 doi:10.20944/preprints202004.0257.v2

Data Analytics: COVID-19 Prediction using Multimodal Data

Parikshit N. Mahalle[1], Nilesh P. Sable[2], Namita P. Mahalle[3], Gitanjali R. Shinde[4]

[1] Senior Member IEEE, Professor and Head, Department of Computer Engineering, STES'S Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India-411041

[2] Chief Examination Officer and Associate Professor, Department of Computer Engineering, JSPM's Imperial College of Engineering and Research, Wagholi, Pune, Maharashtra, India-412207

[3]Consultant Biochemist, Pathology Department, Deenanath Mangeshkar Hospital and Research Center, Pune, Maharashtra, India - 411004

[4] Assistant Professor, Department of Computer Engineering, STES'S Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India-411041

{[1]aalborg.pnm@gmail.com, [2]drsablenilesh@gmail.com,[3] pnmahalle@gmail.com,[4]gr83gita @gmail.com}

Abstract

Globally, there is massive uptake and explosion of data and challenge is to address issues like scale, pace, velocity, variety, volume and complexity of this big data. Considering the recent epidemic in China, modeling of COVID-19 epidemic for cumulative number of infected cases using data available in early phase was big challenge. Being COVID-19 pandemic during very short time span, it is very important to analyze the trend of these spread and infected cases. This chapter presents medical perspective of COVID-19 towards epidemiological triad and the study of state-of-the-art. The main aim this chapter is to present different predictive analytics techniques available for trend analysis, different models and algorithms and their comparison. Finally, this chapter concludes with the prediction of COVID-19 using Prophet algorithm indicating more faster spread in short term. These predictions will be useful to government and healthcare communities to initiate appropriate measures to control this outbreak in time.

Keywords: COVID-19, Predictive Analytics, Machine Learning, Prediction, Pandemic.

 doi:10.20944/preprints202004.0257.v2

1 Introduction

A novel human corona virus was originated from China in December 2019, causing a severe potentially fatal respiratory syndrome (COVID-19). The symptoms of COVID-19 may or may not be visual in infected individual hence spread rate can be faster as individual himself not aware of the infection [1]. Despite of the continuous efforts, the virus has managed to spread in most of the territories in the world; World Health Organization (WHO) has announced COVID-19 as Pandemic [2, 3]. Countries throughout the world working cooperatively and openly with one another and coming together as an united front in regards of efforts to bring this situation under control using available Information, Communication and Technologies (ICT). ICT need to be critically used to bring the situation under control and predictive analytics can be empowered using ICT services, tools and applications. ICT can empower epidemiological study to find the determinants, occurrence, and distribution of health and disease in a defined population in terms of COVID-19.

As study [4] shows that, 5% to 80% of people are tested positive for SARS-COV-2 may be asymptomatic. Predictive analysis using ICT plays an important role as some asymptomatic cases will become symptomatic over a period of time. Artificial Intelligence (AI) can be beneficial tool to fight against pandemic like COVID-19. AI models can be used for estimating and predicting spread rate, AI also used in the past pandemics like Zika-virus in 2015. Due to accurate and fast predictions spread rate can be minimized by taking necessary precautionary action before the time.

In nutshell, taking into consideration the current scenario a sad reality of the COVID-19 pandemic is that many people have been infected. As per the daily situation report of WHO as on April 09, 2020 the COVID-19 transmission scenario reports 1436198 confirmed cases with 85522 deaths globally. The main contribution of this chapter is comparison of various predictive analytics models and algorithms and their applications to appropriate use cases. This study recommends Prophet Machine learning algorithm for prediction due to various reasons which are discussed in section 4 of this chapter. Another contribution of this study is to present various avenues to initiate high-quality research in biomedical science along with integrative approach of predictive analytics and mathematical modeling to control outbreak of any pandemic. In the view of above mentioned related issues we should also promote ecumenical and interfaith collaboration and peaceful coexistence during the COVID-19 pandemic.

