Modeling of Coronavirus Behavior to Predict it's Spread

Shakir Khan[1], Amani Alfaifi[2]*
College of Computer and Information Sciences
Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia

***Abstract*—With the increasing presence and feast of infectious diseases and their fatalities in densest areas, many academics and societies have become fascinated in discovering new behaviors to predict these diseases' feast behaviors. This media will help them to plan and contain the disease better in trivial provinces and thus decrease the beating of human lives. Some cases of an indeterminate cause of pneumonia occurred in Wuhan, Hubei, China, in December 2019, with clinical presentations closely resembling viral pneumonia. In-depth analyzes of the sequencing from lower respiratory tract samples discovered a novel coronavirus, called 2019 novel coronavirus (2019-nCoV). Current events showed us how easily a coronavirus could take root and spread—such viruses transmitted easily between persons. To cure with these infections, we applied time series forecasting model in this paper to predict possible coronavirus events. The forecasting model applied is SIR. The results of the implemented models compared with the actual data.**

Keywords—COVID-19; coronavirus; SIR model; data mining; R Software; forecasting model

I. Introduction

On 30 January 2020, the WHO announces the outbreak of COVID-19 as a public health emergency of international concern (PHEIC) by the WHO [14]. This continuing outbreak has since 3 December 2019 spread to over 50 other countries [13]. There are over 1 million cases of confirmed COVID-19 worldwide and deaths over 50 000 as of 20 February [15]. Together with MERSnCoV and SARS-nCoV, it is the 7th member of the coronavirus group which can spread to humans [1,2,16].

Human coronaviruses, which include hCoV-229E, OC43, NL63, and HKU1, cause light respiratory disease. Fatal coronavirus infections that have occurred over the past two decades are extreme coronavirus acute respiratory syndrome (SARS-CoV) and coronavirus respiratory syndrome in the Middle East [3]. Coronavirus disease (COVID-19) is a recently discovered coronavirus-caused infectious disease. The majority of people diagnosed with COVID-19 have mild to moderate respiratory diseases [4]. This way has drawn considerable interest not only in China but globally.

A brief definition for the term "data mining" is to extract the useful information and patterns from large data sources" [11]. Across different areas, data mining used to help to increase the quality and efficacy of pattern detection and analysis of such events using existing statistical evidence [32]. Data mining technology is the process by which we perform all sorts of analyzes on vast volumes of data [33, 34]. This paper goals to use perceptions made possible from data mining methods to relief in predicting coronavirus feast. Large volumes of data can be cumbersome and repetitive to compile and analyze. However, the underlying patterns in the data identified, such that the predicted occurrence of coronavirus is known beforehand. Forward-thinking techniques joining from the grounds of computer science, mathematics, and data science are necessary for discovering these underlying patterns [12]. Because data mining is a vast assembly of procedures and has extensive solicitation, and agreed data mining concept differs slightly based on the source. The data mining can also be named like "Big Data" or "Data Science" as the alternative name of data mining [19].

Various imminent characteristics of our lives rely on historical data arithmetic analysis. For example, prediction of illness, changes in stock market activities, weather prediction, etc. can be forecasted only if we can find a pattern in historical data due to time and it can be any ways for example daily, weekly, monthly, or annually. This form of forecast is commonly called Forecasting of the Time Series. Observations were sequentially taken in time, usually called time series [5]. Mathematical models have been applied to study a variety of communicable disease outbreaks [8], [9], [10].

The formal description of predictive modeling given by [18] is "the method of evolving a mathematical tool or prototypical that produces a perfect prediction". Predictive demonstrating could be engaged to compute information accessible and to create better conversant result based on different information finding. Modeling techniques offer computational capabilities in circumstances with vast quantities of evidence to build prototypes with the prognostic implication that help in hands-on conclusions. Because of overlapping algorithms to discover unseen facts in data, prognostic modeling closely correlated with data mining [19].

Data spring-cleaning, structure, statistical analysis, prognostic modeling, and statistics picturing performed for this project with R software system and R-studio GUI have used widely. R is a programming language open-source platform with different statistical functionalities. The R environment is interactive computing, graphical display, statistics, visualization, and data manipulation suite. Robert Gentleman and Ross Ihaka of the Department of Statistics at the University of Auckland initially wrote R. Currently, R is a product of user assistances around the world. Apart from users who build innovative features, there are thousands of built

*Corresponding Author

packages open on the Robust R Archive Network (CRAN) with countless features and capabilities [20].

