A Machine Learning Approach in Evaluating Symptom Screening in Predicting COVID-19

John Althom A. Mendoza[1], Geoffrey A. Solano[2], Marc Jermaine Pontiveros[1,2], Jaime DL Caro[1], Peter Martin D. Gomez[4], Conner G. Manuel[4,5], Paulyn Jean Buenaflor Rosell-Ubial[5], Michael Tee[3,5]
[1]Department of Computer Science, College of Engineering
University of the Philippines Diliman, Philippines
[2]Department of Physical Sciences and Mathematics, College of Arts and Sciences
University of the Philippines Manila, Philippines
[3]Department of Physiology, College of Medicine
University of the Philippines Manila, Philippines
[4]Dashlabs.ai, Manila, Philippines
[5]Philippine Red Cross, Manila, Philippines

***Abstract*—COVID-19 is a disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that, to date, has over 245 million confirmed cases and claimed almost 5 million lives. This disease attacks the respiratory system and comes with a number of symptoms. The US Center for Disease Control and Prevention presents a set of symptoms. However, these symptoms only begin to manifest after a number of days, which prevents early detection of this disease. This absence of symptoms during the early stages is what is considered by many to be the very factor that caused the virus into becoming a pandemic. Nonetheless, symptoms checking has been used in practice by commercial and business establishments as an initial screening for COVID-19. The bothersome process of symptom checking are still in place at the entrances of malls and airports. In this study, we determine whether or not symptom screening is an effective system to be employed to assess individuals for COVID-19. Specifically, it aims to determine whether or not one or a set of symptoms are effective predictors of the RT-PCR test results, the gold standard in Covid-19 testing, using machine learning. Using data from the Philippine Red Cross, classification models are developed using LightGBM, AdaBoost, Gaussian Naïve-Bayes, MultiLayer Perceptron, Quadratic Discriminant Analysis and Decision Tree. These models were evaluated using the following metrics: precision, sensitivity, specificity and the type II error rate. Furthermore, for explainability, symptoms are analyzed as to whether or not they are relatively influential on the predicting whether or not a patient has COVID-19. The high type II error rate, low sensitivity and low relative predictor scores of the most significant predictor symptoms clearly show that symptoms do not correlate with the RT-PCR testing results. Thus, we conclude that symptom screening is not a medically suitable process for determining whether an individual has COVID-19. In fact, it even exposes us to the risk of viral transmission as people congregate at the entrances and lobbies of establishments.**

***Index Terms*—symptom screening, machine learning, COVID-19**

I. Introduction

It was in December 2019 when the coronavirus disease 2019 (COVID-19), a condition caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was first identified in the capital of China's Hubei province of Wuhan. In just a few weeks it has swept across the globe. On the 30th of January, 2020, the World Health Organization (WHO) declared the outbreak to be a Public Health Emergency of International Concern. By March 11, 2020, it was declared by WHO as a pandemic [1], [2]. As of the 29th of October, 2021, there are over 245 million confirmed cases and almost 5 million deaths worldwide spanning 222 countries and territories [3], with the Philippines having a total of 2.7 million confirmed cases and over 42 thousand deaths [4].

Various testing types were used to detect COVID-19, the most reliable of which, is the reverse transcriptase-polymerase reaction (RT-PCR) nasopharyngeal (NP) swab testing. It works by identifying viral RNA and is currently the gold standard in detecting whether an individual has COVID-19. One major challenge however is the cost of the procedure and the accessibility of testing centers, especially in less-developed countries. Furthermore, this type of testing has the biggest drawback of having the results available at least a day after the samples were collected. Thus, several faster tests are being administered despite having low confidence on the results. Such tests include lung function testing (LFT) [5], saliva testing [6], and blood testing [7].

