. **Paper No: SC-18** **Smart Computing**

Estimation of the incubation period of COVID-19 using boosted random forest algorithm

P. P. P. M. T. D. Rathnayake*
Department of Industrial Management
University of Kelaniya, Sri Lanka
thidasala.demintha@gmail.com

Janaka Senanayake
Department of Industrial Management
University of Kelaniya, Sri Lanka
janakas@kln.ac.lk

Dilani Wickramaarachchi
Department of Industrial Management
University of Kelaniya, Sri Lanka
dilani@kln.ac.lk

Abstract - **Coronavirus disease was first discovered in December 2019. As of July 2021, within nineteen months since this infectious disease started, more than one hundred and eighty million cases have been reported. The incubation period of the virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), can be defined as the period between exposure to the virus and symptom onset. Most of the affected cases are asymptomatic during this period, but they can transmit the virus to others. The incubation period is an important factor in deciding quarantine or isolation periods. According to current studies, the incubation period of SARS-CoV-2 ranges from2 to 14 days. Since there is a range, it is difficult to identify a specific incubation period for suspected cases. Therefore, all suspected cases should undergo an isolation period of 14 days, and it may lead to unnecessarily allocation of resources. The main objective of this research is to develop a classification model to classify the incubation period using machine learning techniques after identifying the factors affecting the incubation period. Patient records within the age group 5-80 years were used in this study. The dataset consists of 500 patient records from various countries such as China, Japan, South Korea and the USA. This study identified that the patients' age, immunocompetent state, gender, direct/indirect contact with the affected patients and the residing location affect the incubation period. Several supervised learning classification algorithms were compared in this study to find the best performing algorithm to classify the incubation classes. The weighted average of each incubation class was used to evaluate the overall model performance. The random forest algorithm outperformed other algorithms achieving 0.78 precision, 0.84 recall, and 0.80 F1-score in classifying the incubation classes. To fine-tune the model AdaBoost algorithm was used.**

Keywords - AdaBoost, boosted Random Forest, COVID-19, incubation period

I. INTRODUCTION

The Coronavirus disease 2019 (COVID-19) is one of the disastrous infectious diseases identified in late 2019 from a seafood wholesale market in China. Some of the common symptoms of COVID-19 include fever, dry cough, difficulty in breathing, muscle pain, sputum production, diarrhea, and sore throat [1]. While the majority of cases display mild symptoms, some progress to pneumonia and multi-organ failures. As for current findings, the death rate per diagnosed case is 4.4 percent; however, it could range between 0.2%-15% based on the age group and other health problems [2]. The virus typically spreads from one person to another via respiratory droplets released mostly during coughing and sneezing. As of July 2021, the virus has spread over 222 countries and territories resulting in 188,404,542 cases and 4,059,223 deaths [16]. Due to the high rate of diagnosed cases and deaths, the World Health Organization (WHO) has declared the COVID-19 disease as a pandemic on 11th March 2020.

Incubation period of COVID-19 can be defined as the time range a person spends between exposure to the virus and symptom onset. During the incubation period, most of the patients do not show any symptoms of being infected, but they are capable of transmitting the virus to others [17]. It is very important to isolate the suspected cases during this period to avoid virus transmission. Since the incubation period greatly varies among individuals, it is very important to identify the incubation period accurately in order to decide quarantine periods and to allocate limited resources effectively towards controlling the pandemic.

WHO has declared a time range of 2 to 14 days as the incubation period of COVID-19 patients [19]. Since there is a range to the incubation period, every suspected case should undergo a quarantine period of 14 days. During the quarantine period, active monitoring and resource allocation for the suspected cases are mandatory. Although all the suspected cases are quarantined for 14 days, some may have lesser incubation periods than others, because incubation period greatly varies depending on patients' gender, age, chronic disease history, direct/indirect contact with the affected persons, and the residing country. If there is a mechanism to identify the incubation period of each individual based on their characteristics, it will help prevent unnecessary resource allocation for quarantine/active monitoring, and effectively use the limited resources towards controlling the pandemic. The main purpose of this study is to develop a predictive model to classify the incubation period of the COVID-19 suspected cases based on their characteristics.

Section-wise organization of the paper is as follows. Section - II discusses related work. Section – III describes the methodology of the system. Results are discussed in detail in Section -IV. Finally, section – V presents the conclusion and future work directions.

