Journal of Discrete Mathematical Sciences and Cryptography

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tdmc20

Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine

Vijander Singh , Ramesh Chandra Poonia , Sandeep Kumar , Pranav Dass , Pankaj Agarwal , Vaibhav Bhatnagar & Linesh Raja

To cite this article: Vijander Singh , Ramesh Chandra Poonia , Sandeep Kumar , Pranav Dass , Pankaj Agarwal , Vaibhav Bhatnagar & Linesh Raja (2020) Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine, Journal of Discrete Mathematical Sciences and Cryptography, 23:8, 1583-1597, DOI: 10.1080/09720529.2020.1784535

To link to this article: https://doi.org/10.1080/09720529.2020.1784535

Published online: 14 Dec 2020.

Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 23 (2020), No. 8, pp. 1583–1597
DOI : 10.1080/09720529.2020.1784535

Taylor & Francis
Taylor & Francis Group

Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine

Vijander Singh
Department of Computer Science and Engineering
Manipal University Jaipur
Jaipur 303007
Rajasthan
India

Ramesh Chandra Poonia*
Amity Institute of Information Technology
Amity University Jaipur
Jaipur 303002
Rajasthan
India

and

Cyber-Physical Systems Laboratory
Department of ICT and Natural Sciences
Faculty of Information Technology and Electrical Engineering
Norwegian University of Science and Technology
Ålesund
Norway

Sandeep Kumar
Amity School of Engineering & Technology
Amity University Jaipur
Jaipur 303002
Rajasthan
India

E-mail: rameshcpoonia@gmail.com (Corresponding Author)

Pranav Dass[†]
Department of Computer Science and Engineering
Bharati Vidyapeeth's College of Engineering
Paschim Vihar
New Delhi 110063
India

Pankaj Agarwal
Amity School of Engineering & Technology
Amity University Jaipur
Jaipur 303002
Rajasthan
India

Vaibhav Bhatnagar
Linesh Raja
Department of Computer Applications
Manipal University Jaipur
Jaipur 303007
Rajasthan
India

Abstract

Predicting the probability of CORONA virus outbreak has been studied in recent days, but the published literature seldom contains multiple model comparisons or predictive analysis of uncertainty. Time series parameters are the core factors influencing infectious diseases such as severe acute respiratory syndrome (SARS) and influenza. As a global pandemic is imminent, the prediction of real-time transmission of COVID-19 is crucial. The objective of this paper is to produce a real-time forecasts using the SVM model. The purpose of this study is to investigate the Corona Virus Disease 2019 (COVID-19) prediction of confirmed, deceased and recovered cases. This prediction will help to plan resources, determine government policy, and provide survivors with immunity passports, and use the same plasma for care. In this analysis, data including attributes such as location wise confirmed, deceased, recovered COVID-19, longitude and latitude were collected from January 22, 2020 to April 25, 2020 worldwide. Support Vector Machine was used to explore the impact on identification, deceased, and recovery.

Subject Classification: *97R40.*

Keywords: *Pandemic, Support vector machine, COVID-19, Machine learning.*

[†]*E-mail:* **pranav.dass@bharatividyapeeth.edu**

1. Introduction

On 31 December 2019, an outbreak of "unknown cause pneumonia" was reported by the World Health Organization (WHO) in Wuhan City, Hubei Province, China – the 7th largest city in China with more than 10 million people. As of April 26, 2019-nCoV confirmed more than 2,971,936 cases, 205,957 deaths are registered and 873,868 patients are either recovered or discharged globally, including 210 countries and territories worldwide [1]. Corona virus, linked to extreme outbreaks, exacerbates public global health issues with its fast transnational spread inspired by increased trade and global travel [2]. Suitability mapping of Corona virus transmission risk is necessary for making public health plans and disease control policies, which are particularly important in medical facility lacking areas. Wuhan, China, announced a new outbreak of coronavirus disease (COVID-19) caused by extreme acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in December 2019 [3]. In this paper, the first basic work was proposed to try to get some insight into the prediction of the outbreak from time series data. Covid-19 has been reported to have human-to - human transmittable [4] which has subsequently created high interest not only in China but worldwide

