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]

[a] *IMDEA Networks Institute, 28918, Madrid, Spain*
[b] *CEMISID, Universidad de Los Andes, Mérida, 5101, Venezuela*
[c] *CIDITIC, Universidad EAFIT, Medellín, Colombia*
[d] *Universidade do Minho and INESC TEC, Braga, Portugal*
[e] *Inria Rennes, Rennes, France*
[f] *Universidad 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 been also 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 aforementioned 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, on a daily basis, 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, 22, 13, 23]. 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 from self-reported data. Yalçın and Ünaldı [24] 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 Ünaldı 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 [24], 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. [25] 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 [25], 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 [26]. 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 (F1-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** [23] (F_1-score: 55.97%), **Menni** [9] (F_1-score: 55.45%), **Mika** [13] (F_1-score: 53.98%), and **Shoer** [22] (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

Since 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, 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 the past 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
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

Table 1 also includes the test positive rate (TPR) where 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, the number of positive cases, the number of 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. 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.

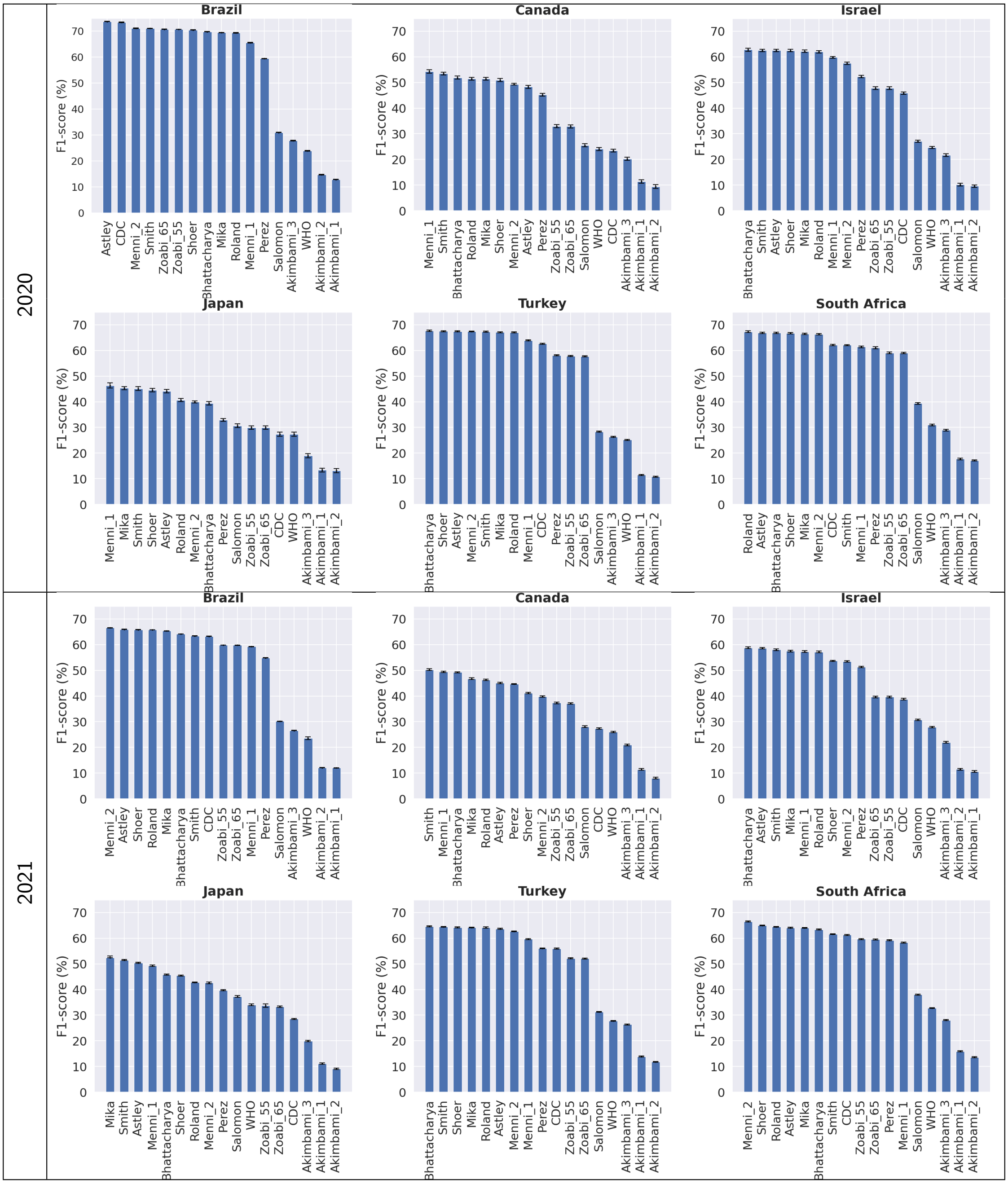

Figure 1: F_1 score in % and the corresponding 95% confidence interval obtained by the various COVID-19 detection methods for the selected countries and for 2020 and 2021.

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:

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{1}$$

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

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

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.

	2020			2021			2020-2021		
Method	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

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 [27], 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. In particular, 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