II. RELATED WORK

A. Findings on incubation period

There are a number of studies to calculate the mean incubation period for the selected populations. One study has calculated the incubation period using 181 cases. This study has referred to patients' residing country, exposure date and time, dates of symptom onset, fever onset and hospitalization and calculated the median incubation period as 5.1 days [3]. The study states that 97.5% of the cases develop symptoms around 11.5 days. Another early analysis has referred to 158 cases outside the Chinese regions and estimated the median incubation period as 5 days which ranges from 2 to 14 days [4]. Authors have estimated the incubation period using lognormal distribution. This study specifies that the median time from illness onset to hospital admission was 3-4 days and the median delay between illness onset to death is 17 days. Another analysis based on 10 confirmed cases in China estimates the mean incubation period as 5.2 days (ranges from 4 to 7 days) [2]. This study specifies that children are less likely to be infected and may show milder symptoms. They have identified that age is one of the crucial factors that decide the incubation period. Their studies specify that 27% of the patients are hospitalized after two days of symptom onset which implies that time available to seek medical attention is generally short. Another analysis on 88 affected cases in Chinese regions outside Wuhan, specifies a mean incubation period of 6.4 days which ranges between 2.1 to 11.1 days [5]. They have obtained the possible values for the incubation period by considering the number of days the person has stayed in Wuhan and the date of symptom onset and fitted three parametric forms for the incubation period: The Weibull distribution, the gamma distribution, and the lognormal distribution.

B. *Factors affecting to the incubation period*

Studies about factors affecting the incubation period of COVID-19 patients are limited. One study has identified that age is directly related to the incubation period. This study was based on 136 patients who had travelled to Hubei, China, and identified the median incubation period as 8.3 days for all patients, 7.6 days for younger adults, and 11.2 days for older adults. This study specifies that elderly patients have a longer incubation period [6]. A study conducted by referring to r Chinese COVID-19 patients specify that men's cases tend to be more serious than women's cases [7]. Using a public dataset of 37 cases, Authors have identified that the number of male deaths from COVID-19 is 2.4 times the number of female deaths. Further, they have identified that the percentage of males were higher in the deceased group than in the survived group. There is strong evidence which suggests that men may have a larger concentration of ACE2 (angiotensin-converting enzyme 2) receptors in their body, which helps coronavirus to latch on and spread inside the body. This is one of the primary reasons why COVID-19 seems to affect men seriously, when compared to women [8]. Centre of disease control and prevention in the United States has identified that the people who have cancer, chronic kidney disease, COPD immunocompromised state (weakened immune system) due to solid organ transplant, obesity, BMI of 30 or higher), serious heart conditions such as heart failure, coronary artery disease or cardiomyopathies, sickle cell disease, type 2 diabetes mellitus have a higher risk of getting severely ill from COVID-19 [9]. Since chronic diseases directly affect the immune system of patients, the incubation period can differ from the immunocompetent people. Studies regarding the factors affecting the incubation period of COVID-19 patients are limited. Out of those studies one study has identified that the age is directly related to the incubation period. Authors have identified that the median incubation period for aset of COVID-19 patients who had traveled to Hubei, China was 8.3 days, and for the younger adults the incubation period was 7.6 days, and for older adults, 11.2 days. This study specifies that elderly patients have a longer incubation period than the younger adults [6]. A study conducted on two populations of COVID-19 patients from two geographic locations to identify the deviation of incubation period across residing location, has proved that there is a deviation of incubation period across two regions. Out of the 181 patients used for the study, 108 patients were diagnosed outside of mainland China with a median incubation period of 5.5 days and 73 patients diagnosed inside China with a median incubation period of 4.8 days [3]. The above literature specifies that the patients Age, Gender, Chronic disease history, and residing country directly affect the incubation period of the COVID-19 patients.

