Performance and explainability of feature selection-boosted tree-based classifiers for COVID-19 detection

Jesús Rufino, Juan Marcos Ramírez, Jose Aguilar, Carlos Baquero, Jaya Champati, Davide Frey, Rosa Elvira Lillo, Antonio Fernández-Anta

Performance and Explainability of Feature Selection-Boosted Tree-based Classifiers for COVID-19 Detection

Jesús Rufino[a], Juan Marcos Ramírez[a,*], Jose Aguilar[a,b,c], Carlos Baquero[d], Jaya Champati[a], Davide Frey[e], Rosa Elvira Lillo[f] and Antonio Fernández-Anta[a]

[a]IMDEA Networks Institute, 28918, Madrid, Spain
[b]CEMISID, Universidad de Los Andes, Mérida, 5101, Venezuela
[c]CIDITIC, Universidad EAFIT, Medellín, Colombia
[d]Universidade do Minho and INESC TEC, Braga, Portugal
[e]Inria, Rennes, France
[f]Universidad Carlos III, Madrid, Spain

ARTICLE INFO

Keywords:
COVID-19 Detection
Explainability Analysis.
Gradient Boosting Classifiers
Random Forest
Recursive Feature Elimination
Shapley Values

ABSTRACT

In this paper, we evaluate the performance and analyze the explainability of machine learning models boosted by feature selection in predicting COVID-19-positive cases from self-reported information. In essence, this work describes a methodology to identify COVID-19 infections that considers the large amount of information collected by the University of Maryland Global COVID-19 Trends and Impact Survey (UMD-CTIS). More precisely, this methodology performs a feature selection stage based on the recursive feature elimination (RFE) method to reduce the number of input variables without compromising detection accuracy. A tree-based supervised machine learning model is then optimized with the selected features to detect COVID-19-active cases. In contrast to previous approaches that use a limited set of selected symptoms, the proposed approach builds the detection engine considering a broad range of features including self-reported symptoms, local community information, vaccination acceptance, and isolation measures, among others. To implement the methodology, three different supervised classifiers were used: random forests (RF), light gradient boosting (LGB), and extreme gradient boosting (XGB). Based on data collected from the UMD-CTIS, we evaluated the detection performance of the methodology for four countries (Brazil, Canada, Japan, and South Africa) and two periods (2020 and 2021). The proposed approach was assessed in terms of various quality metrics: F1-score, sensitivity, specificity, precision, receiver operating characteristic (ROC), and area under the ROC curve (AUC). This work also shows the normalized daily incidence curves obtained by the proposed approach for the four countries. Finally, we perform an explainability analysis using Shapley values and feature importance to determine the relevance of each feature and the corresponding contribution for each country and each country/year.

1. Introduction

During the COVID-19 pandemic, healthcare systems have faced significant challenges in developing surveillance strategies to monitor the spread of the disease. Specifically, these strategies require the collection of high-quality data almost in real-time [1]. In this regard, polymerase chain reaction (PCR) tests have been widely utilized to monitor the spread of the infectious disease. However, many factors affect the accuracy of PCR tests, including the timing of the test relative to the infection [2], the high rate of asymptomatic cases [3], and the limited availability of test kits [4]. To overcome these limitations, numerous approaches have been developed that make use of survey data to track pandemic indicators. For instance, [5] and [4] collected self-reported symptoms provided by individuals tested via PCR to evaluate the performance of COVID-19 detection methods. Similar approaches captured self-reported symptoms through smartphone apps to predict potential COVID-19 cases [6, 7, 8]. Social networks have also been used to publicize online questionnaires about symptoms, social behavior, and isolation measures [9].

Several methods have been developed for the detection of COVID-19-active cases based on individual features extracted from survey data. These methods can be categorized into three classes: prediction rules, logistic regression methods, and machine learning models [10]. Prediction rules identify active cases based on a specific set of symptoms.

juan.ramirez@imdea.org (J.M. Ramírez)
ORCID(s): 0000-0003-0000-1073 (J.M. Ramírez); 0000-0003-4194-6882 (J. Aguilar); 0000-0002-5127-8497 (J. Champati); 0000-0002-6730-5744 (D. Frey); 0000-0003-0802-4691 (R.E. Lillo); 0000-0001-6501-2377 (A. Fernández-Anta)

COVID-like illness (CLI) approaches approved by either the Centers for Disease Control and Prevention (CDC) or the World Health Organization (WHO) are the most representative prediction rules [11, 12, 13]. Additional prediction rules have been reported in [14, 2, 15]. Prediction rules were developed as a simple tool for making decisions about hospitalizations and managing healthcare resources efficiently (hospital beds and intensive care units) when antigen tests were not available. These rules typically consider a small number of symptoms with equal importance. On the other hand, methods based on logistic regression build a linear expression whose parameters represent the contribution of the reported features (symptom, gender, age group) [5, 6, 9, 16, 17, 18]. A reduced number of symptoms is also used in these techniques (usually less than five). Finally, machine learning models optimize supervised classifiers using multiple individual features to predict COVID-19 [1, 4]. Nevertheless, machine learning methods only take into account a limited number of features (symptoms, gender, and age) and ignore information provided by features such as vaccination acceptance, isolation measures, and local community information.

In April 2020, the University of Maryland (UMD), in collaboration with Facebook, launched the Global COVID-19 Trends and Impact Survey (UMD-CTIS), a large health surveillance system based on surveys [19, 1]. More precisely, the purpose of this study was to gather daily information from a representative sample of Facebook's Active User Base (FAUB), who were invited to participate in the survey. This instrument collected information about various COVID-19-related characteristics, including symptoms, PCR test outcomes, vaccination acceptance, isolation measures, local community information, mental health, and demographics. Questionnaires were translated into 56 languages, and data were collected from 114 countries/territories, reaching a wide range of social and economic groups. Note also that the UMD-CTIS data provide fine-grained coverage of pandemic trends, which permits estimating various health indicator trends for different regions.

1.1. Contributions

In this paper, we introduce a machine-learning methodology for detecting COVID-19 cases using tree-based supervised classifiers and feature selection strategies. In contrast to prediction rules and logistic regression models, the proposed methodology takes into account a wide range of individual characteristics for COVID-19 detection. For example, the proposed approach considers other factors besides symptoms, including demographics, vaccination acceptance, local community indicators, and isolation measures. As an alternative to previous machine learning approaches, the proposed approach utilizes a feature selection technique based on Shapley values to reduce dimensionality and minimize overfitting risk. Based on Shapley values, the optimal set of features is identified for the optimal balance between model complexity and detection accuracy [20]. Moreover, compared to prediction rules, supervised tree-based classifiers exhibit outstanding detection performance and allow us to recognize the relevant features contributing to detection tasks. We implemented six versions of the proposed methodology using three different classifiers: random forest (RF), light gradient boosting (LGB), and extreme gradient boosting (XGB). In addition, we evaluated the performance of the developed approach using UMD-CTIS datasets extracted from four countries: Brazil, Canada, Japan, and South Africa, for 2020 and 2021. The performance of the proposed methodology was compared to state-of-the-art techniques based on survey data. In general, our approach has outperformed other state-of-the-art methods in terms of different quality metrics including F1 score, sensitivity, specificity, and precision. Subsequently, we obtained the receiver operating characteristic (ROC) curves for the six versions of the proposed method and calculated the corresponding area under the ROC curves (AUC). As an illustrative example, we determined the normalized COVID-19 daily incidence using the proposed detection approach and compared the generated curves with those generated by official reports covering the four countries from April 2020 to June 2022. A further explanationability analysis has been performed in this study to identify the relevant input features and to outline how they contribute to the detection task[21, 22].

1.2. Related Work

Recent studies have focused on the detection of COVID-19 using explainable machine-learning models. In particular, two machine learning techniques were discussed in [23], the multilayer perceptron artificial neural network and the decision tree. These techniques were used to predict the severity level of COVID-19 patients based on their medical history and laboratory test results. Moreover, a LIME approach was also used to evaluate the explainability of predictions produced by machine learning models. Furthermore, Girardi et al. [24] designed machine learning models (Random Forest, Neural Network, and Time Convolutional) to predict hospitalizations in COVID-19 positive-tested. Additionally, a study of the SHAP values to define the feature importance for the models in different scenarios reveals a high degree of variability across models.

In [25], an effective COVID-19 explanation was developed based on user-centered principles. More precisely, they discussed how to apply an interdisciplinary, user-centered approach based on Design Thinking to develop a prototype of a user-centered explanation regarding people's perception of COVID-19 vaccine development. In [26], machine learning and explainability methodologies were used to construct an aggravation risk score and analyze the effects of COVID-19 features. Age, chest CT severity, and biological variables such as CRP, oxygen saturation, and eosinophil counts were the most important factors. The work reported in [27] discussed the importance of self-organizing maps to interpret hospital data. Particularly, the COVID-19 epidemic was analyzed in detail to understand data patterns and topologies. They determine the most significant variables with networks and topological mapping, which solve this problem by mapping high-dimensional data into lower-dimensional representations based on the overall association.

For classification tasks of CT-COVID-19 images containing clinical findings of COVID-19 from 216 patients, Phongchit et al. [28] studied well-known neural network models (ResNet50V2, DenseNet169, Xception, and EfficientNet B4) to evaluate their performance and explainability. They concluded that the models producing the same COVID-19 classification result might rely on a large number of different features. This implication suggests that although we tend to select the model that performs best in metrics, in a clinical environment, it can be better to assess explanations from them. On the other hand, Aldhahi et al. [29] defined a method to train deep learning models in classifying COVID-19 chest X-rays from normal and pneumonia-related infections, using a training scheme that integrates the cyclic cosine annealing approach with cross-validation and uncertainty quantification. Additionally, they introduced an image processing technique to measure explainability based on ground truth. Ali et al. [30] develop a Convolutional neural network (CNN) model densely connected squeeze convolutional neural network for the classification of X-ray images of COVID-19, pneumonia, normal, and lung opacity patients. Then, to ensure model trust and explainability, they applied two explainable techniques, Grad-CAM and LIME. The goal of the work of Saxena et al. [31] was to detect disease in persons who had an X-ray image. Chest X-rays of COVID-19 patients, viral pneumonia patients, and healthy patients were obtained from different sources. These three groups were classified using deep learning and multiclass classification models. Then, they added a discussion about the explainability of the models. Li et al. [32] developed a multi-task learning framework in which COVID-19 diagnosis and multi-lesion recognition (segmentation of CT images) are achieved simultaneously. The framework is based on an explainable multi-instance multi-task network, which learns task-related features adaptively, and gives explicable diagnosis results by suggesting local CT images with lesions as additional evidence. Finally, severity assessment of COVID-19 and lesion quantification are carried out.

Other recent research in 2023 has continued to study the ability of machine learning models to analyze images of patients with COVID-19, but none consider aspects of explainability. For example, Kathamuthu et al. [33] used several enhanced CNN approaches using transfer learning to detect COVID-19 in chest computed tomography (CT) images. VGG16, VGG19, Densenet121, InceptionV3, Xception and Resnet50 are the basic models used in this work to apply transfer learning. Another work in the same line was proposed by Deeb et al. [34], who propose a CNN, called AdjCNet, that focuses on grayscale variations between adjacent areas within a CT image. The work of Ullah et al. [35] defines a multi-task semi-supervised learning (MTSSL) framework for performing COVID-19 detection in chest x-rays (CXR), which solves the problem of limited amount of labeled data in this domain. MTSSL uses auxiliary tasks for which adequate data are publicly available, specifically, pneumonia, lung opacity, and pleural effusion, which enrich the primary task of COVID-19 detection. MTSSL uses an unsupervised adversarial autoencoder (AAE) to learn and discriminate features and supervised classification networks for COVID-19 detection. Finally, Ershadi et al. [36] considered a special set of characteristics fusing clinical and image data to find treatment plans in groups of patients with COVID-19. They propose a hierarchical model based on expert knowledge to group patients, and then build classifier systems for each group. To design the proposed hierarchical model, they used a fuzzy C-mean (FCM) clustering for clustering tasks and an adaptive neuro-fuzzy inference system (ANFIS) classifier for classification tasks. As can be seen, recent works continue mainly along the same line of image processing, but without carrying out an explainability analysis of the results. Only in the work of Ershadi et al., an FCM-ANFIS approach is proposed that allows an explainability analysis process by using these white box techniques, but this is not considered in the article. To compare this work with the previous one, we proceeded to define four criteria, which are:

- **Criterion 1:** The work performs a feature selection stage.
- **Criterion 2:** The work carries out an explainability analysis using different approaches.
- **Criterion 3:** The work considers the feature explainability during the feature selection stage.

Table 1
Criteria covered by various COVID-19 detection approaches.

Method	Criterion			
	1	2	3	4
Gabbay et al. [23]	✓	✓	✗	✓
Girardi et al. [24]	✗	✓	✗	✗
Novak et al. [25]	✗	✓	✗	✗
Excoffier et al. [26]	✗	✓	✗	✓
Yu et al. [27]	✗	✓	✗	✗
Phongchit et al. [28]	✗	✓	✗	✗
Aldhahi et al. [29]	✗	✓	✗	✗
Ali et al. [30]	✗	✓	✗	✓
Saxena et al. [31]	✗	✓	✗	✓
Li et al. [32]	✗	✓	✗	✗
Ershadi et al. [36]	✗	✗	✗	✓
Our work	✓	✓	✓	✓

- **Criterion 4:** The work considers other factors besides symptoms for COVID-19 detection.

Table 1 shows the criteria covered by the different COVID-19 detection approaches. Notice that the first criterion is satisfied only by [23] and our methodology. In contrast, our approach is the only one that considers feature explainability as a primary consideration during the feature selection stage (third criterion). According to the second criterion, all of the studies meet this standard, which confirms that explainability is one of the most important aspects to consider when exploring medical machine-learning applications. The fourth criterion is met by several studies that utilize a variety of sources in addition to COVID-19 symptoms.

In general, X-ray images of the lungs have been studied for COVID-19 detection; but there are also works that consider other variables such as symptoms, among others. On the other hand, there is an effort to make these models explainable, particularly those based on X-ray images. However, there are no works that seek to select the features automatically (in our case, using recursively the Shapley values), and from there, develop detection models of COVID-19, nor an in-depth analysis of the behavior (both performance and explainability) of self-explaining methods (such as those based on decision trees). In our case, the analysis took into account both the total and the selected characteristics.

Unlike these approaches, our methodology makes COVID-19 predictions and analyzes the explainability of the trained models from survey data that includes a wide range of individual features. This collection strategy enables disease monitoring in almost real-time using limited healthcare resources. Additionally, our methodology selects features based on their explainability, which is determined by the Shapley values.

1.3. Paper organization

This paper is organized as follows. Section 2 describes the tree-based classifiers used to detect COVID-19-positive cases and the explainability analysis approaches used to identify relevant features. We introduce the methodology for detecting COVID-19-active cases in Section 3. Section 4 shows extensive results on the performance of the proposed methodology and the corresponding explainability analysis. The discussion generated by both the performance evaluation and the explainability analysis is summarized in Section 5.

2. Materials and methods

In this section, we describe the UMD-CTIS dataset, machine learning methods used to identify COVID-19-positive cases, and the feature selection technique used.

2.1. Machine Learning Methods

Machine-learning (ML) models used in this work are tree-based classifiers [37, 38]. Specifically, we will use random forest (RF), extreme gradient boosting (XGB), and light gradient boosting (LGB). These models provide rankings of the relevance of input variables, which serves as the basis for performing an explainability analysis process. Furthermore, tree-based models were selected due to their outstanding performance in several applications and the low

training times when the input vector is high-dimensional [39, 40]. It should be noted that our feature selection stage involves iterative training steps that use a large number of variables, so this last aspect is of great relevance in our study. Now, each model is presented:

- **Random Forest (RF)** [41, 37, 38]: RF consists of a set of decision trees, each generated by a bagging algorithm. These trees form a "forest" of trees voting for a specific result. This algorithm uses bootstrapping to fit decision trees with sub-samples of the original dataset. For each tree, the algorithm uses averaging to improve predictive accuracy and control overfitting. In classification algorithms, the most common output will be chosen as the final output of the algorithm.

- **Extreme Gradient-Boosting (XGB)** [42, 37, 38]: It is a class of ensemble machine-learning algorithms that can be used for classification problems. Ensembles are constructed from decision-tree models. The trees are added one at a time to the ensemble, and fitted to correct for prediction errors made by previous models. This is done using data that could not be learned so far. This technique is known as boosting. Moreover, XGB applies a regularization technique to reduce overfitting.

