Modified Machine Learning Techique for Curve Fitting on Regression Models for COVID-19 projections

Andreou Andreas
Department of Computer Science Mobile Systems Laboratory (MoSys Lab)
University of Nicosia and University of Nicosia Research Foundation
Nicosia, Cyprus
andreou.andreas@unic.ac.cy

Constandinos X. Mavromoustakis
Department of Computer Science Mobile Systems Laboratory (MoSys Lab)
University of Nicosia and University of Nicosia Research Foundation
Nicosia, Cyprus
mavromoustakis.c@unic.ac.cy

George Mastorakis
Department of Management Science and Technology
Hellenic Mediterranean University
Agios Nikolaos, Crete, Greece
gmastorakis@hmu.gr

Shahid Mumtaz
Instituto de Telecomunicações
Aveiro, Portugal
smumtaz@av.it.pt

Jordi Mongay Batalla
Warsaw University of Technology and National Institute of Telecommunications
Warsaw, Poland
jordi.mongay.batalla@pw.edu.pl

Evangelos Pallis
Department of Electrical and Computer Engineering
Hellenic Mediterranean University
Heraklion, Crete, Greece
pallis@hmu.gr

***Abstract*—COrona VIrus Disease 2019 (COVID-19) is a disease caused by Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2) and was first diagnosed in China in December, 2019. Dr. Tedros Adhanom Ghebreyesus, World Health Organization (WHO) director-general on March 11th declared the COVID-19 pandemic. The cumulative cases of infected individuals and deaths due to COVID-19 develop a graph that could be interpreted by an exponential function. Mathematical models are therefore fundamental to understanding the evolution of the pandemic. Applying machine learning prediction methods in conjunction with cloud computing to such models will be beneficial in designing effective control strategies for the current or future spread of infectious diseases. Initially, we compare the trendlines of the following three models: linear, exponential and polynomial using R-squared, to determine which model best interprets the prevailing data sets of cumulative infectious cases and cumulative deaths due to COVID-19 disease. We propose the development of an improved mathematical forecasting framework based on machine learning and the cloud computing system with data from a real-time cloud data repository. Our goal is to predict the progress of the curve as accurately as possible in order to understand the spread of the virus from an early stage so that strategies and policies can be implemented.**

Keywords—covid19, coronavirus, machine learning, cloud computing, regression, forecast, epidemic, pandemic, curve fitting

I. INTRODUCTION

Since 31st of December 2019 COVID-19 also known as Coronavirus, first reported in Wuhan, Hubei Province, China, has infected people from 216 countries, and territories worldwide to date [1]. Currently, about five months later, the World Health Organization (WHO) reports that 6,931,000 people have been infected and 400,857 deaths have been reported by COVID-19. Due to the inverse ratio of the large number of infected individuals and the short period of time, cumulative infectious cases are mainly interpreted by exponential functions due to the rapidly increasing values. Mortality occurs mainly in infected patients with a weakened immune system, i.e. to elderly and patients with chronic diseases. The lack of medication to treat the virus has led governments to policies that require citizens to be quarantined and to maintain spatial distance in order to limit the spread as much as possible. The development of innovative solutions would be beneficial in guiding them to policies that would prevent the risk of spreading at an early stage. Thus, the implementation of machine learning and cloud computing could be effective in predicting where and when the disease will spread or be eradicated in order to alert these communities to take appropriate actions.

Mathematical modelling of cumulative infectious cases with linear, exponential and polynomial regression models via Microsoft Excel shows that the 6th degree polynomial interprets the data more accurately, as presented by the R-square evaluation. We developed the 6th degree regression polynomial coefficients using least squares method. In order to minimize the variance between the estimated values from the polynomial function and the actual values from the data set. Furtherance of polynomial's curve can predict the expected progress of the actual curve. Also, from the roots of the 2nd derivatives we will have the opportunity to calculate the turning points which are also known as points of inflection, as these points are milestones in pandemic situations.

© IEEE 2020. This article is free to access and download, along with rights for full text and data mining, re-use and analysis.