The main objectives of this chapter are as follows:

1. To understand medical perspective of COVID-19 towards epidemiological triad.
2. To analyse state-of-the-art for different approaches and models used for forecasting and prediction.
3. To understand various predictive analytics models and algorithm as well as their comparison with respect to the use cases.

 doi:10.20944/preprints202004.0257.v2

4. To study the performance of Prophet Algorithm for prediction of COVID-19.

The remainder of this chapter is organised as below. Section 2 presents medical perspective of COVID-19 in terms of its origin, most infected underline age group and transmission. Section 3 discusses the analysis of different studies available in the literature for predicting COVID-19. Section 4 presents various predictive analytics models, algorithms and their comparison. Finally, section 5 concludes the chapter with future outlook.

2 Medical Perspectives

The emergence of SARS-CoV-2 is confirmed from Wuhan's Huanan Seafood market, China, but specific animal source still remains uncertain. There is uncertainty regarding origin of SARS-CoV-2 [5]. The situation with SARS-CoV-2 is developing faster with the numbers of infected cases and death is increasing exponentially. The unprecedented control measures taken have been effective in preventing spreading of SARS CoV-2. Still, there is continued rise in number of cases with infection of SARS CoV-2. Hence, it is essential to identify that the increase is due to infected cases before lockdown, due to community transmission; hospital acquired infection or spread within family. This should be determined experimentally, which may help in revealing the actual numbers of infected patients and asymptomatic carriers.

Many studies have confirmed transmission amongst human of SARS-CoV-2 [6, 7], but mechanism of transmission and pathogenesis in spreading in humans remains to be fully explored? During transmission from human to human, whether the pathogenicity of this virus is decreased with the increase in rate of transmission? If the transmission of this virus is declined, the outbreak may eventually end. Nevertheless, if there is continuous and effective transmission, SARS-CoV-2 will develop into an additional human coronavirus which is community acquired. It is difficult to recognize and take further actions in patient with undefined and mild symptoms. Studying a group of asymptomatic infected cases, and follow them for their clinical presentation, titers of antibody and viral loads, will help in understanding about the number of subjects have symptoms later, whether viral shedding is actually less robust and how frequently asymptomatic carriers can transmit virus further. A study reported that asymptomatic infection is high (15.8%) in children under 10 years [8].

COVID-19 can be spread through respiratory droplets or due to close contact with the infected patients. SARS-CoV-2 was isolated from fecal samples of infected patients, which supports the significance of feco-oral route in the transmission of SARS-CoV-2, but a WHO-China joint commission report has denied this route of transmission [9]. However, the likelihood of transmission of SARS-CoV-

2 through human waste, contaminated water, aerosols and air conditioners cannot be underestimated, this may have happened in case of Diamond Princess cruise ship, where there was widespread COVID-19 Infection [10] . Still, to confirm the role of feco-oral transmission of SARS-CoV-2, further studies will be required. Severe cases caused by infection of SARS-CoV-2 may develop neurological, respiratory, gastrointestinal and hepatic complications leading to mortality. Many Studies have reported low sense of smell and taste as a manifestation of COVID-19 [11], but whether this is a unique feature of COVID-19 is uncertain. Till date we do not have definite anti-viral drug or vaccine for SARS-CoV-2. However, screening of new drug molecules may prove beneficial in treating COVID-19, which will have therapeutic effect.

Globally there has been lot of progress in monitoring and control of disease spread. It is evident that, there are lot of uncertainties and questions regarding transmission mechanism, asymptomatic or subclinical patient's virus shedding, origin of virus, virus pathogenesis, treatment, symptoms, etc. This highlights the need of integrative approach of predictive analytics and mathematical modeling with biological science, which may help government to take appropriate measures and method for future preparedness in fighting against this outbreak. In spite of rapid progression in research towards this outbreak, most of the studies are unable to suggest and guide effective measures to control this current situation. However, more high-quality research in biomedical science along with predictive analytics and mathematical modeling is warranted to manage public health crisis in short and long term.