2. Related Work

Several cases of pneumonia were identified in several hospitals in Wuhan City, Hubei Province, in December 2019, for a new form of acute respiratory infection caused by coronavirus [5]. This year, coronavirus isolated from the lower respiratory tract of Wuhan in patients with extreme pneumonia is a new form of coronavirus disease called COVID-19. In the paper [6] performed an isolation and quarantine-control feasibility study on COVID-19. They developed a transmission model to parameterize COVID-19. This model is used to assess contact tracing efficacy and case isolation in combating this pandemic. In certain possible cases, case isolation alone would be impossible to control transmission. Isolation was more successful when there was no transmission and quick delay from initiation to isolation prior to symptom onset.

Rajesh Singh-R. Adhikari [7] study on age-structured effects on social distance and contrasted Chinese, Indian, and Italian peoples' social interaction structure and age. They developed a mathematical model to determine the impact of social distancing. They conclude that three-week lockdown is not enough, but prolonged lockdown periods will minimize

COVID-19 cases in India. This will help government and health agencies devote sufficient resources to address the epidemic in the days and weeks ahead, including clinicians, beds and intensive care services. Unless the Italian outbreak follows a similar trend as in Hubei Province , China, the number of newly infected patients will decrease within 3â€“4 days of the exponential trend. Nonetheless, this can not be anticipated at present due to disparities between social distance programs and the potential for rapid development of dedicated facilities in China. Currently, various statistical techniques like time series models [8], multivariate linear regression [9], grey forecasting models [10] [11], back propagation neural networks [12] [13] [14] and simulation models [15] [16] were applied to forecast prevalent cases. There are many factors that affect epidemics and therefore trends and randomness determine the overall spread of the disease more descriptive analysis in a paper [2]. Hence, the quantitative approaches stated are inadequate to determine the randomness of the disorder, and it is difficult to generalize the models. Different models have been used in recent studies to estimate the frequency, prevalence and mortality levels of COVID-19 in China. For example, A method was developed to forecast the the current trend via statistical methods and approximate the COVID-19 outbreak size in China [17]. During this time, investigation of COVID-19 pandemic was carried out in Asia (Mainland China) and few European countries(France and Italy) [18] [19]. The enhanced Adaptive Neuro Fuzzy Inference System (ANFIS) was used to estimate the number of reported cases in China using an Improved Flower Pollination Algorithm using the Swarm Algorithm. [20]. The Patient Information-based Algorithm was designed to forecast the COVID-19 death rate in real time with the data that is publicly available [21]. In brief, there are number of literature studies to estimate COVID-19 spread in China. However, outspread of virus has hit the all over the world. As of April 25, 2020, the Effective Mortality Rate is increased due to rise in the infection. It is therefore essential to analyze the situation of the COVID-19 pandemic and forecast the occurrence pattern.

This paper estimates the occurrence of COVID-19 in the various countries, with the virus spreading quickly and generating tragic results. The details reviewed in this analysis are from 22 January 2020 to 25 April 2020. The main points of reference are as follows: (a) In this paper a prediction model to forecast the confirmed cases of the COVID-19 has been proposed. (b) SVM machine learning technique has been considered for predication. The remaining analysis is structured as follows. In the *section 3*, collected data set is used in proposed SVM machine learning technique to execute and evaluate a case estimate test, and the result discussed is

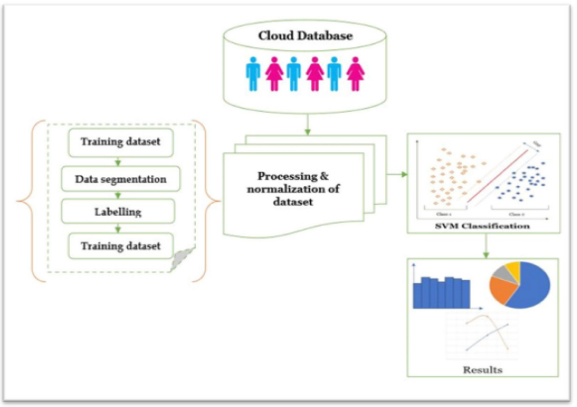


Figure 1

Research Methodology

presented in *Section 4*, with the conclusion mentioned in Section 5 with its future scope.

