Interpretable Machine Learning for COVID-19 Diagnosis Through Clinical Variables

Lucas M. Thimoteo * **Marley M. Vellasco** *
Jorge M. do Amaral ** **Karla Figueiredo** *
Cátia Lie Yokoyama *** **Erito Marques** **

* *Departamento de Engenharia Elétrica, Pontifícia Universidade
Católica do Rio de Janeiro, RJ*
** *Programa de Pós-Graduação em Engenharia Eletrônica (PEL),
Universidade do Estado do Rio de Janeiro, RJ*
*** *Departamento de Biologia Geral, Universidade Estadual de
Londrina, PR*

lucasthim@yahoo.com, marley@ele.puc-rio.br, jamaral@uerj.br ,
catialie@gmail.com, karla@ele.puc-rio.br, mederitomarques@gmail.com

Abstract: This work proposes an interpretable machine learning approach to diagnose suspected COVID-19 cases based on clinical variables. Results obtained for the proposed models have F-2 measure superior to 0.80 and accuracy superior to 0.85. Interpretation of the linear model feature importance brought insights about the most relevant features. Shapley Additive Explanations were used in the non-linear models. They were able to show the difference between positive and negative patients as well as offer a global interpretability sense of the models.

Keywords: COVID-19 Diagnosis, Machine Learning, Explainability, Interpretability, Shapley Additive Explanations.

1. INTRODUCTION

The Corona Virus 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 March 2020, the World Health Organization (WHO) declared the COVID-19 to be a pandemic outbreak WHO (2020b). According to the Situation Report 153 from the WHO WHO (2020a), over 8 million people has been infected and nearly half a million deceased, up to June 2020.

Due to the fact of the COVID-19 highly contagious nature, governments saw the need to implement policies for social distancing, quarantine, and lock-downs. These events brought serious health and economic challenges to several countries, and Brazil was no exception. There is an urgent need to test the majority of the population in order 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 realible 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 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 remarkable approaches are agnostic models that can explain deep learning predictions and used to explain cancer diagnosis through images Palatnik de Sousa et al. (2019) and 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 novel proposes an interpretable machine learning approach to diagnose COVID-19 on suspected cases from the Hospital Israelista Albert Einstein. Interpretation will be carried out by analyzing the feature importance of a linear model and explaining the non-linear models' predictions 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 presents varied clinical specifics, from asymptomatic to Acute Respiratory Distress Syndrome (ARDS). The typical clinical features of this disease are fever, which is the most common symptom among mild to moderate cases. The cough was the second most common symptom observed in recent studies Huang et al. (2020). Also, loss

of smell may, therefore, be a strong predictor of COVID-19 infection. Other symptoms reported include headache, rhinorrhea, sore throat, and fatigue. Chest pain and tightness were reported. Varied gastrointestinal reactions were mentioned, including diarrhea, nausea, vomiting, poor appetite, and anorexia. Ocular reaction, skin rash, and conjunctivitis have also been described. Thus, they are indistinguishable from other respiratory infections. 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 hospitals. Current diagnostic tests for corona virus include the reverse-transcription polymerase chain reaction (RT-PCR). In the laboratory, the amplification of the genetic material extracted from the patient's saliva or mucus sample is through the RT-PCR, which is the synthesis of a double-stranded DNA molecule from an RNA mold. Once the genetic material is satisfactory, the search is for those regions of the genetic code of the CoV that are conserved Cascella et al. (2020). In case of a positive result, it is suggested that the test is repeated for affirmation. In patients with confirmed COVID-19 diagnosis, the laboratory evaluation should be repeated to evaluate for viral clearance before 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).

