Journal of Computer Science

Original Research Paper

Covid-19 Global Spread Analyzer: An ML-Based Attempt

[1]Rana Husni Al Mahmoud, [2]Eman Omar, [3]Khaled Taha, [3]Mahmoud Al-Sharif and [4]Abdullah Aref

[1]Department of Computer Science, University of Jordan, Amman, Jordan
[2]Department of Computer Science, University of the People, USA
[3]Social Media Lab, Trafalgar AI, Reading, UK
[4]Department of Computer Science, Princess Sumaya University for Technology, Jordan

Article history
Received: 18-07-2020
Revised: 28-09-2020
Accepted: 29-09-2020

Corresponding Author:
Rana Husni Al Mahmoud
Department of Computer Science, University of Jordan, Amman, Jordan
Email: rana.husni@gmail.com

Abstract: The novel Coronavirus 2019 (COVID-19) has caused a pandemic disease over 200 countries, influencing billions of humans. In this consequence, it is very much essential to the identify factors that correlate with the spread of this virus. The detection of coronavirus spread factors open up new challenges to the research community. Artificial Intelligence (AI) driven methods can be useful to predict the parameters, risks and effects of such an epidemic. Such predictions can be helpful to control and prevent the spread of such diseases. In this study, we introduce two datasets, each of which consists of 25 country-level factors and covers 137 countries summarizing different domains. COVID-19STC aims to detect the increase of the total cases, whereas COVID-19STD aimed for total death detection. For each data set, we applied three feature selection algorithms (vis. correlation coefficient, information gain and gain ratio). We also apply feature selection by the Wrapper methods using four classifiers, namely, NaiveBayes, SMO, J48 and Random Forest. The GDP, GDP Per Capital, E-Government Index and Smoking Habit factors found to be the main factors for the total cases detection with accuracy of 73% using the J48 classifier. The GDP and E-Government Index are found to be the main factors for total deaths detection with accuracy of 71% using J48 classifier.

Keywords: COVID-19, Coronavirus Disease, Coronavirus, Machine Learning, Prediction, Datasets

Introduction

The word Epidemic, derived from Greek, means the spread of disease rapidly to a large number of people in a short period of time within community, population, or region, whereas a pandemic is an epidemic that spread over multiple countries or continents, such as the H1N1 outbreak in 2009 and Coronavirus Disease 2019 (COVID-19) (Hays, 2005). The history of the epidemic goes long as far back as to the Middle era. The world was suffering from several epidemics (Pyne *et al*., 2015), which brings the need for a scientific and systematic means to study the distribution and determinant causes and risk factors of health-related states and events in specified populations, also known as Epidemiology (Diekmann and Heesterbeek, 2000). During the period of an epidemic, casting and forecasting are of crucial importance for public health planning and control domestically and internationally especially when the number of infections increases exponentially (Wu *et al*., 2020).

COVID-19 has raised serious concerns as its spread has become a global threat (Zhang *et al*., 2020). The virus began to spread widely in China at the end of 2019, before spreading rapidly in other parts of the world (Li *et al*., 2020a), despite the large-scale containment efforts of the Chinese government (Fanelli and Piazza, 2020). It has been declared as a global pandemic by the World Health Organization (WHO) on March 11, 2020 (Zhang *et al*., 2020).

The prediction of new infected cases, deceased cases and healed cases is certainly essential for health policy makers in order to estimate the capacity of a health system to cope with the stress caused by a pandemic (Fanelli and Piazza, 2020). Many scholars are trying to understand the spread dynamics of COVID-19 and to propose effective prevention and control strategies since December 2019 (Fanelli and Piazza, 2020; Zhang *et al*., 2020; Jung *et al*., 2020).