C. *Supervised learning classification algorithms used in COVID-19 domain*

One study has identified factors such as patients' age, residing country, if from Wuhan, if theyy have visited Wuhan and gender directly affect the death/recovery of COVID-19 patients using 100 confirmed laboratory cases in China [10]. This study has used the Naïve Bayes approach to classify the death/ recovery of COVID-19 patients and achieved 93% accuracy. Another study has used the Logistic Regression approach to detect COVID-19 using clinical text data. Authors have labeled 212 clinical records into four categories named COVID, SARS, ARDS, and both (COVID, ARDS). Various text features such as TF/IDF, a bag of words has been extracted from these clinical reports to classify them. This study has reached 94% precision, 96% recall, and 95% f1 score using Logistic regression approach [11]. Support Vector Machine (SVM) has been used in the COVID-19 domain to classify the X-ray images of COVID-19 suspected cases. The study in [12] has used this method to identify the X-ray images of COVID-19 patients by comparing normal X-ray images with X-ray images showing pneumonia. [12] This study has reached an accuracy of 97% by classifying the X-ray images into classes using SVM approach. Another study has used the decision-tree classifier to identify COVID-19 patients by referring to their Chest x-ray (CXR) images [13]. They have used three binary trees to identify the abnormality of the CXR images, identify the symptoms of tuberculosis and to identify COVID-19 symptoms. They have achieved an accuracy of 98% and 80% for the first two decision trees respectively, whereas the average accuracy of the third decision tree has been 95%. One of the studies have used the Random Forest algorithm to identify if a person is infected with the SARS-Cov2 virus and the type of hospitalization (regular ward, semi-ICU, or ICU) needed, based on the hematological parameters such as red blood cells, hemoglobin, neutrophils, lymphocytes, etc. collected from blood tests.. Authors have achieved 92.8% accuracy in identifying the type of hospitalization patients needed based on the hematological parameters from blood tests [14].

III. Methodology

The key purpose of the study is to identify the factors affecting the incubation period and to design a model that can classify the incubation period of the suspected cases based on patients' characteristics. Machine learning techniques were used to build the classification models. Next, the modelling techniques were compared on validation and model accuracy, to select the best technique. At last, the best classification technique was fine-tuned using a boosting algorithm to achieve higher accuracy.

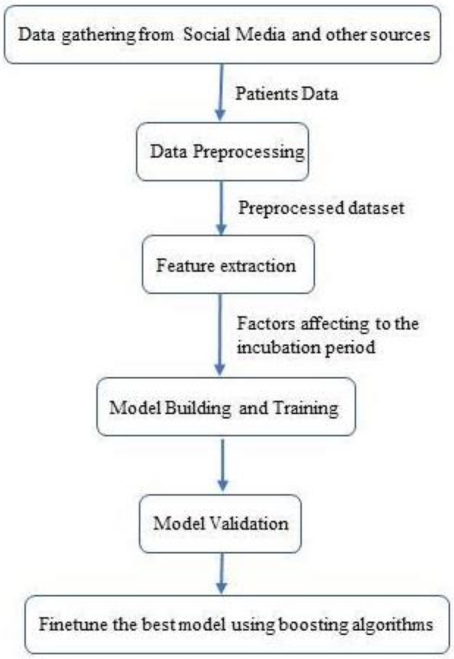
Data gathering from Social Media and other sources
Patients Data
Data Preprocessing
Preprocessed dataset
Feature extraction
Factors affecting to the incubation period
Model Building and Training
Model Validation
Finetune the best model using boosting algorithms

Fig. 1. The methodology of the proposed solution

Publicly available patient data and clinical records were used for this study. The following information about patients was gathered by analyzing the records manually.

i. Age
ii. Gender
iii. Residing Country
iv. Chronic disease history
v. Direct/ indirect contact with the affected cases
vi. Symptom onset date
vii. Exposure date/Travel dates
viii. Hospitalized date

Most of the data were collected from social media posts and status related to the COVID-19 patients. Chinese social media WeChat accounts are one of the major data sources which release daily information on the list of COVID-19 cases. Other than social media and WeChat accounts, following sources were used to collect data.

- Kyodo News
- Weibo.com
- Kaggle

In some of the cases, precise information was not recorded to identify the type (direct/indirect) of contact with the affected persons. If the patients travelled together with affected ones or if they got the virus from a family member, those scenarios are considered direct contact with the affected cases. Otherwise, an assumption was made - that they had indirect contact with the affected persons.

The incubation period was calculated using the date difference between symptom onset date and the exposure date. Since the incubation period of the selected population ranges from 5 to 24 days, it was divided into four classes as below, for classification.

- Class A: 20 - 24 days
- Class B: 15 – 19 days
- Class C: 10 -14 days
- Class D: 5 – 9 days

The incubation class was added to the dataset by creating a new column named ‘Incubation Class’. The median age of each incubation class was used to fill the missing values of the age column. Finally, label encoding was performed on the dataset. For analyzing the data, descriptive statistics were used. Bar charts were used to identify the distribution of the incubation period across patients’ age, gender, residing country, direct/indirect contact with the affected cases and chronic disease history. Next, Pearson’s Correlation Coefficient (PCC) was used to identify the variables which have the strongest relationship with the incubation period.

