Predicting COVID-19 Spread Level using Socio-Economic Indicators and Machine Learning Techniques

Alaeddine Mihoub, Hosni Snoun, Moez Krichen, Montassar Kahia, Riadh Bel Hadj Salah

Predicting COVID-19 Spread Level using Socio-Economic Indicators and Machine Learning Techniques

Alaeddine Mihoub
Department of Management Information System and Production Management
College of Business and Economics, Qassim University
P.O. Box: 6640, Buraidah: 51452, Saudi Arabia
a.mihoub@qu.edu.sa

Hosni Snoun
Departement of Civil Engineering, Modeling in Hydraulics & Environment Labratory
National Engineering School of Tunis
Campus universitaire, BP 37, 1002, Le Bélvédère, 1002 Tunis, Tunisia
hosni.snoun@enit.utm.tn

Moez Krichen
FCSIT, Albaha University, Albaha, Saudi Arabia
ReDCAD Laboratory, University of Sfax, Tunisia
moez.krichen@redcad.org

Riadh Bel Hadj Salah
Department of Digital Transformation STC Solutions
Al Malaz, Riyadh 12641, Saudi Arabia
riadh.bel.hadj2004@gmail.com

Montassar Kahia
Department of Finance & Economics, College of Business and Economics, Qassim University
LAREQUAD & FSEGT, University of Tunis El Manar, Tunisia.
P.O. Box: 6640, Buraidah: 51452, Saudi Arabia
m.kahia@qu.edu.sa

***Abstract*—The new so-called COVID-19 virus is unfortunately founded to be highly transmissible across the globe. In this study, we propose a novel approach for estimating the spread level of the virus for each country for three different dates between April and May 2020. Unlike previous studies, this investigation does not process any historical data of spread but rather relies on the socio-economic indicators of each country. Actually, more than 1000 socio-economic indicators and more than 190 countries were processed in this study. Concretely, data preprocessing techniques and feature selection approaches were applied to extract relevant indicators for the classification process. Countries around the globe were assigned to 4 classes of spread. To find the class level of each country, many classifiers were proposed based especially on Support Vectors Machines (SVM), Multi-Layer Perceptrons (MLP) and Random Forests (RF). Obtained results show the relevance of our approach since many classifiers succeeded in capturing the spread level, especially the RF classifier, with an F-measure equal to 93.85% for April 15th, 2020. Moreover, a feature importance study is conducted to deduce the best indicators to build robust spread level classifiers. However, as pointed out in the discussion, classifiers may face some difficulties for future dates since the huge increase of cases and the lack of other relevant factors affecting this widespread.**

Keywords—covid-19, socio-economic indicators, data preprocessing, spread level prediction, machine learning, country classification, feature importance

I. INTRODUCTION

Considering the continuous COVID-19 pandemic growth across the globe, many worldwide researchers are attempting to estimate accurately its potential spread. According to the World Health Organization, the virus has caused over a quarter million confirmed deaths by Mai 27, 2020 [1]. In order to explain this disturbing spread, socio-economic indicators in each country may be explored to investigate this alarming evolution further. Coupled with machine learning and analytics techniques, these indicators may help in explaining some aspects of the coronavirus crisis around the globe. In fact, a socio-economic policy of a specific country constitutes an interesting source of information and gives relevant insights to predict the number of spread cases. In literature, most machine learning applications developed for spread prediction have tried to forecast national and international statistics concerning total cases, total deaths and total recoveries [2]–[5]. Their overall approach is to build prediction models, essentially based on previous spread data. Our approach in this work is to find a relevant relation between socio-economic indicators and the level of spread in each infected country. Our main contribution consists of predicting the level of spread by proposing different classification models based solely on more than 1000 socio-economic indicators and more than 190 countries. Furthermore, given developed classifiers, an importance study is proposed in order to determine the most influential indicators in the classification process. The rest of the paper is organized as follows: the next section reviews related work to spread prediction and impacting indicators. Our proposed approach is presented in Section III. Models implementations details are exposed in section IV. All results and related discussions are presented in section V. Section VI concludes the paper and gives some perspectives for future work. VIII

II. RELATED WORK

We divide this section into two parts. On the one hand, the first part is dedicated to works that have focused on predicting the spread of the COVID-19 epidemic. On the other hand, the second part focuses on the work concerned with the existing correlation between the spread of certain diseases/pandemics and the socio-economic indicators of different countries.