For mathematical epidemiology, the SIR model is perhaps the most commonly used [23, 24 and 25]. It widely used for a variety of purposes. Often as the core of further multifarious epidemiological prototypes for the transmission of communicable diseases [24, 26, 27], and other times to research the propagation of phenomena in other regions, such as rumor spread [28], computer viruses [29], Dissemination of information in Web fora [30], or the behavior of investors in stock markets [31].

For this paper, we used statistical models of time series to forecast possible cases of coronavirus. Because coronavirus is one of the contagious viruses, it is essential to estimate the number of cases so that the government can take the appropriate steps and measures to avoid its spread to treat or avoid viruses.

II. Related Works

S. Eubank et al. [6] analyzed the algorithmic and structural properties of Portland, Oregon, from extensive social communication networks. A bipartite graph composed of individuals and places created. The individuals are nodes, while edges reflect places. This binary configuration does not provide users with node to node interaction information. Using a CL-model they use the method of a random graph. The CL-model correctly mimics the dataset's critical features. Alternatively, algorithms of rapid approximation developed to measure basic structural properties. Such studies investigated the impact of political decisions on the management of large-scale urban diseases. It study demonstrates the effect of different algorithms that deal with the pattern of spreading the disease but do not visualize its effect.

An epidemiological model developed by K. Wang et al. to explain the flu when its transmission through a network of human contacts. The aim of this study is to build an ABM method combining GIS and civic ecosystems to mimic the spread of influenza. Using JAVA and GIS software, the model was developed using the Repast Symphony framework. A system developed to simulate the spread and control of influenza in a specific area. the model defined influenza using a mathematical relationship among the probability of transmission, the distance between two persons, the latent duration, The time between infections and death, the rate of cure. For example, users could modify the results of the simulation by changing the time-value to the hospital. [7].

The research, as mentioned earlier, and models offer useful information which allows medical leaders to take better outbreak protocol decision making. Nonetheless, each model mentioned above lacks a particular feature, whether it is computational capacity, necessary variables that are essential in calculating an accurate pattern of virus spread, susceptibility to disease expansion, or the incapability to measure over one form of spreading disease.

SEIR points respectively to the Susceptible, Exposed, Infectious, and Removed or Recovered. These results are designed on the basis SIR but gives a variable to the container Exposed. Susceptible relates to persons that may acquire the infection and be carriers if infect, the exposed are persons already infected but asymptomatic, the infectious are persons that display signs of infection. They may spread the disease, eliminate, or recover the virus are previously infected persons who are no longer infectious and who are now immune to the virus [17].

III. Methodology

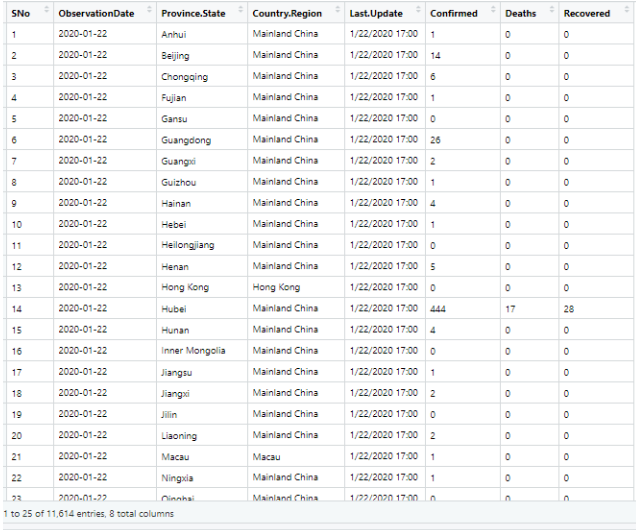
A. Dataset

The dataset used is the "COVID 19 data.csv" file, which the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) periodically updates. The dataset contains data about COVID-19, which is the Province/State and Country/Region also the date and confirmed, deaths, and recovered cases number the data was collected from the starting of the virus spread as shown (screenshot) in Fig. 1.

