IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 23, Issue 3, Ser. I (May – June 2021), PP 17-23
www.iosrjournals.org

Symptoms Based COVID-19 Detection Model Using Logistic Regression Algorithm

Shahadat Hossain[1], Ahmed Imtiaz[1], Dhonita Tripura[1]
[1]Department of Computer Science and Engineering, Rangamati Science and Technology University, Rangamati, Bangladesh.
Corresponding Author: imtiazmain@gmail.com

Abstract: *COVID-19 is a new acute respiratory syndrome disease. World Health Organization has already pointed out major symptoms of this disease. Due to limited test kits available in hospitals, rapid screening of patients is an alternate way to prevent social transmission. Symptom based automatic detection and prediction of COVID-19 is a quick option in this case. A model combining logistic regression algorithm, Scikit-learn, TensorFlow, Pandas and NumPy is applied to the COVID-19 symptoms to predict and detect infected person. Major COVID-19 symptoms are employed by our machine learning model to train itself and for decision making purpose. The proposed method shows satisfactory testing accuracy. It greatly improves the efficiency in clinical practice to detect and predict COVID-19 infection instantly without corona testing kits.*

***Key Word**:COVID-19, symptom, logistic regression, prediction, machine learning.*

Date of Submission: 08-05-2021 Date of Acceptance: 23-05-2021

I. Introduction

COVID-19 is a fast-spreading novel coronavirus. It has spread all over the world within a very short time since detected in December 2019. The COVID-19 virus spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes[1]. Fever, cough, malaise, and sore throat are the most common initial symptoms of coronavirus disease 2019 (COVID-19)[2]. Symptoms may appear 2-14 days after exposure to the virus[3]. Rapid and timely detection of the infected individual is important for both patient's health and prevention to further spreading.

Isolating the infected patients reduces risk of mass transmission. The global supply chain also got hampered due to corona pandemic. Hospitals raise concerns about insufficient storage of COVID-19 testing kits. Sample collection and lab test for corona virus need time. But due to technological advancement, software based COVID-19 detection also has shown promising results. The Food and Drug Administration (FDA) recognizes the extensive variety of actual and potential functions of software applications (apps) and mobile apps, the rapid pace of innovation, and their potential benefits and risks to public health[4].

This paper represents how COVID-19 can be detected by main symptoms without dedicated testing kit. At first a dataset is fed into this model to train itself using logistic regression algorithm and machine learning module: Scikit-learn, TensorFlow, NumPy and Pandas. After that it can detect the possibility of infection of a new patient.

II. Literature Review

AI and natural language processing make easy access to people for gathering information. Besides AI powered medical devices provide promising accuracy. Machine learning based corona detection makes easy for analysis stated by Akib Mohi Ud Din Khanday[5]. The use of artificial intelligence and the deep-learning subtype in particular has been enabled by the use of labeled big data along with markedly enhanced computing power and cloud storage across all sectors[6]. Logistic regression is a supervised learning classification algorithm used to determine the association between very clear dependent variables against the independent variables. Alison Callahan has shown their research to detect COVID-19 using software based on symptom data[7].

III. Symptom Based Detection Model

The proposed flow chart to detect COVID-19 includes following steps:

1. Gathering of symptom data of COVID-19 patients.
2. Train the software with collected symptom dataset.
3. Now the software is ready to predict the possibility of COVID-19 infection for new random patients.

IV. Feature Engineering

Five major symptoms are taken in this model to predict COVID-19 infection. We sit down to find out the best model parameters and such parameters with their numeric weight value are taken as follows:

1. Fever- Continuous value
2. Body Pain- No, Severe: (0/1: Binary value)
3. Age- Discrete value
4. Runny Nose- Yes, No: (0/1: Binary value)
5. Breathing difficulty- No, Mild, Severe: (-1/0/1: Categorical value)

V. Methodology

Python packages helped to build software infrastructure. Firstly a trained dataset is generated using machine learning. Then we take values of fever, age, body pain, runny nose, and breathing difficulty of an infected patient. After comparing with our trained data, logistic regression algorithm suggests the possibility of infection.Figure 1 demonstrates the total work flow of this work.