3. Materials and Method

The methodology layout that is used and evaluated in the current study is hereby presented in Figure 1.

3.1 *Data Collection*

Worldwide data were collected from 22 January 2020 to 25 April 2020, including regular confirms, deceased cases and recovered cases. COVID-19 data were obtained from the Johns Hopkins University Public Repository Center for Systems Science and Engineering (CSSE)[1]. Johns Hopkins CSSE Data repository updated on regular basis.

3.2 *Statistical analysis*

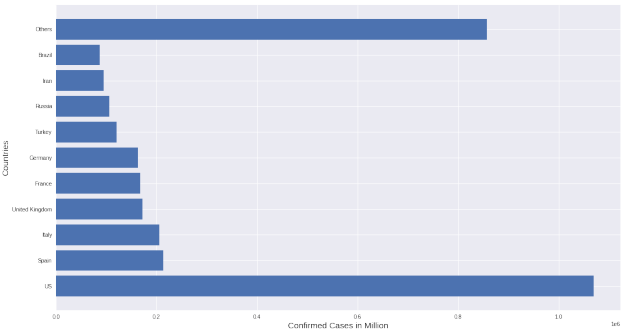
The planet has been put on hold by the ongoing coronavirus epidemic, which has now exceeded the combined death toll of 2003 extreme acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS). The new coronavirus disease 2019 (COVID-19) public health burden is expected to increase and it is time for urgent decisions[4,5]. Mortality rate calculated by dividing the total death by total confirmed

[1] https://systems.jhu.edu/

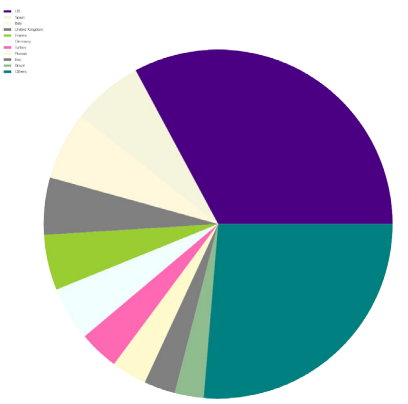
cases in the world. Datewise cases in all the countries are represented by world_cases list.

Mortality Rate = Total Deaths till date / Total confirmed cases in the world

Confirmed, deceased and recovered number of patients till date helps to predict the mortality rate of the COVID-19 which is calculated as shown in Figure 7. On April 26th 2020, more than 29 million confirmed cases are reported, 200 hundred deaths and about 1 million patients are recovered. Generally lockdown is applied for 3 weeks in most of the countries to block the spread chain of the Corona virus. Proposed study 21 days considered



[Case A]



[Case B]

Figure 2

Top 10 affected countries of the world and others in the form of (a) Bar graph(b) Pie chart

to predict the confirm, deceased and recovered cases in the world. The resources may be arranged accordingly on weekly basis and based on the next week data the prediction can be more accurate as this is generalize model and fetching the data directly from repository. Country wise data helps to identify the hot spot for the corona and also to apply or release the lockdown. China shift at 9th position from 1st place in December. USA is struggling at 1st place as shown in Figure 2. This is an alarming situation for the rest of the countries and apply strict and necessary policies for better health.

