



HAL
open science

Consistent comparison of symptom-based methods for COVID-19 infection detection

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

► **To cite this version:**

Jesús Rufino, Juan Marcos Ramírez, Jose Aguilar, Carlos Baquero, Jaya Champati, et al.. Consistent comparison of symptom-based methods for COVID-19 infection detection. *International Journal of Medical Informatics*, 2023, 177, pp.105133. 10.1016/j.ijmedinf.2023.105133 . hal-04406757

HAL Id: hal-04406757

<https://inria.hal.science/hal-04406757v1>

Submitted on 19 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Highlights

Consistent Comparison of Symptom-based Methods for COVID-19 Infection Detection

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

- A consistent comparison of various methods for detecting COVID-19 active cases.
- UMD-CTIS data for six countries and two periods were used to compare detection methods.
- Detection methods were evaluated for countries with different test-positive rates.
- Explainability analysis is conducted to quantify the relevance of symptoms in COVID-19 detection.

Consistent Comparison of Symptom-based Methods for COVID-19 Infection 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

^aIMDEA Networks Institute, 28918, Madrid, Spain

^bCEMISID, Universidad de Los Andes, Mérida, 5101, Venezuela

^cCIDITIC, Universidad EAFIT, Medellín, Colombia

^dUniversidade do Minho and INESC TEC, Braga, Portugal

^eInria Rennes, Rennes, France

^fUniversidad Carlos III, Madrid, Spain

ARTICLE INFO

Keywords:

COVID-19 Detection Methods

Explainability Analysis.

F1-score

Logistic Regression Methods

Rule-based Methods

Tree-based Models

ABSTRACT

Background: During the global pandemic crisis, various detection methods of COVID-19-positive cases based on self-reported information were introduced to provide quick diagnosis tools for effectively planning and managing healthcare resources. These methods typically identify positive cases based on a particular combination of symptoms, and they have been evaluated using different datasets.

Purpose: This paper presents a comprehensive comparison of various COVID-19 detection methods based on self-reported information using the University of Maryland Global COVID-19 Trends and Impact Survey (UMD-CTIS), a large health surveillance platform, which was launched in partnership with Facebook.

Methods: Detection methods were implemented to identify COVID-19-positive cases among UMD-CTIS participants reporting at least one symptom and a recent antigen test result (positive or negative) for six countries and two periods. Multiple detection methods were implemented for three different categories: rule-based approaches, logistic regression techniques, and tree-based machine-learning models. These methods were evaluated using different metrics including F1-score, sensitivity, specificity, and precision. An explainability analysis has also been conducted to compare methods.

Results: Fifteen methods were evaluated for six countries and two periods. We identify the best method for each category: rule-based methods (F1-score: 51.48% - 71.11%), logistic regression techniques (F1-score: 39.91% - 71.13%), and tree-based machine learning models (F1-score: 45.07% - 73.72%). According to the explainability analysis, the relevance of the reported symptoms in COVID-19 detection varies between countries and years. However, there are two variables consistently relevant across approaches: stuffy or runny nose, and aches or muscle pain.


Conclusions: Regarding the categories of detection methods, evaluating detection methods using homogeneous data across countries and years provides a solid and consistent comparison. An explainability analysis of a tree-based machine-learning model can assist in identifying infected individuals specifically based on their relevant symptoms. This study is limited by the self-report nature of data, which cannot replace clinical diagnosis.

1. Introduction

In December 2019, the coronavirus disease 2019 (COVID-19) emerged in China caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. Within a few months, the expansion of this disease triggered a global pandemic crisis that stressed national healthcare systems. In this context, the management of the healthcare resources (hospital beds or intensive care units) was determined by the availability of efficient instruments for tracking the pandemic evolution [2]. In this regard, the antigen test based on reverse transcriptase polymerase chain reaction (RT-PCR) was the standard diagnostic tool for identifying infected people [3]. However, RT-PCR tests required material and human resources that were not always available. These limitations hindered the control of disease expansions and the timely implementation of corrective measures [4].

To overcome these drawbacks, various COVID-19 detection methods based on self-reported health information were developed [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]. In general, these methods identify positive cases based on the most predictive combination of symptoms. Other methods build machine-learning models that evaluate a set of individual features such as symptoms, age groups, and gender. Notice that the techniques have been evaluated using datasets of different sizes and types. In April 2020, the University of Maryland Global COVID-19 Trends and Impact Survey (UMD-CTIS), in partnership with Facebook, launched the largest health surveillance platform to date [19]. More precisely, this project recorded, daily, the responses of invited Facebook users about topics related to the COVID-19 pandemic. This instrument was launched in 56 languages, and it recorded tens of millions of responses from 114 countries or territories worldwide.

This paper presents a consistent comparison of different COVID-19 detection methods based on self-reported

 juan.ramirez@imdea.org (J.M. Ramírez)

ORCID(s): 0000-0003-0000-1073 (J.M. Ramírez)

information. More precisely, we compare the performance of the various detection methods using data extracted from UMD-CTIS for six countries: Brazil, Canada, Israel, Japan, Turkey, and South Africa, and for two periods: 2020 and 2021. We selected countries based on their geographical diversity and the availability of sufficient data samples. In addition, we analyze the performance for 2020 and 2021, which represent different periods during the pandemic: with and without vaccination. Some methods provide either the prediction rules or model parameters [9, 10, 5, 20, 12, 16, 6, 7, 21], so the training phase is not necessary. On the contrary, other methods require a training phase to optimize the detection engines based on machine-learning models [11, 14, 13, 15]. The performance of each method is evaluated using four metrics: F_1 -score, sensitivity, specificity, and precision. Since imbalanced classes affect the estimation of the F_1 -score, in addition to our comparative analysis on each country and period, we also evaluate the methods for three groups of countries: the entire set of the six countries, the countries with a high-test positive rate (TPR), and the countries with a low TPR. Lastly, an explainability analysis is conducted on the best detection method per category.

There are few studies comparing COVID-19 detection techniques with self-reported data. Yalçın and Ünalı [22] examined the performance of various machine-learning models using a dataset with symptoms (e.g., fever, dry cough, and breathing problems) and other features such as contact with infected people, and mask-wearing. Specifically, Yalçın and Ünalı built detectors based on the K-nearest neighbor, multilayer perceptron neural networks, logistic regression, gated recurrent unit, support vector machines, long short-term memory, and deep learning algorithms. This approach is limited by the fact that it does not elaborate on the optimization of the machine learning models or model architectures. In contrast to [22], our approach compares the performance of methods widely used for COVID-19 detection at early pandemic stages. Moreover, we analyze the explainability of the most relevant features for detecting COVID-19 positives. Moreover, Sedik et al. [23] proposed two data-augmentation models to study the learnability of both Convolutional Neural Networks and Convolutional Long Short-Term Memory-based deep learning models. The method proposed by Sedik et al. detects positive cases by applying deep learning techniques to different medical imaging modalities. Unlike [23], our approach compares the performance of various COVID-19 detection methods based on self-reported information.

In this paper, we perform a comparative study of various detection methods based on self-reported information using the UMD-CTIS data [24]. The main contributions are twofold:

- We compare the performance of COVID-19 detection techniques based on self-reported information using UMD-CTIS data extracted from six countries for 2020 and 2021. These methods are consistently examined using quality metrics (F_1 -score, sensitivity, specificity, and precision).

- The comparison includes an explainability analysis that considers the response provided by the best detection technique of each category (rule-based approaches, regression techniques, and tree-based classifiers). The explainability analysis identifies the relevant features in COVID-19 detection.

In general, the detection methods exhibiting the best performances across different groups and metrics are **Smith** [10] (F_1 -score: 56.59%), **Astley** [15] (F_1 -score: 55.97%), **Menni** [9] (F_1 -score: 55.45%), **Mika** [13] (F_1 -score: 53.98%), and **Shoer** [14] (F_1 -score: 53.35%). Individual features associated with the best detection methods are loss of smell, loss of taste, cough, and fever.

The article is organized as follows. Section 2 describes the information to carry out the experiments (datasets, quality metrics, and the experimental protocol). Section 3 shows the results yielded by each method using the same datasets, as well as an explainability analysis of the best detection technique per category. Finally, Section 4 makes a general analysis of the achievements, and summarizes conclusions and future work.

2. Experiments

2.1. Dataset

From April 23, 2020, Facebook worldwide users outside the USA were invited to participate in the UMD-CTIS by displaying a banner on the user page. Users who accepted the invitation were moved to a web-survey platform, where potential participants must report age ≥ 18 and consent of data use before responding to the survey. The survey, designed by the University of Maryland, consists of a questionnaire collecting information on gender, age groups, symptoms, and COVID-19 testing, among others. These questionnaires were translated into 56 languages for 114 countries and territories. Furthermore, the survey instrument was continuously updated. Finally, UMD organized and stored daily microdata that was further processed to develop our comparative study.

Based on the UMD-CTIS data, we compare the performance of different COVID-19 detection methods in six countries: Brazil, Canada, Israel, Japan, Turkey, and South Africa. These countries are selected based on geographical diversity and a large amount of available data. Furthermore, we compare the performance yielded by the various methods for two periods: (2020) from April 23 to December 31, 2020, and (2021) from January 1 to December 31, 2021. Notice that the end of 2020 matches the start of the first COVID-19 vaccination campaigns. Therefore, we analyze the detection methods without and with information on vaccination acceptance. We extract samples from respondents who reported at least one symptom within 24 hours and a test result (positive or negative) within the preceding 14 days. As can be seen in Table 1, 83,238 respondents from Brazil reported a test outcome and at least one symptom in 2020. In this cohort, 44,963 participants reported a positive test result, and 38,275 respondents had a negative test outcome. Table 1 also includes the test positive rate (TPR) where

Table 1

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

Characteristic	Brazil		Canada		Israel		Japan		Turkey		South Africa	
	2020	2021	2020	2021	2020	2021	2020	2021	2020	2021	2020	2021
1. Tested symptomatic, N	83238	262683	8927	33997	5944	19063	4698	41010	15952	28896	7883	23038
2. Test outcome												
(a) Positive, N	44963	106471	838	3433	1238	2869	532	4011	6167	9228	2866	8459
(b) Negative, N	38275	156212	8089	30564	4706	16194	4166	36999	9785	19668	5017	14579
(c) TPR, %	54.02	40.53	9.39	10.10	20.83	15.05	11.32	9.78	38.66	31.94	36.35	36.71
3. Gender												
(a) Female, N	45357	130235	5438	19472	2941	9290	1679	14283	3939	7185	3923	11291
(b) Male, N	24928	76689	2315	9824	2199	6746	2388	20791	8920	15292	2525	6730
4. Age groups												
(a) 18-24, N	8270	27474	1136	3248	583	1498	179	871	1716	2267	739	1580
(b) 25-34, N	19596	56227	2337	7172	1144	3069	577	3797	4375	5756	2252	4889
(c) 35-44, N	21061	57452	1750	6688	1041	3333	997	7527	4043	7110	1801	4721
(d) 45-54, N	13776	39122	1210	5215	933	3115	1216	10413	2071	4594	1141	3878
(e) 55-64, N	6968	22190	954	4478	880	2634	828	8724	862	2400	491	2124
(f) 65-74, N	140	6016	308	2421	510	1957	479	3529	158	719	1667	799
(g) 75+, N	233	1025	126	825	143	627	66	846	21	134	27	230
5. Average number of symptoms among positive	5.37	5.16	5.25	5.27	4.99	5.13	4.38	4.45	5.39	5.36	5.51	5.61
6. Symptoms among positive												
(a) Fever, %	22.56	21.92	22.43	22.63	22.70	24.22	39.28	38.49	22.86	25.12	32.55	30.77
(b) Cough, %	54.73	57.46	63.01	67.46	54.93	59.99	61.65	64.47	51.55	55.93	58.89	65.96
(c) Difficulty breathing, %	30.72	28.17	23.74	22.80	24.47	22.55	18.79	16.62	24.58	24.65	29.03	27.61
(d) Fatigue, %	60.51	57.58	69.33	71.13	72.78	73.20	51.50	57.06	69.66	67.51	65.24	67.88
(e) Stuffy or runny nose, %	57.86	57.33	62.29	68.62	50.89	62.39	49.24	47.31	56.22	59.44	55.02	62.59
(f) Aches or muscle pain, %	58.90	58.01	55.13	53.10	55.17	53.29	41.35	44.45	65.02	62.82	57.43	58.73
(g) Sore throat, %	35.06	34.37	34.84	39.67	32.79	33.04	37.21	35.27	40.21	39.04	36.14	38.78
(h) Chest pain, %	32.00	30.03	22.19	21.52	26.90	25.27	20.67	22.88	32.16	30.57	39.25	35.57
(i) Nausea, %	29.94	28.34	26.61	25.08	25.04	24.33	11.65	10.17	26.53	24.60	27.84	28.41
(j) Loss of smell or taste, %	54.15	46.25	53.34	42.67	49.35	49.11	40.22	39.99	52.21	48.41	51.70	45.89
(k) Headache, %	65.74	63.73	60.14	58.86	58.08	56.81	41.35	44.40	58.81	57.26	64.68	65.72
(l) Chills, %	34.96	33.31	32.21	33.46	26.17	28.76	25.56	24.28	39.13	40.86	33.67	33.75
7. Average number of symptoms among negative	3.12	2.88	3.19	2.83	2.69	2.55	2.73	2.28	3.10	3.01	2.85	2.99
8. Symptoms among negative												
(a) Fever, %	6.12	5.79	4.61	4.58	4.99	4.59	19.23	11.61	5.65	6.57	10.94	12.13
(b) Cough, %	34.17	32.75	38.45	32.24	33.09	28.05	37.57	28.55	31.32	32.21	33.57	35.98
(c) Difficulty breathing, %	13.71	11.50	12.34	10.10	11.58	9.52	4.70	3.25	14.62	14.49	10.94	11.10
(d) Fatigue, %	33.46	30.02	53.05	48.95	54.63	57.42	35.29	30.48	44.34	42.29	36.06	38.81
(e) Stuffy or runny nose, %	48.86	47.88	55.09	49.82	42.65	40.31	46.35	44.60	41.79	44.39	40.82	44.61
(f) Aches or muscle pain, %	41.67	40.19	39.85	37.05	26.86	27.58	34.28	35.19	42.10	39.76	33.59	35.87
(g) Sore throat, %	23.76	21.83	27.83	21.90	23.06	18.33	28.11	20.40	26.78	23.81	22.06	22.30
(h) Chest pain, %	15.11	12.97	10.97	8.09	10.43	9.97	10.01	7.24	16.52	14.62	15.15	15.34
(i) Nausea, %	15.37	13.42	16.27	12.99	13.15	12.54	7.97	6.47	14.64	12.87	13.85	14.94
(j) Loss of smell or taste, %	10.70	5.97	4.56	3.54	3.74	3.50	3.48	2.10	8.70	6.60	8.11	7.33
(k) Headache, %	50.90	49.47	43.92	42.75	36.00	34.40	34.49	30.58	43.73	41.76	48.79	47.52
(l) Chills, %	18.15	16.31	11.82	10.77	9.12	8.73	12.00	7.78	20.37	21.34	11.36	12.66

TPR = $(100 \times \text{positive}) / (\text{Tested symptomatic})$. For example, the TPR for Brazil 2020 is 54.02%. For Brazil 2021, the dataset was extracted from 262,683 participants. In this case, 106,471 respondents reported a positive test result, and 156,212 individuals informed a negative test outcome with a TPR of 40.53%. The number of tested symptomatic, positive cases, negative cases, and the TPR in % for the remaining countries in 2020 and 2021 are displayed in Table 1. Additionally, Table 1 provides information on other characteristics such as gender, age groups, the average number of reported symptoms among positives and negatives, and the frequency of symptoms among positives and negatives.

