Original Research

COVID-19 diagnosis prediction by symptoms of tested individuals: a machine learning approach

Yazeed Zoabi[1], Noam Shomron[1,*]

[1]Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, 6997801, Israel.

*Correspondence should be addressed to Noam Shomron (nshomron@tauex.tau.ac.il).

Abstract

Motivation: Effective screening of SARS-CoV-2 enables quick and efficient diagnosis of COVID-19 and can mitigate the burden on healthcare systems. Prediction models that combine several features to estimate the risk of infection have been developed in hopes of assisting medical staff worldwide in triaging patients when allocating limited healthcare resources.

Results: We established a machine learning approach that trained on records from 51,831 tested individuals (of whom 4,769 were confirmed COVID-19 cases) while the test set contained data from the following week (47,401 tested individuals of whom 3,624 were confirmed COVID-19 cases). Our model predicts COVID-19 test results with high accuracy using only eight features: gender, whether age is above 60, known contact with an infected individual, and five initial clinical symptoms.

Summary: Overall, based on the nationwide data publicly reported by the Israeli Ministry of Health, we developed a model that detects COVID-19 cases by simple features accessed by asking basic questions. Our framework can be used, among other considerations, to prioritize testing for COVID-19 when allocating limited testing resources.

Availability: All data used in this study was retrieved from the Israeli Ministry of Health website.
Contact: yazeed@tauex.tau.ac.il, nshomron@tauex.tau.ac.il

1 Introduction

The novel coronavirus disease 2019 (COVID-19) pandemic caused by the newly emerged SARS-CoV-2 is a critical and urgent threat to global health. The outbreak in early December 2019 in the Hubei province of the People's Republic of China has spread worldwide. As of May 2020, the overall number of patients confirmed to have the disease has exceeded 3,580,000 in more than 180 countries, the number of people infected is probably much higher, and more than 250,000 people have died from COVID-19. [1]

This pandemic continues to challenge medical systems worldwide in many aspects, including sharp increases in demands for hospital beds and critical shortages in medical equipment, while many healthcare workers have themselves been infected. Thus, the capacity for immediate clinical decisions and effective usage of healthcare resources is crucial. The most

validated diagnosis test for COVID-19, using reverse transcriptase polymerase chain reaction (RT-PCR), is currently in shortage in developing countries. This contributes to increased infection rates and delays critical preventive measures.

Effective screening enables quick and efficient diagnosis of COVID-19 and can mitigate the burden on healthcare systems. Prediction models that combine several features to estimate the risk of infection have been developed in hopes of assisting medical staff worldwide in triaging patients when allocating limited healthcare resources. These models use features such as computer tomography (CT) scans [2–5], information available at hospital admission including clinical symptoms [6], and laboratory tests. [7]

In Israel, all diagnostic laboratory tests for COVID-19 are performed according to criteria determined by the Israeli Ministry of Health. While subject to change, these currently include the presence and severity of clinical symptoms, possible exposure to confirmed patients, geographical area, the risk of complications if infected, and other factors. [8]

2 Methods

Study Data and Features

The Israeli Ministry of Health recently publicly released data of individuals who were tested for SARS-CoV-2 via RT-PCR assay of a nasopharyngeal swab [9]. The dataset contains initial records, on a daily basis, for all citizens tested for COVID-19 nationwide. In addition to the test date and result, various information is available, including clinical symptoms, gender and a binary indication as to whether the tested individual is above age 60 years. Based on this data, we developed a model that predicts COVID-19 test results using eight features: gender, whether age is above 60, known contact with an infected individual, and five initial clinical symptoms.

The training set consisted of records from 51,831 tested individuals (of whom 4,769 were confirmed COVID-19 cases), from the period March 22th, 2020 through March 31st, 2020. The test set contained data from the following week, April 1nd through April 7th (47,401 tested individuals of whom 3,624 are confirmed COVID-19 cases).

The following list describes each of the features used by the model:

A. Basic information:

   1. Gender (male/female).
   2. Age ≥ 60 (true/false)

B. Symptoms:

3. Cough (true/false).
4. Fever (true/false).
5. Sore throat (true/false).
6. Shortness of breath (true/false).
7. Headache (true/false).

C. Other information:

8. Known contact with a confirmed COVID-19 individual (true/false).

Table 1 Characteristics of the dataset and the features used by the model in this study.

