SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
VOL. 3, NO. 3, DECEMBER 2020
DOI: 10.35377/saucis.03.03.771501

The Data Science Met with the COVID-19: Revealing the Most Critical Measures Taken for the COVID-19 Pandemic

Abdullah Talha Kabakus[1]
[1]Duzce University; talhakabakus@duzce.edu.tr; 903805421036

Received 19 July 2020; Revised 27 September 2020; Accepted 20 October 2020; Published online 31 December 2020

Abstract

The whole world has been fighting against the novel coronavirus 2019 (COVID-19) for months. Despite the advances in medical sciences, more than 235,000 people have died so far. And, despite all the measures taken for it, more than 3 million people have become sick of the COVID-19. The measures taken for the COVID-19 vary through countries. So, revealing the most critical measures is necessary for a better fight against both the COVID-19 and possible similar pandemics in the future. To this end, an analysis of the worldwide measures, which were taken so far, for the COVID-19 pandemic was proposed within this paper. Since it is still early days, for the best of our knowledge, there does not exist a single dataset contains all the features utilized within this study. Therefore, a novel global dataset containing the data regarding the COVID-19 for 52 countries around the world was constructed by combining various datasets. Then, the feature importance techniques were employed to reveal the importance of the utilized features which means revealing the most important measures taken for the COVID-19 pandemic for our case. Within the analysis, four features were utilized, namely, the population density, the walking mobility, the driving mobility, and the number of lockdown days. According to the experimental result, the population density was found as the most important feature which means the most critical measure in terms of increasing the spread of the COVID-19 pandemic. The order of the importance of the other features was found as the walking mobility, the driving mobility, and the number of lockdown days, respectively.

Keywords: COVID-19, coronavirus, pandemic, feature analysis, feature importance

1. Introduction

The novel coronavirus 2019 (COVID-19) emerged in Wuhan, China in December 2019 [1], [2], and it has rapidly spread into other provinces in China, and eventually, 212 countries as well by the 1st May 2020 [3]. The total number of confirmed global COVID-19 cases, and the total number of deaths have reached to 3,336,680, and 235,245, respectively, by the 1st May 2020 [3]. Since the COVID-19 is thought to be primarily transmitted by respiratory droplets [4], the primary goal of the whole world is to prevent the person-to-person spread of disease. To this end, the three common measures that are put into practice are (i) isolation, (ii) quarantine, and (iii) community containment. 'Isolation' is the separation of the infected people from non-infected people to prevent the person-to-person spread of disease, and generally occurs in a hospital or a designated facility. 'Quarantine' means the movement restriction of non-infected or suspicious people, as they may be still in the incubation period, to control communicable disease outbreaks. Quarantine generally involves restriction to the home or a designated facility and may be applied at the individual or group level and may be voluntary or mandatory. One of the points to take into consideration is that people in quarantine should monitor themselves for the occurrence of any symptoms of the disease. 'Community containment' is an intervention applied to an entire community, city, or region in order to minimize person interactions, except for the necessary minimal interaction to ensure vital supplies. Community containment starts with social distancing which involves keeping the distance between other people and usage of facemasks at all times in a broader community to reduce interaction between people. When social distancing is deemed to be insufficient, further practices such as the closure of schools, public markets, office buildings, and shopping malls, shutting down of the public transportation, and declaring a lockdown are put into practice as most of them have been applied in many countries. The three common measures that are put into practice in order to prevent pandemic, namely, (i) isolation, (ii) quarantine, and (iii) community containment, are illustrated in Figure 1.