2.2. Experimental Protocol

For every country and period, we build a dataset by picking the answers reporting a lab test done in the last 14 days and at least one potential COVID-19 symptom, i.e., we select the tested and symptomatic cases. We select symptomatic cases because rule-based methods typically aim at finding the most predictive combination of symptoms.

In addition, we choose the tested individuals with the aim of obtaining the ground truth that allows us to build machine-learning models. Since questionnaires contain categorical data, we apply binary encoding such that every potential choice aggregates a column to the dataset. This leads to datasets with 201 features (attributes, columns, or variables) for 2020, and the datasets have between 431 and 452 columns for 2021 depending on the selected country. For each dataset, this study obtains the performance of the various COVID-19 detection methods under test. A brief description of each method is included in the Supplementary Material A. It is important to mention that references to most methods under test provide the parameters to build the detection models [10, 20, 16, 7, 21, 9, 12, 6]. On the other hand, other methods (such as Zoabi [5]) have repositories containing the codes and files to reproduce the reported detection engines. For the remaining methods [14, 11, 15], we carefully follow the procedures outlined in the references to build the detection models, including the hyperparameter

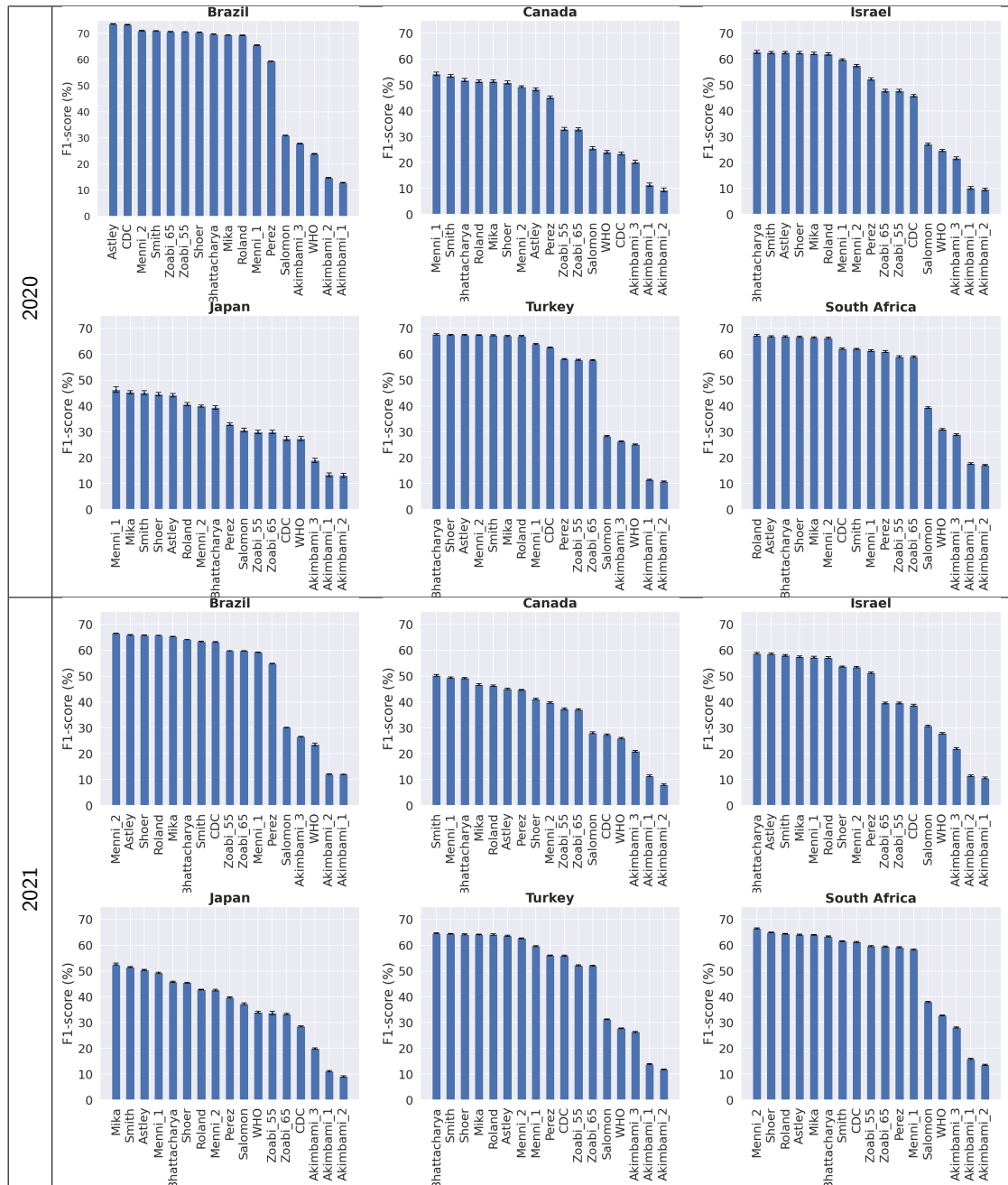


Figure 1: F_1 score in % and the corresponding 95% confidence interval obtained by the various COVID-19 detection methods for the selected countries and for 2020 and 2021.

optimization stage to avoid overfitting. Our study divided every dataset into 100 partitions. For each trial, 80% of the dataset rows (questionnaires or samples) were randomly selected as training samples, and the remaining 20% were used to test the detection methods. On the other hand, verification was made to self-assess the quality of our contribution in the field of machine learning applied to the medical area, using the checklist proposed in [25]. This self-assessment is presented in section C of the supplementary material.

2.3. Metrics

We use the F_1 -score to quantitatively assess the performance of the various detection methods. To this end, our procedure first obtains the predictions over the test set for each trial. From the predicted estimates and the ground truth data, the procedure identifies the number of true positives TP, false positives FP, true negatives TN, and false negatives FN. Then, the F_1 -score is obtained as follows:

$$F_1 = \frac{2TP}{2TP + FP + FN}. \quad (1)$$

We also compute for each trial the sensitivity, specificity, and precision. These metrics are defined as follows:

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (3)$$

$$\text{precision} = \frac{TP}{TP + FP} \quad (4)$$

In this work, laboratory test results provided by survey respondents are used as gold standard outcomes to obtain either TP or TN samples. In particular, sensitivity measures the ability of a detection method to correctly identify infected individuals. Furthermore, specificity quantifies the ability of the detection method to identify healthy individuals correctly. Moreover, precision basically provides the proportion of positive results. For example, a high value of precision indicates that an important proportion of the detected positives are true positives (infected people). Finally, F_1 -score merges precision and sensitivity by computing the harmonic mean to compare the combined performance of them [26, 27, 28].

3. Results

3.1. General Results

Figure 1 displays the F_1 in % scores and the 95% confidence intervals (CIs) yielded by COVID-19 detection methods for the six countries and for 2020 and 2021. Table SM1 in the supplemental material B also shows the F_1 scores and their 95% CIs for the six countries and for 2020. Specifically, every value in this table is obtained by averaging 100 realizations of the corresponding experiment, where for each realization a different test set is evaluated. For 2020, the methods generating the best F_1 scores for each country are: Brazil (**Astley**: 73.72%), Canada (**Menni_1**: 54.33%), Israel (**Bhattacharya**: 62.78%), Japan (**Menni_1**: 46.33%), Turkey (**Bhattacharya**: 67.67%), and South Africa (**Roland**: 67.32%). Additionally, the methods that produce the lowest F_1 scores for each country are: Brazil (**Akinbami_1**: 12.85%), Canada (**Akinbami_2**: 9.41%), Israel (**Akinbami_2**: 9.59%), Japan (**Akinbami_2**: 13.16%), Turkey (**Akinbami_2**: 10.81%), and South Africa (**Akinbami_2**: 17.14%). The F_1 score in % and the CIs obtained for 2021 are displayed in Table SM5 in the supplemental material B. For 2021, the best F_1 scores for each country are: Brazil (**Menni_2**: 66.54%), Canada (**Smith**: 50.28%), Israel (**Bhattacharya**: 58.76%), Japan (**Mika**: 52.41%), Turkey (**Bhattacharya**: 64.61%), and South Africa (**Menni_2**: 66.50%). In 2021, the worst F_1 scores for every country are: Brazil (**Akinbami_1**: 12.02%), Canada (**Akinbami_2**: 8.03%), Israel (**Akinbami_1**: 10.60%), Japan (**Akinbami_2**: 9.10%), Turkey (**Akinbami_2**: 11.80%), and South Africa (**Akinbami_2**: 13.61%). Fig SM1 in the supplemental material B shows the F_1 score yielded by each detection method across the six countries for 2020 and 2021. As can be seen in this figure, detection methods generally

are better for Brazil, Turkey, and South Africa compared to those yielded by Canada, Israel, and Japan.

It is worth noting that in Table 1, the TPR values exhibited by Brazil, Turkey, and South Africa are at least two-fold those shown by Canada, Israel, and Japan. Since the F_1 score is highly affected by imbalanced classes [29], we also evaluate the performance of the various detection methods for three groups: the broad set of the six countries, the set of countries with high TPR (Brazil, Turkey, and South Africa), and the countries with low TPR (Canada, Israel, and Japan). Table 2 displays the average of the F_1 score for the overall five countries (overall), for the countries with high TPR (High TPR), and for the countries with low TPR (Low TPR) for 2020, 2021, and the entire interval 2020-2021. As can be observed, countries with low TPR exhibit lower F_1 scores than countries with high TPR: (a) 2020 ($\beta = -2.32$, $p < 0.05$), and (b) 2021 ($\beta = -2.06$, $p < 0.05$). The detection techniques generating the best F_1 scores for the overall six countries are 2020 (**Astley**: 60.49%), 2021 (**Mika**: 58.35%), 2020-2021 (**Mika**: 59.33%). The methods that yield the best F_1 scores for the countries with low TPR are 2020 (**Smith**: 53.67%), 2021 (**Smith**: 53.25%), and 2020-2021 (**Smith**: 53.46%). Finally, the methods with the best performance according to the F_1 score for the countries with high TPR are 2020 (**Astley**: 69.34%), 2021 (**Menni_2**: 65.25%), and 2020-2021 (**Astley**: 66.95%).

Radar charts of sensitivity, specificity, and precision in % for the different detection methods are shown in Fig 2. Radar charts are presented for each country and for 2020 and 2021. Among the most relevant things to highlight from the radar figures, it can be observed that there is no method that is simultaneously better in all three metrics. On the other hand, the precision values are much better than those obtained with sensitivity and specificity. In the supplementary material B, Tables SM2, SM3, and SM4 show the averages and the CI for 2020 for sensitivity, specificity, and precision, respectively. In addition, the averages and CI for sensitivity, specificity, and precision for 2021 are displayed in the supplementary material B, in Tables SM6, SM7, and SM8, respectively. Note that blue lines, orange lines, and green lines correspond to sensitivity, specificity, and precision, respectively.

Specifically, the highest values in terms of sensitivity for each country and 2020 are: Brazil (**Zoabi_65**: 90.68%), Canada (**CDC**: 88.44%), Israel (**CDC**: 85.84%), Japan (**CDC**: 86.99%), Turkey (**CDC**: 89.11%), and South Africa (**CDC**: 88.57%). In addition, the lowest sensitivity values for each country and 2020 are: Brazil (**Akinbami_1**: 6.94%), Canada (**Akinbami_2**: 5.07%), Israel (**Akinbami_2**: 5.11%), Japan (**Akinbami_2**: 7.21%), Turkey (**Akinbami_2**: 5.77%), and South Africa (**Akinbami_2**: 9.51%). For 2021, the highest sensitivity values for each country and 2021 are: Brazil (**Shoer**: 87.72%), Canada (**CDC**: 88.39%), Israel (**CDC**: 85.05%), Japan (**Bhattacharya**: 88.79%), Turkey (**CDC**: 87.19%), and South Africa (**Astley**: 89.09%). Furthermore, the lowest sensitivity values for each country and 2021 are: Brazil (**Akinbami_2**: 6.49%), Canada (**Akinbami_2**: 4.29%), Israel (**Akinbami_1**: 5.82%), Japan (**Akinbami_2**:

Table 2

F_1 score (in %) and its 95% confidence interval for three different groups of countries: the overall five countries (overall), the countries with high TPR (High TPR: Brazil and South Africa), and the countries with low TPR (Low TPR: Canada, Germany, and Japan) for 2020, 2021, 2020-2021.

Method	2020			2021			2020-2021		
	Overall	Low TPR	High TPR	Overall	Low TPR	High TPR	Overall	Low TPR	High TPR
Menni_1	58.55	53.47	63.63	55.52	51.98	59.06	57.03	52.73	61.34
Menni_2	58.61	48.91	68.30	55.27	45.29	65.25	56.94	47.10	66.78
Roland	59.64	51.35	67.92	56.76	48.75	64.77	58.20	50.05	66.34
Smith	60.25	53.67	66.82	58.19	53.25	63.12	59.22	53.46	64.97
Zoabi_55	49.72	36.89	62.54	47.04	36.88	57.20	48.38	36.89	59.87
Zoabi_65	49.67	36.85	62.48	46.91	36.70	57.13	48.29	36.78	59.81
CDC	49.13	32.22	66.05	45.86	31.58	60.14	47.50	31.90	63.10
Shoer	60.44	52.64	68.23	55.86	46.73	64.99	58.15	49.69	66.61
Bhattacharya	59.72	51.36	68.08	57.66	51.27	64.06	58.69	51.32	66.07
WHO	26.02	25.35	26.68	28.68	29.33	28.04	27.35	27.34	27.36
Perez	51.50	43.47	59.53	50.96	45.23	56.68	51.23	44.35	58.11
Mika	60.30	52.96	67.64	58.35	52.22	64.48	59.33	52.59	66.06
Akinbami_1	12.83	11.64	14.01	12.48	11.05	13.91	12.65	11.35	13.96
Akinbami_2	12.47	10.72	14.21	11.02	9.54	12.51	11.75	10.13	13.36
Akinbami_3	23.99	20.29	27.69	23.97	20.94	27.01	23.98	20.62	27.35
Salomon	30.33	27.76	32.89	32.59	32.02	33.16	31.46	29.89	33.03
Astley	60.49	51.63	69.34	57.96	51.36	64.56	59.22	51.50	66.95

4.79%), Turkey (**Akinbami_2**: 6.36%), and South Africa (**Akinbami_2**: 7.41%). Observe that **CDC** produces the highest sensitivity values. Specifically, the average sensitivities obtained by **CDC** along countries in 2020 and 2021 are 87.84% in 2020 and 79.53%, respectively. In addition, **Akinbami_2** yields the lowest sensitivity results, with only 6.78% and 5.92% for 2020 and 2021. In contrast to the sensitivity metric, the **Akinbami_2** method produces the highest specificity values, whereas the **CDC** method yields the lowest specificity results. For 2020 and 2021, the specificity averages along countries obtained by **Akinbami_2** exceed 99 percent. Instead, the specificity averages along countries obtained by **CDC** are 42.81% and 54.85% in 2020 and 2021, respectively. Regarding precision, the **Akinbami_2** method also yields the highest values, while the **CDC** method also obtains the lowest results. In summary, in terms of sensitivity, **CDC** provides the best performance, whereas **Akinbami_2** produces the worst. Contrarily, in terms of specificity and precision, **Akinbami_2** yields the best performance, while **CDC** generates the worst. For this study, we use the F_1 -score to rank the detection methods under test. Notice that this metric offers a better trade-off between sensitivity and specificity.

