ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.020

Risk Analysis and Prediction of Covid-19 Infection Using Machine Learning Techniques

H L Gururaj [1+], Hong Lin [2], B R Sunil Kumar [1], Janhavi V [1] and Manu M N [3]

[1] Department of CSE, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India.

[2] Department of CS, University of Houston, Downtown, USA.

[3] Department of ISE, SJBIT, Bengaluru, Karnataka, India.

Abstract. Coronavirus (COVID-19), the lethal contagious virus which has caused a pandemic, has metastasized all over the world starting from China. The figures observed of the number of casualties, is in millions and billions. This new malicious virus has caused panic amongst pubic, implanted fear and number of doubts in people's minds. There is lack of information as scientists are working on eradicating this deadly virus, less information has instilled doubts and people are panicking being helpless about how to cope up with the virus. Ways to protect oneself from getting infected, how could and where could one seek medical help when needed, these kinds of queries should be sorted out and the public needs to be educated about the virus. This will help calm down the public. This would also aid in keeping tranquil environment and even help in health and government sector workers to carry on with their duties without any obstacles.

Keywords: global pandemic, coronavirus (COVID-19), Asymptomatic; symptoms, diagnosis, prognosis, k-means, naïve bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), self-assessment, posterior probability.

1. Introduction

Since the first case of the Coronavirus (COVID-19), which was reported in December 2019, by the officials in China, in the city of Wuhan, the spread of this infectious virus has been very rapid. It has spread its wings across the globe causing turbulence in everyone's lives and has caused massive casualties. Because of this new situation faced by mankind, there is a lack of information, which has in turn induced fear, doubts, panic in people's minds [1]. Myths are spreading like a wildfire; false information is being circulated which needs to be stopped. It is crucial to contain the virus and educate people about the virus. A system is needed which could aid in acknowledging people's fears, doubts; a system is needed which acts as a myth buster and prevents the spreading of false information. A system where any individual can anonymously get themselves tested is needed to break the taboos present in society is very much needed [2].

There are a few pre-defined symptoms faced by a COVID-19 infected person given out by World Health Organization (WHO) such as fever, dry cough, tiredness, body pain, sore throat, etc. The symptoms might vary in a wide range, but any person infected with the Coronavirus (COVID-19), might or might not face the pre-defined symptoms like dry cough, fever, tiredness, etc. This case cannot be concluded as Coronavirus negative case; This kind of case is considered asymptomatic and they too require medical care. A system that would predict vital parameters which play a very important role in treating and deciding whether the person is infected or not. Since in the existing system the results issued are late, there is a need for a system that gives faster results, which helps any user at risk of infection to take further steps[3].

If a report on where one stands with the coronavirus (COVID-19) infection is given, that is whether the risk of infection is low, medium, or high, will help him/her take right decisions calmly and not in haste. If a person is at high risk of infection might need instant emergency care, so a system where one can call emergency help provided by the government or any health care institution is needed. A person with medium risk must be guided to isolate themselves and take medications[4].

+ Corresponding author. Tel.: +91- 9686418942; fax: +91- 9686418942
E-mail address: gururaj1711@vvce.ac.in

An individual with low risk must be suggested to boost their immunity (by administering their daily day activities like physical work-out, changing food habits, etc.) in order to prevent infection in the future. There is a requirement for a system with these features. Right medical treatment at right time is very important in saving the victim's life or avoiding worsening of the victim's health condition. Hence this creates a necessity to provide the public about the nearest health center in their current location[5]. This helps in easy access to the medical help made available to them by the government. The non-accepting attitude of society towards infected individuals is also a major issue that needs to be addressed and solved. Because of this reason people might back down to seek medical help. This calls for a system that promises confidentiality of one's data as the data of the patient should be kept between the patient himself/herself and the doctor he/she is consulting[6].

Further paper is organized as follows. In Section 2, a literature survey of mechanism of action is described. In Section 3, the methodology of the proposed system is given. Section 4, depicts the analysed results of the system which is being put forward. The Section 5, the paper is concluded accordingly.

2. Literature Survey

Machine Learning (ML) gives several approaches, ways, and gadget that helps in solving identification and prediction problems in various medical domains. ML is used for the examination of the significance of clinical parameters and their combinations for diagnosis, e.g., prediction of disease progression, extraction of medical knowledge for outcome research, therapy planning and support, and for the overall management of the patient [7]. ML is used for data analysis too, for instance, observation of regularities in the data by appropriately dealing with the data which is not perfect, explanation of continuous data used in the Intensive Care Unit, and intelligent alarming resulting in helpful and systematic monitoring. It is declared that the successful execution of ML methods can help the combination of computer-based systems in the healthcare environment giving chances to ease and amplify the work of medical experts and eventually to improve the productivity and quality of medical care [8].