Since this disease attacks the respiratory system and comes with a number of symptoms, many have resorted to symptom checking has been used as an initial screening for COVID-19. The US Center for Disease Control and Prevention presents the following, among others, as symptoms which may appear 2-14 days after exposure to the virus, ranging from mild to severe cases [8]:

- Fever or chills
- Cough
- Shortness of breath or difficulty breathing
- Fatigue
- Muscle or body aches
- Headache
- New loss of taste or smell
- Sore throat
- Congestion or runny nose

- Nausea or vomiting
- Diarrhea

All over the globe, airlines have required Health Declaration Forms (HDFs) to be filled out by those who will be taking flights. This is essentially declaring whether or not one is manifesting any of the above-mentioned symptoms. Companies have also employed the use of online symptom self-checking systems for their employees. Commercial and business establishments have required individuals to fill out an online health declaration form prior to entering the building premises. The bothersome process of symptom checking are still in place at the entrances of malls and airports. These have resulted in queues at the building entrance. Aside from the hassle and delays, this brings the risk of exposure to the virus due to human aggregation. Furthermore, these symptoms only begin to manifest 2-14 days after exposure, which prevents early detection of this disease [2], [9], [10]. In fact, this absence of symptoms during the early stages is what is considered by many to be a major factor that caused the virus into becoming a pandemic [7].

All these bring about the question of whether symptoms checking is indeed an effective means for detecting COVID-19. This question is what is investigated in this paper. This study aims to determine whether or not one or a set of symptoms are effective predictors of the RT-PCR test results using machine learning. Using data from the Philippine Red Cross, classification models are developed and symptoms will be analyzed as to whether or not they are relatively influential on the predicting whether a patient has COVID-19. The paper is organized as follows: On the next section we explore related work. This is followed by the Methodology where we present the series of steps were taken in order to accomplish the objectives of this study. In Section IV and V, we present the results of this study and the discussion of these results, respectively. Finally, the conclusion and suggestions for future work are in Section VI.

II. Related Work

Many supervised learning and feature extraction approaches have been used focusing on COVID-19 with different objectives. In a review by Alyasseri, et al., some of the objectives are: (1) to determine how the COVID-19 pandemic will end, (2) to predict how the coronavirus gets transmitted over regions, (3) to correlate the effect of weather conditions on coronavirus and (4) to diagnose COVID-19 based on symptoms and various X-ray and CT scan images [11].

Among the reviewed approaches in [11], one study conducted by Fayyoumi et. al have similar focus on diagnosing COVID-19 status based on signs and symptoms using 64 positive and 41 negative PCR tests of patients in Jordan, and reported an accuracy of 91.67% using Multilayer Perceptron (MLP). [12]. In their work, several attributes have been used to build the statistical models, namely age, smoker status, positive x-ray chest, fever, sore throat, aches and pain, dry cough, nasal congestion, absence of smell, diarrhea, vomiting, and breathing difficulty. The data was collected by answering questionnaires.

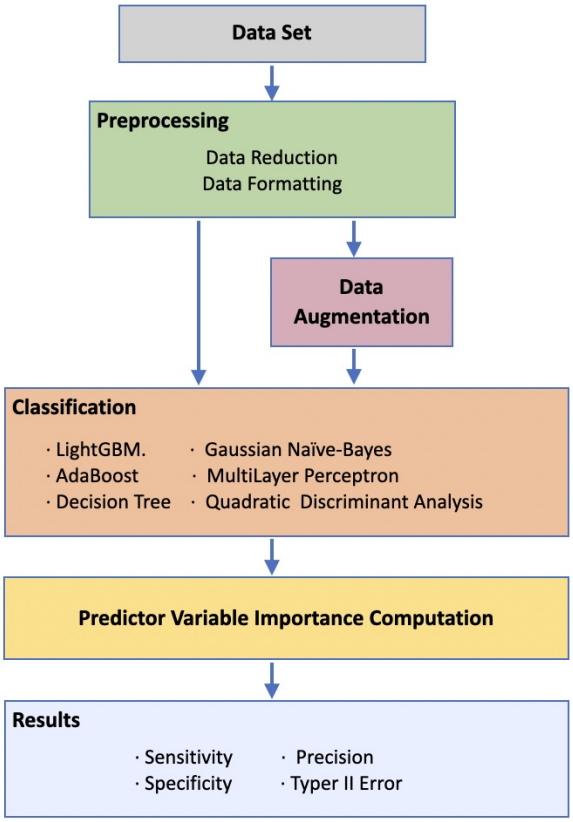

Fig. 1. The general workflow of the study

Among the attributes, only age were considered numeric and the rest are binary categorical variable. The authors mentioned that applying the technique in much larger dataset must be done in future studies.