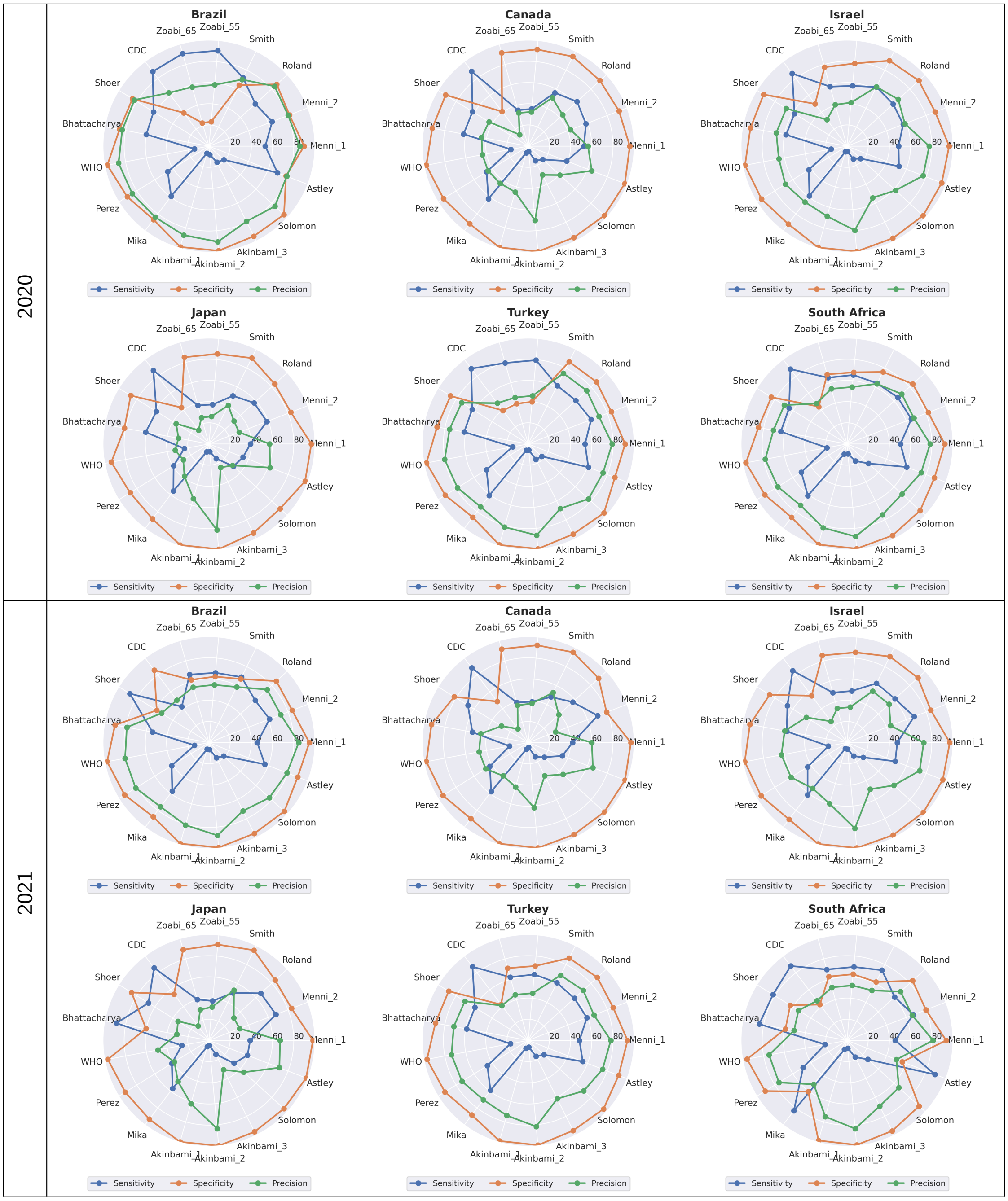

Figure 2: Radar chart of sensitivity (blue circles), specificity (orange circles), and precision (green circles) in % exhibited by the various methods for the entire set of countries and for 2020 and 2021. The closer the distance to the center, the worse the performance of the corresponding method.

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. Notice that blue lines, orange lines, and green lines correspond to sensitivity, specificity, and precision, respectively. Finally, the best methods by category of estimation detection methods are Smith for the rules-based methods, Astley for the machine learning methods, and Menni for the regression technique.