Epidemiology modeling allows for epidemiological parameters estimation from data, identification of

patterns, assessment of the relative merits of alternative control strategies and prediction of epidemiological or evolutionary dynamics. It helps in gaining insights into infectious as well as in designing control strategies (Bauch *et al*., 2005). In this study, we introduce two datasets, each of which consists of 25 country-level factors and covers 137 countries summarizing Geographic, Demographic, Economic, Healthcare System, Transportation, Technological, Social, Cultural, Religious and Political domains. COVID-19STC aims to detect the increase of the total cases, whereas COVID-19STD aimed for total death detection. Then we analyzed the two datasets using different machine learning algorithms with various feature selection sachems. Four of these factors found to be able to create models comparable to those models created based of the twenty-five potential factors analyzed.

The paper is organized as follows. Related work is presented next followed by a description of the data sources used and the proposed method. Experiments and results' analysis are given in section 4, followed by discussion of the research results. Conclusions and directions for future work are presented in the last section.

Related works

Epidemics of viruses have been studied with the aid of graphs and random graphs for decades (Kephart and White, 1992). A directed random graph used in (Khelil *et al*., 2002) to extend epidemiological models to investigate the spreading of computer viruses. Cellular automata used to model the spread of diseases in small-world networks in (Verdasca *et al*., 2005). The small-world network model is found to be better than the classical Susceptible, Infected and Recovered (SIR) for describing the local variability. A 4-state model was used to simulate the SARS transmission under a small-world topology in (Anghel *et al*., 2007) using the Hong Kong SARS data. The model takes into account those who were infected but not yet infectious. The research suggests that outbreaks could be prevented if the patients with symptoms were isolated as soon as possible (Costa *et al*., 2011).

Classical mathematical epidemiology found to be successful in informing public health policy makers. Such models focus on rate-based differential equation models, where the population is partitioned into subgroups based on various criteria and uses differential equation models to describe the disease dynamics across these groups (Marathe and Ramakrishnan, 2013). A mathematical model for the spread of Ebola fever epidemics was built in (Legrand *et al*., 2007) based on data from Democratic Republic of Congo (DRC) in 1995 and in Uganda in 2000, concentrating on the rapid institution of control measures. Other researchers analyzed Ebola outbreak with different strategies including (Sau, 2017; Pandey and Karthikeyan, 2011; Pigott *et al*., 2014). Nevertheless a potential weakness of such approach is its inability to capture the complexity of human interactions and behaviors (Marathe and Ramakrishnan, 2013).

Artificial Intelligence tools are proposed for predicting outbreak for some diseases (Philemon *et al*., 2019; Abdulkareem *et al*., 2020). For example, Diarrhea outbreak (Machado *et al*., 2019) and cardiovascular diseases (Mezzatesta *et al*., 2019; Jhuo *et al*., 2019).

Recently, advances in machine learning, data mining and data science make it possible to develop an indispensable solution to treat using data. It is used to predicate epidemiological characteristics and control the spatiotemporal transmission of disease throughout the world (Hamer *et al*., 2020). The use of machine learning and reasoning methods in support of computational epidemiology is a rich area with many significant research challenges (Marathe and Ramakrishnan, 2013).

Machine learning was used in (Forna *et al*., 2019) to study the epidemiological characteristics of the Ebola virus outbreak in West Africa. The research presented in (Sadilek *et al*., 2012) explores how individuals contribute to the global spread of disease. Using the Support Vector Machine learning algorithm (SVM), scholars predicted if users were sick based on their tweets. Geo-tagged tweets are used to infer user locations and the move of individuals between cities and the timelines of target users are used to infer their interactions with others. Machine learning techniques are used to evaluate the performance of the time series forecasting of casualties in the case of Ebola Outbreak in