(#) Feature	Total n = 99232		COVID-19 negative n = 90839		COVID-19 positive n = 8393	
	n	%	n	%	n	%
(1) Gender						
Male	50350	50.74	45545	50.1	4805	57.2
Female	48882	49.26	45294	49.8	3588	42.7
(2) Age 60+						
True	15279	15.4	13619	14.9	1660	19.7
False	83953	84.6	77220	85	6733	80.2
(3) Cough						
True	14768	14.88	10715	11.8	4053	48.2
False	84223	84.87	79909	87.9	4314	51.4
(4) Fever						
True	8122	8.18	4387	4.83	3735	44.5
False	90868	91.5	86237	94.9	4631	55.1
(5) Sore throat						
True	1273	1.28	96	0.11	1177	14
False	95062	95.8	88059	96.9	7003	83.4
(6) Shortness of breath						
True	930	0.94	71	0.08	859	10.2
False	95405	96.14	88084	96.9	7321	87.2
(7) Headache						
True	1799	1.81	68	0.07	1731	20.6
False	94536	95.27	88087	96.9	6449	76.8
(8) Known contact with a confirmed COVID-19 case						
True	5507	5.55	1455	1.6	4052	48.2
False	93725	94.45	89384	98.4	4341	51.8

Statistical Analysis

Predictions were generated using a gradient-boosting machine model built with decision-tree base-learners [10]. Gradient boosting is widely considered state of the art in predicting tabular data [11] and is used by many successful algorithms in the field of machine learning [12]. As suggested by previous studies [13], missing values were inherently handled by the gradient-boosting predictor [14]. We used the gradient-boosting predictor trained with the LightGBM [15] Python package.

To identify the principal features driving model prediction**,** SHAP (SHapley Additive exPlanations) values [16] were calculated. These values are suited for complex models such as artificial neural networks and gradient-boosting machines [17]. Originating in game theory, SHAP values partition the prediction result of every sample into the contribution of each constituent feature value. This is done by estimating the difference between models with subsets of the feature space. By averaging across samples, SHAP values estimate the contribution of each feature to overall model predictions.

3 Results

For the prospective tests set, the model predicted with 0.90 auROC (area under the receiver operating curve) with 95% CI: 0.892-0.905 (Figure 1.a). Possible working points are: 87.3% sensitivity and 72% specificity, or 85.7% sensitivity and 79% specificity.

Our framework also provides ranking of the most important features that were used to define the decisions (Figure 1.b). Presenting with fever and cough were key features in predicting contraction of the disease. As expected, close contact with a confirmed COVID-19 individual was also an important feature, thus corroborating the disease's high transmissibility [18]. In addition, 'male' gender was revealed as a predictor of a positive result by the model, concurring with the observed gender bias [19,20].

