Covid-19 Forecasting using Supervised Machine Learning Techniques – Survey

P. Lakshmi Sruthi[1] *and Dr. K.* Butchi Raju[2]

[1] M.Tech Student, C.S.E, GRIET, Hyderabad, Telangana, India
[2] Associate Professor, C.S.E, GRIET, Hyderabad, Telangana, India

ABSTRACT: COVID-19 is a global epidemic that has spread to over 170 nations. In practically all of the countries affected, the number of infected and death cases has been rising rapidly. Forecasting approaches can be implemented, resulting in the development of more effective strategies and the making of more informed judgments. These strategies examine historical data in order to make more accurate predictions about what will happen in the future. These forecasts could aid in preparing for potential risks and consequences. In order to create accurate findings, forecasting techniques are crucial. Forecasting strategies based on Big data analytics acquired from National databases (or) World Health Organization, as well as machine learning (or) data science techniques are classified in this study. This study shows the ability to predict the number of cases affected by COVID-19 as potential risk to mankind.
Keywords: pandemic, COVID-19, corona virus, exponential smoothing, R2 score adjusted, machine learning supervised

1. Introduction

Machine learning (ML) has become a popular research subject in the previous decade, handling a variety of complex and sophisticated problems. ML algorithms often learn through trial and error, in contrast to traditional algorithms, which computer instructions based on decision statements such as if-else. Forecasting is one of the most important aspects of machine learning [1]. In this field, a variety of typical machine learning methods have been applied to direct future activities as shown in below Fig 1.

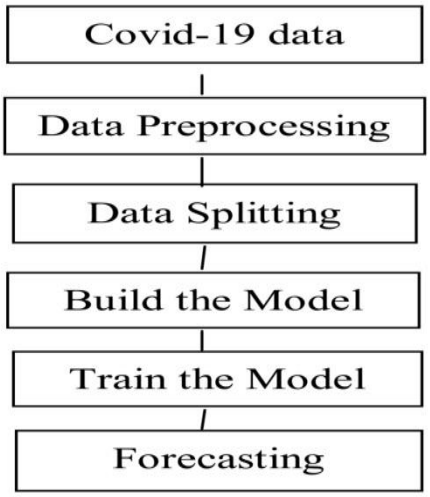

Figure 1: *Supervised Machine learning workflow*

The researchers' primary goals were to produce a study that could be beneficial for future decision-making models. Historical data is evaluated to gain perspective during the decision-making process. However, having access to data in such a short length of time is insufficient to build Artificial Intelligence (AI) models [12]. Time-series data requires AI models that can be effectively trained (During the early phases of an epidemic's spread, there is a scarcity of data). The time series analysis can help enhance forecasting efficiency.

Time series analysis is a large field that has been used to solve a wide range of issues, from econometrics to earthquakes and weather forecasting. A time series is a collection of measurements taken at regular intervals over time. A time series might be yearly, quarterly, monthly, or weekly, depending on the frequency [3]. There are two ways in which Time-series differs from a traditional regression problem. The first is time-related; in linear regression analysis, variables are independent. However, in this case, they are dependent on time. Seasonality trends, on the other hand, are fluctuations that are specific to a given span of time [4].

2. COVID (2019) OVERVIEW

COVID-19 (Corona virus) is a novel virus that causes inflammation. The disease induces a respiratory illness (such as cold, cough, fever and difficulty breathing in more severe cases).

Pandemics have posed a threat to the world on many occasions throughout history. Every pandemic's impact has always had a massive influence on the entire world, and it has also flipped the roles. Corona virus (2019), the latest destructive outbreak, is currently sweeping the globe. Not only are economics collapsing, but so are the countries' entire strengths and morale.