In terms of the F_1 -score, the most efficient methods by category are Smith for rule-based methods, Menni for regression-based methods, and Astley for tree-based ML models. The Smith method developed a clinical prediction rule to diagnose COVID-19 in symptomatic patients [10]. A multivariable logistic regression method was used to identify independent predictors of COVID-19 and estimate a weighted prediction rule. In this case, the multivariate logistic regression approach was decisive in the development of the diagnostic rule. In the context of logistic regression techniques, the Menni method exhibits high performance

because the detection engine is built using a large dataset collected from many users (more than 2.5 million participants) who reported COVID-19 symptoms and a COVID-19 antigen test result [9]. Notice that the developed model combines these symptoms to predict infected people. For the development of the model, the Menni method used a stepwise modeling process with logistic regression. This was done to identify the symptoms strongly correlated with COVID-19 active cases, for which multiple adjustments were made. Due to the combination of the dataset and the logistic regression modeling process, the model yields good performance. Finally, the work of Astley et al. uses the UMD-CTIS, which is the largest global health survey to date [15]. This work performed deep data analysis of these data (multivariate analysis, Pearson, and Spearman correlations, etc.). Then, this method built a prediction model using the machine learning technique based on decision trees known as Light Gradient Boosting Machine (LightGBM). For the construction of the prediction model, the Astley method used different data samples grouped according to different criteria (region, age, etc.) to predict COVID-19 trends. The high performance of this approach is due to the combination of different samples to build the model using LightGBM. In summary, in terms of the F_1 -score, both the Smith method and the Astley method yield similar performances, while the Menni method has worse quality.

3.2. Explainability Analysis

For the explainability analysis, we focus on three methods: Smith for rule-based methods, Menni for regression-based methods, and Astley for tree-based models. The methods chosen were those that gave the best results in each category.

In particular, the Smith method defines a prediction rule to identify COVID-19 positives in symptomatic individuals

Consistent Comparison of Symptom-based Methods for COVID-19 Infection Detection

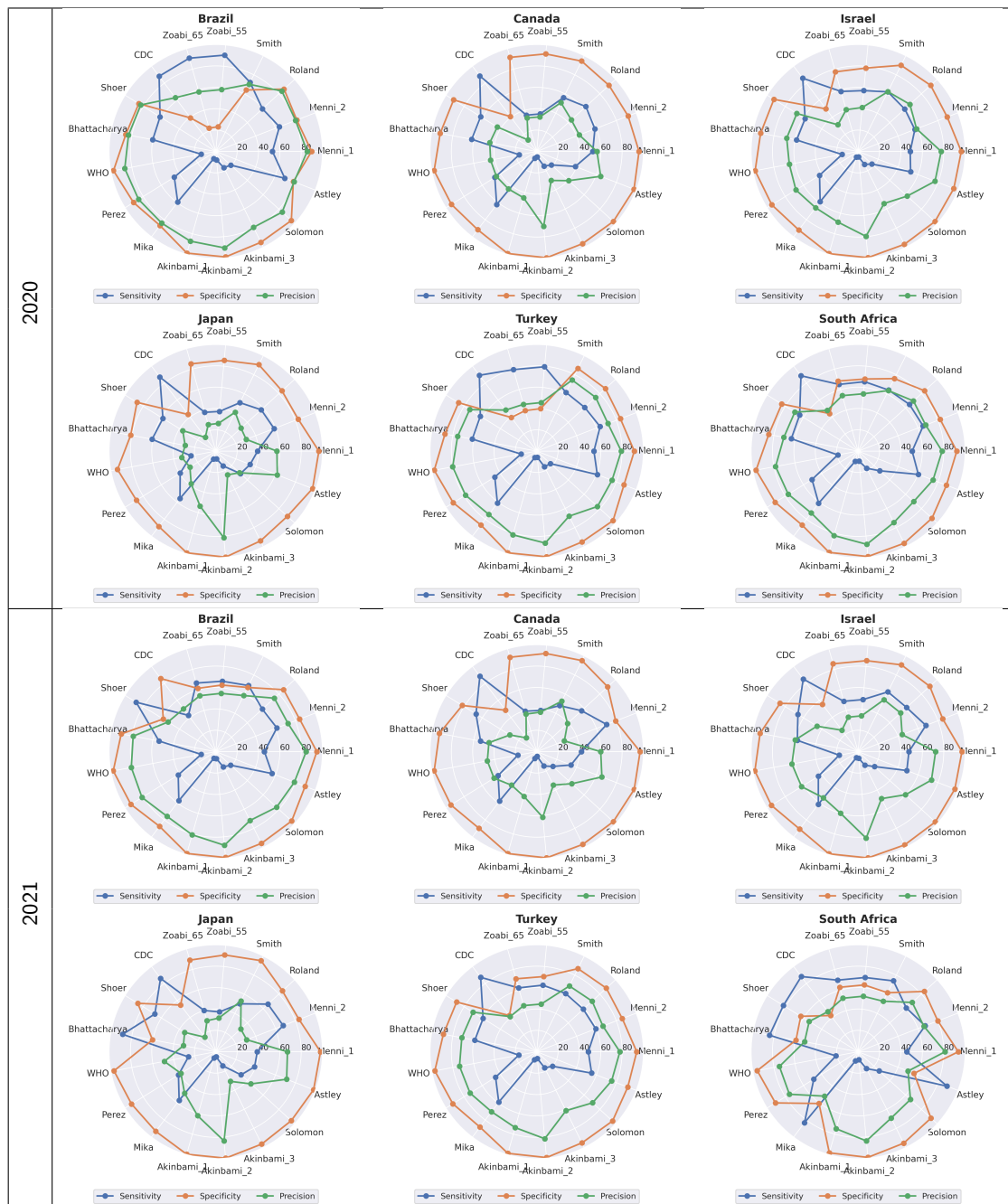


Figure 2: Radar chart of sensitivity (blue circles), specificity (orange circles), and precision (green circles) in % exhibited by the various methods for the entire set of countries and for 2020 and 2021. The closer the distance to the center, the worse the performance of the corresponding method.

using the following symptoms and their respective weights: *loss of smell/taste* (2), *fever and cough* (1) and *chest pain* (-1). Thus, *odor/taste loss* has a higher weight, while *chest pain* has a negative score because they consider it to be caused by another virus. In the case of Menni, the variables considered by the best logistic regression model are *age* (0.01), *gender* (0.44), *odor/taste loss* (1.75), *cough* (0.31), *fatigue* (0.49) and *skipped meals* (0.39). We see that in Menni, the one with the greatest weight/relevance is *loss of smell/taste* and then *fatigue*.

In the case of Astley, they used a LightGBM technique, so we can use the ranking of feature importance given by this technique for explainability analysis. In this case, Figures 3 and 4 show the most relevant variables for the six countries and for 2020 and 2021. For this case, there is no common most relevant variable for all cases, or in one year, or even for the same country for different years. That made us create a table to establish the 5 most relevant characteristics provided by this model for each year for all countries (see Table 3).

Among the most relevant things of Table 3 and Figs 3 and 4 is that there are variables with a very different

Table 3
Most relevant characteristics in the Astley method

Variables	2020	2021
Cough	0.074	0.071
Stuffy or runny nose	0.084	0.082
Aches or muscle pain	0.077	0.078
Headache	0.076	0.073
Sore Throat	0.077	0.073
Fever	0.063	0.073

behavior between countries (for example, *Fever*), sometimes being among the most relevant and in others with very little relevance. Also, there are two variables that are consistently among the most relevant which are *Stuffy or runny nose* and *Aches or muscle pain*. There are some variables that sometimes appear on the list and then never appear, such as *nausea*, or that appear rarely in the top 5 list but always appear as *Difficulty breathing*.

Regarding the symptoms by country, the same order of relevance is different between countries for the same year, but many of the most relevant variables coincide in some cases (for example, see the first 5 most relevant characteristics between Canada and Israel). Nor do the 5 most relevant characteristics for the same country coincide between different years, although almost always for all countries their 5 most relevant characteristics are very similar for each year, although in a different order (for example, see Canada and Turkey).

Table 4 summarizes information about methods under test. In that table, *low* refers to a variable with poor importance, *high* denotes significant importance, and so on for the rest. These labels are determined by the weight/importance the method assigns to the variable. We can see that *Loss of smell/taste* and *Cough* are common symptoms for the methods, although in some cases they appear with low importance. We can also see that the most relevant characteristics are very different between the methods, but *Cough* coincide among the most important of all of them. In general, we can observe that *Loss of smell/taste*, *Fever*, *Cough*, *Chest pain* and *Fatigue* appear in at least two methods.

Regarding the methods, the main individual features considered by these methods are (a) Smith: *Loss of taste and smell* (b) Menni: *Loss of smell and taste*, and (c) Astley: *Stuffy or runny nose* and *Aches or muscle pain*. Thus, there is also no complete coincidence between the methods. Also, we observe that the number of symptoms reported with the Astley method is very large. LightGBM gives a lot of information about the relevance of the characteristics, even by country, which allows a better decision-making process considering the specific relevance in each context. Thus, it allows a detailed analysis by country and year. We can also see that there is not a great common characteristic/symptom between the models, but that is highly variable, which is also the case for LightGBM when the analysis is done by country and year.

Table 4
Methods vs used variables

Used variables vs Methods	LightGBM	Smith	Menni
Gender			Normal
Age			Low
Stuffy or runny nose	High		
Loss of smell/taste	Low	High	High
Fever	Normal	Normal	
Cough	Normal	Normal	Normal
Chest pain	Low	Normal	
Fatigue	Low		Normal
Skipped meals			Normal
Aches or muscle pain	High		
Headache	Normal		
Sore Throat	Normal		

If we consider the explainability allowed by year and/or country as the main criterion for comparing the explainability analysis that the methods studied in the work allow, the best technique is LightGBM. So, a great conclusion of this section is that methods like LightGBM allow better explainability, being able to be used to give more details and better reason for the decisions. The other methods are more general and are more difficult to consider if it is necessary to reason a decision in a specific context.

4. Discussion

First, it is worth noticing that the TPR of the study population is a parameter to be considered to evaluate the performance of the various detection methods. More precisely, the TPR affects performance metrics such as the F_1 score and precision that assess the performance of the detection method for the positive class. In essence, prediction rules will likely detect more active cases and therefore will exhibit larger F_1 scores and precision values, when the TPR of the dataset is high. For example, as can be seen in Tables SM1, SM5, and 2, the F_1 scores generated by the different detection techniques for the countries with high TPR are, in general, larger than scores obtained for the countries with low TPR. Indeed, when assuming that all cases are positive, the F_1 scores yielded for the countries with high TPR are at least two times larger than those obtained for the countries with low TPR. Similarly, as can be observed in Tables SM4 and SM8 in the Supplementary Material B, the precision values outputted by different methods for the countries with high TPR are larger than those obtained for the countries with low TPR. Hence, this comparative study considers the TPR of every dataset as a source of bias that can introduce confounding.

One may compare the performance of various methods and select the best model for detecting COVID-19 active cases. Nevertheless, as can be seen in Tables SM1, SM5, and 2, none of the methods achieve an F_1 score above 75% indicating that no model has a good enough performance. Although no single method exhibits outstanding performance, we attempt to extract the techniques showing the

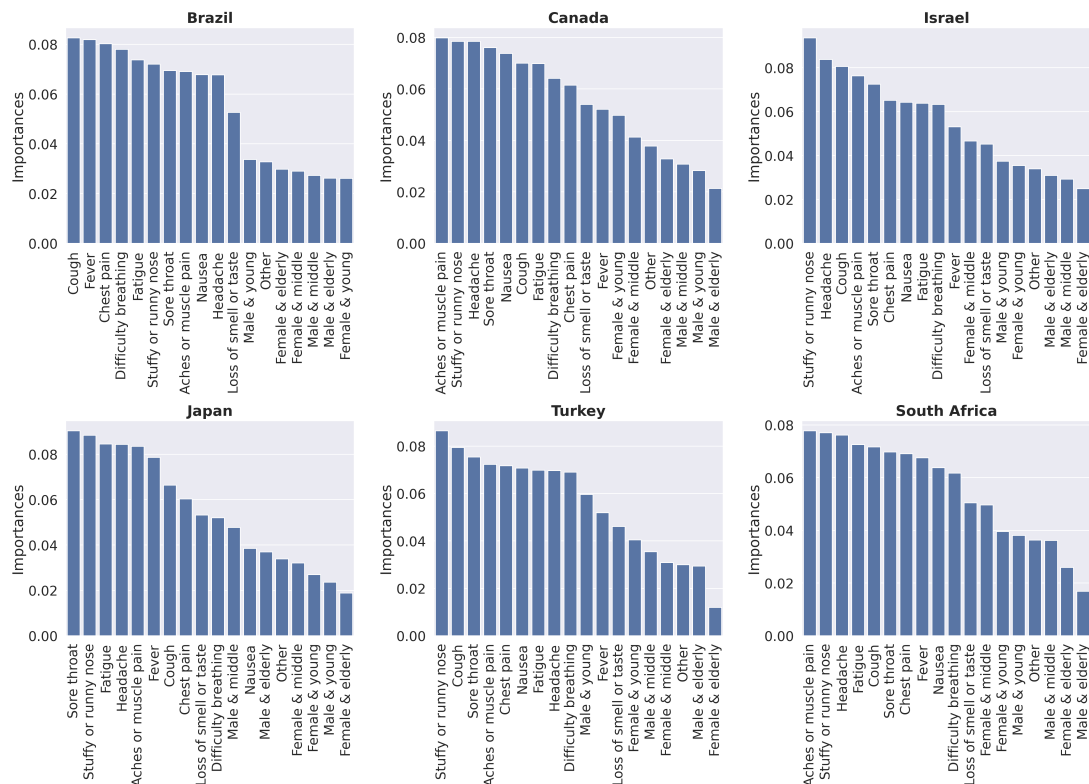


Figure 3: Feature importances of the Astley method for 2020 and for the entire set of countries.

best indicators among the considered metrics. Notice that the knowledge of the TPR influences the selection of the best detection method. For example, if the TPR is unknown, the Smith method provides the best performance (Table 2, Overall and 2020-2021: 56.59%). Instead, if the TPR is known, the best performances are provided by Menni_1 and Astley methods for low TPR (Table 2, 2020-2021: 51.67%) and high TPR (Table 2, 2020-2021: 67.64%), respectively.

For 2020, when there was no vaccination yet, the best detection methods are Mika (Table 2, Overall, 2020: 58.47%), Menni_1 (Table 2, Low TPR, 2020: 53.77%), and Astley (Table 2, High TPR, 2020: 70.28%). In particular, the Mika method detects a COVID-19 active case by considering fever, cough, loss of taste and smell, and gastrointestinal problems. As can be seen, positive cases have a strong association with loss of smell and taste, cough, and fever for 2020. On the other hand, the best methods for 2021 (when vaccination started and new variants have appeared) are Smith (Table 2, Overall, 2021: 54.99%), Smith (Table 2, Low TPR, 2021: 49.98%), and Shoer (Table 2 High TPR, 2021: 65.39%). Notice that the Shoer method considers individual features such as age, gender, prior medical conditions, and self-reported symptoms. It is important to note that both F_1 scores and precision values are lower for 2021 than those obtained for 2020. In 2021, new variants of COVID-19 appeared and the intensity of symptoms in vaccinated people was reduced. Therefore, the effectiveness of the methods under test is affected by the presence of new variants and the exponential increase in the number of vaccinated people.