Another related work by Zoabi et. al utilized a total of 99,232 COVID-19 test results, with a focus of diagnosing COVID-19 based on symptoms [13]. Their model utilizes the following attributes: sex, binarized age (greater than or equal 60 years), known contact with an infected individual, and appearance of five clinical symptoms (cough, fever, sore throat, shortness of breath, and headache). Their data is based on published data by Israel Ministry of Health of individuals who were tested for SARS-CoV-2 via RT-PCR assay of a nasopharyngeal swab. It contains initial records of all the residents who were tested for COVID-19 nationwide on daily basis. The study mentioned some shortcomings, and one would be missing information in some of the features, another would be the lack of records in reported symptoms by the Ministry of Health, such as the loss of smell and loss of taste.

Both studies are affirming the use of signs and symptoms in prioritizing testing and triaging for COVID-19, especially when the resources are limited.

III. Methodology

A series of steps were taken in order to accomplish the objectives of this study. These steps are illustrated in a

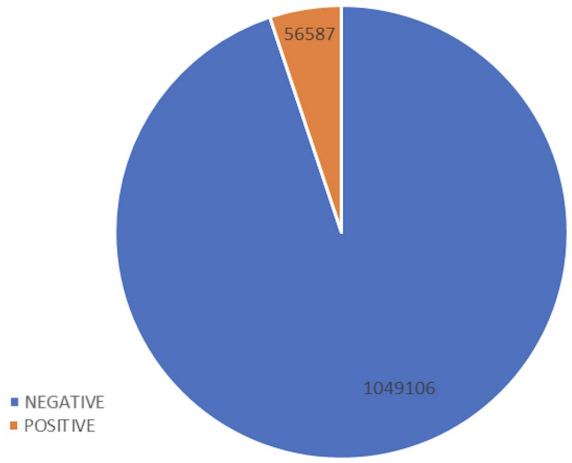

Fig. 2. The ratio of the positive to the negative cases prior to augmentation

flowchart in Figure 1.

A. The Data Set

The data used in the study was collected by the Philippine National Red Cross between June 2020 – January 2021 totaling 1,434,868 records.

B. Preprocessing

1) Reduction: This included observations which were still "in-process", and thus, whose positivity to COVID-19 were not yet identified. These observations were therefore removed and those that remained totaled to 1,105,693 records.

2) Formatting: The dataset was originally in text with each row representing an observation. Part of each row is the positivity value, along with a list of symptoms separated by commas. These were all processed by transforming positivity along with the most frequent symptoms into columns or features with boolean values. Positivity became the target variable with boolean values indicating whether the patient was positive or otherwise. Similarly each of the most frequent symptoms all became features (column) whose boolean values indicated whether that particular patient (row) had that symptom. All the other symptoms that were present in less than ten observations were fused into one feature/column "others", and so if at least one of these were present in an observation, that the value of the column for the said observation is 1 or "true". The summary of frequencies of these features are seen in Table I. Clearly, the fused symptom "others" was the most frequent one, being present in 21,827 cases, whereas cough was present in 17,421 cases. Only 18 cases experienced appetite loss. On the other hand in Table II we see the symptom count on the observations. Around 1,046,932 cases were asymptomatic. There were 46,411 cases who had exactly one symptom and there was one case that had 10 symptoms.

TABLE I
SUMMARY OF FEATURES (SYMPTOMS) AND THEIR RESPECTIVE FREQUENCIES.

Symptom	Count
others	21,827
cough	17,421
colds	15,789
sore throat	6,989
fever	6,741
difficulty breathing	3,651
headache	1,163
smell loss	1,073
taste loss	1,066
body malaise	1,066
diarrhea	644
body pain	539
anosmia	367
sense loss	271
ageusia	172
agnosia	29
apetite loss	18

TABLE II
SUMMARY OF SYMPTOM COUNTS.

