Covid 19 diagnosis prediction using Machine Learning

Sunit Kumar Jaiswal
Department of Computer Science and Engineering
BIT, Gorakhpur-273209
India
sunit007mmmec@gmail.com

Sudhir Agarwal
Department of Computer Science and Engineering
BIT, Gorakhpur-273209
India
sagarwal22@bit.ac.in

***Abstract*—SARS--CoV--2 virus, COVID--19, prompted a massive wide variety of deaths and monetary problems. COVID-19 may be detected directly and easily, easing the load on already beaten healthcare systems. With the help of machine-learning algorithm, a version for predicting COVID--19 patients is being advanced, which is trained on 81,312 examined patients, 4430 of whom have been licensed positive. The facts for the subsequent week`s take a look at batch consists of 48,539 those who have been examined, with 4420 of them being discovered to be positive. We used the subsequent 8 binary traits in our version to broaden a version with excessive accuracy: age, sex, touch trace, and 5 scientific symptoms. We need to alleviate the pressure on an already overburdened healthcare system. The facts are primarily based totally on statistics freely to be had in India. Using easy statistics obtained through primary questions, we built a version that detects COVID-19 situations.**

Keywords—ROC, auROC, PPV, COVID-19, SHAP, SARS

I. INTRODUCTION

Corona virus infection was on the rise in 2019(COVID--19) The whole world was on the threat due to this pandemic. In the month of December of this year, an outbreak began within China. After spreading there had been greater than 39,700,000 showed times of the ailment in greater than one hundred eighty nations as of October 2020, with the authentic range of humans inflamed probable significantly higher. Nearly 1,120,000 humans have died because of COVID-19. With powerful screening, COVID-19 may be recognized unexpectedly and efficiently, lowering the weight on healthcare systems. Prediction fashions that integrate a lot of signs to expect the chance of contamination were advanced withinside the hopes of helping clinical practitioners round the arena in triaging sufferers, mainly in nations wherein healthcare sources are limited. CT scans[1,2,4,5,6] symptoms[7], lab tests [9] and a combination of all aspects are included in these styles[10]. The majority of preceding fashions, on the opposite hand, had been primarily based totally on information from hospitalised sufferers and so is inefficient for screening for SARS-CoV-2 withinside the preferred public. We offer an eight-query machine-learning knowledge of version that appropriately forecast a tremendous COVID--19 All SARS-CoV-2 checking out achieved in India in starting months of COVID--19 outbreak had been.

As a result, our generation is probably implemented international for virus screening and prioritisation withinside the popular population. The dilemma or suspension of non-critical sports has been used as an emergency non-pharmaceutical preventative method in numerous nations to decrease each the price of latest infections and the hazard of surpassing health facility capability. Without a doubt, the capacity to directly discover high-threat sufferers and successfully assign health priorities is tough, each for boosting health facility capability making plans and for supplying sufferers with early treatment. Artificial intelligence processes were appeared as a effective and promising generation that may help now no longer handiest in figuring out a affected person`s mortality threat even as looking for scientific attention, however additionally withinside the prognosis process, disorder spreading dynamics prediction, and prognosis process. Machine learning is an synthetic intelligence subfield that attempts to offer computers "gaining knowledge of capability" via way of means of the use of well-described algorithms to enhance overall performance or make correct predictions. These algorithms often examine from formerly to be had information withinside the shape of labelled schooling units. In order to minimise the loss function, supervised learning algorithms use those labelled units to alter the parameters of a statistical version. The skilled version can then make correct predictions with information that become in no way utilised as enter at some point of the learning phase. Naturally, the first-class and amount of the information units used are crucial in making sure the algorithm's right operation. Machine gaining knowledge has been used to increase a couple of algorithms focused at detecting sufferers who're possibly to grow to be inflamed at an early level at some point of the prevailing outbreak.

Copyright @ 2022 Authors

II. APPROACHES

Data Collection And Analysis

The Indian Ministry of Health[11] tracked Individuals who were screened for SARS--CoV--2 by use of a nasopharyngeal swab RT-PCR assay. The proposed model that predicts COVID-19 take a look at outcomes the use of 8 binary characteristics. The training data set was produced using data from 51,832 people who were surveyed between March 2nd and March 9th, 2020. (of whom 4769 have been showed to have COVID-19).The coming weak data, April 1st to 10th, was included in the test set (3603 of the 47,300 individuals who underwent testing had COVID-19 verified.). Using a 5:1 ratio, the training-test set was separated into training and test sets (Table 1)

The following list of attributes of the datasets are used

1. Ary You Male(yes or no)

2. Are you above 60 years (yes or no)?

3. Are you having cough problem (yes or no)?

4. Are you having fever(yes or no)?

5. Are you having throat problem(yes or no)?

6. Are you having breath problem(yes or no).

7. Are you having headache problem(yes or no)?

8. Have you come in touch with anyone who was verified positive(yes or no)

III. THE MODEL'S DEVELOPMENT

Predictions were made using a gradient-boosting machine model created with decision-tree base-learners. Gradient boosting is commonly regarded as the state-of-the-art in tabular data forecasting[17], and it is employed in a number of successful machine learning algorithms[16].