Consequently, for the overall period 2020-2021, we can choose Smith, Astley, Menni_1, Mika, and Shoer methods as the best detection techniques under the F_1 score criterion.

A comprehensive review of the benefits and limitations of UMD-CTIS data has been reported in [19, 15]. Note that the countries were selected based on geographical diversity and the availability of sufficient samples per country. Further, the countries were selected to observe the response of different detection methods in countries with low and high TPR. A limitation of the dataset for this study lies in the fact that UMD-CTIS data do not collect all the symptoms needed by the different methods under test. For example, the Zoabi method uses one feature indicating ages older than 60, and the UMD-CTIS records ages older than 55 and 65. To overcome this drawback, we evaluate two versions of the Zoabi method, one version using a feature with ages older than 55 (**Zoabi_55**) and the other version using a feature with ages older than 65 (**Zoabi_65**). In a similar manner, we evaluate two versions of the Menni method and three versions of the Akimbami method (see Appendix A). These and other limitations will be studied in future works.

Also, in future work, a selection of different machine learning techniques will be made for the use of the different variables included in the CTIS database, which are not present in these studied methods. The main goal of future work is to attempt to improve the methods studied in this "Consistent Comparison of Symptom-based Methods for COVID-19 Infection Detection" by improving the F_1 score and presenting the ROC curves for each model. In addition,

Consistent Comparison of Symptom-based Methods for COVID-19 Infection Detection

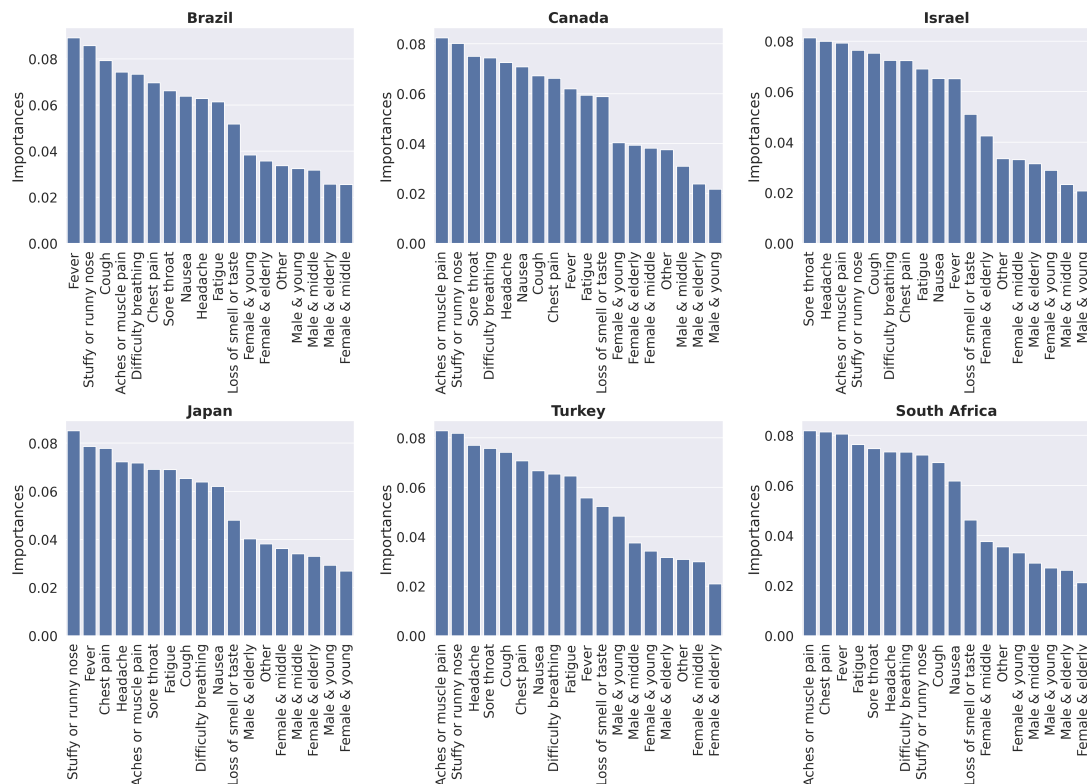


Figure 4: Feature importances of the Astley method for 2021 and for the entire set of countries.

a study of the most important variables based on the models obtained previously will be carried out for the same countries as in this report: Brazil, Canada, Israel, Japan, Turkey, and South Africa.

5. Summary table

What was already known on the topic?

- Several COVID-19 detection methods based on information collected from patients have been proposed during the global pandemic crisis.
- Normally, these methods have been developed and evaluated using specific datasets.

What does this study add to our knowledge?

- This paper provides a solid and consistent comparison among multiple COVID-19 detection methods using homogeneous data across six countries and two years.
- This comparison is based on a wide variety of performance metrics and the explainability analysis of the different COVID-19 detection methods.

6. 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 ethics and privacy guidelines and regulations.

7. 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 micro-data 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).

References

- [1] R. Wölfel, V. M. Corman, W. Guggemos, M. Seilmaier, S. Zange, M. A. Müller, D. Niemeyer, T. C. Jones, P. Vollmar, C. Rothe, et al., Virological assessment of hospitalized patients with COVID-2019, *Nature* 581 (2020) 465–469.

- [2] S. Whitelaw, M. A. Mamas, E. Topol, H. G. C. Van Spall, Applications of digital technology in COVID-19 pandemic planning and response, *The Lancet Digital Health* 2 (2020) e435–e440.
- [3] M. P. Cheng, J. Papenburg, M. Desjardins, S. Kanjilal, C. Quach, M. Libman, S. Dittrich, C. P. Yansouni, Diagnostic testing for severe acute respiratory syndrome-related coronavirus 2: a narrative review, *Annals of internal medicine* 172 (2020) 726–734.
- [4] H. Tian, Y. Liu, Y. Li, C.-H. Wu, B. Chen, M. U. G. Kraemer, B. Li, J. Cai, B. Xu, Q. Yang, B. Wang, P. Yang, Y. Cui, Y. Song, P. Zheng, Q. Wang, O. N. Bjornstad, R. Yang, B. T. Grenfell, O. G. Pybus, C. Dye, An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China, *Science* 368 (2020) 638–642.
- [5] 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.
- [6] 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).
- [7] 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.
- [8] A. Maharaj, J. Parker, J. Hopkins, E. Gournis, I. Bogoch, B. Rader, C. Astley, N. Ivers, J. Hawkins, L. Lee, A. Tuite, D. Fisman, J. Brownstein, L. Lapointe-Shaw, Anticipating the curve: can online symptom-based data reflect COVID-19 case activity in Ontario, Canada?, *medRxiv* (2021).
- [9] 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.
- [10] 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.
- [11] 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, *International Forum of Allergy & Rhinology* 10 (2020) 832–838.
- [12] 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).
- [13] 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.
- [14] S. Shoer, T. Karady, A. Keshet, S. Shilo, H. Rossman, A. Gavrieli, T. Meir, A. Lavon, D. Kolobkov, I. Kalka, et al., A prediction model to prioritize individuals for a SARS-CoV-2 test built from national symptom surveys, *Med* 2 (2021) 196–208.
- [15] C. Astley, G. Tuli, K. Mc Cord, E. Cohn, B. Rader, T. Varrelman, S. Chiu, X. Deng, K. Stewart, T. 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).
- [16] 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.
- [17] J. Álvarez, C. Baquero, E. Cabana, J. P. Champati, A. F. Anta, D. Frey, A. Garcia-Agundez, C. Georgiou, et al., Estimating Active Cases of COVID-19, *medRxiv* (2021).
- [18] J. Fan, Y. Li, K. Stewart, A. R. Kommareddy, A. Bradford, S. Chiu, F. Kreuter, N. Barkay, et al., Covid-19 world symptom survey data api., <https://covidmap.umd.edu/api.html>, 2020.
- [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 Facebook on a university-based rapid turn-around global survey, *Survey Research Methods: SRM* 14 (2020) 159–163.
- [20] C. for Disease Control, Prevention, Coronavirus Disease 2019 (COVID-19) 2020 Interim Case Definition, Approved April 5, 2020, National Notifiable Diseases Surveillance System (NNDSS) (2020).
- [21] 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).
- [22] N. Yalçın, S. Ünal, Symptom based covid-19 prediction using machine learning and deep learning algorithms, *Journal of Emerging Computer Technologies* 2 (2022) 22 – 29.
- [23] A. Sedik, A. Iliyasa, B. Abd El-Rahiem, M. Abdel Samea, A. Abdel-Raheem, M. Hammad, J. Peng, F. Abd El-Samie, A. Abd El-Latif, Deploying machine and deep learning models for efficient data-augmented detection of covid-19 infections, *Viruses* 12 (2020).
- [24] The University of Maryland Social Data Science Center, The University of Maryland Social Data Science Center Global COVID-19 Trends and Impact Survey in partnership with Facebook, <https://covidmap.umd.edu/>, 2021. Accessed: 2022-01-10.
- [25] F. Cabitza, A. Campagner, The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical ai studies, *International Journal of Medical Informatics* 153 (2021) 104510.
- [26] F. Ali, S. Ahmed, Z. N. K. Swati, S. Akbar, Dp-binder: machine learning model for prediction of dna-binding proteins by fusing evolutionary and physicochemical information, *Journal of Computer-Aided Molecular Design* 33 (2019) 645–658.
- [27] F. Ali, S. Akbar, A. Ghulam, Z. A. Maher, A. Unar, D. B. Talpur, Afp-cmbpred: Computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information, *Computers in Biology and Medicine* 139 (2021) 105006.
- [28] S. Akbar, M. Hayat, M. Tahir, S. Khan, F. K. Alarfaj, cacp-deepgram: classification of anticancer peptides via deep neural network and skip-gram-based word embedding model, *Artificial Intelligence in Medicine* 131 (2022) 102349.
- [29] H. He, Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*, Wiley-IEEE Press, 1st edition, 2013.
- [30] M. Pollán, B. Pérez-Gómez, R. Pastor-Barriuso, J. Oteo, M. A. Hernán, M. Pérez-Olmeda, J. L. Sanmartín, A. Fernández-García, et al., Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study, *The Lancet* 396 (2020) 535–544.

Supplemental Materials: Consistent Comparison of Symptom-based Methods for COVID-19 Infection Detection

A. Materials and Methods

A.1. Estimation Detection Methods

Detection methods were grouped into three types. More precisely, the rules-based methods give weights to the symptoms using some criteria (physician, from some organization, etc.) to identify positive cases. In addition, machine-learning-based techniques build tree-based classifiers to detect infected people from datasets containing information on symptoms. Finally, the regression-based approaches build prediction models using logistic regression techniques.

A.1.1. Rule-based methods

The rule-based methods used in this work are:

- Smith

This work builds a clinical prediction rule to identify COVID-19 active cases in symptomatic individuals [10]. To this end, this method implemented a multivariable logistic regression analysis to identify the independent predictors of COVID-19 active cases. Specifically, the Smith method selects a few symptoms associated with positive cases and assigns them different coefficients: loss of smell/loss of taste (2), fever and cough (1), and chest pain (-1). The chest pain variable has a negative score because this symptom being likely caused by another virus. The dataset used was obtained from a standardized clinical questionnaire that was administered to patients before applying the RT-PCR test. Moreover, the performance of the Smith method was tested using a dataset with 120 SARS-CoV-2-positive cases and 120 SARS-CoV-2-negative cases for training, and 40 cases for validation of the classification model.

- Centers for Disease Control and Prevention (CDC)

In August 2020, the Centers for Disease Control and Prevention approved the COVID-like illness (CLI) metric [20]. This metric declares a COVID-19 positive case if an individual presents at least two of the following symptoms: fever, chills, rigors, myalgia (muscle aches and pain), headache, sore throat. This metric also identifies an active case if the individual has at least one of the following symptoms: cough, shortness of breath, difficulty breathing, loss of smell, or loss of taste. Notice that the UMD-CTIS survey does not register myalgia and rigors. Hence, we estimate the CLI metric without those symptoms.

- WHO

Under the context of continuous monitoring of the COVID-19 pandemic, the World Health Organization (WHO) published a COVID-like illness¹ metric that

declares a potential active case when individuals report the following symptoms: fever, cough, and fatigue [16]. This metric was used in [17] to estimate COVID-19 active cases in various countries such as Spain, Peru, Ecuador, UK, Greece, and India considering the information extracted from the UMD-CTIS.

- Akinbami

The Akinbami method uses a combination of three symptoms to declare a COVID-19 positive case [7]. To this end, the study built three classification models depending on the combination of the symptoms: 1) **Akinbami_1** which uses fever, shortness of breath, and chills, 2) **Akinbami_2** which uses fever, shortness of breath, and anosmia/ageusia, and 3) **Akinbami_3** which uses fever, shortness of breath, and headache. The dataset was provided by a serologic survey collected in Detroit and New York from May 17 to July 2, 2020. The extracted dataset contains 40,938 tested individuals of which 6,631 are positive.

- Salomon

This method defines a metric also referred to as COVID-like illness (CLI)² different from those specified in [20] and [17]. Specifically, this metric identifies a positive case if the participant reports fever, cough, or shortness of breath/difficulty breathing [21, 18, 17]. This method was evaluated on datasets obtained from the CMU-CTIS survey that was collected in the United States from April 2020 to April 2021 by the Delphi Group at the Carnegie Mellon University (CMU) and the CTIS in partnership with Facebook. Like the UMD-CTIS project, the CMU-CTIS survey also obtained information about individual features (such as age groups, gender, testing, and vaccination) and COVID-19 symptoms.

A.1.2. Machine Learning methods

The machine learning techniques used in this work are:

- Astley

This study focused on building COVID-19 diagnostic models based on machine learning techniques [15]. More precisely, this approach selected the Light Gradient Boosting Machine (LightGBM) engine to build the COVID-19 diagnostic models considering two individual features (age groups and gender) and twelve symptoms. Furthermore, this method extracted the datasets from the UMD-CTIS collected in 114 countries from April to December 2020. Note that the training set used to build the classification models for each country is a subset of rows derived from questionnaires reporting a laboratory test.

- Zoabi

The Zoabi method considers eight features: gender, age (≥ 60), cough, fever, sore throat, shortness of

¹Aka UMD CLI WHO in [17]

²Aka CLI in [21, 18] and UMD CLI in [17]

breath, headache, and known contact with an individual COVID-19 confirmed positive case. It builds a classification model based on a machine learning approach [5]. In essence, this method builds a gradient-boosting classification model with decision-tree base-learners. This approach trained and evaluated the classification model from data released by the Israeli Ministry of Health. This dataset contains information on individuals with RT-PCR tests. Specifically, the training set consists of 51,831 individuals of whom 4,769 are confirmed positive cases. On the other hand, the test set consists of 47,401 individuals of whom 3,624 are positive cases. In this case, it is worth noting that UMD-CTIS data ranges of ages does not have a boundary at 60. The boundary is either at 55 or 65. Hence, we have created 2 different models, one for each range of age labeled **Zoabi_55** and **Zoabi_65**, to go around this difference in the data.