A number of supervised learning classification models were compared in this study to identify the best model for this particular problem. Models were implemented using Google Collab platform which provides a Jupyter notebook environment that requires no setup and runs entirely in the cloud with the accessibility of powerful computing resources from the browser. Classification algorithms such as multiple regression, support vector machine, random forest, K- nearest neighbor algorithm, naive bayes, and decision tree were compared to find the best model with highest accuracy, to classify the incubation period class based on patients’ demographics and other characteristics. In order to validate the classification models, percentage split technique was used. The dataset was divided into two categories randomly, mainly 20% for testing and 80% for training. Furthermore, performance metrics such as Precision, Recall and F1 Score were used to compare model performance.

Boosting algorithms were used in this study to achieve higher accuracy in machine learning algorithms. Boosting algorithms are very useful to create high accuracy models by combining low accuracy models. AdaBoost algorithm was used in this study to improve the accuracy of the best performing classification model.[18] The main concept of AdaBoost is that it assigns weights to classifiers and training the data samples in each iteration such that it ensures the accurate predictions of unusual observations.

IV. Results And Discussion

This section mainly describes the details related to the results obtained from the implementation process and the discussion of the results.

The gathered dataset for the study consists of 500 patient records with the age ranging from 5-80 years. Out of those records, 285 were male and 215 were female. The dataset includes patients’ information from most of the countries around the world with the majority of cases from China Singapore, France, Germany, Taiwan, Japan, Malaysia, United States, and South Korea. Following is the incubation period distribution for the dataset.

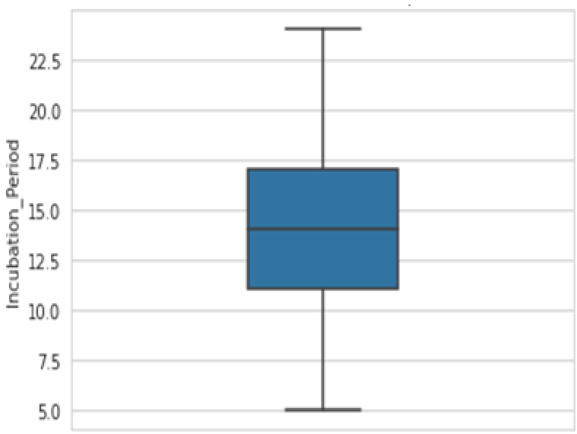

Fig. 2. Incubation period distribution for the dataset

The incubation period of the selected population ranges from 5-24 days with a median value of 13.86 days. The highest number of patients (51) have their incubation period as 14 days. Out of the 500 patient records, 31 of them (7.3% of the overall population) have their incubation period more than or equal to 20 days. 79 patients (15.8% of the overall population) have their incubation period less than or equal to 9 days. Majority of the patients have their incubation period between 10-19 days which is 76.8% of the overall population.

Correlation analysis was used in this study to identify the variables which have the strongest relationship with the incubation period. Based on the results of the correlation analysis, patients' age and the incubation class have a very strong positive relationship which is 0.819. When it comes to the direct contact with the affected cases, it also has a moderate positive relationship with the incubation class which is 0.360. Having a history of chronic diseases such as cardiac, respiratory and metabolic diseases also have a strong positive relationship with the incubation class. Patients' residing country also has a weak relationship with the incubation class which is 029.

Results based on descriptive statistics and the correlation analysis suggest that men's COVID-19 cases tend to decrease as the incubation period increases. This implies that men's COVID-19 cases tend to show symptoms quickly than women's cases do. Patients with chronic disease history such as Serious heart conditions, heart failures, coronary artery disease, cardiomyopathies, sickle cell disease, type 2 diabetes mellitus tend to show symptoms quicker than others. The different incubation periods can be the result of different types of inflammation and immune responses. When it comes to the method of exposure to the virus, results specify that patients who got direct exposure to the virus have a shorter incubation period than others. This implies that, if the patients had close contact with someone who has COVID-19 and got exposed to the virus directly, they tend to show symptoms very quickly than others who have got indirect exposure to the virus.

Number of supervised learning classification algorithms were compared in this study to identify the best model to classify the incubation class based on patients age, gender, chronic disease history, direct/indirect exposure to the virus and the residing country. The following figure explains the accuracy of each model in classifying the incubation class.

