JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF COVID CASES BY APPLYING CNN ALGORITHM ANALYZING COVID-19 DATA

Dr. M. Dhanalakshmi. Professor Of Information Technology Department, Jawaharlal Nehru Technology University Hyderabad College Of Engineering, Nachupally (Kondagattu), Jagityal Dist-505501.

Damalla. Mamatha Rani. Pg Scholar, Department Of Information Technology, Jawaharlal Nehru Technology University Hyderabad Collage Of Engineering, Nachupally (Kondaguttu), Jagityal,Dist-505501.

Abstract - The Corona virus (COVID-19) is creating panic all over the world fast growing cases. There are various datasets available which provide information of world-wide affected information.

Covid-19 has affected all counties with a large number of cases with variation of numbers under death, survived, effected. In this project we are using a New York data set which has country wise details of cases with various combined features and labels. Covid-19 data analysis and case prediction project provide solutions for data analysis of various countries at various time and data factors and creating models for survival and death cases and prediction cases in future.

Machine learning provides deep learning methods like Convolution neural networks which are used for model creation and prediction for next few days are done using this project.

1. INTRODUCTION

1.1 Introduction

The new corona virus first occurred in the Chinese's city of Wuhan in Two thousand and nineteen and was reported to the World Health Organization (W.H.O) IN December 31, 2019. W.H.O. called the virus COVID-19 ON February 11, 2020, after it established a global danger. Depending on respiratory syndrome, a COVID-19infected individual; will develop symptoms between 2-14 days (SARS). Dry cough, exhaustion, and fever are indications and symptoms of mild to moderate instances, according to the World Health Organization, while shortness of breath, fever and tiredness are signs and symptoms of severe cases. People who have other illnesses such as asthma, diabetes, or heart disease, are more susceptible to the virus and may become seriously ill as a result. The person's symptoms and travel history are used to get a diagnosis. The vital signs of the client who is experiencing symptoms are being closely monitored. Regularly wash hands with soap for 30 seconds and abiding close contact with people by

keeping a distance of approximately 1m may lower the risk of infection. Covering the mouth and nose with disposable tissue during sneezing and avoiding contact with the nose, ear, and mouth can help prevent it. SARS is an airborne disease that first arose in China in 2003 and has since spread to 26 countries, with million cases reported in the same year and spread from person to person. ARDS (acute respiratory distress syndrome) is defined by the quick development of inflammation in the lungs that leads to respiratory failure, with blue skin colour, exhaustion, and shortness of diagnosing diseases using picture and textual data.

For the detection of new corona viruses, machine learning can be applied. It may also predict the virus's characteristics to identify or forecast illnesses. For categorizing text or images into distinct categories, supervised machine learning algorithms require annotated data. Over the last decade, significant progress has been achieved in this sector in order to complete certain key projects. Data in the form of X-ray pictures was supplied by John Hopkins University, and multiple researchers developed a machine learning model that identifies X-ray images as CPVID-19 or not.

Using Methods and Algorithms:

Convolutional neural networks (CNN) are a form of artificial neural network used in computer vision. CNN was used to decrease the size of pictures using convolution; and pooling layers before sending the reduced input to fully connected layers.

The information consists of clinical records in the form of written reports, which we categorize into four illness categories: confirmed cases, deaths cases, recovered cases, and active cases. It can aid in the prediction of corona virus based on previous clinical findings. On COVID to COVID-19 Next 10 Day Prediction World Wide Cases, we employed supervised machine learning approaches to give deep learning in CNN model for classifying the text into four separate categories.

1.2 Problem Statement

A large number of individuals were affected. The domestic outbreak is currently under control, but the new corona virus is fast spreading in other locations. A new pneumonia epidemic was proclaimed a "global pandemic" by the World Health Organization (WHO) on March 11.

With the rising number of COVID-19 cases throughout the world, daily forecasts and analyses are needed to keep the pandemic under control.

1.3 Objective

Data pre-processing and data analysis are done on the dataset, and a machine-learning model is constructed for future case prediction using data from Kaggle and NEW YORK dataset.