A.1.3. Regression-based methods

The regression-based methods used in this work are:

- **Menni**

This method performs logistic regressions to build the optimal COVID-19 classification model for a set of individual features such as age, gender, loss of smell and taste, cough, fatigue, and loss of appetite [9]. The building and evaluation of the classification model used a dataset extracted from a symptom tracker based on a smartphone app launched in the United Kingdom and the United States in March 2020. Specifically, this method was evaluated over responses from 2, 618, 862 participants voluntarily recording their symptoms. In this study, the best classification model according to the Akaike information criterion (AIC) was described as

$$x = -1.32 - (0.01 \times \text{age}) + (0.44 \times \text{gender}) \\ + (1.75 \times \text{loss of smell and taste}) \\ + (0.31 \times \text{cough}) + (0.49 \times \text{fatigue}) \\ + (0.39 \times \text{skipped meals}),$$

where symptoms represent binary variables. More precisely, every feature coded as 1 reports the presence of the symptom, while each variable coded as 0 indicates its absence. The gender variable also stands for a binary variable, the one-value indicates a male, and the zero-value represents a female. Afterward, this method identified a COVID-19 active case whether $\frac{e^x}{1+e^x} \geq 0.5$. It is worth noting that UMD-CTIS data did not register the skipped meal variable. Therefore, we modified the Menni method by computing the x score with the skipped meals variable fixed to zero. This approach is labeled as **Menni_1**. Furthermore, we followed the procedure reported in [9] to build the logistic regression model from individual features available in our dataset (**Menni_2**). In other words,

we built a logistic regression model that considers the features: age, gender, loss of smell and taste, cough, and fatigue.

- **Roland**

This study performs logistic regression analysis to build a classification model based on five symptoms: loss of taste and smell, body aches, fever or chills, shortness of breath, and sore throat [11]. This method uses a dataset extracted from an anonymous electronic survey publicized with ads on social networks (Facebook, Twitter, Reddit, and Nextdoor) from March 31 to April 10, 2020. Specifically, the Roland method was evaluated from a dataset provided by 620 participants of whom 339 reported COVID test outcomes. This work built a stepwise logistic model whose training set was obtained by randomly extracting 25% of the rows belonging to the COVID-tested individuals. The remaining rows were used to examine the performance of the classification model.

- **Mika**

This method is similar to the Roland method. In other words, the Mika method fits a logistic regression model with the following symptoms: fever $> 38^\circ\text{C}$, cough, loss of taste and smell, and gastro-intestinal (GI) [13]. The data set consisted of 3114 participants of which 778 were tested positive. The UMD-CTIS survey does not have a question on GI symptoms. Therefore, we use the answer for the presence of nausea instead, as it is the closest related symptom.

- **Shoer**

This research builds two models based on logistic regression analysis to estimate the probability of individuals testing positive for SARS-CoV-2 [14]. In particular, this study obtained the datasets from two surveys launched in Israel in 2020. On one hand, the online survey registered various individual features such as age, gender, prior medical conditions, and self-reported symptoms. On the other hand, the shortened survey captured the information by means of an interactive voice response (IVR) platform. Specifically, the IVR version collected information on variables such as age group, prior medical conditions, general feeling, and a shortened list of symptoms. To generate the first model, an integrated dataset is constructed from both the features collected by the online version and the reduced set of attributes acquired by the IVR version. The second model used the information provided by the online version only. The UMD-CTIS survey does not have questions on prior medical conditions and general feeling, and therefore, we do not include them in the model.

- **Bhattacharya**

The Bhattacharya method proposes a clinical symptom-based score [12] given by

$$\begin{aligned} \text{score} = & 41.7 \times \text{Fever}(> 100^\circ\text{F}) + (13.5 \times \text{Cough}) \\ & +(15.8 \times \text{Headache}) + (10 \times \text{Myalgia}) \\ & +(94.7 \times \text{Loss of smell}). \quad (\text{SM1}) \end{aligned}$$

If the score is greater than 41.7, then the individual is declared a COVID-19 positive case. The method was examined on responses registered in a clinical screening applied to individuals with suspicion of having COVID-19. The number of participants in this study was 378 of which 125 individuals reported a positive COVID test result.

- Perez

This method builds a classifier based on logistic regression that considers the following symptoms: anosmia (loss of smell), ageusia (loss of taste), shortness of breath, digestive symptoms, fever, tiredness, sore throat absence, headache, and cough [6]. Then, the Perez method defined different risk scores and assigned them to four symptoms: severe tiredness (1), absence of sore throat (1), fever (2), and anosmia/ageusia (5). This approach declares an individual COVID-19 positive case whether the number of symptoms present is at least 4 and the cumulative score is at least 3. This study considers the data of the seroepidemiological study performed in Spain from April to June 2020. More precisely, more than 61000 participants nationwide completed a questionnaire on symptoms along with SARS-CoV-2 antibodies assays [30]. Notice that the number of positive cases are 2669, out of which 781 (approx 30%) are asymptomatic. The UMD-CTIS survey does not have a question on digestive symptoms. Furthermore, we consider that *severe tiredness* is equivalent to *fatigue*, and *shortness of breath* is equivalent to *difficulty in breathing*.

B. Results

The following tables present the quality of the methods for the different studied metrics.

Table SM1

F₁ score and its 95% confidence interval for the selected countries for 2020, in %.

Method	Brazil	Canada	Israel	Japan	Turkey	South Africa
Menni_1	65.56 (65.48 - 65.64)	54.33 (53.66 - 54.99)	59.76 (59.16 - 60.36)	46.33 (45.33 - 47.33)	63.93 (63.68 - 64.17)	61.39 (61.07 - 61.70)
Menni_2	71.13 (71.01 - 71.24)	49.33(48.77 - 49.88)	57.50 (57.04 - 57.97)	39.91 (39.27 - 40.54)	67.41 (67.21 - 67.60)	66.36 (66.10 - 66.62)
Roland	69.38 (69.30 - 69.46)	51.44 (50.86 - 52.02)	61.93 (61.46 - 62.41)	40.68 (39.98 - 41.39)	67.06 (66.87 - 67.26)	67.32 (67.05 - 67.58)
Smith	71.11 (71.05 - 71.18)	53.43 (52.85 - 54.01)	62.47 (61.98 - 62.97)	45.12 (44.42 - 45.82)	67.30 (67.11 - 67.49)	62.06 (61.80 - 62.32)
Zoabi_55	70.71 (70.65 - 70.77)	32.96 (32.37 - 33.54)	47.76 (47.32 - 48.20)	29.95 (29.29 - 30.60)	57.86 (57.69 - 58.03)	59.05 (58.80 - 59.31)
Zoabi_65	70.73 (70.67 - 70.79)	32.86 (32.28 - 33.44)	47.79 (47.36 - 48.23)	29.91 (29.27 - 30.55)	57.72 (57.55 - 57.88)	59.00 (58.74 - 59.25)
CDC	73.42 (73.36 - 73.48)	23.43 (23.14 - 23.72)	45.84 (45.46 - 46.21)	27.38 (27.00 - 27.75)	62.60 (62.42 - 62.78)	62.13 (61.88 - 62.39)
Shoer	70.45 (70.39 - 70.52)	50.95 (50.37 - 51.54)	62.41 (61.93 - 62.89)	44.57 (43.86 - 45.28)	67.49 (67.30 - 67.69)	66.76 (66.52 - 67.00)
Bhattacharya	69.77 (69.70 - 69.83)	51.90 (51.31 - 52.50)	62.78 (62.30 - 63.26)	39.41 (38.84 - 39.97)	67.67 (67.48 - 67.87)	66.81 (66.52 - 67.10)
WHO	23.92 (23.83 - 24.01)	24.08 (23.45 - 24.70)	24.69 (24.15 - 25.24)	27.29 (26.52 - 28.06)	25.14 (24.90 - 25.38)	30.97 (30.59 - 31.35)
Perez	59.47 (59.39 - 59.55)	45.20 (44.56 - 45.83)	52.27 (51.71 - 52.82)	32.93 (32.23 - 33.64)	58.12 (57.89 - 58.35)	61.00 (60.70 - 61.30)
Mika	69.43 (69.37 - 69.49)	51.43 (50.86 - 52.01)	62.16 (61.68 - 62.63)	45.29 (44.65 - 45.94)	67.08 (66.89 - 67.28)	66.40 (66.13 - 66.68)
Akinbami_1	12.85 (12.77 - 12.94)	11.33 (10.72 - 11.93)	10.22 (9.82 - 10.62)	13.38 (12.58 - 14.18)	11.48 (11.26 - 11.70)	17.70 (17.34 - 18.07)
Akinbami_2	14.69 (14.60 - 14.78)	9.41 (8.89 - 9.92)	9.59 (9.16 - 10.01)	13.16 (12.35 - 13.98)	10.81 (10.60 - 11.03)	17.14 (16.80 - 17.49)
Akinbami_3	27.84 (27.73 - 27.94)	20.23 (19.66 - 20.81)	21.67 (21.14 - 22.19)	18.98 (18.22 - 19.73)	26.31 (26.05 - 26.56)	28.93 (28.57 - 29.29)
Salomon	30.97 (30.87 - 31.07)	25.52 (24.84 - 26.20)	27.12 (26.58 - 27.66)	30.64 (29.93 - 31.35)	28.36 (28.10 - 28.61)	39.35 (38.98 - 39.72)
Astley	73.72 (73.65 - 73.78)	48.29 (47.58 - 49.00)	62.47 (61.98 - 62.97)	44.13 (43.32 - 44.93)	67.45 (67.24 - 67.65)	66.85 (66.61 - 67.09)

Table SM2

Sensitivity and its 95% confidence interval for the selected countries for 2020, in %

Method	Brazil	Canada	Israel	Japan	Turkey	South Africa
Menni_1	53.11 (53.02 - 53.21)	52.53 (51.78 - 53.28)	48.66 (48.02 - 49.30)	39.15 (38.09 - 40.21)	53.60 (53.30 - 53.90)	50.47 (50.13 - 50.81)
Menni_2	63.94 (63.72 - 64.16)	58.46 (57.75 - 59.16)	56.81 (56.05 - 57.58)	58.57 (57.32 - 59.81)	63.70 (63.43 - 63.98)	65.12 (64.82 - 65.43)
Roland	59.21 (59.08 - 59.33)	62.28 (61.55 - 63.00)	58.96 (58.36 - 59.56)	57.59 (56.31 - 58.87)	60.82 (60.55 - 61.09)	64.96 (64.65 - 65.27)
Smith	72.10 (72.02 - 72.17)	56.26 (55.54 - 56.98)	62.24 (61.65 - 62.84)	50.75 (49.86 - 51.65)	61.52 (61.05 - 61.98)	63.97 (62.01 - 65.93)
Zoabi_55	90.43 (90.18 - 90.67)	35.24 (34.00 - 36.47)	57.15 (56.26 - 58.04)	37.36 (35.86 - 38.86)	79.44 (78.45 - 80.43)	65.39 (64.72 - 66.07)
Zoabi_65	90.68 (90.50 - 90.87)	35.28 (33.87 - 36.69)	58.32 (57.55 - 59.09)	37.82 (36.26 - 39.38)	79.49 (78.34 - 80.63)	65.04 (64.50 - 65.57)
CDC	88.09 (88.03 - 88.16)	88.44 (88.03 - 88.85)	85.84 (85.43 - 86.24)	86.99 (86.36 - 87.63)	89.11 (88.96 - 89.25)	88.57 (88.35 - 88.79)
Shoer	61.40 (61.30 - 61.50)	61.75 (61.05 - 62.46)	58.22 (57.67 - 58.76)	58.24 (57.27 - 59.20)	62.13 (61.86 - 62.41)	64.24 (63.93 - 64.56)
Bhattacharya	60.08 (60.00 - 60.16)	61.97 (61.24 - 62.70)	58.45 (57.91 - 58.99)	60.72 (59.92 - 61.53)	61.40 (61.13 - 61.67)	63.50 (63.18 - 63.82)
WHO	13.88 (13.82 - 13.94)	16.63 (16.15 - 17.11)	15.26 (14.88 - 15.65)	23.66 (22.98 - 24.35)	14.91 (14.75 - 15.08)	19.32 (19.04 - 19.59)
Perez	45.71 (45.62 - 45.80)	46.14 (45.40 - 46.87)	42.38 (41.81 - 42.95)	39.22 (38.37 - 40.07)	46.18 (45.92 - 46.44)	50.64 (50.31 - 50.98)
Mika	59.17 (59.09 - 59.25)	62.25 (61.52 - 62.98)	58.85 (58.30 - 59.39)	55.67 (54.80 - 56.54)	61.08 (60.82 - 61.35)	61.18 (60.85 - 61.51)
Akinbami_1	6.94 (6.89 - 6.99)	6.51 (6.15 - 6.87)	5.53 (5.30 - 5.76)	7.70 (7.22 - 8.18)	6.18 (6.05 - 6.30)	9.93 (9.71 - 10.16)
Akinbami_2	7.99 (7.94 - 8.05)	5.07 (4.78 - 5.35)	5.11 (4.88 - 5.35)	7.21 (6.74 - 7.68)	5.77 (5.65 - 5.89)	9.51 (9.31 - 9.72)
Akinbami_3	16.88 (16.81 - 16.95)	15.28 (14.80 - 15.76)	13.56 (13.20 - 13.92)	15.49 (14.83 - 16.14)	16.31 (16.13 - 16.49)	17.95 (17.70 - 18.20)
Salomon	18.98 (18.91 - 19.05)	18.70 (18.14 - 19.26)	17.40 (16.99 - 17.81)	31.32 (30.58 - 32.05)	17.38 (17.20 - 17.56)	27.36 (27.05 - 27.68)
Astley	69.34 (69.24 - 69.44)	38.82 (38.04 - 39.59)	52.58 (52.04 - 53.12)	34.51 (33.71 - 35.31)	60.87 (60.58 - 61.16)	60.34 (60.01 - 60.67)

Table SM3

Specificity and its 95% confidence interval for the selected countries for 2020, in %