978-1-7281-7407-5/20/$31.00 ©2020 IEEE
DOI 10.1109/SMART-TECH49988.2020.00041

A. Studies about COVID-19 spread prediction

Following the onset of the COVID 19 epidemic, several researchers and scientists have taken an interest in studying the possibilities of the spread of this pandemic. These different studies used different techniques and focused on specific geographic areas. In what follows, we offer a brief overview of some of these studies. First of all, since the pandemic initially appeared in China, several research works [5]–[8] have concentrated on the study of its spread in this country of origin. First in [5], the authors presented a new prediction model called FPASSA-ANFIS, which is an extension of a neuro-fuzzy inference method for forecasting the number of confirmed COVID-19 cases over ten days based on previously reported cases observed in China. Second in [6], the authors proposed the so-called SUQC model to describe the COVID-19 dynamics and parameterize the intervention impacts of quarantine and control measures. Third in [7], the authors proposed a hybrid AI (Artificial-Intelligence) model which combines an ISI (Improved Susceptible–Infected), an NLP (Natural Language Processing) module and an LTSM (Long Short-Term Memory) network for COVID-19 prediction. Fourth in [8], the authors adopted a sub-epidemic wave model, the Richards growth model, and generalized logistic growth model for generating five-and ten-day ahead predictions of COVID-19 spread in two Chinese cities namely Guangdong and Zhejiang.

In what follows, we consider a collection of works which focused on the spread of the COVID-19 pandemic outside of China. In [9], a simple heuristic was proposed in order to identify the date at which the number of confirmed cases outside China will reach one million. The proposed heuristic consists of approximating the number of cases using an exponential curve. In [10], an econometric model based on the ARIMA (Auto Regressive Integrated Moving Average) model was proposed in order to predict the spread of COVID-19. In the study presented in [11], the authors proposed a multivariate prediction model to approximate pandemic trajectories in sixteen countries, from different continents and economic categories (High-income, Upper-middle income, Lower-middle income and Low-income) and with respect to different prevention scenarios. More precisely, the proposed analysis was based on an SEIR (Susceptible Exposed Infected Recovered) compartmental model.

In [12], data-driven prediction techniques such as curve fitting and LSTM (Long Short-Term Memory) were explored in order to approximate the number of COVID-19 cases in India 30 days ahead. The study presented in [13] allowed to use a segmented Poisson model to make a statistical forecasting about the attack rate, duration and turning point for COVID 19 for six Western countries (USA, UK, Italy, Germany, France and Canada). Similarly, the authors of [14] proposed an analysis for predicting the duration of the pandemic and the number of infections in eight western countries (USA, UK, France, Spain, Italy, Germany, the Netherlands, Greece) using a Gaussian hypothesis for propagation. In [15], the propagation of the pandemic in six African countries (Kenya, Senegal, Nigeria, Algeria, Egypt and South Africa) was simulated and estimated using the customized SEIR (Susceptible Exposed Infectious Recovered) Model and MH (Maximum-Hasting) parameter prediction technique under three intervention situations (mildness, mitigation and suppression).

It is worth mentioning that all the cited works in this subsection only considered as input for propagation prediction the previously identified and registered numbers of confirmed cases in different countries and none of them explicitly considered socio-economic indicators for achieving that purpose.

B. Studies about the correlation between the spread of diseases and socio-economic indicators.

As previously mentioned in the previous subsection, not many works in the literature studied the correlation between the spread of COVID-19 and socio-economic indicators. For this reason, we enlarge our perspective and consider works that studied this correlation for both COVID-19 and previous pandemics, which appeared in the modern history during the last few decades. For instance, the authors of [16] explained that their goal was to study the impact of social and economic factors on the propagation of COVID-19 in China. However, they were limited to healthcare measures, weather characteristics, geographic proximity and similarity in economic conditions and did not consider other socio-economic factors. Moreover, the previous study was restricted to the case of China. In another study [17], perspectives from behavioral economics were explored, focusing attention on how to encourage people to take part in preventive behaviors with COVID-19. In this previous work, the presented study was qualitative and was not based on any mathematical modeling or calculation.