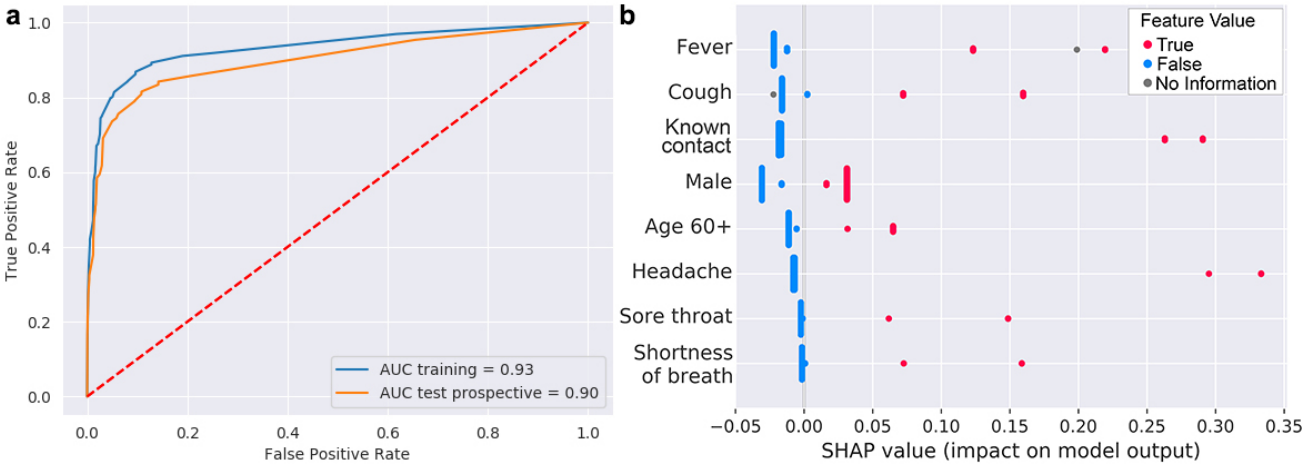


Figure 1. a. ROC curves of the predictive model. The blue line reflects training and testing via cross-validation. The orange line reflects testing the model on the prospective dataset. **b. SHapley Additive exPlanations** (SHAP) summary plots for COVID-19 diagnosis prediction show the SHAP values for the most important features of the model. Features in the summary plots (y-axis) are organized by their mean absolute SHAP values (x-axis), which represent the importance of that feature in driving the classifier's prediction. Values of those features for each patient (i.e. fever) are colored by their relative value.

4 Discussion

This research is not without shortcomings. We relied on the data reported by the Israeli Ministry of Health, which has limitations and biases. For instance, symptom reporting was more comprehensive in the positive test result group and validated with a directed epidemiological effort [21]. This can be reflected by the percentage of COVID-19 positive patients from the overall individuals positive for each symptom, with which we identified features with biased reporting (headache 96.2%, sore throat 92.3% and shortness of breath 92.4%) and symptoms with balanced reporting (cough 27.4% and fever 45.9%). We should also note that all symptoms were self-reported, and a negative value for a symptom can also mean that the symptom was not reported. If we train and test our model while filtering out symptoms of high bias in advance, we get an auROC of 0.862 with a slight change in the SHAP summary plot (Supplementary Figure 1).

However, we hope that readers will appreciate the rapid rate at which the pandemic scenario has evolved over the past weeks and understand the limitations of this research while also acknowledging that unusual times call for unusual solutions. We highlight the need for more robust data to complement our framework while also acknowledging the fact that self-reporting of symptoms is always subject to bias. As the COVID-19 pandemic progresses, it is crucial for public

organizations and associations to continue recording and sharing robust data with the scientific community that is eager to contribute to the ongoing scientific effort.

Overall, based on the nationwide data reported by the Israeli Ministry of Health, we developed a model that detects COVID-19 cases by simple features accessed by asking eight basic questions. Our framework can be used, among other considerations, to prioritize testing for COVID-19 when allocating limited testing resources.

Acknowledgements

We thank Professor David Gurwitz, Shomron lab members Artem Danilevsky and Guy Shapira for their comments on this work. Y.Z. is partially supported by the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University.

Funding

There is **NO** Competing Interest.

No external funding was received for this project.

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*. Published online February 19, 2020. doi:10.1016/S1473-3099(20)30120-1

2. Gozes O, Frid-Adar M, Greenspan H, et al. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis. *arXiv e-prints*. 2020;2003:arXiv:2003.05037. Accessed May 4, 2020. http://adsabs.harvard.edu/abs/2020arXiv200305037G

3. Song Y, Zheng S, Li L, et al. Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images. *medRxiv*. Published online February 25, 2020:2020.02.23.20026930. doi:10.1101/2020.02.23.20026930

4. Wang S, Kang B, Ma J, et al. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *medRxiv*. Published online April 24, 2020:2020.02.14.20023028. doi:10.1101/2020.02.14.20023028

5. Jin C, Chen W, Cao Y, et al. Development and Evaluation of an AI System for COVID-19 Diagnosis. *medRxiv*. Published online March 27, 2020:2020.03.20.20039834. doi:10.1101/2020.03.20.20039834

6. Tostmann A, Bradley J, Bousema T, et al. Strong associations and moderate predictive value of early symptoms for SARS-CoV-2 test positivity among healthcare workers, the Netherlands, March 2020. *Eurosurveillance*. 2020;25(16):2000508. doi:10.2807/1560-7917.ES.2020.25.16.2000508

7. Feng C, Huang Z, Wang L, et al. A Novel Triage Tool of Artificial Intelligence Assisted Diagnosis Aid System for Suspected COVID-19 pneumonia In Fever Clinics. *medRxiv*. Published online March 20, 2020:2020.03.19.20039099. doi:10.1101/2020.03.19.20039099