2. LITERATURE SURVEY

2.1 Used Methods

- Two thousand and nineteen Pneumonia in Wuhan, China: Risk Factors Associated with (ARDS) Deaths in Patients with Corona virus Disease A machine learning algorithm that can predict if a person has COVID-19 and is at risk of developing acute respiratory distress syndrome (ARDS). The proposed model was shown to be 80% percent accurate.
- The findings of diabetes diagnosis using machine learning and ensemble learning approaches showed that the ensemble methodology ensured 98.60% percent accuracy. These functions may be useful in diagnosing and predicting COVID-19.
- By providing an accurate diagnosis, machine learning and deep learning can replace people. X-rays and computed tomography (CT) scans may be used to train the machine learning model, which can save radiologists' time and be more cost-effective than traditional testing for COVID-19.
- COVID-Net, a deep convolutional neural network that can diagnose COVID-19 from chest radiography pictures, was developed. When COVID-19 is found in a person, he or she is considered to have COVID-19.

2.2 Existing System

- For pattern detection, explanation, and prediction, big data-based models are used in machine learning and natural language processing. In recent years, NLP has attracted a lot of attention. Classification is one of the most important jobs in test mining, and it may be accomplished using a variety of the methods [6]. For mining unstructured data, Kumar et al conducted a SWOT analysis of several supervised and unstructured text classification systems.

- Sentiment analysis, fraud detection, and spam detection are some of the uses of text categorization. Opinion mining is mostly utilized in elections, advertising, and business.

- Sarwar et al. [10] used machine learning and ensemble learning approaches to diagnose diabetes. The results showed that the ensemble methodology had 98.60% percent accuracy rate. The functions may be useful in diagnosing and predicting COVID-19. From the COVID-19 we can save millions of lives and generate a massive amount of data on which machine learning (ML) models can be trained if it is diagnosed correctly. Machine learning might in this sense, especially when establishing diagnosis based on clinical writing, radiographic images, and other data.

- Machine learning and learning, according to Bullock et al. [11], can replace people by providing correct diagnosis. A precise diagnosis can save radiologists time and money when compared to routine COVID-19 testing. The machine learning model may be trained using X-rays and computed tomography (CT) data. In the regard, a number of initiatives are in the works.

- COVID-Net, a deep convolutional neural network built by Wang Wong [12], can diagnose COVID-19 from chest radiography pictures. When COVID-19 is found in a person, the question is whether or not that individual will be impacted, and if so, how severely. Not all COVID-19-positive individuals will require close monitoring. Knowing who would be affected more severely can aid in directing assistance and planning the allocation and usage of medical resources.

- Using data from (only) 29 patients at Tongli Hospital in Wuhan, China, Yan et al. [13] developed a prognostic prediction algorithm to forecast the mortality risk of a person who has been infected.

- Jiang et al. [14] created a machine learning model that can predict whether or not a person is infected with COVID-19 and acute respiratory distress syndrome (ARDS).

- The suggested model achieved an accuracy of 80%. A total of patients was utilized to train their model, and they were only allowed to use samples from two Chinese hospitals. ML can be used to diagnose COVID-19, a virus that requires a lot of study but is not yet extensively utilized. We employed machine learning and ensemble learning models to categorize the clinical reports into four types of viruses because there is less work being done on diagnosis and prediction using language.

2.3 Drawbacks

- Using a machine learning model that can predict whether or not a person is impacted by Covid-19. However, they label data, which is constant data, and they can simply forecast the data because it is in numerical format.
- In diabetic patients who are easily infected with COVID-19, machine learning can be used to diagnose diabetes. As a result, they solely forecast diabetes patients.
- COVID-19 pictures, like X-rays and CT scans, can only determine whether or not a person is infected with corona virus

3. PROPOSED SYSTEM

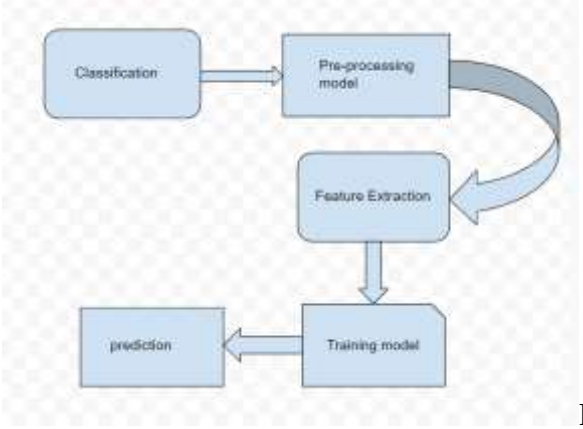


Fig 3:

Pre-processing methods

A CNN model prediction is incorporated in our model to capture a common knowledge of viral transmission pathways, including COVID-19. Furthermore, the anticipated indices may be derived as model features from them. The goal of event trends detection is to compensate for the CNN model's shortcomings and forecast the outcomes for the following 10 days.