B. Data Visualization

From Fig. 2 to 7, data were visualized as most of the countries where COVID19 is spread, the rates of Infect, recovery, and death caused by the epidemic, as well as the rates of new cases in the most affected countries.

By using the R program, we visualize major outbreaks for the top 10 countries in Fig. 2.



SNo	ObservationDate	Province.State	Country.Region	Last.Update	Confirmed	Deaths	Recovered
1	2020-01-22	Anhui	Mainland China	1/22/2020 17:00	1	0	0
2	2020-01-22	Beijing	Mainland China	1/22/2020 17:00	14	0	0
3	2020-01-22	Chongqing	Mainland China	1/22/2020 17:00	6	0	0
4	2020-01-22	Fujian	Mainland China	1/22/2020 17:00	1	0	0
5	2020-01-22	Gansu	Mainland China	1/22/2020 17:00	0	0	0
6	2020-01-22	Guangdong	Mainland China	1/22/2020 17:00	26	0	0
7	2020-01-22	Guangxi	Mainland China	1/22/2020 17:00	2	0	0
8	2020-01-22	Guizhou	Mainland China	1/22/2020 17:00	1	0	0
9	2020-01-22	Hainan	Mainland China	1/22/2020 17:00	4	0	0
10	2020-01-22	Hebei	Mainland China	1/22/2020 17:00	1	0	0
11	2020-01-22	Heilongjiang	Mainland China	1/22/2020 17:00	0	0	0
12	2020-01-22	Henan	Mainland China	1/22/2020 17:00	5	0	0
13	2020-01-22	Hong Kong	Hong Kong	1/22/2020 17:00	0	0	0
14	2020-01-22	Hubei	Mainland China	1/22/2020 17:00	444	17	28
15	2020-01-22	Hunan	Mainland China	1/22/2020 17:00	4	0	0
16	2020-01-22	Inner Mongolia	Mainland China	1/22/2020 17:00	0	0	0
17	2020-01-22	Jiangsu	Mainland China	1/22/2020 17:00	1	0	0
18	2020-01-22	Jiangxi	Mainland China	1/22/2020 17:00	2	0	0
19	2020-01-22	Jilin	Mainland China	1/22/2020 17:00	0	0	0
20	2020-01-22	Liaoning	Mainland China	1/22/2020 17:00	2	0	0
21	2020-01-22	Macau	Macau	1/22/2020 17:00	1	0	0
22	2020-01-22	Ningxia	Mainland China	1/22/2020 17:00	1	0	0
23	2020-01-22	Qinghai	Mainland China	1/22/2020 17:00	0	0	0

1 to 25 of 11,614 entries, 8 total columns

Fig. 1. Dataset.

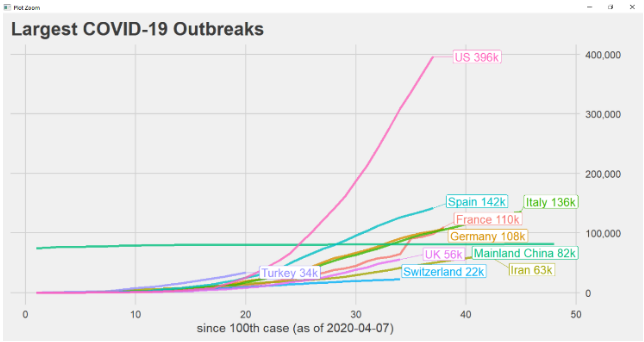


Fig. 2. Major Outbreaks.

And daily new cases for the top 20 counties in Fig. 3:

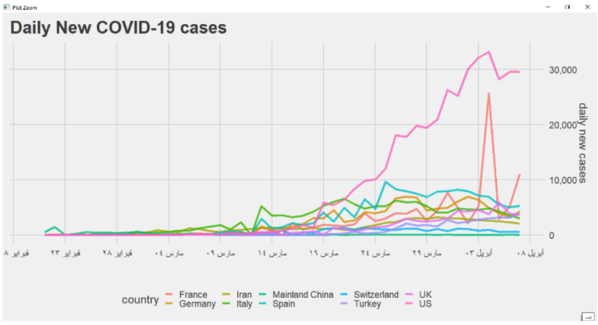


Fig. 3. Daily New Cases.

