JCRES
Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 4- Issue 2, Paper 12, July 2021

MACHINE LEARNING MODEL TO PREDICT COVID-19 DIAGNOSIS BASED ON SYMPTOMS

OmChavan, Aleem Tajir,
Department of Computer Engineering,
M. H. SabooSiddik College of Engineering, Mumbai, India

Abstract*— Coronavirus is a virus which may cause illness in humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered covid causes coronavirus disease (COVID-19). It is difficult to handle the large amount of data of the patients. It's easier to handle this data through ML algorithms. There are tons of procedures for the treatment of multiple diseases across the planet . Machine Learning is an arising approach that helps in the prediction, diagnosis of a disease. This paper depicts the prediction of disease supported symptoms using machine learning. Machine Learning algorithm (Logistic Regression) is employed on the provided dataset and predict the disease. Its implementation is completed through the python programing language . The research indicates the best algorithm based on its accuracy. The accuracy of an algorithm is inferred by the performance of the given dataset.*

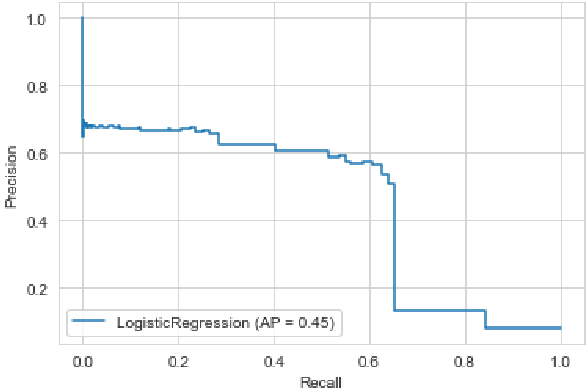
I. INTRODUCTION

2019 Novel Coronavirus (2019-nCoV) may be a virus (more specifically, a coronavirus) identified because the explanation for an epidemic of respiratory disease first detected in Wuhan, China. Early on, many of the patients within the outbreak in Wuhan, China reportedly had some link to an outsized seafood and animal market, suggesting animal-to-person spread. However, a growing number of patients reportedly haven't had exposure to animal markets, indicating person-to-person spread is happening . At this time, it's unclear how easily or sustainably this virus is circulating between people. The Israeli Ministry of Health publicly published data of all individuals who were tested for SARS-CoV-2 via RT-PCR

assay of a nasopharyngeal swab. During the primary months of the COVID-19 pandemic in Israel, all diagnostic laboratory tests for COVID-19 were performed consistent with criteria determined by the Israeli Ministry of Health. While subject to vary , the standards implemented during the study period included the presence and severity of clinical symptoms, possible exposure to individuals confirmed to possess COVID-19, certain geographical areas, and therefore the risk of complications if infected. apart from alittle minority who were tested under surveys among healthcare workers, all the individuals tested had indications for testing. Thus, there was no apparent referral bias regarding the overwhelming majority of the themes within the dataset utilized in this study; this contrasts with previous studies, that such bias was a drawback. additionally , all negative and positive COVID-19 cases this dataset were confirmed via RT-PCR assay.

In this paper, we indicate a machine-learning model that predicts a positive SARS-CoV-2 infection during a RT-PCR test by asking eight basic questions. The model was trained on data of all individuals in Israel tested for SARS-CoV-2 during the primary months of the COVID-19 pandemic. Thus, our model are often implemented globally for effective screening and prioritization of testing for the virus within the general population.

II. Results

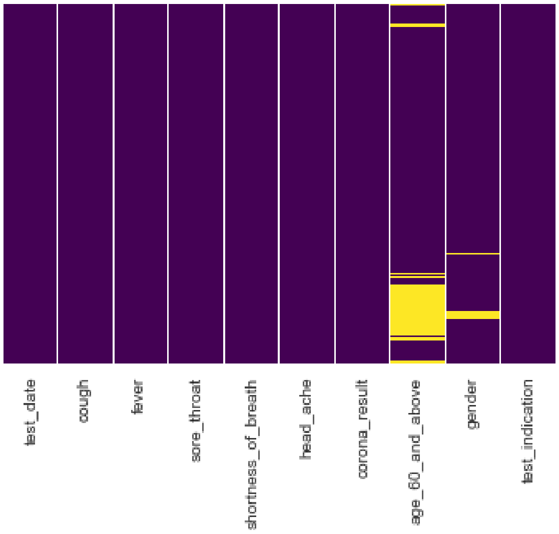


The data that were reported by the Israeli Ministry of Health has boundaries and biases. as an example , symptom reporting was more comprehensive among those that tested positive for COVID-19, and validated with a directed epidemiological effort13. Thus, mislabeling of symptoms among those that tested negative for COVID-19 is predicted . this is often reflected within the proportion of persons who were COVID-19 positive from the entire number of people who were positive for every symptom. Accordingly, we identified aspects with biased reporting (headache 96.2%, pharyngitis 92.3% and shortness of breath 92.4%) and symptoms with balanced reporting (cough 27.4% and fever 45.9%). Mislabeling of symptoms can also arise from an understatement and underreporting of symptoms among persons who tested negative.