Advantages of Proposed System

- COVID-19-related event patterns are being detected and shown in a new way.
- By merging an epidemic dynamic model with top event trends, graph statistics from the study's findings are used to forecast the number of confirmed cases and fatalities.
- Finally, the suggested prediction method has been demonstrated to be useful; in real-world practice by determining the best day to return to work.

3.1 CLASSIFICATION

3.1.1 Working with CNN:

Deep Learning Models in CNN was the first to employ convolutional networks for zip code recognition in their present form. Convolutional, pooling, fully connected layers are the most common layers on CNNs. In the convolutional layers, a set of feature maps, also known as activation maps, is created. Each neuron in the feature map is connected to the input layer corresponding to reducing the number of parameters significantly when compared to a fully connected neural network. In The most common CNN architectures, pooling layers alternate with convolutional layers to decrease the spatial dimension of the feature maps in preparation for the succeeding computational stages, decreasing computing cost and minimizing overfitting.

Fully connected layers combine the generated feature maps and produce a classification measure at the network's conclusion after an arbitrary number of previous levels.

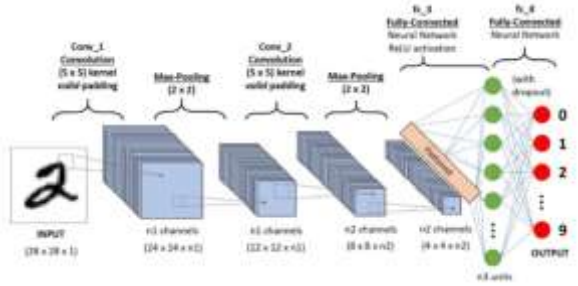


Fig 3.1: CNN Architecture

3.1.2 Function Model:

The following is the quick summary of each layer: Input layer: Receive input of COVID-19confirmed cases. The first Cov1D layer scans over the input sequence, acquiring new information and dealing with noise in the data before projection the results onto feature maps.

The second Conv1D layer repeats the procedure on the feature maps created by the first, attempting to improve any notable features. To read the input sequences, we used 64 feature maps per convolutional layer and a kernel size of three-time steps. The max pooling layer streamlines the process by deleting particular values from the convolved features. The feature maps and creates a smaller-dimensional matrix.

Dropout layer: The dropout layer was introduced to the network to prevent the model from overfitting. The random subsampling of a layer's outputs under dropout has the effect of

reducing the network's capacity of thinning it during training.

Flatten layer: The distilled feature maps are flattened into a single long vector that may be used as input to the decoding process after the dropout layer.

3.2 PRE-PROCESSING DATA:

The data was then analysed in a Jupyter notebook using a python project that used the libraries Pandas, NumPy, SciPy, and Matplotlib. We filtered data from Morocco first, then chose the characteristics 'total cases,' 'population,' 'total deaths,' and 'new deaths.' We kept the final one.

- Because it is the objective of our forecasts, we have chosen to highlight 'new cases.' Because contamination might occur from sick persons who are constantly alive, the characteristics describing the overall number of deaths and cases are relevant to prognosis.
- The next step was feature scaling, which included normalizing the range of independent variables or data characteristics so that each feature contributes roughly and proportionally to the ML algorithm. The Min-Max scaler, which converts all values between 0 and 1, was then applied.
- The fourth job of the pre-processing procedure is to change our model to learn from previous time-steps in order to forecast positive COVID-19 cases for the next 10 time-steps, where the dataset is split into 80% and training and 20% testing. Two days are used as inputs, and the model predicts ten days as outputs
- The data was then analysed in a Jupyter notebook using a Python project that used the libraries Pandas, NumPy, SciPy, and Matplotlib. We started by filtering data from India, then choosing the characteristics 'total cases,' 'population,' 'total deaths,' and 'new deaths.' The last feature, 'new cases,' has been kept because it is the goal of our forecasts. Because contamination might occur from sick persons who are constantly alive, the characteristics describing the overall number of deaths and cases are relevant to prognosis. The next step was feature scaling, which included normalizing the range of independent variables or data characteristics such that each one contributes roughly and proportionally to the algorithm.
- The fourth task of the pre-processing process is to adapt our model to learn from previous time-steps, with total confirmed cases as features and total log of 10 exponential count as label, and these values used to fit the algorithm and train the model.
- The model takes prior days as inputs and returns future forecasts.