* Corresponding author: plsruthi@gmail.com

The global effect of the novel corona virus (COVID-19) necessarily requires detailed forecasting of confirmed patients as well as analysis of death and recovery rates. Forecasting, on the other hand, needs a large amount of past data. At the same time, no prediction can be made with certainty because the future rarely repeats itself. This study details the timetable of a live forecasting exercise with significant implications for planning and decision-making, as well as objective projections for COVID-19 cases that have been confirmed[5]. The discovery of the disease and its categorization as a pandemic by the World Health Organization are important milestones [6]

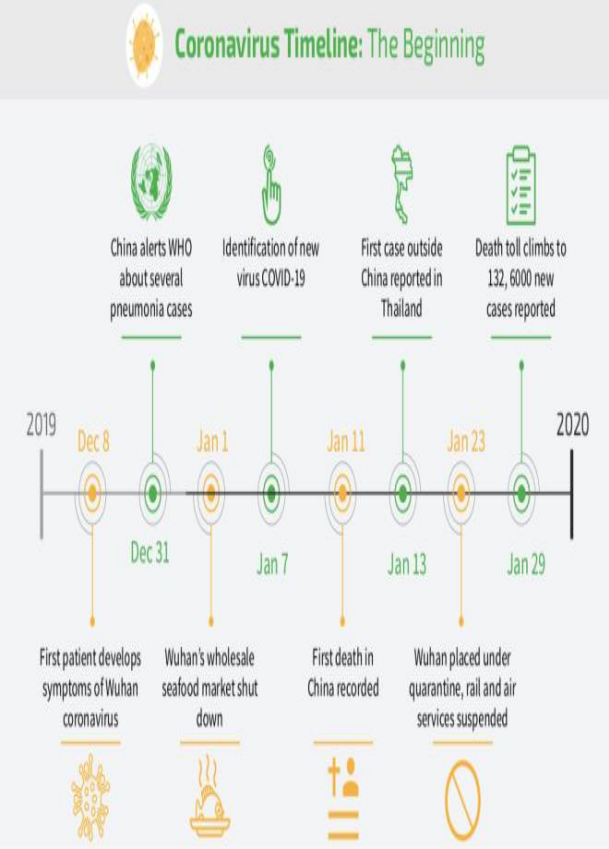

Figure 2: The Origins of the Corona virus

As shown in above Fig 2,WHO is responsible for human disease planning and response, hence diseases in the International Classification of Diseases are formally named by WHO. On February 11, 2020, ICTV declared the new virus's name as "(SARS-CoV-2) Severe Acute Respiratory Syndrome corona virus 2" and the (WHO) World Health Organization named this new disease "COVID-19". Because the virus is genetically linked to the corona virus that caused the SARS outbreak in 2003, it was given this name. The two viruses are related, but they are not the same. Corona viral infections (COVID-19) erupted in Wuhan, China, has rapidly expanded across the country [7].
COVID-19, SARC, PLAGUE, and other acquired diseases are examples. It signifies that diseases are transmitted by pathogenic agents (bacteria or virus or any micro-organism).To defend against the novel corona-virus, the **WHO [8]** recommends the following **basic precautions**.

- Keep up to current on the COVID-19 outbreak by checking out WHO updates or your local and national public health authority.
- Hand hygiene should be done on a regular basis, either with an alcohol-based hand massage.
- Keep your hands away from your eyes, nose, and mouth.
- Coughing or sneezing into a bent elbow or tissue, then discarding the tissue, is a good way to strengthen respiratory hygiene.
- If you've breathing difficulties, put on a surgical mask and wash your hands carefully after removing it.
- People who are experiencing respiratory problems have to maintain safe distance (about 2 m).
- If you've a fever, a cough, or are having trouble breathing, visit a doctor.