Method	Brazil	Canada	Israel	Japan	Turkey	South Africa
Menni_1	89.77 (89.71 - 89.84)	95.87 (95.79 - 95.96)	96.37 (96.26 - 96.48)	96.59 (96.48 - 96.70)	91.44 (91.31 - 91.57)	92.17 (91.99 - 92.34)
Menni_2	81.41 (81.23 - 81.58)	91.83 (91.71 - 91.94)	89.23 (88.83 - 89.63)	82.83 (82.32 - 83.34)	84.09 (83.91 - 84.27)	82.23 (82.00 - 82.45)
Roland	86.54 (86.44 - 86.64)	91.68 (91.56 - 91.81)	91.71 (91.45 - 91.97)	83.91 (83.16 - 84.66)	87.08 (86.90 - 87.25)	84.01 (83.81 - 84.21)
Smith	64.18 (64.09 - 64.27)	94.37 (94.28 - 94.47)	90.02 (89.68 - 90.36)	90.56 (90.39 - 90.72)	86.63 (86.22 - 87.04)	75.98 (73.43 - 78.53)
Zoabi_55	23.21 (22.64 - 23.79)	91.56 (90.65 - 92.48)	78.31 (77.69 - 78.93)	85.44 (84.14 - 86.73)	40.05 (38.43 - 41.67)	67.92 (66.89 - 68.95)
Zoabi_65	22.73 (22.31 - 23.15)	91.46 (90.45 - 92.47)	77.39 (76.93 - 77.85)	85.05 (83.71 - 86.38)	39.53 (37.65 - 41.41)	68.31 (67.53 - 69.08)
CDC	39.32 (39.21 - 39.42)	40.87 (40.68 - 41.06)	49.90 (49.59 - 50.22)	43.06 (42.75 - 43.37)	39.57 (39.38 - 39.77)	44.13 (43.84 - 44.41)
Shoer	84.86 (84.78 - 84.95)	91.60 (91.46 - 91.75)	92.53 (92.35 - 92.71)	86.83 (86.39 - 87.27)	86.18 (86.02 - 86.34)	83.90 (83.67 - 84.14)
Bhattacharya	85.81 (85.74 - 85.88)	92.01 (91.88 - 92.13)	92.67 (92.52 - 92.83)	81.11 (80.89 - 81.33)	87.43 (87.28 - 87.58)	84.95 (84.73 - 85.16)
WHO	97.51 (97.48 - 97.54)	97.79 (97.74 - 97.85)	97.86 (97.78 - 97.94)	93.69 (93.55 - 93.84)	97.68 (97.62 - 97.73)	96.96 (96.88 - 97.05)
Perez	90.64 (90.58 - 90.70)	93.98 (93.88 - 94.08)	94.79 (94.66 - 94.92)	87.46 (87.26 - 87.66)	91.96 (91.84 - 92.08)	91.13 (90.96 - 91.30)
Mika	86.75 (86.68 - 86.82)	91.69 (91.57 - 91.82)	91.96 (91.81 - 92.11)	88.55 (88.37 - 88.74)	86.78 (86.64 - 86.93)	86.83 (86.57 - 87.10)
Akinbami_1	98.80 (98.78 - 98.82)	99.18 (99.14 - 99.21)	99.33 (99.29 - 99.38)	99.17 (99.11 - 99.22)	99.13 (99.09 - 99.16)	98.76 (98.69 - 98.82)
Akinbami_2	99.02 (99.00 - 99.04)	99.78 (99.75 - 99.80)	99.66 (99.62 - 99.69)	99.80 (99.77 - 99.83)	99.44 (99.42 - 99.47)	99.24 (99.19 - 99.29)
Akinbami_3	94.85 (94.81 - 94.89)	96.30 (96.21 - 96.40)	96.97 (96.87 - 97.06)	94.02 (93.88 - 94.17)	95.15 (95.06 - 95.23)	96.50 (96.41 - 96.59)
Salomon	95.80 (95.76 - 95.84)	97.14 (97.08 - 97.20)	97.17 (97.08 - 97.26)	90.72 (90.56 - 90.88)	96.74 (96.68 - 96.80)	93.36 (93.20 - 93.51)
Astley	77.93 (77.81 - 78.05)	97.75 (97.68 - 97.82)	95.86 (95.75 - 95.98)	97.26 (97.15 - 97.38)	87.67 (87.51 - 87.83)	88.50 (88.29 - 88.72)

Table SM4

Precision and its 95% confidence interval for the selected countries for 2020, in %

Method	Brazil	Canada	Israel	Japan	Turkey	South Africa
Menni_1	85.64 (85.55 - 85.73)	56.41 (55.64 - 57.18)	77.58 (76.92 - 78.24)	57.27 (56.13 - 58.42)	79.22 (78.94 - 79.49)	78.42 (77.97 - 78.87)
Menni_2	80.17 (80.06 - 80.29)	42.76 (42.18 - 43.34)	58.58 (57.76 - 59.39)	30.51 (29.86 - 31.17)	71.59 (71.33 - 71.86)	67.69 (67.33 - 68.05)
Roland	83.80 (83.70 - 83.89)	43.89 (43.30 - 44.48)	65.41 (64.70 - 66.12)	32.00 (31.04 - 32.96)	74.77 (74.51 - 75.03)	69.90 (69.54 - 70.26)
Smith	70.17 (70.09 - 70.26)	50.99 (50.35 - 51.64)	62.45 (61.61 - 63.29)	40.75 (40.05 - 41.46)	74.46 (73.99 - 74.93)	63.42 (61.33 - 65.51)
Zoabi_55	58.07 (57.94 - 58.19)	32.48 (31.33 - 33.63)	41.25 (40.69 - 41.82)	26.07 (25.15 - 26.99)	45.70 (45.27 - 46.13)	54.01 (53.53 - 54.49)
Zoabi_65	57.98 (57.86 - 58.09)	32.55 (31.35 - 33.76)	40.62 (40.12 - 41.12)	25.84 (24.92 - 26.75)	45.57 (45.09 - 46.05)	54.09 (53.68 - 54.50)
CDC	62.94 (62.86 - 63.02)	13.51 (13.33 - 13.70)	31.28 (30.96 - 31.61)	16.26 (16.01 - 16.51)	48.25 (48.06 - 48.45)	47.87 (47.59 - 48.15)
Shoer	82.66 (82.58 - 82.75)	43.48 (42.84 - 44.11)	67.35 (66.74 - 67.96)	36.36 (35.54 - 37.17)	73.90 (73.65 - 74.14)	69.54 (69.17 - 69.91)
Bhattacharya	83.19 (83.11 - 83.27)	44.74 (44.13 - 45.36)	67.88 (67.31 - 68.45)	29.23 (28.73 - 29.74)	75.40 (75.16 - 75.64)	70.53 (70.15 - 70.91)
WHO	86.76 (86.60 - 86.92)	43.98 (43.01 - 44.95)	65.37 (64.37 - 66.36)	32.43 (31.45 - 33.41)	80.16 (79.76 - 80.55)	78.42 (77.85 - 79.00)
Perez	85.11 (85.01 - 85.20)	44.42 (43.73 - 45.12)	68.34 (67.67 - 69.01)	28.49 (27.80 - 29.17)	78.42 (78.14 - 78.70)	76.76 (76.36 - 77.16)
Mika	84.00 (83.92 - 84.08)	43.91 (43.32 - 44.50)	65.93 (65.38 - 66.48)	38.29 (37.65 - 38.94)	74.42 (74.19 - 74.64)	72.68 (72.25 - 73.11)
Akinbami_1	87.17 (86.95 - 87.39)	45.02 (43.20 - 46.84)	68.68 (66.88 - 70.48)	53.66 (51.23 - 56.09)	81.61 (80.88 - 82.33)	82.30 (81.48 - 83.13)
Akinbami_2	90.55 (90.37 - 90.73)	70.32 (68.01 - 72.64)	79.68 (77.85 - 81.52)	81.58 (79.19 - 83.97)	86.57 (85.93 - 87.22)	87.77 (87.04 - 88.50)
Akinbami_3	79.35 (79.20 - 79.51)	30.27 (29.42 - 31.13)	54.35 (53.29 - 55.41)	24.78 (23.78 - 25.77)	68.12 (67.66 - 68.58)	74.82 (74.21 - 75.43)
Salomon	84.15 (84.00 - 84.30)	40.49 (39.57 - 41.41)	61.96 (61.09 - 62.82)	30.12 (29.35 - 30.89)	77.02 (76.65 - 77.40)	70.22 (69.67 - 70.76)
Astley	78.69 (78.59 - 78.79)	64.33 (63.51 - 65.16)	77.06 (76.49 - 77.63)	61.82 (60.67 - 62.97)	75.65 (75.39 - 75.91)	75.02 (74.63 - 75.41)

Table SM5

F₁ score and its 95% confidence interval for the selected countries for 2021, in %

Method	Brazil	Canada	Israel	Japan	Turkey	South Africa
Menni_1	59.24 (59.18 - 59.31)	49.38 (49.02 - 49.74)	57.31 (56.96 - 57.65)	49.24 (49.16 - 49.83)	59.65 (59.44 - 59.87)	58.28 (58.06 - 58.50)
Menni_2	66.54 (66.49 - 66.59)	39.82 (39.59 - 40.05)	53.46 (53.21 - 53.70)	42.60 (42.37 - 42.84)	62.71 (62.56 - 62.85)	66.50 (66.33 - 66.68)
Roland	65.76 (65.71 - 65.82)	46.28 (46.03 - 46.53)	57.16 (56.86 - 57.46)	42.82 (42.62 - 43.03)	64.13 (63.96 - 64.31)	64.41 (64.23 - 64.59)
Smith	63.37 (63.32 - 63.42)	50.28 (49.99 - 50.57)	58.00 (57.68 - 58.33)	51.48 (51.23 - 51.74)	64.38 (64.21 - 64.55)	61.62 (61.45 - 61.80)
Zoabi_55	59.83 (59.79 - 59.88)	37.31 (37.01 - 37.60)	39.63 (39.28 - 39.98)	33.71 (33.45 - 33.98)	52.14 (51.88 - 52.40)	59.62 (59.47 - 59.77)
Zoabi_65	59.78 (59.74 - 59.83)	37.10 (36.81 - 37.39)	39.64 (39.29 - 39.99)	33.36 (33.11 - 33.62)	52.06 (51.80 - 52.31)	59.54 (59.38 - 59.69)
CDC	63.22 (63.17 - 63.26)	27.41 (27.28 - 27.55)	38.78 (38.59 - 38.97)	28.54 (28.40 - 28.68)	55.96 (55.81 - 56.11)	61.25 (61.10 - 61.39)
Shoer	65.81 (65.76 - 65.87)	41.10 (40.84 - 41.36)	53.67 (53.37 - 53.97)	45.42 (45.07 - 45.78)	64.18 (64.01 - 64.35)	64.97 (64.80 - 65.15)
Bhattacharya	64.16 (64.11 - 64.22)	49.22 (48.96 - 49.49)	58.76 (58.48 - 59.03)	45.82 (45.59 - 46.05)	64.61 (64.44 - 64.78)	63.40 (63.22 - 63.59)
WHO	23.62 (23.56 - 23.68)	26.01 (25.66 - 26.35)	27.92 (27.59 - 28.24)	34.05 (33.74 - 34.37)	27.72 (27.49 - 27.94)	32.78 (32.58 - 32.98)
Perez	54.85 (54.79 - 54.90)	44.70 (44.40 - 45.00)	51.27 (50.93 - 51.61)	39.72 (39.45 - 40.00)	56.03 (55.86 - 56.21)	59.17 (58.98 - 59.35)
Mika	65.33 (65.28 - 65.38)	46.76 (46.40 - 47.12)	57.50 (57.22 - 57.79)	52.41 (51.73 - 53.09)	64.13 (63.96 - 64.31)	63.98 (63.81 - 64.15)
Akinbami_1	12.02 (11.96 - 12.07)	11.43 (11.17 - 11.70)	10.60 (10.33 - 10.88)	11.11 (10.82 - 11.39)	13.86 (13.69 - 14.03)	15.86 (15.66 - 16.06)
Akinbami_2	12.02 (12.05 - 12.16)	8.03 (7.79 - 8.27)	11.48 (11.20 - 11.75)	9.10 (8.83 - 9.31)	11.80 (11.64 - 11.96)	13.61 (13.44 - 13.79)
Akinbami_3	26.59 (26.00 - 26.11)	20.96 (20.64 - 21.27)	21.96 (21.62 - 22.30)	19.90 (19.63 - 20.17)	26.35 (26.12 - 26.58)	28.08 (27.85 - 28.31)
Salomon	30.15 (30.11 - 30.24)	28.06 (27.70 - 28.43)	30.72 (30.39 - 31.05)	37.27 (36.97 - 37.57)	31.31 (31.09 - 31.53)	38.03 (37.83 - 38.23)
Astley	65.95 (65.90 - 66.01)	45.07 (44.74 - 45.40)	58.62 (58.29 - 58.94)	50.39 (50.08 - 50.70)	63.67 (63.50 - 63.85)	64.06 (63.88 - 64.24)

Table SM6

Sensitivity and its 95% confidence interval for the selected countries for 2021, in %

Method	Brazil	Canada	Israel	Japan	Turkey	South Africa
Menni_1	45.56 (45.49 - 45.62)	41.96 (41.59 - 42.33)	47.46 (47.11 - 47.81)	38.87 (38.53 - 39.22)	48.34 (48.09 - 48.59)	45.44 (45.20 - 45.68)
Menni_2	61.38 (61.33 - 61.43)	70.30 (69.94 - 70.65)	67.92 (67.50 - 68.33)	67.74 (67.46 - 68.02)	59.35 (59.11 - 59.58)	67.11 (66.88 - 67.34)
Roland	59.02 (58.96 - 59.09)	56.82 (56.49 - 57.15)	61.26 (60.81 - 61.72)	66.10 (65.82 - 66.38)	59.02 (58.81 - 59.24)	60.86 (60.60 - 61.12)
Smith	69.00 (68.94 - 69.05)	48.06 (47.73 - 48.39)	62.43 (62.09 - 62.78)	50.07 (49.78 - 50.37)	60.80 (60.25 - 61.35)	74.04 (73.85 - 74.24)
Zoabi_55	66.13 (65.66 - 66.59)	38.77 (37.62 - 39.91)	48.89 (48.07 - 49.71)	37.39 (36.37 - 38.40)	62.51 (62.14 - 62.88)	69.70 (69.44 - 69.96)
Zoabi_65	66.66 (65.75 - 67.57)	39.34 (38.05 - 40.62)	48.95 (48.16 - 49.74)	40.18 (38.51 - 41.86)	62.07 (61.71 - 62.42)	69.61 (69.30 - 69.92)
CDC	42.56 (42.51 - 42.61)	88.39 (88.18 - 88.61)	85.05 (84.82 - 85.29)	85.90 (85.69 - 86.12)	87.19 (87.06 - 87.32)	88.09 (87.96 - 88.23)
Shoer	87.72 (87.68 - 87.76)	66.90 (66.53 - 67.27)	66.13 (65.70 - 66.57)	67.12 (66.67 - 67.57)	59.08 (58.85 - 59.30)	82.11 (81.92 - 82.31)
Bhattacharya	54.12 (54.06 - 54.18)	53.51 (53.17 - 53.85)	57.80 (57.44 - 58.16)	88.79 (88.73 - 88.85)	59.01 (58.79 - 59.22)	84.16 (84.03 - 84.29)
WHO	13.84 (13.80 - 13.88)	17.97 (17.71 - 18.24)	17.95 (17.71 - 18.19)	26.08 (25.81 - 26.36)	17.08 (16.92 - 17.24)	20.99 (20.84 - 21.15)
Perez	41.39 (41.33 - 41.44)	42.56 (42.24 - 42.89)	43.68 (43.35 - 44.02)	41.14 (40.82 - 41.46)	45.31 (45.11 - 45.51)	48.66 (48.45 - 48.87)
Mika	57.42 (57.36 - 57.48)	57.87 (57.35 - 58.40)	61.65 (61.30 - 62.00)	56.97 (56.54 - 57.41)	59.02 (58.81 - 59.24)	83.17 (83.05 - 83.28)
Akinbami_1	6.49 (6.46 - 6.52)	6.58 (6.42 - 6.75)	5.82 (5.66 - 5.98)	6.10 (5.93 - 6.27)	7.65 (7.55 - 7.75)	8.87 (8.75 - 8.99)
Akinbami_2	6.50 (6.47 - 6.53)	4.29 (4.16 - 4.43)	6.18 (6.02 - 6.34)	4.79 (4.66 - 4.93)	6.36 (6.27 - 6.45)	7.41 (7.31 - 7.51)
Akinbami_3	15.91 (15.87 - 15.95)	14.98 (14.74 - 15.23)	14.16 (13.92 - 14.41)	14.67 (14.46 - 14.89)	16.77 (16.61 - 16.94)	17.61 (17.44 - 17.77)
Salomon	18.76 (18.71 - 18.81)	20.50 (20.20 - 20.80)	20.68 (20.43 - 20.94)	32.02 (31.73 - 32.32)	20.11 (19.93 - 20.28)	26.68 (26.50 - 26.86)
Astley	56.60 (56.53 - 56.67)	34.44 (34.12 - 34.76)	48.80 (48.42 - 49.19)	38.99 (38.67 - 39.32)	55.19 (54.96 - 55.41)	89.09 (88.96 - 89.21)