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.

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

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

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		

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.

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 a better explainability, being able to be used to give more details and better reason 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 to notice 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

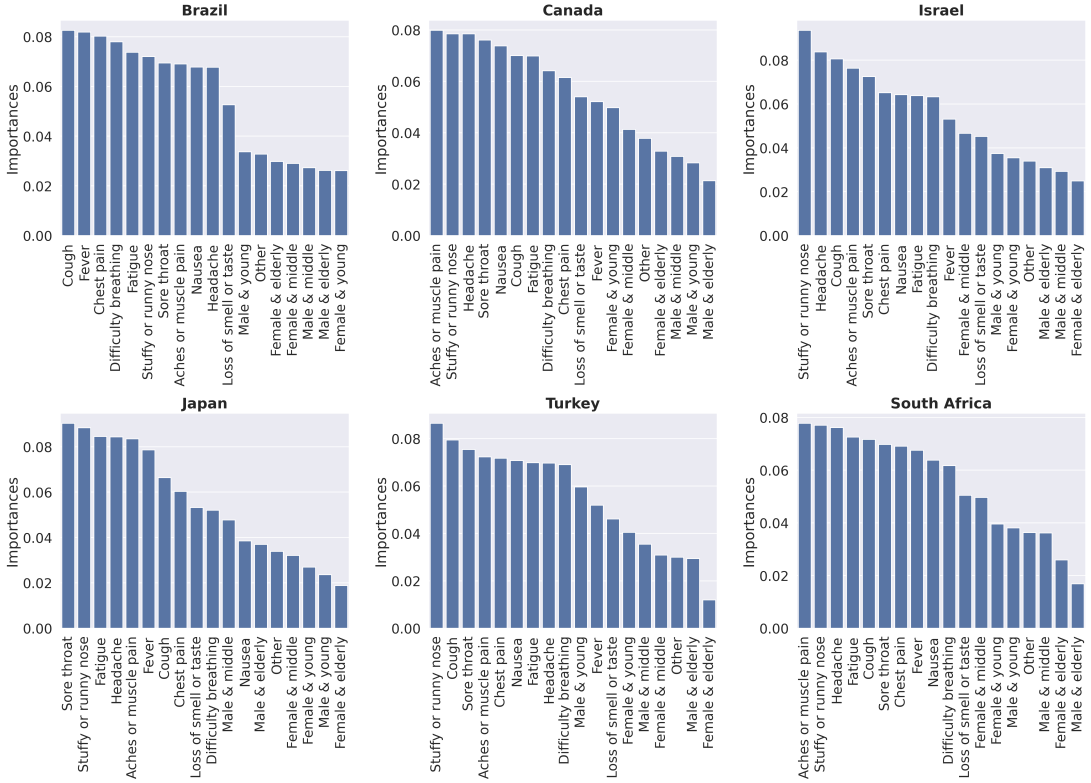

Figure 3: Feature importances of the Astley method for 2020 and for the entire set of countries.