In [18], the authors argue that existing pandemic transmission models generally do not take into account the specific nature of a society and location-specific parameters. For this reason, they attempted to identify the underlying spatial attributes which may influenced SARS transmission in Hong Kong in 2003. In another study [19], the authors analyzed the spatial-temporal mortality trends in Spain due to the 1918-1919 influenza pandemic in Spain. However, they did not take into account any other socio-economic indicators. In [20], the author studied the effects of economic activities and social interactions on infection spread. For this purpose, data describing the occurrence of three main viral infections were exploited. The data extends up to a quarter of a century through geographical areas in France, at a weekly pace.

In contrast with all these previous cited works, our goal in this work is to study the correlation between COVID-19 spread and as many socio-economic indicators as possible in as many countries as possible using artificial intelligence techniques [21].

III. APPROACH

A. Overview

The key idea of this paper is to estimate, at a specific date, the spread level of each country on the basis of multiple socio-economic indicators. Indeed, with regards to spread intensity, four levels of spread are proposed in our work:

- Level 1: Total number of confirmed cases is less than 1000 cases.
- Level 2: Total number of confirmed cases is between 1000 and 10000 cases.
- Level 3: Total number of confirmed cases is between 10000 and 50000 cases.
- Level 4: Total number of confirmed cases is above 50000 cases.

This way, each infected country around the globe will be assigned to one of these four levels. These proposed levels correspond to four output classes, as we will detail in the classifiers section. The ranges of levels 3 and 4 were set intentionally quite large in order to make the corresponding classes consistent in terms of number of instances. For instance, on May 1st, the number of countries belonging to level 3 was 24 and only 12 countries belong to level 4. More split on these classes makes them irrelevant for training and testing processes.

In our classification approach and unlike previous works, we rely uniquely on socio-economic indicators without any use of historical spread data. Actually, our methodology tries to map the socio-economic situation of a country to a potential level of COVID-19 spread. Concretely, almost 1429 socio-economic indicators were used to categorize each infected country around the globe. These indicators span over a wide range of socio-economic related field such as Economy and Growth, Demography, Agriculture & Rural Development, Energy, Environment, Education, Financial Sector, Public & Private Sector, Health, Infrastructure, Poverty, Science & Technology, Social Development, Social Protection, Trade, Urban Development, etc. Therefore, our philosophy is to find relevant models that map those multimodal indicators to the expected spread level.

To reach this goal, our approach can be split into three big steps: (1) Data Preprocessing (2) Feature Selection and (3) Classification and Evaluation (see Fig. 1). First, before any classification procedure, a data preprocessing step is required to denoise and prepare data. In this work, many preprocessing manipulations were undertaken, especially treating missing values, normalizing indicators values and changing the format of the target variable. More details about preprocessing are presented in the Implementation section. Afterward, the second step is feature selection. Feature selection consists of selecting the best features fitting the tackled problem. Concretely, it is the process of selecting the subset of indicators that contribute the most in predicting the spread level. Irrelevant indicators should be discarded in this step since it may even decrease the model accuracy. In this paper, feature selection was applied using the Univariate Feature Selection approach that will be shortly described in the next subsection. Once raw data are preprocessed and relevant indicators are selected, we can move toward the classification process. Three classifiers are proposed and tested in this research, which are Support Vectors Machines (a.k.a. SVM), Neural Networks (especially Multi-Layer Perceptrons a.k.a. MLP) and Random Forests (a.k.a. RF).

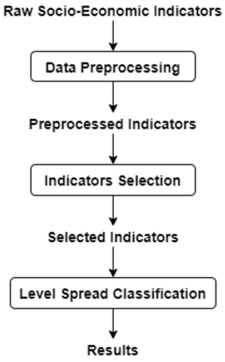

Fig. 1:Overview of level spread prediction using socio-economic indicators

B. Univariate Feature Selection

The idea of Univariate feature selection is to select the most relevant indicators on the basis of univariate statistical tests computed between each input indicator and the output variable. It allows us to keep indicators having the highest scores according to specific ANOVA tests [22]. Concretely, using our quantitative indicators and our target spread levels, we can compute the ANOVA F-value score between each indicator and the desired output classes. Hence, selecting the best k indicators is actually selecting the k indicators having the highest ANOVA F-value scores.