3. RELATED WORK

In the academic literature, machine learning (ML) methods have been offered as time-series forecasting alternative solutions to statistical approaches. However, there is a scarcity of information about their respective performance and computational needs. Using a subset of (1045) monthly data sets from the M-3 Competition, this study's purpose is to evaluate such performance over a variety of predicting horizons. When we compared the post sample accuracy of 8 prominent algorithms of ML to that of 8 classic statistical methods, study discovered that the first consistently outperformed the latter across all accuracy measures and forecasting horizons. Furthermore, we discovered that they had far higher computational requirements than statistical approaches. The study describes the findings, explains why models of ML are less accurate than statistical models, and suggests some possible next steps. Our study's empirical findings underscore the need for unbiased and fair approaches to assess the efficacy of predicting methodologies, which can also be done via major, multinational events that allow for significant comparisons and conclusions. Artificial Intelligence (AI) has gained in popularity in recent years, thanks to various elevated applications in intelligent robotics, voice recognition, image recognition, legal, medical, social applications, and even defeating winners in games such as chess and cards. The success of AI is dependent upon its usage of techniques that can learn by experimentation and improve their ability over time, rather than the typical programming domain of coding directions based on reasoning, if then principles, and Decision Trees [1].

The study's purpose is to increase machine learning (ML) algorithms' interoperability with Internet Of Things (IoT) technology in engaging with public and its surroundings in order to reduce COVID-19.Furthermore, the research looks at and examines different solution frameworks that use machine learning techniques to generate, capture, store, and analyze data. These algorithms can detect, prevent, and trace the transmission of COVID-19 in smart cities, as well as provide a better understanding of the virus. Similarly, the report highlighted case studies on the use of ML in hospitals around the world to aid in the fight against COVID-19. The research offers a thorough examination of the primary components required for integrating machine learning with other AI-based solutions. As shown in below Fig 3 and Fig 4, The study's framework provides a complete overview of the essential components required for integrating machine learning with other AI-based solutions [9] .

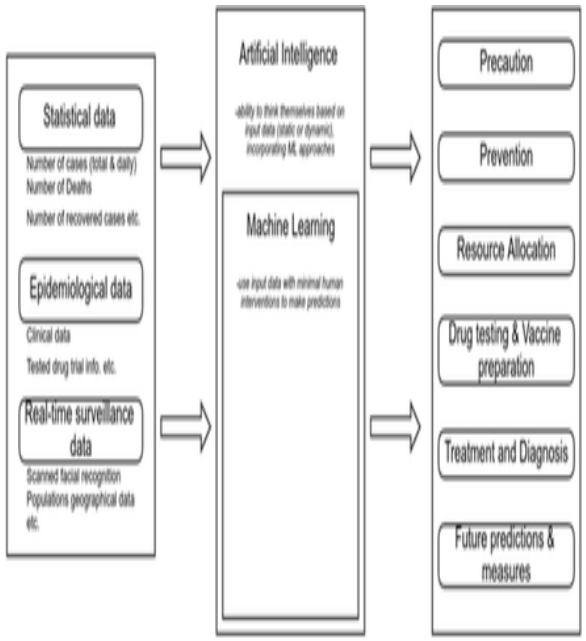

Figure 3 : Application and workflow of ML and other sub-sections of AI in tackling corona virus

The information and communication technology equipment incorporated in smart cities generates a variety of data kinds. The first type of data is **statistical data**, which often includes daily statistics such as the number of recognised cases, positive cases, deaths, and recovered cases. The second sort of data is **epidemiological data**, which mostly consists of all clinical test results for various medications, various drug trials, the patient's medical history, the patient's response to various medications, and so on. The third form of data is real-time surveillance data created by smart city sensors and cameras. Fever is one of the first symptoms of COVID-19 that can be detected. People's body temperatures and other personal information are examples of data that can help stop the spread of COVID-19 [9].

