Archives of Computational Methods in Engineering (2021) 28:2645–2653
https://doi.org/10.1007/s11831-020-09472-8

ORIGINAL PAPER

The Number of Confirmed Cases of Covid-19 by using Machine Learning: Methods and Challenges

Amir Ahmad[1] · Sunita Garhwal[2] · Santosh Kumar Ray[3] · Gagan Kumar[4] · Sharaf Jameel Malebary[5] · Omar Mohammed Barukab[5]

Received: 2 July 2020 / Accepted: 23 July 2020 / Published online: 4 August 2020

Abstract
Covid-19 is one of the biggest health challenges that the world has ever faced. Public health policy makers need the reliable prediction of the confirmed cases in future to plan medical facilities. Machine learning methods learn from the historical data and make predictions about the events. Machine learning methods have been used to predict the number of confirmed cases of Covid-19. In this paper, we present a detailed review of these research papers. We present a taxonomy that groups them in four categories. We further present the challenges in this field. We provide suggestions to the machine learning practitioners to improve the performance of machine learning methods for the prediction of confirmed cases of Covid-19.

1 Introduction

Coronavirus disease 2019 (Covid-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. It was first reported in Wuhan, China in December 2020 [2]. Since then it has been spreading all over the world, it spreads differently in countries (Figs. 1 and 2). It was declared pandemic by World Health Organization (WHO) on 11th March 2020 [3]. As of 4th June 2020, there were more than 6.5 million Covid-19 confirmed cases across 188 countries [4].

The fast spread of the Covid-19 has put a lot of burden on healthcare systems of countries. Knowing the number of confirmed cases in future has become an important task for the public health policy makers so that they can increase medical facilities accordingly. Different governments propose various public health interventions such as lockdown, social distancing, closing of schools etc. to slowdown the spread of Covid-19. The effect of various health policies should be estimated with accurate prediction models so that the health policies can be modified to be more effective.

Different approaches have been proposed to model the spread of infectious diseases. Susceptible-Infectious-Recovered model and its extensions such as Susceptible-Infectious-Recovered-Deceased-model, Susceptible-Exposed-Infectious-Recovered-model etc. are established epidemiological models to predict the spread of infectious diseases [6]. The models are generally based on ordinary differential equations. These models can also be used to study the effect of public health interventions on the spread of infectious diseases. Recently, machine learning methods have been applied to model the spread of infectious diseases [7] .

In machine learning, models are built using historical data and these models are used to predict the new outcome. Regression, classification, clustering, deep learning etc. [8, 9] are some of the machine learning methods which have been successfully used in various domains such as image analysis, speech recognition, health informatics etc.

Many applications of machine learning methods for Covid-19 have been proposed such as diagnosis and prognosis, patient outcome prediction, tracking and predicting the outbreak, drug development, vaccine discovery, false news prediction, etc. [10–14]. Many machine learning models

✉ Amir Ahmad
amirahmad@uaeu.ac.ae

1 College of Information Technology, United Arab Emirates University, Al Ain, UAE

2 Department of Computer Science and Engineering, Thapar University, Patiala, India

3 Department of Information Technology, Khawarizmi International College, Al Ain, UAE

4 Department of Physics, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India

5 Faculty of Computing and Information Technology, King Abdulaziz University, P.O. Box 411, Rabigh, Jeddah 21911, Saudi Arabia

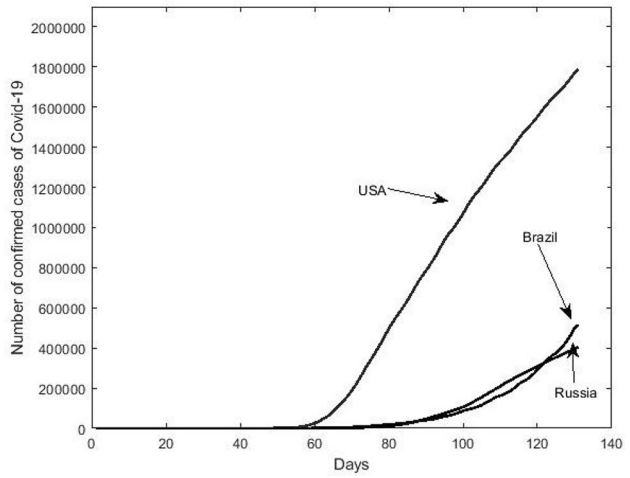

Fig. 1 Confirmed cases of three countries (USA, Brazil and Russia) where the number of confirmed cases is increasing steadily. From 22nd January 2020 to 31st May 2020 [5]