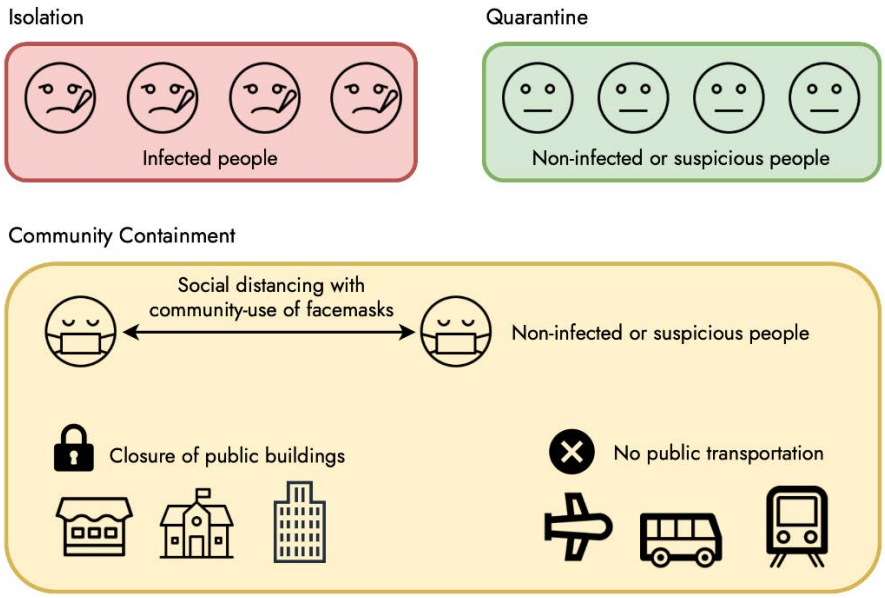

Figure 1 An illustration of the three common measures that are put into practice in order to prevent pandemic, namely, (i) isolation, (ii) quarantine, and (iii) community containment

The whole world has been fighting against the COVID-19 pandemic by taking various measures. So, it is critical to reveal the most critical measures taken for preventing the spread of the COVID-19 pandemic, and possible similar pandemic in the future. To this end, an analysis of the worldwide measures, which were taken so far, for the COVID-19 pandemic was proposed within this paper. To the best of our knowledge, there does not exist a single dataset contains all the features utilized within this study. Therefore, a novel global dataset was constructed by combining various sources. Then, machine learning algorithms were employed to reveal the most critical measures taken for preventing the spread of the COVID-19 pandemic. The rest of the paper is organized as follows: Section 2 presents the limited related work as it is still early days for completely understanding the COVID-19 and proposing approaches to prevent its spread. Section 3 describes the proposed study with implementation detail. Section 4 presents the experimental result and discussion. Finally, Section 5 concludes the paper with future directions.

2. Related Work

Jiang *et al.* [5] proposed a tool with AI (Artificial Intelligence) capabilities in order to predict patients at risk for more severe illness on initial presentation after algorithmically identifying the combinations of clinical characteristics of the COVID-19. They noted that key characteristics such as fever, lymphopenia, and chest imaging were not as predictive as severity. In addition to this, they reported that epidemiologic risks such as age and gender were not as predictive. They utilized six machine learning algorithms, namely, Logistic Regression, kNN (k-Nearest Neighbors), Decision Tree based on Gain Ratio, Decision Tree based on Gini Index, Random Forest, and SVM (Support Vector Machine), for the predictions. When it comes to the utilized algorithms' performance, SVM outperformed the other algorithms by providing an accuracy of 80%. Unlike this study, the proposed study aims to reveal the most critical features (measures) instead of prediction. Strzelecki and Rizun [6] proposed an infodemiological study based on Google Trends [7], which is a service that analyzes the popularity of the given terms or topics. According to the experimental result, Google Trends was able to forecast the rise of new cases. Unlike this study, the proposed study utilizes several data sources to make us various features from various sources. Fang *et al.* [8] proposed a radiomic signature to screen the COVID-19 from CT (Computed Tomography) images. They segmented the lung lesions from the CT images and extracted 77 radiomic features from the lesions. The four of these features, which were found as highly associated with the COVID-19 by utilizing the unsupervised consensus clustering and multiple cross-validations, were used as the inputs of *SVM* to build the radiomic signature. According to the result of the conducted experiments, the proposed model achieved an accuracy of 82.6% for the test set. Unlike

this study, the proposed one utilized publicly available data such as the number of lockdown days, the driving and walking mobility trends, and the population densities of countries instead of medical data.