Furthermore, Serological testing for SARS-CoV-2 is now becoming widely available. Serum samples should be screened for the presence of COVID-19 virus-specific antibodies using serological essays Krammer and Simon (2020). Serological testing may be carried out using enzyme-linked immunosorbent assay (ELISA), immunoflu-orescence (IFA) or, in case of limited lab capacity, Rapid Diagnostic Tests (RDT). The RDT is becoming an essential tool for the early diagnosis of SARS-CoV-2, particularly in situations with limited access to molecular methods Porte et al. (2020). Despite the various antibody tests designed, to this date, the serologic diagnosis has limitations in both specificity and sensitivity. Serologic essays, however, can have a crucial role in broad-based surveillance of COVID-19 Cascella et al. (2020).

Until now, there is no specific antiviral treatment that has been proven to be useful for COVID-19. Also, there is no vaccine available. Non-invasive (NIV) and invasive mechanical ventilation (IMV) may be required in cases of respiratory failure refractory to oxygen therapy Huang et al. (2020). Therefore, diagnostic testing is critical to an effective response to the novel coronavirus, and intensive care is needed to deal with complicated forms of the disease to avoid further aggravations Cascella et al. (2020).

3. METHODS

Three models were chosen to access the COVID-19 prediction problem: Random Forest, Support Vector Machine (SVM) and the Logistic Regression.

To address the interpretability analysis, Shapley Additive Explanations (SHAP) method will be used on the non-linear classifiers and the logistic regression features weights will be evaluated separately.

3.1 Logistic Regression Classifier

The Logistic Regression 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 the it fits the input features into linear a regression problem. The decision of the model is given by:

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

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

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 analysing the values of the β feature weights.

3.2 Random Forest

The Random Forest Classifier is Random Forest (RF) 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's output by individual trees, while in the later, it presents the mean of the individual trees 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 ≤ 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.

3.3 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 non-linear 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 is 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).

3.4 Shapley Additive Explanations

In order to address the interpretability of models, Lundberg and Lee (2017) proposed the Shapley Additive Explanations (SHAP) as a local explanation to single predictions or a group of predictions.

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. SHAP specifies the explanation Molnar (2020) for a prediction as:

$$g(x') = \phi_0 + \sum_{n=1}^{M} \phi_n x'_n \qquad (3)$$

Where ϕ are the weights (SHAP values) for each feature of each sample.

The SHAP values estimation steps are the following:

(1) Randomnly 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 or SHAP Tree

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

$$L(f, g, \pi_x) = \sum [f(h_x(z')) - g(z')]^2 \pi_x(z') \qquad (4)$$

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 (2020).

The SHAP Tree Explainer is a simplified version of the SHAP Kernel and can only be used with tree based methods Lundberg et al. (2020). 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.

4. EXPERIMENTAL ASSESSMENT

In this section, the specific procedure of applying the methods mentioned above is presented with more technical details. The main steps consisted of data acquisition, pre-processing, feature selection, model selection, and interpretation of models and predictions. All the algorithms implemented were built-in Python 3.7, and code is available at the url: www.github.com/ lucasthim/covid19-prediction.

4.1 Data Acquisition

The experiment was executed with data provided by the Hospital Israelita Albert Einstein (São Paulo, Brazil), through a charitable Kaggle competition, and it is available at www.kaggle.com/einsteindata4u/covid19.

Since the primary goal was to help Brazil fight the COVID-19 pandemic, there was no prize, and no leaderboard with top results was present. The major competition tasks were to confirm suspected cases and also predict admission to the general ward, semi-intensive unit or intensive care unit among confirmed COVID-19 cases.

The database consisted of clinical spectrum data of 5647 suspected COVID-19 cases that were admitted to the hospital during March of 2020. Initially, there were a total of 107 features in the database and a flag identifying the result of the COVID-19 exam. The features comprised of blood, pressure, urine and hormonal tests, age, and the presence of other diseases such as other types of corona virus and influenza.

All the features were already normalized with mean equal to zero and standard deviation equal to one. Therefore, the real sense of the range of the variables was lost, especially for the age.

Despite having plenty of features, most of them had missing values in over 80% of the entire database. It happens because the data was a gathering of different clinical trials performed by several different doctors under diverse circumstances. Thus, by not having a mandatory clinical trial script, each doctor requested the most suitable exams and tests for their patients under their care.