Based on 6th degree polynomial model and the daily reporting from affected countries through an open access data repository entitle "Our World in Data" [2], we propose an innovative forecasting model, which will use machine learning and cloud computing to predict the course of the virus. The cutting-edge system will be updated daily and will recalculate the future course of the curve. The goal is to have a real-time updated curve for more accurate and early forecasts.

II. Literature Review

This section will highlight some of the related work on pandemic control approaches. Now more than ever, we need to find ways to balance the exponential curve of cumulative new cases of COVID-19 and pave the way for controlling future epidemics or pandemics. Liang Fang and Zhi Dong Cao et al. [3] have designed and developed an online real-time system using the technology of ArcGIS and Mashup to collect and display information about new outbreaks according to geographic location, time, and infectious agent. C. X. Mavromoustakis and G. Mastorakis et al. proposed a novel offloading methodology that hosts a "resource-aware" recommendation scheme, which allows the efficient monitoring of energy draining applications that run in an IoT ecosystem [4]. Various models have been developed to predict the stability and remediation of MERS-CoV infection by Isra Al-Turaiki et al. [5] using Naive Bayes and J48 decision tree classification algorithms. Zhaoyang Zhang et al. [6] proposed an epidemiological control based on clusters through social networking sites that could collect both vital points and social interaction. The spread of the virus could be effectively reduced due to real-time social contact and health information that could determine the optimal number or set of nodes to be removed. Sareen S. Sood, S.K. and Gupta, S.K. et al. [7] proposed a cloud-based architecture for predicting and preventing Ebola using Temporal Network Analysis (TNA) and wearable body sensor technology. Zika virus (ZikaV) is currently one of the most important emerging viruses in the world that have caused outbreaks and epidemics and has also been associated with severe clinical manifestations and congenital malformations. Sanjay Sareen et al. [8] introduced a cloud-based system for detecting and controlling the ZikaV using a mobile device. The authors in the article [9] presented and identified the different ways to implement the edge computing paradigm, by using M2M communications in dense networked systems via social connectivity from two different perspectives: the offered reliability for delay-tolerant (delay-sensitive) services and the energy conservation over reliability provision. Both perspectives introduce significant application execution optimization when using delay-sensitive data.

An overview of the current state of technology solutions shows that the development of machine learning, cloud computing and artificial intelligence can contribute to tackle the COVID-19 pandemic. Alibaba Cloud 2020 deployed machine learning and deep learning to establish a modified Susceptible - Exposed - Infectious - Recovered (SEIR) model to predict the prevalence of COVID-19 and assess the increased risk of contamination in a specific area [10]. This innovative solution can provide information on the course of the COVID-19 pandemic with 98% accuracy by submitting primary data such as flight information, number of new infectious cases, number of close contacts and number of quarantined people. A powerful biomedical tool that could also contribute to the fight is the genomic sequence of machine modeling to predict possible virus responses to different drugs or to monitor the spread of COVID-19 [11]. An artificial intelligence system has been developed for the automatic evaluation of computed tomography images that detects the characteristics of COVID-19 pneumonia to monitor infected patients [12].

III. Regression models and Performance comparison

Applying regression analysis, which is a prediction technique through mathematical modelling, allows us to examine the relationship between the dependent variable (y) and the independent variable (x). Linear, exponential and 6th degree polynomial regressions have been used to develop interpretive models as shown in Fig. 1, Fig. 2 and Fig. 3 respectively, for the cumulative infectious cases according to daily reports, beginning January 22, 2020.

To determine the coefficients of linear, exponential and 6th degree polynomial regression, we applied the method of least squares and we concluded to the following functions, respectively:

$$y = 52205x - 2 * 10^6 \tag{1}$$

$$y = 9{,}784.261e^{0.055x} \tag{2}$$

$$\begin{aligned} y = {} & 0.00003x^6 - 0.011161x^5 \\ & + 1.478908x^4 \\ & - 78.482575x^3 \\ & + 1{,}720.083148x^2 \\ & - 12{,}628.242181x \\ & + 51000 \end{aligned} \tag{3}$$

The variable (y) (vertical axis) represents the cumulative infectious cases from COVID-19 worldwide according to the daily measurements reported by the variable (x) (horizontal axis). Linear (1), exponential (2) and 6th degree polynomial (3) functions are graphically represented by the dashed lines in Fig. 1, Fig. 2 and Fig. 3 respectively. All models interpret 105 data days worldwide as of January 22, 2020 and predict a 30-day perspective as shown by the extended dashed line in the figures.