The (MLP) multi_layer perceptron is a fully-linked, (ANN) artificial neural network made up of layers of neuron like processing units feed forwarded. MLP is used for producing high quality models and also requiring less training period than more sophisticated approaches. Hyper parameters (Example: The learning rate for training a neural network) are settings that specify the ANN model's architecture. Correct hyper parameter settings are critical for producing a high-quality model. The grid search technique was used to find the optimal hyperparameter combination. A multi_layer perceptron (MLP) artificial neural network (ANN) is trained using a time series data source that is turned into a regression data source. The goal of training is to create a global model that includes the maximum patients from all locations in each time unit. With a total of 5376 hyperparameter combinations, the MLP's hyperparameters are modified using a grid-search technique. ANNs 48384 are trained using these combinations, and each model is evaluated using the determination coefficient (Zlantan Car, 2020) When cross-validation is used, the scores for confirmed, recovered, and deceased patient models drop to 0.94, 0.781, and 0.986, respectively. The deceased patient model has a high level of robustness, whereas the confirmed patient model has a decent level of robustness and the recovered patient model has a low level of robustness [10].

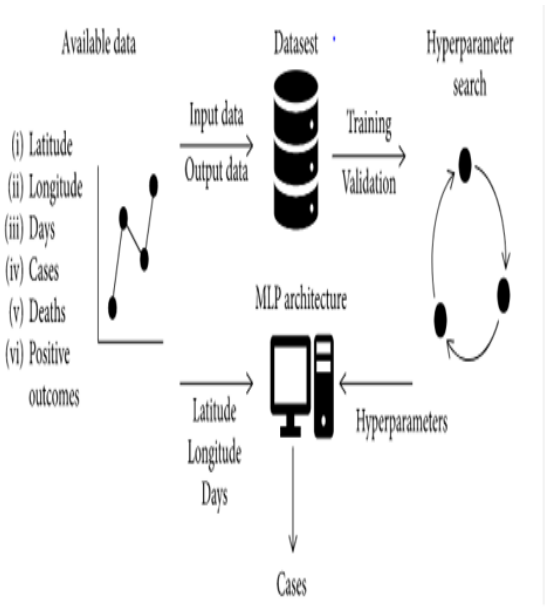

Figure 4 : *Modeling the spread of covid-19 using MLP*

The major proceedings of this paper: Comparison of the ML forecasting techniques accuracy with normal statistical ones. As highlighted in the Table 1 comparison in their study the performance of eight families of the ML model regarding their accuracy: 1) (MLP) Multi_Layer Perceptron 2) (BNN) Bayesian Neural Network 3) (RBF) Radial Basis Function 4) (GRNN) Generalized Regression Neural Networks 5) (KNN) K-Nearest Neighbor regression 6) (CART) Classification and Regression tree 7) (SVR) Support Vector Regression, and 8) (GP) Gaussian Processes. The sMAPE (Symmetric Mean Absolute Percentage Error) and ordering of these '8' methods can be represented in Table 1. From the Observation, the MLP got the highest accuracy, then after the BNN and the GP. The remaining methods' sMAPE is in the double digits, indicating a significant variation in accuracy. Investigating the grounds for the variations in performance among the different ML approaches and developing guidelines for picking the most appropriate one for new sorts of forecasting applications would be of significant research value. [11].

Table 1 Predicting Accuracy of ML methods (sMAPE)

Rank	Method	sMAPE(%)
1	MLP	8.34
2	BNN	8.58
3	GP	9.62
4	GRNN	10.33
5	KNN	10.34
6	SVR	10.40
7	CART	11.72
8	RBF	15.79

According to 'WHO' globally 634,835 confirmed cases have been reported worldwide, to date, 29,891 deaths have been confirmed. Figure 5 shows the statistics broken down by region. The following are the Regions: The Western Pacific Region, the European Region, the South_East Asian Region, the Eastern Mediterranean Region, the American Region, and the African Region are all part of the Western Pacific Region. China, France, Spain, Italy and the United States are among the heavily impacted regions.