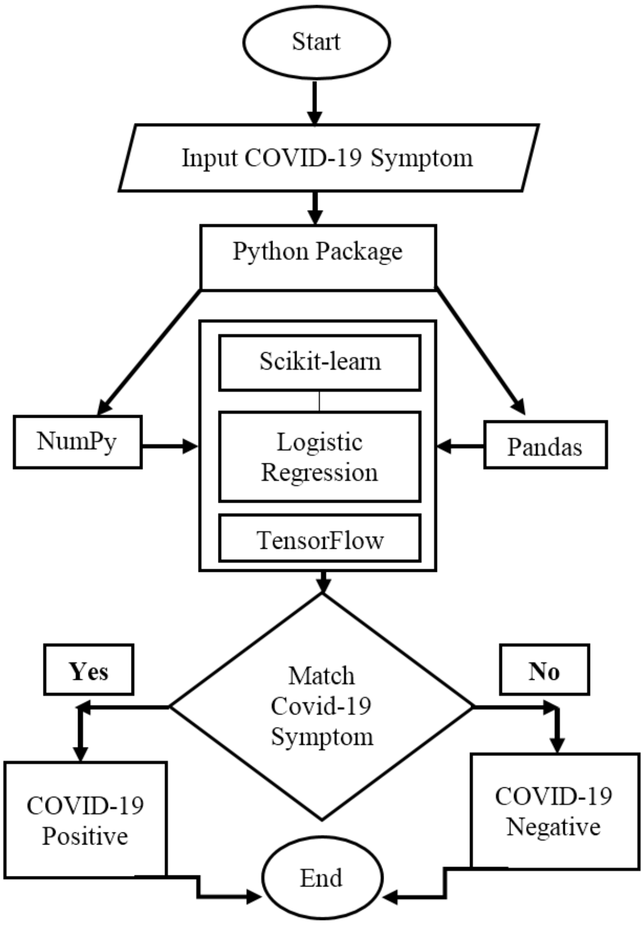


Figure 1. Logistic regression based COVID-19 detection process

VI. Training Procedure with Logistic Regression Algorithm

Machine learning algorithm like logistic regression acts on the concept of probability. It uses predictive analysis to classify problems. It is a statistical method for predicting into binary classes. The outcome or target variable is dichotomous in nature indicating there are only two possible classes[8]. So using this logistic regression, this model can predict the two possible classes that is COVID-19 positive or negative.

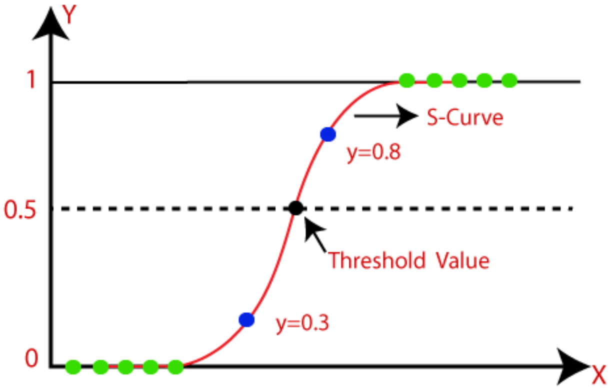


Figure 2. Logistic regression curve

The proposed method needs to input the random patients' symptoms data. After that these data will be processed and split. The split data will be two category one is training data and another is testing data. Training data is used by logistic regression algorithm and testing data is classified as 0 meaning COVID-19 negative and as 1 meaning COVID-19 positive.

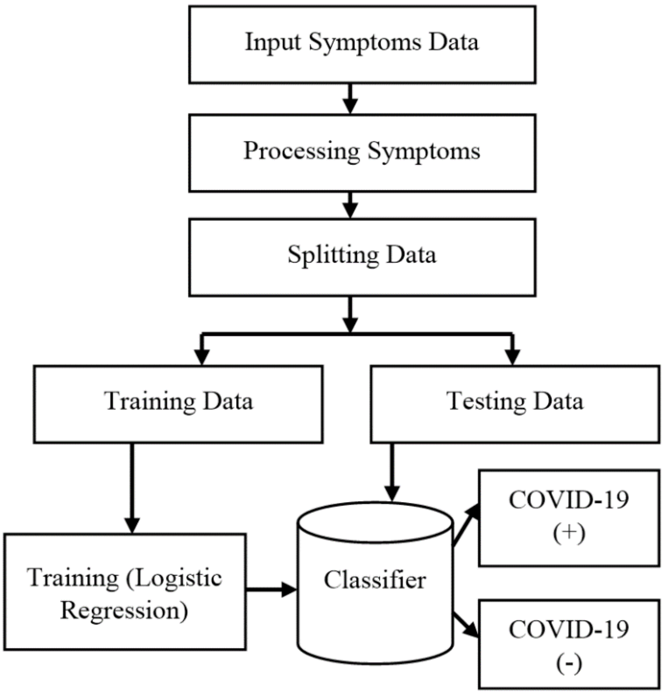


Figure 3. Workflow of proposed detection model

Here logistic regression algorithm provides a probabilistic value between 0 and 1. For decision making purpose, value of 50 is set as threshold level shown in figure 2. Any output reading greater than threshold value is denoted as higher possibility of COVID-19 infection.