Daily new COVID-19 cases in the 12 most-affected countries in Fig. 4:

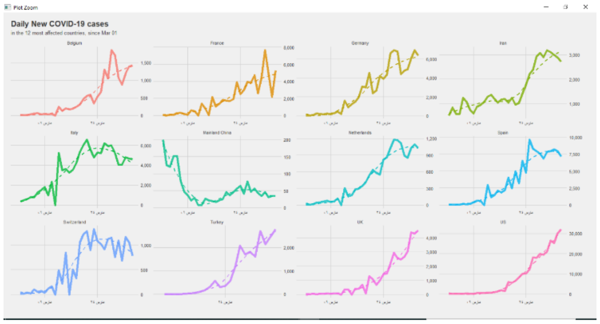


Fig. 4. Daily New COVID-19 Cases in the 12 Most-Affected Countries.

Daily new deaths in the 12 most-affected countries in Fig. 5:

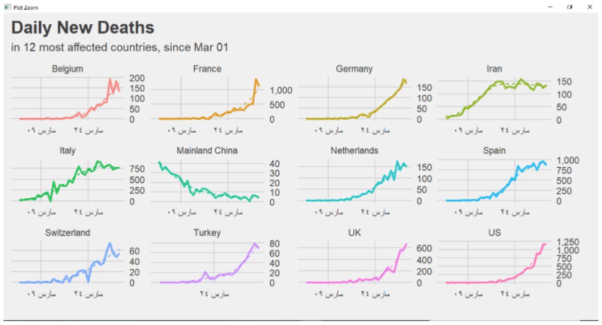


Fig. 5. Daily New Deaths in 12 Most Affected Countries.

Finally, daily recovered in 12 most affected countries in Fig. 6:

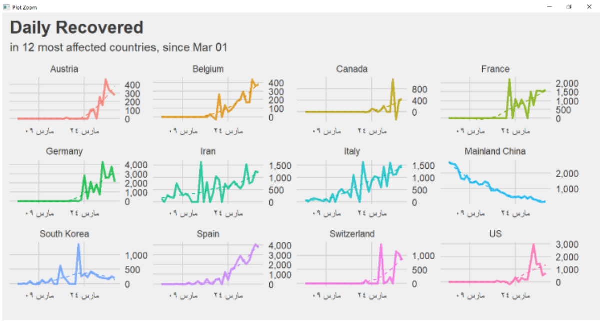


Fig. 6. Daily Recovered in 12 Most Affected Countries.

C. SIR Model

Here a briefly explain the features of the essential Susceptible-Infected- Recovered (SIR) system which used define the recent COVID-19 outbreak. The original SIR model, in which Kermack and McKendrick modified a Malthusian growth model, is a model known to simulate epidemic growth using ODE.

The SIR model describes three-stage rules for the infection. The first rule is the unsafe condition (S) in which an agent is likely to become contaminated at any given point in time. The second is infection (I) when many neighbors are also in this state. The agent will switch to that state. An agent moves to the third state that is recovered (R) after a given period [21]. The SIR diagram in Fig. 7 shows how individuals move.

According to [22], S, I, and R are susceptible, infectious, and removed individuals, and where parameters β and π are the rate of infection and the rate of recovery. The equations for every time t are defined as follows:

$dS\ (t)\ /\ dt = -\ \beta S\ (t)\ I\ (t),$

$dI\ (t)\ /\ dt = \beta S\ (t)\ I\ (t) - \gamma\ I\ (t),$

$dR\ (t)\ /\ dt = \gamma\ I\ (t)$

We first created a time series of that data, divided into infected and recovered data. We can visualize the generated time-series in Fig. 7.

It is interesting to note that there two sub-waves of infection, which may be the time the coronavirus left China and spread worldwide. The two different growths in the number of reported cases (indicating spread) can be more clearly seen in the graph in Fig. 8 after calculating the rate of infection and recovery.

The first wave considered to run from the beginning of the dataset until 14 February, and the second wave is from 15 February onwards until four days before the current date. This cut-off lets us later compare the output of the model with the actual data. At the 95% confidence interval, both the two-sided t-test and the KS-test conducted with the null hypothesis that the two infection growth rates are the same as the output in Table I.