3. Material and Method

In this section, the detail regarding the constructed dataset and performed analysis was described. Since it is still early days, for the best of our knowledge, there does not exist a single dataset contains all the features utilized within this study. Therefore, a novel global dataset was constructed programmatically through the implemented Python scripts by combining various sources. Similarly, the analysis including the exported plot was performed programmatically on the constructed data. Eventually, an end-to-end analysis based on the pipeline architecture was proposed within this study that accepts the raw CSV data as the input and generates the analysis result as the output.

3.1 Dataset Construction and Feature Extraction

The John Hopkins University Center for Systems Science and Engineering (JHU CSSE) provides the data of their interactive web-based dashboard [9] that tracks the COVID-19 global cases in real-time. The data, which is stored as CSV (comma-separated values) files, is hosted on *GitHub* [10] and is being updated daily through the reported global cases by starting the 22nd of January 2020. In order to read and query CSV files, a widely-used Python library, namely *Pandas* [11], was utilized which creates data frames from the raw CSV data. This data contains a CSV file per each day that contains (i) the number of confirmed cases, (ii) the number of deaths, (iii) the number of recovers, and (iv) the number of active cases for the countries/regions around the world. Since the aim of this study is revealing the best measures taken to minimize the spread of COVID-19, the only feature that is associated with the aim of this study is the number of confirmed cases as the other features are strongly associated with the health care services provided by countries which are out of the scope of this study. Therefore, the features available in the CSV file except for the number of confirmed cases were not included in the feature set of the proposed study. To better reflect the spread of the COVID-19, the ratio of the number of confirmed cases to the population of the country was utilized instead of the number of confirmed cases. The populations of countries were retrieved through the implemented Python script that utilizes a Python library, namely, *countryinfo* [12]. The countries, whose populations were not provided by this module, were eliminated from the analysis. Apple reports the COVID-19 mobility trends in countries/regions and cities on a daily basis by starting the 13th of January 2020 through the requests for directions in *Apple Maps* [13]. This report contains mobility trends in three transportation types, namely, (i) driving, (ii) walking, and (iii) transit. Since the report does not contain the transit mobility trend for some countries, this feature was not utilized within this study. The mobility trends in the 30 days after the first case was confirmed were considered, and the averages of the mobility trends during the 30 days were regarded as the final mobility trends. Similar to the population retrieval, the countries, whose mobility trends were not included in the report provided by Apple, were eliminated from the analysis. The other features utilized within this study are the population density of each country, which was retrieved from an up-to-date report [14] based on the reports of both the United Nations and The World Bank, and the number of lockdown days in the 30 days after the first case was confirmed. The date of the first confirmed case for each country was determined programmatically by sequentially inspecting the daily reports provided by the *JHU CSSE*. The start dates of the lockdowns for the countries, that declared a lockdown, were retrieved through a dataset hosted on *Kaggle* [15]. Eventually, the constructed dataset contains the aforementioned features for the 52 countries from all around the world which are highlighted in the world map* in Figure 2.

* The world map was generated thanks to the https://mapchart.net web portal.