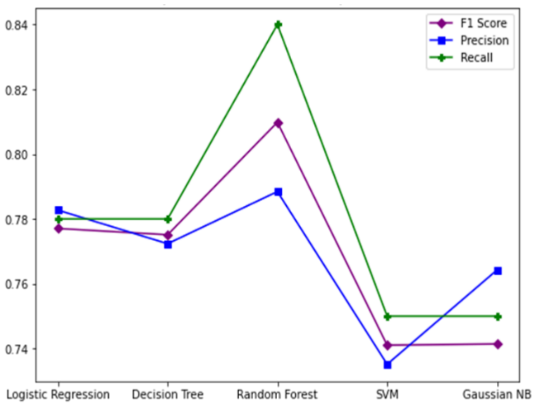

Fig. 3. Comparison of model performance without boosting algorithms

The above figure specifies that the Random forest algorithm performed better in classifying the incubation class by achieving higher precision, recall, and F1 score. Since the F1 score provides the harmonic mean between precision and recall, it was considered the best performance metric to evaluate the models. Following is the model performances in tabular format.

TABLE I. COMPARISON OF MODEL PERFORMANCE IN TABULAR FORMAT

Classifier	Precision	Recall	F1 Score
Naïve Bayes	0.764	0.750	0.741
SVM	0.735	0.750	0.741
Logistic Regression	0.780	0.782	0.777
Random Forest	0.788	0.840	0.809
Decision Tree	0.772	0.780	0.775

AdaBoost algorithm was used in this study to improve the accuracy of the classification algorithms. Since the AdaBoost algorithm needs a base classifier, random forest was used as the base classifier since it outperforms other classification algorithms.

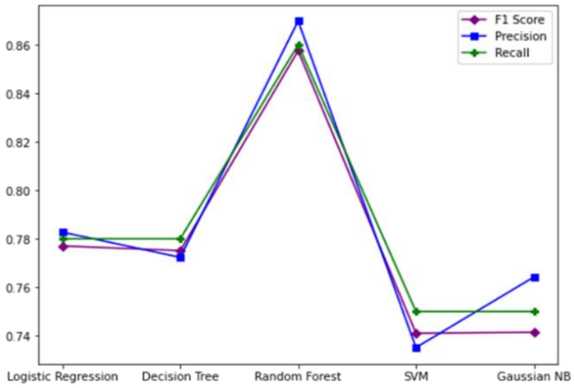

Fig. 4. Comparison of model performance with AdaBoost algorithm

Figure 4 displays the model performance after implementing the AdaBoost algorithm with the random forest algorithm as the base classifier

From figure 4, we can identify that the performance of the random forest algorithm increased after the application of the AdaBoost algorithm. Before applying the AdaBoost algorithm Random Forest algorithm outperformed other algorithms achieving a 0.78 Precision, 0.84 Recall, and a 0.80 F1 score. After applying the AdaBoost algorithm the performance metrics of the Random Forest algorithm increases up to 0.87 Precision score, 0.86 Recall Score, and a 0.86 F1 score.

V. Conclusion

This study implies that patients' age, gender, residing country, the method of exposure to the virus (direct/indirect exposure), and the history of chronic diseases such as cancer, chronic kidney disease, COPD, serious heart conditions, type 2 diabetes directly affect the incubation period of the SARS-CoV-2 virus. When it comes to age, older people tend to show symptoms quicker than younger people and they have a shorter incubation period compared to others. Gender wise, male cases tend to show symptoms quicker than others. Patients who have chronic diseases and immunocompromised states have a shorter incubation period than others and show symptoms quicker. The people who got direct exposure to the virus and who had a closer relationship with the affected cases tend to show symptoms quicker than people who got indirect exposure to the virus.

In this study, several supervised learning classification algorithms such as SVM, naïve nayes, logistic regression, random forest, and decision tree were compared to find the best model with the highest accuracy to classify the incubation period. Random forest algorithm outperformed in classifying the incubation period achieving higher precision, recall, and F1 score. Finally, boosting algorithms such as the AdaBoost algorithm was integrated with the random forest algorithm to achieve 0.87 Precision, 0.86 Recall, and a 0.86 F1 score in classifying the incubation period.