- **Light Gradient-Boosting (LGB)** [42, 37, 38]: LGB is a gradient-boosting algorithm based on decision trees which decreases memory usage and improves model efficiency. It uses two techniques: the classic XGB based on Exclusive Feature Bundling, and the Gradient-driven One Side Sampling (GOSS). GOSS keeps those instances with large gradients (they will contribute more to information gain) and randomly drops those instances with small gradients to improve accuracy estimation. It is faster than the XGB algorithm because GOSS filters out the data instances to find a suitable split value. This makes the process faster.

2.2. Explainability Analysis Methods

In medicine, there is an increasing demand for AI approaches that are both efficient and transparent, as well as easily explainable by a human expert [21, 22]. Currently, it is difficult to find explanations as to why a result occurs or how a model describes the underlying biological process [43]. In COVID-19 studies that use machine learning models, explainable AI is urgently needed to understand and retrace the machine's decision-making process. It is critical, for example, to analyze the relationships between symptoms, age groups, gender, and COVID-19 cases. On the other hand, explaining, interpreting, or understanding, are synonymous in the context of explainable AI, and various approaches have been proposed [43]. As a first categorization, the explanator must describe the model or result (e.g., classification, prediction). Therefore, this classification is whether the explicability is *global* to provide insights into the inner workings of the entire model for some specific dataset or *local* for a single test input x and its corresponding result y [44, 43]. There are two types of explicability: *ante-hoc* consists of building it directly from the beginning of model creation (the model can be understood immediately), and *post-hoc* consists of building it after model creation using a technique to extract the explanation [44, 43].

Global explanators try to reveal certain properties of the model independently of results. An example is the tree-based approach (e.g., decision trees and ensembles of decision trees, such as RF) [41, 45, 42]. In this case, the information gained from a variable accumulated over all trees can be used as a relevance measure. Another example of a feature importance metric for tree-based methods is the feature's depth in the tree. Local explanations are only valid near a result. The classic methods relate the model result to the feature vector by ranking the explanatory power, i.e. the salience of each feature. There are two main families of methods. The first, *Attention Based Models*, examines the most promising parts of input features that lead to a certain output for a given task. For a given output, they try to find out whether input features with high attention weights were responsible for the outcome. The second is *feature-attribution approaches* that explicitly try to quantify the contribution of each individual feature to the results. In this work, we analyze a global ante-hoc approach based on the tree-based methods using the feature ranking provided by them and a local and post-hoc approach that extracts the feature importance for a given input based on the fact that output can be written as the sum of bias and single feature contributions (Shapley values).

- **SHapley Additive exPlanations (SHAP values) [20, 46]:** Shapley values are an example of local approaches [20, 46], which are created by means of a method from coalitional game theory, assuming that each feature value of the instance is a player in a game where the output is the payoff. Shapley values dispense the payoff among the features. The goal of SHAP is to explain the output of a model by computing the contribution of each feature to this output. For that, SHAP computes Shapley values. A Shapley value assigns each feature an importance

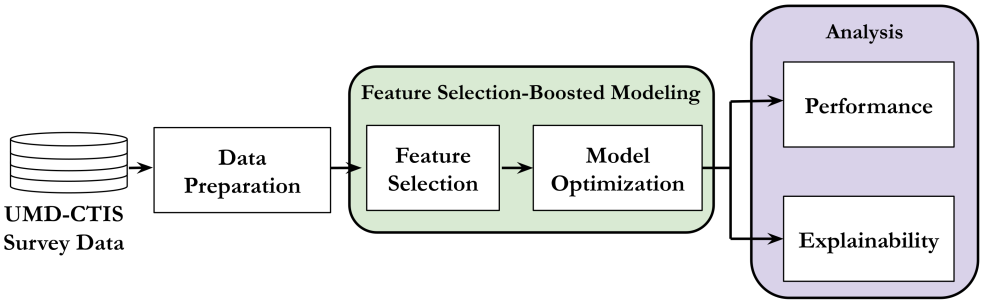

Figure 1: Flowchart illustrating the methodology for detecting COVID-19-active cases based on classifiers and feature selection and the corresponding performance evaluation and explainability analysis.

value for a particular output to define the explanation. This value, for feature i, is the unified measure of additive feature attributions (ϕ_i) [20, 46]:

$$\phi_i = \sum_{S \in F \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{1}$$

Where F is the set of input features, S is a subset of input features, M is the number of input features and $f_S(.)$ represents the output function of the model (e.g. prediction). In this equation, $\sum_{S \in F \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!}$ represents the weighted average of all possible subsets of S in F. In addition, this equation considers the difference between when this feature is present in the output ($f_{S\cup\{i\}}(x_{S\cup\{i\}})$ and when it is absent ($f_S(x_S)$). With these SHAP values, we are able to select the variables that give the model the highest contribution. The calculation time increases exponentially according to the number of features. To avoid this, one solution is to determine contributions for only a few samples of the possible coalitions. We have used a sample of 100 to compute the importance of each feature using the kernel-explainer function.

- **Tree-based methods [41, 45]:** The explainability analysis for tree-based ML methods is possible by their capabilities to rank their features/variables. These methods make it possible to compute various feature/variable importance measures to be used in an explainability analysis. For example, MDA (mean decrease in accuracy) determines feature importance (ranking) as the mean decrease of accuracy over all predictions, when a given variable is permuted after training [41, 45]. Thus, MDA calculates the average decrease of accuracy against random permutations of feature values in different cases. The cases use a trained model, and each tree of the model is permuted along the m-th feature and the average of these differences for all decision trees gives the m-th feature's MDA. Also, it can use the Gini value, which measures the average gain of purity by splits of a given variable [41, 45]. If the variable is useful, then it tends to split labeled nodes into single classes. The permuted variables tend neither to increase nor to decrease node purities. Permuting a useful variable tends to yield a large decrease in mean Gini gain. Gini importance is normally inferior to variable importance because it is more unstable and biased.

Finally, in some cases, the large number of features could be a problem because of the noise produced by some of the variables or their low significance. There are some ML algorithms that already have some regularization mechanisms that reduce the number of features. However, the techniques being used in this study do not have any of these mechanisms. For this reason, we also use SHAP values as a feature-reduction technique that we apply to each of them. Ultimately, we end up with 6 different models, 3 of them with all the features (RF, LGBM and XGB), and 3 with the features selected according to the Shapley values (RF.SHAP, LGBM.SHAP and XGB.SHAP).

3. Experimental Protocol

Figure 1 illustrates the flowchart of the methodology for identifying COVID-19-positive cases from self-reported information using feature selection-boosted tree-based models. The datasets to be evaluated are extracted from the UMD-CTIS survey records. As shown in this figure, the methodology is divided into three stages: data preparation, feature selection-boosted modeling, and analysis. Data preparation involves data preprocessing and data understanding.

Table 2
Characteristics of the study population for the various countries and two non-overlapped periods (2020 and 2021).

Characteristic	Brazil		Canada		Japan		South Africa	
	2020	2021	2020	2021	2020	2021	2020	2021
1. Number of responses	3470298	1669105	627813	282914	2132918	918147	329528	122629
2. Tested symptomatic	83238	262683	8927	33997	4698	41010	7883	23038
(a) Training size	66590	210146	7142	27198	3758	32808	6306	18430
(b) Test size	16648	52537	1785	6799	940	8202	1577	4608
3. Test outcome								
(a) Positive	44963	106471	838	3433	532	4011	2866	8459
(b) Negative	38275	156212	8089	30564	4166	36999	5017	14579
(c) TPR	54.02%	40.53%	9.39%	10.10%	11.32%	9.78%	36.35%	36.71%
(d) Minimum test size	1527	1482	523	558	617	542	1422	1428
4. Gender								
(a) Female	45357	130235	5438	19472	1679	14283	3923	11291
(b) Male	24928	76689	2315	9824	2388	20791	2525	6730
5. Age groups								
(a) 18-24	8270	27474	1136	3248	179	871	739	1580
(b) 25-34	19596	56227	2337	7172	577	3797	2252	4889
(c) 35-44	21061	57452	1750	6688	997	7527	1801	4721
(d) 45-54	13776	39122	1210	5215	1216	10413	1141	3878
(e) 55-64	6968	22190	954	4478	828	8724	491	2124
(f) 65-74	140	6016	308	2421	479	3529	1667	799
(g) 75+	233	1025	126	825	66	846	27	230
6. Average number of reported symptoms	4.33	3.80	3.38	3.07	2.92	2.49	3.82	3.95
7. Average number of reported symptoms among positive	5.37	5.16	5.25	5.27	4.38	4.45	5.51	5.61

Data preprocessing includes filtering techniques to extract the target datasets from UMD-CTIS data for the countries and periods of interest. Moreover, we perform a descriptive analysis of the dataset to gain more insight into the main characteristics of the study population. Feature selection and model optimization are the two steps in feature selection-boosted modeling. We implement a feature selection technique based on the RFE approach to reduce the number of variables without compromising performance. Then, the selected variables are used to optimize the tree-based supervised classification models. Finally, we conduct performance evaluation and explainability analysis from the outcomes yielded by the tree-based classifiers.

3.1. Dataset Preparation

The University of Maryland (UMD), in partnership with Facebook, launched the Global COVID-19 Trends and Impacts Survey (UMD-CTIS), an extensive remote health surveillance system to monitor the COVID-19 pandemic evolution. More precisely, the UMD-CTIS collected self-reported data on approximately 120 indicators related to COVID-19, such as symptoms, age groups, gender, demographics, isolation measures, vaccination acceptance, and mental health, among others. In addition, the survey was run daily from April 23, 2020, to June 25, 2022 [1] (the questionnaire of the UMD-CTIS survey is shown in Supplementary Material 1). In this study, we extracted UMD-CTIS data from four countries: Brazil (BR), Canada (CA), Japan (JP), and South Africa (ZA). Geographic diversity and the availability of sufficient samples were considered when selecting these countries. Moreover, we considered data for two periods: 2020 (April 23 - December 31, 2020) and 2021 (January 1 - December 31, 2021). We selected these periods to observe the impact of vaccination campaigns on both the feature selection stage and the tree-based model optimization. Table 2 provides a summary of the population characteristics for the four countries and 2020 and 2021. As can be seen, there are 9,553,352 survey responses for all countries and periods (BR: 3,470,298, CA: 627,813, JP: 2,132,918, ZA: 329,528, for 2020; and BR: 1,669,105, CA: 282,914, JP: 918,147, ZA: 122,629, for 2021).

Since survey data contain categorical data only, we first apply binary encoding to the dataset extracted from each country and period. Therefore, we have a column with binary elements for each potential response. For 2020 and 2021, binary encoding generated datasets with 417 and 614 columns, respectively. Surveys for 2021 include additional questions related to the vaccine campaigns. Then, we extracted samples from participants who had reported at least one symptom during the previous 24 hours and who had provided a test result within the past 14 days. Samples with symptomatic reports were selected to compare the performance of the proposed approach with respect to previously developed methods based on self-reported symptoms. In addition, we considered the samples with antigen test results

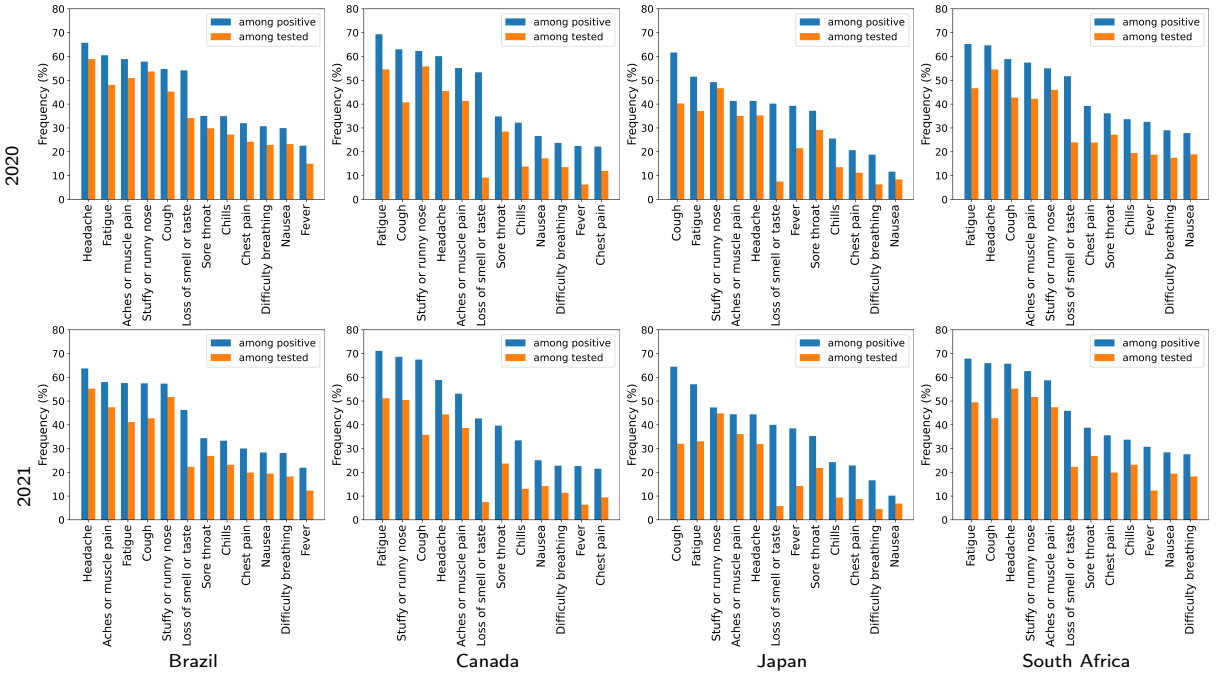

Figure 2: Rate of tested positives reporting a particular symptom in 2020 and 2021 for the four countries. Bar plots also show the percentage of tested symptomatic reporting each symptom.

to have a ground truth to train and test the detection models. For 2020, we analyzed 104,746 respondents (BR: 83,238, CA: 8,927, JP: 4,698, and ZA: 7,883) who reported at least one symptom in the last 24 hours and provided a test result within the past 14 days (Tested symptomatic). Similarly, we extracted data from 370,728 individuals in 2021 (BR: 262,683, CA: 33,997, JP: 41,010, and ZA: 23,038 ZA).

Table 2 also includes the number of *positive* and *negative tests* among *tested symptomatic*, as well as the *test positive rate* (TPR), calculated as follows: TPR = (100×positive)/(Tested symptomatic), for each country and period. Notice that the TPR values obtained for Brazil and South Africa are at least three times larger than those yielded by Canada and Japan for 2020 and 2021. Table 2 displays information on other individual features such as gender, age group, the average number of reported symptoms per questionnaire, and the average number of reported symptoms per questionnaire among positives. In contrast to previous approaches that take into account a reduced set of individual features, our approach considers the full set of features collected by the UMD-CTIS questionnaires. Figure 2 depicts the percentage of tested positives who reported a particular symptom for each country and period in descending order. In addition, Figure 2 illustrates the corresponding rate of tested symptomatic reporting each symptom. As can be seen in this figure, fatigue is the most common symptom among positives, with the highest rates in the bar plots. The first conclusion is that, in general, the symptom patterns vary among countries and for 2020 and 2021. Finally, it is important to observe that the selected datasets are subject to potential bias sources, which may affect both the accuracy of the results as well as the identification of significant variables [1, 15]. As an example, the set of respondents is not a random sample of the population, since invitations to participate in the survey were sent to Facebook users. Furthermore, since CTIS-UMD provides COVID-19 signals based on self-reports, some indicators may differ from those obtained from more objective tests due to various sources of measurement error, such as recall bias and social desirability bias. However, we assume that these bias sources do not affect the results because of the amount of data used and they do not change rapidly over time, therefore, the signals reflect mean behavior during periods of interest (2020 and 2021).

3.2. Feature Selection-Boosted Modeling

3.2.1. Feature Selection

As seen in Section 3.1, our methodology considers all variables extracted from each country and period. Notice that a large number of variables do not necessarily lead to performance improvement of the detection models and are typically associated with problems such as long training times and model overfitting [47]. Therefore, we include a

feature selection stage to reduce the number of variables without compromising the detection performance. Moreover, this stage enables us to exclude irrelevant and redundant variables, thus simplifying models and boosting their explainability [48]. To implement the feature selection stage, we use the recursive feature elimination (RFE) approach proposed in [49] based on Shapley values. Algorithm 1 shows the pseudocode of the RFE method.