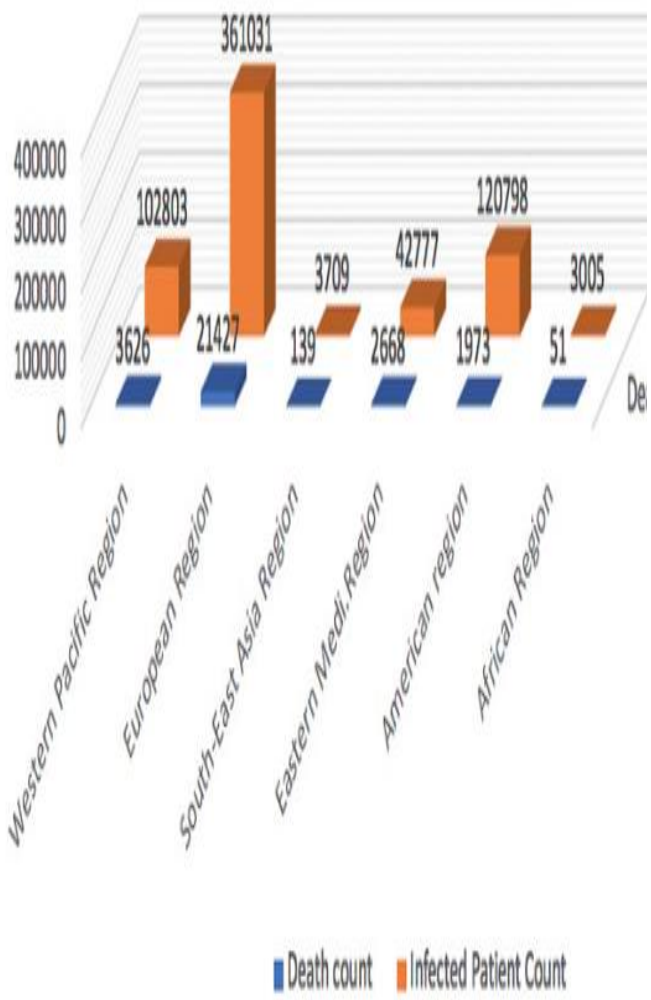

Figure 5: Regional wise Statistics

These figures are from a WHO report published March 29, 2020 [12]. This pandemic is keep expanding throughout all regions, as evidenced by the figure. Prediction can be accomplished using a variety of approaches from the fields of statistics, data science, machine learning, and artificial intelligence [12].

4. ANALYSIS

Forecasting has been done in the research using a variety of forecasting methodologies and data sources. To understand existing forecasting models for better analysis

4.1 Data Science/Machine Learning Techniques

Table 2Analysis of Covid (2019) prediction on ML Techniques or Data Science

Work ref.	Studied regions	Data source	Parameters	Remark
[13]	China	Small dataset	Corrective feedbacks of model	Forecasting suspected numbers of COVID-19
[14]	China	Chinese Center for Disease Control and Prevention	Cost of isolation, cost of treatment, no of suspects, no of confirmed COVID patients	Recommendation for decision making
[15]	China	WHO	Daily death count	Forecasting of death count
[16]	102 countries	WHO	Degree of intervention and starting intervention time	Impact of a public health intervention on the global-wide spread
[17]	China	2003 SARS Data	Death count	Forecasting of death numbers
[18]	China and European countries	WHO	Infection rate	Prediction of infection rate
[19]	Global Data	International Classification of Diseases	Preexisting medical conditions	Identify individuals who are at the greatest risk

Due to their precision, ML techniques are now utilized for forecasting all over the world. However, there are a few limitations to the use of machine learning (ML) approaches because there's very little data accessible. The optimal parameter selection and selecting models of ML are two issues involved in training a model for forecasting.