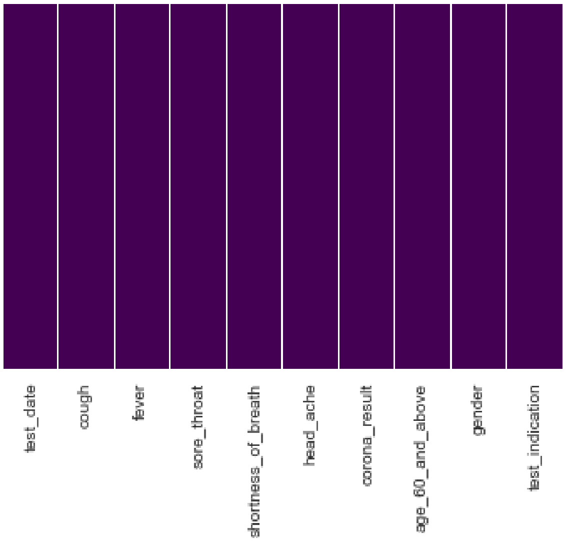
III. Discussion

This research isn't without shortcomings. We relied on the info reported by the Israeli Ministry of Health, which has limitations, biases and missing information regarding a number of the features. for instance , for patients labeled as having had contact with an individual confirmed to possess COVID-19, additional information like the duration and site (indoors/outdoors) of the contact wasn't available. Some symptoms (such as lack of smell and taste) were recognized as being very predictive of a COVID-19 infection by previous studies19, but weren't recorded by the Israeli Ministry of Health. We showed that training and testing a model while filtering out symptoms of high bias beforehand still achieved very high accuracy. We also note that each one the symptoms were self-reported and a negative value for a symbol might mean that the symptom wasn't reported. it's therefore important to assess the model’s performance within the circumstance that more values are unreported or missing instead of with negative values.

We highlight the necessity for more robust data to enrich our framework, while also acknowledging that self-reporting of symptoms is usually subject to bias. Because the COVID-19 pandemic progresses, ongoing recording and sharing of strong data between public organizations and therefore the scientific community are crucial. In parallel to increasing understanding of the contribution of varied symptoms to diagnosing the disease, additional symptoms could be integrated into future models.



Missing_Data1



Data_After_Cleaning1

In conclusion, supported nationwide data reported by the Israeli Ministry of Health, we developed a model for predicting COVID-19 diagnosis by asking eight basic questions. Our frameworks are often used, among other considerations, to prioritize testing for COVID-19 when testing resources are limited. additionally , the methodology presented during this study may benefit the health system response to future epidemic waves of this disease and of other respiratory viruses generally .

IV. Methods

The Israeli Ministry of Health publicly released data of people who were tested for SARS-CoV-2 via RT-PCR assay of a nasopharyngeal swab11. The dataset contains initial records, on a day to day , of all the residents who were tested for COVID-19 nationwide. Additionally to the test date and result, various information is out there , including clinical symptoms, sex and a binary indication on whether the tested individual is aged 60 years or above. supported these data, we formulated a model that predicts COVID-19 test results using eight binary features: sex, age 60 years or above, known contact with an infected individual, and five initial clinical symptoms.

The training-validation set consisted of records from 51,831 tested individuals (of whom 4769 were confirmed to possess COVID-19), from the amount March 22th, 2020 through March 31st, 2020. The test set contained data from the next week, April 1st through April 7th (47,401 tested individuals, of whom 3624 were confirmed to possess COVID-19).

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

A. Basic information:
   a) Sex (male/female).
   b) Age ≥60 years (true/false)

B. Symptoms:
   c) Cough (true/false).
   d) Fever (true/false).
   e) Sore throat (true/false).
   f) Shortness of breath (true/false).
   g) Headache (true/false).

C. Other information:
   h) Known contact with an individual confirmed to have COVID-19 (true/false).

V. Development of the model

Predictions were generated using a Logistic Regression model. Logistic Regression is widely considered state of the art in predicting binary regression data and is used by many successful algorithms in the field of machine learning. As suggested by previous studies, missing values were inherently handled by the seaborn. We used Seaborn to find the missing values.

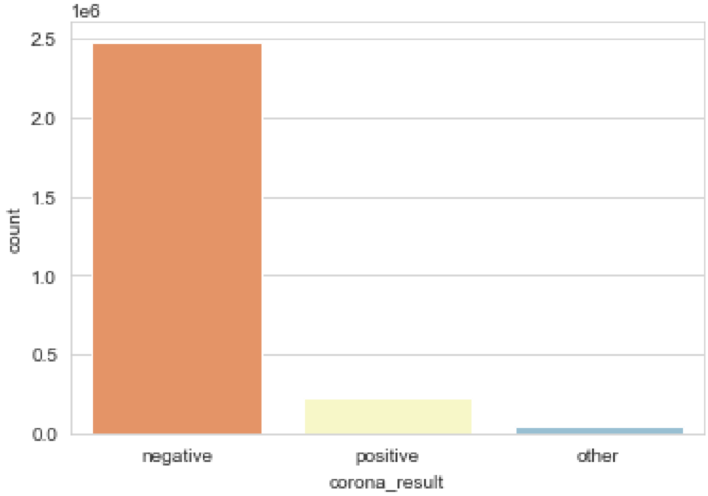
VI. Evaluation of the model

The model was scored on the test set using the auROC. In addition, plots of the PPV against the sensitivity (precision–recall curve) were drawn across different thresholds. Metrics were calculated for all the thresholds from all the ROC curves, including sensitivity, specificity, PPV and negative predictive value, false-positive rate, false-negative rate, false discovery rate and overall accuracy.