3.3 *Proposed Prediction using SVM Model*

3.3.1 SVM Hyperplane

A line can be used mathematically to distinguished linearly [22] separable data as shown in in Figure 3. The line equation 1 is as follows.

$$y = m \times x + c \tag{1}$$

Renaming x with ψ_1 and y with ψ_2, the equation will change to $m \times \psi_1 - \psi_2 + C = 0$. If $X = (\psi_1, \psi_2)$ and $w = (m, -1)$, then equation 2) of the hyperplane:

$$w \times x + C = 0 \tag{2}$$

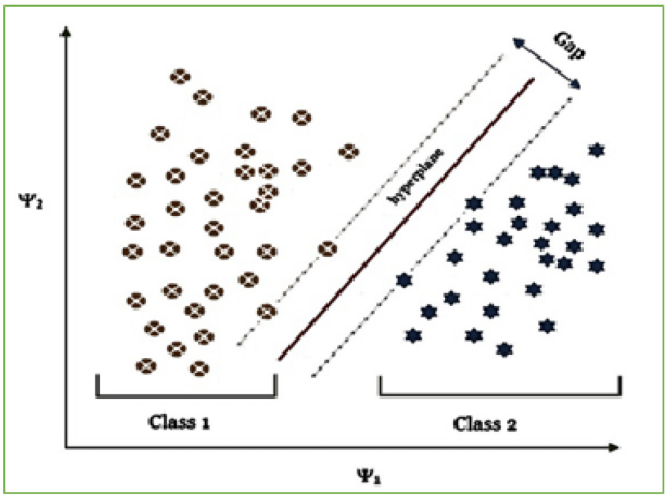


Figure 3

SVM 2D Coordinate system

3.3.2 SVM Classifier

To make prediction hyperplane is defined by hypothesis function h_0 as shown in Eq. 3.

$$h_0(x_i) = \begin{cases} +1, & \text{if } w \times x + C \geq 0 \\ -1, & \text{otherwise.} \end{cases} \tag{3}$$

As mentioned above the hyperplane will be classified as class +1, as class -1.

3.3.3 Derivation of SVM optimization problem

The SVM learning algorithm generate optimum hyperplane which separate the data accordingly. To calculate w and C of hyperplane, following equation need to be maximize.

$$\max_{w,C} M \text{ subject to } \alpha_i \geq M, i = 1 \ldots z \tag{4}$$

where α and M are defined as,

$$\alpha = y \times \left(\frac{w}{\| w \|} \times + \frac{c}{\| w \|} \right)$$

$$M = \min_{i=1 \ldots z} y_i \left(\frac{w}{\| w \|} \times x + \frac{c}{\| w \|} \right)$$

It is also known that,

$M = \frac{F}{\|w\|}$ the above optimization equation can be re-defined as,

$$\max_{w,C} M \text{ subject to } \beta_i \geq M, i = 1 \ldots z \tag{5}$$

Maximizing M does not change the optimized result, therefore by putting $F = 1$, the above equation can be defined as

$$\max_{w,C} \frac{1}{\| w \|} \text{ subject to } \beta_i \geq M, i = 1 \ldots z \tag{6}$$

This maximization equation can be written as minimization equation:

$$\min_{w,C} \| w \| \text{ subject to } \beta_i \geq M, i = 1 \ldots z \tag{7}$$

As we know that $l2$ optimization is more stable than $l1$, so again above-mentioned equation can be re-written as

$$\min_{w,C} \frac{1}{2} \| w \|^2 \text{ subject to } y_i (m \times x + C) - 1, i = 1 \ldots z \tag{8}$$

The above equation h, is defined as SVM Optimization Equation.

This paper adopts the concept of Support Vector Machine (SVM), which is likely to be a suitable technique for time series dataset with world's total population. Support vector machine (SVM) model was given by Vapnik et. at. for pattern classification. [23]. SVM uses a technique to the upper limit for error by minimizing the boundary distance between the training data and hyperplane opposed to traditional ways of reducing observational testing errors [24]. Train to SVM is equal to solve a quadratic programming problems with linear restrictions while training of other networks which involves non-linear optimization with the risk of getting stuck in the local minima. In SVM, the problem-solving process only depends on a set of training data, which is called a support vector.