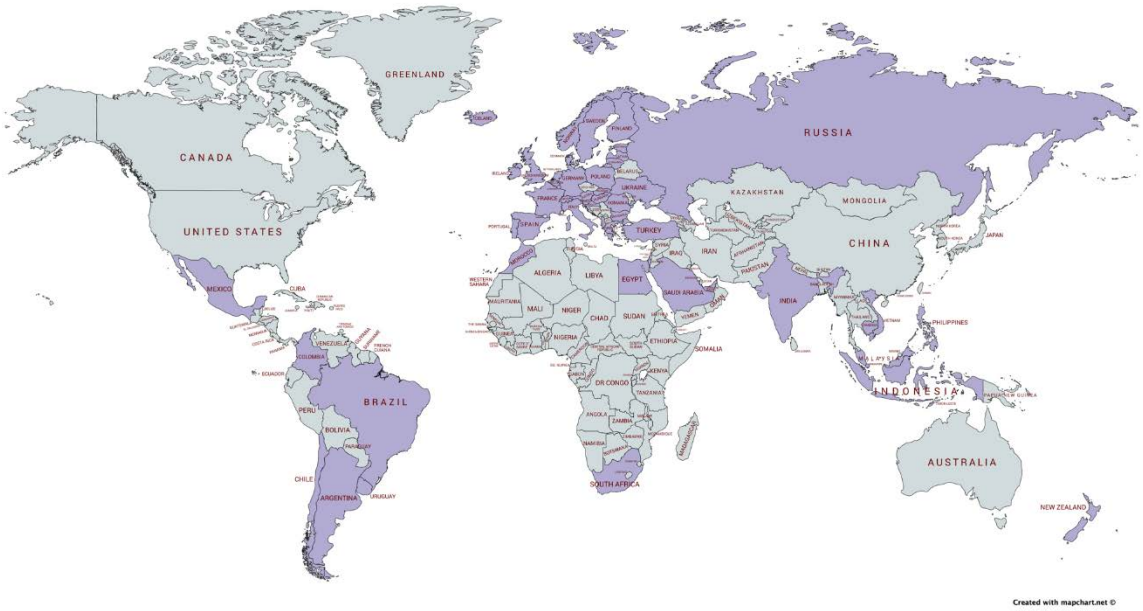

Figure 2 The world map that highlights the countries which are included in the analysis within the proposed study

After the dataset was constructed, it was mapped into the range of (0,1), which is also known as *Min-Max Normalization*, through the *MinMaxScaler* functionality of the *scikit-learn* [16], which is a widely-used machine learning library for the Python programming language. Data normalization is a critical process for all analyses based on data as it minimizes unwanted biases and experimental variance [17], [18]. It is very useful when the features of the data are on widely different scales [18] which was the case for the constructed dataset. An overview of the dataset construction phase of the proposed study including the extracted features, which are highlighted in bold, is presented in Figure 3.