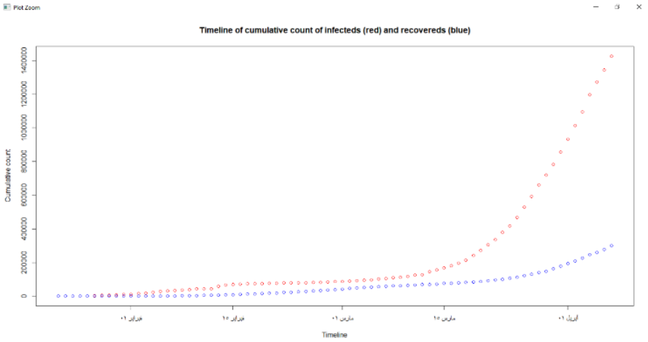


Fig. 7. Timeline of Cumulative Count of Infected and Recoverds.

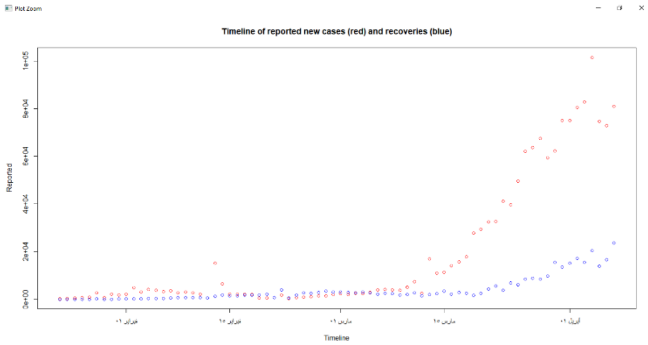


Fig. 8. Timeline of Reportes New Cases and Recoveries.

TABLE I. T-TEST, KS-TEST

t-test	Welch Two Sample t-test data: log(wave.1.results$rate) and log(wave.2.results$rate) t = -4.3879, df = 61.827, p-value = 4.543e-05 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -2.1750871 -0.8135133 sample estimates: mean of x mean of y 7.469233 8.963533
KS-test	Two-sample Kolmogorov-Smirnov test data: log(wave.1.results$rate) and log(wave.2.results$rate) D = 0.45652, p-value = 0.001595 alternative hypothesis: two-sided

So, the inference is that the second wave is more aggressive on average than the first (rejecting H0 at 5 percent t-test significance) since the KS-Test's p-value is above 0.05. In contrast, the t-test is only marginally lower, while the overall shape of the distribution is similar (not enough evidence to reject H0 at 5% concerning the KS-test).

IV. Applying the SIR Model and Discussion

N reflects the total considered population. One condition the SIR equations have to satisfy for all t is:

$$N = S+I+R$$

We will observe linear growths in I in order to evaluate b and c, hence the motivation to divide the data into two waves, as discussed earlier. As we have found that both waves behave similarly, we carry out the modeling using the second wave only. The output is shown in Table II.

The p-value is small for the coefficient, which shows that we have sufficient evidence to reject the null hypothesis and conclude that there is a linear correlation between growth and time. The R-squared values are also very similar to 1 which means that the above regression also provides a good fit for the variance.

TABLE II. REGRESSION SUMMARY

The regression summary of infected	Residuals: Min 1Q Median 3Q Max -29196 1494 4987 8254 10342 Coefficients: Estimate Std. Error t value Pr(>|t|) x -0.18938 0.00549 -34.49 <2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 7588 on 49 degrees of freedom Multiple R-squared: 0.9604, Adjusted R-squared: 0.9596 F-statistic: 1190 on 1 and 49 DF, p-value: < 2.2e-16
The regression summary of recovery	Residuals: Min 1Q Median 3Q Max -5552.2 94.7 793.4 1667.3 3634.0 Coefficients: Estimate Std. Error t value Pr(>|t|) x 0.019180 0.000649 29.55 <2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 1628 on 49 degrees of freedom Multiple R-squared: 0.9469, Adjusted R-squared: 0.9458 F-statistic: 873.3 on 1 and 49 DF, p-value: < 2.2e-16

So, the c and b are determined, which means we have the necessary coefficients to build the model. That is where the last equation will come in. The model is building on the already present data, predicting 150 days in the future. Fig. 9 shows the output of the SIR model along with the data.