2.1. Challenges faced by Existing System

Coronavirus disease is a viral infection with a high rate of transmitting among individuals all around the world. The very first coronavirus disease case was announced in December 2019, in China and then it rapidly outspread to all other countries. As of today, in the world, the total COVID-19 case count is +187,889,550 and the death count is +4,0521700 according to the Worldometer. Even after the release of vaccinations for the betterment of the current situation, there is no sign that the number of cases and death counts will be getting to a minimal number and to get the situation under control. . Due to this serious situation, WHO has increased the COVID-19 risk assessment to the highest level and declared it as a global pandemic [9].

3. Methodology

3.1. Prediction of Covid-19 using SVM and Naïve Bayes

Support Machine Vector Machine (SVM) being a supervised machine learning algorithm, is implemented using Python, scikit-learn (sklearn).

Initially, we import the scikit-learn library and an object is created for further. The datasets are split into training and testing datasets. Then we create a classifier that uses the Radial Basis Kernel since the kind of data is non-linearly separable data. We set the values of the hyper-parameters, which are C and Gamma values. The C determines how tolerant the model is towards misclassification and the gamma parameter defines how far the influence of a single training instance reaches. After this, we try and fit the training data and predict the outcome.

The proposed model will be trained by feeding the datasets. The Support Vector Machine will use the RBF Kernel (Radial Basis Kernel) and will plot the data points in infinite dimension just like its trivial character and segregates different classes. The user will register by entering his/her age and gender, further, the user provides input of the symptoms he/she is facing. This input is taken and plotted placed in one of the classes segregated earlier. When the user enters all the symptoms, he/she is facing, they will be used and

classified but, where the user stands with the infection, this is given by Naïve Bayes Classifier Algorithm [10].

Prior Probability is calculated first and we find the likelihood probability from the trained dataset using users given attributes for each value. Later, using the Bayes formula, we calculate the posterior probability using both new data and trained data. The class with the highest probability is the outcome of the prediction. This helps in deciding the user to know whether he/she has a low, medium, or high risk of coronavirus infection [11].

3.2. Classification and Prediction of Vital Factors Using K-Means

The classification of the Covid-19 vitals factors is done using the above-mentioned method that is using the K-means clustering algorithm. Vital factors are the clinical measurements of the factors, that indicate the state of the patient's essential body functions [12].

Some of the vital factors that majorly causing the increase in mortalities are:

1. Heart rate: A normal resting heart rate should be between 60 to 100 beats per minute but according to the researcher's covid-19 can cause an irregular or high heart rate over hundred beats per minute.
2. Blood pressure: The normal blood pressure level is less than 120 mm Hg or 180 mm Hg, therefore a further reduction in the level can increase mortality.
3. Oxygen saturation: Oxygen should be around 95% or higher than that, i.e 75 to 100 mm Hg, A person below 60 mm Hg may require an oxygen supplement.
4. Respiratory rate: Regarding the respiratory rate, 12 to 16 breaths per minute are the ideal range but an infected person can have more than 23 words per minute.

The rate of transmission of the virus is increasing day by day and it has become important to know the vital factors associated and comorbidities. Any fluctuation in these factors range may cause greater difficulties and can cause the death of an individual. Therefore, having knowledge and understanding becomes important for each individual to take necessary caring in the time.

1. The data consists of each patient record that has all the inputs given by users
2. Each patient data is a node which is considered as the one data point
3. The factors like age, gender, and symptoms are taken as the clustering parameter
4. Distance measure between these data points is done by using the Euclidean distance measure
5. Allocation of each data point is done to the cluster that has similarity with that specific parameter
6. The iterative process continues until the groups become stable. The final groups are classified which are compared with trained data values and prediction of the vital parameters is given with ranges

The implementation of the K-means clustering is done using the java packages which have hundreds of classes and in our proposed system we are using few classes like BufferedReader, Iterator, FileReader, HashMap, Locale.

Using the HashMap values for storing the input data and Iterator for looping through all the data points, we retrieve the data points, to which the Euclidean distance method is applied and classification of final clusters is obtained.