3.2.1 What we are using in data:

The WHO COVID-19 dashboard the data for this analysis. It contains information on corona virus cases in each specific country, such as the number of confirmed, dead, cumulative confirmed, and cumulative deaths (defined by the of the country, country code, and WHO region) every day from the start of the COVID-19 infections (4/1/2020) to the end of the COVID-19 infections (24/9/2020). At the time of this research, the dataset contained 62,510 records for 216 different nations and 265 days, totalling 31,798,308 new cases and 973,653 deaths cases. The regional distribution of cases accumulated three distinct timestamps is depicted. COVID-19 confirmed and mortality cases are presented globally.

3.2.2 Data collection:

The World Health Organization has designated the Corona virus pandemic a public health emergency. The data pertaining to this pandemic is freely available from researchers and hospitals. We gathered information from GitHub, an open-source data repository. This is the location where patient data is kept. The data includes the country's name, the date, the source URL, the source label, the number of observations, the cumulative total, the cumulative total per thousand, and a description. The data sets supplied by Johns

Hopkins CSSE1 include worldwide pandemic figures beginning on January 22nd. Different nations and areas are represented in the data sets. The data set is divided into three categories: daily confirmed cases in each nation or area, daily fatalities in each country or region, and daily recovered cases in each country or region.

3.2.3What are processing Data:

Preparing time series data, creatine predictive models, and implementing the predictive model are the three steps of the suggested research technique. The time series "New COVID-19 cases" was used. There are three phases in the first phase, "Preparing time series data": Convert a dataset into a time series, normalize data in a time series, and divide data in a time series. Three phases make up the second phase, "Building the Predictive Model": Optimize the models, train the models, and assess the models. Models were optimized to obtain the best hyperparameter, models were then trained on a train set using the best hyperparameter, which began on January 4, 2020, when the first instance of COVID-19 began, and ended on July 17, 2020.

The trained models were then put to the test on s test set, which runs from July 18, 2020 to August 14, 2020. Forecasting was approximated and compared to real values between July 18, 2020 to August 14, 2020. Third step, "Applying the Predictive Model," includes COVID-19 data from August 15, 2020 to September 18, 2020 end. Forecasting was estimated and compared to real values between September 12, 2020 and September18, 2020.

3.2.3 Dataset

From the data set or data collection, we first define confirmed cases from all states, countries, or regions, and then we take death cases from all states, countries, or regions, as well as recovered cases. We will not be active cases any more. Calculate the data you will need.

Active cases=confirmed cases-death cases-recovered cases then we had active case data on pre-processing.

At the time of this research, the dataset had 62,510 records for 216 different nations and 265 days, totaling 973,653 deaths. Cases that have accumulated at three separate time stamps are distributed spatially. COVID-19 instances are expected to be widespread all over the world

3.3 FEATURE EXTRACTION

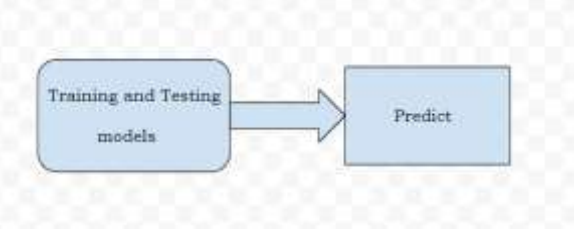
The dataset we utilized was "New cases" (COVID-19 confirmed cases). We set the date as parse dates when opening the csv file to make the dataset a time series. The loaded dataset was resampled and grouped by day using the resample () function with the argument "D." A total of 265 data points were divided into 37 weeks of data time-series data from COVID-19 confirmed cases were normalized using min-max normalization within the range [0,1] normalization must be reversed. Various aspects of India, Sri Lanka, New York, and other countries may be found in the pre-processed reports. We have more than the entire count of 400 persons in India (100), Sri Lanka (200), and New York (100).

They split the 400 traits of persons into two categories, with 80% and 20% picked. Deep learning in CNN models is achieved by assigning the appropriate weight to the feature and feeding the same input to machine learning algorithms.

3.3.1 Extraction

Total confirmed cases are used as features and total log of 10 exponential count is used as a label, with this data being used to fie the algorithm and train the model. The model takes prior days as inputs and returns future forecasts.

3.4 TRAINING AND TESTING MODEL



3.4(a) Trainig and Testing models

To prepare the time series for model development, the NumPy split () function was used to separate 37 weeks of normalized time series data into training, test, and forecast sets in a sequential manner. Table 3.2.3 illustrates the outcome of the splitting method in detail, whereas Figure 3.4.2 depicts it as a graph. The above-mentioned CNN Algorithm is used for training.