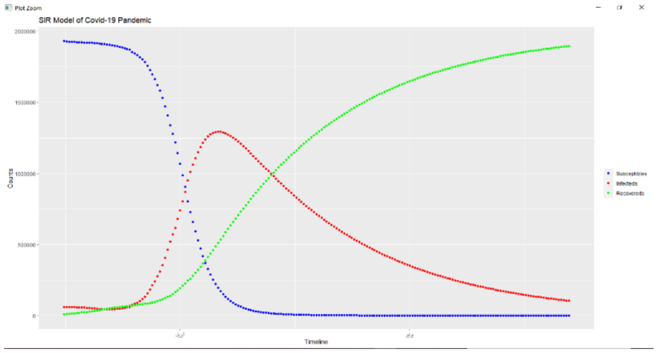


Fig. 9. SIR Modle of COVID-19.

The result below shows the model predictions of five days in the past and five days in the future:

```
Dates Susceptible Infected Recovered

50 2020-04-04 802595.0 951253 246152.0

51 2020-04-05 727885.0 1012103 260012.0

52 2020-04-06 658128.7 1062447 279424.2

53 2020-04-07 591920.1 1108278 299802.0

54 2020-04-08 529803.4 1149138 321058.9

55 2020-04-09 472155.6 1184745 343099.5

56 2020-04-10 419188.5 1214989 365823.0

57 2020-04-11 370962.9 1239911 389126.6

58 2020-04-12 327410.0 1259682 412908.1

59 2020-04-13 288357.5 1274574 437069.0

60 2020-04-14 253556.4 1284928 461515.4
```

In Table III, a comparison between actual data and the data from the SIR model application, which proved the efficiency of the model as the numbers are very close to each other, while some numbers constitute a perfect match. It is important to note that the recovery rate tends to be reliable in the short term, at least. In contrast, the infected rate causes underestimation when this model previously tested. With more data available due to disease progression, the predictions will get better. The results are generally reassuring that the epidemic, according to the model, will soon be limited.

TABLE III. COMPARING PREDICTED INFECTIONS AND RECOVERY NUMBERS WITH THE ACTUAL NUMBERS

	Dates	Predicted_Infecteds	Predicted_Recovereds	Actual_Infecteds	Actual_Recovereds
1	2020-04-05	1012103	260012.0	1012103	260012
2	2020-04-06	1062447	279424.2	1068586	276515
3	2020-04-07	1108278	299802.0	1126042	300054

V. CONCLUSION

Coronavirus has become a concern of many recently, and the focus on it became intense due to the rapid spread and lethality of people. In this paper, the SIR model for predicting 150 days in the future was applied, and the numbers compared with the real numbers, which proved the efficiency of the model. As the numbers are very close to each other, while some numbers constitute a perfect match. The results are generally reassuring that the epidemic, according to the model, will soon be limited.

ACKNOWLEDGMENT

The researcher is master student of big data course and wants to thank instructor Dr. Shakir Khan for his valuable support in this research.

REFERENCES

[1] WHO. Novel coronavirus – China. Jan 12, 2020. http://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/ (Jan 12, 2020).

[2] WHO. Novel coronavirus – Thailand (ex-China). Jan 14, 2020. http://www.who.int/csr/don/14-january-2020-novel-coronavirusthailand /en/ (Jan 14 ,2020).

[3] Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China [published correction appears in Lancet. 2020 Jan 30;]. *Lancet*. 2020;395(10223):497-506. doi:10.1016/S0140-6736(20)30183-5

[4] WHO. Novel coronavirus – https://www.who.int/health-topics/coronavirus#tab=tab_1 (2020).

[5] S. Nashreen, and N. Sharma, "Statistical Models for Predicting Swine F1u Incidences in India." 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC). IEEE, 2018.

[6] S. Eubank, V.S. Kumar, M. V. Marathe, A. Srinivasan, N. Wang, "Structural and algorithmic aspects of massive social networks." Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms.718-727, 2004

[7] J.Wang, J. Xiong, K. Yang, S. Peng, Q. Xu, "Use of GIS and Agent-Based Modeling to Simulate the Spread of Influenza," 2010 18th International Conference on Geoinformatics. IEEE, pp. 1-6, 2010

[8] R. M. Anderson, R. M, "Infectious Diseases of Humans", Oxford Science Publications, May 1992.