Algorithm 1 Recursive Feature Elimination (RFE) Algorithm

Inputs: $\mathbf{y}, \mathbf{X}, f_{\boldsymbol{\theta}}(\cdot), \gamma$
$M, N \rightarrow$ number of rows and columns of the input dataset $\mathbf{X}$
$\mathcal{L}_0 = \{i \in \{1, \ldots, N\} | \mathcal{L}_{0_i} = i\}$
$f_{\boldsymbol{\theta}^{(0)}} = \text{optimize}\left(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{y})\right)$
$\varepsilon \rightarrow 1$
$k \rightarrow 1$
while $k < N$ and $\varepsilon > \gamma$ **do**
  $\mathbf{m} = \text{shapley\_values}\left(f_{\boldsymbol{\theta}^{(k)}}\right)$
  $\mathcal{L}_{k+1} = \{i \in \{1, \ldots, N\} | \mathcal{L}_{0_i} \neq \min_i m_i\}$
  $f_{\boldsymbol{\theta}^{(k+1)}} = \text{optimize}\left(f_{\boldsymbol{\theta}}\left(\mathbf{X}^{(\mathcal{L}_k+1)}, \mathbf{y}\right)\right)$
  $\varepsilon = \text{F}_1\text{−score}\left(f_{\boldsymbol{\theta}^{(k)}}\left(\mathbf{X}^{(\mathcal{L}_k)}\right)\right) - \text{F}_1\text{−score}\left(f_{\boldsymbol{\theta}^{(k+1)}}\left(\mathbf{X}^{(\mathcal{L}_k+1)}\right)\right)$
  $k \rightarrow k + 1$
end while
Output: $\mathcal{L}_k$

As seen in Algorithm 1, the input dataset set $\mathbf{X}$ with dimensions $M \times N$, where M is the number of samples and N is the number of features; the label vector $\mathbf{y}$ with M elements; the supervised machine learning model used for feature selection $f_{\boldsymbol{\theta}}(\cdot)$; and the minimum performance loss γ. Initially, the procedure creates an index vector $\mathcal{L}_0$ pointing to the selected features. Then, the machine learning model is optimized with the input dataset $\mathbf{X}$ and the label vector $\mathbf{y}$. Moreover, the initial performance loss is set to one. For each iteration, the algorithm computes the Shapley values of the machine learning model and updates the set of indices by removing the least important feature. The machine learning model is then optimized based on the updated set of selected features. Notice that the model optimization stage is conducted using a hyperparameter optimization (HPO) based on the random search strategy. Moreover, due to that the classification models are based on trees, this optimization phase evaluates the performance of the trained models for different numbers of estimators (`n_estimators`) and numbers of leaves (`num_leaves`). The procedure computes the performance loss based on the F_1-score. Finally, the algorithm returns the selected features after a loss criterion is achieved.

3.2.2. Model Optimization

We obtained the set of selected features for each country, period, and machine learning model. In this stage, we use the same supervised machine learning model as in the feature selection stage. For each country, period, and learning model, the extracted dataset was split into 100 partitions. For the training set, 80% of the samples were randomly selected from each partition. The remaining samples (20%) are included in the test set. The training and test sizes for the various countries and periods are shown in the third (2(a)) and fourth (2(b)) rows of Table 2. In addition, we estimate the minimum test sample size based on the TPR values for each country and period. We set the confidence interval at 95% and the marginal error at 2.50% [50, 51]. The ninth row (3(d)) of Table 2 displays the minimum test sample size for each country and period. As seen in this table, test sizes are larger than required for all countries and periods. For analyses with shorter durations (monthly, weekly, or daily), some countries and periods exhibit smaller test sample sizes than required. Therefore, for the selected countries, a yearly analysis is the minimum duration that satisfies the sample size requirements. Furthermore, as mentioned above, annual analyses (2020 and 2021) allow us to observe the impact of vaccination campaigns on the detection task and the identification of relevant variables. During the training phase, machine learning models applied an HPO based on the random search strategy. Then, we obtained the metrics results by evaluating the trained model on the test set. In general, performance results are calculated by averaging 100 realizations of the corresponding partitions, and the explainability analysis is performed on the model generating the best F_1-score.

Table 3
Performance metrics in percentage and the 95% confidence intervals (CIs) obtained by the proposed COVID-19 detection methods for Brazil and for 2020 and 2021.

Year	Method	F_1-scores	Specificity	Sensitivity	Precision
2020	RF	**84.24 (84.19 - 84.30)**	**82.88 (82.79 - 82.98)**	**83.43 (83.35 - 83.51)**	**85.08 (85.00 - 85.15)**
	XGB	80.56 (80.50 - 80.62)	78.96 (78.86 - 79.06)	79.59 (79.50 - 79.67)	81.57 (81.49 - 81.65)
	LGB	80.28 (80.22 - 80.34)	78.53 (78.43 - 78.63)	79.37 (79.29 - 79.44)	81.22 (81.14 - 81.30)
	RF SHAP	84.23 (84.18 - 84.28)	82.87 (82.78 - 82.97)	83.41 (83.33 - 83.49)	85.07 (84.99 - 85.15)
	XGB SHAP	80.57 (80.51 - 80.63)	78.97 (78.87 - 79.07)	79.59 (79.51 - 79.66)	81.58 (81.50 - 81.66)
	LGB SHAP	80.26 (80.21 - 80.32)	78.53 (78.42 - 78.64)	79.34 (79.26 - 79.41)	81.21 (81.13 - 81.30)
2021	RF	**80.43 (80.39 - 80.47)**	**91.40 (91.37 - 91.43)**	**75.74 (75.68 - 75.80)**	**85.74 (85.68 - 85.79)**
	XGB	73.89 (73.85 - 73.94)	88.75 (88.71 - 88.79)	68.25 (68.19 - 68.31)	80.55 (80.48 - 80.61)
	LGB	73.50 (73.45 - 73.54)	88.54 (88.50 - 88.58)	67.86 (67.80 - 67.91)	80.16 (80,09 - 80.23)
	RF SHAP	79.11 (79.07 - 79.15)	90.15 (90.11 - 90.18)	74.90 (74.84 - 74.96)	83.85 (83.79 - 83.91)
	XGB SHAP	72.53 (72.49 - 72.58)	88.26 (88.22 - 88.30)	66.69 (66.62 - 66.76)	79.50 (79.43 - 79.57)
	LGB SHAP	73.49 (73.44 - 73.53)	88.54 (88.50 - 88.58)	67.84 (67.78 - 67.90)	80.16 (80.10 - 80.22)

4. Results

4.1. Performance analysis

Table 3 illustrates different performance metrics in percentage and the 95% confidence intervals (CIs) generated by the various implementations of the proposed approach for Brazil, in 2020 and 2021. In particular, this table includes F_1-score, specificity, sensitivity, and precision obtained by the proposed techniques. A bold font and an underlined value indicate the best and second-best values for each metric and year. For Brazil 2020, the method based on the RF classification method generates the best performance results, i.e., F_1-score (RF: 84.24%, 95% CI: 84.19% to 84.30%), specificity (RF: 82.88%, 95% CI: 82.79% to 82.98%), sensitivity (RF: 83.43%, 95% CI: 83.35% to 83.51%), and precision (RF: 85.08%, 95% CI: 85.00% to 85.15%). In addition, the second best values are generated by the RF model optimized using the recursive feature elimination stage based on Shapley values: F_1-score (RF SHAP: 84.23%, 95% CI: 84.18% to 84.28%), specificity (RF SHAP: 82.87%, 95% CI: 82.78% to 82.97%), sensitivity (RF SHAP: 83.41%, 95% CI: 83.33% to 83.49%), and precision (RF SHAP: 85.07%, 95% CI: 84.99% to 85.15%). For 2021, equally, the RF and RF SHAP classifiers generate the best and second-best metric values, respectively. It is important to note that methods based on Shapley values, that apply a significant dimensionality reduction, exhibit a negligible performance loss in comparison with methods that do not include the recursive feature elimination step. Tables SM2, SM3, and SM4 in Supplemental material 2 show the performance metrics in percentage and the 95% CIs yielded by the proposed detection methods for Canada, Japan, and South Africa, respectively.

Figure 3 presents the receiver operating characteristic (ROC) curves and 95% CIs produced by the implemented machine learning models for the four countries and 2020. More precisely, each ROC curve is derived by averaging ten realizations of the respective experiment, where different training and test sets are randomly generated at each trial. The training set contains 80% of the samples, while the test set contains the 20% remaining ones. Every ROC curve includes the area under the ROC curve (AUC) and its 95% CI. For 2020, the detection methods obtaining the best auROCs for each country are Brazil (RF: 0.884, 95% CI: $0.845-0.923$), Canada (LGB_Shap: 0.913, 95% CI: $0.892-0.934$), Japan (LGB: 0.880, 95% CI: $0.835-0.925$), and South Africa (RF_Shap: 0.919, 95% CI: $0.871-0.967$). Note that the lowest AUC value is obtained for Brazil (XGB_Shap: 0.854, 95% CI: $0.839-0.869$).

On the other hand, Figure 4 illustrates the ROC curves and their 95% CIs obtained by the proposed COVID-19 detection models for the four countries and 2021. The detection methods yielding the best AUC values for every country are Brazil (RF: 0.879, 95% CI: $0.805-0.953$), Canada (LGB: 0.903, 95% CI: $0.889-0.917$), Japan (RF: 0.918, 95% CI: $0.881-0.955$), and South Africa (RF: 0.918, 95% CI: $0.870-0.966$). For 2021, the lowest AUC value is also obtained for Brazil (XGB: 0.817, 95% CI: $0.716-0.918$).

Figure 5 displays the F_1-scores and the corresponding 95% CIs obtained by different COVID-19 detection methods for the four countries and for 2020 and 2021. Notice that F_1-scores are presented in descending order to identify the best performance. Particularly, we display the F_1-scores produced by the detection methods based on RF, XGB, LGB, RF_Shap, XGB_Shap, and LGB_Shap. For comparison purposes, we also included the F_1-scores obtained by previously reported detection techniques such as Menni [6], Smith [5], Shoer [16], Mika [17], and Astley [1]. In Supplemental Material 2, Table SM1 shows the numerical values of the F_1-scores and the 95% CIs obtained by the

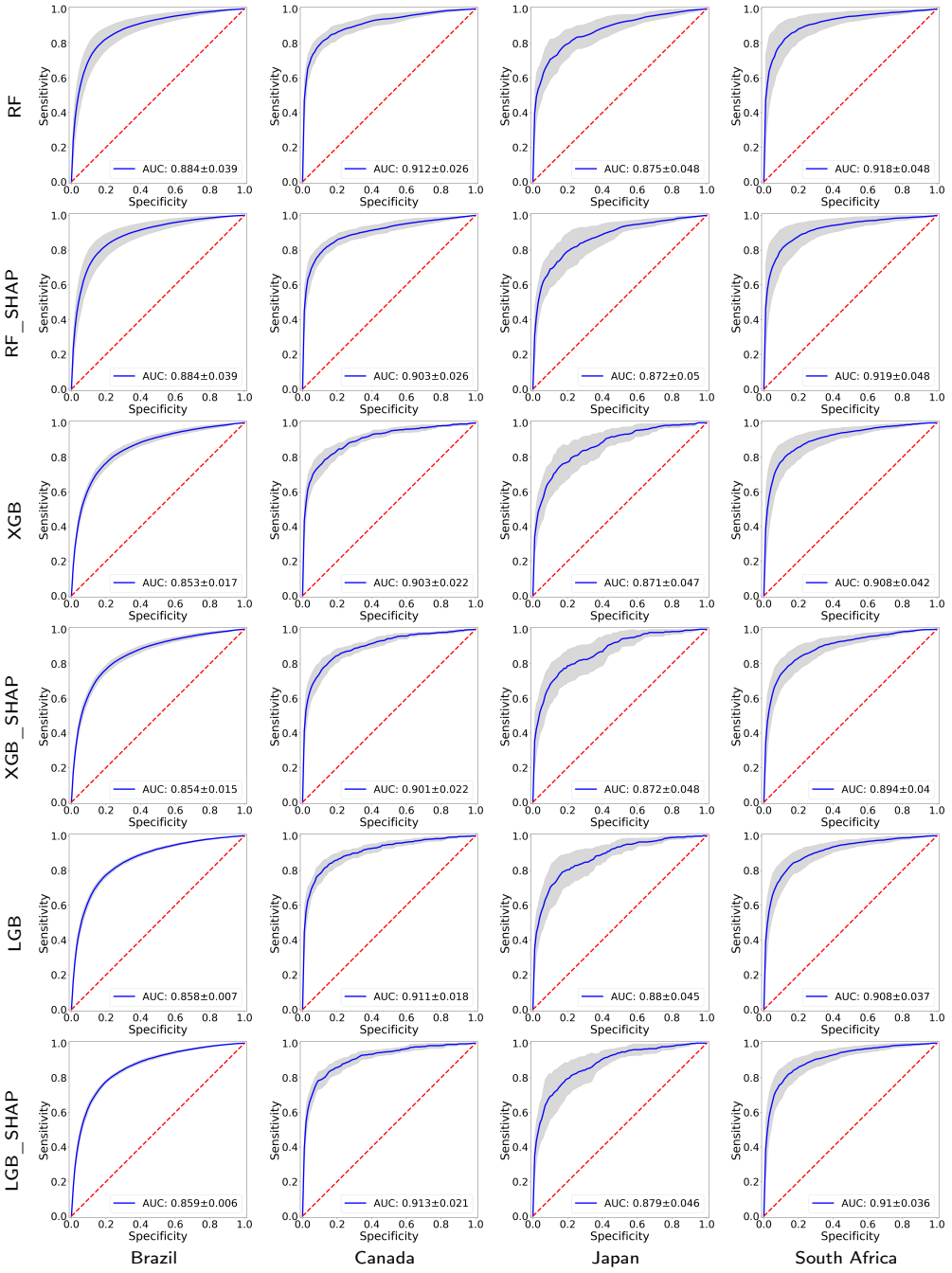

Figure 3: ROC curves and their 95% confidence intervals for the four countries and for 2020 using the proposed approach with different classifiers. The AUC value is included in each ROC curve.

different detection methods for the four countries and for 2020 and 2021. Specifically, for 2020, the detection methods that yield the best F_1-scores for every country are: Brazil (RF: 84.24%, 95% CI: 84.19% to 84.29%), Canada (XGB: 62.53%, 95% CI: 61.98% to 63.09%), Japan (XGB_Shap: 59.70%, 95% CI: 58.82% to 60.57%), and South Africa (RF_Shap: 81.88%, 95% CI: 81.68% to 82.09%). For 2021, the methods yielding the best F_1-scores for every country are: Brazil (RF: 80.43%, 95% CI: 80.39% to 80.47%), Canada (RF: 63.80%, 95% CI: 63.52% to 64.08%), Japan (RF: 70.11%, 95% CI: 69.84% to 70.37%), and South Africa (RF: 77.69%, 95% CI: 77.53% to 77.85%). It is worth noting that the proposed COVID-19 detection methods outperform previously reported techniques for the four countries and

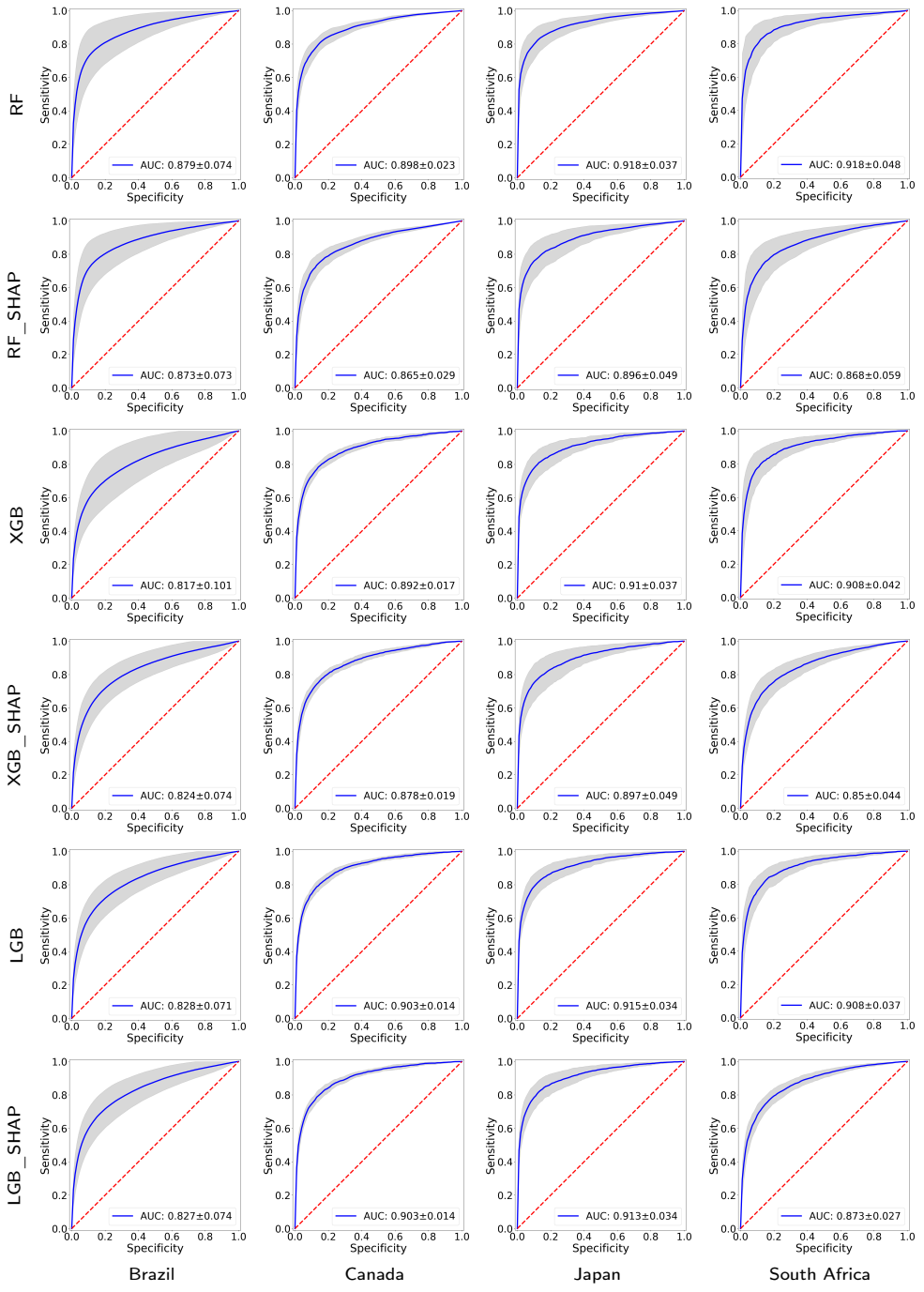

Figure 4: ROC curves and their 95% confidence intervals for the four countries and for 2021 using the proposed approach with different classifiers. The AUC value is included in each ROC curve.

for the two periods under test. F_1 scores for 2020 and 2021 are presented in Figure SM1 in the Supplemental Material to compare the performance of each detection method across the countries under test. As shown in this figure, each method generates the best F_1 scores for Brazil or South Africa. Table 2 highlights that these countries exhibit TPR values that are at least three times larger than those of Canada and Japan.