C. Classifiers

1) Support Vectors Machines

Support vector machines (SVM) represent a powerful technique widely used and successfully applied for treating classification and regression problems [23]. This method is based on two key principles formally combined by Vapnik [24] in 1995. The first principle is the maximum-margin hyperplane principle. The main idea here is to find the optimal hyperplane that separates the classes with the maximum margin. If data is linearly separable, this is a classic quadratic optimization problem. However, data is often linearly inseparable. The kernel function which represents the second key idea, gives the solution by transforming the initial data space to a higher-dimensional space, where it is likely to get a linear separator. This way, nonlinear classification problems are handled efficiently by the SVM concept.

2) Multi-Layer Perceptrons

Multi-Layer Perceptrons (MLP) are part of Artificial Neural Networks (ANN) and Deep Learning models [25]. ANNs are computational models that are inspired from the human brain and composed of several interconnected and successive layers. Each layer is composed of a set of artificial neurons called nodes. These nodes are connected to the next layer via links representing their impact on the next layer node. The first layer is called the input layer since it injects input data to the network. The intermediate layers are called hidden layers and the last one is the output layer. Basic topologies of ANNs are also known as Multi-Layer Perceptrons or Feed-Forward Networks.

3) Random Forests

Random Forests (RF) [26] is another widely applied technique for classification. It is part of the Ensemble Learning approaches [27] since it forms an ensemble of decision trees [28] and merges their multiple classification decisions according to a voting system. It represents an interesting approach since it maintains decision trees advantages and prevents their over-fitting issues. Compared to classic decision trees, RF results in a more stable and accurate classification output. In the next section, we present the implementation details of our approach.

IV. IMPLEMENTATION

A. Data Preprocessing

We remind that our goal is to predict the spread level of COVID-19 for each infected country. To this end, 1429 socio-economic indicators were used as raw data to our models. Indicators data were downloaded from the Word Bank official website[1]. Because of many missing indicators

[1] https://databank.worldbank.org/source/world-development-indicators#

for the last two years, we precise that used indicators concerned the year 2017. For COVID-19 spread data - especially the total number of cases to a specific date - three key dates are used: April 1st, April 15th and May 1st, 2020. For the spread data, they are easy to find and freely available on many websites[2]. All used data in this work can be downloaded using this link: "http://167.114.185.168:8800/". Once we get the indicators and the spread data of a certain date we merge the two databases with an inner join on the country code. For instance, for May 1st, this manipulation gives a dataset with shape (196, 1430): 196 countries on rows and 1430 columns (1429 indicators in addition to the total number of cases column).

First, a major issue with this dataset is missing values in some columns. To treat this problem, a missing value filter was applied to keep only the indicators having at least 120 non-missing values. Because of this mandatory operation, almost half of the indicators were deleted and the new shape of our dataset decreased to (196, 682). Second, to normalize data a min-max scaler is applied. This way, all indicators values are now rescaled and ranged between 0 and 1. We remind that this normalization operation is recommended (and even mandatory in some cases) for efficient classification. Lastly, the output variable i.e. the total number of cases was discretized in order to obtain a categorical output variable for the classification. This new output variable contains as previously described (refer to subsection III.A) four levels of spread. Table I gives, for instance, the number of countries for each level on May 1st, 2020:

TABLE I: OUTPUT VARIABLE (LEVEL OF SPREAD) DISTRIBUTION ON MAY 1ST , 2020

Level of spread	Number of countries
Level 1	*109*
Level 2	*51*
Level 3	*24*
Level 4	*12*

B. Indicators Selection and Classifiers

As mentioned before, three types of classifiers were applied in this paper, namely SVM, MLP and RF. Before presenting these classifiers results, we precise that: three big classification experiments were explored since we try to estimate the level of spread for three different dates which are April 1st, April 15th and May 1st, 2020. For each tested date, the corresponding dataset was split into a training set (67% of data) and a testing set (33% of data). Moreover, please note that before classification, we apply the univariate selection approach to keep only relevant indicators. From the preprocessed indicators (681 indicators, for instance, on May 1st), many numbers were tested, especially 20, 40, 60, 120, 200, 300 and 500.