3.3. Classification and Prediction of Nearest Testing Center Using K-NN

K nearest neighbor algorithm can be used in controlling the spread of the coronavirus by locating the nearest test centers. The government has organized many test centers in each area to help people and to reduce the crowd. But people are not aware of the different test centers due to which the crowd in common centers will increase and the spread of COVID-19 might increase. To prevent and control this situation, it is necessary to provide and locate the nearest Covid-19 test centers. The trained data will have the information of existing testing centers based on longitude and latitude coordinates.

The datasets having patient's details are loaded and then region vise specifications including unidentified and recorded regions are provided. And the data is randomly split into testing and training sets using cross validation method and the K value is selected. If the data is large splitting of the data will help in finding the

k value better. Then the distance from the new test data to each trained data is calculated using Euclidean distance.

Then the model is fitted to training and target values after the K-NN classifier. After the fitting and region vise classification, the radians i.e., latitudes and longitudes of the test center and user from the map are selected and the distance between the latitudes and longitudes is calculated by bearing distance formula by implementing cosine function and the opposite of bearing distance is calculated using the horizontal distance formula. Then the difference between the default values altitude and opposite distance are calculated which provides the original distance for prediction

$$\theta = \text{atan2}(\sin \Delta\Lambda \,.\, \cos\varphi 2, \cos\varphi 1 \,.\, \sin\varphi 2 - \sin\varphi 1 \,.\, \cos\varphi 2 \,.\, \cos\Delta\Lambda) \quad \text{eq(5)}$$

where, $\varphi 1, \Lambda 1$ is the start point,

$\varphi 2, \Lambda 2$ is the end point,

$\Delta\Lambda$ is the difference in longitude and latitude

Bearing Distance Formula

$$\text{Base2} = \text{Base} * 1000$$

$$\text{distance} = \text{Base} * 2 + \text{Oppsite} * 2 / 2$$

$$\text{Caltitude} = \text{Oppsite} - \text{Aaltitude} \quad (6)$$

Horizontal Distance Formula

Then the prediction is made based upon the classification and calculations. Based on the selected latitude and longitude range, the prediction is done and the nearest test centers for that range are fetched and located.

The working flow of the developed application:

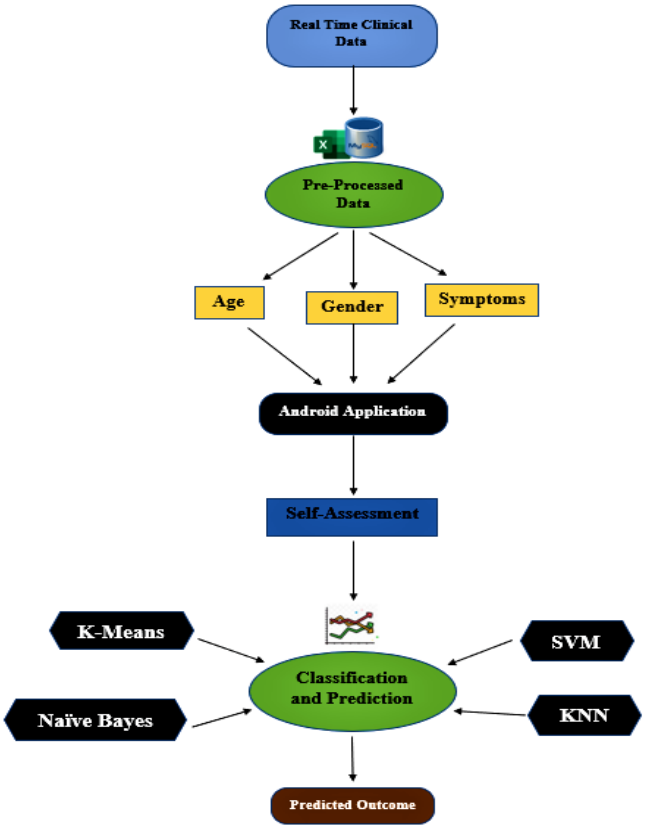

Fig. 1: Architecture Application Flowchart

We consider real-time COVID-19 infected patient's data; we pre-process the data to remove empty fields and unwanted fields as they can hinder the accuracy of the proposing model. The user enters the system anonymously by providing just their age and gender. Then the user undergoes self-assessment by answering the questions by providing appropriate details of symptoms, comorbidities along intensities. Based on the user's inputs, the Support Vector Machine algorithm classifies whether the user is infected with virus or not. Naïve Bayes algorithm helps to categorize the level of risk of the infection user is facing, that is low, mild, or high-level of risk. If the user is at low risk of the coronavirus infection, then they are suggested to isolate themselves and tips are given to boost their immunity. Else, the user's vital parameters are approximately predicted based on the generalized vital parameter range provided in collected datasets (trained data) using the K-means algorithm. Mild and high risk users are asked to register themselves to further book doctor's appointment in the nearest health center based on their current location which is given by K Nearest Neighbor Algorithm as shown in Fig. 1.