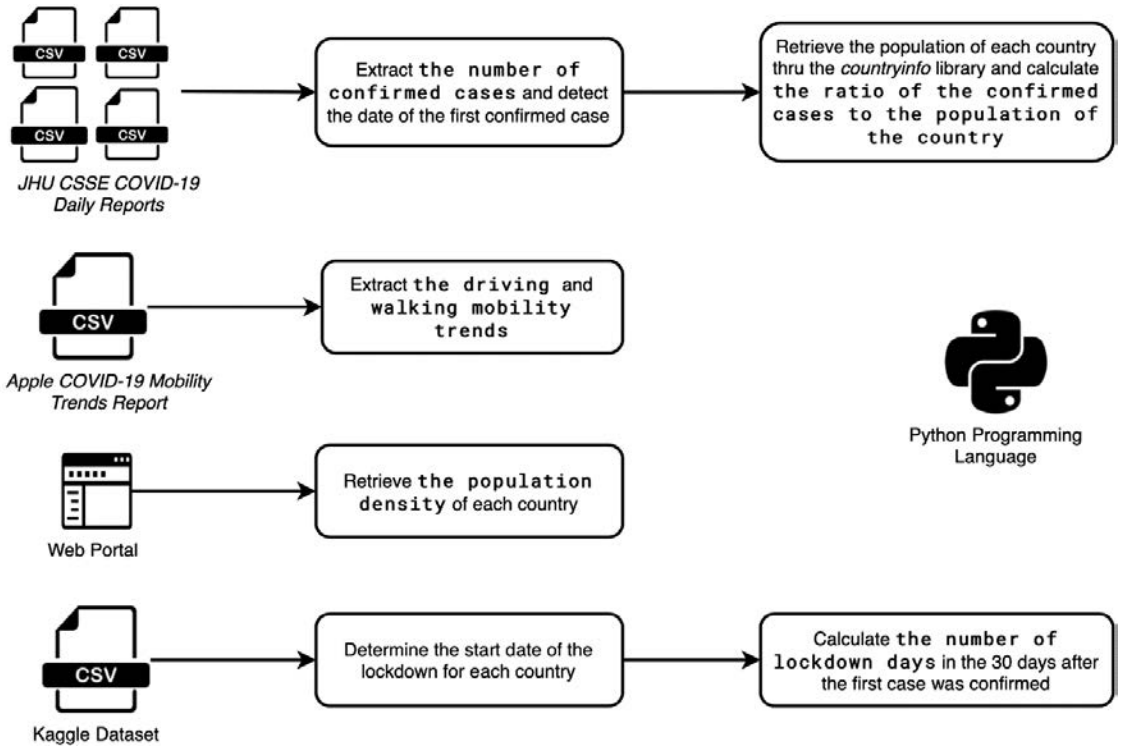

Figure 3 An overview of the dataset construction phase of the proposed study including the extracted features, which are highlighted in bold

3.2 Revealing the Importance of Features

The aim of the feature importance is revealing the relative importance of the features when making predictions for the target feature. In other words, feature importance reveals the inductiveness of features in terms of predicting the target feature [19]. To this end, both the Decision Tree Regressor and the Random Forest Regressor implementations, which are provided by the scikit-learn library, were utilized since these machine learning techniques have proven their efficiencies in the similar problems [20], [21]. Decision Tree Regressor is a non-parametric supervised learning method that follows recursive binary splitting technique to find the best prediction [22]. Random Forest Regressor is an ensemble method that is assembly of various Decision Tree Regressors which are combined using ensemble and predictions of each tree are averaged to find the best prediction [23]. The ratio of the confirmed cases to the population of the country was the target feature of the proposed study. The averages of the scores calculated through these regression techniques were regarded as the final scores of the features. The reason behind utilizing both of these regression techniques instead of employing one of them is to minimize the fluctuations of the scores that are calculated through each machine learning algorithm. Let *dt_score*, and *rf_score* denote the importance scores calculated by the Decision Tree Regressor, and Random Forest Regressor, respectively, the final importance score of a feature, which was denoted by importance_score, was calculated as seen in Equation 1:

$$importance_{score} = \frac{dt_{score} + rf_{score}}{2} \quad (1)$$

4. Experimental Result and Discussion

The proposed feature importance approach performed on the constructed dataset to reveal the importance of the features, which means the importance of the measures taken for the COVID-19 pandemic in our case. According to this experimental result, the utilized features were ranked as follows, respectively: the population density, the walking mobility, the driving mobility, and the number of lockdown days as the calculated importance scores of the utilized features are presented in Figure 4. When the experimental result was inspected, the following outcomes were deduced:

- *The population density* of the country, which is denoted by density in Figure 4, was by a wide margin the most important feature that increases the rate of incidence of COVID-19 as both the health professionals and researchers emphasize the critical importance of not being in close contact with people in a broader community (*a.k.a.* social distancing) [4] during the pandemic which is naturally more common in the countries with high population densities. Hence, a low population density is a natural way of ensuring social distancing.
- *The walking mobility* trend in the country, which is denoted by *avg_walking* in Figure 4, was the second most important feature after *the population density* in terms of affecting the rate of incidence of the COVID-19. This result is reasonable as while being in the broader community, the distance between people becomes critical to reduce contact with people, who may be still in the incubation phase. According to the reports, a distance of at least 1 meter between people is needed to be maintained for the COVID-19 [24]–[26].
- *The number of lockdown days*, which is denoted by *lockdown_days* in Figure 4, came after *the walking mobility* trend. During the COVID-19 pandemic, while some countries declared full lockdown, some others declared partial lockdown. In addition to this, some countries such as Sweden even did not declare any types of lockdown. The number of cases was relatively high in the United Kingdom (UK) despite declaring a full lockdown on the second day after the first case was confirmed.
- *The driving mobility* trend, which is denoted by *avg_driving* in Figure 4, was found as less critical compared to the other features as the driving already ensures some level of social distancing.