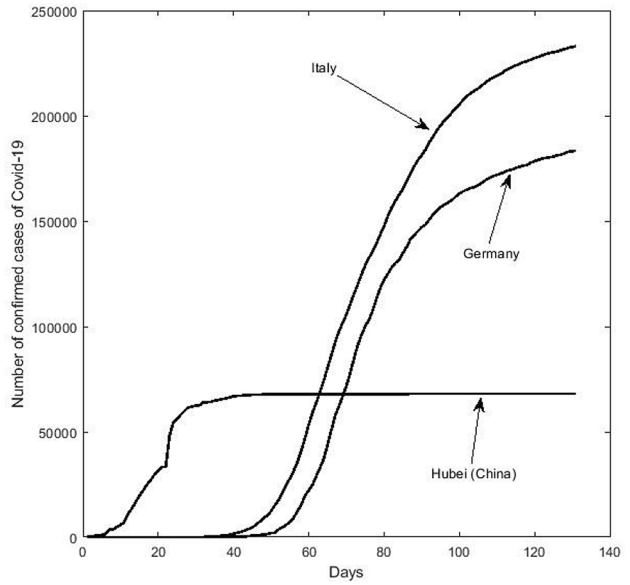

Fig. 2 Confirmed cases of two countries (Italy and Germany) and Hubei (China) where the number of confirmed cases has almost reached to its peak. From 22nd January 2020 to 31st May 2020 [5]

have been used to model the number of confirmed cases of Covid-19 [11–13]. In this paper, we will review these machine learning models. There are other review papers on the related topics, we present the detailed analysis of these review papers and differentiate them from our paper. In this paper, we present a taxonomy to identify four broad machine learning methodologies for predicting confirmed cases of Covid-19. Using this taxonomy, we present the comprehensive review of related published papers. We also present the challenges that have impacted this area. We further present suggestions to improve the performance of the machine learning models for the prediction of confirmed cases of Covid-19.

The paper is organized in following way. In the next section, we will discuss published reviews on the applications of machine learning methods for Covid-19. Section 3 will discuss various machine learning models to predict the number of confirmed cases of Covid-19. Challenges are presented in Section 4. The paper ends with suggestions to improve the accuracy of machine learning methods for the prediction of Covid-19 confirmed cases.

2 Related Review Papers

A few review papers have been published that discuss applications of machine learning methods or artificial intelligence (AI) techniques for Covid-19. In this section, we will discuss these review papers. Naude [12] presents an early review of the applications of AI techniques for Covid-19. Prediction of the confirmed cases is one of the applications discussed in the paper. However, few related papers are discussed. Bullock et al. [13] discuss some of the studies that apply machine learning methods for handling Covid-19. They discuss some of the papers that apply machine learning methods to forecast the spread of Covid-19. Vaishya et al. [14] discuss some of the applications of AI techniques for Covid-19 pandemic. They suggest that AI techniques can be used to predict the number of cases, however, no related papers are reviewed. Pham et al. [11] review the applications of AI and big data for Covid-19 pandemic. In that paper, they review eight papers related to prediction of spread of Covid-19. The study suggests that all the related review papers focus on many applications of machine learning methods for Covid-19. Prediction of the confirmed cases is not extensively covered in these review papers. None of these papers present any taxonomy to review the papers related to the prediction of the confirmed cases. Our paper concentrates on only one application of machine learning methods for Covid-19. We try to include all the related papers which are not covered in aforementioned review papers.

3 Taxonomy of Machine Learning Methodologies for the Prediction of Confirmed Cases of Covid-19

The world is facing Covid-19 pandemic. The health policy makers need reliable estimates of the confirmed cases in future to make informed decisions about the required health facilities. They also want to estimate the effect of public health interventions to mitigate the epidemic. Machine

learning methods have been applied to forecast the confirmed cases.

In this section, we will present a taxonomy to review the related published papers. The taxonomy identifies four research themes- *traditional machine learning regression, deep learning regression, network analysis, and social media and search queries data-based methods.* There are some papers which may belong to more than one research themes, we take utmost care to place them in the most related and relevant research theme.

3.1 Traditional Machine Learning Regression

Regression analysis is a supervised machine learning approach which estimates the relationship between a dependent variable and independent variables. The independent variables are also called predictor variables [8]. These relationships are learned from the given data and are used for prediction. Some of the regression analysis methods assume models and try to find the best parameters that fit the data to those models; for example multivariate linear regression (1) it is assumed that the dependent variable y is linearly dependent on j predictor variables x_i ($i = 1$ to j). The task is to find out the values of a_i and a_0 so that the equation best fit the given data.

$$y = \sum_{i=1}^{j} a_i x_i + a_0 \tag{1}$$

where a_i ($i = 1$ to j) and a_0 are constant.

These relationships can be nonlinear. For example polynomial regression of order more than one, S-shaped logistic curve (2);

$$N_t = \frac{N_{max}}{1 + e^{-c(t-t_o)}} \tag{2}$$

where t is a predictor variable, N_{max} is the maximum value of the curve, t_0 is the t value of the sigmoid's midpoint, and c is the logistic growth rate. With the given data, the best values of N_{max}, t_0 and c are estimated. Then the equation is used to predict the number of curve, N_t, at given t.

These methods work well if the assumed models are correct otherwise the performance may not be very accurate. Some of the machine learning methods do not assume any relationships and can learn complex relationships between predictor variables and dependent variable. Random forests [15] and neural networks [8] come under this category.

In the prediction of confirmed cases of Covid-19 using regression methods two approaches have been used.

1. Time series analysis - In this approach the confirmed cases against the days curve is used for the prediction in the future. Two methodologies have been used for this purpose.
   - m previous days confirmed cases are used to predict the next day confirmed cases. A relationship is learned using multivariate regression analysis which is used to predict the confirmed cases in the future [16].
   - A relationship such as logistic curve is learned to predict the number of confirmed cases at a given day [17, 18].
2. The relationships between confirmed cases of Covid-19 and the other factors such as temperature, humidity etc. are leaned from the data and the relationships are used to predict the number of confirmed cases with the new values of the factors [19].