VII. Training Data Collection

Due to the present worldwide pandemic, hospitals have opened special corona care unit. Symptoms begin at one to fourteen days after exposure of COVID-19 virus. Firstly 2000 COVID-19 patients were selected locally. Their data about five major symptoms (fever, body pain, age, runny nose and breathing difficulty) were collected for this model.

VIII. Supervised Learning of Symptom Based Prediction Model

Training dataset

In order to detect COVID-19 at an early stage, this study gives insight on how machine learning method can be used. Our software infrastructure is based on python package- TensorFlow, NumPy, Panda and

Scikit-learn. Symptom parameters with their respective numeric value also take part in it. Table 1 represents a snapshot of our training dataset of 2000 COVID-19 patients. This training data is fed into logistic regression algorithm to train our model. Finally logistic regression algorithm predicts COVID-19 positive or negative by classifying data for a new patient.

Table 1. Training dataset for detection process.

Input					Output
Fever	**Body pain**	**Age**	**Runny nose**	**Breathing difficulty**	**Infection probability**
95.93	0	36	1	-1	0
96	1	91	0	0	1
96.27	0	50	1	1	1
96.54	1	25	1	1	0
96.81	1	33	0	1	1
97.08	0	47	1	0	0
97.35	0	42	1	-1	0
97.62	1	18	0	-1	1
97.89	1	36	0	0	0
98.16	1	91	0	0	1
98.43	0	50	1	-1	1
98.70	1	25	1	0	0
...	...	...	...	...	...

Detection process

Here conditions of a patient are dependent variable. Binary regression is used as it has the binary outcome such as 0 and 1 or yes and no or true and false. Logistic regression model with a set of variables for prediction is written mathematically as:

$$p(x) = \log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \qquad \text{...............} \ (1)$$

Here p(x) is the predicted response. The variables b_0, b_1, b_2, b_n are predicted weight coefficient for the respective symptoms. x_1, x_2, x_n are patient's symptom input value. Our goal is to make p(x) as close as actual response. Logistic regression determines the best predicted weights b_0, b_1, b_2, b_n such that the function p(x) is as close as possible to all actual response[9]. Python based web application is introduced where patient needs to submit exact body temperature, age, body pain (severe or no), runny nose (yes or no) and breathing difficulty (severe or little or no). Figure 4 shows the designed symptom based systemic approach to detect COVID-19.

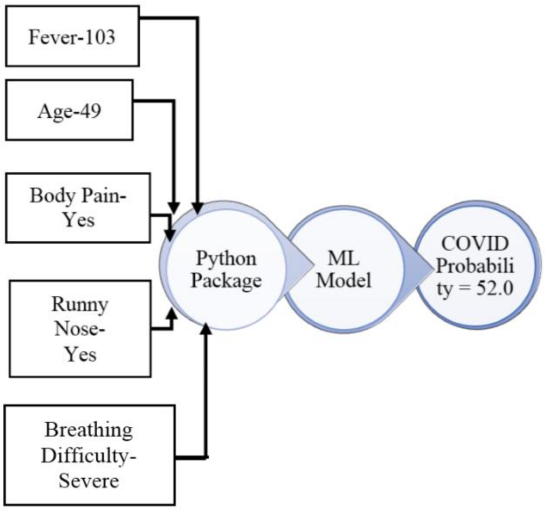
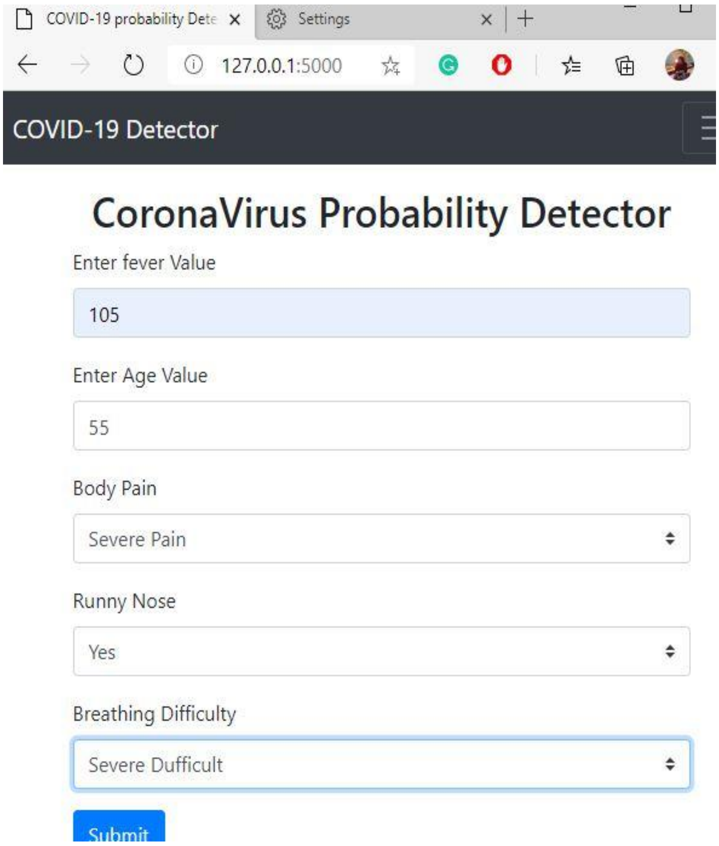