8. The Novel Coronavirus - Israel Ministry of Health. Accessed May 2, 2020. https://govextra.gov.il/ministry-of-health/corona/corona-virus-en/

9. COVID-19 - Government Data. Accessed May 2, 2020. https://data.gov.il/dataset/covid-19

10. Hastie T, Tibshirani R, Friedman J. Boosting and Additive Trees. In: Hastie T, Tibshirani R, Friedman J, eds. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer; 2009:337-387. doi:10.1007/978-0-387-84858-7_10

11. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*. 2014;15(90):3133-3181. Accessed April 30, 2020. http://jmlr.org/papers/v15/delgado14a.html

12. Omar KBA. XGBoost and LGBM for Porto Seguro ' s Kaggle challenge : A comparison Semester Project. In: ; 2018.

13. Josse J, Prost N, Scornet E, Varoquaux G. On the consistency of supervised learning with missing values. *arXiv:190206931 [cs, math, stat]*. Published online March 25, 2019. Accessed April 30, 2020. http://arxiv.org/abs/1902.06931

14. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. Association for Computing Machinery; 2016:785–794. doi:10.1145/2939672.2939785

15. Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.; 2017:3146–3154. Accessed April 30, 2020. http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf

16. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. *arXiv:170507874 [cs, stat]*. Published online November 24, 2017. Accessed April 30, 2020. http://arxiv.org/abs/1705.07874

17. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*. 2018;2(10):749-760. doi:10.1038/s41551-018-0304-0

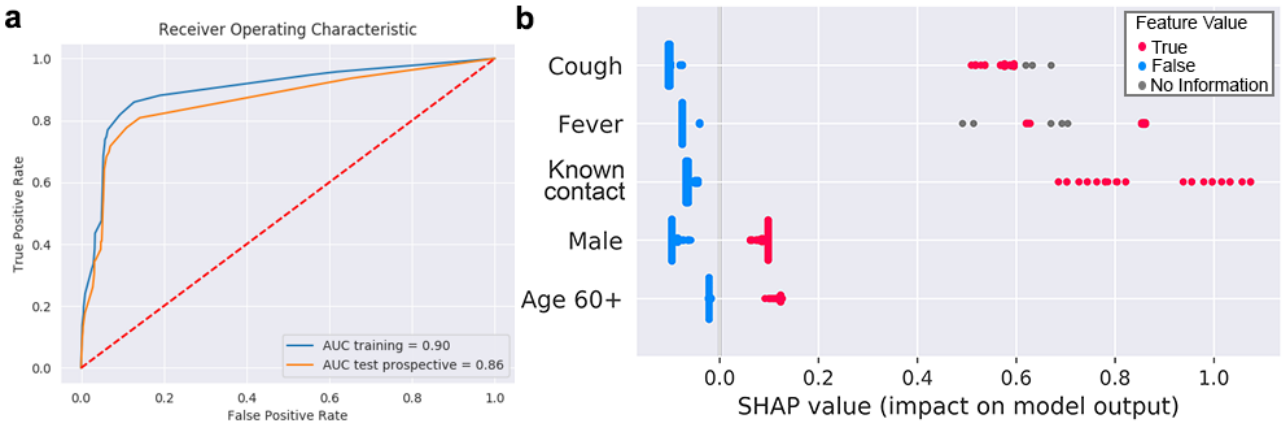
18. Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J Travel Med*. 2020;27(2). doi:10.1093/jtm/taaa021

19. Jin J-M, Bai P, He W, et al. Gender Differences in Patients With COVID-19: Focus on Severity and Mortality. *Front Public Health*. 2020;8. doi:10.3389/fpubh.2020.00152

20. BMJ GH Blogs. Sex, gender and COVID-19: Disaggregated data and health disparities. BMJ Global Health blog. Published March 24, 2020. Accessed May 8, 2020. https://blogs.bmj.com/bmjgh/2020/03/24/sex-gender-and-covid-19-disaggregated-data-and-health-disparities/

21. COVID-19 - Government Data Information. Accessed May 2, 2020. https://data.gov.il/dataset/covid-19/resource/3f5c975e-7196-454b-8c5b-ef85881f78db/download/-readme.pdf

Supplementary Information



Supplementary Figure 2 a. ROC curves and b. SHAP summary plots for training and testing using only balanced features.