4.2 Pre-processing

Since there was no prior knowledge about the original nature and range of each feature, deleting missing data was performed without imputing or replacing any value. Features (columns) with over 90% missing data were dropped, and patients (rows) with over 10% of missing data. Also, features containing only one value were removed. The goal of this step was to eliminate most of the missing values,

but keeping as many features from different natures as possible, i.e., from blood tests, urine, diabetes, etc.

Thereby, a database with 362 patients and 28 features was obtained with no missing values. Among those, approximately 10% of the suspected cases are indeed a positive confirmation of COVID-19. These features consisted of age, blood test variables, and also disease exams.

As mentioned, there was no prior knowledge about the real values of the features; hence they were normalized with a maximum value of 1 and a minimum of 0. Thus, it is possible to have a better sense of the feature's limits. This last step was essential to help interpret the models and the predictions.

4.3 Feature Selection

First, a linear correlation analysis was done to eliminate highly correlated features. Then, a feature selection process was implemented with the Recursive Feature Elimination (RFE) Guyon et al. (2002).

The RFE consists of ranking features based on a chosen type of feature importance and recursively eliminating the least important features. This process is done until no more features can be removed, and a subset of features is chosen based on the grouping with the highest ranking.

To mitigate the effect of feature interactions and also variance in data, the RFE was executed with a 5-fold Cross-Validation, and the optimal set was chosen based on the average of all folds.

4.4 Model Selection

Several hyper-parameter values were tested using a 10-fold cross-validation for the SVM, Random Forest and Logistic Regression. The chosen metric for hyper-parameter tuning, as well as scoring the RFE process, was the F-2 measure.

Given the nature of the COVID-19 infectious spread, it is important that the subject tested positive remains in social isolation and under quarantine as long as necessary. A false positive case that stays under quarantine until the suspicion of the virus is gone is way less dangerous than a false negative case walking freely on the streets.

Hence, F-2 measure is reasonable to guide the model selection to pay more attention to identifying true positives rather than getting more overall predictions correct.

4.5 Interpretation

After selecting and testing models, the interpretation of the results and models was carried with the SHAP method. However, since the Logistic Regression classifier can be directly interpreted by analysing its feature weights (β), the SHAP method will be applied only in the non-linear models of this work, the Random Forest and the SVM.

SHAP values for the Random Forest are calculated with the Tree Explainer technique, since it is a faster way of doing it. For the SVM, SHAP values are calculated with the Kernel Explainer.

The SHAP values offer a wide variety of interpretations. Therefore, the following visualizations were chosen to help interpret the models and explain the predictions: Force plots for single predictions, summary plot of all SHAP values and dependence plots. The summary plot and the dependence plots offer a broader view of how the model interprets a handful of values. Hence, it is possible to have a grasp on the most relevant features for a model, as well as if there is a range of feature value that contributes more for an outcome.

Moreover, inspecting specific prediction paths can help understand possible similarities between different instances. For example, true positives and false positives might have similar paths.

5. RESULTS

As mentioned before, the classifiers were tuned with a 5-fold cross validation on a training set of 253 samples and tested on a holdout test set containing 109 samples.

After the RFE and correlation analysis steps, there were 18 remaning features. The eliminated features were the ones that either showed to be the least important among all three classifiers or, as mentioned before, presented a high linear correlation.

Table 1 presents the summary of the results for the test set on all three classifiers. The confusion matrices for the test set are shown in Figure 1 for the Logistic Regression, Random Forest and SVM, respectively.

Table 1. Results of experiments

Classifier	Accuracy	F2	AUC
Logistic Regression	0.872	**0.844**	**0.899**
Random Forest	**0.908**	0.834	0.896
SVM	0.853	0.813	0.866

Although all the classifiers were optimized for the F2 metric, the Random Forest balanced its false positive and false negative occurrences. In contrast, the Logistic Regression and the SVM classifiers had considerably more false positives. Due to the nature of the problem, one could say it is a good trade-off to have more false positives rather than false negatives. A false-positive patient could easily stay under quarantine for 15 days without having to go on other clinical trials.