Table SM7

Specificity and its 95% confidence interval for the selected countries for 2021, in %

Method	Brazil	Canada	Israel	Japan	Turkey	South Africa
Menni_1	94.67 (94.65 - 94.70)	96.88 (96.84 - 96.91)	96.98 (96.92 - 97.04)	98.12 (98.09 - 98.15)	93.73 (93.65 - 93.80)	93.61 (93.52 - 93.70)
Menni_2	84.24 (84.21 - 84.28)	79.32 (79.11 - 79.53)	84.77 (84.57 - 84.96)	83.62 (83.48 - 83.77)	85.95 (85.81 - 86.09)	79.86 (79.70 - 80.02)
Roland	86.03 (85.99 - 86.06)	89.97 (89.90 - 90.04)	90.61 (90.38 - 90.84)	84.46 (84.39 - 84.53)	88.24 (88.15 - 88.34)	83.69 (83.53 - 83.85)
Smith	67.04 (67.00 - 67.09)	95.14 (95.08 - 95.19)	90.65 (90.44 - 90.86)	95.15 (95.11 - 95.20)	86.82 (86.34 - 87.31)	61.61 (61.44 - 61.79)
Zoabi_55	62.52 (61.75 - 63.29)	92.17 (91.54 - 92.81)	85.42 (84.91 - 85.92)	90.86 (90.30 - 91.41)	70.58 (70.34 - 70.81)	62.74 (62.40 - 63.09)
Zoabi_65	61.52 (60.02 - 63.03)	91.75 (91.04 - 92.46)	85.40 (84.93 - 85.87)	89.05 (88.07 - 90.02)	70.91 (70.71 - 71.12)	62.68 (62.24 - 63.11)
CDC	85.62 (85.58 - 85.66)	48.56 (48.44 - 48.68)	55.38 (55.23 - 55.54)	54.73 (54.63 - 54.83)	42.38 (42.23 - 42.53)	42.44 (42.26 - 42.63)
Shoer	57.88 (57.81 - 57.94)	82.04 (81.80 - 82.28)	85.79 (85.51 - 86.07)	85.96 (85.63 - 86.29)	88.25 (88.13 - 88.38)	62.96 (62.68 - 63.23)
Bhattacharya	90.14 (90.12 - 90.17)	92.79 (92.73 - 92.85)	93.13 (93.04 - 93.22)	60.31 (60.00 - 60.61)	89.02 (88.93 - 89.12)	59.11 (58.88 - 59.35)
WHO	97.73 (97.72 - 97.74)	97.72 (97.68 - 97.75)	98.12 (98.08 - 98.17)	97.05 (97.01 - 97.08)	97.12 (97.07 - 97.17)	95.88 (95.81 - 95.94)
Perez	93.57 (93.54 - 93.60)	94.60 (94.55 - 94.65)	95.30 (95.23 - 95.37)	92.82 (92.77 - 92.87)	92.41 (92.33 - 92.49)	90.86 (90.76 - 90.96)
Mika	87.47 (87.44 - 87.50)	89.78 (89.28 - 90.28)	90.67 (90.49 - 90.84)	93.26 (92.81 - 93.71)	88.24 (88.15 - 88.34)	60.61 (60.47 - 60.90)
Akinbami_1	98.97 (98.96 - 98.98)	99.04 (99.02 - 99.06)	99.31 (99.29 - 99.34)	99.60 (99.58 - 99.61)	98.73 (98.70 - 98.76)	98.24 (98.21 - 98.28)
Akinbami_2	99.38 (99.37 - 99.39)	99.70 (99.69 - 99.71)	99.75 (99.73 - 99.76)	99.90 (99.89 - 99.91)	99.33 (99.31 - 99.36)	99.16 (99.13 - 99.19)
Akinbami_3	95.80 (95.78 - 95.82)	96.84 (96.81 - 96.88)	97.41 (97.36 - 97.46)	96.43 (96.40 - 96.47)	95.13 (95.07 - 95.19)	95.50 (95.42 - 95.58)
Salomon	96.20 (96.18 - 96.22)	97.11 (97.07 - 97.15)	97.54 (97.48 - 97.59)	95.66 (95.62 - 95.70)	96.11 (96.06 - 96.17)	92.10 (92.01 - 92.20)
Astley	89.74 (89.70 - 89.78)	97.92 (97.88 - 97.96)	96.88 (96.82 - 96.94)	98.28 (98.25 - 98.31)	91.48 (91.40 - 91.57)	56.00 (55.76 - 56.24)

Table SM8

Precision and its 95% confidence interval for the selected countries for 2021, in %

Method	Brazil	Canada	Israel	Japan	Turkey	South Africa
Menni_1	84.66 (84.66 - 84.74)	59.99 (59.68 - 60.29)	72.38 (71.90 - 72.87)	67.16 (67.89 - 68.31)	77.92 (77.69 - 78.15)	81.23 (81.15 - 81.32)
Menni_2	72.65 (72.6 - 72.7)	27.78 (27.61 - 27.95)	44.12 (43.80 - 44.43)	31.07 (30.88 - 31.27)	66.50 (66.27 - 66.73)	65.90 (65.79 - 66.03)
Roland	74.24 (74.21 - 74.28)	39.04 (38.84 - 39.24)	53.71 (53.16 - 54.26)	31.67 (31.51 - 31.83)	70.23 (70.01 - 70.44)	68.40 (68.32 - 68.48)
Smith	58.59 (58.55 - 58.64)	52.72 (52.47 - 52.96)	54.26 (53.71 - 54.82)	52.97 (52.77 - 53.19)	68.70 (68.07 - 69.33)	52.77 (52.62 - 52.93)
Zoabi_55	54.70 (54.50 - 54.90)	37.09 (36.22 - 37.96)	33.59 (33.05 - 34.13)	31.67 (30.82 - 32.52)	44.75 (44.47 - 45.03)	52.11 (51.90 - 52.32)
Zoabi_65	54.44 (54.05 - 54.84)	36.51 (35.55 - 37.46)	33.55 (33.03 - 34.07)	30.17 (29.16 - 31.18)	44.85 (44.58 - 45.12)	52.04 (51.80 - 52.29)
CDC	50.14 (50.10 - 50.18)	16.22 (16.14 - 16.31)	25.12 (24.96 - 25.28)	17.11 (17.02 - 17.20)	41.21 (41.05 - 41.36)	46.95 (46.81 - 47.07)
Shoer	52.66 (52.61 - 52.72)	29.66 (29.46 - 29.86)	45.25 (44.80 - 45.71)	34.32 (34.04 - 34.62)	70.27 (70.02 - 70.52)	53.75 (53.60 - 53.91)
Bhattacharya	78.77 (78.75 - 78.83)	45.57 (45.37 - 45.78)	59.80 (59.44 - 60.15)	30.88 (30.68 - 31.08)	71.42 (71.21 - 71.62)	50.86 (50.67 - 51.05)
WHO	80.52 (80.48 - 80.56)	47.07 (46.56 - 47.45)	62.91 (62.25 - 63.57)	49.04 (48.70 - 49.37)	73.59 (73.25 - 73.93)	74.79 (74.61 - 74.84)
Perez	81.28 (81.25 - 81.31)	47.07 (46.79 - 47.33)	62.09 (61.64 - 62.55)	38.39 (38.17 - 38.64)	73.44 (73.20 - 73.67)	75.47 (75.36 - 75.55)
Mika	75.77 (75.74 - 75.80)	39.23 (38.96 - 39.49)	53.95 (53.50 - 54.41)	48.53 (47.67 - 49.37)	70.23 (70.01 - 70.44)	51.99 (51.81 - 52.17)
Akinbami_1	80.80 (80.48 - 81.13)	43.47 (42.94 - 43.87)	59.89 (58.75 - 61.02)	62.18 (61.69 - 62.77)	73.71 (73.20 - 74.22)	74.83 (74.47 - 75.20)
Akinbami_2	87.91 (87.60 - 88.23)	61.62 (61.14 - 62.10)	81.22 (80.28 - 82.16)	83.71 (83.45 - 83.98)	81.64 (81.09 - 82.18)	83.73 (83.26 - 84.20)
Akinbami_3	71.91 (71.89 - 71.93)	34.89 (34.42 - 35.25)	49.06 (48.39 - 49.73)	30.93 (30.55 - 31.25)	61.52 (61.15 - 61.90)	69.26 (69.09 - 69.58)
Salomon	77.08 (77.07 - 77.08)	44.45 (44.06 - 44.9)	59.81 (59.19 - 60.43)	44.58 (44.28 - 44.86)	70.84 (70.51 - 71.16)	66.19 (66.08 - 66.29)
Astley	79.01 (78.99 - 79.04)	65.49 (64.96 - 66.02)	73.47 (73.04 - 73.90)	71.21 (71.04 - 71.35)	75.27 (75.06 - 75.49)	50.01 (49.83 - 50.19)

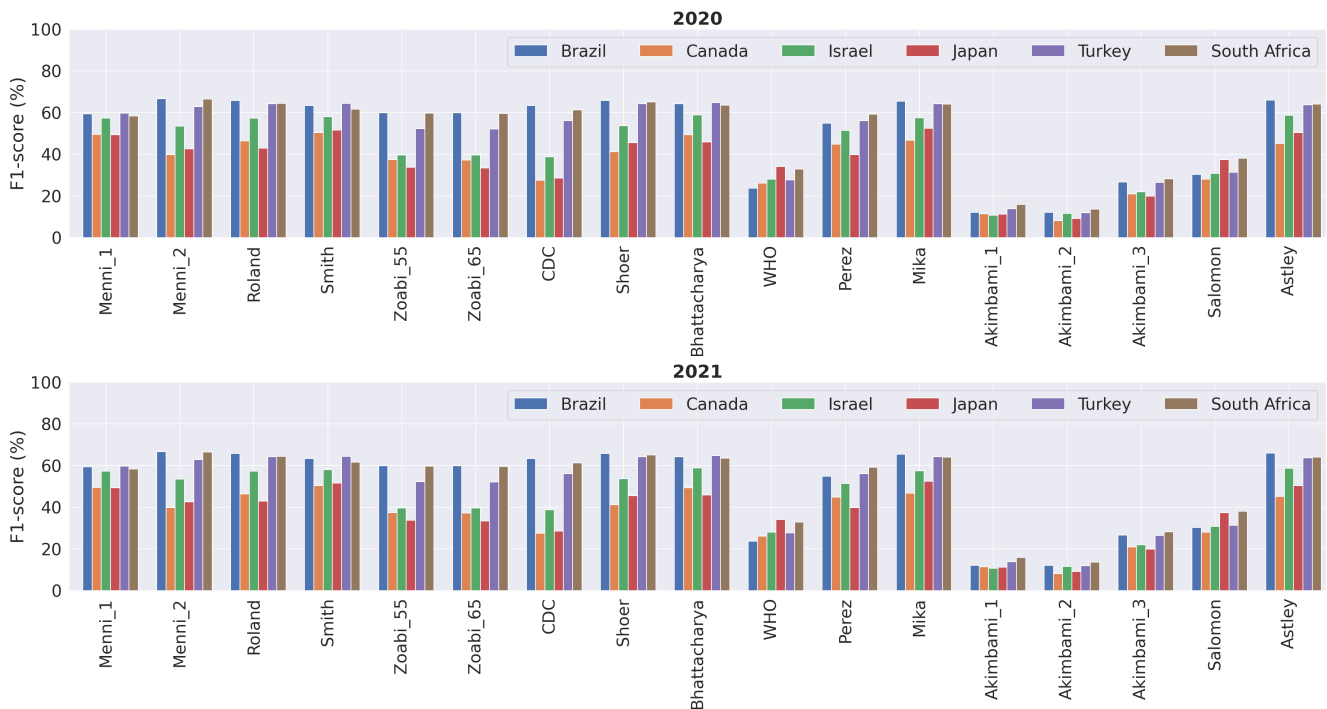


Figure SM1: F₁ scores in % obtained by each COVID-19 detection method across the six countries for 2020 and 2021.

C. Checklist for assessment of requirements and recommendations for sound medical ML contributions.

I. Problem understanding

1. Is the study population described, also in terms of inclusion/exclusion criteria (e.g., patients older than 18 tested for COVID-19; all inpatients hospitalized for 24 or more hours)?

Response: Yes. In section 2.1 (Dataset), we first describe the data collection process for the UMD-CTIS data. This description mentions that Facebook users were invited to participate in a web-based survey in which participants must report an age above or equal to 18 years. We describe that the performance comparison is performed using datasets extracted from UMD-CTIS for six countries (Brazil, Canada, Israel, Japan, Turkey, and South Africa) and two periods (2020 and 2021). Furthermore, we indicate that for each country and period, we select answers reporting a lab test done in the last 14 days and at least one potential COVID-19 symptom (tested symptomatic). In addition, Table 1 shows quantitatively the characteristics of the study population such as gender, test positivity rate, and age groups.

2. Is the study design described? (e.g., retrospective, prospective, cross-sectional, observational, randomized control trial)

Response: We use previously collected datasets. Datasets have their characteristics and the UMD-CTIS collection process is detailed in the citation below. In this work, we compare the performance of various state-of-the-art methods for detecting COVID-19-infected people using UMD-CTIS datasets for six countries and two periods. We describe the experimental protocol to compare detection methods from UMD-CTIS datasets in Section 2.2.

Kreuter, N. Barkay, A. Bilinski, A. Bradford, S. Chiu, R. Eliat, J. Fan, T. Galili, D. Haimovich, B. Kim, et al., Partnering with Facebook on a university-based rapid turn-around global survey, *Survey Research Methods: SRM 14* (2020) 159–163.

3. Is the study setting described? (e.g., teaching tertiary hospital; primary care ambulatory, nursing home, medical laboratory, R&D laboratory)

Response: In section 2.1, it is explained that it is a direct survey carried out through Facebook where its users were invited to participate, and those who accepted were transferred to a web platform.

4. Is the source of data described? (e.g., electronic specialty registry; laboratory information system; electronic health record; picture archiving and communication system)

Response: Yes. Section 2.1 describe the source of data.

5. Is the medical task reported? (e.g., diagnostic detection, diagnostic characterization, diagnostic staging, prognosis (on which endpoint), event prediction, risk stratification, anatomical structure segmentation, treatment selection and planning, monitoring)

Response: UMD-CTIS was created as a surveillance tool to track COVID-19 indicators trends among 114 countries/territories. To this end, this survey collects information about multiple characteristics such as symptoms, demographics, age groups, gender, isolation measures, mental health, and vaccination acceptance. In this work, we compare the performance of the various previously reported COVID-19 detection methods using the UMD-CTIS information for six countries and two periods. The methods under test typically consider a reduced set of variables such as symptoms, age, and gender (which are collected by the UMD-CTIS).

6. Is the data collection process described, also in terms of setting-specific data collection strategies (e.g., whether body temperatures are measured only in the morning; whether some blood tests are performed only in light of a specific diagnostic hypothesis)? Any consideration about data quality is appreciated, e.g., in regard to completeness, plausibility, and robustness with respect to upcoding or downcoding practices.