This study uses the following properties of the dataset and the attributes

(#) Feature	Total n = 99,232		COVID-19 negative n = 90,839		COVID-19 positive n = 8393	
	n	%	n	%	n	%
(1) Sex						
Male	50,350	50.74	45,545	50.1	4805	57.2
Female	48,882	49.26	45,294	49.8	3588	42.7
(2) Age 60+						
True	15,279	15.4	13,619	14.9	1660	19.7
False	83,953	84.6	77,220	85	6733	80.2
(3) Cough						
True	14,768	14.88	10,715	11.8	4053	48.2
False	84,223	84.87	79,909	87.9	4314	51.4
(4) Fever						
True	8122	8.18	4387	4.83	3735	44.5
False	90,868	91.5	86,237	94.9	4631	55.1
True	1273	1.28	96	0.11	1177	14
False	95,062	95.8	88,059	96.9	7003	83.4
(5) Shortness of breath						
True	930	0.94	71	0.08	859	10.2
False	95450	96.14	88084	96.9	7321	87.2
(6)Headache						
True	1799	1.81	68	0.07	1731	20.6
False	94536	95.27	88027	96.9	6449	76.8
(7)Contact with a person who has been proven to have COVID-19						
True	5507	5.55	1455	1.6	4052	48.2
False	93725	94.45	89384	98.4	4341	51.8

As advised by studies done previously, those value which are missed were dealt with the gradient-boosting predictor inherently. LightGBM library of Python is used to train a gradient-boosting predictor. auRoc curve is used for performance checking. The SHAP values were chosen to emphasise the essential characteristics that influence model prediction. These parameters help complex models like artificial neural networks and gradient-boosting machines. SHAP values are generated from game theory and divide the effect of each and every ingredient feature value in each sample's prediction outcome. This is accomplished by comparing models to subsets of the feature space and calculating differences. Average of all the sample are taken then SHAP values are used to measure the effect of feature to model predictions in totality.

IV. DISCUSSION

Many ongoing research are focusing into the pathogenic approaches of SARS-Cov-2, in addition to the signs and symptoms related to it. This development of a COVID-19 test model is entirely focused on primary medical indicators and symptoms. By optimizing administration of healthcare assets in the whole SARS-Cov-2 pandemic waves that may come in future , improving medical priorities may also help to ease the burden presently encountered by fitness systems[18].This is particularly significant in less developed nations where resources are less available. This investigation has a number of flaws. The used data from the Indian Ministry of Health, which includes flaws, biases, and gaps in data for a variety of characteristics. For sufferers recognized as having had touch with someone tested to have COVID-19, extra facts along with the duration and location (indoors/outdoors) of interplay with someone showed to have COVID-19 changed into now no longer available. Some signs and symptoms (along with a lack of odor and taste) were located to be extensively predictive of a COVID-19 contamination in preceding studies19.The model may yield excessive accuracy if we filter out some of the important characteristics, so required care is taken in this implementation. Missing values can also lead to mislead. As a result, because there are many unreported or missing values than there are unsatisfactory values, evaluating the version's