Furthermore, since the Logistic Regression presented results as good as the other classifiers, it suggests a good linear relationship between the features and the output.

6. INTERPRETATION OF RESULTS

6.1 Logistic Regression Classifier

As mentioned, the Logistic Regression is an intrinsic interpretable model and the feature weights of the trained model are presented Figure 2. The negative contribution to the model means that the higher the value of a feature, the less likely it will be for a positive case. Hence, the positive contribution means that the higher the value of a feature, the more likely it will be for a positive case. Since all the features from the dataset were normalized between zero and one, feature weights have a direct sign impact on the probability of the outcome. That is, the contribution
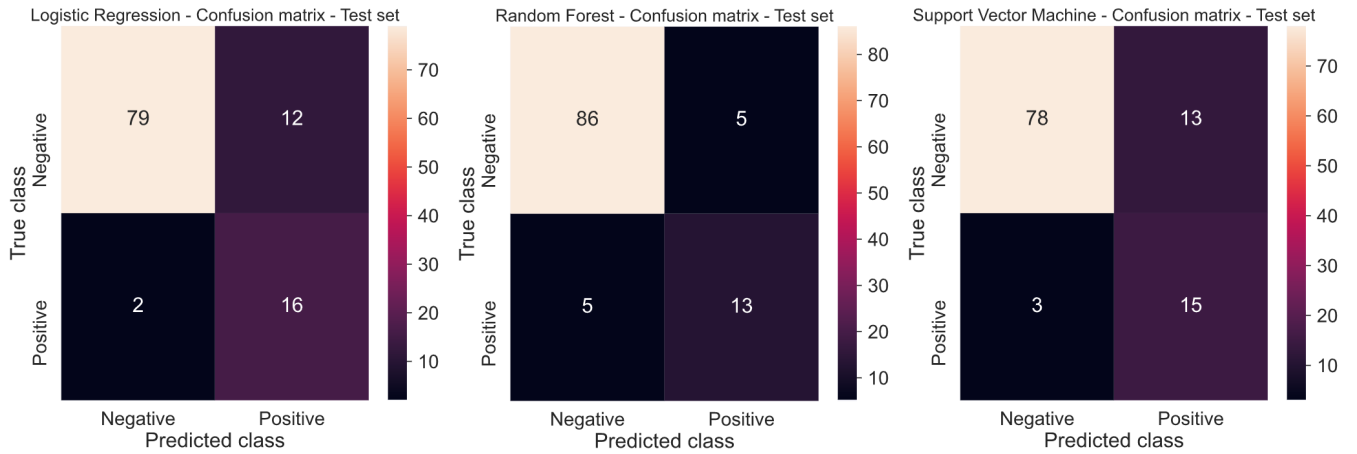
Figure 1. Confusion matrices for test set

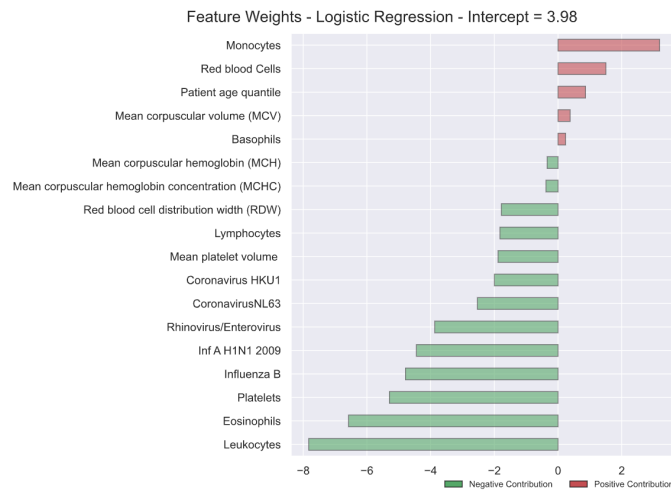of each feature does not switch from positive to negative and vice-versa.