Figure 4. Detection of COVID-19 using machine learning process

Figure 5 represents the graphical user interface part of the software. Analyzed by logistic regression algorithm the model provides probability of corona infection. Based on that probability we measure possibility to get infected.



CoronaVirus Probability Detector

Enter fever Value

105

Enter Age Value

55

Body Pain

Severe Pain

Runny Nose

Yes

Breathing Difficulty

Severe Dufficult

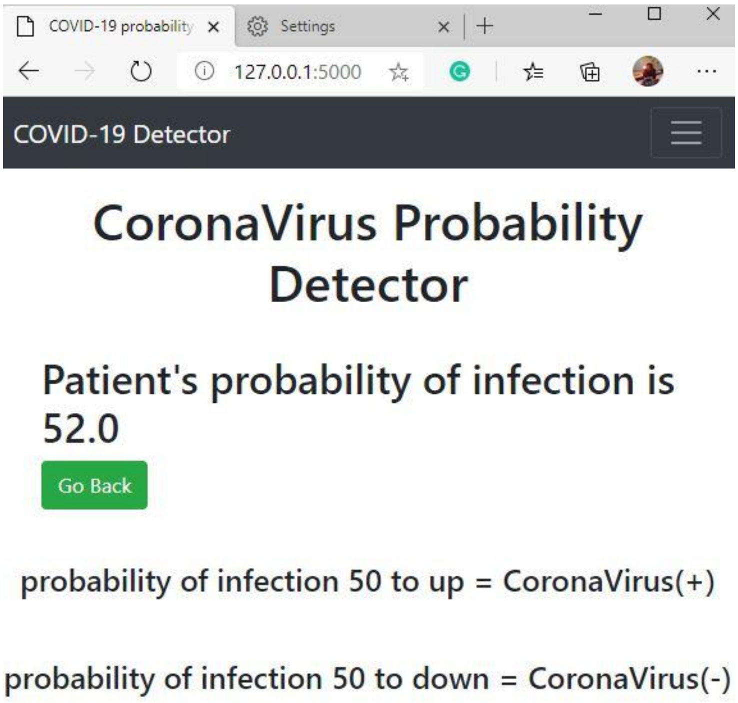
Figure 5. Web based symptom information input screen in COVID-19 detection software

The numeric weight values of respective symptom data are processed by logistic regression algorithm to find probability of infection.

Table 2. Measuring infection probability and detection of COVID-19 for new patients.

New Patient's Symptom Input					Output
Fever	**Body pain**	**Age**	**Runny nose**	**Breathing difficulty**	**Infection probability**
103	Severe	49	Yes	Severe	52.0
98	No	23	Yes	No	49.0
104	Severe	60	Yes	Severe	52.0
90	No	30	No	Little	46.0
106	Severe	80	Yes	Severe	54.0
…	…	…	…	…	…

If result shows value ≥ 50%, then patient should be physically diagnosed with corona testing kit. Otherwise the patient does not need a physical test. This is how hospitals can efficiently utilize their limited testing kits. Besides it helps professionals to set priority among people to diagnose them physically at an emergency basis.