Response: The reference below describes the data collection strategy :

Kreuter, N. Barkay, A. Bilinski, A. Bradford, S. Chiu, R. Eliat, J. Fan, T. Galili, D. Haimovich, B. Kim, et al., Partnering with Facebook on a university-based rapid turn-around global survey, *Survey Research Methods: SRM 14* (2020) 159–163.

II. Data understanding

7. Are the subject demographics described in terms of average age (mean or median); age variability (standard deviation (SD) or inter-quartile range (IQR)); gender breakdown (e.g., 55% female, 44% male, 1% not reported); main comorbidities; ethnic group (e.g., Native American, Asian, South East Asian, African, African American, Hispanic, Native Hawaiian or Other Pacific Islander, European or American White); socioeconomic status?

Response: For each country and period, we report the number of tested symptomatic from gender and age groups (see Table 1). For some COVID-19 detection methods under test, gender and age groups are input variables.

8. If the task is supervised, is the gold standard described? (e.g., “100 manually annotated clinical notes and pain scores recorded in EHR, Death, re-admission and International Classification of Disease (ICD) codes in discharge letters”). In particular, the authors should describe the process of ground truthing described in terms of: Number of annotators (raters) producing the labels; Their profession and expertise (e.g., years from specialization or graduation); Particular instructions given to annotators for quality control (e.g., which data were discarded and why); Inter-rater agreement score (e.g., Alpha, Kappa, Rho); Labeling technique (e.g., majority voting, Delphi method, consensus iteration).

Response: Notice that, for each country and period, we extracted answers reporting a lab test done in the last 14 days. We also extract answers reporting at least one potential COVID-19 symptom. As mentioned in Section 2.2 (Experimental Protocol), we selected answers from tested symptomatic individuals to obtain a ground truth set to build the machine learning models. Ground truth selection using responses from tested people has been used in various methods under test.

- (a) 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.
 - (b) C. M. Astley, G. Tuli, K. A. M. Cord, E. L. Cohn, B. Rader, T. J. Varrelman, S. L. Chiu, X. Deng, K. Stewart, T. H. Farag, K. M. Barkume, S. LaRocca, K. A. Morris, F. Kreuter, J. S. Brownstein, 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).
 - (c) 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).
9. In the case of tabular data, are the features described (also in regard to how they were used in the model in terms of categories or transformation)? This description should be done for all, or, in the case that the

features exceed 20, for a significant subset of the most predictive features in the following terms: name, short description, type(nominal, ordinal, continuous), and

- If continuous: unit of measure, range (min, max), mean and standard deviation (or median and IQR). Violin plots of some relevant continuous features are appreciated. If data are hematochemical parameters, also mention the brand and model of the analyzer equipment.
- If nominal, all codes/values and their distribution. Feature transformation (e.g., one-hot encoding) should be reported if applied. Any terminology standard should be explicitly mentioned (e.g., LOINC, ICD-11, SNOMED) if applied.

Response: Note that every method under test considers a subset of variables that includes gender, age groups, or symptom. Therefore, the characteristics of the variables used are defined in Table 1, which are described in either numbers or percentages. Anyway, in the cases that were required, an explanation was added (See section 2.2): “Since questionnaires contain categorical data, we apply binary encoding such that every potential choice aggregates a column to the data”.

III. Data preparation

10. Is outlier detection and analysis performed and reported? If the answer is yes, the definition of an outlier should be given and the techniques applied to manage outliers should be described (e.g., removal through application of an Isolation Forest model).

Response: This dataset does not require an outlier detection method since the input variables are categorical with a limited number of possible responses.

11. Is missing-value management described? This description should be reported in the following terms:
- The missing rate for each feature should be reported;
 - The technique of imputation, if any, should be described, and reasons for its choice should be given. If the missing rate is higher than 10%, a reflection about the impact on the performance of a technique with respect to others would be appreciable.

Response: This dataset does not require a missing value detection method since the input variables are categorical with a limited number of possible responses. Non-reported categorical variables encode all possible responses as zero after one-hot encoding.

12. Is feature pre-processing performed and described? This description should be reported in terms of scaling transformations (e.g., normalization, standardization, log-transformation) or discretization procedures applied to continuous features, and encoding of categorical or ordinal variables (e.g., one-hot encoding, ordinal encoding).

Response: We do not implement a feature engineering process.

13. Is data imbalance analysis and adjustment performed and reported? The authors should describe any imbalance in the data distribution, both in regard to the target (e.g., only 10% of the patients were affected by a given disease); and in regard to important predictive features (e.g., female patients accounted for less than 10% of the total cases). The authors should also report about any technique (if any) applied to adjust the above mentioned imbalances (e.g., under- or oversampling, SMOTE).

Response: Yes. We reported imbalances in the test positivity rate (TPR) across countries for the two periods. In Section 3, we indicate that the TPR values exhibited by Brazil, Turkey, and South Africa are at least twofold those shown by Canada, Israel, and Japan. Furthermore, we point out that F_1 scores are highly affected by imbalanced classes. To address this issue, we evaluate the performance of the various detection methods for three groups: the broad set of six countries, the set of countries with high TPR (Brazil, Turkey, and South Africa), and the countries with low TPR (Canada, Israel, and Japan). Table 2 displays the average of the F_1 score for the overall five countries (overall), for the countries with high TPR (High TPR), and for the countries with low TPR (Low TPR) for 2020, 2021, and the entire interval 2020-2021.

IV. Modeling

14. Is the model task reported? (e.g., classification, multi-label classification, ordinal regression, continuous regression, clustering, dimensionality reduction, segmentation).

Response: Yes. We compare various methods for detecting COVID-19-positive cases. These methods attempt to solve a binary classification problem.

15. Is the model output specified? (e.g., disease positivity probability score, probability of infection within 5 days, postoperative 3-month pain scores).

Response: Yes. The output of every model under test obtains a binary detection for each valid data sample.

16. Is the model architecture or type described? (e.g., SVM, Random Forest, Boosting, Logistic Regression, Nearest Neighbors, Convolutional Neural Network).

Response: Yes. Supplemental Material describes all COVID-19 detection methods under comparison with the corresponding references.

V. Validation

17. Is the data splitting [60] described (e. g., no data splitting; k-fold cross-validation (CV); nested k-fold CV; repeated CV; bootstrap validation; leave-one-out CV; 80%/10%10% train/validation/test)? In the case of data splitting, the authors must explicitly state that splitting was performed before any pre-processing steps (e.g., normalization, standardization, missing value imputation, feature selection) or model construction steps (training, hyper-parameter optimization), so to avoid data leakage and overfitting.

Response: In section 2.2 is explained this process: Our study divided every dataset into 100 partitions. For each trial, 80% of the dataset rows (questionnaires or samples) were randomly selected as training samples, and the remaining 20% were used to test the detection methods.

18. Is the model training and selection described? In particular, the training procedure, hyper-parameter optimization or model selection should.

- Range of hyper-parameters;
- Method used to select the best hyper-parameter configuration (e.g., Hyper- parameter selection was performed through nested k-fold CV based grid search);
- Full specification of the hyper-parameters used to generate results;
- Procedure (if any) to limit over-fitting, in particular as related to the sample size.

Response: Supplemental Material A describes this information. Specifically, the references provide the parameter set of the majority of the detection models under comparison (Smith, Menni, WHO, CDC, Zoabi, Solomon, Akimbami, Solomon, Bhattacharya, and Perez). For the remaining detection methods, we carefully follow the procedure outlined in the corresponding reference to build the corresponding model. For example, Atsley et al describe both the input variable set and the training stage.

C. M. Astley, G. Tuli, K. A. M. Cord, E. L. Cohn, B. Rader, T. J. Varrelman, S. L. Chiu, X. Deng, K. Stewart, T. H. Farag, K. M. Barkume, S. LaRocca, K. A. Morris, F. Kreuter, J. S. Brownstein, 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).

19. (classification models) Is the model calibration described? If the answer is yes, the Brier score should be reported, and a calibration plot should be presented.

Response: No. We do not include a model calibration. The idea is to compare various previously reported COVID-19 detection methods. On one hand, references provide the parameter set of the majority of detection models. On the other hand, references do not include model calibration to enhance performance or explainability.

20. Is the internal/internal-external model validation procedure described (e.g., internal 10-fold CV, time-based cross-validation)? The sets have been split before normalization, standardization and imputation, to avoid data leakage (also refer to item 17 of this guideline). If possible, the authors should also comment on the adequacy of the available sample size for model training and validation. Moreover, the authors should try to choose the test set so that it is the most diverse with respect to the remainder of the sample [66] (w.r.t. some multivariate similarity function) and how this choice relates to conservative (and lower-bound) estimates of the model's accuracy (and performance).

Response: No. We do not include an internal/external model validation. The idea is to compare various previously reported COVID-19 detection methods. On one hand, references provide the parameter set of the majority of detection models. On the other hand, references do not include model validation.

21. Has the model been externally validated? If the answer is yes, the characteristics of the external validation set(s) should be described. For instance, the authors could comment about the heterogeneity of the data with respect to the training set (e.g., degree of correspondence Ψ , Data Representativeness Criterion) and the cardinality of the external sample. If the performance on external datasets is found to be comparable with (or better than) that on training and internal datasets, the authors should provide some explanatory conjectures for why this happened (e.g., high heterogeneity of the training set, high homogeneity of the external dataset).

Response: This paper is an effort to externally validate the models under comparison. More precisely, we aim to evaluate the performance of different COVID-19 detection methods using the same UMD-CTIS data

extracted from six countries.

22. Are the main error-based metrics used?

- Classification performance should be reported in terms of: Accuracy, Balanced accuracy, Specificity, Sensitivity (recall), Area Under the Curve (if the positive condition is extremely rare – as in case of stroke events – authors could consider the “Area under the Precision-Recall Curve” [70]). Optionally also in terms of: positive and negative predictive value, F1 score, Matthew coefficient [71], F score of sensitivity and specificity, the full confusion matrix, Hamming Loss (for multi-label classification), Jaccard Index (for multi-label classification).
- Regression performance should be reported in terms of: R2; Mean Absolute Error (MAE); Root Mean Square Error (RMSE); Mean Absolute Percentage Error (MAPE) or the Ratio between MAE (or RMSE) and SD (of the target);
- Clustering performance should be reported in terms of: External validation metrics (e.g., mutual information, purity, Rand index), when ground truth labels are available, and Internal validation metrics (e.g., Davies-Bouldin index, Silhouette index, Homogeneity). The reported results of internal validation metrics should be discussed [72]
- Image segmentation performance, depending on the specific task, should be reported in terms of metrics like [73]: accuracy-based metrics (e.g., Pixel accuracy, Jaccard Index, Dice Coefficient), distance-based metrics (e.g., mean absolute, or maximum difference), or area-based metrics (e.g., true positive fraction, true negative fraction, false positive fraction, false negative fraction).
- 5. Reinforcement learning performance, depending on the specific task, should be reported in terms of metrics like [74]: Fixed-Policy Regret, Dispersion across Time, Dispersion across Runs, Risk across Time, Risk across Runs, Dispersion across Fixed-Policy Roll-outs, Risk across Fixed-Policy Rollouts.

The above estimates should be expressed, 90% confidence intervals (CI), or with other indicators of variability, with respect to the evaluation metrics reported. In this case, the authors should report which methods were applied for the computation of the confidence intervals (e.g., whether k-fold CV or bootstrap was applied, normal approximation). When comparing multiple models, the authors should discuss the statistical significance of the observed differences (e.g., through CI comparisons, or hypothesis testing). When comparing multiple regression models, a Taylor diagram could be reported and discussed.

Response: Section 2.3 describes the quality metrics used to compare the performance of the various COVID-19 detection methods: specificity, sensitivity, F1 score, and area under the curve (AUC). Notice that the results are estimated with a 95% confidence interval (see Section 3).

23. Are some relevant errors described? characteristic of some noteworthy classification errors or cases for which the regression prediction was much higher ($>2\times$) than the MAE. If these cases represent statistical outliers for some covariates, the authors should comment on that. To detect relevant cases, the authors could focus on those cases on which the inter-rater agreement (either re ground truth or by comparing human vs. model's performance) is lowest.

Response: Section 3.1 presents a general description of the results, but also highlights some values of interest.

VI. Deployment

24. Is the target user indicated? (e.g., clinician, radiologist, hospital management team, insurance company, patients)

Response: Does not apply.

25. (classification models) Is the utility of the model discussed? The authors should report the performance of a baseline model (e.g., logistic regression, Naive Bayes). Additionally, the authors could report the Net Benefit or similar metrics and present utility curves. In particular, the authors are encouraged to discuss the selection of appropriate risk thresholds; the relative value of benefits (true positives/negatives) and harms (false positives/negatives); and the clinical utility of the proposed models.

Response: The article makes an in-depth comparison of various COVID 19 detection techniques, so the objective is to comparatively analyze the behavior between them.

26. Is information regarding model available [80] (e.g., feature importance, interpretable surrogate models, information about the model parameters)? Claims of "high" or "adequate" model interpretability (e.g., by means of visual aids like decision trees, Variable Importance Plots or Shapley Additive Explanations Plots (SHAP)) or model causability should always be supported by some user study, even qualitative or questionnaire-based. In the case surrogate models were applied, the authors should report about their fidelity.

Response: There is a whole section (section 3.2) dedicated to an analysis of the explainability of the best models by category of detection technique considered in the work.

27. Is there any discussion regarding model fairness, ethical concerns or risks of bias (for a list of clinically relevant biases, refer to)? If possible, the authors should report the model performance stratified for particularly relevant population strata (e.g., model performance on male vs. female subjects, or on minority groups).

Response: The study considers two fundamental dimensions of analysis: year and country. There is a discussion of the quality of the models for these two dimensions.

28. Is any point made about the environmental sustainability of the model, or about the carbon footprint, of either the training phase or inference phase (use) of the model? If the answer is yes, then such a footprint should be expressed in terms of carbon dioxide equivalent (CO₂eq) and details about the estimation method should be given. Any efforts to this end will be appreciated, including those based on tools available online, as well as any attempts to popularise this concept, e.g., through equivalences with the consumption of everyday devices such as smartphones or kilometers traveled by a fossil-fuelled car.

Response: We do not make any analysis on the environmental sustainability or carbon footprint of the model.

29. Is code and data shared with the community? § If not, are reasons given? If code and data are shared, institutional repositories such as Zenodo should be preferred to private-owned repositories (arxiv, GitHub). If code is shared, specification of dependencies should be reported and a clear distinction between training code and evaluation code should be made. The authors should also state whether the developed system, either as a sand-box or as fully-operating system, has been made freely accessible on the Web.

Response: Section 7 describes where the data and code are available.

30. Is the system already adopted in daily practice? If the answer is yes, the authors should report on where (setting name) and since when. Moreover, appreciated additions would regard: the description on the digitized workflow integrating the system; any comment about the level of use; a qualitative assessment of the level of efficacy of the system's contribution to the

clinical process; any comment about the technical and staff training effort actually required. If the answer is no, the authors should be explicit in regard to the point in the clinical workflow where the ML model should be applied, possibly using standard notation (e.g., BPMN). Moreover, the authors should also propose an assessment of the technology readiness of the described system, with explicit reference to the Technology Readiness Level framework or to any adaptation of this framework to the AI/ML domain. In either above cases (yes/no), the authors should report about the procedures (if any) for performance monitoring, model maintenance and updating.

Response: Does not apply