Figure 2. Feature weights - Logistic Regression

For this model, the most critical features are some types of white blood cells (Leukocytes and Eosinophils) responsible for maintaining the immunological system. The platelets, being the third most important variable, are blood cells responsible for stop bleeding. Moreover, it is not clear for a non-expert on the field, why the Monocytes and Red Blood Cells have a considerable positive contribution for this model. Therefore, urther investigation, with the help of an expert on the field, should be carried to understand it.

The features with average importance are the tests for other diseases. When analysing the training set, it can be observed that approximately 37% of the negative cases have at least one of these comorbities (influenza or other type of coronavirus), while all the positive cases for the COVID-19 tested negative for these exams. Therefore, this suggests that testing positive for these comorbities might help on testing negative for COVID-19.

The least important variables are the ones related to the red blood cells (RDW, MCV, MCH and MHCH), suggesting a weak linear relationship between these features and the outcome. It is interesting to notice that the age of

the patient also had little impact on diagnosing suspected cases.

Following next, Random Forest and SVM will be analysed together, since both will take advantage of the SHAP method.

6.2 Random Forest and SVM Classifiers

Since the SHAP method requires a dataset, the entire dataset was used to calculate the SHAP values for both classifiers. It is important to mention that in the determination of the SHAP values, there is no standard recommendation about the dataset to be used. Moreover, there was no significant difference between SHAP values obtained with the entire dataset and the ones calculated using only the training dataset. The best way to start the interpretation using SHAP is to visualize a force plot. Figures 3 and 4 show the predictions for all four distinct cases of True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) for both Random Forest and SVM. A SHAP value can be interpreted as the force (or weight) that each feature has for a single prediction. The objective of a force plot, is to show how a prediction deviates from the base value, which is the average of all predictions, based on the forces (SHAP values) that act upon a sample. Therefore, the blue arrows are the forces that decrease the prediction, and the values in pink arrows increase the predictions. The values above the ruler are graded in the SHAP values scale, and the colored values beside each feature name are the given feature value for a sample. Only the features that contribute the most have their names shown in the plots for better visualization.

The TP and TN cases appear pretty distinct from each other, showing an apparent distinction between positive and negative COVID-19 cases. For the the FP, the forces acting on these plots form a similar path to the TP. This implies that these FP samples might lie inside an uncertainty area where the models struggle to decide which class the sample belongs to. The same stands to the FN compared to the TN cases.