The ANOVA test was conducted to assess whether there are statistically significant differences in performance among all detection methods based on the work [52]. Specifically, the ANOVA test was applied to the results of 100

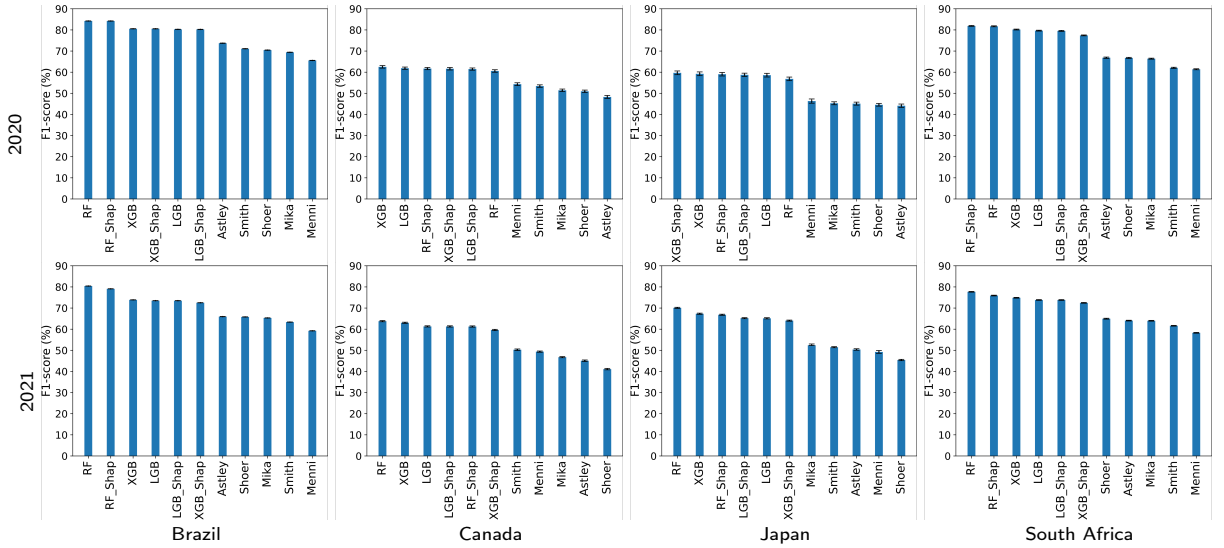

Figure 5: F_1-scores and the 95% CIs yielded by various COVID-19 detection methods for the four countries and for 2020 and 2021.

Table 4
Results of the ANOVA test for five classification metrics for the four countries for 2020 and 2021.

Year	Classification Metric	Brazil F	Brazil p	Canada F	Canada p	Japan F	Japan p	South Africa F	South Africa p
2020	F_1-score	4874.00	$p < 0.001$	5.01	$p < 0.001$	5.31	$p < 0.001$	234.10	$p < 0.001$
	Sensitivity	2789.36	$p < 0.001$	38.80	$p < 0.001$	28.89	$p < 0.001$	116.71	$p < 0.001$
	Specificity	1817.85	$p < 0.001$	114.33	$p < 0.001$	76.30	$p < 0.001$	79.14	$p < 0.001$
	Precision	2202.13	$p < 0.001$	75.64	$p < 0.001$	38.29	$p < 0.001$	110.62	$p < 0.001$
	AUC	4894.43	$p < 0.001$	27.88	$p < 0.001$	19.61	$p < 0.001$	222.14	$p < 0.001$
2021	F_1-score	24479.56	$p < 0.001$	103.08	$p < 0.001$	245.99	$p < 0.001$	429.85	$p < 0.001$
	Sensitivity	17558.30	$p < 0.001$	73.83	$p < 0.001$	128.96	$p < 0.001$	255.35	$p < 0.001$
	Specificity	3849.87	$p < 0.001$	221.76	$p < 0.001$	166.94	$p < 0.001$	184.18	$p < 0.001$
	Precision	6395.48	$p < 0.001$	204.08	$p < 0.001$	210.32	$p < 0.001$	230.96	$p < 0.001$
	AUC	26173.41	$p < 0.001$	76.38	$p < 0.001$	156.77	$p < 0.001$	460.19	$p < 0.001$

realizations of the test set. Statistically significant differences were observed between the performance of the proposed COVID-19 detectors for 2020 and 2021 for all countries. As an illustrative example, Table 4 displays the results computed by the ANOVA test for five classification metrics (F_1-score, sensitivity, specificity, precision, and AUC) for the four countries and for 2020 and 2021. Furthermore, we used the Friedman test to compare the performance yielded by the proposed classifiers [53]. The Friedman test was also applied to the results of the 100 realizations of the test set. It is worth noting that this test shows statistically significant performance differences between the proposed machine learning models, with $p < 0.001$ for all countries for all metrics, for 2020 and 2021. Additionally, a pairwise analysis was conducted using the Wilcoxon signed-rank test and different significance levels: $\alpha = 0.05$, $\alpha = 0.01$, and $\alpha = 0.001$ [52]. The pairwise comparison results using different significance levels for the AUC for the four countries in 2020 are displayed in Table 5. The Wilcoxon signed-rank test shows statistically significant differences in the AUC performance between RF and LGB, RF and LGB_SHAP, RF and XGB, and RF and XGB_SHAP. On the other hand, this test does not find statistically significant differences between RF and RF_SHAP, LGB and LGB_SHAP, and XGB and XGB_SHAP. Thus, notice that RF_SHAP, LGB_SHAP, and XGB_SHAP models, which are built using a reduced set of variables generated by the feature selection stage, exhibit similar performances to those yielded by RF, LGB, and XGB classifiers.

Finally, to evaluate the proposed COVID-19 detection methods in a practical problem, we estimate the normalized COVID-19 daily case curves for the four countries from January 1, 2021, to June 25, 2022. Using the proposed detection methods based on RF, LGB, and XGB, we first identify the daily COVID-19 active cases in the interval of interest. Afterward, the estimated daily case curves are normalized with respect to the maximum value. A comparison of the daily case curves obtained by the proposed detection methods for the four countries is shown in Figure 6. In addition, we include the normalized COVID-19 daily case curve provided by the respective national healthcare

Table 5
Results of the Wilcoxon signed-rank test using different significance levels for the AUC metric for the four countries in 2020.

Method Pairs	Brazil				Canada				Japan				South Africa			
	p	$p < 0.05$	$p < 0.01$	$p < 0.001$	p	$p < 0.05$	$p < 0.01$	$p < 0.001$	p	$p < 0.05$	$p < 0.01$	$p < 0.001$	p	$p < 0.05$	$p < 0.01$	$p < 0.001$
RF vs RF_SHAP	0.2342	✗	✗	✗	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0119	✓	✗	✗
RF vs LGB	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0000	✓	✓	✓
RF vs LGB_SHAP	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0000	✓	✓	✓
RF vs XGB	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0000	✓	✓	✓
RF vs XGB_SHAP	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0000	✓	✓	✓
RF_SHAP vs LGB	0.0000	✓	✓	✓	0.0008	✓	✓	✓	0.0168	✓	✗	✗	0.0000	✓	✓	✓
RF_SHAP vs LGB_SHAP	0.0000	✓	✓	✓	0.0069	✓	✓	✗	0.1061	✗	✗	✗	0.0000	✓	✓	✓
RF_SHAP vs XGB	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.2070	✗	✗	✗	0.0000	✓	✓	✓
RF_SHAP vs XGB_SHAP	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0082	✓	✓	✗	0.0000	✓	✓	✓
LGB vs LGB_SHAP	0.3154	✗	✗	✗	0.8615	✗	✗	✗	0.3358	✗	✗	✗	0.0837	✗	✗	✗
LGB vs XGB	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0000	✓	✓	✓
LGB vs XGB_SHAP	0.0000	✓	✓	✓	0.0006	✓	✓	✓	0.0000	✓	✓	✓	0.0000	✓	✓	✓
LGB_SHAP vs XGB	0.0000	✓	✓	✓	0.0000	✓	✓	✓	0.0005	✓	✓	✓	0.0000	✓	✓	✓
LGB_SHAP vs XGB_SHAP	0.0000	✓	✓	✓	0.0007	✓	✓	✓	0.0000	✓	✓	✓	0.0000	✓	✓	✓
XGB vs XGB_SHAP	0.6180	✗	✗	✗	0.0388	✓	✗	✗	0.0327	✓	✗	✗	0.0000	✓	✓	✓

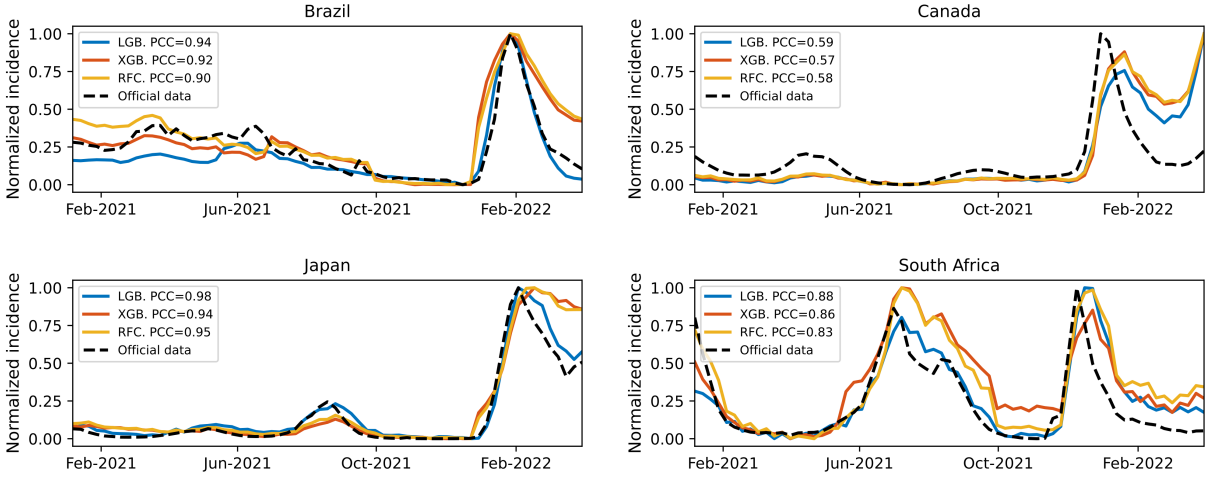

Figure 6: Normalized daily incidence curves generated by the proposed detection methods based on RF, LGB, and XGB. The normalized daily incidence curve determined from official reports is also displayed for each country.

system for comparison. In each country, we present a Pearson correlation coefficient between the curve obtained by each proposed detection methodology and the curve provided by the national healthcare system. The best correlation coefficients are generated by the proposed approach based on the LGB classifier, i.e., Brazil (LGB: 0.94), Canada (LGB: 0.59), Japan (LGB: 0.98), and South Africa (LGB: 0.88). In general, the estimated curves follow the trends reported in official statistics for Brazil, Canada, Japan, and South Africa. For the winter of 2022, all the curves present some difficulty in following the trends of the official ones. Something important to highlight to explain this is that the UMD-CTIS collected information on the same variables across the entire period between April 2020 and June 2022, with a few minor modifications. However, these variables did not capture information on different indicators related to the dynamic behavior of the pandemic, such as the loss of effectiveness of vaccines over time or the emergence of new variants such as Omicron [54], which affected, in turn, the accuracy of the corresponding estimates. As an important note, the daily case curves generated by these detection methods have been used by the CoronaSurveys Project (`https://coronasurveys.org`), a collaboration between several academic institutions aimed at providing global pandemic surveillance based on surveys, to estimate daily active cases for more than 150 countries/territories [55].

4.2. Explainability analysis

This work considers a local and post-hoc approach based on Shapley values and a global ante-hoc approach based on the RF because of its superior accuracy with respect to the other tree-based methods considered in this work.

4.2.1. Explainability analysis using Shapley values

We now proceed to analyze the explainability based on the results given by the different techniques, using (or not) the Shapley-based feature removal, to estimate positive cases, for the four countries and for 2020 and 2021 (see Figures 7, 8). We will delve deeper into the model that gave the best results, both in its full version and with the Shapley-based feature reduction. In the analysis, we have considered that the variables with Shapley values greater than 0.05 are relevant to consider, as stipulated in the literature [20, 46].

- Year 2020: According to Figure 7, in the case of RF the variable B1.10.1 is relevant in all cases (countries), while in other cases, the variable B1.10.2 is also relevant as C0.2.1 (it is the most relevant for Japan but with a low relevance value of less than 0.03). This makes sense because these variables refer to the loss of smell/taste (B1.10.1 and B1.10.2), and the fact of having gone to a market, store or pharmacy (C0.2.1). By 2020, the most widespread symptom was loss of smell and taste. B1.10.1, even in the other methods, had an even higher relevance value, reaching values greater than 0.1 in LGB and LGB.SHAP for Brazil and South Africa, and 0.08 in XGB and XGB.SHAP. Also, variable B1bx10.1, which means that a usual symptom was the loss of taste or smell, appears relevant in the case of South Africa for RF (close to 0.03). This is very much in agreement with what happened for the variant that prevailed in 2020.

  In general, for 2020, the most important variable in every model is the loss of smell or taste. Similarly to what was determined by [56, 57], this was the most representative symptom of COVID-19 with the first variant, the only one present in 2020. Along with the variables described in the previous paragraph, other variables somewhat relevant (in some cases, with Shapley values greater than 0.04) were B5.1 (spent time with COVID-19 infected people), C6.1 (Not spend time with someone outside your household) or C5.6 (Not going out for a week) to determine if an individual is COVID-19 positive.

  To continue with the explainability analysis, and as an example, we did our own ranking to determine the relevance of the features in general, according to the order using the Shapley value in which a feature appears in each technique for each country. For this ranking, the first 10 variables were considered according to the Shapley value for each technique in each country, and a 10 was assigned to the one with the highest Shapley value, 9 to the next, and so on. Then, at the end, the values obtained by each feature in each country-technique pair are added to obtain its position in our ranking.

  Table 3 lists only the first 10 variables according to that ranking. Each column contains the ranking of the ten most important features for each constructed model based on Shapley values. As the value of the cell increases, the importance of the feature increases, with one being the least important feature and ten being the most important. A zero value indicates that the variable is not on the list of the ten most important features. In the last column, you will find the sum of importance across the models. We can see that the variable B1.10.1 has a value much higher than the rest (218), and with C0.2.1 (192) they are very far from the rest. They are the same variables that we had determined before as the most relevant, which corroborates our previous explainability analysis. B1.10.1 has the highest value in almost all cases, and only in Canada, for some techniques, it has a low value (for example, XGB with 3). This happens less in C0.2.1, since its lowest value is 5 (also for XGB and Canada). The rest of the techniques have at least once a value of 0, which means that they are not among the first 10 variables according to the Shapley value in that technique and country.