We will present the detailed review of traditional machine learning regression methods for the confirmed cases of Covid-19. Gupta et al. [16] use polynomial regression to predict the number of confirmed cases in India. They use data from 30th January till 25th March 2020 as the training data and predict the number of Covid-19 patients in India for next two weeks. Gu et al. [20] apply cubic regression equations, which use the number of days as the input variable to predict the conformed Covid 19 cases in China and world.

Pavlyshenko [17] apply logistic curve to model Covid-19 spread. Model parameters are computed using Bayesian regression approach. The author argues that in the Bayesian inference, prior distributions can be set up by a Covid-19 expert which can be useful for small historical Covid-19 datasets. Predictions are made for confirmed Covid-19 cases in different countries. Batista [21] apply logistic growth regression model to predict the the final size of the Covid-19 epidemic. The final value is predicted around 90000, which is way below the current value. Batista [22] does the similar calculation for final size of the second phase of the Covid-19 using logistic model, the parameters of the logistic model are computed using regression analysis. Ensembles are combinations of accurate and diverse models, they generally perform better than single model [23]. Buizza [24] create an ensembles of logistic curves to estimate the confirmed cases. These diverse logistic curves are generated by perturbing the training dataset. The model works well for China.

Petropoulos and Makridakis [25] predict the global spread of Covid-19 using models from the exponential smoothing family. These models have shown great prediction accuracy for short time series and are useful to capture various types of trends. This approach is opposite to S-Curve (logistic curve) approach that assumes convergence. Stubinger and Schneider [26] argue that the Covid-19 spread in China first and then other countries got

affected by it. Therefore, there are lead-lag effects between spreads of Covid-19 in different countries. The relationships can be exploited to predict the spread of Covid-19 in other countries using the data of China. The relationships are computed using dynamic time warping [27].

Tobias et al. [28] predict the number of the confirmed cases in Italy and Spain under lockdown using quasi-Poisson regression model. Interaction model is applied to compute the variations in confirmed cases trends. Xu et al. [29] apply a rolling growth curve approach (RGCA) to estimate the spread in USA. The model uses the number of daily hospitalized Covid-19 patients as the independent variable. The authors argue that the number of confirmed deaths due to Covid-19 or the number of daily hospitalized COVID-19 patients is more reliable than the reported number of COVID-19 cases. Li et al. [30] propose a regression equation to predict the early spread of Covid-19 in China. In this equation, *log* value of the summation of confirmed cases and 34 is linearly dependent on the days. An adaptive neuro-fuzzy inference system (ANFIS) is applied to forecast the number of conformed cases in China in the time-series framework [31]. Enhanced flower pollination algorithm (FPA) and salp swarm algorithm (SSA) are used for selecting parameters of ANFIS. The results suggest that the FPA and SSA perform better than the other parameter selection methods.

Maier and Brockmann [18] demonstrate that the scaling law, number of confirmed cases is proportional to $(days)^x$, is universal for confirmed cases in affected provinces in Mainland China universal, with a range of exponents $(x) = 2.1 \pm 0.3$. They suggests that effective containment strategies can be the reason for this behaviour. It is shown that all country-specific infection rates follow a power law growth behaviour [32]. Different countries have different scaling exponents. The authors calculate scaling exponents for different countries. It is shown [33] that cumulative distribution function [34] can be used to predict the spread of the Covid-19 using days as the predictor variable.

Gupta and Pal [35] investigate the application of ARIMA (Auto-Regressive Integrated Moving Average) time series method for the prediction of confirmed covid 19 cases in India. ARIMA model is used to estimate the spread in China, Italy, South Korea, Iran and Thailand [36]. The model predicts a stable trend in China and Thailand in future, whereas the model predicts that Iran and Italy will have unstable trends in future. Using time-series framework, Chakraborty and Ghosh [37] combine ARIMA model and Wavelet-based forecasting model to generate ten days ahead estimates of Covid-19 spread in various countries.

Perc et al. [38] develop an iteration method to estimates the transmission of Covid-19 that requires the daily values of confirmed cases as input. It accounts for expected recoveries and deaths using a parameter. Covid-19 spread is predicted for various countries using the proposed method.

Lua et al. [39] apply linear regression to suggest that countries with lower Healthcare Access and Quality (HAQ) Index may have larger number of unreported cases of Covid-19. Pirouz et al. [19] apply a group method of data handling (GMDH) type of neural network [40] to investigate the relationship between environmental and urban factors and the number of confirmed cases. It is also shown using 42 datasets in four countries, including China, Japan, South Korea, and Italy to show that there is a very low correlation between them, therefore different models should be created for these datasets. Zhao et al. [41] propose that the daily traffic from Wuhan and the total traffic in this period can be used to estimate the spread in Chinese cities in January 2020. Multiple regression models are developed which use number of passengers and local population as predictor variables to explain the variance of the number of cases in the infected cities. Ensembles of ten different machine learning algorithms are used to estimate the spread of Covid-19 using climate variables (monthly mean temperature, interaction term between monthly minimum temperature and maximum temperature, monthly precipitation sum, downward surface short-wave radiation, and actual evapotranspiration) [42].