When aggregating all SHAP values from all samples, a summary of the SHAP values can be visualized. Therefore, it is possible to have a grasp on the classifier feature
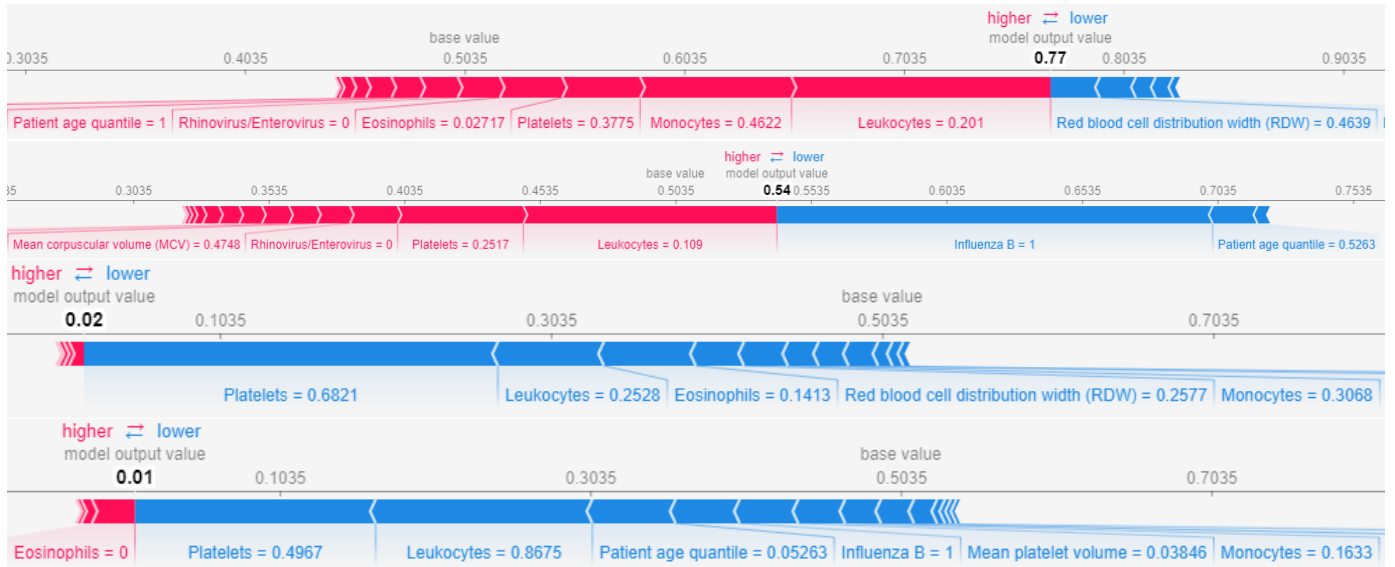
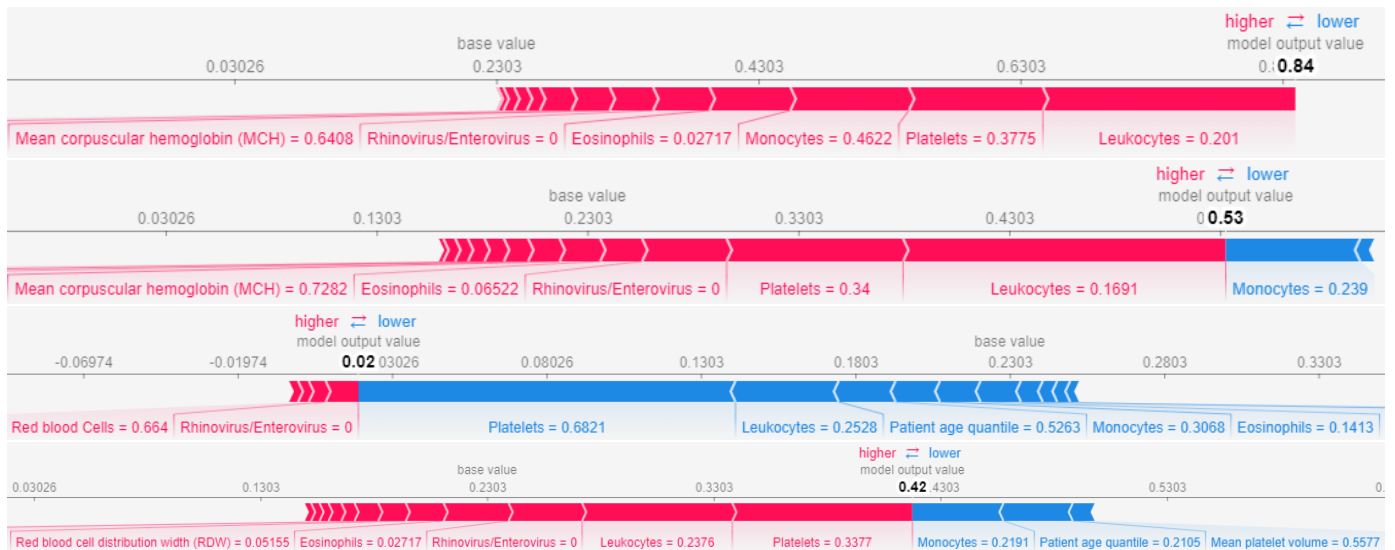Figure 3. Random Forest Single Predictions. From top to bottom: TP, FP,TN,FN

Figure 4. SVM Single Predictions. From top to bottom: TP, FP,TN,FN

importance on a more global level. The summaries of SHAP values are presented in Figure 5 for the Random Forest and SVM, respectively. Each dot represent the SHAP value of a feature of a sample.