The coefficient of determination commonly known as R-squared (or R^2) is defined as the percentage of variance for the dependent variable (y) explained by the independent variable (x) in the regression models. When the values range from 0 to 1 it means that they range from 0% to 100% of the variation in the vertical axis and depends on the values of the horizontal axis.

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2}\right] \quad (4)$$

The second term of (4) compares the unexplained variance with the total variance. TABLE I presents the values of R-square for the evaluation of the three regression models.

TABLE I. R-SQUARED VALUES

Regression Models	R-Square
Linear	$R^2 = 0.8609$
Exponential	$R^2 = 0.922$
6th degree Polynomial	$R^2 = 0.999336$

Comparison of the regression model with the coefficient of determination R-squared as shown in TABLE I leads to the conclusion that the 6th degree polynomial regression fits the data better as the value is closer to 1.

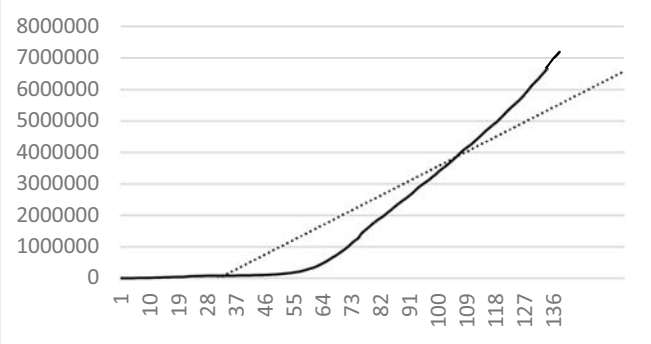

Fig. 1. Linear Regression Model

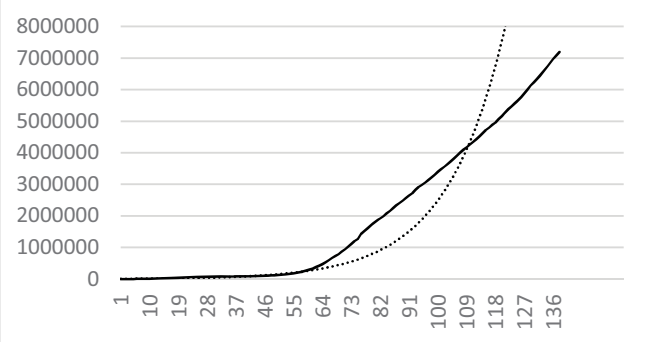

Fig. 2. Exponential Regression Model

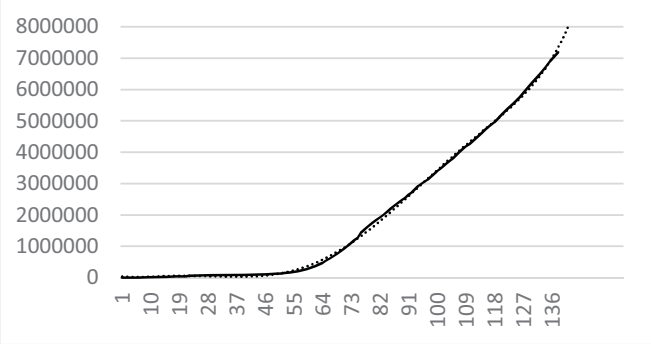

Fig. 3. Polynomial Regression Model

IV. Concavity and Points of Inflection

The inflection point is the point where the curve changes cavity. Due to the symmetry that appears at this point based on the sigmoid function from the cumulative infectious cases the turning point can be used to study the course of the pandemic. If we have $F''(x) > 0$ then the curve turns the concaves upwards while if we have $F''(x) < 0$ the curve turns the concaves downwards. Therefore, if we first match the 6th degree polynomial function (5) that best interprets the data, we can easily calculate the second derivative and by substituting its root into the original function we will determine the coordinate of the inflection point.