The 18 forecasting models developed in this study were chosen because picking the best and most accurate forecasting model for anticipating the COVID-19 pandemic is a difficult procedure.

Models were trained on a train set, which is a time series that begins on January 4, 2020 and ends on July 17, 2020 using the optimal hyperparameter.

3.4.1 Evaluate the models

The trained models were then put to the test on a test set, which runs from July 18, 2020 to August 14, 2020. The forecasting horizon is crucial for an intelligent model's prediction accuracy from a forecasting model analysis when predicting the next daily verified case is referred to as the forecasting horizon.

The train and prediction procedures were conducted ten times to account for unpredictability. Average forecasting between July 18, 2020 and August 14, 2020 was computed and compared to the true values. The models' performance was then evaluated using testing data.

In the training model split data set to labels of 80% total data for training and 20% of total data on testing. In that we take a deep learning classifier on using CNN modal.

For Example, on feature extraction of data we take that by training on each training segment of test and training label are having testing taken 20% and training 80% data taken from data set on web site.

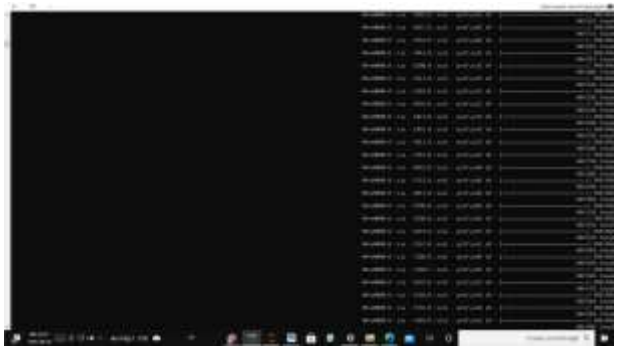


Fig 3.4(b) Training Data

3.4.2 Testing model:

In the test model we are testing the dataset we've gathered to the test. Divide the data into confirmed cases, death cases, recovered cases, and active cases in the dataset. And the data may be found at

We forecast the model based on the data set supplied to the train mode.

3.4.3 Validation and Cross validation:

This data does not allow validation or cross validation. The predicted data in our daily scenarios varies from one another since the dataset is in the format of a textual data collection. As a result, cross-validation isn't possible.

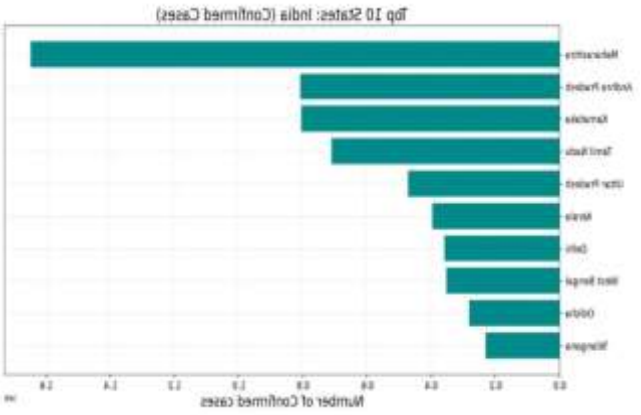
3.5 PREDICTION:

Predictions from the CNN model, as well as highlights from recent event patterns. State (City), Country (Region), and Date are all included in the original data. The month and day are retrieved and turned into numeric codes for date. Susceptibility prediction, confirmed prediction, recovered prediction, and fatality prediction are all features of the CNN model. This model is used to provide forecasts for the next 10 days

4. RESULTS

Using data from the COVID-19 India page on GitHub.com, we predicted the total number of confirmed cases in India. And we're forecasting India's top 10 states.

Graph fig:

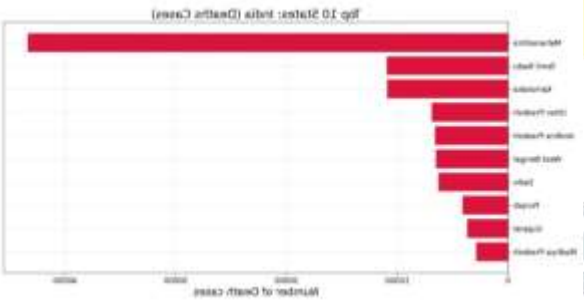


4.1 Number of confirmed cases top 10 states

The graph below depicts confirmed instances in India. And Maharashtra got afflicted with covid-19 in large numbers. Telangana, meantime, was gradually infected with covid-19.