Furthermore, some of the important parameters of classifiers are as follows: for the SVM, Radial Basis Function (RBF) kernel is used and a one-vs-one strategy is applied to deal with the multiclass problem. For MLP architecture, many topologies were explored. The optimal configuration was two hidden layers with a number of nodes equal to 1.5*number of selected indicators. For RF, the number of decision trees was 100 trees. Please note that many other hyper-parameters were finely tuned based on literature review and empirical results in order to optimize classification performance. For information, indicators preprocessing, feature selection, models training, models testing and evaluation metrics were all computed using Python Data Science packages, especially Numpy, Pandas, Matplotlib and Scikit-learn [29].

[2] https://opendatawatch.com/

V. RESULTS AND DISCUSSION

All the metrics presented in this section represent F-measures. F-measure represents the harmonic mean of recall and precision. Recall, precision and F-measure are reminded in the following formulas where:

- TP stands for True Positive: number of instances classified in a class and actually belonging to that class.
- FP for False Positive: number of instances classified in a class and actually not belonging to that class.
- FN for False Negative: number of instances belonging to a class but classified in another class.

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$F-measure = \frac{2*Recall*Precision}{Recall+Precision} \quad (3)$$

As mentioned in the implementation section, three dates are tested: April 1st, April 15th and May 1st, 2020. For each date, 3 classifiers are applied: SVM, MLP and RF. For each classifier many numbers of indicators were explored in the selection process especially: 20, 40, 60, 120, 200, 300 and 500. Best classifiers results and selection configurations are exposed bellow in Table II.

TABLE II: BEST CLASSIFIERS RESULTS AND SELECTION CONFIGURATIONS FOR ALL TESTED DATES.

Testing date	Best classifier	F-measure	Optimal number of selected indicators
April 1st	*SVM*	*89.06 %*	*40*
April 15th	*RF*	*93.85 %*	*60*
May 1st	*SVM, MLP*	*81.54 %*	*300*

As shown in Table II, for April 1st, the best classifier was the SVM with an F-measure equal to 89.06%. This optimal result is obtained with the optimal number of 40 indicators representing the best 40 indicators according to the univariate selection approach. This result is emphasized in Fig. 2 in which we observe increasing scores till 40 selected indicators and then decreasing scores when more indicators were selected (60 indicators and higher numbers). This is explained by the fact that adding irrelevant features may disturb the classification process.

The overall best result was 93.85% obtained by the Random Forest (RF) classifier for the tested date of April 15th. If we take a close look at the confusion matrix of this classification (see Fig. 3), we observe quite interesting performance even for classes having a few numbers of training and testing examples. For instance, the RF classifier has succeeded to detect 4 of 5 instances for class 3 (Level 3)

and the 3 instances of class 4 (Level 4). This result was computed using the best 60 indicators according to the univariate selection approach.

Since the RF classifier resulted to the highest overall score (93.85%), we tried to compute the importance of each indicator for this particular classifier (for date April 15th). By crossing univariate selection results and RF feature importance results, some indicators with contrast to others seem to have an impact on the classification process. Among these indicators, we find essentially those related to touristic activities in countries such as "International tourism, number of arrivals", "International tourism, receipts" and "International tourism, expenditures". Here, we would like to insist that our study was intended only to predict countries levels of spread by examining the socio-economic situation of each country. We do not claim any direct causation between the aforementioned indicators and the virus spread. Causality studies need further and deeper investigations. However, importance scores may serve as a means to explore and search informative variables that may be used to build robust predictors [30].

For May 1st, the best F-measure was 81.54% obtained by two classifiers: SVM and MLP. For this particular result, we observe that classifiers needed a large number of indicators to reach such a score (300 indicators). It may be explained by the unexpected high increase of cases since April 1st. In fact, one month after that date, the virus has widely spread throughout the world and the socio-economic indicators alone may not be sufficient to explain such high spread levels. We even expect that our results may decrease below 80% for future dates since other factors - not included in our study - should also have a significant effect on the spread such as human attitudes, political decisions, meteorological aspects, etc.