Data is plotted into train and test set for train and test the model to measure the accuracy of the model. Support Vector model is suitable for time series dataset and we have applied SVM model to predict the confirm, deaths and recovered patients. The SVM kernel is a tuning parameter that takes input space of low dimensions and transforms it into a higher dimensional space, i.e. translates nonseparable problem into a separable problem. Here 'poly', 'sigmoid' and 'rbf' has been used as the problem is linear problem.

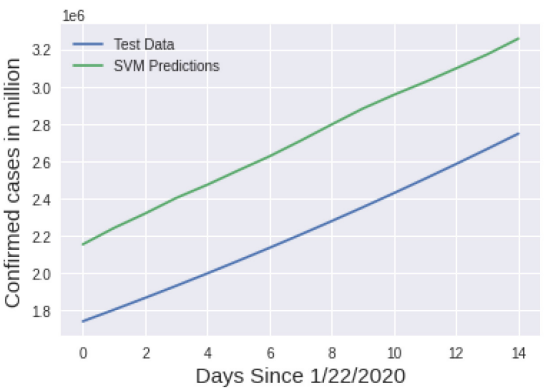


Figure 4

Test the SVM trained model

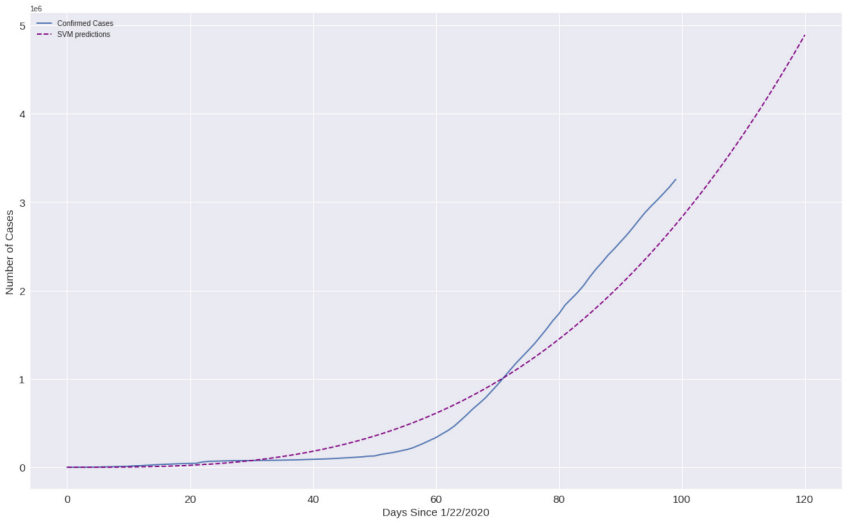


Figure 5

Confirmed versus Predicted cases using SVM model

After train the model the test dataset has been used to calculate the errors. Which shows both lines are parallel as shown in Figure 4.

Confirmed cases and SVM predicted cases are mostly walking together and also the SVM prediction is ahead of the confirmed data available which is upto 25th April, 2020 as shown in graph Figure 5.

```
print('SVM future predictions:')
set(zip(future_forcast_dates[-21:],svm_pred[-21:]))

SVM future predictions:
{('04/27/2020', 2198998.510365022),
 ('04/28/2020', 2268392.3243681914),
 ('04/29/2020', 2339231.7910792357),
 ('04/30/2020', 2411531.8141382462),
 ('05/01/2020', 2485307.2971788174),
 ('05/02/2020', 2560573.143835736),
 ('05/03/2020', 2637344.257747484),
 ('05/04/2020', 2715635.5425509936),
 ('05/05/2020', 2795461.90188129),
 ('05/06/2020', 2876838.23937471),
 ('05/07/2020', 2959779.4586650855),
 ('05/08/2020', 3044300.4633943564),
 ('05/09/2020', 3130416.1571934945),
 ('05/10/2020', 3218141.4437003857),
 ('05/11/2020', 3307491.226548744),
 ('05/12/2020', 3398480.4093771703),
 ('05/13/2020', 3491123.8958256966),
 ('05/14/2020', 3585436.589524103),
 ('05/15/2020', 3681433.3941101558),
 ('05/16/2020', 3779129.2132240064),
 ('05/17/2020', 3878538.9504965073)}
```