Wang et al. [43] apply linear regression framework to suggest that high temperature and high humidity significantly reduce the transmission of Covid-19. They use data from China in their experiments. Oliveiros et al. [44] apply a linear regression model to study the effect of temperature, humidity, precipitation, and wind speed on the doubling rate of Covid-19 spread in China.

3.2 Deep Learning Regression

Deep learning is generally related with artificial neural networks which mimic human brain. Deep neural networks have large number of hidden layers [9]. Deep neural networks have shown excellent performance in various domains image analysis, speech recognition, text analysis etc. Various types of deep learning neural networks have been applied to predict the spread of Covid-19.

Long short-term memory (LSTM) is an artificial recurrent neural network architecture which is used for time series forecasting [9]. Tomar and Gupta [45] investigate the use of LSTM for the prediction of the number of confirmed cases in India. The predicted cases are very close to the official number of cases. Hu et al. [46] use modified auto-encoders [47] to model Covid-19 time series of confirmed cases in various Chinese cities. They use the trained model to predict six-step to ten-step forecasting. Yang et al. [48] use LSTM to predict the spread in China in Feb. 2020. The LSTM is trained using the SERS 2003 data. COVID-19 epidemiological parameters such as transmission probability, incubation

rate, etc. are incorporated in the model. The model predicts that the number of confirmed cases in China should peak by late February. Fong et al. [49] demonstrate that for small datasets polynomial neural network with corrective feedback method outperforms linear regression, support vector machines [8] and ARIMA methods. Deep learning-based Composite Monte-Carlo simulation [50] is used in conjunction of fuzzy rule induction techniques to predict the spread in China. Deep learning produces better fitted Monte Carlo outputs which lead to a better prediction. It is demonstrated that the combination of LSTM and gated recurrent unit perform better than individual methods to predict the confirmed cases in the world [51]. Nonlinear autoregressive artificial neural networks are used to predict the spread in many countries [52]. Huang et al. [53] demonstrate that convolutional neural networks [9] outperform other deep learning models such as gated recurrent unit, LSTM and multilayer perceptron in the prediction of confirmed cases of Covid-19 in Chinese cities.

3.3 Network Analysis

Networks or graphs consist of nodes and edges [54]. Nodes represent the entities. Edges represent the connections between nodes. Analysis of graphs or networks is an important research area of machine learning [54]. Web mining, social networks analysis, community detection etc. are some of the applications of the networks analysis [54]. Humans are connected with other humans, therefore they can be represented by networks. These networks have been used to study the spread of the infectious diseases [54]. When one of nodes of a network gets infected, it can infect other nodes which are connected to the infected node. This process continues and the disease spread to the other nodes of the network. In this section, we will discuss those papers that use network analysis to predict the number of confirmed cases of Covid-19.

It has been observed that virological transmission usually satisfies the Gaussian distribution. Therefore, this is used to predict the Covid-19 transmission [55], it is assumed that an infected person can infect 1 to ∞ persons. The model is used to estimate the transmission in China and the other countries. Zhuang et al. [56] use a stochastic model to estimate the confirmed cases in Republic of Korea and Italy. In this model, it is assumed that the number of secondary cases associated with a primary Covid-19 case follows a negative binomial distribution. The mean parameter represents the basic reproduction number of Covid-19, whereas the dispersion parameter represents the likelihood of occurrence of other factors that can effect spread like super-spreading events. The current growth closely follows power-law kinetics in China, indicative of an underlying fractal or small-world network (a small average shortest path length, and a large clustering coefficient [54]) of connections between susceptible and infected individuals [57]. Li et al [58] demonstrate that the spread of Covid-19 closely follows a power-law kinetics in China during January 2020 to February 2020. It suggests that the underlying network has small-world property.

Herrmann and Schwartz [59] propose that the network of interactions can be used to estimate the spread of Covid-19. They use scale-free networks (a scale-free network's degree distribution follow a power law) with Susceptible-Infected-Susceptible model [60], with the model parameters computed for Covid-19. The results indicate that directly targeting hubs in the network is far more effective than randomly decreasing the number of connections between individuals.

It is shown that by transforming the time-series of Covid-19 infection curve to a visibility graph one can study the time-series as a complex network [61]. Complex-network-based splines regression method is proposed for the prediction of confirmed cases in Greece [62]. The proposed method outperforms both the cubic regression model and the randomly-calibrated splines regression model.

Pujari and Shekatkar [63] propose a hybrid model to predict the spread of Covid-19 in Indian cities. Susceptible-Infectious-Recovered models are created for individual cities. The migration among cities are modelled using transportation networks. Indian aviation and railway networks are used as transportation networks. Biswas and Sen [64] use Susceptible-Infected-Removed model of epidemic spreading on Euclidean networks to study space-time dependence of Covid-19 spread in many countries. It is assumed that the disease can be transmitted to a nearest neighbour and to some random other agent who is connected with a probability decaying algebraically with the Euclidean distance separating them.

Gross et al. [65] study the spatial dynamics of the Covid-19 in Hubei and other provinces of China. It is demonstrated that power laws hold for the number of confirmed cases in each province as a function of the province population and the distance from Hubei.