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 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 takes into account 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. As a consequence, 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.

In future work, a selection of different machine learning techniques will be made for the use of the different variables

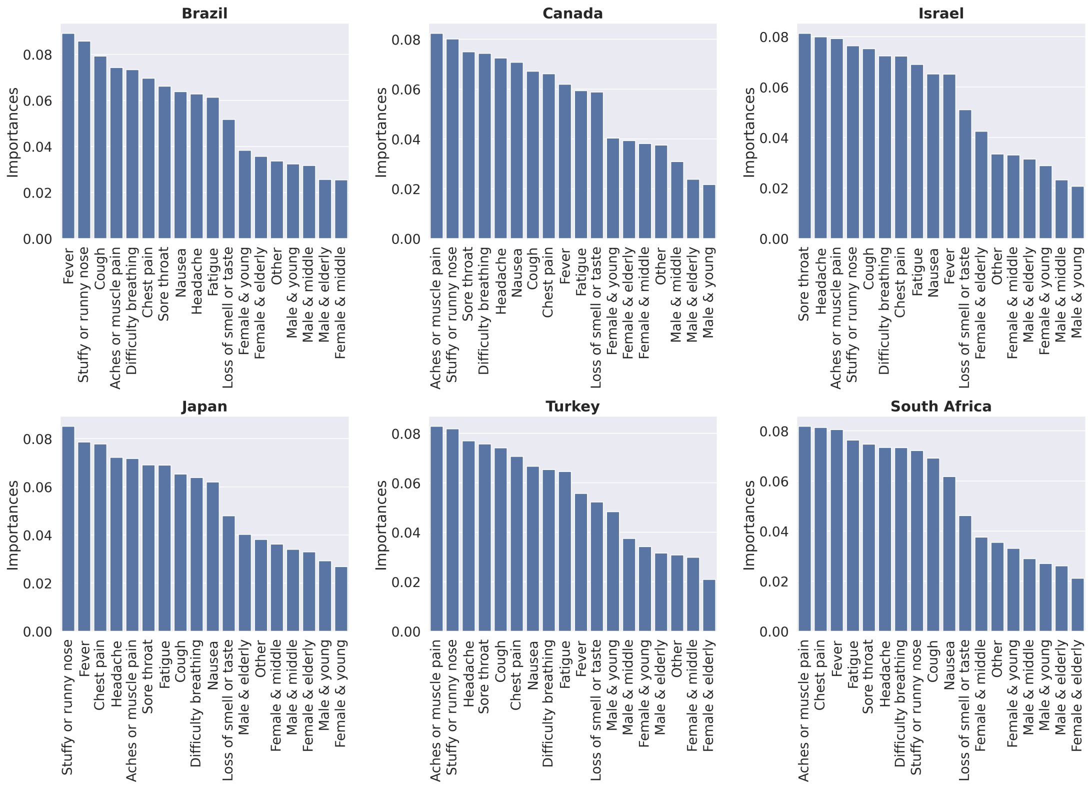

Figure 4: Feature importances of the Astley method for 2021 and for the entire set of countries.

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 F1 score and presenting the ROC curves for each model. In addition, 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 of 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 microdata of the CTIS survey from which the aggregated data was obtained cannot be shared, as per the Data Use Agreements signed with Facebook, Carnegie Mellon University (CMU) and the University of Maryland (UMD).

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. 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) e2111455118.

[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, M. Goessens, H. Hernández, R. Lillo, R. Menezes, R. Moreno, N. Nicolaou, O. Ojo, A. Ortega, E. Rausell, J. Rufino, E. Stavrakis, G. Jeevan, C. Glorioso, Estimating Active Cases of COVID-19, medRxiv (2021).

[18] J. Fan, Y. Li, K. Stewart, A. R. Kommareddy, A. Bradford, S. Chiu, F. Kreuter, N. Barkay, A. Bilinski, B. Kim, R. Eliat, T. Galili, D. Haimovich, S. LaRocca, S. Presser, K. Morris, J. A. Salomon, E. A. Stuart, R. Tibshirani, T. A. Barash, C. Cobb, A. Garcia, A. Gros, A. Isa, A. Kaess, F. Karim, O. E. Kedosha, S. Matskel, R. Melamed, A. Patankar, I. Rutenberg, T. Salmona, D. Vannette, 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] S. Shoer, T. Karady, A. Keshet, S. Shilo, H. Rossman, A. Gavrieli, T. Meir, A. Lavon, D. Kolobkov, I. Kalka, et al., Who should we test for covid-19? a triage model built from national symptom surveys, Medrxiv (2020).