Researchers made predictions based on publicly accessible datasets and utilised the best machine learning model for each dataset [13, 14,15,16,17]. To determine rates of infection in Italy and China, [18] Research proposed a model based on the Logistic-equation, Weibull-equation, and the Hill-equation. Data analysis is conducted in this study to determine the environmental factors impact on the spread of COVID (2019). This model focused on three environmental factors: relative humidity, maximum environmental temperature, and wind speed. The results demonstrated that there is no correlation between COVID-19 spread and humidity or wind speed. The study [19] proposed a model that included a hybrid model, gradient boost trees, and logistic regression that used Medical data. The results of above models will aid in the development of management planning and the implementation of remedies in order to reduce the spread. **Table 2** summarises the results of this research.

4.2 Big Data

Table 3 Analysis of covid (2019) prediction on Big Data

Ref.	Studied regions	Data source	Parameters	Remark
[20]	China, Japan, Korea, European countries, and North America	Johns Hopkins University, GitHub repository	Transmission rate, Infection rate, and recovery rate	Recommendations for decision making
[21]	Italy, Portugal	Italy national data	Number of susceptible, exposed, asymptomatic infected, mild-to-severe infected patients	Forecasting numbers of COVID-19 patients
[22]	US	US Centers for Disease Control	Disease control interventions and traffic restrictions	Impact of disease control interventions and traffic restrictions on spread rate
[23]	Brazil	WHO	Number of susceptible, exposed, infectious and recovered patients	Suggested policy-making for avoiding outbreak in metropolitan cities

Researchers have forecasted using data from recognised national and international sources, according to the literature. Various methodologies, such as mathematical equations or machine learning algorithms, are used to analyse a large dataset.

Research [20] has given decision-making systems based on the COVID-19 data collected from Johns Hopkins University for countries such as China, European countries, Japan, Korea and North America. [21] Research used WHO COVID-19 databases, Italian national data and Johns Hopkins data to forecast death rates. The impact of disease management actions and transportation limitations on the spread rate was described [22]. The study was based on a dataset obtained from the US-CDC (Centers for Disease Control and Prevention). [23] Study has discussed the key tasks of Isolation in reducing COVID-19 dissemination rates. **Table 3** summaries the results of the literature review.

5. CONCLUSION

The spread and reproduction number should be predicted using a variety of datasets. For more accurate worldwide forecasting, the models described in the literature should be evaluated internationally. On similar considerations, several peaks must be considered in the model not just for short-term forecasting but also for forecasting the outbreak later in the year.

We expect that by analysing multiple COVID-19 forecasting models, we will be able to better modify intervention measures and, more importantly, we will be able to reduce the pandemic's worrying effect. In this study many publications analysed are preprints, which means they are not subjected to rigorous review. Though, given COVID-19's rapid global expansion, a detailed survey of comparison is urgently needed for the mankind.

References

[1] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "*Statistical and machine learning forecasting methods: Concerns and ways forward*," PloS one, vol. **13**, no. 3, (2018).

[2] L. van der Hoek, K. Pyrc, M. F. Jebbink, W. Vermeulen-Oost, R. J. Berkhout, K. C. Wolthers, P. M. Wertheim-van Dillen, J. Kaandorp, J. Spaargaren, and B. Berkhout, "*Identification of a new human coronavirus*," Nature medicine, vol. **10**, no. 4, pp. 368–373, (2004).

[3] Dimitris Effrosynidis"*Time Series Analysis with Theory, Plots, and Code Part 1*" Apr 5, (2020) [online] https://towardsdatascience.com/time-series-analysis-with-theory-plots-and-code-part-1-dd3ea417d8c4

[4]Bhanuka Dissanayake, "An introduction to time series, and basic concepts and modelling techniques related to time series analysis and forecasting", Jul 14, 2020[online]https://towardsdatascience.com/introduction-to-time-series-forecasting-7e03c4bd83e0

[5]F. Petropoulos and S. Makridakis, "*Forecasting the novel coronavirus covid-19*," Plos one, vol. **15**, no. 3, p. e0231236, (2020).