[9] N. G. Becker, L. R. Egerton, "A Transmission Model of HIV", Mathematical Biosciences, vol. 119, pp. 205-224, 1994.

[10] J. D. Murray, "Mathematical Biology Second Corrected Edition", Berlin: Springer- Verlag, 1993.

[11] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Mag., vol. 17, no. 3, pp. 37, 1996.

[12] M. Campion et al., "Predicting West Nile Virus (WNV) occurrences in North Dakota using data mining techniques," 2016 Future Technologies Conference (FTC), San Francisco, CA, pp. 310-317, doi: 10.1109/FTC.2016.7821628, 2016.

[13] World Health Organization (WHO), "Coronavirus disease 2019 (COVID-19) Situation Report - 35," WHO, 2020.

[14] World Health Organization (WHO), "Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCov)," WHO, 2020.

[15] World Health Organization (WHO), " Coronavirus disease 2019 (COVID-19): Situation Report - 75 (4 April 2020)", WHO, 2020.

[16] D. Hui et al, "The continuing 2019-nCoV epidemic threat of novel coronavirus to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China," International Journal of Infectious Diseases, vol. 91, pp. 264-266, 2020.

[17] A. Rachah and D. F. M. Torres, "Analysis, simulation and optimal control of a SEIR model for Ebola virus with demographic effects," Commun. Fac. Sac. Univ. Ank. Series A1, vol. 67, no. 1, pp. 179-197, 2018.

[18] M. Kuhn, K. Johnson, Applied Predictive Modeling, New York City: Springer New York Heidelberg Dordrecht London, 2013.

[19] Campion, Mitch, et al. "Predicting West Nile Virus (WNV) occurrences in North Dakota using data mining techniques." 2016 Future Technologies Conference (FTC). IEEE, (pp. 310-317), 2016.

[20] R. Gentleman, R. Ihaka, "R: The R Project for Statistical Computing", 2016, https://www.r-project.org/.

[21] N. Guizani, and A. Ghafoor. . Modeling and evaluation of disease spread behaviors. In 2014 International Wireless Communications and Mobile Computing Conference (IWCMC) (pp. 996-1003). IEEE. 2014.

[22] Y. Maki, & H. Hirose, "Infectious disease spread analysis using stochastic differential equations for SIR model". In 2013 4th International Conference on Intelligent Systems, Modelling and Simulation (pp. 152-156). IEEE. 2013.

[23] B. Norman, "The mathematical theory of infectious diseases and its applications. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE. 1975.

[24] R. M. Anderson, and M. M. Robert," Infectious diseases of humans: dynamics and control". Oxford university press, 1992.

[25] K. Matt, and P. Rohani, "Modeling infectious diseases in humans and animals". Princeton University Press, 2011.

[26] K. Sattar, T. Ahmad, Abdulghani HM, Khan S, John J, Meo S. Social networking in medical schools: medical student's viewpoint. Biomed Res 2016; 27: 1378-84

[27] H. Hans, et al. "Modeling infectious disease dynamics in the complex landscape of global health." Science 347.6227: aaa4339, 2015.

[28] D.J. Daley, D.G. Kendall, "Epidemics and rumours. Nature", 204(4963), 1118-1118. 1964.

[29] B.K Mishra, D Saini. "Mathematical models on computer viruses". Applied Mathematics and Computation, 187(2), 929-936. 2007

[30] J. Woo, H. Chen. Epidemic model for information diffusion in web forums: experiments in marketing exchange and political dialog. Springer Plus, 5(1), 66. 2016

[31] S. Shive, “An epidemic model of investor behavior,” Journal of Financial and Quantitative Analysis, 45(1), 169-198. 2010.

[32] S. Khan. "How Data Mining Can Help Curb City Crime", International Journal of Control Theory and Applications (IJCTA) 9 (23), 483-488, 2016

[33] M.F. AlAjmi, & S. Khan. "Data Mining–Based, Service-Oriented Architecture (SOA) in e-Learning". In ICERI2012 Proceedings pp. 3023-3023, IATED. 2012

[34] S. Khan, M. F. AlAjmi. "Impact of Medical Technology on Expansion in Healthcare Expenses", International Journal of Advanced Computer Science & Applications 4 (4), pp. 150-152. 2013.