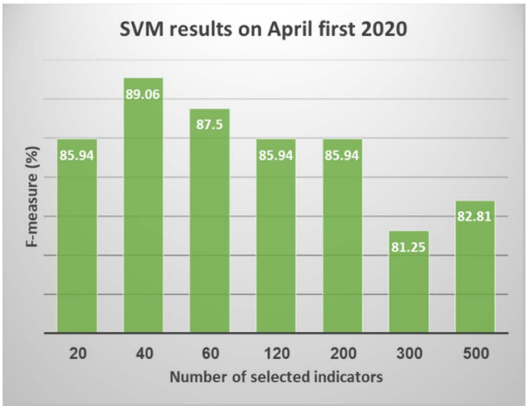

Fig. 2: SVM results (F-measure) for each selected number of indicators. The tested date is April 1st, 2020.

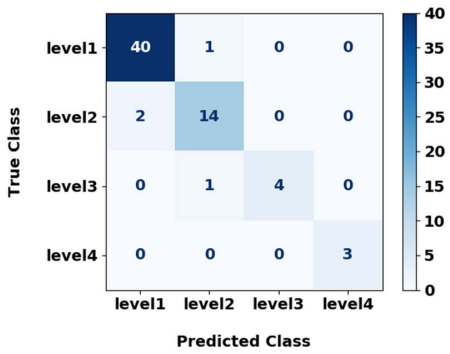

Fig. 3: Confusion matrix for RF classifier (classification date is April 15th, the number of selected indicators is 60 and F-measure is equal to 93.85%).

VI. Conclusion

This paper suggests innovative methods for measuring the spread level of the COVID-19 for each infected country around the world based on their socio-economic indicators through performing the best classifiers. To this end, more than 1000 socio-economic indicators and more than 190 countries were processed in this study. Moreover, to find the class level of each country for three different dates between April and May 2020, many classifiers were proposed based especially on Support Vectors Machines (SVM), Multi-Layer Perceptrons (MLP) and Random Forests (RF). The results provide evidence that our approach has promising performance since many classifiers succeeded in capturing the spread level, especially the RF classifier, with an F-measure equal to 93.85% for April 15th, 2020. For this specific date and classifier, a feature importance investigation is performed showing that indicators, especially those associated with touristic activities, seem to have an impact on the classification process. Other probable extensions of this paper may consist of developing a novel tool to overcome the difficulties faced by classifiers for future dates regarding the considerable increase of cases and the lack of additional relevant factors explaining this widespread pandemic.

References

[1] W. H. Organization, 'Coronavirus disease (COVID-19) outbreak situation', *World Health Organization. Available at https://www. who. int/emergencies/diseases/novel-coronavirus-2019. Accessed March*, vol. 23, 2020.

[2] J. Bullock, K. H. Pham, C. S. N. Lam, and M. Luengo-Oroz, 'Mapping the landscape of artificial intelligence applications against COVID-19', *arXiv preprint arXiv:2003.11336*, 2020.

[3] S. K. Bandyopadhyay and S. Dutta, 'Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release', *medRxiv*, 2020.

[4] C.-J. Huang, Y.-H. Chen, Y. Ma, and P.-H. Kuo, 'Multiple-input deep convolutional neural network model for covid-19 forecasting in china', *medRxiv*, 2020.

[5] M. A. Al-qaness, A. A. Ewees, H. Fan, and M. Abd El Aziz, 'Optimization method for forecasting confirmed cases of COVID-19 in China', *Journal of Clinical Medicine*, vol. 9, no. 3, p. 674, 2020.

[6] S. Zhao and H. Chen, 'Modeling the epidemic dynamics and control of COVID-19 outbreak in China', *Quant Biol*, vol. 8, no. 1, pp. 11–19, Mar. 2020, doi: 10.1007/s40484-020-0199-0.

[7] N. Zheng *et al.*, 'Predicting COVID-19 in China Using Hybrid AI Model', *IEEE Transactions on Cybernetics*, pp. 1–14, 2020, doi: 10.1109/TCYB.2020.2990162.

[8] K. Roosa *et al.*, 'Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China: February 13–23, 2020', *Journal of*

Clinical Medicine, vol. 9, no. 2, p. 596, Feb. 2020, doi: 10.3390/jcm9020596.

[9] W. W. Koczkodaj *et al.*, '1,000,000 cases of COVID-19 outside of China: The date predicted by a simple heuristic', *Global Epidemiology*, vol. 2, p. 100023, Nov. 2020, doi: 10.1016/j.gloepi.2020.100023.