[6]Harry Kretchmer,"Key milestones in the spread of the coronavirus pandemic", 22 Apr 2020 [online] https://www.weforum.org/agenda/2020/04/coronavirus-spread-covid19-pandemic-timeline-milestones/

[7] "*WHO. Naming the coronavirus disease (covid-19) and the virus that causes it*". [Online]. Available: https://www.who.int/emergencies/diseases/novelcoronavirus-2019/technical-guidance/naming-the-coronavirus-disease- (covid-2019)-and-the-virus-that-causes-it

[8] "*World Health Organization. Coronavirus disease (COVID-19) advice for the public*".[Online]. Available:https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public

[9] Ezugwu, Absalom & Abaker, Ibrahim & Oyelade, Olaide & Chiroma, Haruna & Al-Garadi, Mohammed & Abdullahi, Idris & Otegbeye, Olumuyiwa & Shukla, Amit & Almutari, Mubarak, "*A Novel Smart City Based Framework on Perspectives for application of Machine Learning in combatting COVID-19*", (2020): doi:

10.1101/2020.05.18.20105577.
[10]Zlatan Car, Sandi Baressi Šegota, Nikola Anđelić, Ivan Lorencin, Vedran Mrzljak, "*Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron*", Computational and Mathematical Methods in Medicine, vol. 2020, Article ID 5714714, 10 pages, (2020). https://doi.org/10.1155/2020/5714714
[11]Ahmed NK, Atiya AF, Gayar NE, El-Shishiny H. "*An Empirical Comparison of Machine Learning Models for Time Series Forecasting. Econometric Reviews*", (2010); 29(5–6):594–621. https://doi.org/10.1080/07474938.2010.481556
[12]Shinde, Gitanjali R et al. "*Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art.*" SN computer science vol**. 1**, 4 (2020): 197. doi:10.1007/s42979-020-00209-9
[13]Fong SJ, Li G, Dey N, Crespo RG Herrera-Viedma E. "*Finding an accurate early forecasting model from small dataset: a case of 2019-ncov novel coronavirus outbreak*". arXiv preprint arXiv:2003.10776. (2020)
[14]Fong SJ, Li G, Dey N, Crespo RG, Herrera-Viedma E. "*Composite monte carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction*". Appl Soft Comput. (2020);106282.
[15]Batista M. "*Estimation of the final size of the second phase of the coronavirus COVID-19 epidemic by the logistic model*". 2020.03.11.20024901; doi: https://doi.org/10.1101/2020.03.11.20024901
[16]Hu Z, Ge Q, Li S, Jin L, Xiong M. "*Evaluating the effect of public health intervention on the global-wide spread trajectory of Covid-19*". medRxiv. (2020).
[17]Jia L, Li K, Jiang Y, Guo X. "*Prediction and analysis of coronavirus disease 2019*". arXiv preprint https://arXiv:2003.05447. (2020).
[18]Kumar J, Hembram KPSS. "*Epidemiological study of novel coronavirus (COVID-19)*". arXiv preprint https://arXiv:2003.11376. (2020).
[19]DeCaprio D, Gartner J, Burgess T, Kothari S, Sayed S. "*Building a COVID-19 vulnerability index*". arXiv preprint https://arXiv:2003.07347. (2020).
[20]Toda AA. "Susceptible-*infected-recovered (sir) dynamics of covid-19 and economic impact*". arXiv preprint arXiv:2003.11221. (2020).
[21]Teles P. "*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 https://arXiv:2003.10047. (2020).
[22]Liu P, Beeler P, Chakrabarty RK. "*COVID-19 progression timeline and effectiveness of response-to-spread interventions across the United States*". medRxiv. (2020).
[23]Rocha Filho TM, dos Santos FSG, Gomes VB, Rocha TA, Croda JH, Ramalho WM, Araujo WN "*Expected impact of COVID-19 outbreak in a major metropolitan area in Brazil*". medRxiv.(2020).