Copyright @ 2022 Authors

performance is significantly more important. Poor ratings of all 5 indications are chosen and deleted the bad figures to create less biased condition in the prospective check set. The version produced positive results (Fig. 1) when applied to those simulated check sets, supporting our faith in the technique. While variations in reporting signs and symptoms might be a weak spot in our version, all and sundry who changed into examined (aside from a small institution who had been evaluated as a part of healthcare personnel surveys) had a purpose to be tested[13].For the vast majority of the participants in our study, our suggests that there was no referral bias. Cough and fever had been recognized as good sized signs and symptoms withinside the Health Department of India guidelines, and we consider that even in sufferers who examined poor for SARS-Cov-2, those signs and symptoms are hard to overlook. Furthermore, It is considered that the excessive pattern length aids withinside the removal of COVID-19-poor institution biases. The emphasis was at the importance of more robust data that supplement the methods used, while also emphasising that self-reporting of symptoms are inevitably biased. Because pandemic COVID--19 is spreading, now the importance is increased to continue gathering and sharing solid records with public groups and the medical community. Additional signs and symptoms can be integrated into destiny fashions as we benefit a higher know-how of the function of numerous. Many ongoing research are focusing into the pathogenic techniques of SARS-Cov-2, in addition to the signs and symptoms related to it. This model is created by looking at screening paradigm primarily based totally on primary scientific symptoms and signs and symptoms. Finally, we created a version for predicting COVID--19 analysis primarily based totally on statewide facts furnished with the aid of using the Health Dept. of India with the aid of using asking 8 essential questions. When checking out sources are restricted, we will utilise our approach to prioritise COVID--19 checking out, amongst different things. Furthermore, these methods used in the study should aid the fitness-care system in responding to future epidemics of the disease and also other respiratory viruses in general. Figure 2 illustrates the most important features. COVID--19 analysis prediction SHapley Additive exPlanations (SHAP) beeswarm graphic, with SHAP values for the version's maximum applicable functions. The suggest absolute SHAP values are used to organization the functions withinside the précis plots (y-axis). Each factor represents a completely unique player withinside the observe. The effect of every feature at the classifier's prediction for a given character is proven with the aid of using the placement of every factor at the x-axis. The hue represents the value of certain characteristics (for example, fever).

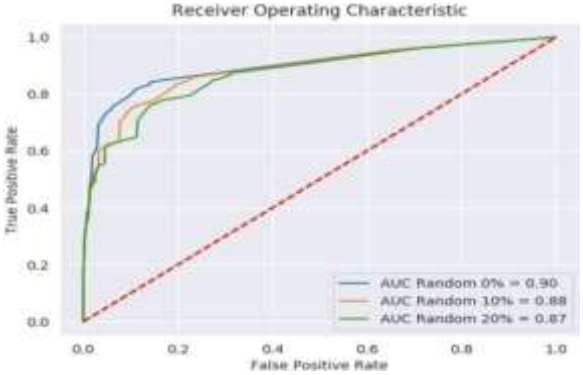

Figure 1: Shows Test set results for triggered tests

ROC curves illustrating the model's performance on driven testing data, where negative reports for each of the five symptoms were replaced with a blank value at random. To show the ROC curve for the actual testing dataset blue color is chosen. For each of the five symptoms with random substitution of 10% and 20% of the -ve values orange and green curves in ROC is used to display . While inconsistencies in reporting indications constitute a flaw in our model, everyone who was tested (with the exception of a small institution that was evaluated as part of healthcare workforce surveys) had a reason to be tested[13]. According to our findings, the vast majority of the participants in our study were not subjected to referral bias. Cough and fever had been diagnosed as sizable signs withinside the Health Department of India guidelines, and we trust that even in sufferers who examined poor for SARS--Cov--2, those signs are hard to overlook. Furthermore, we trust that the excessive pattern length aids withinside the removal of COVID--19-poor institution biases. We emphasise the importance of extra sturdy statistics to complement our technique, in addition to the reality that self-reporting of signs is inherently biased. The persistent recording and sharing of strong statistics among public companies and the medical network is important because the COVID--19 pandemic proceeds. Additional signs can be integrated into destiny fashions as we benefit a higher understanding of the position of diverse signs in diagnosing the condition. Finally, with the aid of using asking 8 key questions, a version is proposed for predicting COVID--19 analysis primarily based totally on statewide statistics furnished with the aid of using the Indian Health Department. Since the sources are restricted, the proposed approach is to prioritise COVID--19 checking out, amongst different things. Additionally ,this study and its approaches may aid the health-support in responding to future outbreaks of this pandemic, as well as some other respiratory viruses.

Copyright @ 2022 Authors

V. RESULTS

On new test set, prediction of this model is 0.90 for auROC on confidence interval of 0.892–0.905 with 95 percent. (See Fig. 2a) On the basis of projections on test set, possible working points are 71.98 percent specificity and 87.30 percent sensitivity or 79.18 percent specificity and 85.76 percent sensitivity. With an auPRC (area under the precision-recall curve) of 0.66 and a 95 percent confidence interval of 0.647–0.678, Figure 2b is showing the PPV(Positive Predicted Value) for diagnosis of Covid with respect to sensitivity.