According to both summaries, the Leukocytes are an essential feature for these models. Also, the Platelets and Eosinophils are ranked in the top four, which is consistent with the previous findings in the Logistic Regression.

The coronavirus and influenza tests appear with little importance for both models. However, the Rhinovirus/Enterovirus appears in the top four ranks for both models. It is interesting to notice that for the low value of this feature, that is 0, the SHAP values have very little value. However, for a positive value of the feature, the SHAP values gain a more extensive range and higher values. The same happens for Influenza B, but with a narrower range of values.

Contrary to the Logistic Regression, the red blood cell features appeared to be moderately important for both models, suggesting that non-linear models can extract additional information about these features.

Figure 6 show the dependence plots for the Leukocytes and the monotonic decreasing relationship with the SHAP values, as well as the interaction with the Platelets. Also, for higher values of Platelets, the impact of this feature on SHAP values decreases, and for low values of the Platelets, the impact on SHAP values depends more on the Platelets.

7. CONCLUSION

This work presented an interpretable machine learning approach for the COVID-19 diagnosis through clinical variables. Despite the obstacles faced with the impossibility to access the original values of the dataset, and also its limited nature, the results obtained showed to be entirely satisfactory.

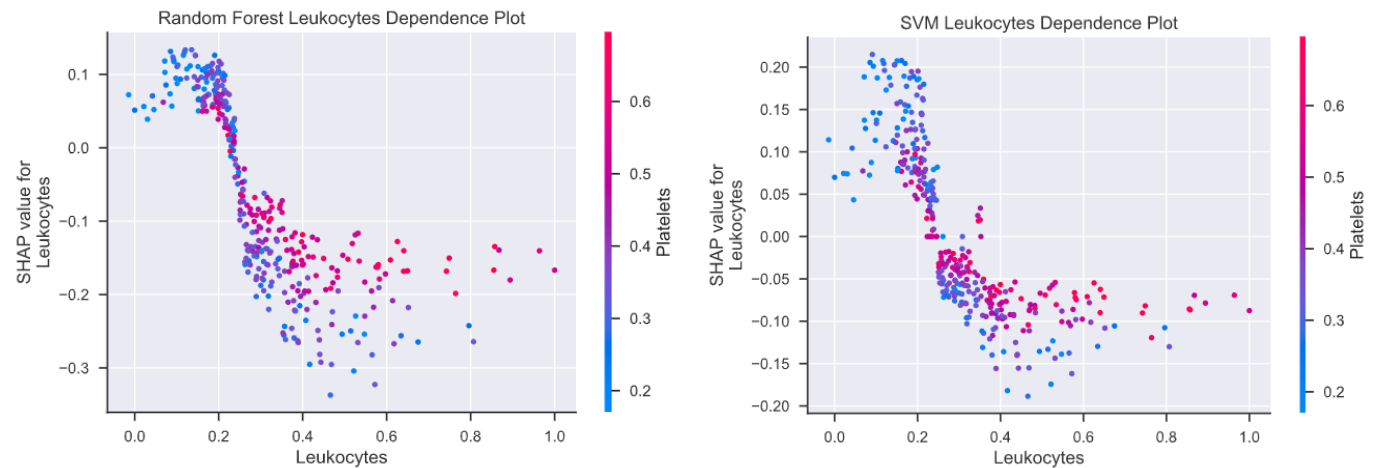Figure 5. Summary of SHAP values for Random Forest and SVM

Figure 6. Leukocytes SHAP dependence plots for Random Forest and SVM

Also, the interpretability analysis shows to be an essential step when applying machine learning to diagnostic medicine. Even when analyzing just the weights of the Logistic Regression, it is possible to get some insights about the features. Moreover, the SHAP values analysis brought specific insights into the models' predictions. It is clear the difference in the profile prediction of a positive and negative patient. It is interesting to notice how close the false cases could be to their relative classifications. These findings might orient doctors on looking for more specific traces of the COVID-19 in the patients. Also, the SHAP values summary plots brought broader interpretability to the models, and the most relevant features among the Logistic Regression, Random Forest and Support Vector Machine were consistent.