Asymptomatic	1,046,932
1	46,411
2	7,884
3	2,514
4	1,057
5	627
6	174
7	63
8	23
9	7
10	1

C. Data Augmentation

It was noticed that with respect to the target variable, the data set has a class imbalanc, with the ratio of those positive to those negative being approximately 1:18.54. This disparity can be seen in Figure 2. Thus, for this step, each of the positive observations was copied 18 times in order for the data set to be balanced. Thus, after augmentation, there were 1,018,566 positive cases.

Furthermore, there are therefore two sets of data used in the succeeding sections of this study : one with data augmentation and one without data augmentation.

TABLE III
PERFORMANCE OF CLASSIFICATION MODELS (WITHOUT AUGMENTATION)

Classifier	Precision	Type II Error	Sensitivity	Specificity
LightGBM	0.5862	0.4138	0.4215	0.6949
AdaBoost	0.5417	0.4583	0.3223	0.7203
GaussianNaiveBayes	0.5000	0.5000	0.9339	0.0424
MultiLayerPerceptron	0.4861	0.5139	0.2893	0.6864
QuadraticDiscriminant	0.5063	0.4937	1.0000	0.0000
DecisionTree	0.1818	0.8182	0.1322	0.3898

TABLE IV
PERFORMANCE OF CLASSIFICATION MODELS (WITH AUGMENTATION)

Classifier	Precision	Type II Error	Sensitivity	Specificity
LightGBM	0.7521	0.2479	0.0921	0.9697
AdaBoost	0.7521	0.2479	0.0921	0.9697
GaussianNaiveBayes	0.7521	0.2479	0.0921	0.9697
MultiLayerPerceptron	0.7572	0.2428	0.0902	0.9711
QuadraticDiscriminant	0.7521	0.2479	0.0921	0.9697
DecisionTree	0.7536	0.2464	0.0921	0.9699

D. Development of Classification Models

There were six machine learning models developed for both working data sets. These are LightGBM, AdaBoost, Gaussian Naïve-Bayes, MultiLayer Perceptron, Quadratic Discriminant Analysis and Decision Tree.

E. Computation of Predictor Variable Importance Scores

One important task in interpreting classification models is understanding which predictor variables are relatively influential on the predicted outcome. Thus, aside from measuring evaluation metrics, variable importance was determined for both data sets. This variable importance measure was computed via permutation which included the following steps:

```
For any given loss function do
1: compute loss function for full model (denote _full_model_)
2: randomize response variable, apply given ML, and compute loss function
3: for variable j
     | randomize values
     | apply given ML model
     | compute & record loss function
   end
```

This model agnostic variable importance measure computed via permutation [14] is essential in the explainability of the classification models developed.

F. Evaluation of Classification Models

For both the augmented and non-augmented data sets, the six models were evaluated using the following metrics : precision (positive predictive value), sensitivity (true positive rate or probability of detection), specificity (true negative rate) and the type II error rate (false negative rate or error of omission). Type II error rate is also chosen as the main metric since this is a health science study.

IV. RESULTS

In Figure 3 we see the importance of the variables and how they fare with each other with respect to determining the target variable for the data set where augmentation was not applied. *Smell loss* outperforms the rest at 3.7%, followed by *fever* (1.4%) and *colds* (1.1%). Also, the features *body malaise, ageusia* and *others* do not seem to be contributing as good predictors of COVID-19 positivity. In Table III we see the comparison of the performance of the models on the non-augmented data set based on the metrics that were used in this study. LightGBM performed best with respect to Precision and type II error. The quadratic discriminant provides 100% sensitivity while Gausian Naive Bayes provides 93.4%. However, their specificity results are 0% and 4.24%, respectively. Adabost, on the other hand, provides the highest specificity at 72.03%. However, all the six models for the non-augmented data have a very high type II error at $\geq 41.38\%$.

On the other hand, for the augmented data set, we can see the variable importance in Figure 4. Clearly, all of the variables performed better with *smell loss* still leading at 13.7%, followed by *fever* (13%) and *colds* (12.4%). *Sense loss, agnosia* and *others* appeared to be the least important variables. In Table IV we see the comparison of the performance of the models on the augmented data set based on the metrics that mentioned. Multilayer perceptrons performed better than the rest with respect to precision, type II error and specificity at 75.72%, 24.28% and 97.11% respectively. However, it dipped a bit compared to the other models with sensitivity. The Decision Tree model, however, was just second to Multilayer perceptrons with respect to precision, type II error and specificity at at 75.36%, 24.64% and 96.99% respectively, but it performed slightly better with sensitivity.