Initially we apply the data set of the cumulative infectious cases of a country, for example Italy. As presented in Fig. 4 we calculated the trendline with the coefficient of determination $R^2 = 0.9995$, which mean a high interpretation accuracy as it is very close to 1. Thereafter, we determined the function of the trendline and applied the second derivative as shown below:

$$F(x) = -8 * 10^{-6}x^6 + 0.0026x^5 - 0.2901x^4 + 13.321x^3 - 162.55x^2 + 408.28x + 914.82 \quad (5)$$

$$F''(x) = -0.00024x^4 + 0.052x^3 - 3.4812x^2 + 79.926x - 325.1 \quad (6)$$

$$F''(x) = 0 \Leftrightarrow x_1 \approx 5.12668, x_2 \approx 41.6206, x_3 \approx 55.4674$$

$$F(x_1) = 339, F(x_2) = 109362, F(x_3) = 182844$$

TABLE II: CONCAVITY

x		x_1		x_2		x_3	
$F''(x)$	-		+		-		+
$F(x)$	⌒		⌣		⌒		⌣

Turning points could benefit the accuracy of data interpretation and the furtherance of the curve. Concavity change is important because it shows the progression of the

curve at the next turning point and how infectious cases will develop.

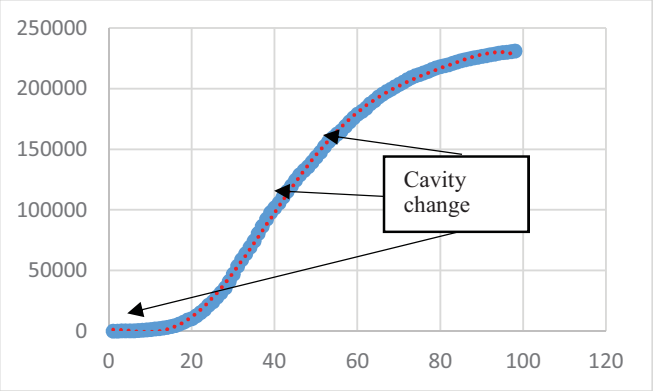

Fig. 4. Italy's cumulative infectious cases

V. Proposed Cloud Framework

To predict the growth and the trend of COVID-19 pandemic, we propose a cloud framework developed with machine learning models based on the regression models that we already analysed in section III. Regression models can contribute by forecasting the spread of the virus, as well as any other relevant parameter that could be beneficial to governments in aligning their policy strategies against the invisible enemy. Since 22nd of January 2020 government hospitals, private health-centres and laboratories around the world report daily positively infected cases, cumulative mortality cases due to COVID-19, number of tests performed, total recovered cases and serious critical cases in a cloud-based environment and the data are stored in an open data repository. Our proposal is to integrate machine learning through cloud computing to evaluate this data and accurately predict in real time the trend of the virus between countries. The aim is to inform governments with a warning about the progress of the virus, as well as possible critical cases that may arise.

Fig. 6 is a visual representation of the proposed system model where the raw data from laboratories and hospitals worldwide are stored in an open data repository. The data will then be driven and analysed using the Levenberg–Marquardt algorithm, which is a machine learning technique capable of interpreting data accurately [13]. The algorithm of Levenberg [1944] and Marquardt [1963] which was a modification of the Gauss-Newton method, gives the solution to the problem of the least quadratic determination for nonlinear coefficients of equations. As the proposed polynomial in section III operates by substituting the coefficients determined by the minimum value of (7).

$$F(x) = \frac{1}{2}\sum_{i=1}^{m}[f_i(x)]^2 = \frac{1}{2}\|f(x)\| \qquad (7)$$

The infinite increase of the graph, as well as the outliers and the noise of the data gave us the incentive to develop an iterative weighting strategy in order to normalize the graph of the curve and reduce the error. Initially, the progression of the infinite graph will flatten over time and then we will reconstruct the regression model to achieve a better curve fit by reducing the data distances from the curve. Composition of the SoftMax function [14] and the function which is the difference between the distances of all values along y axis from the curves of Sigmoid function, Arc-tangent function and Hyperbolic-tangent function respectively develop three weights as follow:

$$\dot{w}_i^{n+1} = \frac{e^{\left[1-\frac{d_i^n-\left(1+e^{-d_i^n}\right)^{-1}}{\max_i d_i^n-\left(1+e^{-d_i^n}\right)^{-1}}\right]}}{\sum_i e^{\left[1-\frac{d_i^n-\left(1+e^{-d_i^n}\right)^{-1}}{\max_i d_i^n-\left(1+e^{-d_i^n}\right)^{-1}}\right]}} \qquad (8)$$

$$\ddot{w}_i^{n+1} = \frac{e^{\left[1-\frac{d_i^n-\left(\frac{2}{\pi}\right)\tanh^{-1} d_i^n}{\max_i d_i^n-\left(\frac{2}{\pi}\right)\tanh^{-1} d_i^n}\right]}}{\sum_i e^{\left[1-\frac{d_i^n-\left(\frac{2}{\pi}\right)\tanh^{-1} d_i^n}{\max_i d_i^n-\left(\frac{2}{\pi}\right)\tanh^{-1} d_i^n}\right]}} \qquad (9)$$

$$\ddot{w}_i^{n+1} = \frac{e^{\left[1-\frac{d_i^n-\tanh d_i^n}{\max_i d_i^n-\tanh d_i^n}\right]}}{\sum_i e^{\left[1-\frac{d_i^n-\tanh d_i^n}{\max_i d_i^n-\tanh d_i^n}\right]}} \qquad (10)$$

The differences between the distances of all the coordinates along the y-axis from the curves are divided by the maximum value and subtracted from 1, the SoftMax function standardized the results corresponding to each coordinate. First, we enter three unary weights for all data points to fit the curve from the Levenberg–Marquardt algorithm. Thereafter, we substitute the weight calculated by the equations (8), (9) and (10) corresponding to each point, for the next iteration. Finally, we apply the Levenberg–Marquardt algorithm with the new weights and evaluate the adjustment of the curve by the three different methods. The sum of the deviation of all weights should be lower than a threshold value for the convergence of the algorithm.

In addition, we developed a 10th degree polynomial in MATLAB for curve fitting corresponding to the cumulative infectious cases of Italy. Next, we integrated the findings with python encoding in order to reduce the error. Our findings are shown in Fig. 5 where we have $R^2 = 0.999934933$.

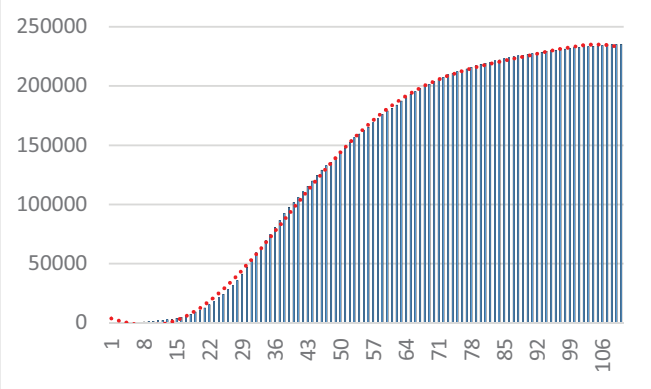

Fig. 5. Italy's Curve Fitting output from Python Program

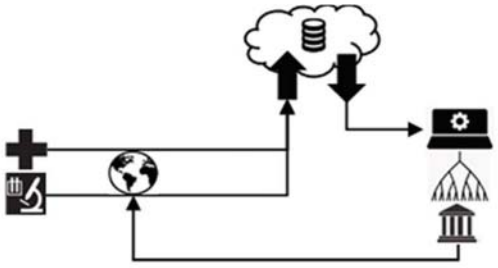

Fig. 6. Proposed Model

VI. Distribution Fitting

To determine the most appropriate distribution model for the data corresponding to daily new COVID-19 cases, we used datasets from countries where the new cases show a declining course of the curve. We determined the distributions with the best performance for each case using version 5.6 of EasyFit Standard. Fig. 7 and Fig. 8 presents the distribution Johnsons SB, as the best performance of goodness of fit for Italy and Spain respectively, as assessed by the Kolmogorov Smirnov and Anderson Darling tests. The Johnsons SB distribution shows the best fitting performance compared to other distributions in two countries, Italy (Fig. 7) and Spain (Fig. 8). From the weighted iterative approach in section V, we conclude that the distributions fit the curve better than without the weighted iterations.