This study mainly focused on the symptomatic transmission of COVID-19. Symptomatic transmission refers to transmission from a person while they are experiencing symptoms such as fever, cough, tiredness, etc. In a symptomatic case, we are able to track the incubation period by the date difference, between exposure to symptom onset. There are some cases showing asymptomatic transmission of COVID-19. Asymptomatic transmission can be defined as the transmission of virus from person to person, without showing symptoms of being infected. Very few asymptomatic transmission cases have been reported as a result of contact tracing efforts in some countries. Since asymptomatic patients do not show symptoms, it is relatively difficult to identify the incubation period. This study was conducted using only 500 patient records from several countries around the world. If there is larger number of patient records representing all the countries around the world with patients' clinical information, a comprehensive study can be carried out. Further, unsupervised machine learning algorithms such as artificial neural networks can be implemented with a larger dataset in order to achieve higher accuracy.

As future work, chest X-ray images of COVID-19 affected persons can be combined with geographic and healthcare data processing models which will then be integrated into applications that will support the decision-making process for the authorities and for the growth of the healthcare systems. This will finally lead to the development of semi-autonomous classification systems that can provide the facility to detect the incubation period of COVID-19 patients accurately and prepare us for future outbreaks.

References

[1]. Symptomps of Coronavirus, Retrieved from https://www.cdc.gov/coronavirus / 2019-ncov/symptoms-testing/symptoms.html, September 2020

[2]. X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, M Leung, E. Lau, J. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia", The New England Journal of Medicine, 2020.

[3]. K. Grantz, Q. Bi, F. Jones, Q. Zheng, H. Meredith, A. Azman, N. Reich, J. Lessler, "The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application", American College of Physicians Public Health Emergency Collection, 2020

[4]. N. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A. Akhmetzhanov, S. Jung, B. Yuan, R. Kinoshita, H. Nishiura, "Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data", Journal of Clinical Medicine, 2020

[5]. J. Backer, D. Klinkenberg, J. Wallinga, "Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travelers from Wuhan, China, 20–28 January 2020", Europe's journal on infectious disease surveillance, epidemiology, prevention and control, 2020

[6]. T. Kong, "Longer incubation period of coronavirus disease 2019 (COVID-19) in older adults" Aging Medicine journal, 2020

[7]. J. Jin, P. Bai, W. He, F. Wu, X. Liu, D. Han, S. Liu, J. Yang, "Gender Differences in Patients With COVID-19: Focus on Severity and Mortality", Frontiers in Public Health Journal, 2020

[8]. Coronavirus: Why Men May Suffer From Severe Symptoms Of COVID-19 Than Women, According To Studies, Retrieved from https://timesofindia.indiatimes.com/, January 2020

[9]. People with Certain Medical Conditions, Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions /people -with-medical-conditions.html, September 2020

[10]. I. Sudirman and D. Nugraha, "Naive Bayes classifier for predicting the factors that influence death due to COVID-19 in china.", Journal of Theoretical and Applied Information Technology, 2020

[11]. A. handay, S. Rabani, Q. Khan, N. Rouf, M. Din, "Machine learning-based approaches for detecting COVID-19 using clinical text data", Nature Public Health Emergency Collection, 2020.

[12]. D. Novitasari, R. Hendradi, R. Caraka, Y. Rachmawati, "Detection of COVID-19 chest X-ray using support vector machine and convolutional neural network", Communications in Mathematical Biology and Neuroscience, 2020

[13]. S. Yoo, H. Geng, T. hiu, S. Yu, D. Cho, J. Heo, M. Choi, I. Choi, C. Van, N. Nhung, B. Min, H. Lee, "Deep Learning-Based Decision-Tree Classifier for COVID-19 Diagnosis From Chest X-ray Imaging", University of Medicine and Health Sciences, United Arab Emirates, 2020

[14]. V. Barbosaa, J. Gomesb, M. Santanab, C. Limab, R. Caladoe, "Covid-19 rapid test by combining a random forest-based web system and blood tests", Department of Mechanical Engineering, Federal University of Pernambuco, Recife, Brazil, 2020

[15]. Transmission of COVID-19 by asymptomatic cases, Retrieved from http://www.emro.who.int/health-topics/coronavirus/transmission-of-covid-19-by-asymptomatic-cases.html, January 2020

[16]. Covid-19 Coronavirus Pandemic, Retrieved from https://www.worldometers.info/coronavirus/, July 2020

[17]. Transmission of SARS-CoV-2: implications for infection prevention precautions, https://www.who.int /news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions, July 2020

[18]. E. Prabhakar, C. Nalini "Boosted Adaboost to Improve the Classification Accuracy", Department of Information Technology, Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India, 2012

[19]. Coronavirus disease (COVID-19), Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19, October 2020