basel, Switzerland)*. https://doi.org/10.3390/s19132969

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:170208608 [cs, stat].

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2019). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research, 9*, 31.

Gangloff, C., Rafi, S., Bouzillé, G., Soulat, L., & Cuggia, M. (2021). Machine learning is the key to diagnose COVID-19: A proof-of-concept study. *Scientific Reports, 11*(1), 7166. https://doi.org/10.1038/s41598-021-86735-9

Ghaderzadeh, M., & Asadi, F. (2021). Deep learning in the detection and diagnosis of COVID-19 using radiology modalities: A systematic review. *Journal of Healthcare Engineering, 2021*, e6677314. https://doi.org/10.1155/2021/6677314

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*(3), 297–310. https://doi.org/10.1214/ss/1177013604

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J., & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet (London, England), 395*(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Jha, P. K., Cao, L., & Oden, J. T. (2020). Bayesian-based predictions of COVID-19 evolution in Texas using multispecies mixture-theoretic continuum models. *Computational Mechanics*. https://doi.org/10.1007/s00466-020-01889-z

Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., Shi, J., Dai, J., Cai, J., Zhang, T., & Wu, Z. (2020). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Computers, Materials and Continua, 62*(3), 537–551. https://doi.org/10.32604/cmc.2020.010691

Krammer, F., & Simon, V. (2020). Serology assays to manage COVID-19. *Science, 368*(6495), 1060–1061. https://doi.org/10.1126/science.abc1227

Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., Cheng, L., Li, J., Wang, X., Wang, F., Liu, L., Amit, I., Zhang, S., & Zhang, Z. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nature Medicine, 26*(6), 842–844. https://doi.org/10.1038/s41591-020-0901-9

Lin, W. C., & Tsai, C. F. (2019). Missing value imputation: A review and analysis of the literature (2006–2017). *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-019-09709-4

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P. Q., Corrado, G.S., Hipp, J. D., Peng, L., & Stumpe, M. C. (2017). Detecting cancer metastases on gigapixel pathology images. *MICCAI Tutorial* (2017) arXiv:1703.02442v2

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Association for Computing Machinery, New York, NY, USA, KDD'13* (pp. 623–631). https://doi.org/10.1145/2487575.2487579

Lundberg, S. M. (2020). SHAP—A game theoretic approach to explain the output of any machine learning model. https://github.com/slundberg/shap, library Catalog: github.com.

Lundberg, S. M., & Lee, S. I. (2017). 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., pp. 4765–4774.

Maggi, E., Canonica, G. W., & Moretta, L. (2020). COVID-19: Unanswered questions on immune response and pathogenesis. *Journal of Allergy and Clinical Immunology, 146*(1), 18–22. https://doi.org/10.1016/j.jaci.2020.05.001

Martinez, E. Z., Aragon, D. C., & Nunes, A. A. (2020a). Short-term forecasting of daily COVID-19 cases in Brazil by using the Holt's model. *Revista da Sociedade Brasileira de Medicina Tropical*. https://doi.org/10.1590/0037-8682-0283-2020

Martinez, F. O., Combes, T. W., Orsenigo, F., & Gordon, S. (2020b). Monocyte activation in systemic Covid-19 infection: Assay and rationale. *EBioMedicine*. https://doi.org/10.1016/j.ebiom.2020.102964

Mehta, P., McAuley, D. F., Brown, M., Sanchez, E., Tattersall, R. S., & Manson, J. J. (2020). COVID-19: Consider cytokine storm syndromes and immunosuppression. *The Lancet, 395*(10229), 1033–1034. https://doi.org/10.1016/S0140-6736(20)30628-0

Meidaninikjeh, S., Sabouni, N., Marzouni, H. Z., Bengar, S., Khalili, A., & Jafari, R. (2021). Monocytes and macrophages in COVID-19: Friends and foes. *Life Sciences, 269*, 119010. https://doi.org/10.1016/j.lfs.2020.119010

Merad, M., & Martin, J. C. (2020). Pathological inflammation in patients with COVID-19: A key role for monocytes and macrophages. *Nature Reviews Immunology, 20*(6), 355–362. https://doi.org/10.1038/s41577-020-0331-4

Michelen, M., Jones, N., & Stavropoulou, C. (2020). In patients of COVID-19, What are the symptoms and clinical features of mild and moderate cases? Library Catalog: www.cebm.net.

Mohammad-Rahimi, H., Nadimi, M., Ghalyanchi-Langeroudi, A., Taheri, M., & Ghafouri-Fard, S. (2021). Application of machine learning in diagnosis of COVID-19 through X-Ray and CT images: A scoping review. *Frontiers in Cardiovascular Medicine*. https://doi.org/10.3389/fcvm.2021.638011

Molnar, C. (2019). Interpretable machine learning. Lulu.com, https://christophm.github.io/interpretable-ml-book/.

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. arXiv:190909223 [cs, stat].