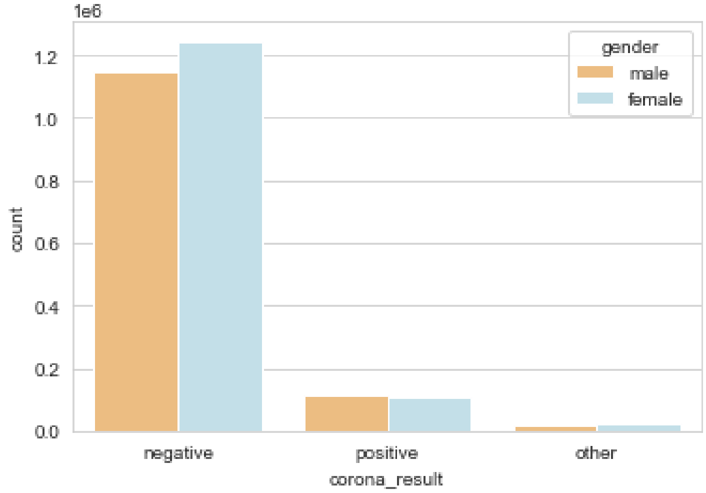
Journal of Current Research in Engineering and Science
JCRES
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 4- Issue 2, Paper 12, July 2021

VII. FIGURES



Testing_Result1



Testing_Result2

REFERENCES

[1] Feng, C. et al. A novel triage tool of artificial intelligence assisted diagnosis aid system for suspected COVID-19 pneumonia in fever clinics. *medRxiv*, https://doi.org/10.1101/2020.03.19.20039099 (2020).

[2] Punn, N. S., Sonbhadra, S. K. & Agarwal, S. COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms. *medRxiv*, https://doi.org/10.1101/2020.04.08.20057679 (2020). 10. Mei, X. et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nat. Med*. **26**, 1224–1228 (2020).

[3] COVID-19-Government Data. https://data.gov.il/dataset/covid-19 (2020).

[4] The Novel CoronavirusIsrael Ministry of Health. https://govextra.gov.il/ministry-of-health/corona/corona-virus-en/ (2020).

[5] COVID-19-Government Data Information. https://data.gov.il/dataset/covid-19/resource/3f5c975e-7196-454b-8c5b-ef85881f78db/download/-readme.pdf (2020).

[6] Struyf, T. et al. Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has COVID-19 disease. *Cochrane Database Syst. Rev.*, https://doi.org/10.1002/14651858.CD013665 (2020).

[7] Liu, Y., Gayle, A. A., Wilder-Smith, A. &Rocklöv, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J. Travel Med.* **27** (2020).

[8] Jin, J.-M. et al. Gender Differences in Patients With COVID-19: Focus on Severity and Mortality. *Front. Public Health* **8** (2020).

[9] BMJ GH Blogs. Sex, gender and COVID-19: Disaggregated data and health disparities. *BMJ Global Health blog* https://blogs.bmj.com/bmjgh/2020/03/24/sex-gender-and-covid-19-disaggregated-data-and-health-disparities/ (2020).

[10] Whittington, A. M. et al. Coronavirus: rolling out community testing for COVID-19 in the NHS. *BMJ Opinion* https://blogs.bmj.com/bmj/2020/02/17/coronavirus-rolling-out-community-testing-for-covid-19-in-the-nhs/ (2020).

[11] Menni, C. et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat. Med.* **26**, 1037–1040 (2020).

[12] Hastie, T., Tibshirani, R. & Friedman, J. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (eds. Hastie, T., Tibshirani, R. & Friedman, J.) 337–387 (Springer, 2009).

[13] Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15**, 3133–3181 (2014).

[14] Omar, K. B. A. *XGBoost and LGBM for Porto Seguro's Kaggle challenge: A comparison Semester Project* (ETH Zurich, 2018).

[15] Josse, J., Prost, N., Scornet, E. &Varoquaux, G. On the consistency of supervised learning with missing values. *arXiv:1902.06931 [cs, math, stat]* (2019).

[16] 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* 785–794 (Association for Computing Machinery, 2016).

[17] Ke, G. et al. In *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 3146–3154 (Curran Associates, Inc., 2017).

[18] Raskutti, G., Wainwright, M. J. & Yu, B. Early stopping for non-parametric regression: An optimal data-dependent stopping rule. in *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* 1318–1325 (2011).

[19] Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]* (2017).