V. DISCUSSION

It can be observed that augmenting the data generally helps in achieving better models. One possible reason is that training biases over the uneven distribution of the observations are solved. In non-augmented data, none of the symptoms has an importance of more than 5%. However, after augmentation, 11 symptoms obtained variable importance which were higher than 5%, with *smell loss* leading with an improvement from 3.7% to 13.7%. Furthermore, all the models improved after data augmentation with three of the four metrics. However, the sensitivity scores in the augmented data are quite curious as all the models scored very low at sensitivity, with Multilayer perceptrons scoring only 9.02% and the rest of the models scoring 9.21%.

The most significant symptoms are *smell loss*, *fever* and *colds*. However, their scores as relative predictors for the target variable are still low at at 13.7%, 13% and 12.4%, respectively.

It can be observed based on the results that COVID-19 cannot be predicted accurately with just the symptoms. The

two better models are Multilayer perceptrons and Decision Trees. Multilayer perceptrons had precision, type II error and specificity at 75.72%, 24.28% and 97.11% respectively. However, the sensitivity score is at 9.02% only. Decision Trees on the other hand had precision, type II error and specificity scores of 75.36%, 24.64% and 96.99% respectively, but the sensitivity score is at 9.21% only. Furthermore, all the models still had a high Type II error, which is a big issue for health prediction classifier models.

Given the hightype II error, low sensitivity and low relative predictor scores of the most significant predictor symptoms, thus there are no symptoms, whether one or a set, that is an effective predictor of the RT-PCR test results based on the data set. This study finds that symptom screening is not an effective system to be employed to monitor and assess individuals for COVID-19.

VI. Conclusion

In this study, data from the Philippine Red Cross was used to determine whether or not symptom screening is an effective system to be employed to assess individuals for COVID-19. Classification models were developed using LightGBM, AdaBoost, Gaussian Naïve-Bayes, MultiLayer Perceptron, Quadratic Discriminant Analysis and Decision Tree and were evaluated using the following metrics: precision, sensitivity, specificity and the type II error rate. Furthermore, for explainability, symptoms were analyzed as to whether or not they are relatively influential on the predicting whether or not a patient has COVID-19. Across all models, the high type II error rate ($\geq 24.28\%$), low sensitivity ($\leq 9.21\%$) and low relative predictor scores of the most significant predictor symptoms ($\leq 13.7\%$) clearly there are no symptoms, whether one or a group of symptoms, that is an effective predictor of the RT-PCR test results based on the Red Cross data set. Thus, we conclude that symptom screening is not a medically suitable process for determining whether an individual has COVID-19. In fact, it even exposes us to the risk of viral transmission as people congregate at the entrances and lobbies of establishments.

Acknowledgement

The authors would like to thank the Philippine Red Cross, particularly Senator Richard J. Gordon and Ms. Elizabeth Zavalla.

References

[1] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, "A novel coronavirus outbreak of global health concern," *The Lancet*, vol. 395, pp. 470–473, 2020. [Online]. Available: https://doi.org/10.1016/S0140-6736(20)30185-9

[2] S. Chauhan, "Comprehensive review of coronavirus disease 2019 (covid-19)," *Biomedical Journal*, vol. 43, no. 4, pp. 334–340, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2319417020300871

[3] "Who coronavirus (covid-19) dashboard." [Online]. Available: https://covid19.who.int/

[4] "Covid-19 tracker: Department of health website." [Online]. Available: https://doh.gov.ph/covid19tracker

[5] J. Hull, J. Lloyd, and B. Cooper, "Lung function testing in the covid-19 endemic," *The Lancet Respiratory Medicine*, vol. 8, 05 2020.

[6] L. M. Czumbel, S. Kiss, N. Farkas, I. Mandel, A. Hegyi, A. Nagy, Z. Lohinai, Z. Szakacs, P. Hegyi, M. C. Steward, and G. Varga, "Saliva as a candidate for covid-19 diagnostic testing: A meta-analysis," *Frontiers in Medicine*, vol. 7, p. 465, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmed.2020.00465

[7] J. Giesecke, "The invisible pandemic," *The Lancet*, vol. 395, no. 10238, May 2020.