4. Result

The primary step to analyse any medical condition is to note down the symptoms faced by the patient and then the diagnosis of the condition the patient is suffering from. The system which we are proposing follows the same precedence.

Machine learning model will be trained with datasets containing symptoms and results of whether the patient has or not has COVID-19 infection. Further, the user input is taken of what symptoms the user is facing and those symptoms severities. These symptoms taken as input will be classified by the Support Vector Machine (SVM) as to whether the patient is infected with Coronavirus or not. The classification of whether the patient has COVID-19 or not is done by Support Vector Machine, whereas where does the patient stand with the infection, that is if the user is at low, medium, or high risk of infection is prediction job is done by Naïve Bayes Algorithm.

Support Vector Machine (SVM) classifies the new instance given as COVID-19 infected or not by drawing a hyper-plane and placing the new instance in one of the segregated classes. Naïve Bayes uses the prior probability and gives output as to whether the patient is at low, medium, or high risk of infection

Fig. 2: Patient inputting data

The above fig.2 shows the user symptoms input where the user will register by entering and inputs the symptoms and severities they are facing.

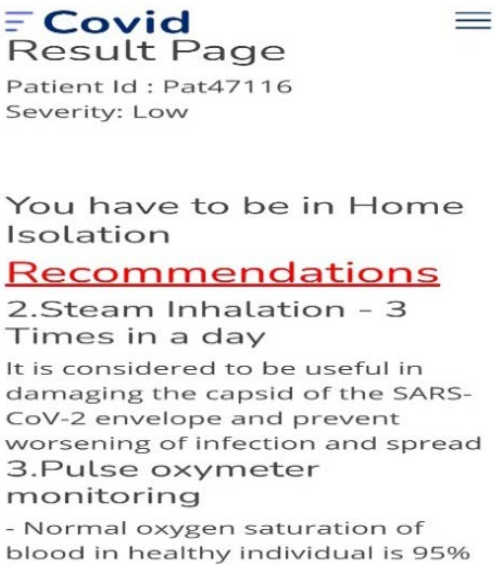

Fig. 3: User classification page

The above fig.3 shows the result of a user with a low risk of infection.

Based on the training datasets with which we have trained our system, symptoms are divided into dependent and independent symptoms. Any symptom which could alone contribute to the risk of infection is considered an independent symptom. In our system, cough, breathing problem, Diabetes, loss of taste and smell, hypertension, travel history, contact with an individual who is COVID-19 positive; these are the independent symptoms as they are alone enough to classify an individual as COVID-19 positive or not. Whereas, symptoms that might be or might not be faced only by a COVID-19 positive patient and those which also depend on other symptoms are considered as dependent symptoms. Symptoms such as age, gender are dependent symptoms because just by knowing the age and gender of a person we cannot conclude that the person is COVID-19 positive or not.

The above-proposed system, was partially developed and few users tested and based on user experience, feedback was taken. Fig. 4 shows the results of rating the prediction accuracy of the partially developed system.

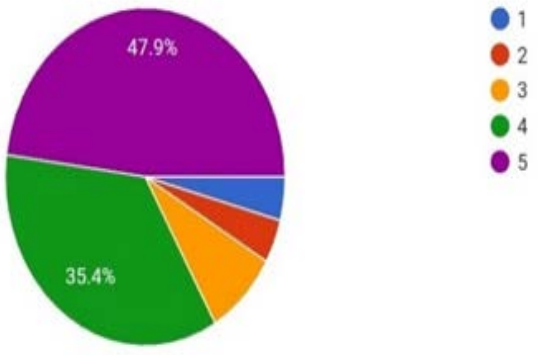

Fig. 4: Pie chart of the prediction accuracy rate

In the feedback, 47.9% of users gave the five ratings and 35.4% gave four ratings; a total of 83.3%. Looking at the numbers, we can conclude that the accuracy of the system is satisfactory. Further after classifying the user, based on the risk of infection user is directed to take different steps.