  We also see some cases where the techniques in some countries only use a few of the best ranked variables according to our ranking. For example, the case of Brazil with XGB using Shapley, or Japan with XGB, which use only 5 of the first 10 variables of our ranking. That implies that they have other more relevant variables, in addition to those 5, to the first 10 established by our ranking (for example, V1.1 has the second largest Shapley value in Japan with XGB). We also see that there is no technique for any country that uses those 10 best ranked variables according to our criteria, only RF for Canada used 9 of them.

- Year 2021: This year was characterized by the appearance of several variants of COVID-19 (Delta, Omicron, etc.) and by the massive vaccination of the population against COVID-19. In this case, the variable B1.10.1 continued to appear as one of the most relevant variables, and in some cases B1.10.2. However, other variables that also appeared with great relevance were V1.1 and B7.1. These variables refer to whether the person has been vaccinated against COVID-19 (V1.1), and if in the last 4 weeks the person did any paid work (B7.1). Particularly, V1.1 is relevant in Brazil (despite being a country where some government instances promoted the denial of the

Table 6
Ranking of the features based on the Shapley values for the entire set of countries for 2020.

Feature	BR RF 2020	BR RF SHAP 2020	BR XGB 2020	BR XGB SHAP 2020	BR LGB 2020	BR LGB SHAP 2020	CA RF 2020	CA RF SHAP 2020	CA XGB 2020	CA XGB SHAP 2020	CA LGB 2020	CA LGB SHAP 2020	JP RF 2020	JP RF SHAP 2020	JP XGB 2020	JP XGB SHAP 2020	JP LGB 2020	JP LGB SHAP 2020	ZA RF 2020	ZA RF SHAP 2020	ZA XGB 2020	ZA XGB SHAP 2020	ZA LGB 2020	ZA LGB SHAP 2020	SUM
B1.10.1	10	10	10	10	10	10	10	9	9	3	4	4	10	9	10	10	10	10	10	10	10	10	10	10	218
C0_2.1	7	7	8	9	9	8	7	6	7	5	8	8	9	10	9	9	9	9	7	8	7	9	8	9	192
B1b.x10.1	8	8	0	0	0	0	8	8	6	7	7	7	5	4	6	6	6	5	8	7	6	7	9	5	139
B1.10.2	9	9	7	7	6	6	9	10	10	10	10	10	6	7	0	0	0	0	9	9	0	0	0	0	134
B5.1	6	6	9	8	8	9	4	7	8	9	9	9	0	0	3	3	0	1	0	4	2	0	3	8	116
C0_2.2	5	4	0	0	0	0	2	4	0	0	0	0	8	7	0	0	0	0	6	8	7	9	6	9	75
C5.6	0	0	1	0	0	0	6	5	5	7	6	6	0	0	0	0	0	0	5	5	8	6	5	2	67
B3.2	0	0	0	0	0	0	5	0	0	0	0	0	7	8	7	8	8	8	4	1	3	4	0	4	67
B6.1	4	5	6	6	7	7	1	3	4	6	5	5	0	0	0	0	0	0	0	0	0	0	0	0	59
C0_1.1	3	3	3	0	3	3	0	0	0	1	3	3	0	0	0	0	0	0	3	0	0	5	0	3	33

positive effect of vaccination), and B7.1 is the most relevant on several occasions (for example, for RF and XGB in Canada, South Africa and Japan, although with low relevance values, of the order of 0.015). Other variables that appear with some relevance in some cases are B3.1 that deals with whether someone in the local community is known to have been ill with fever, cough, or difficulty breathing (for example, the most relevant for RF.SHAP in Canada and Japan, among others, but with a very low relevance (less than 0.004)), B15.2 which is if the person has had an appointment to receive a COVID-19 vaccine (the most relevant for XGB.SHAP in South Africa), and V1.2 which deals with whether to have a COVID-19 vaccine (the most relevant for XGB.SHAP in South Africa with 0.05).

Variable B3.1 is the most relevant for all SHAP-based methods for Canada and Japan, which indicates whether anyone in the local community was known to have fever and cough or shortness of breath as one of the causes of seropositivity (rapid spread of the virus). Similarly, variable B7.1 is the most relevant for all methods without SHAP for Japan, Canada and South Africa (contagion from going to work, also linked to the rapid spread of the virus). Finally, V1.2 is the most relevant for all methods with SHAP in South Africa, indicating the fact of not having been vaccinated as one of the reasons for high seropositivity in that country. In the case of Brazil, the most relevant variables were again B1.10.1, B1.10.2 and V1.1. Also to note that C0.2.1 ceased to be relevant in 2021.

For 2021, again, one of the most important features is the loss of smell or taste. However, with less relevance in some countries as it was in 2020 due to the COVID-19 variants during this year along with the vaccination. In addition, variables linked to facilitating the spread (such as going to work (B7.1) or if the person had acquaintances with symptoms (B3.1)) appear as reasons for seropositivity, or the case of not having been vaccinated yet (South Africa and V1.2). Thus, 2021 has some of the same important features as 2020 but the variables related to the vaccines and the rapid spread of the virus also play a key role.

4.2.2. Explainability analysis using the range of features given by RF

In this part, we carry out an explainability analysis for RF becasue it is the technique that showed the best quality (see Figures 9 and 10). With RF, various measures of feature importance can be used for an explainability analysis. In this work, we have used MDA for being one of the best feature-importance measures according to the literature [41, 45].

- 2020: It is interesting to see that again, the most relevant variables are B1.10.1, B1.10.2, as well as B1bx10.1. The relevance order changes in some cases, as for RF (B1.10.1 is the variable that appears as relevant most frequently for RF using the Shapley values, and in this case is shared with B1.10.2). In general, the relevance values of the most relevant variables are always high, and occasionally very high (for example, B1.10.1 for Canada close to 0.8).

  Another aspect to note is how in some cases the most frequent variables change, as in the case of Japan with C0.2.1 (it was among the most relevant variables in the Shapley values and was no longer among the three most

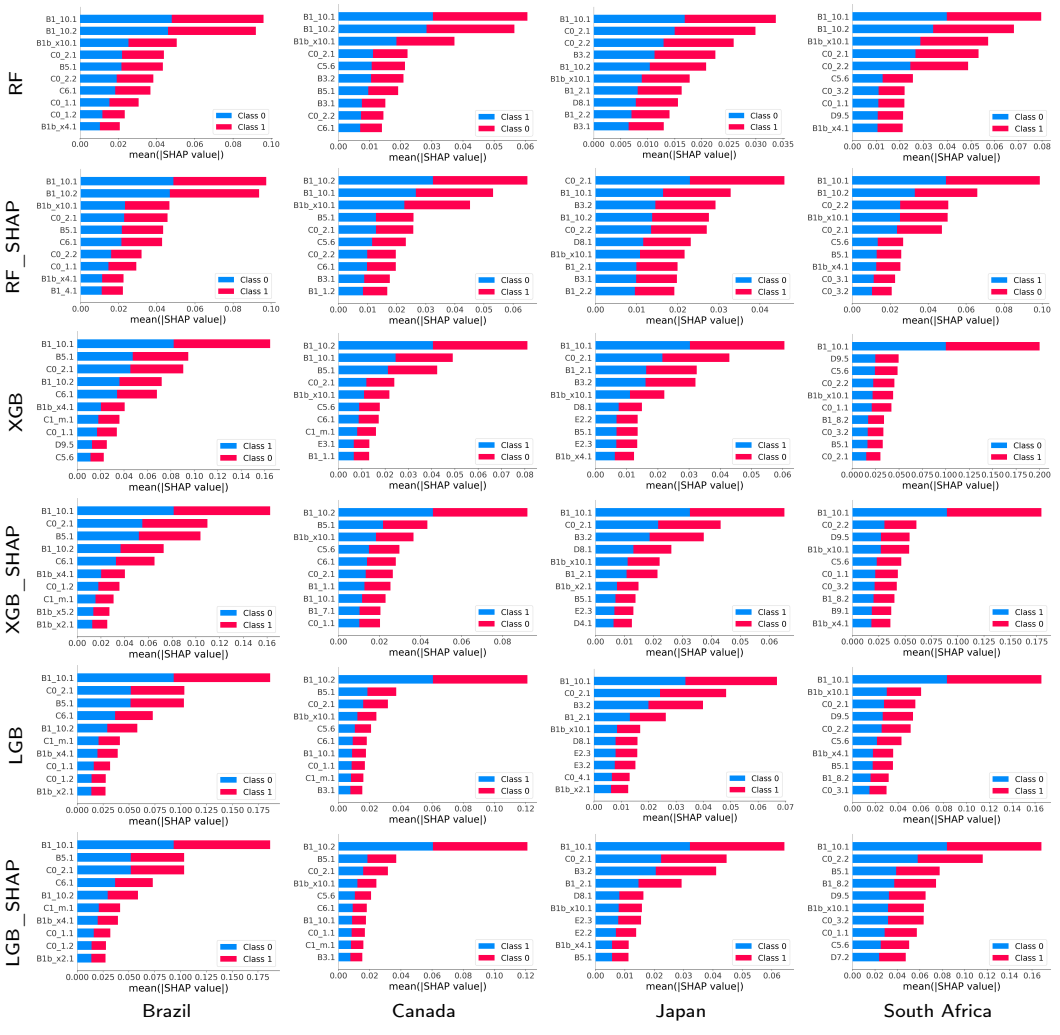

Figure 7: SHAP values and their impact on the detection output of the 10 most relevant features obtained by the proposed approach using different classification models for the four countries and for 2020.

relevant in the RF ranking). Finally, B1.10.1 and B1.10.2 are always the most relevant, with values greater than 0.05 regardless of whether RF or RF.SHAP is used. This clearly indicates that the variables linked to loss of smell/taste are the fundamental ones for estimating seropositivity to COVID-19 in the case of RF and RF.SHAP.

- 2021: In the case of Brazil, the most relevant variables were once again B1.10.1, B1.10.2, with B1.10.2 now being more relevant. On the other hand, the variable V1.1 disappears from the relevant group (which makes sense, because Brazil was a country where the denial of the positive effect of vaccination was promoted from some of the government instances). Also, variable B7.1 continues to be the most relevant for all methods without SHAP for Japan, Canada and South Africa, being very decisive in Japan and Canada (values greater than 0.08, and the next with relevance values around 0.04). Another variable that is no longer relevant is B3.1, with two variables appearing as highly relevant, B1bx10.1 and C5.6, particularly for SHAP-based techniques. B1bx10.1 is relevant in Japan, which means that a common symptom was the loss of taste or smell; and C5.6 in South Africa and Canada, and it has to do with not having been in public during the last 7 days (self-care of people).

Note that the most important features selected for RF with the full set of input variables differ from those selected for RF.SHAP for each country and period. The difference is due to the fact that the feature selection method extracts the

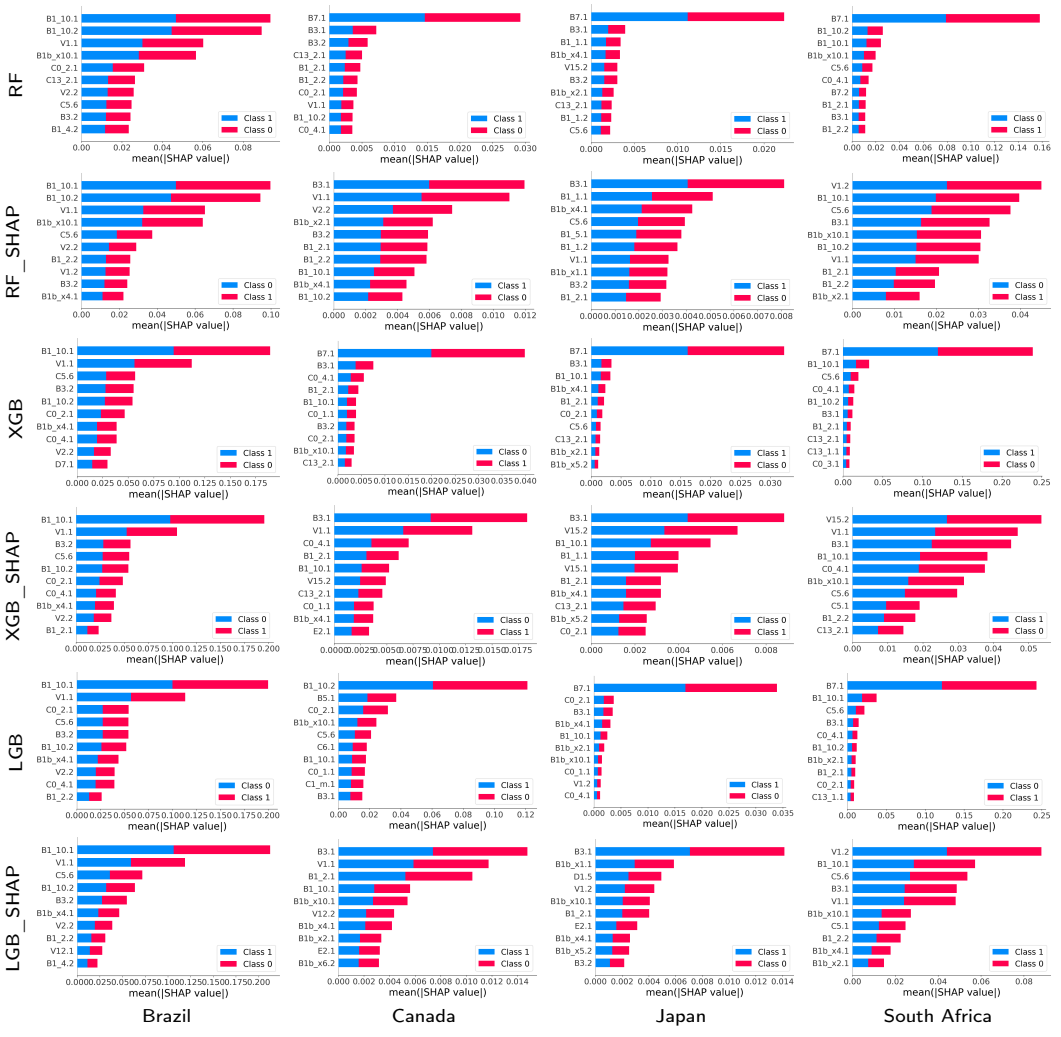

Figure 8: SHAP values and their impact on the detection output of the 10 most relevant features obtained by the proposed approach using different classification models for the four countries and for 2021.

Shapley values while the feature rankings are computed using the MDA method. As can be seen in the explainability analysis, this approach provides information on additional features that contribute to active case detection.

5. Discussion

5.1. Performance quality of the machine-learning approaches

This study presents a machine-learning approach to detecting COVID-19-active cases based on three classification models: random forest (RF), light gradient boosting (LGB), and extreme gradient boosting (XGB). More precisely, the proposed detection approach predicts active cases using the entire set of variables collected from the UMD-CTIS questionnaires. These questionnaires record a wide range of individual features such as gender, age group, vaccination acceptance, and isolation measures. In addition, we introduce a feature-reduction approach that uses the RFE strategy to train the classification model. A key objective of the RFE algorithm is to identify and keep relevant features based on Shapley values without compromising detection accuracy. It is pertinent to mention that the proposed method is evaluated on UMD-CTIS data extracted from four countries: Brazil, Canada, Japan, and South Africa, and two periods: 2020 and 2021. Specifically, we consider experiments with at least one symptom reported within the past 24 hours and a test result within the past 14 days. Extracted datasets may contain biases, limitations, and missing values. In countries where demographic information can be a significant factor in detecting active cases, such as isolated communities and

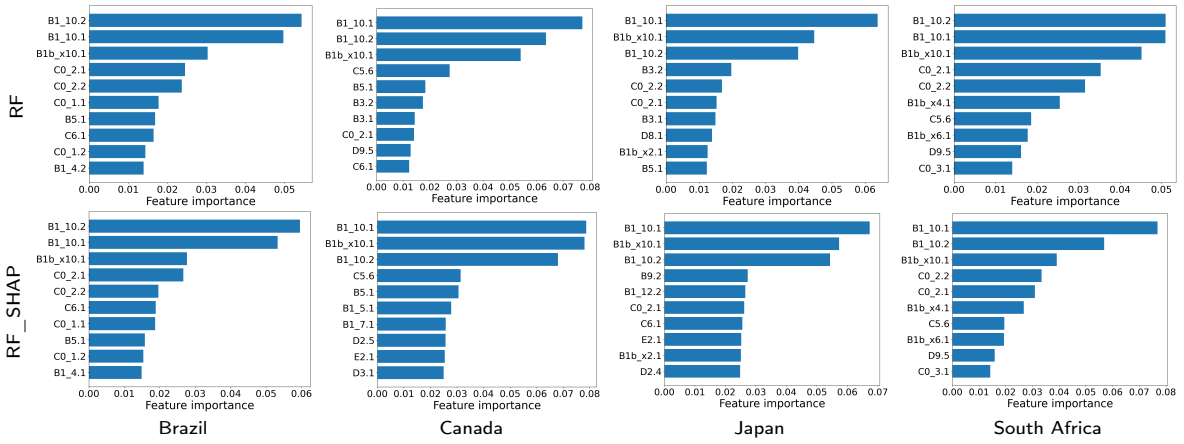

Figure 9: Feature importance of the 10 most relevant input variables obtained by classification models based on the random forest method for the four countries and for 2020.