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

[24] N. Yalçın, S. Ünaldı, Symptom based covid-19 prediction using machine learning and deep learning algorithms, Journal of Emerging Computer Technologies 2 (2022) 22 – 29.

[25] A. Sedik, A. M. Iliyasu, B. Abd El-Rahiem, M. E. Abdel Samea, A. Abdel-Raheem, M. Hammad, J. Peng, F. E. Abd El-Samie, A. A. Abd El-Latif, Deploying machine and deep learning models for efficient data-augmented detection of covid-19 infections, Viruses 12 (2020).

[26] 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.

[27] H. He, Y. Ma, Imbalanced Learning: Foundations, Algorithms, and Applications, Wiley-IEEE Press, 1st edition, 2013.

[28] M. Pollán, B. Pérez-Gómez, R. Pastor-Barriuso, J. Oteo, M. A. Hernán, M. Pérez-Olmeda, J. L. Sanmartín, A. Fernández-García, I. Cruz, N. Fernández de Larrea, M. Molina, F. Rodríguez-Cabrera, M. Martín, P. Merino-Amador, J. León Paniagua, J. F. Muñoz-Montalvo, F. Blanco, R. Yotti, F. Blanco, R. Gutiérrez Fernández, M. Martín, S. Mezcua Navarro, M. Molina, J. F. Muñoz-Montalvo, M. Salinero Hernández, J. L. Sanmartín, M. Cuenca-Estrella, R. Yotti, J. León Paniagua, N. Fernández de Larrea, P. Fernández-Navarro, R. Pastor-Barriuso, B. Pérez-Gómez, M. Pollán, A. Avellón, G. Fedele, A. Fernández-García, J. Oteo Iglesias, M. T. Pérez Olmeda, I. Cruz, M. E. Fernandez Martinez, F. D. Rodríguez-Cabrera, M. A. Hernán, S. Padrones Fernández, J. M. Rumbao Aguirre, J. M. Navarro Marí, B. Palop Borrás, A. B. Pérez Jiménez, M. Rodríguez-Iglesias, A. M. Calvo Gascón, M. L. Lou Alcaine, I. Donate Suárez, O. Suárez Álvarez, M. Rodríguez

Pérez, M. Cases Sanchís, C. J. Villafáfila Gomila, L. Carbo Saladrigas, A. Hurtado Fernández, A. Oliver, E. Castro Feliciano, M. N. González Quintana, J. M. Barrasa Fernández, M. A. Hernández Betancor, M. Hernández Febles, L. Martín Martín, L.-M. López López, T. Ugarte Miota, I. De Benito Población, M. S. Celada Pérez, M. N. Vallés Fernández, T. Maté Enríquez, M. Villa Arranz, M. Domínguez-Gil González, I. Fernández-Natal, G. Megías Lobón, J. L. Muñoz Bellido, P. Ciruela, A. Mas i Casals, M. Doladé Botías, M. A. Marcos Maeso, D. Pérez del Campo, A. Félix de Castro, R. Limón Ramírez, M. F. Elías Retamosa, M. Rubio González, M. S. Blanco Lobeiras, A. Fuentes Losada, A. Aguilera, G. Bou, Y. Caro, N. Marauri, L. M. Soria Blanco, I. del Cura González, M. Hernández Pascual, R. Alonso Fernández, P. Merino-Amador, N. Cabrera Castro, A. Tomás Lizcano, C. Ramírez Almagro, M. Segovia Hernández, N. Ascunce Elizaga, M. Ederra Sanz, C. Ezpeleta Baquedano, A. Bustinduy Bascaran, S. Iglesias Tamayo, L. Elorduy Otazua, R. Benarroch Benarroch, J. Lopera Flores, A. Vázquez de la Villa, Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study, The Lancet 396 (2020) 535–544.