Hybrid non-linear cellular automata (HNLCA) classifier is trained on different parameters such as movement of the people, various transmission rates, vulnerable people in the region, etc. to estimate the spread in India [66]. Regression analysis is carried out to evaluate the relationship between migration and the spread of Covid-19 in China [67]. Bivariate correlation analysis is used to extract the strength of these relationships in various cities of China.

3.4 Social Media and Search Queries Data-Based Methods

Internet search queries and social media have emerged as rich sources of data. This data has been used to predict and monitor infectious diseases [68]. It is easy to collect this

data, however there is a problem of noisy data. Internet search queries and social media have been used to predict the number of confirmed cases of Covid-19, we will review these research works in this section.

Qin et al. [69] predict the number of Covid-19 cases by using social media search indexes (SMSI) of Covid-19 symptoms (dry cough, fever, chest distress, coronavirus, and pneumonia) collected from Baidu search engine data as the independent variables. Five regression methods (subset selection, forward selection, lasso regression, ridge regression, and elastic net) are used to determine the relationships between the number of Covid-19 cases and independent variables. The subset selection method produces the best result. Jahanbin and Rahmanian [70] conclude that tweets extracted from twitter can be used to estimate the spread of Covid-19. Fuzzy rule-based evolutionary algorithm called Eclass1-MIMO is used to model the spread. The results suggest that geographical origins of tweets posted about Covid-19 are consistent with the number of confirmed cases of Covid-19.

ARGOnet [68] combines AutoRegression with General Online information (ARGO) with spatio-temporal information about influenza activity to predict influenza transmission. Liu et al. [71] apply AGROnet to predict Covid-19 cases, the model uses the data from official health reports from *Chinese Center Disease for Control and Prevention*, Covid-19 related internet search activity from Baidu, media activity reported by Media Cloud, and daily forecasts of COVID-19 activity by an agent-based mechanistic model. They apply clustering to get spatio-temporal Covid-19 information across Chinese provinces required for AGROnet. ARGOnet outperforms an autoregressive model trained only on historical confirmed cases.

Ayyoubzadeh et al. [72] use search statistics collected from Google Trends for search queries related to Covid-19 such as Corona, Covid-19, Coronavirus, hand washing, Antiseptic, etc. to predict the number of confirmed cases in the next days in Iran. Linear regression and LSTM models are used for the prediction. Linear regression model performs better.

Lampos et al. [73] use the data pertaining to google search queries and basic news media coverage metric associated with Covid-19. Elastic net models are trained using the data. The authors also study the transferability of their models between countries. It is proposed that Twitter sentiment analysis can be used to predict the spread of Covid-19 outbreak [74].

4 Challenges

Many machine learning methods have been used to predict the number of confirmed cases of Covid-19. However, there are many challenges for the accurate prediction by machine learning methods. In this section, we will discuss these challenges.

Time series framework is very popular in prediction of the confirmed cases. In many countries the first case of Covid-19 came in January 2020 or after. Therefore till 30th May 2020, the number of data points is less 150. It is difficult to train accurate machine learning models with such small datasets. Deep learning methods are successful because of large training data which is not available for Covid-19 confirmed cases prediction task. It is difficult to select appropriate architectures and parameters for deep learning neural networks with small datasets.

The lack of historical data is a major problem. Pandemics are rare and the characteristics of Covid-19 are different than those of other coronaviruses such as SARS and MERS [75]. Therefore, their data cannot be used for Covid-19.

It is argued that many countries are not doing enough testing [76]. Therefore, it is impossible to have correct number of confirmed cases in these countries. Training machine learning algorithms with datasets of poor quality will lead to misleading conclusions.

Governments take different preventive steps such as lockdown, social distancing etc. to slowdown the spread of Covid-19. Inclusion of the effects of these measures in machine learning models is a challenging task.

It has been estimated that about 80% of people with COVID-19 are mild or asymptomatic cases [77]. Many of these people do not get tested, therefore the numbers of confirmed cases in countries are not accurate. However, these people contribute to the new cases. It makes creating accurate machine learning models for Covid-19 confirmed cases a difficult task.

Social stigma attached to Covid-19 [78] in many countries force suspected Covid-19 patients to stay away from medical facilities. Therefore, many confirmed cases are not recorded. However, they contribute to new confirmed cases.

Logistic curves have been applied to predict the number of the confirmed cases. It is shown that by using logistic curves, a real-world epidemic outbreak can be predicted reliably only in the short term [79]. Therefore, logistic curves may not produce accurate prediction for long term.

Social media data has been applied for the prediction of Covid-19 cases. This data is huge, however, the data is noisy. Search queries have been used to predict the number of confirmed cases. However, many normal people also try to find the information about Covid-19. Therefore, it is difficult to predict the confirmed cases only on the basis of search queries. Furthermore, there is a vast difference in Internet penetration is various countries. The number of Internet users are vastly different in countries [80]. The amount of social media data and search queries data is quite different in countries.Therefore, it is difficult to find prediction models which may work for many countries.

Predicting the number of confirmed cases of Covid-19 using network analysis requires accurate estimation of the transmission of Covid-19 from one person to another person. As Covid-19 is a new pandemic, all the characteristics of this pandemic are not known. Hence, the predicted confirmed cases may not be accurate.