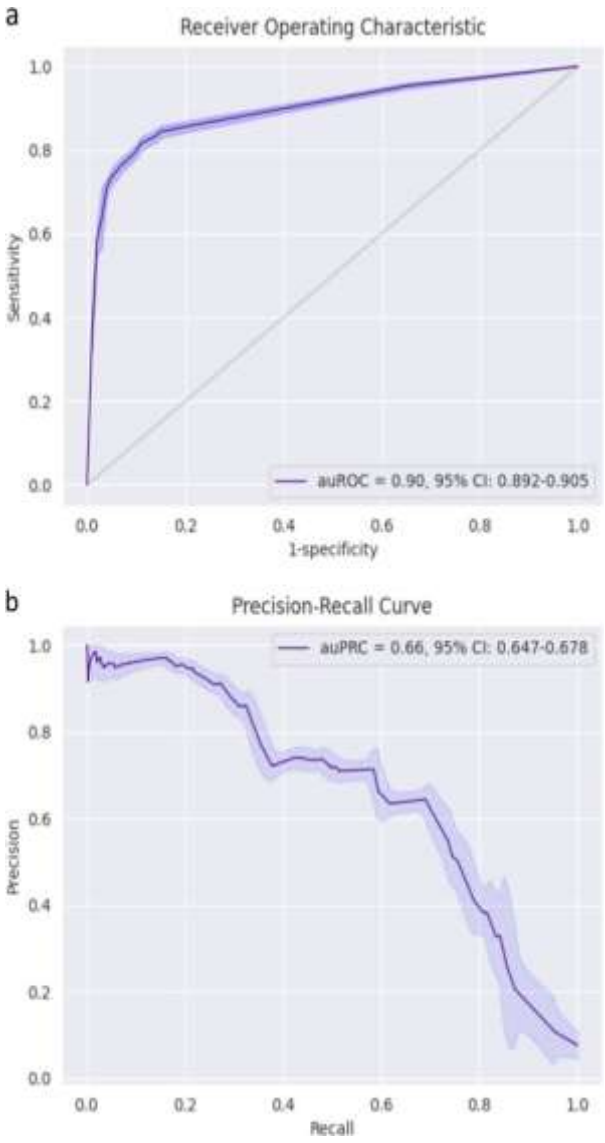

Figure 2: Shows the performance

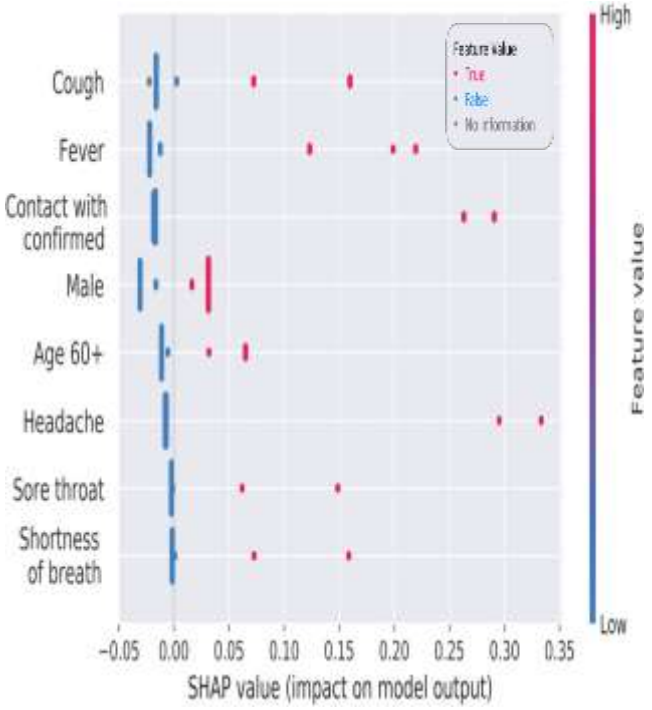

Figure 3: Shows a ranking of the model's most important attributes

VI. USING IMPARTIAL FEATURES TO TRAIN

In India, the information provided by the Health Department has errors and biases. Individuals who tested positive for COVID--19, for example, reported more detailed symptoms, which was supported by a focused epidemiological effort. As a result, those who test negative for COVID--19 are likely to mislabel their symptoms. This is seen by the percentage of people who tested positive for COVID--19 out of the total number of people who tested positive for each symptom. As a result, the features discovered with biased information (96.2 percent headache, 92.3 percent sore throat, and 92.4 percent breath shortage) and features with balanced reporting (96.2 percent headache, 92.3 percent sore throat, and 92.4 percent shortness of breath) (cough 27.4 percent and fever 45.9 percent). Incorrect categorization may occur as a result of misinterpretation and failure to report of symptoms among those who tested negative. In this proposed model training and testing is done by filtering indicators of high bias in advance. So that an auROC of .862 with small adjustment to SHAP(SHapley Additive exPlanations) can be attained as per graphic shown in fig 4.