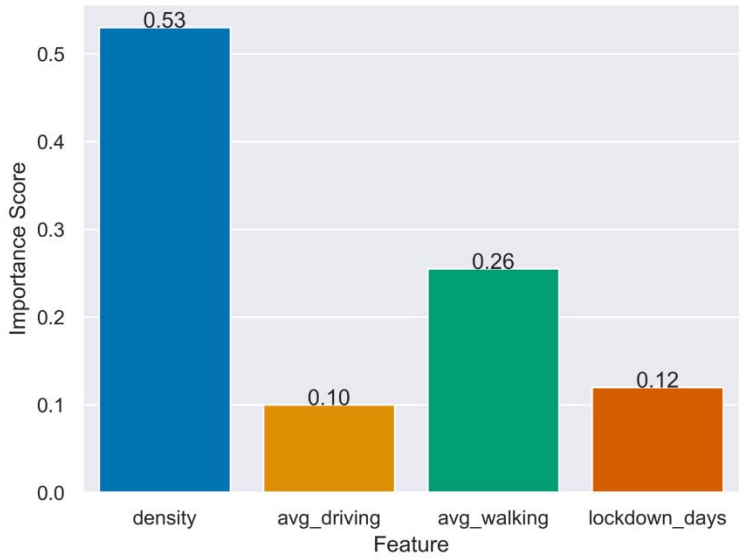

Figure 4 The calculated importance scores of the utilized features

The plots presented in Figure 5 were obtained when the model was employed with the Decision Tree Regressor and Random Forest Regressor, respectively. *The population density* was remained the most important feature amongst all features for both techniques. The remaining features were ranked when the model was employed with the Decision Tree Regressor as follows: *The walking mobility*, *the driving mobility*, and *the number of lockdown days*. When it was employed with the Random Forest Regressor, the rank of the remaining features was obtained as follows: The number of lockdown days, *the driving mobility*, and *the walking mobility*.