Structures of networks play an important role in prediction accuracy. In a country the connections between nodes are dependent on the culture, environment, population density etc. Some of these factors such as culture are difficult to quantify. Therefore, it is difficult to estimate the connections between nodes accurately. Furthermore, models in one country cannot be applied to other countries easily as they have different factors.

5 Suggestions

We present some suggestions that can be used to address the challenges for machine learning methods for the prediction of confirmed cases of Covid-19.

Machine learning methods have not been very successful for the prediction of confirmed cases of Covid-19 because of the challenges discussed in the last section. Established epidemiological models, such as Susceptible-Infectious-Recovered models, have been successfully used for modelling infectious diseases. The hybrid models of machine learning algorithms and epidemiological models will be a promising research area.

Machine learning experts should work with epidemiologists to understand the data well and to select the parameters of machine learning methods for the prediction of confirmed cases of Covid-19.

As it is difficult to predict the number of confirmed cases of Covid-19 accurately from one type of data. Models trained on different types of data such as times series data, social media data etc. may be combined to predict the number of confirmed cases.

Countries are at different stages of the pandemic. Therefore, data from one country that are ahead in the epidemic curve may be used for other countries in earlier stages of the epidemic curve. There have been some attempts in this research direction [26]. Transfer learning deals with training a model for a given problem and using it to related but different problem [81]. Transfer learning has been applied for the prediction of confirmed cases by using the model trained on data from one country in which is ahead in the epidemic curve to other countries still in earlier stages of the epidemic curve [73]. More research is required in this direction.

Some countries have better facilities of collecting data. This data can be used for countries with less facilities of collecting data. The data should be modified to the local context for the better representation of a country.

In countries where social stigma attached with Covid-19 scare people from taking medical help, new sources for data such as medicine buying pattern, online-purchasing pattern, shopping pattern etc. should be investigated for the prediction of Covid-19 cases.

In this paper, we identified four major research themes to predict the number of confirmed cases of Covid-19. We presented a comprehensive state-of-the-art review of the related research papers within them. We discussed the challenges in this research area and presented suggestions to address them. We believe that the review paper will be helpful to the researchers to develop an in-depth understanding of the research area. This paper will help to generate novel ideas of using machine learning methods for accurate prediction of the number of confirmed cases of Covid-19. This will be very beneficial to mankind.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Naming the coronavirus disease (covid-19) and the virus that causes it. World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. (Accessed on 26 May 2020)
2. Novel coronavirus in China. World Health Organization. https://www.who.int/csr/don/12-january-2020-novel-coronavirus-China/en/. (Accessed on 26 May 2020)
3. WHO Director-General's opening remarks at the media briefing on Covid-19. World Health Organization (press release). 11 march 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. (Accessed on 26 May 2020)
4. Covid-19 dashboard by the center for systems science and engineering (csse) at Johns Hopkins University, USA. https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6. (Accessed on 26th May 2020)
5. Novel coronavirus (covid-19) cases, provided by Johns Hopkins University, USA. https://github.com/CSSEGISandData/COVID-19. (Accessed on 31 May 2020)
6. Hethcote HW (2000) The mathematics of infectious diseases. SIAM Rev 42(4):599–653
7. Hao K (2020) This is how the cdc is trying to forecast coronavirus spread. https://www.technologyreview.com/2020/03/13/905313/cdc-cmu-forecasts-coronavirus-spread/. (Accessed on 25 May 2020)
8. Bishop CM (2008) Pattern recognition and machine learning. Springer, Berlin
9. Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press. http://www.deeplearningbook.org
10. McCall B (2020) Covid-19 and artificial intelligence: protecting health-care workers and curbing the spread. Lancet Dig Health 2(4):e166–e167