Figure 6

Prediction for next 3 weeks using SVM model

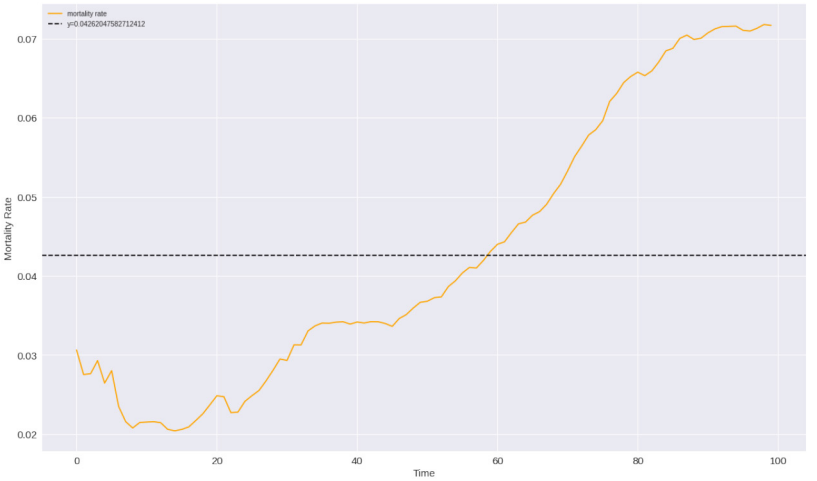


Figure 7

Mortality Rate over the world

Similar to data available at JH University repository, SVM predict the date wise confirmed, deceased and recovered numbers as shown in Figure 6 date wise.

This is also shown, that the outbreak rate will be significantly and quickly decreased in order to achieve an appreciable decrease in the disease peak and mortality rate as seen in Figure 6.