CoronaVirus Probability Detector

Patient's probability of infection is 52.0

probability of infection 50 to up = CoronaVirus(+)

probability of infection 50 to down = CoronaVirus(-)

Figure 6. Output screen of COVID-19 detection software

IX. Result and Discussion

People of local hospitals from the south-eastern part of Bangladesh have participated in this study from July 2020 to February 2021. A patient should do a physical COVID-19 test if the COVID-19 probability ≥ 50. Total 675 new COVID-19 positive patients were selected for this experiment. Among them the model classifies 523 patients as positive. The proposed classification model to detect COVID-19 achieved about 77% accuracy.

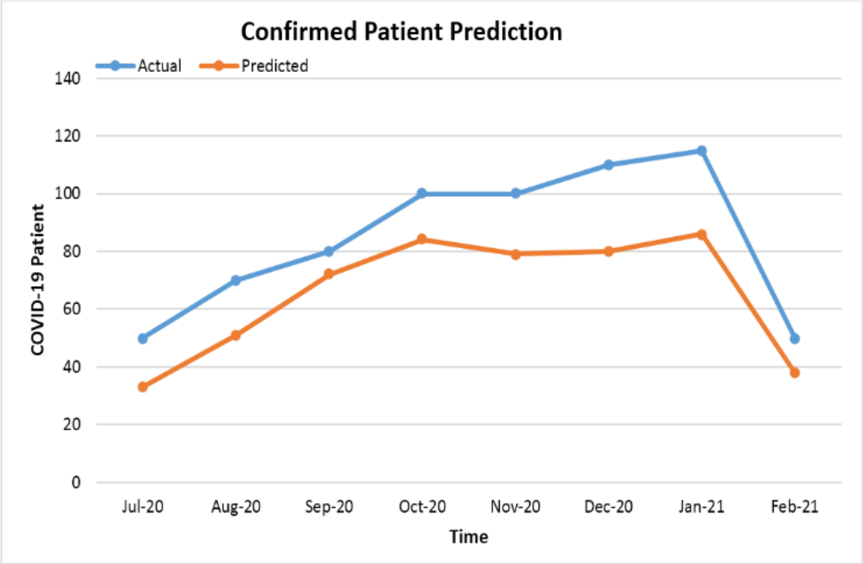


Figure 7. Performance of the proposed model: detection comparison between predicted patients and actual patients

The main advantage of this study can be summarized as follows: symptoms can be obtained from all patients very easily. This method is completely an end-to-end system. Professionals can use it at any place and the test is performed at an instant basis.

X. Conclusion

Shortage of corona testing kit hampers the faster isolation of patients. Other people can be infected with close contact of him. World Health Organization admits it as one of the key reasons of spreading across the world. Since a big range of sufferers are attending outdoor or emergency medical service, doctors' diagnosis time has become limited. This paper is a starting solution of COVID-19 detection by computer aided symptom based prediction system. The proposed method is hoped to be useful for professionals to detect COVID-19 immediately without wasting testing kit and also for making decisions in clinical practice.

References

[1]. World Health Organization, (2021, April 22). Corona virus: Overview, Available: https://www.who.int/health-topics/coronavirus#tab=tab_1

[2]. J. Komagamine, T. Yabuki. "Initial symptoms of patients with coronavirus disease 2019 in Japan: A descriptive study," Journal of General and Family Medicine, vol. 22(1), pp. 61-64, 2021.

[3]. Centers for Disease Control and Prevention. (2021, April 22). Considerations for Owners and Operators of Multifamily Housing Including Populations at Increased Risk for Complications from COVID-19, Available: https://www.cdc.gov/coronavirus/2019-ncov/community/multifamily-housing.html

[4]. US FOOD & DRUG ADMINISTRATION, (2021, April 22), Policy for Device Software Functions and Mobile Medical Applications Guidance for Industry and Food and Drug Administration Staff, Available: https://www.fda.gov/media/80958/download

[5]. A. Khanday, S. Rabani, Q. Khan, N. Rouf, M. Din, "Machine learning based approaches for detecting COVID-19 using clinical text data," International Journal of Information Technology, vol. 12(3), pp.731–739, 2020.

[6]. E. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, vol. 25, pp. 44–56, 2019.

[7]. A. Callahan, E. Steinber, 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, no. 95, 2020.

[8]. GOOD AUDIENCE, (2021, April 22), Machine Learning using Logistic Regression in Python with Code, Available: https://blog.goodaudience.com/machine-learning-using-logistic-regression-in-python-with-code-ab3c7f5f3bed

[9]. Real Python, (2021, April 22), Logistic Regression in Python, Available: https://realpython.com/logistic-regression-python/

Ahmed Imtiaz, et. al. "Symptoms Based COVID-19 Detection Model Using Logistic Regression Algorithm." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 23(3), 2021, pp. 17-23.