11. Pham Q, Nguyen D C, Huynh-The T, Hwang W, Pathirana P (xxxx) Artificial intelligence (ai) and big data for coronavirus (covid-19) pandemic: a survey on the state-of-the-arts. https://www.preprints.org/manuscript/202004.0383/v1
12. Artificial intelligence against covid-19: an early review. https://www.iza.org/publications/dp/13110/artificial-intelligence-against-covid-19-an-early-reviewBatis
13. Bullock J, Luccioni A, Pham KH, Lam CSN, Luengo-Oroz M (2020) Mapping the landscape of artificial intelligence applications against covid-19
14. Vaishya R, Javaid M, Khan IH, Haleem A (2020) Artificial intelligence (ai) applications for covid-19 pandemic. Diabetes Metabolic Syndrome Clin Res Rev 14(4):337–339
15. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
16. Gupta R, Pandey G, Chaudhary P, Pal SK (2020) Seir and regression model based covid-19 outbreak predictions in India. medRxiv
17. Pavlyshenko BM (2020) Regression approach for modeling covid-19 spread and its impact on stock market
18. Maier BF, Brockmann D (2020) Effective containment explains subexponential growth in recent confirmed covid-19 cases in China. Science 368(6492):742–746
19. Pirouz B, Haghshenas SS, Haghshenas SS, Piro P (2020) Investigating a serious challenge in the sustainable development process: analysis of confirmed cases of covid-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis. Sustainability 12(6):2427
20. Gu C, Zhu J, Sun Y, Zhou K, Gu J (2020) The inflection point about covid-19 may have passed. Sci Bull 5:98
21. Batista M (2020) Estimation of the final size of the covid-19 epidemic. medRxiv
22. Batista M (2020) Estimation of the final size of the second phase of the coronavirus epidemic by the logistic model. medRxiv
23. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley-Interscience, London
24. Buizza R (2020) Weather-inspired ensemble-based probabilistic prediction of covid-19
25. Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus covid-19. PLoS ONE 15:1–8, 03
26. Stubinger J, Schneider L (2020) Epidemiology of coronavirus covid-19: Forecasting the future incidence in different countries. Healthcare 99(8):2
27. Müller M (2007) Dynamic time warping. Springer, Berlin, pp 69–84
28. Tobias A (2020) Evaluation of the lockdowns for the sars cov 2 epidemic in italy and spain after one month follow up. Sci Total Environ 725:138539
29. Xu S, Clarke C, Shetterly S, Narwaney K (2020) Estimating the growth rate and doubling time for short-term prediction and monitoring trend during the covid-19 pandemic with a sas macro. medRxiv
30. Li Y, Liang M, Yin X, Liu X, Hao M, Hu Z, Wang Y, Jin L (2020) Covid-19 epidemic outside China: 34 founders and exponential growth. medRxiv
31. Ganess MAAA, Ewees AA, Fan H, Aziz MAE (2020) Optimization method for forecasting confirmed cases of covid-19 in China. J Clin Med 9(3):100
32. Singer H M (2020) Short-term predictions of country-specific covid-19 infection rates based on power law scaling exponents
33. Cassaro FA, Pires LF (2020) Can we predict the occurrence of covid-19 cases? considerations using a simple model of growth. Sci Total Environ 728:138834
34. Zandbergen P, Chakraborty J (2006) Improving environmental exposure analysis using cumulative distribution functions and individual geocoding. Int J Health Geograph 5(23):62
35. Gupta R, Pal SK (2020) Trend analysis and forecasting of covid-19 outbreak in India. medRxiv
36. Dehesh T, Fard H, Dehesh P (2020) Forecasting of covid-19 confirmed cases in different countries with arima models. medRxiv
37. Chakraborty T, Ghosh I (2020) Real-time forecasts and risk assessment of novel coronavirus (covid-19) cases: a data-driven analysis. medRxiv
38. Perc M, Miksic NG, Slavinec M, Stozer A (2020) Forecasting covid-19. Front Phys 8:127
39. Lau H, Khosrawipour V, Kocbach P, Mikolajczyk A, Ichii H, Schubert J, Bania J, Khosrawipour T (2020) Internationally lost covid-19 cases. J Microbiol Immunol Infect 5:67
40. Ivakhnenko AG (1988) Self-organizing methods in modelling and clustering: Gmdh type algorithms. Syst Anal Simul I:86–88
41. Zhao X, Liu X, Li X (2020) Tracking the spread of novel coronavirus (2019-ncov) based on big data. medRxiv
42. Araujo M B, Naimi B (2020) Spread of sars-cov-2 coronavirus likely to be constrained by climate. medRxiv
43. Wang J, Tang K, Feng K, Lv W (2020) High temperature and high humidity reduce the transmission of covid-19
44. Oliveiros B, Caramelo L, Ferreira N C, Caramelo F (2020) Role of temperature and humidity in the modulation of the doubling time of covid-19 cases. medRxiv
45. Tomar A, Gupta N (2020) Prediction for the spread of covid-19 in India and effectiveness of preventive measures. Sci Total Environ 728:138762
46. Hu Z, Ge Q, Li S, Jin L, Xiong M (2020) Artificial intelligence forecasting of covid-19 in China
47. Charte D, Charte F, Garcia S, Jesus MJD, Herrera F (2018) A practical tutorial on autoencoders for nonlinear feature fusion: taxonomy, models, software and guidelines. Inf Fusion 44:78–96
48. Yang Z, Zeng Z, Wang K, Wong S-S, Liang W, Zanin M, Liu P, Cao X, Gao Z, Mai Z, Liang J, Liu X, Li S, Li Y, Ye F, Guan W, Yang Y, Li F, Luo S, Xie Y, Liu B, Wang Z, Zhang S, Wang Y, Zhong N, He J (2020) Modified seir and ai prediction of the epidemics trend of covid-19 in China under public health interventions. J Thoracic Dis 12(3):52
49. Fong SJ, Li NDG, Crespo RG, Herrera-Viedma E (2020) Finding an accurate early forecasting model from small dataset: a case of 2019-ncov novel coronavirus outbreak. Int J Interact Multimedia Artif Intell 6:132–139
50. Fong SJ, Li G, Dey N, Crespo RG, Herrera-Viedma E (2020) Composite monte carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. Appl Soft Comput 93:106282
51. Bandyopadhyay SK, Dutta S (2020) Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release. medRxiv
52. Ghazaly NM, Abdel-Fattah MA, El-Aziz AAA (2020) Novel coronavirus forecasting model using nonlinear autoregressive artificial neural network. Int J Adv Sci Technol 29(5s):68
53. Huang C-J, Chen Y-H, Ma Y, Kuo P-H (2020) Multiple-input deep convolutional neural network model for covid-19 forecasting in China. medRxiv
54. Barabasi AL, Posfai M (2016) Network science. Cambridge University Press, Cambridge
55. Li L, Yang Z, Dang Z, Meng C, Huang J, Meng H, Wang D, Chen G, Zhang J, Peng H, Shao Y (2020) Propagation analysis and prediction of the covid-19. Infect Dis Modell 5:282–292
56. Zhuang Z, Zhao S, Lin Q, Cao P, Lou Y, Yang L, Yang S, He D, Xiao L (2020) Preliminary estimates of the reproduction number of the coronavirus disease (covid-19) outbreak in republic of korea and italy by 5 March 2020. Int J Infect Dis 95:308–310
57. Ziff AL, Ziff RM (2020) Fractal kinetics of covid-19 pandemic. medRxiv
58. Li M, Chen J, Deng Y (2020) Scaling features in the spreading of covid-19