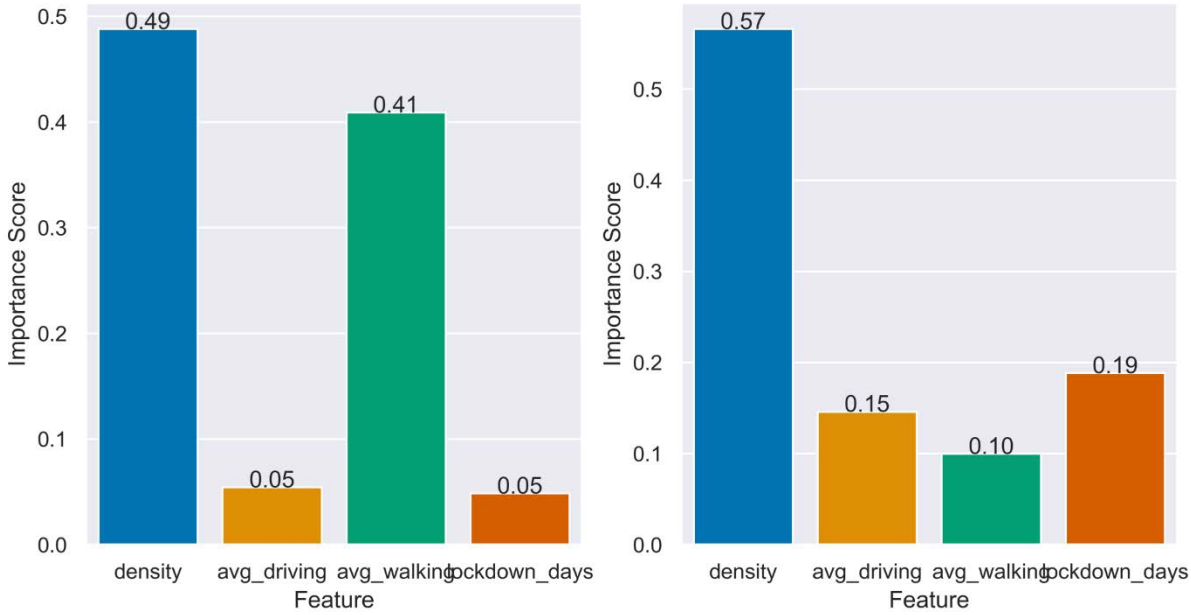

Figure 5 The calculated importance scores of the utilized features when the model was employed with the *Decision Tree Regressor* (left) and *Random Forest Regressor* (right), respectively

5. Conclusion

The COVID-19 pandemic has dramatically changed daily life worldwide as it has influenced 212 countries at the time of writing this paper. The countries that have been fighting against the COVID-19 have taken various measures against it to minimize the spread. So, it has become critical to reveal the most critical measures in terms of preventing the spread of the COVID-19. To this end, a novel global

dataset containing the data regarding the COVID-19 for 52 countries around the world was constructed and analysis was performed on it within this study. Within the analysis, four features were utilized, namely, the population density, the walking mobility, the driving mobility, and the number of lockdown days. According to the experimental result, the population density was found as the most important feature which means the most critical measure in terms of increasing the spread of the COVID-19 pandemic. The order of the importance of the other features was found as the walking mobility, the driving mobility, and the number of lockdown days, respectively.

As future work, the duration of the analysis may be extended as it was set to 30 days since it is still early days of COVID-19 pandemic. In addition to this, the lockdown may be detailed as there are some types of lockdown as the rules during a lockdown vary through the country it is declared by. Finally, the mobility trends would be retrieved from the official sources when they are revealed for worldwide. This would provide more inclusive data regarding the mobility trends in countries.

Acknowledgments

We would like to thank Professor of Pediatrics *Nimet Kabakus* for his insightful comments.

References

[1] N. Zhu *et al.*, "A novel coronavirus from patients with pneumonia in China, 2019," *N. Engl. J. Med.*, vol. 382, pp. 727–733, 2020, doi: 10.1056/NEJMoa2001017.

[2] J. F. W. Chan *et al.*, "A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster," *Lancet*, vol. 395, pp. 514–523, 2020, doi: 10.1016/S0140-6736(20)30154-9.

[3] "COVID-19 CORONAVIRUS PANDEMIC," *Worldometer*, 2020. https://www.worldometers.info/coronavirus/ (accessed May 05, 2020).

[4] A. Wilder-Smith and D. O. Freedman, "Isolation, quarantine, social distancing and community containment: Pivotal role for old-style public health measures in the novel coronavirus (2019-nCoV) outbreak," *J. Travel Med.*, vol. 27, no. 2, pp. 1–4, 2020, doi: 10.1093/jtm/taaa020.

[5] X. Jiang *et al.*, "Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity," *Comput. Mater. Contin.*, vol. 63, no. 1, pp. 537–551, 2020, doi: 10.32604/cmc.2020.010691.

[6] A. Strzelecki and M. Rizun, "Infodemiological Study Using Google Trends on Coronavirus Epidemic in Wuhan, China," *Int. J. Online Biomed. Eng.*, vol. 16, no. 4, pp. 139–146, 2020, doi: 10.3991/ijoe.v16i04.13531.

[7] "Google Trends," *Google*, 2020. https://trends.google.com/trends (accessed Aug. 02, 2020).

[8] M. Fang *et al.*, "CT radiomics can help screen the Coronavirus disease 2019 (COVID-19): a preliminary study," *Sci. China Inf. Sci.*, vol. 63, no. 7, pp. 1–8, 2020, doi: 10.1007/s11432-020-2849-3.