[8] "Center for disease control and prevention: Symptoms of covid-19." [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

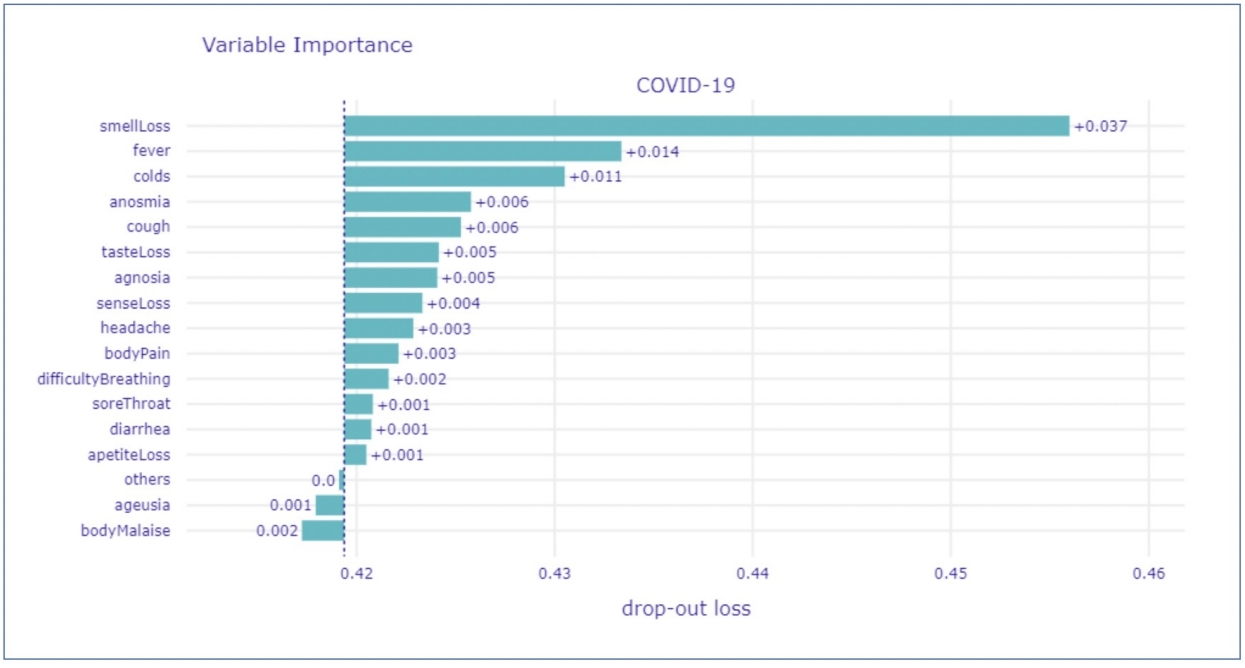

Fig. 3. Variable importance (without augmentation)