[10] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, and M. Ciccozzi, 'Application of the ARIMA model on the COVID-2019 epidemic dataset', *Data in Brief*, vol. 29, p. 105340, Apr. 2020, doi: 10.1016/j.dib.2020.105340.

[11] The Global Dynamic Interventions Strategies for COVID-19 Collaborative Group *et al.*, 'Dynamic interventions to control COVID-19 pandemic: a multivariate prediction modelling study comparing 16 worldwide countries', *Eur J Epidemiol*, vol. 35, no. 5, pp. 389–399, May 2020, doi: 10.1007/s10654-020-00649-w.

[12] A. Tomar and N. Gupta, 'Prediction for the spread of COVID-19 in India and effectiveness of preventive measures', *Science of The Total Environment*, vol. 728, p. 138762, Aug. 2020, doi: 10.1016/j.scitotenv.2020.138762.

[13] X. Zhang, R. Ma, and L. Wang, 'Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries', *Chaos, Solitons & Fractals*, vol. 135, p. 109829, Jun. 2020, doi: 10.1016/j.chaos.2020.109829.

[14] G. D. Barmparis and G. P. Tsironis, 'Estimating the infection horizon of COVID-19 in eight countries with a data-driven approach', *Chaos, Solitons & Fractals*, vol. 135, p. 109842, Jun. 2020, doi: 10.1016/j.chaos.2020.109842.

[15] Z. Zhao, X. Li, F. Liu, G. Zhu, C. Ma, and L. Wang, 'Prediction of the COVID-19 spread in African countries and implications for prevention and control: A case study in South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya', *Science of The Total Environment*, vol. 729, p. 138959, Aug. 2020, doi: 10.1016/j.scitotenv.2020.138959.

[16] Y. Qiu, X. Chen, and W. Shi, 'Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID-19) in China', *J Popul Econ*, May 2020, doi: 10.1007/s00148-020-00778-2.

[17] M. Soofi, F. Najafi, and B. Karami-Matin, 'Using Insights from Behavioral Economics to Mitigate the Spread of COVID-19', *Appl Health Econ Health Policy*, vol. 18, no. 3, pp. 345–350, Jun. 2020, doi: 10.1007/s40258-020-00595-4.

[18] K. Kwong and P. Lai, 'Spatial Components in Disease Modelling', in *Computational Science and Its Applications – ICCSA 2010*, Berlin, Heidelberg, 2010, pp. 389–400, doi: 10.1007/978-3-642-12156-2_30.

[19] G. Chowell, A. Erkoreka, C. Viboud, and B. Echeverri-Dávila, 'Spatial-temporal excess mortality patterns of the 1918–1919 influenza pandemic in Spain', *BMC Infect Dis*, vol. 14, no. 1, p. 371, Jul. 2014, doi: 10.1186/1471-2334-14-371.

[20] J. Adda, 'Economic Activity and the Spread of Viral Diseases: Evidence from High Frequency Data*', *The Quarterly Journal of Economics*, vol. 131, no. 2, pp. 891–941, May 2016, doi: 10.1093/qje/qjw005.

[21] W. Naudé, 'Artificial intelligence vs COVID-19: limitations, constraints and pitfalls', *AI & Soc*, Apr. 2020, doi: 10.1007/s00146-020-00978-0.

[22] R. G. M. Jr., *Beyond ANOVA : Basics of Applied Statistics*. Chapman and Hall/CRC, 1997.

[23] A. J. Smola and B. Schölkopf, 'A tutorial on support vector regression', *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004, doi: 10.1023/B:STCO.0000035301.49549.88.

[24] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[25] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[26] L. Breiman, 'Random Forests', *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[27] R. Polikar, 'Ensemble learning', *Scholarpedia*, vol. 4, no. 1, p. 2776, 2009, doi: 10.4249/scholarpedia.2776.

[28] S. B. Kotsiantis, 'Decision trees: a recent overview', *Artif Intell Rev*, vol. 39, no. 4, pp. 261–283, Jun. 2011, doi: 10.1007/s10462-011-9272-4.

[29] F. Pedregosa *et al.*, 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, vol. 12, p. 2825−2830, Oct. 2011.

[30] A. Hjerpe, 'Computing Random Forests Variable Importance Measures (VIM) on Mixed Continuous and Categorical Data', 2016.