The characteristic of how the models interpreted the variables and predicted the cases make this approach more similar to the forementioned COVID-19 rapid tests. These tests seek to identify the disease's presence by analyzing some changes in the immunological system and other variables.

Future works will seek partnerships with the public healthcare system to obtain more clinical variables such as blood pressure, sugar levels, hormones, and urine exams. Also, other interpretability techniques and interpretable models should be applied. Interesting models to the diagnostic medicine are ruled based models, such as the SENFIS Alves et al. (2019). They are fuzzy systems ensemble capable of dealing with a high number of dimensions and offering a small set of rules with a low number of antecedents to a classification problem.

REFERENCES

Alves, E.D.S., Vellasco, M., and Tanscheit, R. (2019). SENFIS - Selected Ensemble of Fuzzy Inference Systems. Library Catalog: www.researchgate.net.

Amaral, J.L.M., Lopes, A.J., Jansen, J.M., Faria, A.C.D., and 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. doi:10.1016/j.cmpb.2011.09.009.

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

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

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

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S.C., and Di Napoli, R. (2020). Features, Evaluation and Treatment Coronavirus (COVID-19). In *StatPearls*. StatPearls Publishing, Treasure Island (FL).

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

Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., and Lin, C.J. (2019). LIBLINEAR: A Library for Large Linear Classification. 31.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1), 389–422. doi:10.1023/A:1012487302797.

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., and Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet (London, England)*, 395(10223), 497–506. doi:10.1016/S0140-6736(20)30183-5.

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

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., and Stumpe, M.C. (2017). Detecting Cancer Metastases on Gigapixel Pathology Images.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. doi:10.1038/s42256-019-0138-9. Number: 1 Publisher: Nature Publishing Group.

Lundberg, S.M. and 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, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, 4765–4774. Curran Associates, Inc.

Molnar, C. (2020). *Interpretable Machine Learning*. URL https://christophm.github.io/interpretable-ml-book.

Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, M., and Costa da Silva, E. (2019). Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases. *Sensors (Basel, Switzerland)*, 19(13). doi:10.3390/s19132969.

Porte, L., Legarraga, P., Vollrath, V., Aguilera, X., Munita, J.M., Araos, R., Pizarro, G., Vial, P., Iruretagoyena, M., Dittrich, S., and 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*. doi:10.1016/j.ijid.2020.05.098.

Potie, N., Giannoukakos, S., Hackenberg, M., and 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)*, 1–6. doi:10.1109/FUZZ-IEEE.2019.8858976. ISSN: 1558-4739.

Singh, P., Singh, S.P., and 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*, 1–6. doi:10.1109/CICT48419.2019.9066250.

Singhal, T. (2020). A Review of Coronavirus Disease-2019 (COVID-19). *Indian Journal of Pediatrics*, 87(4), 281–286. doi:10.1007/s12098-020-03263-6.

Tjoa, E. and Guan, C. (2019). A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI.

Vapnik, V.N. (2000). *The Nature of Statistical Learning Theory*. Springer New York, New York, NY. doi:10.1007/978-1-4757-3264-1.

WHO (2020a). Coronavirus Disease (COVID-19) Situation Report 153. URL https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports. Library Catalog: www.who.int.

WHO (2020b). WHO announces COVID-19 outbreak a pandemic. URL 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.

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

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*, 9. doi:10.3389/fgene.2018.00515.