3 Related Works

As per the Italy official release, there are total 27980 infected cases and 2158 deaths of people who were positive of COVID-19 [12]. Due to rapid spread of COVID-19, in short time many studies have been carried out for prediction of trend and its impact. This section briefs about all such recent studies which are essentially related to predictive analytics. Giulia Giordano et al. [12] proposes epidemic prediction model that compares infected density and the level of symptoms. Authors have proposed a SIDARTHE Model which helps to redefine reproduction number and simulation results also shows that the proposed model gives accurate results after comparing the findings with real data on the COVID-19 epidemic in Italy. Melanie Bannister at al. [13] presented an interesting study to establish the correlation of temperature and evidence of COVID-19 in Europe. Authors claim that, the seasonal variation; essentially in the temperature greatly impact the spread of COVID-19. Study states that, higher average temperature is potential candidate to limit the spread of COVID-19. Lucia Russo et al. [14] presented a mechanism to find the first day of infections and predictions of COVID-19 in Italy. Depending upon proposed work, authors are able to estimate that the actual count of exposed cases of COVID-19. Vitaly Volpert et al. [15] nicely presented

 doi:10.20944/preprints202004.0257.v2

the effect of quarantine model on the spread of coronavirus infection using data analytics. The main aim of this work is to present the assessment of placed quarantine mechanism using mathematical modeling.

Albertine Weber et al. [16] presented the trend analysis of COVID-19 pandemic in China using globally accepted SIR model in this study. The dataset used in this study is taken from Johns Hopkins University site for analysis and it is found that epidemic was contained in China. The basic aim of the study presented by Feng Zhang et al. [17] is to provide control measures to be considered internationally for global control of this pandemic. The time frame of dataset is from 3-10 February, 2020 and authors presented a time-series model to predict number of infected cases and the turning point where the spread is at peak. Feasibility analysis of controlling COVID-19 spread by isolating infected cases and quarantine is presented by Joel Hellewell et al. [18]. The proposed probabilistic model presented in this study considered varied scenarios like initial infections, basic reoccur number, and probability of contacts traced and rate of clinical infections. The results show that, in epidemic situation, isolation of infected people and contact tracing is not sufficient to minimize the rate of spread. Modeling of COVID-19 epidemic in China for cumulative number of infected cases using data available in early phase based on differential equation is presented by Z. Liu et al. [19]. . Simulation results show that, if the restrictions would have been applied one week before, then there would have been significant reduction in the number of infected cases.

Various ML models are discussed in the literature however for better accuracy deep leaning models can be used for better predictions [20, 21, 22]. Furthermore, predictions can be more accurate using active learning models in this multitudinal and multimodal data used for predictions instead of single type of data [23]. In addition to this, early forecasting of COVID-19 from small dataset is presented by Fong et al [24]. Simon James Fong et al. [25] have also proposed to use Composite Monte-Carlo simulation forecasting method for helping government to initiate critical actions and decisions to control spread of novel coronavirus. Experimental results using deep learning-based composite Monte-Carlo with fuzzy rule induction shows that decision makers are benefited more in the form of better fitted Monte-Carlo outputs.

All the studies discussed above are centric towards prediction and forecasting of COVID-19 based on short term data available on this pandemic. Literature shows that various mathematical and stochastic theory based approaches are used for estimation and prediction of spread rate of COVID-19. Most of the studies are giving expected predictions. There are so many predictive analysis models, such as Susceptible-Infection-Recovered (SIR) [26] and Hospital Impact Model for Epidemics (CHIME) [27] which has been working from decades. The SIR models work best in the case where data is not dynamic. In COVID-19 there is frequent change in data hence learning model can be suitable for analysis of pandemic data like COVID-19. Prediction of number of hospitals and facilities i.e. hospital beds, ventilators is also equally important. In the view of this, predictive Healthcare team developed COVID-19 CHIME model at Penn Medicine. These predictions can help hospitals to be prepared for worst case scenarios.

 doi:10.20944/preprints202004.0257.v2

4 Predictive Analytics

Predictive analytics is specialized branch of data analytics for making better predictions using past data and using analysis techniques which includes statistical and learning methods. Discovery of patterns in input data and anticipating what is likely to happen is the man objective of predictive analytics. Statistical analysis, predictive modeling and machine learning are three main pillars of predictive analytics. The main capabilities of predictive analytics are *statistical analysis, predictive modeling, linear regression and logistic model.*

Selection of appropriate predictive model and algorithm decide how efficiently we can make the better insights and useful decisions. Use case like hospital interested in prediction of number of patients likely to be admitted in intensive care unit in next seven days and prediction of fraud transaction for online banking provider might require different predictive analytics model than for predicting defaulter applicant for loan provider and predicting number of COVID-19 infected patient in next 10 days. Selection of appropriate predictive model is based on what predictive question would you like to address and how optimization can be carried out using predictive algorithms. The major pillars of predictive analytics are listed below:

1. Predictive Analytics Models
2. Predictive Analytics Algorithms

Predictive Analytics Models

Classification models are best for decision problems where the answer is merely Yes or No. This model classifies data into multiple categories using past data and prediction of fraud transaction for online banking provider will come into this model. *Clustering model* arranges data into multiple logical groups based on some common attributes. An interesting use case for this model might be grouping of students into different logical buckets based on marks, city they come from in order to decide the distribution of amount of effort for improving performance. *Forecast model* is another most popular predictive model and mainly applied to the use case where past numerical data is available to predict the value performance metrics or new value using learning from past data. As stated earlier, forecasting number of COVID-19 infected patient in next 10 days will fit into this model. When dataset contains inconsistent data records, *outlier models* are most useful as these models can identify these inconsistent entries. Finding strange records in insurance claim can be solved by this model. *Time series model* are used for short term predictions using data points collected from the past in time domain (i.e. based on time as input parameter). Collecting short term data from China epidemic and predicting the same for India can be solved using these models.

Predictive Analytics Algorithms

 doi:10.20944/preprints202004.0257.v2

Predictive analytics algorithms are either based on machine learning or deep learning. Machine learning algorithms are used when there is a need of classification or clustering for prediction, decision or analysis. These algorithms are more suitable for structured data and can be linear or nonlinear in nature. Deep learning algorithms are subset of machine learning algorithms and more useful when there is a need of identification or to recognize something. These algorithms are more useful to bigger data like audio, video and images where machine learning algorithms start underperforming. The predictive analytics is mainly driven by learning techniques and there are wide ranges of applications for disease prediction in healthcare community [28, 29]. *Random forest* algorithm is based on decision trees and used for both classification and regression purposes. This algorithm is more suitable for big data and uses bagging to avoid the errors. This model can address over fitting more effectively. *Gradient boosted model* is ensemble model of decision tress and used for classification. This model uses incremental model by building one tree at each time by correcting errors made by previously trained tree. In contrast, in random forest there is no relation amongst trees. *K-means* algorithm works on unlabeled data and places new incoming data into logical groups based on some common feature. Consider the COVID-19 example where clusters are formed of various patients based on some severity of infection. K-mean model is useful to put new incoming patient into appropriate cluster. This method is extremely useful in this growing pandemic of COVID-19 due to large number of cases. Prediction of mortality and spread rate plays very important role in pandemic disease like COVID-19, as based on this prediction precautionary measures can be taken by public, government and heath care systems. WE have used FBProphet [30] algorithm for training the model and predicting number of infected cases in next three months. We agree that there are many machine learning algorithms present in the literature. However, this study recommends Prophet Algorithm for better prediction because it is mainly open source algorithm giving more accurate prediction. As we are aware that in sudden pandemic likes COVID -19, adequate data is not available due to various reasons like duration and lack of required parameters for better prediction. Prophet algorithm enables better forecast and does not require dataset training in time series methods. The key features of this algorithm are it works more accurate for time series data and mainly used for prediction and capacity planning. Dataset can be referred from widely accepted sources like John Hopkins University and WHO. In this study, the dataset is referred from Kaggle where the statistics for this COVID-19 pandemic is given in the form of features like State/province, country, latitude, longitude, Date, Confirmed infected, Deaths and Recovered. Out of these eight fields in the dataset and another feature of Profet is it does not require splitting of dataset wherein for fitting it takes whole dataset for accurate results. Figure 1 and 2 shows the short term prediction of number of infected cases.

Figure 1 and 2 shows the prediction of spread of COVID-19. The numbers of confirmed cases of COIVD-19 within respective duration are presented in the graph, X-axis presents the duration and Y-axis shows the number of CVOID-19 confirmed cases. ML model is trained for prediction based on the worldwide dataset retrieved from Github. Predictions shown in Figure 1 and shows that the con-

 doi:10.20944/preprints202004.0257.v2

firmed COVID-19 infected cases would be 1.6 million and 2.3 million by end of May and June respectively and hence can be concluded that with increasing duration spread of COVID-19 increasing and government should initiate appropriate control measures in time to regulate this pandemic.