59. Herrmann HA, Schwartz J (2020) Using network science to propose strategies for effectively dealing with pandemics: the covid-19 example. medRxiv
60. Dezső Z, Barabási A-L (2002) Halting viruses in scale-free networks. Phys Rev E 65:055103
61. Lacasa L, Luque B, Ballesteros F, Luque J, Nuño JC (2008) From time series to complex networks: the visibility graph. Proc Nat Acad Sci 105(13):4972–4975
62. Demertzis K, Tsiotas D, Magafas L (2020) Modeling and forecasting the covid-19 temporal spread in greece: an exploratory approach based on complex network defined splines
63. Pujari B S, Shekatkar S M (2020) Multi-city modeling of epidemics using spatial networks: Application to 2019-ncov (covid-19) coronavirus in India. medRxiv
64. Biswas K, Sen P (2020) Space-time dependence of corona virus (covid-19) outbreak
65. Gross B, Zheng Z, Liu S, Chen X, Sela A, Li J, Li D, Havlin S (2020) Spatio-temporal propagation of covid-19 pandemics. medRxiv
66. Pokkuluri KS, Nedunuri ND, Usha US (2020) A novel cellular automata classifier for covid-19 prediction. J Health Sci 10:34–38
67. Fan C, Cai T, Gai Z, Wu Y (2020) The relationship between the migrant population migration network and the risk of covid 19 transmission in China empirical analysis and prediction in prefecture level cities. Int J Environ Res Public Health 17:82
68. Lu FS, Hattab MW, Clemente CL, Biggerstaff M, Santillana M (2019) Improved state-level influenza nowcasting in the united states leveraging internet-based data and network approaches. Nat Commun 10:192
69. Qin L, Sun Q, Wang Y, Wu K, Chen M, Shia B, Wu S (2020) Prediction of number of cases of 2019 novel coronavirus (covid-19) using social media search index. Int J Environ Res Public Health 17(7):72
70. Jahanbin K, Rahmanian V (2020) Using twitter and web news mining to predict covid-19 outbreak. Asian Pac J Trop Med
71. Liu D, Clemente L, Poirier C, Ding X, Chinazzi M, Davis JT, Vespignani A, Santillana M (2020) A machine learning methodology for real-time forecasting of the 2019-2020 covid-19 outbreak using internet searches, news alerts, and estimates from mechanistic models
72. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, Niakan KSR (2020) Predicting covid-19 incidence through analysis of google trends data in Iran: Data mining and deep learning pilot study. JMIR Public Health Surveill 6:e18828
73. Lampos V, Moura S, Yom-Tov E, Edelstein M, Majumder M, Hamada Y, Rangaka MX, McKendry RA, Cox IJ (2020) Tracking covid-19 using online search
74. Dubey AD (2020) Twitter sentiment analysis during covid19 outbreak
75. Petrosillo N, Viceconte G, Ergonul O, Ippolito G, Petersen E (2020) Covid-19, sars and mers: are they closely related? Clin Microbiol Infect 26(6):729–734
76. Coronavirus cases. https://www.worldometers.info/coronavirus/#countries. (Accessed on 31 May 2020)
77. Coronavirus disease 2019 (covid-19) situation report 46. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.pdf?sfvrsn=96b04adf_4. (Accessed on 31 May 2020)
78. Social stigma associated with covid-19. https://www.who.int/docs/default-source/coronaviruse/covid19-stigma-guide.pdf?sfvrsn=226180f4_2. (Accessed on 31st May 2020)
79. Bastian PMAA, Mieghem PV (2020) Fundamental limits of predicting epidemic outbreaks. https://www.nas.ewi.tudelft.nl/people/Piet/papers/TUD2020410_prediction_limits_epidemic_outbreaks.pdf. (Accessed on 20 May 2020)
80. Cia world factbook. https://www.cia.gov/library/publications/the-world-factbook/rankorder/2153rank.html. (Accessed on 26 May 2020)
81. Zou B, Lampos V, Cox IJ (2019) Transfer learning for unsupervised influenza-like illness models from online search data. In: Liu L, White RW, Mantrach A, Silvestri E, McAuley JJ, Baeza-Yates R, Zia L(eds.) The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, pp 2505–2516, ACM