Pardi, N., & Weissman, D. (2020). Development of vaccines and antivirals for combating viral pandemics. *Nature Biomedical Engineering, 4*(12), 1128–1133. https://doi.org/10.1038/s41551-020-00658-w

Porte, L., Legarraga, P., Vollrath, V., Aguilera, X., Munita, J. M., Araos, R., Pizarro, G., Vial, P., Iruretagoyena, M., Dittrich, S., & Weitzel, T. (2020). Evaluation of novel antigen-based rapid detection test for the diagnosis of SARS-CoV-2 in respiratory samples. *International Journal of Infectious Diseases*. https://doi.org/10.1016/j.ijid.2020.05.098

Potie, N., Giannoukakos, S., Hackenberg, M., & Fernandez, A. (2019). On the need of interpretability for biomedical applications: Using fuzzy models for lung cancer prediction with liquid biopsy. In *2019 IEEE international conference on fuzzy systems (FUZZ-IEEE)* (pp. 1–6). https://doi.org/10.1109/FUZZ-IEEE.2019.8858976, ISSN: 1558-4739.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., Rudd, J. H. F., Sala, E., & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence, 3*(3), 199–217. https://doi.org/10.1038/s42256-021-00307-0

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer.

Schmidt, D., Niemann, M., & Trzebiatowski, G. (2015). The handling of missing values in medical domains with respect to pattern mining algorithms. In *CS&P*.

Singh, P., Singh, S. P., & Singh, D. S. (2019). An introduction and review on machine learning applications in medicine and healthcare. In *2019 IEEE conference on information and communication technology* (pp. 1–6). https://doi.org/10.1109/CICT48419.2019.9066250

Singhal, T. (2020). A review of coronavirus disease-2019 (COVID-19). *Indian Journal of Pediatrics, 87*(4), 281–286. https://doi.org/10.1007/s12098-020-03263-6

Skevaki, C., Fragkou, P. C., Cheng, C., Xie, M., & Renz, H. (2020). Laboratory characteristics of patients infected with the novel SARS-CoV-2 virus. *Journal of Infection, 81*(2), 205–212. https://doi.org/10.1016/j.jinf.2020.06.039

Teijaro, J. R., Walsh, K. B., Rice, S., Rosen, H., & Oldstone, M. B. A. (2014). Mapping the innate signaling cascade essential for cytokine storm during influenza virus infection. *Proceedings of the National Academy of Sciences, 111*(10), 3799–3804. https://doi.org/10.1073/pnas.1400593111

Thimoteo, L. M. (2020). COVID-19 Prediction. https://github.com/lucasthim/covid19-prediction, original-date: 2020-03-30T22:42:48Z.

Thimoteo, L. M., Vellasco, M. M., Amaral, J. M. D., Figueiredo, K., Yokoyama, C. L., & Marques, E. (2020). Interpretable machine learning for COVID-19 diagnosis through clinical variables. *Congresso Brasileiro de Automática - CBA*. https://doi.org/10.48011/asba.v2i1.1590

Tjoa, E., & Guan, C. (2019). A survey on explainable artificial intelligence (XAI): Towards medical XAI. IEEE Transactions on Neural Networks and Learning Systems, arXiv:1907.07374v4.

Vapnik, V. N. (2000). Methods of pattern recognition. In V. N. Vapnik (Ed.), *The nature of statistical learning theory, statistics for engineering and information science*. Springer, New York, NY (pp. 123–180). https://doi.org/10.1007/978-1-4757-3264-1_6

Viguerie, A., Lorenzo, G., Auricchio, F., Baroli, D., Hughes, T. J. R., Patton, A., Reali, A., Yankeelov, T. E., & Veneziani, A. (2021). Simulating the spread of COVID-19 via a spatially-resolved susceptible-exposed-infected-recovered-deceased (SEIRD) model with heterogeneous diffusion. *Applied Mathematics Letters*. https://doi.org/10.1016/j.aml.2020.106617

Wallis, S. (2013). Binomial confidence intervals and contingency tests: Mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics, 20*(3), 178–208.

WHO. (2020). WHO announces COVID-19 outbreak a pandemic. https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic, library Catalog: www.euro.who.int.

WHO. (2021). WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int.

Zhai, P., Ding, Y., Wu, X., Long, J., Zhong, Y., & Li, Y. (2020). The epidemiology, diagnosis and treatment of COVID-19. *International Journal of Antimicrobial Agents, 55*(5), 105955. https://doi.org/10.1016/j.ijantimicag.2020.105955

Zhang, D., Guo, R., Lei, L., Liu, H., Wang, Y., Wang, Y., Qian, H., Dai, T., Zhang, T., Lai, Y., Wang, J., Liu, Z., Chen, T., He, A., O'Dwyer, M., & Hu, J. (2020). COVID-19 infection induces readily detectable morphological and inflammation-related phenotypic changes in peripheral blood monocytes, the severity of which correlate with patient outcome. medRxiv p 2020.03.24.20042655, https://doi.org/10.1101/2020.03.24.20042655

Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digital Medicine, 4*(1), 1–5. https://doi.org/10.1038/s41746-020-00372-6

Zohdi, T. I. (2020). Modeling and simulation of the infection zone from a cough. *Computational Mechanics, 66*(4), 1025–1034. https://doi.org/10.1007/s00466-020-01875-5

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2018.00515