The next result in the proposed system is the classification of the Covid-19 vital factors with ranges, The increase in positive cases around the country is because of the transmission rate of the virus which is high.

Nowadays the virus is getting transmitted even with shorter contact time because of the new strain of the virus.

The vital factors are mainly involved in the matter of mortality, It is seen that the patient's age, decreased oxygen level which causes more respiratory rate and the comorbidities like diabetes have oddly increased mortality. Every individual should know if any fluctuations occur to the ideal ranges so that needed action can be taken. Therefore, the classification of the Covid-19 vital factors and its prediction with ranges is important as it also creates the needed awareness among the common people about the factors that are primarily causing the death of the individuals.

The prediction of the vital factors is done only for the user whose infection prediction comes at mild-level and high-level. The users predicted with low-level risk, are suggested to isolate themselves

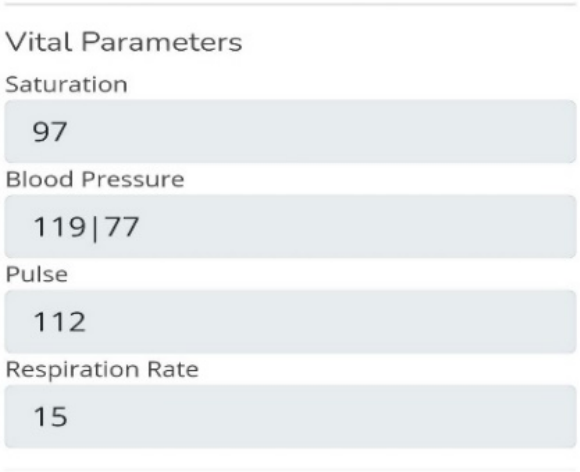

Fig. 5: Vital factors prediction of a user

The above fig.5 shows the classification and the prediction of the user, who is classified as the high-level risk with ranges for vitals factors, oxygen saturation with 97%, blood pressure between 77 mm Hg and 119mm Hg, pulse rate with 112 beats per minute and respiration rate with 15 breaths per minute

The k-means algorithm is used on a large set of data, and yet the accuracy rate is high as it assigns every data point to a cluster using the distance measure method. Thus, the accuracy of the K means clustering in this system is 91% - 95%.

In the proposed system, the next important step after diagnosing the condition, classifying users into different risk levels, and predicting the vitals of patients with a medium to high risk is locating the nearest test centers. Providing the nearest test centers location to the users will help in reducing the crowd at common centers and help in controlling the spread of coronavirus. The model will be trained with the datasets containing all required data for the prediction such as region, address, latitude, and longitude data.

After training the model, the radians i.e., latitudes and longitudes of the registered user are accessed from the map and the nearest test centers for their location or radians are predicted and located using K nearest neighbor(KNN). Further, the user will be also provided with the option of booking an appointment in the selected centers to seek further medical help.

For the new test data, the K nearest neighbor will select the k value and calculate the Euclidean distance from the new test data to each train data point and shortlist the nearest k neighbors. Then the distance between the latitudes and longitudes are calculated using bearing distance and horizontal distance and then with the obtained final distance value, the new test data is assigned to the category having the maximum number of votes, that is the prediction of nearest test centers.

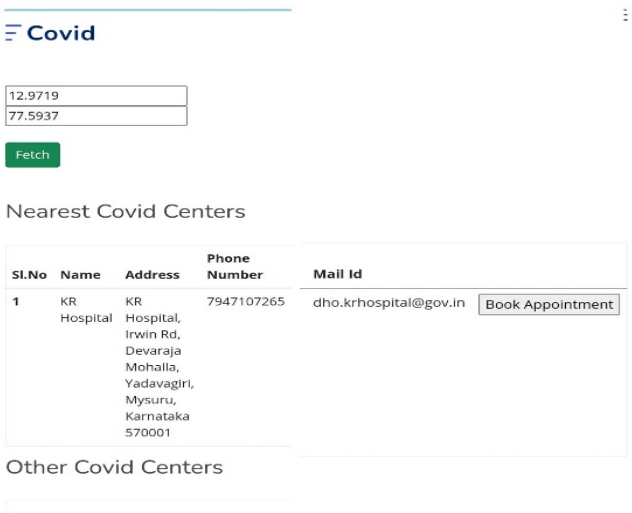

Fig. 6: Nearest Covid-testing center

The above fig.6 shows the user's latitudes and longitudes taken from the map. Then after the usage of KNN, it fetches and lists the nearest test centers for that particular radians.