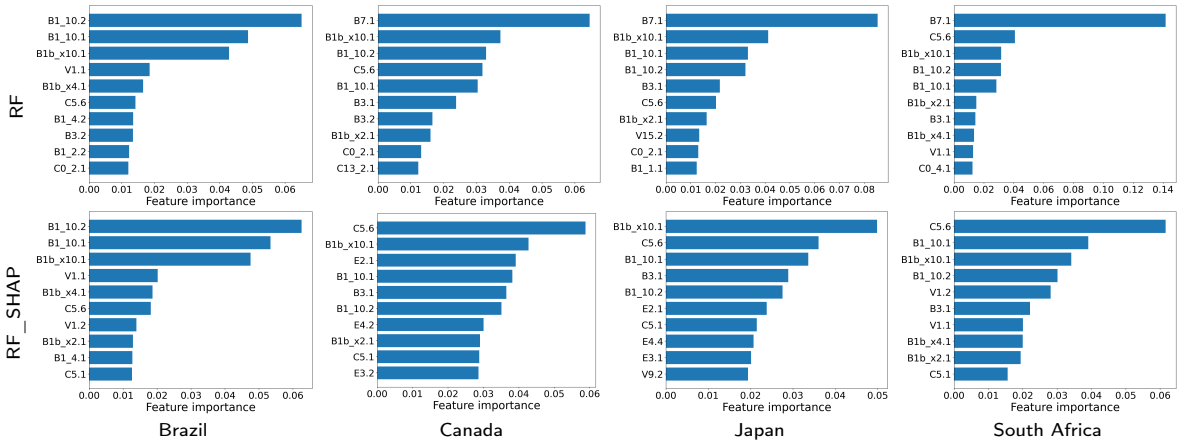

Figure 10: Feature importance of the 10 most relevant input variables obtained by classification models based on the random forest method for the four countries and for 2021.

islands, the data may be affected by biases due to the homogenization of the population. In addition, biases can arise from assuming that all populations have Internet access uniformly. To reduce biases, we randomly select a limited set of experiments as the training set to optimize the classification models.

The proposed approach has shown competitive performance for the four countries for 2020 and 2021. In particular, the feature selection stage removes a large number of irrelevant variables with a negligible impact on classification accuracy. According to different quality metrics (such as F_1-score, sensitivity, specificity, precision, and AUC), RF and RF SHAP models exhibit the most accurate detection performances across the four countries for 2020 and 2021. We also compare the performance of the developed technique to those yielded by previously reported approaches based on surveys. The proposed detection methodology outperforms the state-of-the-art methods for the four countries in terms of F_1-score. The RF-based approach obtained the highest results, regardless of whether or not feature selection was used. In the final step, we use the developed detection approach to construct the normalized daily-case curve for the four countries between January 1, 2021, and June 25, 2022, to observe pandemic trends. In comparison with official records, these estimated curves provide consistent tracking of pandemic evolution. Therefore, taking into account both the proposed detection approaches and the massive amount of data provided by the UMD-CTIS questionnaires, we can reliably track pandemic indicator trends in a similar way to that provided by public healthcare systems.

5.2. Explainability Analysis

Regarding the explainability analysis, the variables that appear most frequently are the loss of smell or taste (B1.10.1/B1.10.2), regardless of the year, country, the technique used for the explainability analysis, the prediction technique, or dataset reduced or not by the Shapley method. Other variables appear in some specific cases (countries, forecasting techniques, etc.). For example, using the Shapley method for explainability analysis in 2020, the variable C0.2.1 (the fact of having gone to a market, store or pharmacy) is relevant in some countries. Also, in 2021, the variable V1.1 is relevant in Brazil (whether the person has been vaccinated against COVID-19), the variable B7.1 is relevant in some countries (if in the last 4 weeks the person did any paid work) for the cases without data reduction, and B3.1 (it indicates whether anyone in the local community was known to have fever and cough) for cases with data reduction using the Shapley method. The same happens using the RF ranking for the explainability analysis: new variables appear (such as C5.6) or some existing ones disappear (such as C0.2.1 and B3.1).

In any case, attribute-based explainability analysis shows the relevant variables for decision makers to detect seropositivity very quickly. This is valid both for the case of using the ranking given by RF or the Shapley values. However, it is important to highlight that although they present common variables, between the two techniques there are some differences between those that are considered to be more relevant. For example, V1.1 appears as relevant in the Shapley method for Brazil and disappears in the RF ranking, which makes more sense because in Brazil, the vaccination campaign did not have a strong support from the government.

Thus, RF seems sufficient to achieve good results and explain the results obtained (explainability analysis). But although explainability is aimed at the understanding by experts and non-experts, there are no designs or formal evaluations on the human usability of the methods analyzed in this work. This is pending work, which goes beyond simple representations of explanations. At the same time, the analysis of the variables by classes has rarely been carried out (in the case of Shapley, the values by class are similar/symmetrical), which opens up a space for research for the development of techniques that allow analyzing the possibility of explainability by classes (which characteristics/variables are relevant for each class).

Now, in this last part, we define a variable pattern by combining the results of the two explainability analysis techniques (Shapley value and feature importance) for each year and country/year. According to the explainability analysis results, we build symptomatic patterns using the relevance of the features in both techniques. Thus, the relevant features for each year are:

- **2020:** Loss of smell and taste in the last 24h (B1.10.1)
- **2021:** Loss of smell and taste in the last 24h (B1.10.1); and COVID-19 vaccination in the last 24h (V1.1).

For all countries in 2020, the most frequent variable is loss of smell and taste in the last 24 hours (variable B1.10.1). Additionally, vaccination campaigns in 2021 become a relevant factor in the positive prediction of COVID-19. For country/year, the relevant features are:

- **Brazil 2020:** Loss of smell and taste in the last 24h (B1.10.1); Loss of smell and taste after 24 hours (B1.x10.1); Have you gone to a market, grocery store, or pharmacy in the last 24 hours (C02.1), Have you spent time with any of these people in the last 7 days? (B5.1).
- **Brazil 2021:** Loss of smell and taste in the last 24h (B1.10.1); Loss of smell and taste after 24 hours (B1.x10.1), Have Fatigue? (B1b.x4.1).
- **Canada 2020:** Loss of smell and taste in the last 24h (B1.10.1); Loss of smell and taste after 24 hours (B1.x10.1); Have you spent time with any of these people in the last 7 days? (B5.1); In the last 7 days, have you not been in public? (C5.6).
- **Canada 2021:** Loss of smell and taste in the last 24h (B1.10.1); Loss of smell and taste after 24 hours (B1.x10.1); Do you personally know anyone in your local community who is sick with a fever and either a cough or difficulty breathing? (B3.1).
- **Japan 2020:** Loss of smell and taste after 24 hours (B1.x10.1); Loss of smell and taste in the last 24h (B1.10.1); have you gone to a market, grocery store, or pharmacy in the last 24 hours (C02.1).

- **Japan 2021:** Do you have a cough? (B1b.x2.1); Do you personally know anyone in your local community who is sick with a fever and either a cough or difficulty breathing? (B3.1).
- **South Africa 2020:** Loss of smell and taste after 24 hours (B1.x10.1); Loss of smell taste in the last 24h (B1.10.1); Have you gone to a market, grocery store, or pharmacy in the last 24 hours (C02.1); In the last 7 days, have you not been in public? (C5.6).
- **South Africa 2021:** Loss of smell and taste after 24 hours (B1.x10.1); Loss of smell and taste in the last 24h (B1.10.1); Do you personally know anyone in your local community who is sick with a fever and either a cough or difficulty breathing? (B3.1); In the last 7 days, have you not been in public? (C5.6); Have you had a COVID-19 vaccination? (V1.1).

As can be seen above, the most relevant variables are loss of smell and taste after 24 hours (B1.x10.1) and loss of smell and taste in the last 24 hours (B1.10.1), which appear for all countries. The rest of the variables are very specific to each country. During the pandemic, each country experienced different health, geographical, or economic conditions (for example, the bad South Africa's vaccination campaign in 2021 turned off the variable V1.1; or the fact that the Brazilian authorities denied the pandemic during 2020 affected the variables "Have you been to the market, grocery store, or pharmacy in the last 24 hours?" (C02.1) or" Have you spent time with any of these people within the last week?" (B5.1)). In summary, as we have mentioned before, the only variable that appears in all explainability analyses, regardless of explainability technique, or machine learning method used, is loss of smell and taste after 24 hours.

6. Declarations

6.1. Ethical Declaration

The Ethics Board (IRB) of IMDEA Networks Institute gave ethical approval for this work on 2021/07/05. IMDEA Networks has signed Data Use Agreements with Facebook, Carnegie Mellon University (CMU) and the University of Maryland (UMD) to access their data, specifically UMD project 1587016-3 entitled C-SPEC: Symptom Survey: COVID-19 and CMU project STUDY2020_00000162 entitled ILI Community-Surveillance Study. The data used in this study was collected by the University of Maryland through The University of Maryland Social Data Science Center Global COVID-19 Trends and Impact Survey in partnership with Facebook. Informed consent has been obtained from all participants in this survey by this institution. All the methods in this study have been carried out in accordance with relevant of ethics and privacy guidelines and regulations.

6.2. Availability of Data and Materials

The data presented in this paper (in aggregated form) and the programs used to process it will be openly accessible at https://github.com/GCGImdea/coronasurveys/. The microdata of the CTIS survey from which the aggregated data was obtained cannot be shared, as per the Data Use Agreements signed with Facebook, Carnegie Mellon University (CMU) and the University of Maryland (UMD). Thus, the authors do not have permission to share the microdata of the CTIS survey.

6.3. Funding Declaration

This work was partially supported by grant CoronaSurveys-CM, funded by IMDEA Networks and Comunidad de Madrid, Spain, grants COMODIN-CM and PredCov-CM, funded by Comunidad de Madrid and the European Union through the European Regional Development Fund (ERDF), grants TED2021-131264B-I00 (SocialProbing) and PID2019-104901RB-I00, funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR, and individual donations to the CoronaSurveys Project https://coronasurveys.org.

References

[1] C. M. Astley, G. Tuli, K. A. Mc Cord, E. L. Cohn, B. Rader, T. J. Varrelman, S. L. Chiu, X. Deng, K. Stewart, T. H. Farag, et al., Global monitoring of the impact of the covid-19 pandemic through online surveys sampled from the facebook user base, Proceedings of the National Academy of Sciences 118 (2021).

[2] L. J. Akinbami, L. R. Petersen, S. Sami, N. Vuong, S. L. Lukacs, L. Mackey, J. Atas, B. J. LaFleur, Coronavirus Disease 2019 Symptoms and Severe Acute Respiratory Syndrome Coronavirus 2 Antibody Positivity in a Large Survey of First Responders and Healthcare Personnel, May-July 2020, Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 73 (2021) e822–e825.

[3] M. Klompas, Coronavirus disease 2019 (covid-19): protecting hospitals from the invisible, 2020.

[4] Y. Zoabi, S. Deri-Rozov, N. Shomron, Machine learning-based prediction of covid-19 diagnosis based on symptoms, npj digital medicine 4 (2021) 1–5.

[5] D. S. Smith, E. A. Richey, W. L. Brunetto, A symptom-based rule for diagnosis of covid-19, SN comprehensive clinical medicine 2 (2020) 1947–1954.

[6] C. Menni, A. M. Valdes, M. B. Freidin, C. H. Sudre, L. H. Nguyen, D. A. Drew, S. Ganesh, T. Varsavsky, M. J. Cardoso, J. S. E.-S. Moustafa, et al., Real-time tracking of self-reported symptoms to predict potential COVID-19, Nature medicine 26 (2020) 1037–1040.

[7] A. T. Chan, J. S. Brownstein, Putting the public back in public health—surveying symptoms of covid-19, New England Journal of Medicine 383 (2020).

[8] W. E. Allen, H. Altae-Tran, J. Briggs, X. Jin, G. McGee, A. Shi, R. Raghavan, M. Kamariza, N. Nova, A. Pereta, et al., Population-scale longitudinal mapping of covid-19 symptoms, behaviour and testing, Nature human behaviour 4 (2020) 972–982.

[9] L. T. Roland, J. G. Gurrola, P. A. Loftus, S. W. Cheung, J. L. Chang, Smell and taste symptom-based predictive model for covid-19 diagnosis, in: International forum of allergy & rhinology, Wiley Online Library, pp. 832–838.

[10] J. Rufino, J. M. Ramírez, J. Aguilar, C. Baquero, J. Champati, D. Frey, R. E. Lillo, A. Fernández-Anta, Consistent comparison of symptom-based methods for covid-19 infection detection, International Journal of Medical Informatics 177 (2023) 105133.

[11] Coronavirus Disease 2019 (COVID-19) 2020 Interim Case Definition, Approved April 5, 2020, National Notifiable Diseases Surveillance System (NNDSS) (2020).

[12] World Health Organization, Coronavirus disease (COVID-19) Q&A, https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19, 2020. Accessed: 2021-06-02.

[13] J. Álvarez, C. Baquero, E. Cabana, J. P. Champati, A. F. Anta, D. Frey, A. Garcia-Agundez, C. Georgiou, M. Goessens, H. Hernández, R. Lillo, R. Menezes, R. Moreno, N. Nicolaou, O. Ojo, A. Ortega, E. Rausell, J. Rufino, E. Stavrakis, G. Jeevan, C. Glorioso, Estimating Active Cases of COVID-19, medRxiv (2021).

[14] B. Pérez-Gómez, R. Pastor-Barriuso, M. Pérez-Olmeda, M. A. Hernán, J. Oteo-Iglesias, N. F. de Larrea, A. Fernández-García, M. Martín, P. Fernández-Navarro, I. Cruz, et al., Ene-covid nationwide serosurvey served to characterize asymptomatic infections and to develop a symptom-based risk score to predict covid-19, Journal of clinical epidemiology (2021).

[15] J. A. Salomon, A. Reinhart, A. Bilinski, E. J. Chua, W. La Motte-Kerr, M. M. Rönn, M. B. Reitsma, K. A. Morris, S. LaRocca, T. H. Farag, et al., The US COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination, Proceedings of the National Academy of Sciences 118 (2021).

[16] S. Shoer, T. Karady, A. Keshet, S. Shilo, H. Rossman, A. Gavrieli, T. Meir, A. Lavon, D. Kolobkov, I. Kalka, et al., Who should we test for covid-19? a triage model built from national symptom surveys, Medrxiv (2020).

[17] J. Mika, J. Tobiasz, J. Zyla, A. Papiez, M. Bach, A. Werner, M. Kozielski, M. Kania, A. Gruca, D. Piotrowski, et al., Symptom-based early-stage differentiation between sars-cov-2 versus other respiratory tract infections—upper silesia pilot study, Scientific reports 11 (2021) 1–13.

[18] A. Bhattacharya, P. Ranjan, A. Kumar, M. Brijwal, R. M. Pandey, N. Mahishi, U. Baitha, S. Pandey, A. Mittal, N. Wig, Development and validation of a clinical symptom-based scoring system for diagnostic evaluation of covid-19 patients presenting to outpatient department in a pandemic situation, Cureus 13 (2021).

[19] F. Kreuter, N. Barkay, A. Bilinski, A. Bradford, S. Chiu, R. Eliat, J. Fan, T. Galili, D. Haimovich, B. Kim, et al., Partnering with a global platform to inform research and public policy making, in: Survey Research Methods, pp. 159–163.

[20] J. Chen, S. Yuan, D. Lv, Y. Xiang, A novel self-learning feature selection approach based on feature attributions, Expert Systems with Applications 183 (2021) 115219.

[21] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Interdiscip Rev Data Min Knowl Discov. 9 (2019).

[22] R. Nyrup, D. Robinson, Explanatory pragmatism: a context-sensitive framework for explainable medical ai, Ethics Inf Technol. 24 (2022).

[23] F. Gabbay, S. Bar-Lev, O. Montano, N. Hadad, A lime-based explainable machine learning model for predicting the severity level of covid-19 diagnosed patients, Applied Sciences 11 (2021).