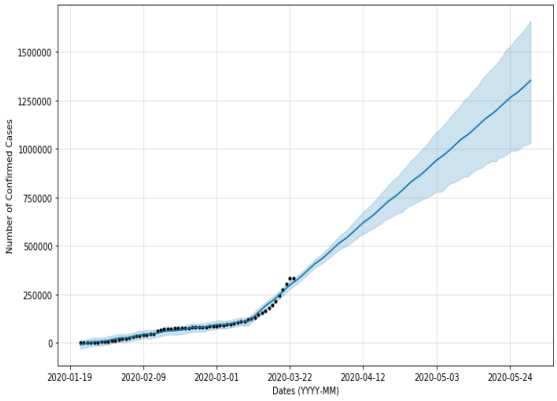

Fig. 1 Prediction of Confirmed cases till end of May

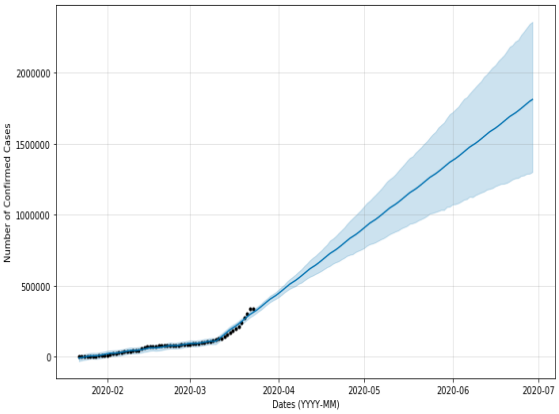

Fig. 2 Prediction of Confirmed cases till end of June

5 Conclusion

Due to pandemic of Coronavirus and COVID-19, all countries are looking towards mitigation plan to control the spread with the help some modeling techniques. This research works aims to understand the complete medical perspective of this COVID-19 pandemic and how predictive analytics will empower the pre-

dictions. Analysis of various predictive analytics methods available in the literature is presented in this chapter. We have also discussed and presented the comparative analysis of various predictive analytics models and algorithm by suggesting more appropriate use cases for application. Out study indicate that, there is a need of thorough assessment of these predictive analytics algorithm based on type of question to be answered. Application of Prophet predictive analytics algorithm on Kaggle dataset its predictions are also presented in this chapter. Simulation result of this model shows that the confirmed COVID-19 infected cases would be 1.6 million and 2.3 million by end of May and June respectively. We hope that these predictions will be also helpful to pharmaceutical companies to manufacture drugs in faster rate.

References

1. The Novel Corona virus Pneumonia Emergency Response Epidemiology Team**.** The Epidemiological Characteristics of an Outbreak of 2019 Novel Corona virus Diseases (COVID-19)—China, 2020. China CDC Weekly. 2020;2: 113-22.
2. WHO. Novel Corona virus (2019-nCoV) Situation Report - 39. 2020 [cited 2020 March]; Available from: https://www.who.int/docs/default-source/coronaviruse/situationreports/20200228-sitrep-39-covid-9.pdf.
3. WHO. Novel Corona virus (2019-nCoV) Situation Report - 52. 2020 [cited 2020 March]; Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200312-sitrep-52-covid-19.pdf.
4. The center of Evidence-Based Medicine Develops, Promotes and disseminates better evidence for healthcare CEBM-University of OXFORD report [cited 2020 April 6] https://www.cebm.net/covid-19/covid-19-what-proportion-are-asymptomatic/
5. Singhal T.Indian J Pediatr. 2020 Apr;87(4):281-286.
6. Rothe C, Schunk M, Sothmann P, Bretzel G, Froeschl G, Wallrauch C, Zimmer T, et al. Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany. N Engl J Med. 2020 Mar 5; 382(10):970-971.
7. Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet. 2020; 395(10223):514–23. 2.
8. Xiaoxia L, Liqiong Z, Hui D, Jingjing Z, Yuan L, Jingyu Q,et al. SARS-CoV-2 Infection in Children. New England Journal of Medicine. 2020. 10.1056/NEJMc2005073.
9. https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf.
10. Moriarty LF, Plucinski MM, Marston BJ, Kurbatova EV, Knust B, Murray EL, et al. Public health responses to COVID-19 Outbreaks on Cruise Ships-Worldwide, February-March 2020.MMWR Morb Mortal Wkly Rep 2020;69:347-352.
11. Russell Beth, Moss Charlotte, Rigg Anne, Hopkins Claire, Papa Sophie, Van Hemelrijck Mieke Anosmia and ageusia are emerging as symptoms in patients with COVID-19: What does the current evidence say?2020 ecancer 14 ed98.
12. Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, M. (2020). A SIDARTHE Model of COVID-19 Epidemic in Italy. arXiv preprint arXiv:2003.09861.
13. Bannister-Tyrrell, M., Meyer, A., Faverjon, C., & Cameron, A. (2020). Preliminary evidence that higher temperatures are associated with lower incidence of COVID-19, for cases reported globally up to 29th February 2020. medRxiv.