Copyright @ 2022 Authors

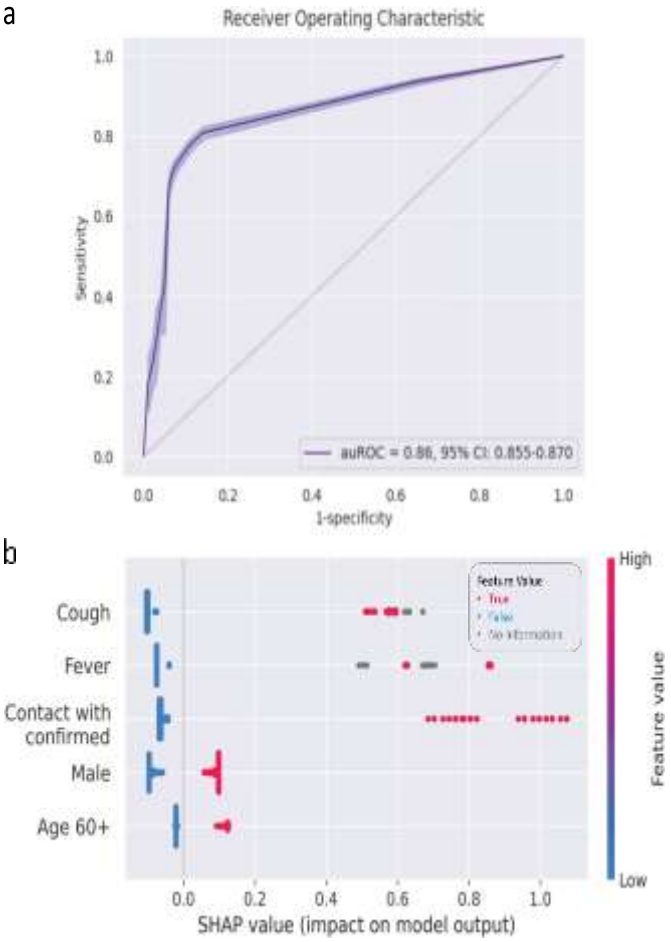

Figure 4: Shows Performance using only balanced Attributes

VII. MODEL ANALYSIS

In this model auROC is used to determine the version on the test set. Various thresholds were used to create plots of PPV vs. sensitivity (precision–recall curve) .For all sensitivity, thresholds, specificity, PPV, bad predictive value and fake-great rate, fake-poor rate, fake discovery rate, universal accuracy, ROC curves, PPV, bad predictive value sensitivity, specificity ,self belief intervals (CI) for several overall performance metrics were calculated using the bootstrap percentile technique with 1,000 repeats.

.

REFERENCES

[1] Gozes, O. 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 2003, arXiv:2003.05037 (2020).

[2] Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. https://doi.org/10.1016/S1473-3099(20)30120-1 (2020).

[3] Wang, S. et al. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). Springer, https://link.springer.com/article/10.1007/s00330-021-07715-1 (2021).

[4] Song, Y. et al. Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images. IEEE, https://ieeexplore.ieee.org/document/9376253 (2021).

[5] Jin, C. et al. Development and evaluation of an AI system for COVID-19 diagnosis.

medRxiv, https://doi.org/10.1101/2020.03.20.20039834 (2020).

[6] Punn, N. S. & Agarwal, S. Automated diagnosis of COVID-19 with limited pos- teroanterior chest X-ray images using fine-tuned deep neuralnetworks.Springer https://link.springer.com/article/10.1007/s10489-020-01900-3(2021)

[7] Tostmann, A. et al. Strong associations and moderate predictive value of early symptoms for SARS-CoV-2 test positivity among healthcare workers, the Neth- erlands, March 2020. Eurosurveillance 25, 2000508 (2020).

[8] Feng, C. et al. A novel triage tool of artificial intelligence assisted diagnosis aid Mei, X. et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat. Med. 26, 1224–1228 (2020).

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

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

[11] 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).

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

[13] 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).

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

[15] Omar, K. B. A. XGBoost and LGBM for Porto Seguro's Kaggle challenge: A compar- ison Semester Project (ETH Zurich, 2018).

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

[17] 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).

[18] Cong Feng et. al. A novel artificial intelligence-assisted triage tool to aid in the diagnosis of suspected COVID-19 pneumonia cases in fever clinics

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7940949/(2021)

Copyright @ 2022 Authors