This graph depicts the CNN model's training process, after which the model is saved after being trained with a global dataset of all nations' survival and death instances.

Graph fig:

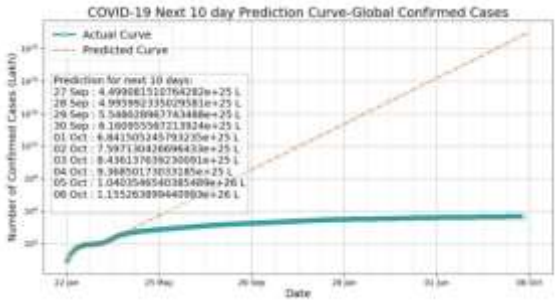


4.2 Number of death cases in top 10 states

The graph below depicts mortality cases in India. And Maharashtra has a high number of death cases linked to covid-19 infection. And on Covid-19, Madhya Pradesh had the fewest infection-related deaths.

The graph depicts future case predictions throughout the work in the form of a graph, which is forecasted using a CNN trained model.

Graph Fig:



4.3 Covid-19 Next 10-day prediction World Wide cases

Because many factors influence the spread of the COVID-19 virus, such as city size, heavy population density, and total population, different models for the COVID-19 pandemic period in different countries and regions should be established, so the data is divided according to each country and region for simulation.

4.1 Result and Discussion

We started by pre-processing the data set on a non-relevant data set that isn't acceptable for testing and training on a pre-processing model. A list format in which each element in a single document is listed. In addition, the papers CC model the document with similarity score from the data set or supplied query after pre-processing many pre-processing approaches.

Treatment of number of confirmed:

The number of confirmed cases in the data pertains to the total number of confirmed instances each day, hence it has to be processed. In the event of an epidemic, after a verified patient dies, his or her body will be disposed of swiftly, ensuring that the dead patient is no longer contagious to normal people. As a result, the data's proven case is handled as follows:

Active cases= confirmed cases - recovered cases - deaths cases.

5. CONCLUSION

Data and Information about the novel corona virus and the outbreak's progress become available at an unprecedented rate in today's digital and globalized society. Still, critical concerns remain unsolved, and precise solutions for anticipating the outbreak's dynamics are just impossible to provide at this time. We emphasize the ambiguity of available official statistics, especially in terms of the real baseline number of infected patients.

We covered the whole global case count in this project and forecasted future cases based on global data in the following days.

We used an Indian dataset, obtained from the COVID-19 India website, to train a CNN model and analyse the mortality and survival rates in India.

5.1 Future scope:

In future study, we may utilize a global dataset to analyse each nation's data and forecast confirmed and fatality cases by country, as well as assess the accuracy of each model for data from various Indian states and predict cases.

REFERENCES

- Johns Hopkins Researchers Publish COVID-19 'Prediction Model' https://www.hopkinsmedicine.org/news/newsroom/news-releases/johns-hopkins-researchers-publishers.
- World health organization: https://www.who.int/new-room/g-adetail/q-a-corronaviruses#:/text=symptoms. Accessed 10 Apr 2020
- Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ (2020) A new coronavirus associated with human respiratory disease in China. Nature 44(59):265–269
- Medscape Medical News, The WHO declares public health emergency for novel coronavirus (2020) https://www.medscape. com/view article/924596
- Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y, Xia J, Yu T, Zhang X, Zhang L (2020) Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 395(10223):507–513
- World health organization: https://www.who.int/new-room/g-adetail/q-a-corronaviruses#:/text=symptoms. Accessed 10 Apr 2020
- Wikipedia coronavirus Pandemic data: https://en.m.wikipedia. org/wiki/Template:2019%E2%80%9320_coronavirus_pan Demic data. Accessed 10 Apr 2020
- Khan Day, A.M.U.D., Amin, A., Manzoor, I., & Bashir, R., ''Face Recognition Techniques: A Critical Review'' 2018
- Kumar A, Dabas V, Hooda P (2018) Text classification algorithms for mining unstructured data: a SWOT analysis. Int J Inf Technol. https://doi.org/10.1007/s41870-017-0072-1
- Verma P, Khan Day AMUD, Rabani ST, Mir MH, Jamwal S (2019) Twitter Sentiment Analysis on Indian Government Project using R. Int J Recent Tech Eng. https://doi.org/10.35940/ijrte.C6612.098319

.