14. Russo, L., Anastassopoulou, C., Tsakris, A., Bifulco, G. N., Campana, E. F., Toraldo, G., & Siettos, C. (2020). Tracing DAY-ZERO and Forecasting the Fade out of the COVID-19 Outbreak in Lombardy, Italy: A Compartmental Modelling and Numerical Optimization Approach. medRxiv.
15. Volpert, V., Banerjee, M., & Petrovskii, S. (2020). On a quarantine model of coronavirus infection and data analysis. *Mathematical Modelling of Natural Phenomena*, *15*, 24.
16. Weber, A., Ianelli, F., & Goncalves, S. (2020). Trend analysis of the COVID-19 pandemic in China and the rest of the world. *arXiv preprint arXiv:2003.09032*.
17. Zhang, F., Zhang, J., Cao, M., & Hui, C. (2020). A simple ecological model captures the transmission pattern of the coronavirus COVID-19 outbreak in China. medRxiv.
18. Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., ... & Flasche, S. (2020). Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health.
19. Liu, Z., Magal, P., Seydi, O., & Webb, G. (2020). Predicting the cumulative number of cases for the COVID-19 epidemic in China from early data. arXiv preprint arXiv:2002.12298.
20. H Mukherjee, S Ghosh, A Dhar, Sk. Md. Obaidullah, K.C. Santosh, Kaushik Roy. Shallow Convolutional Neural Network for COVID-19 Outbreak Screening using Chest X-rays https://doi.org/10.36227/techrxiv.12156522.v1.
21. Rajinikanth, V., Dey, N., Raj, A. N. J., Hassanien, A. E., Santosh, K. C., & Raja, N. (2020). Harmony-Search and Otsu based System for Coronavirus Disease (COVID-19) Detection using Lung CT Scan Images. arXiv preprint arXiv:2004.03431.
22. D Das, K.C. Santosh, U Pal. Truncated Inception Net: COVID-19 Outbreak Screening using Chest X-rays, DOI:https://doi.org/10.21203/rs.3.rs-20795/v1.
23. Santosh, K.C. AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data. J Med Syst 44, 93 (2020). https://doi.org/10.1007/s10916-020-01562-1.
24. Fong, S. J., Li, G., Dey, N., Crespo, R. G., & Herrera-Viedma, E. (2020). Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak. arXiv preprint arXiv:2003.10776.
25. Fong, S. J., Li, G., Dey, N., Crespo, R. G., & Herrera-Viedma, E. (2020). Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. Applied Soft Computing, 106282.
26. Teles, P. (2020). Predicting the evolution Of SARS-Covid-2 in Portugal using an adapted SIR Model previously used in South Korea for the MERS outbreak. arXiv preprint arXiv:2003.10047.
27. http://predictivehealthcare.pennmedicine.org/2020/03/14/accouncing-chime.html.
28. M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017.
29. Shinde, Gitanjali R;Kalamkar, Asmita B.; Mahalle, Parikshit N; Dey, Nilanjan; Chaki, Jyotismita; Hassanien, Aboul ella; . (2020): Forecasting Models for Coronavirus (COVID-19): A Survey of the State-of-the-Art. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.12101547.v1
30. https://facebook.github.io/prophet/docs/quick_start.html.