Fig. 7: Other Covid-testing centers

The fig.7 shows, the other test centers are listed for users' convenience.

There is also an option provided to the users to book an appointment in the selected test center. This feature helps in reducing the crowd as a particular slot at a particular time will be provided to the users for medical help. The user will get an alert after booking an appointment.

K nearest neighbor performs well and provides highly accurate outcomes for a small-scale dataset. The prediction and the outcomes will be more precise and accurate as the quality of prediction depends on the distance measure. Thus, the accuracy of the K nearest neighbor in this system is 92% - 94%.

5. Conclusion

Above we are putting forward a system, which helps public to take up self-assessments if they have a doubt of infection and obtain quick results and help them take up further steps to avoid spreading of virus or deterioration of their health. Real-time data collected from a Government medical health organization and

transformed it into datasets, favorable for input to the algorithms, K-means, Naïve Bayes, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). The Machine Learning model developed using these algorithms, analyses the levels of infection of Coronavirus (COVID-19) in a person, and generates a report. On the grounds of the predicted outcome, the person is suggested to isolate themselves or to consult medical help. The above-proposed system resolves the issue of searching for medical help when needed and also how to take precautions, as it also informs the user of the medical health center nearest to their location and gives immunity advancement suggestions respectively. The real-time data that is fed to train the models implemented helps in predicting the risk of infection. Predicted vital parameters of each individual play a significant role in improving the accuracy of our system. Providing nearest test center location, reduces the crowd and help people and nation in controlling the pandemic situation. In total, this proposed system will improve the healthcare sector of the country and help containment of the contagious, lethal Coronavirus (COVID-19), by creating awareness amongst the public.

6. References

[1] Gururaj H L, H. L. Gururaj; B. C. Soundarya; V. Janhavi, "Machine Learning Algorithm for Covid-19 Prediction" *IEEE Technology Policy and Ethics* (Volume: 7, Issue: 1, Jan. 2022), **DOI:** 10.1109/NTPE.2022.9778144

[2] An Efficient Approach For COVID-19 Using Machine Learning Techniques' *International Conference on Innovative Computing and Communication.*

[3] D. Pelleg, A. Moore (2000): "X-means: Extending K-means with Efficient Estimation of the Number of Clusters"; ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning, pp. 727-734.

[4] Department of Economics,Payame Noor University, Tehran, Iran(2013) : Application of K-Nearest Neighbor (KNN) Approach forPredicting Economic Events: Theoretical Background6.Thanh Thi Nguyens

[5] Susmita Ray,Department of Computer Science & Technology Manav Rachna UniversityFaridabad, India(2019):A quick Review of machine learning.

[6] Vibhatha Abeykoon, Supun Kamburugamuve, Kannan Govindrarajan, Pulasthi Wickramasinghe, Chathura Widanage, Niranda Perera, Ahmet Uyar, Gurhan Gunduz,Selahattin Akkas and Gregor Von Laszewski,USA(2019):Streaming Machine Learning Algorithms with Big Data Systems

[7] Ziqing Guo, Qizheng Ye, Feixing Li and Yuwei Wang (2019): Study on Corona Discharge Spatial Structure and Stages Division Based on Visible Digital Image.

[8] Thanh Thi Nguyen Deakin University(2020):Artificial Intelligence in the Battle against Coronavirus (COVID-19): A Survey and Future Research Directions,DOI:10.13140/RG.2.2.36491.23846/1

[9] Quoc-Viet Pham1, Dinh C. Nguyen2, Thien Huynh-The3,Won-Joo Hwang4,5, And Pubudu N. Pathirana(2020):Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic:A Survey on the State-of-the-Arts,DOI10.1109/ACCESS.2020.DO

[10] Gros, C., Valenti, R., Schneider, L., Valenti, K., & Gros, D. (2020).Containment efficiency and control strategies for the Corona pandemic costs.

[11] Shreyas Setlur Arun,Ganesh Neelakanta Iyer(2020):On the Analysis of COVID19 – Novel Corona Viral Disease Pandemic Spread Data Using Machine Learning Techniques,DOI: 10.1109/ICICCS48265.2020.9121027

[12] Gros, C., Valenti, R., Valenti, K., & Gros, D. (2020). Strategies for controlling the medical and socio-economic costs of the Corona pandemic.