[24] I. Girardi, P. Vagenas, D. Arcos-Diaz, I. Bessa, A. Bu-Sser, L. Furlan, R. Furlan, M. Gatti, A. Giovannini, E. Hoeven, C. Marchiori, On the explainability of hospitalization prediction on a large covid-19 patient dataset, in: AMIA Annu Symp Proc., pp. 526–535.

[25] J. Novak, T. Maljur, K. Drenska, Transferring ai explainability to user-centered explanations of complex covid-19 information, in: J. Y. C. Chen, G. Fragomeni, H. Degen, S. Ntoa (Eds.), HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence, Springer Nature Switzerland, Cham, 2022, pp. 441–460.

[26] J.-B. Excoffier, N. Salaün-Penquer, M. Ortala, M. Raphaël-Rousseau, C. Chouaid, C. Jung, Analysis of covid-19 inpatients in france during first lockdown of 2020 using explainability methods, Medical & Biological Engineering & Computing 60 (2022) 1647–1658.

[27] Z. Yu, A. Sohail, T. A. Nofal, J. M. R. Tavares, Explainability of neural network clustering in interpreting the covid-19 emergency data, Fractals 30 (2022) 2240122.

[28] N. Phongchit, P. Taeprasartsit, Prediction performance and explainability of covid-19 classification models, in: 2021 25th International Computer Science and Engineering Conference (ICSEC), IEEE, pp. 383–387.

[29] W. Aldhahi, S. Sull, Uncertain-cam: Uncertainty-based ensemble machine voting for improved covid-19 cxr classification and explainability, Diagnostics 13 (2023) 441.

[30] S. Ali, A. Hussain, S. Bhattacharjee, A. Athar, Abdullah, H.-C. Kim, Detection of covid-19 in x-ray images using densely connected squeeze convolutional neural network (dcscnn): Focusing on interpretability and explainability of the black box model, Sensors 22 (2022).

[31] P. Saxena, S. K. Singh, G. Tiwary, Y. Mittal, I. Jain, An artificial intelligence technique for covid-19 detection with explainability using lungs x-ray images, in: 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), IEEE, pp. 1–6.

[32] M. Li, X. Li, Y. Jiang, J. Zhang, H. Luo, S. Yin, Explainable multi-instance and multi-task learning for covid-19 diagnosis and lesion segmentation in ct images, Knowledge-Based Systems 252 (2022) 109278.
[33] N. D. Kathamuthu, S. Subramaniam, Q. H. Le, S. Muthusamy, H. Panchal, S. C. M. Sundararajan, A. J. Alrubaie, M. M. A. Zahra, A deep transfer learning-based convolution neural network model for covid-19 detection using computed tomography scan images for medical applications, Advances in Engineering Software 175 (2023) 103317.
[34] A. Deeb, A. Debow, S. Mansour, V. Shkodyrev, Covid-19 diagnosis with deep learning: Adjacent-pooling ctscan-covid-19 classifier based on resnet and cbam, Biomedical Signal Processing and Control 86 (2023) 105285.
[35] Z. Ullah, M. Usman, J. Gwak, Mtss-aae: Multi-task semi-supervised adversarial autoencoding for covid-19 detection based on chest x-ray images, Expert Systems with Applications 216 (2023) 119475.
[36] M. Ershadi, Z. Rise, Fusing clinical and image data for detecting the severity level of hospitalized symptomatic covid-19 patients using hierarchical model, Research on Biomedical Engineering 39 (2023) 209—232.
[37] A. Arabameri, S. C. Pal, F. Rezaie, R. Chakrabortty, A. Saha, T. Blaschke, M. D. Napoli, O. Ghorbanzadeh, P. T. T. Ngo, Decision tree based ensemble machine learning approaches for landslide susceptibility mapping, Geocarto International 37 (2022) 4594–4627.
[38] M. Yasir, A. M. Karim, S. K. Malik, A. A. Bajaffer, E. I. Azhar, Application of decision-tree-based machine learning algorithms for prediction of antimicrobial resistance, Antibiotics 11 (2022).
[39] X. Y. Liew, N. Hameed, J. Clos, An investigation of xgboost-based algorithm for breast cancer classification, Machine Learning with Applications 6 (2021) 100154.
[40] J. M. Ramirez, J. I. M. Torre, H. Arguello, Feature fusion via dual-resolution compressive measurement matrix analysis for spectral image classification, Signal Processing: Image Communication 90 (2021) 116014.
[41] K. U. Birant, Multi-view rank-based random forest: A new algorithm for prediction in esports, Expert Systems 39 (2022) e12857.
[42] A. Delgado-Panadero, B. Hernández-Lorca, M. T. García-Ordás, J. A. Benítez-Andrades, Implementing local-explainability in gradient boosting trees: Feature contribution, Information Sciences 589 (2022) 199–212.
[43] N. Burkart, M. F. Huber, A survey on the explainability of supervised machine learning, J. Artif. Int. Res. 70 (2021) 245–317.
[44] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in: IJCAI-17 workshop on explainable AI (XAI), pp. 8–13.
[45] M. Z. Alam, M. S. Rahman, M. S. Rahman, A random forest based predictor for medical data classification using feature ranking, Informatics in Medicine Unlocked 15 (2019) 100180.
[46] A. Messalas, Y. Kanellopoulos, C. Makris, Model-agnostic interpretability with shapley values, in: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), pp. 1–7.
[47] G. James, D. Witten, T. Hastie, R. Tibshirani, et al., An introduction to statistical learning, volume 112, Springer, 2013.
[48] A. Jović, K. Brkić, N. Bogunović, A review of feature selection methods with applications, in: 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), Ieee, pp. 1200–1205.
[49] R. Kohavi, G. H. John, Wrappers for feature subset selection, Artificial intelligence 97 (1997) 273–324.
[50] I. Guyon, J. Makhoul, R. Schwartz, V. Vapnik, What size test set gives good error rate estimates?, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 52–64.
[51] R. D. Riley, J. Ensor, K. I. Snell, F. E. Harrell, G. P. Martin, J. B. Reitsma, K. G. Moons, G. Collins, M. Van Smeden, Calculating the sample size required for developing a clinical prediction model, Bmj 368 (2020).
[52] N. Japkowicz, M. Shah, Evaluating learning algorithms: a classification perspective, Cambridge University Press, 2011.
[53] J. Demšar, Statistical comparisons of classifiers over multiple data sets, The Journal of Machine learning research 7 (2006) 1–30.
[54] J. Rufino, C. Baquero, D. Frey, C. A. Glorioso, A. Ortega, N. Reščič, J. C. Roberts, R. E. Lillo, R. Menezes, J. P. Champati, et al., Using survey data to estimate the impact of the omicron variant on vaccine efficacy against covid-19 infection, Scientific Reports 13 (2023) 900.
[55] C. Baquero, P. Casari, A. Fernandez Anta, A. García-García, D. Frey, A. Garcia-Agundez, C. Georgiou, B. Girault, A. Ortega, M. Goessens, et al., The coronasurveys system for covid-19 incidence data collection and processing, Frontiers in Computer Science 3 (2021) 641237.
[56] J. Mullol, I. Alobid, F. Mariño-Sánchez, A. Izquierdo-Domínguez, C. Marin, L. Klimek, D. Wang, Z. Liu, The loss of smell and taste in the covid-19 outbreak: a tale of many countries, Curr Allergy Asthma Rep. 20 (2020).
[57] M. Hannum, R. Koch, V. Ramirez, S. Marks, A. Toskala, R. Herriman, C. Lin, P. Joseph, R. DR., Taste loss as a distinct symptom of covid-19: A systematic review and meta-analysis, Chem Senses (2022).

Supplemental Materials: Feature Selection for an Explainability Analysis in Detection of COVID-19 Active Cases in Countries from Facebook User-Based Online Surveys

1. Questions of the survey

- **B1 In the last 24 hours, have you had any of the following?** Fever (B1_1), Cough (B1_2), Difficulty breathing (B1_3), Fatigue (B1_4), Stuffy or runny nose (B1_5), Aches or muscle pain (B1_6), Sore throat (B1_7), Chest pain (B1_8), Nausea (B1_9), Loss of smell or taste (B1_10), Headache (B1_12), Chills (B1_13).
- **B1b Are any of these symptoms unusual for you?** Fever (B1b_x1), Cough (B1b_x2), Difficulty breathing (B1b_x3), Fatigue (Bb1_x4), Stuffy or runny nose (B1b_x5), Aches or muscle pain (B1b_x6), Sore throat (B1b_x7), Chest pain (B1b_x8), Nausea (B1b_x9), Loss of smell or taste (B1b_x10), Headache (B1b_x12), Chills (B1b_x13).
- **B3 Do you personally know anyone in your local community who is sick with a fever and either a cough or difficulty breathing?** Yes(1), No(2).
- **B5 Have you spent time with any of these people in the last 7 days?** Yes(1), No(2)[1]
- **B6 Have you ever been tested for coronavirus (COVID-19)?** Yes(1), No(2)[1]
- **B7 Have you been tested for COVID-19 in the past 14 days?** Yes(1), No(2).
- **B8 Did your most recent test find that you had COVID-19?** Yes(1), No(2), I don't know(3).
- **B9 Did you have to pay anything out-of-pocket for this test?** Yes(1), No(2), I don't know(3).[1]
- **B10 Have you or your household had to reduce spending on things you need (such as food,housing, or medication) because of the cost you paid to get the coronavirus (COVID-19) test?** Yes(1), No(2), I don't know(3).[1]
- **B10 Have you or your household had to reduce spending on things you need (such as food, housing, or medication) because of the cost you paid to get the coronavirus (COVID-19) test?** Yes(1), No(2), I don't know(3).[1]
- **B12_1 Do any of the following reasons describe why you haven't been tested for coronavirus (COVID-19) in the last 14 days? I tried to get a test but was not able to get one** Yes(1), No(2).[1]
- **B12_2 Do any of the following reasons describe why you haven't been tested for coronavirus (COVID-19) in the last 14 days? I don't know where to go** Yes(1), No(2).[1]
- **B12_3 Do any of the following reasons describe why you haven't been tested for coronavirus (COVID-19) in the last 14 days? I can't afford the cost of the test** Yes(1), No(2).[1]
- **B12_4 Do any of the following reasons describe why you haven't been tested for coronavirus (COVID-19) in the last 14 days? I don't have time to get tested** Yes(1), No(2).[1]
- **B12_5 Do any of the following reasons describe why you haven't been tested for coronavirus (COVID-19) in the last 14 days? I am unable to travel to a testing location (including because of transportation cost, safety, or physical limitations)** Yes(1), No(2).[1]
- **B12_6 Do any of the following reasons describe why you haven't been tested for coronavirus (COVID-19) in the last 14 days? I am worried about bad things happening to me or my family (including discrimination, government policies, and social stigma)** Yes(1), No(2).[1]
- **B13_1 In the last 30 days, was there any time when you needed any of the following health services or products but could not get it? Emergency transportation services or emergency rescue** Yes(1), No(2).[1]

[1] Only for 2020

- **B13_2 In the last 30 days, was there any time when you needed any of the following health services or products but could not get it? Medical care with overnight stay in any type of facility** Yes(1), No(2).[1]
- **B13_3 In the last 30 days, was there any time when you needed any of the following health services or products but could not get it? Medical or dental care or treatment without an overnight stay** Yes(1), No(2).[1]
- **B13_4 In the last 30 days, was there any time when you needed any of the following health services or products but could not get it? Preventive health services (including immunization/vaccination, family planning, prenatal/postnatal care, routine check-up services)** Yes(1), No(2).[1]
- **B13_5 In the last 30 days, was there any time when you needed any of the following health services or products but could not get it? Medication** Yes(1), No(2).[1]
- **B13_6 In the last 30 days, was there any time when you needed any of the following health services or products but could not get it? Mask, medical gloves, or other protective equipment** Yes(1), No(2).[1]
- **B13_7 In the last 30 days, was there any time when you needed any of the following health services or products but could not get it? Eyeglasses, hearing aid, crutches, band-aids/plasters, thermometer, or any other health product** Yes(1), No(2).[1]
- **B14_1 In the last 30 days, were you unable to get needed treatment, services, medicine, or medical products for any of the following reasons? I didn't know where to go** Yes(1), No(2).[1]
- **B14_2 In the last 30 days, were you unable to get needed treatment, services, medicine, or medical products for any of the following reasons? I couldn't afford the treatment, service, or product** Yes(1), No(2).[1]
- **B14_3 In the last 30 days, were you unable to get needed treatment, services, medicine, or medical products for any of the following reasons? I was unable to travel to the health care provider (including because of transportation cost, safety, or physical limitations)** Yes(1), No(2).[1]
- **B14_4 In the last 30 days, were you unable to get needed treatment, services, medicine, or medical products for any of the following reasons? I was afraid of being infected at the health care provider** Yes(1), No(2).[1]
- **B14_5 In the last 30 days, were you unable to get needed treatment, services, medicine, or medical products for any of the following reasons?The treatment, service, or product was not available** Yes(1), No(2).[1]
- **C0_1 In the last 24 hours, have you done any of the following? Gone to work outside the place where you are currently staying** Yes(1), No(2)
- **C0_2 In the last 24 hours, have you done any of the following? Gone to a market, grocery store, or pharmacy** Yes(1), No(2)
- **C0_3 In the last 24 hours, have you done any of the following? Gone to a restaurant, cafe, or shopping center** Yes(1), No(2)
- **C0_4 In the last 24 hours, have you done any of the following? Spent time with someone who isn't currently staying with you** Yes(1), No(2)
- **C0_5 In the last 24 hours, have you done any of the following? Attended a public event with more than 10 people** Yes(1), No(2)
- **C0_6 In the last 24 hours, have you done any of the following? Used public transit** Yes(1), No(2)
- **C1_m In the last 24 hours, have you had direct contact with anyone who is not staying with you? Direct contact means spending longer than one minute within two meters of someone or touching, including shaking hands, hugging, or kissing.** Yes(1), No(2)[1]
- **C2 How many people, who are not staying with you, have you had this kind of direct contact with in the last 24 hours?** 1-4 people(1), 5-9 people(2), 10-19 people (3), 20 or more(4).[1]

- **C3 Do you have access to soap and water for washing your hands at the place where you are currently staying?** Yes(1), No(2)[1]
- **C5 In the last 7 days, how often did you wear a mask when in public?** All of the time(1), Most of the time(2), Some of the time(3), A little of the time(4), None of the time(5), I have not been in public during the last 7 days(6)
- **C6 In the last 7 days, how many days have you spent time with people who aren't staying with you?** 0 days(1), 1 day(2), 2-4 days(3),5-7 days(4).[1]
- **C7 In the last 24 hours, about how many times have you washed your hands with soap and water or used hand sanitizer?** 0 times (1), 1-2 times (2), 3-6 times (3),7 or more times (4).[1]
- **C8 Do you have access to soap and water for washing your hands at the place where you are currently staying?**Yes (1), No (2)[1]
- **C13a During which activities in the past 24 hours did you wear a mask? Please select all that apply.** Gone to work or school indoors, outside the place where you are currently staying (1), Gone to an indoor market, grocery store, or pharmacy (2), Had a drink or meal indoors at a bar, restaurant, or cafe (3), Spent time indoors with someone who isn't currently staying with you (4), Attended an indoor event with more than 10 people (5),Used public transit (6). [2]
- **D1 During the last 7 days, how often did you feel so nervous that nothing could calm you down?** All of the time(1), Most of the time(2), Some of the time(3), A little of the time(4), None of the time(5).
- **D2 During the last 7 days, how often did you feel so depressed that nothing could cheer you up?** All of the time(1), Most of the time(2), Some of the time(3), A little of the time(4), None of the time(5).
- **D3 How worried are you that you or someone in your immediate family might become seriously ill from coronavirus (COVID-19)?** Very worried(1), Somewhat worried(2), Not too worried(3), Not worried at all(4).[1]
- **D4 How worried are you about having enough to eat in the next week?** Very worried(1), Somewhat worried(2), Not too worried(3), Not worried at all(4).
- **D5 How worried are you about your household's finances in the next month?** Very worried(1), Somewhat worried(2), Not too worried(3), Not worried at all(4).
- **D7 In the past 4 weeks, did you do any work for pay? By work for pay, we mean any kind of business, farming, or other activity to earn money, even if only for one hour.** Yes (1), No(2).
- **D8 Before February 2020, were you working for pay, or doing any kind of business, farming, or other activity to earn money?** Yes(1), No(2)[1]
- **D9 Why did you stop working?** My employer closed for coronavirus-related reasons (1), My employer closed for another reason (2), I was laid off or furloughed (3), I am a seasonal worker (4), I was ill or quarantined (5), I needed to care for someone (6), Other (7)[1]
- **D10 What is the main activity of the business or organization in which you work?**Agriculture (1),Buying and selling (2), Construction (3),Education (4),Electricity / water / gas / waste (5),Financial / insurance / real estate services (6), Health (7), Manufacturing (8), Mining (9), Personal services (10), Professional / scientific / technical activities (11), Public administration (12), Tourism (13), Transportation (14), Other (15).
- **E2 Which of these best describes the area where you are currently staying?** City(1), Town(2), Village or rural area(3).
- **E3 What is your gender?** Male(1), Female(2), Other(3), Prefer not to answer(4)