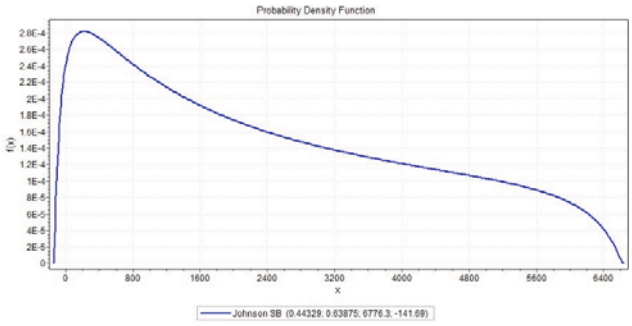

Fig. 7. Italy

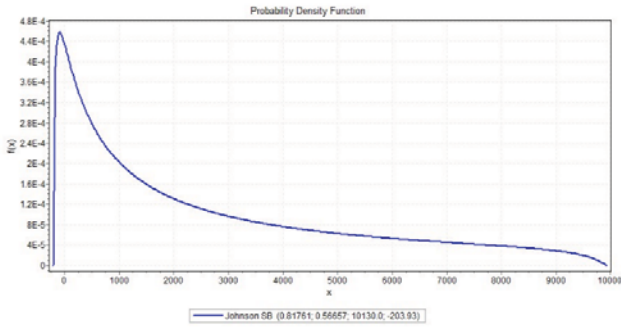

Fig. 8. Spain

Algorithm

Requirements:

x: Input sequence of days from first reported case

y: Input number of cases corresponding to each day in x

t: Threshold parameter (earliest time a failure may occur)

Process:

$w_0 \leftarrow 1 * x$

for iteration n from 0, step 1 **do**

$f \leftarrow$ Levenberg Marquardt (input: x, y, w^n)

$d_i \leftarrow |f(x_i) - y_i|, \forall i \in \mathbb{N}$

Apply one of the following:

$w_i^{n+1} \leftarrow (8)$

$w_i^{n+1} \leftarrow (9)$

$w_i^{n+1} \leftarrow (10)$

if $\sum_i |w_i^n - w_i^{n+1}| < t$ **then**

break

end for

end procedure

VII. Discussion & Concussion

This paper proposes a modified machine learning technique for interpreting data provided by countries around the world regarding infectious cases of COVID-19 as accurately as possible. The aim is to anticipate the course of the spread and to inform governments at an early stage to formulate their policies. Forecast models that scientific teams among countries use to predict the progress of the virus could lead to contradiction because a premature uplifting of the lockdown, may cause adverse effect on the management of the pandemic situation. Thus, the goal of our model is the implementation of a technique that best fits regression models to the actual data curve. Daily updated data will be driven to construct a new polynomial function for error reduction. Due to the updated polynomial function, the resulting turning points will also be updated to contribute with the prediction of curve's furtherance. As presented in Fig. 12 quarantine measures began on March 23 in most countries but the new infectious cases continue to show an increasing trend. Therefore, the early lockdown strategy could lead to a reduction in new cases.

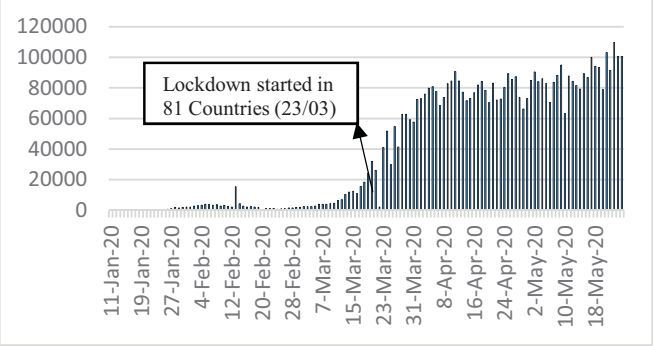

Fig. 12. Confirmed Cases

ACKNOWLEDGMENT

This research work was undertaken under the project SHAPES – Smart and Health Ageing through People Engaging in Supportive Systems – is funded by the Horizon 2020 Framework Programme of the European Union for Research Innovation. Grant agreement number: 857159 - SHAPES – H2020 – SC1-FA-DTS – 2018-2020. Additionally, the work presented in this article was partially funded by the Ambient Assisted Living (AAL) project vINCI: "Clinically-validated INtegrated Support for Assistive Care and Lifestyle Improvement: The Human Link", funded by the Research and Innovation Foundation in Cyprus (under the AAL framework with Grant Nr. vINCI /P2P/AAL/0217/0016) and the Polish National Centre for Research and Development in Poland."