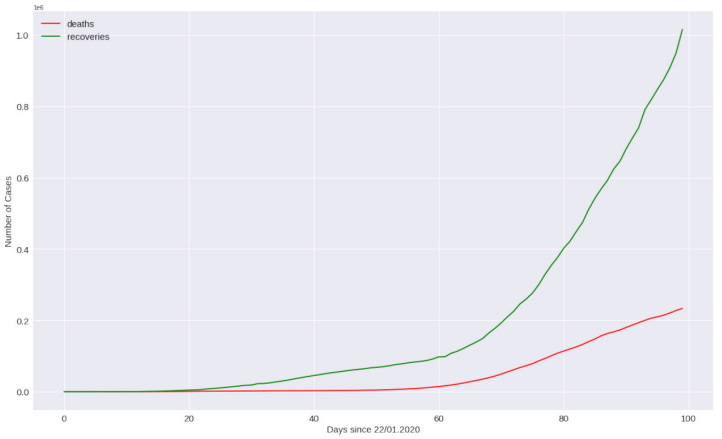


Figure 8

Deaths and recovery

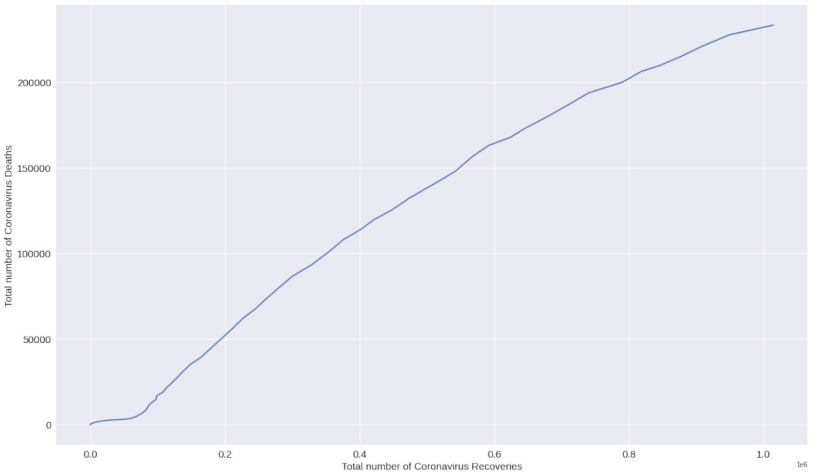


Figure 9

Deaths per recovery

4. Results and Discussion

The forecasting of predicted patient of COVID-19 depends on the scores of mentioned attributes by using SVM calculations. In addition, we find the ideal hyperplane using RBF and C. It helped us to compare the hyperplane parameters to better examine support vectors. At the same time, we performed statistical analysis through bar charts to distinguish groups of subjects. SVM produces optimized performance values to forecast the predicted COVID-19 cases using kernel function.

The outcomes generated by the machine learning system and Hyperplane are matched to prove the theory of support vector algorithms.

Forecasting of COVID-19 pandemic confirmed cases was forecasted by the SVM machine learning algorithm. It requires effective strategies to prevent and monitor disease spread. It is very much crucial for the government to carry out the mandatory resource allocation of medical supplies, healthcare equipment's, agriculture and control of manufacturing activities to maintain the national economy growth. Therefore, it is very much important to develop as effective forecasting model which can provide the advantage to the government to decide on the macroeconomic emergency strategies. SVM model is one of the good quality machine learning technique used for prediction because of its simplicity. In this analysis the present COVID-19 pandemic situation all around the world

is highlighted, and the Continuing pattern and severity of outbreak have been calculated by the SVM method.The proposed research is an attempt to introduce SVM models to forecast COVID-19 cases all over the world.

5. Conclusion

The aim of the paper is to examine and forecast the registered, deceased, deaths and death numbers per reported case (Mortality Rate) based on Corona-19's world health data for the world population. Findings from the author indicate that COVID-19's regular mortality is positively correlated with number of confirmed cases. It may also be dependent upon the population's dietary routine and immune system. This study suggests that an emergency situation can awaken before the proper vaccine is invented. Some critical elements were considered by writers, including individual countries, provinces and reported, deaths, and patients recovered, updated daily. If more information is available, the model can be extended to predict more parameters and can also be implemented in individual cities. This model predict that the number of cases in western Europe and Spain will increase in coming weeks and COVID-19 epidemic would continue[25]. In the end, the study carried out on the prediction of COVID 19 proposes the suggestive methods which can be employed for minimization of COVID 19 transmission among human by preventing them from migrating from one place to another and also reposing various group communal activities.[26].

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Coronavirus Cases:, https://www.worldometers.info/coronavirus/, accessed: 2020-04-10.

[2] Bhatnagar, V., Poonia, R. C., Nagar, P., Kumar, S., Singh, V., Raja, L., & Dass, P. (2020). Descriptive analysis of COVID-19 patients in the context of India. *Journal of Interdisciplinary Mathematics*, 1-16.

[3] A. E. Gorbalenya, Severe acute respiratory syndrome-related coronavirus–the species and its viruses, a statement of the coronavirus study group, BioRxiv (2020).

[4] C. Wang, P. W. Horby, F. G. Hayden, G. F. Gao, A novel coronavirus outbreak of global health concern, The Lancet 395 (10223) (2020) 470-473.

[5] K. Su, L. Xu, G. Li, X. Ruan, X. Li, P. Deng, X. Li, Q. Li, X. Chen, Y. Xiong, et al., Forecasting inuenza activity using self-adaptive ai model and multi-source data in chongqing, china, EBioMedicine 47 (2019) 284-292.

[6] K. Liu, Y. Chen, R. Lin, K. Han, Clinical features of covid-19 in elderly patients: A comparison with young and middle-aged patients, Journal of Infection (2020).

[7] R. Singh, R. Adhikari, Age-structured impact of social distancing on the covid-19 epidemic in india, arXiv preprint arXiv:2003.12055 (2020).