[9] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *Lancet Infect. Dis.*, 2020, doi: 10.1016/S1473-3099(20)30120-1.

[10] "Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE," *Johns Hopkins University*, 2020. https://github.com/CSSEGISandData/COVID-19 (accessed May 05, 2020).

[11] "pandas: Python Data Analysis Library," 2020. https://pandas.pydata.org (accessed May 05, 2020).

[12] P. Chandro, "porimol/countryinfo: A Python module for returning data about countries, ISO info and states/provinces within them," 2019. https://github.com/porimol/countryinfo (accessed May 05, 2020).

[13] "COVID-19 Mobility Trends Reports - Apple," *Apple*, 2020. https://www.apple.com/covid19/mobility (accessed May 05, 2020).

[14] "Countries by Population Density 2020 - StatisticsTimes.com," *StatisticsTimes.com*, 2020. http://statisticstimes.com/demographics/countries-by-population-density.php (accessed May 05, 2020).

[15] J. Papastylianou, "COVID-19 Lockdown dates by country | Kaggle," 2020. https://www.kaggle.com/jcyzag/covid19-lockdown-dates-by-country (accessed May 05, 2020).

[16] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[17] H. U. Zacharias, M. Altenbuchinger, and W. Gronwald, "Statistical Analysis of NMR Metabolic Fingerprints: Established Methods and Recent Advances," *Metabolites*, vol. 8, no. 3, pp. 1–10, 2018, doi: 10.3390/metabo8030047.

[18] S. C. Nayak, B. B. Misra, and H. S. Behera, "Impact of Data Normalization on Stock Index Forecasting," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 6, pp. 257–269, 2014.

[19] A. Zien, N. Krämer, S. Sonnenburg, and G. Rätsch, "The Feature Importance Ranking Measure," in *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2009)*, 2009, pp. 694–709, doi: 10.1007/978-3-642-04174-7_45.

[20] K. B. Prakash, S. S. Imambi, M. Ismail, T. Pavan Kumar, and Y. V. R. Naga Pawan, "Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 5, pp. 2199–2204, 2020, doi: 10.30534/ijeter/2020/117852020.

[21] A. K. M. B. Haque, T. H. Pranto, A. A. Noman, and A. Mahmood, "Insight about Detection, Prediction and Weather Impact of Coronavirus (COVID-19) using Neural Network," *Int. J. Artif. Intell. Appl.*, vol. 11, no. 4, pp. 67–81, 2020, doi: 10.5121/ijaia.2020.11406.

[22] S. Patil, A. Patil, and V. M. Phalle, "Life Prediction of Bearing by Using Adaboost Regressor," in *Proceedings of the TRIBOINDIA-2018 An International Conference on Tribology*, 2018, pp. 1–8, doi: 10.2139/ssrn.3398399.

[23] S. Patil, S. Desai, A. Patil, V. M. Phalle, V. Handikherkar, and F. S. Kazi, "Remaining Useful

Life (RUL) Prediction of Rolling Element Bearing Using Random Forest and Gradient Boosting Technique," in *Proceedings of the ASME International Mechanical Engineering Congress and Exposition, Proceedings (IMECE 2018)*, 2018, pp. 1–7, doi: 10.1115/IMECE2018-87623.

[24] V. C. C. Cheng *et al.*, "Escalating infection control response to the rapidly evolving epidemiology of the Coronavirus disease 2019 (COVID-19) due to SARS-CoV-2 in Hong Kong," *Infect. Control Hosp. Epidemiol.*, pp. 1–6, 2020, doi: 10.1017/ice.2020.58.

[25] "Rational use of personal protective equipment for coronavirus disease 2019 (COVID-19)," *World Health Organization (WHO)*. pp. 1–7, 2020.

[26] M. Lazzerini and G. Putoto, "COVID-19 in Italy: momentous decisions and many uncertainties," *Lancet Glob. Heal.*, vol. 8, no. 5, pp. e641–e642, 2020, doi: 10.1016/S2214-109X(20)30110-8.