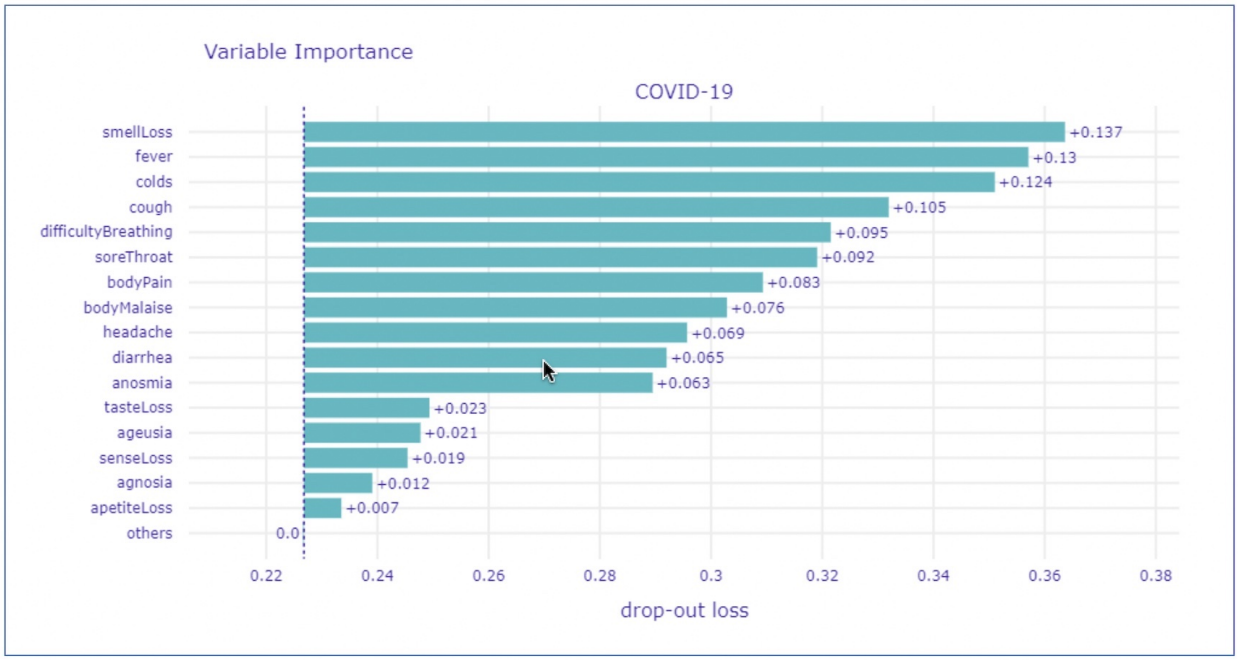

Fig. 4. Variable importance (with augmentation)

[9] L. Wang, Y. Wang, D. Ye, and Q. Liu, "Review of the 2019 novel coronavirus (sars-cov-2) based on current evidence," *International Journal of Antimicrobial Agents*, vol. 55, no. 6, p. 105948, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924857920300984

[10] R. Wang, M. Pan, X. Zhang, M. Han, X. Fan, F. Zhao, M. Miao, J. Xu, M. Guan, X. Deng, X. Chen, and L. Shen, "Epidemiological and clinical features of 125 hospitalized patients with covid-19 in fuyang, anhui, china," *International Journal of Infectious Diseases*, vol. 95, pp. 421–428, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1201971220302034

[11] Z. A. A. Alyasseri, M. A. Al-Betar, I. A. Doush, M. A. Awadallah, A. K. Abasi, S. N. Makhadmeh, O. A. Alomari, K. H. Abdulkareem, A. Adam, R. Damasevicius, M. A. Mohammed, and R. A. Zitar, "Review on COVID -19 diagnosis models based on machine learning and deep learning approaches," Jul. 2021. [Online]. Available: https://doi.org/10.1111/exsy.12759

[12] E. Fayyoumi, S. Idwan, and H. AboShindi, "Machine learning and statistical modelling for prediction of novel COVID-19 patients case study: Jordan," vol. 11, no. 5, 2020. [Online]. Available: https://doi.org/10.14569/ijacsa.2020.0110518

[13] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," vol. 4, no. 1, Jan. 2021. [Online]. Available: https://doi.org/10.1038/s41746-020-00372-6

[14] "Model interpretability with dalex." [Online]. Available: https://uc-r.github.io/dalex?fbclid=IwAR2vtKGe6Eht5sDgkHHtiLkXVR88K3OgN0Adbk1cWDLhfGEYEOPGeEf02bovi

[15] T. Zitek, "The appropriate use of testing for covid-19," *Western Journal of Emergency Medicine*, vol. 21, 04 2020.

[16] F. Zeng, L. Li, J. Zeng, Y. Deng, H. Huang, B. Chen, and G. Deng, "Can we predict the severity of coronavirus disease 2019 with a routine blood test?" *Polish archives of internal medicine*, vol. 130, no. 5, May 2020.

[17] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of covid-19 diagnosis based on symptoms," *npj Digital Medicine*, vol. 4, 01 2021.

[18] A. Callahan, E. Steinberg, J. Fries, S. Gombar, B. Patel, C. Corbin, and N. Shah, "Estimating the efficacy of symptom-based screening for covid-19," *npj Digital Medicine*, vol. 3, 12 2020.