[8] V. Kurbalija, M. Radovanović, M. Ivanović, D. Schmidt, G. L. von Trzebi-atowski, H.-D. Burkhard, C. Hinrichs, Time-series analysis in the medical domain: A study of tacrolimus administration and influence on kidney graft function, Computers in biology and medicine 50 (2014) 19-31.

[9] M. C. Thomson, A. M. Molesworth, M. H. Djingarey, K. Yameogo, F. Belanger, L. E. Cuevas, Potential of environmental models to predict meningitis epidemics in africa, Tropical Medicine & International Health 11 (6) (2006) 781-788.

[10] Y.-w. Wang, Z.-z. Shen, Y. Jiang, Comparison of arima and gm (1, 1) models for prediction of hepatitis b in china, PloS one 13 (9) (2018).

[11] L. Zhang, L. Wang, Y. Zheng, K. Wang, X. Zhang, Y. Zheng, Time prediction models for echinococcosis based on gray system theory and epidemic dynamics, *International journal of environmental research and public health* 14 (3) (2017) 262.

[12] Q. Liu, Z. Li, Y. Ji, L. Martinez, U. H. Zia, A. Javaid, W. Lu, J. Wang, Forecasting the seasonality and trend of pulmonary tuberculosis in jiangsu province of china using advanced statistical time-series analyses, *Infection and drug resistance* 12 (2019) 2311.

[13] H. Ren, J. Li, Z.-A. Yuan, J.-Y. Hu, Y. Yu, Y.-H. Lu, The development of a combined mathematical model to forecast the incidence of hepatitis e in shanghai, china, BMC infectious diseases 13 (1) (2013) 421.

[14] X. Zhang, Y. Liu, M. Yang, T. Zhang, A. A. Young, X. Li, Comparative study of four time series methods in forecasting typhoid fever incidence in china, PloS one 8 (5) (2013).

[15] C. Orbann, L. Sattenspiel, E. Miller, J. Dimka, Defining epidemics in computer simulation models: How do definitions influence conclusions?, Epidemics 19 (2017) 24-32.

[16] E. O. Nsoesie, R. J. Beckman, S. Shashaani, K. S. Nagaraj, M. V. Marathe, A simulation optimization approach to epidemic forecasting, PloS one 8 (6) (2013).

[17] Q. Li, W. Feng, Y.-H. Quan, Trend and forecasting of the covid-19 outbreak in china, *Journal of Infection* 80 (4) (2020) 469-496.

[18] K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. Hyman, P. Yan, G. Chowell, Real-time forecasts of the covid-19 epidemic in china from february 5th to february 24th, 2020, Infectious Disease Modelling 5 (2020) 256-263.

[19] D. Fanelli, F. Piazza, Analysis and forecast of covid-19 spreading in china, italy and france, Chaos, Solitons & Fractals 134 (2020) 109761.

[20] M. A. Al-qaness, A. A. Ewees, H. Fan, M. Abd El Aziz, Optimization method for forecasting con_rmed cases of covid-19 in china, Journal of Clinical Medicine 9 (3) (2020) 674.

[21] Z. Ceylan, Estimation of covid-19 prevalence in italy, spain, and france, Science of The Total Environment (2020) 138817.

[22] R. Gandhi, Support vector machine | introduction to machine learning algorithms, Towards Data Science (2018).

[23] V. Vapnik, The nature of statistical learning theory, springer science, & business media (2013).

[24] B.-C. Kuo, H.-H. Ho, C.-H. Li, C.-C. Hung, J.-S. Taur, A kernel-based feature selection method for svm with rbf kernel for hyperspectral image classification, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7 (1) (2013) 317-326.

[25] J. Yuan, M. Li, G. Lv, Z. K. Lu, Monitoring transmissibility and mortality of covid-19 in europe, *International Journal of Infectious Diseases* (2020).

[26] A. Remuzzi, G. Remuzzi, Covid-19 and italy: what next?, The Lancet (2020).

Received May, 2020

Revised August, 2020