[2]Only for 2021

- **E4 What is your age?** 18-24 years(1), 25-34 years(2), 35-44 years(3), 45-54 years(4), 55-64 years(5), 65-74 years(6), 75 years or older(7).
- **V1 Have you had a COVID-19 vaccination?**Yes (1), No (2),I don't know (3)[2]
- **V2 How many COVID-19 vaccinations have you received?**1 vaccination or dose (1), 2 vaccinations or doses (2),I don't know (3)[2]
- **V3 If a vaccine to prevent COVID-19 were offered to you today, would you choose to get vaccinated?** Yes, definitely (1), Yes, probably (2), No, probably not (3), No, definitely not (4)[2]
- **V5a Which of the following, if any, are reasons that you definitely wouldn't choose to get a COVID-19 vaccine? Please select all that apply.** I am concerned about possible side effects of a COVID-19 vaccine (1), I don't know if a COVID-19 vaccine will work (2), I don't believe I need a COVID-19 vaccine (3),I don't like vaccines (4), I plan to wait and see if it is safe and may get it later (5), I think other people need it more than I do right now (6), I am concerned about the cost of a COVID-19 vaccine (7), It is against my religious beliefs (8), I don't trust the government (10), Other (9)[2]
- **V5b Which of the following, if any, are reasons that you probably wouldn't choose to get a COVID-19 vaccine? Please select all that apply.** I am concerned about possible side effects of a COVID-19 vaccine (1), I don't know if a COVID-19 vaccine will work (2), I don't believe I need a COVID-19 vaccine (3), I don't like vaccines (4), I plan to wait and see if it is safe and may get it later (5),I think other people need it more than I do right now (6), I am concerned about the cost of a COVID-19 vaccine (7), It is against my religious beliefs (8), I don't trust the government (10), Other (9)[2]
- **V5c Which of the following, if any, are reasons that you probably wouldn't choose to get a COVID-19 vaccine? Please select all that apply.** I am concerned about possible side effects of a COVID-19 vaccine (1), I don't know if a COVID-19 vaccine will work (2), I don't believe I need a COVID-19 vaccine (3), I don't like vaccines (4), I plan to wait and see if it is safe and may get it later (5),I think other people need it more than I do right now (6), I am concerned about the cost of a COVID-19 vaccine (7), It is against my religious beliefs (8), I don't trust the government (10), Other (9).[2]
- **V6 Why don't you believe that you need a COVID-19 vaccine? Please select all that apply.**I already had COVID-19 (1), I do not spend time with any high-risk people (2), I am not a member of a high-risk group (3), I plan to use masks or other precautions instead (4), I don't believe COVID-19 is a serious illness (5), I don't think vaccines are beneficial (6), Other (7).[2]
- **V10 Have you ever been told by a doctor, nurse, or other health professional that you have any of the following medical conditions? Please select all that apply.** Asthma (1), Chronic lung disease such as COPD, chronic bronchitis, or emphysema (2), Cancer (3), Diabetes (4), High blood pressure (5), Kidney disease (6), Weakened or compromised immune system (7), Heart attack, heart disease, or other heart condition (8), Obesity (9), one of these (10).[2]
- **V15 Do you have an appointment to receive a COVID-19 vaccine?** Yes (1),No (2)[2]
- **V16 Have you tried to get an appointment to receive a COVID-19 vaccine?** Yes (1),No (2)[2]

2. Results

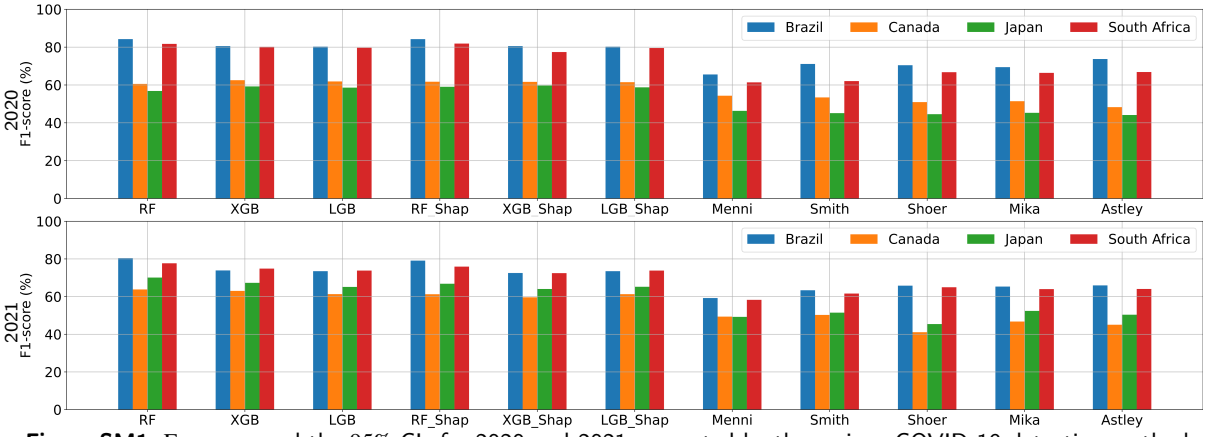

Figure SM1: F_1-scores and the 95% CIs for 2020 and 2021 generated by the various COVID-19 detection methods.

Table SM1

F_1-scores and the 95% CIs yielded by various COVID-19 detection methods for the four countries and for 2020 and 2021.

Year	Method	Brazil	Canada	Japan	South Africa
2020	RF	84.24 (84.19 - 84.29)	60.57 (59.96 - 61.17)	56.84 (55.96 - 57.72)	81.74 (81.54 - 81.94)
	XGB	80.56 (80.50 - 80.62)	62.53 (61.98 - 63.09)	59.24 (58.36 - 60.13)	80.19 (79.96 - 80.42)
	LGB	80.28 (80.22 - 80.33)	61.87 (61.36 - 62.38)	58.60 (57.76 - 59.43)	79.66 (79.45 - 79.87)
	RF_SHAP	84.23 (84.17 - 84.28)	61.72 (61.12 - 62.31)	59.05 (58.27 - 59.83)	81.88 (81.68 - 82.09)
	XGB_SHAP	80.56 (80.51 - 80.62)	61.65 (61.12 - 62.17)	59.70 (58.82 - 60.57)	77.41 (77.19 - 77.64)
	LGB_SHAP	80.26 (80.20 - 80.31)	61.48 (60.88 - 62.07)	58.76 (57.91 - 59.61)	79.54 (79.34 - 79.75)
	Menni [6]	65.56 (65.48 - 65.64)	54.33 (53.66 - 54.99)	46.33 (45.33 - 47.33)	61.39 (61.07 - 61.70)
	Smith [5]	71.11 (71.05 - 71.18)	53.43 (52.85 - 54.01)	45.12 (44.42 - 45.82)	62.06 (61.80 - 62.32)
	Shoer [16]	70.45 (70.39 - 70.52)	50.95 (50.37 - 51.54)	44.57 (43.86 - 45.28)	66.76 (66.52 - 67.00)
	Mika [17]	69.43 (69.37 - 69.49)	51.43 (50.86 - 52.01)	45.29 (44.65 - 45.94)	66.40 (66.13 - 66.68)
	Astley [1]	73.72 (73.65 - 73.78)	48.29 (47.58 - 49.00)	44.13 (43.32 - 44.93)	66.85 (66.61 - 67.09)
2021	RF	80.43 (80.39 - 80.47)	63.80 (63.52 - 64.08)	70.11 (69.84 - 70.37)	77.69 (77.53 - 77.85)
	XGB	73.89 (73.85 - 73.94)	63.03 (62.75 - 63.32)	67.31 (67.03 - 67.60)	74.87 (74.69 - 75.05)
	LGB	73.50 (73.45 - 73.54)	61.32 (61.04 - 61.61)	65.14 (64.87 - 65.41)	73.82 (73.64 - 74.00)
	RF_SHAP	79.11 (79.07 - 79.15)	61.26 (60.97 - 61.56)	66.83 (66.58 - 67.08)	75.92 (75.75 - 76.11)
	XGB_SHAP	72.53 (72.49 - 72.58)	59.63 (59.33 - 59.93)	64.05 (63.77 - 64.34)	72.46 (72.27 - 72.64)
	LGB_SHAP	73.50 (73.45 - 73.54)	61.30 (61.01 - 61.59)	65.22 (64.96 - 65.48)	73.82 (73.64 - 74.00)
	Menni [6]	59.24 (59.18 - 59.31)	49.38 (49.02- 49.74)	49.24 (49.16 - 49.83)	58.28 (58.06 - 58.50)
	Smith [5]	63.37 (63.32 - 63.42)	50.28 (49.99 - 50.57)	51.48 (51.23 -51.74)	61.62 (61.45 - 61.80)
	Shoer [16]	65.81 (65.76 - 65.87)	41.10 (40.84 - 41.36)	45.42 (45.07 - 45.78)	64.97 (64.80 - 65.15)
	Mika [17]	65.33 (65.28 - 65.38)	46.76 (46.40 - 47.12)	52.41 (51.73 - 53.09)	63.98 (63.81 - 64.15)
	Astley [1]	65.95 (65.90 - 66.01)	45.07 (44.74 - 45.40)	50.39 (50.08 - 50.70)	64.06 (63.88 - 64.24)

Table SM2
Performance metrics in % and the 95% CIs obtained by the proposed COVID-19 detection methods for Canada and for 2020 and 2021.

Year	Method	F_1-score	Specificity	Sensitivity	Precision
2020	RF	60.57 (59.97 - 61.18)	**98.80 (98.75 - 98.85)**	48.50 (47.80 - 49.19)	**80.97 (80.29 - 81.66)**
	XGB	**62.54 (61.99 - 63.09)**	97.96 (97.89 - 98.03)	**54.41 (53.71 - 55.10)**	73.77 (73.07 - 74.47)
	LGB	61.87 (61.36 - 62.39)	98.17 (98.11 - 98.24)	52.65 (52.03 - 53.27)	75.24 (74.54 - 75.95)
	RF SHAP	61.72 (61.13 - 62.32)	98.43 (98.37 - 98.50)	51.39 (50.67 - 52.11)	77.57 (76.83 - 78.31)
	XGB SHAP	61.65 (61.13 - 62.17)	97.79 (97.72 - 97.86)	54.03 (53.37 - 54.69)	72.03 (71.27 - 72.79)
	LGB SHAP	61.48 (60.88 - 62.08)	98.05 (97.97 - 98.12)	52.72 (52.01 - 53.42)	74.00 (73.20 - 74.80)
2021	RF	**63.81 (63.53 - 64.08)**	**98.76 (98.73 - 98.79)**	52.03 (51.71 - 52.35)	**82.53 (82.15 - 82.92)**
	XGB	63.04 (62.75 - 63.32)	98.18 (98.15 - 98.22)	**53.48 (53.16 - 53.79)**	76.82 (76.43 - 77.21)
	LGB	61.33 (61.04 - 61.61)	98.18 (98.14 - 98.22)	51.41 (51.08 - 51.73)	76.05 (75.65 - 76.46)
	RF SHAP	61.27 (60.97 - 61.57)	98.57 (98.53 - 98.60)	49.82 (49.48 - 50.15)	79.64 (79.20 - 80.08)
	XGB SHAP	59.63 (59.33 - 59.93)	98.11 (98.07 - 98.14)	49.66 (49.32 - 50.01)	74.68 (74.29 - 75.08)
	LGB SHAP	61.33 (61.03 - 61.62)	98.18 (98.14 - 98.22)	51.40 (51.07 - 51.73)	76.07 (75.66 - 76.48)

Table SM3
Performance metrics in % and the 95% CIs obtained by the proposed COVID-19 detection methods for Japan and for 2020 and 2021.

Year	Method	F_1-score	Specificity	Sensitivity	Precision
2020	RF	56.85 (55.97 - 57.73)	**98.89 (98.82 - 98.97)**	43.33 (42.45 - 44.21)	**83.22 (82.12 - 84.31)**
	XGB	59.25 (58.36 - 60.13)	97.82 (97.72 - 97.92)	49.51 (48.57 - 50.44)	74.17 (73.04 - 75.30)
	LGB	71.54 (71.24 - 71.85)	98.23 (98.19 - 98.26)	63.99 (63.61 - 64.38)	81.16 (80.81 - 81.50)
	RF SHAP	59.06 (58.28 - 59.84)	97.95 (97.85 - 98.04)	48.86 (48.04 - 49.68)	75.03 (73.96 - 76.09)
	XGB SHAP	59.70 (58.83 - 60.58)	97.82 (97.73 - 97.91)	50.05 (49.13 - 50.98)	74.31 (73.23 - 75.39)
	LGB SHAP	**71.55 (71.25 - 71.85)**	98.21 (98.17 - 98.24)	**64.09 (63.72 - 64.46)**	81.03 (80.68 - 81.38)
2021	RF	70.11 (69.85 - 70.38)	**98.95 (98.93 - 98.98)**	59.24 (58.91 - 59.56)	**85.94 (85.61 - 86.27)**
	XGB	67.32 (67.03 - 67.60)	98.48 (98.44 - 98.51)	57.92 (57.58 - 58.25)	80.42 (80.03 - 80.81)
	LGB	80.28 (80.22 - 80.34)	78.53 (78.43 - 78.63)	**79.37 (79.29 - 79.44)**	81.22 (81.14 - 81.30)
	RF SHAP	66.80 (66.54 - 67.06)	98.55 (98.52 - 98.58)	56.91 (56.58 - 57.25)	80.92 (80.57 - 81.28)
	XGB SHAP	64.06 (63.77 - 64.34)	98.40 (98.36 - 98.43)	54.13 (53.80 - 54.47)	78.50 (78.09 - 78.90)
	LGB SHAP	**80.26 (80.21 - 80.32)**	78.53 (78.42 - 78.64)	79.34 (79.26 - 79.41)	81.21 (81.13 - 81.30)

Table SM4
Performance metrics in % and the 95% CIs obtained by the proposed COVID-19 detection methods for South Africa and 2020 and 2021.

Year	Method	F_1-score	Specificity	Sensitivity	Precision
2020	RF	81.75 (81.54 - 81.95)	**92.55 (92.38 - 92.72)**	78.06 (77.75 - 78.38)	85.83 (85.54 - 86.12)
	XGB	80.19 (79.96 - 80.42)	91.30 (91.13 - 91.48)	77.03 (76.69 - 77.38)	83.66 (83.36 - 83.95)
	LGB	79.66 (79.45 - 79.88)	91.23 (91.06 - 91.40)	76.27 (75.95 - 76.58)	83.41 (83.13 - 83.69)
	RF SHAP	**81.89 (81.68 - 82.10)**	92.53 (92.36 - 92.70)	**78.31 (78.01 - 78.62)**	**85.84 (85.55 - 86.13)**
	XGB SHAP	77.42 (77.19 - 77.64)	90.57 (90.36 - 90.77)	73.49 (73.14 - 73.84)	81.83 (81.50 - 82.17)
	LGB SHAP	79.55 (79.35 - 79.75)	91.25 (91.08 - 91.42)	76.07 (75.77 - 76.36)	83.40 (83.12 - 83.68)
2021	RF	**77.70 (77.54 - 77.86)**	**91.83 (91.73 - 91.93)**	**72.47 (72.24 - 72.70)**	**83.76 (83.57 - 83.95)**
	XGB	74.87 (74.69 - 75.05)	90.08 (89.96 - 90.20)	70.06 (69.83 - 70.29)	80.41 (80.18 - 80.64)
	LGB	73.83 (73.65 - 74.01)	90.10 (89.98 - 90.22)	68.50 (68.25 - 68.74)	80.08 (79.86 - 80.31)
	RF SHAP	75.68 (75.52 - 75.84)	90.70 (90.57 - 90.82)	70.63 (70.42 - 70.84)	81.53 (81.29 - 81.76)
	XGB SHAP	72.46 (72.28 - 72.64)	89.52 (89.41 - 89.63)	67.07 (66.82 - 67.32)	78.81 (78.59 - 79.03)
	LGB SHAP	73.82 (73.64 - 74.00)	90.14 (90.02 - 90.26)	68.44 (68.19 - 68.69)	80.13 (79.91 - 80.36)