REFERENCES

[1] I. Ghinai, T. McPherson, J. Hunter, H. Kirking, D. Christiansen, K. Joshi, R. Rubin, S. Morales-Estrada, S. Black, M. Pacilli, M. Fricchione, R. Chugh, K. Walblay and S. Ahmed, "First known person-to-person transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the USA," *Lancet,* vol. 395, no. 10230, pp. 1137-1144, 04 Apr. 2020.

[2] H. Ritchie, "Our World in Data," University of Oxford, [Online]. Available: https://ourworldindata.org/coronavirus-source-data. [Accessed 2005].

[3] L. Fang and Z. Cao, "An Online Real-Time System to Detect Risk for Infectious Diseases and Provide Early Alert," in *Lecture Notes in Computer Science*, vol. 6749, Berlin, Springer, 2011, pp. 101-107.

[4] C. Mavromoustakis, G. Mastorakis, J. M. Batalla, J. Rodrigues and J. Sahalos, "Edge Computing for Offload-Aware Energy Conservation Using M2M Recommendation Mechanisms," in *2019 IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, 2019.

[5] I. A. Turaiki, M. Alshahrani and T. Almutairi, "Building predictive models for MERS-CoVinfections using data mining techniques," *Journal of Infection and Public Health,* vol. 9, pp. 744-748, Sep. 2016.

[6] Z. Zhang, H. Wang, C. Wang and H. Fang, "Cluster-based Epidemic Control Through Smartphone-based Body Area Networks," *IEEE Trans Parallel Distrib Syst.,* vol. 26, no. 3, pp. 681-690, 02 Mar. 2015.

[7] S. Sareen, S. Sood and S. K. Gupta, "IoT-based cloud framework to control Ebola virus outbreak," *Journal of Ambient Intelligence and Humanized Computing,* vol. 9, pp. 459-476, 20 Oct. 2016.

[8] S. Sareen, S. Sood and S. K. Gupta, "Secure Internet of Things based Cloud Framework to control ZIKA virus outbreak," *International Journal of Technology Assessment in Health Care,* vol. 33, no. 1, pp. 11-18, 2017.

[9] C. Mavromoustakis, J. M. Batalla, G. Mastorakis, E. Markakis and E. Pallis, "Socially Oriented Edge Computing for Energy Awareness in IoT Architectures," *IEEE Communications Magazine,* vol. 56, no. 7, pp. 139-145, Jul. 2018.

[10] "Alibaba Cloud Helps Fight COVID-19 Through Technology," Alibaba Cloud, 2020.

[11] K. Kupferschmidt, "Genome analyses help track coronavirus' moves," *Science,* vol. 367, no. 6483, pp. 1176-1177, 13 Mar. 2020.

[12] S. Jin, B. Wang, H. Xu, C. Luo, L. Wei, W. Zhao, X. Hou, W. Ma, Z. Xu, Z. Zheng, W. Sun, L. Lan, W. Zhang, X. Mu, C. Shi, Z. Wang, J. Lee, Z. Jin, M. Lin, H. Jin and L. Zhang, "AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system in four weeks," *medRxiv,* 23 Mar. 2020.

[13] J. Moré, "The Levenberg-Marquardt algorithm: Implementation and theory," *Numerical Analysis. Lecture Notes in Mathematics,* vol. 630, pp. 105-116, 27 Aug. 2006.

[14] S. Horiguchi, D. Ikami and K. Aizawa, "Significance of Softmax-Based Features in Comparison to Distance Metric Learning-Based Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 42, no. 5, pp. 1279-1285, 05 May 2020.