Recently many scholars are attracted to find a way to predict and recover, either based on data analysis or on health models, epidemic predictions of COVID-19 (Peng *et al*., 2020; Zhao *et al*., 2020a; 2020b; Chen *et al*., 2020b; Li *et al*., 2020b; Wu *et al*., 2020; Imai *et al*., 2020; Hilton and Keeling, 2020; Kastner *et al*., 2020; Jia *et al*., 2020; Zeng *et al*., 2020; Buizza, 2020). Methods to predict COVID-19 patients, using a mobile phone, was presented in (Rao and Vazquez, 2020) and a prognostic prediction model based on XGBoost machine learning algorithm built in (Yan *et al*., 2020) to identify early detection of high-risk patients before they transmitted from mild to critically ill. The work in (Ivanov, 2020) shows how that epidemic outbreaks represent one specific case of supply chain disruptions. This type of supply chain risks is distinctively characterized by long-term disruption existence and its unpredictable scaling, simultaneous disruption propagation and epidemic outbreak propagation and simultaneous disruptions in supply, demand and logistics infrastructure. An online/mobile GIS and mapping dashboards and applications for tracking the COVID-19 epidemic and associated events described in (Boulos and Geraghty, 2020). The work in (Killeen *et al*., 2020) presents aggregated out-of-home activity information for

various points of interest for each county of the US, as well as providing tools to read them, to help researchers investigating how the disease spreads. The metrics they are working on include demographics ethnicity, housing, education, employment, income, climate, transit scores, healthcare system-related. To predict the country-specific risk of (COVID-19), a shallow Long Short-Term Memory (LSTM) based neural network optimized using Bayesian optimization presented in (Pal *et al*., 2020). Observed spread of COVID 19 found to be correlated with climatological temperatures, latitude, travel, population density and sociological trends as pointed out in (Poole, 2020). Similar findings presented in (Sajadi *et al*., 2020).

An attempt to forecast the number of deaths in China due to COVID-19 China is presented in (Gao *et al*., 2020) based on official accumulated the number of deaths using Boltzmann function and the Richards function. The generalized additive model used in recent study on death rates in Wuhan. The study suggested that the temperature variation and humidity are factors affecting death rates due to the COVID-19 (Ma *et al*., 2020). Based on statistical analysis of data from 54 countries, the it was suggested in (Chen *et al*., 2020a) that temperature, wind speed and relative humidity combined together could predict the epidemic situation, which could help decision maker on COVID-19 outbreak control.

An attempt to statistically analyze COVID-19 infections based on data obtained form WHO is presented in (Kumar and Hembram, 2020) and found that the infection curve of China and Republic of Korea almost saturated. No solid reasoning provided for such findings as the aim of the work is to provide statistical analysis.

As with (Yang *et al*., 2020), the work in (Jia *et al*., 2020) attempted to predict the epidemic curve in China. However, they adopted three mathematical models: Logistic model, Bertalanffy model and Gompertz model. They based their work on SARS data and found that Logistic model outperforms the other two models and that the accumulative number of infections in Chain would between 80261 and 85140 (Pandey and Karthikeyan, 2011).

Machine learning modeling used in (DeCaprio *et al*., 2020) for identify individuals who are at the greatest risk due to COVID-19 based on data for complications due to other upper respiratory infections to address limited COVID-19 specific information. They used a feature set derived from medical insurance claims. A variant of the Susceptible-Exposed-Infectious-Removed (SEIR) model used in (Yang *et al*., 2020) with Long-Short-Term-Memory (LSTM) recurrent neural network to derive the epidemic curve in China based on SARS data. They emphasized that the adopted control measures in January 2020 in China was necessary for reducing the spread of COVID-19. Whereas the work of (Fong *et al*., 2020b) proposed Polynomial Neural Network with corrective feedback (PNN + cf) to help forecasting the number of infections even with small data set. The method found to useful in generating acceptable forecast for a novel disease such as COVID-19. The work further elaborated in (Fong *et al*., 2020a) and a deep learning-based Composite Monte-Carlo (CMC) is used in conjunction of fuzzy rule induction techniques and validated based on COVID-19 data from The Chinese Center for Disease Control and Prevention1 (CDCP) considering factors such as infection rates and death rates. The work focuses fusing on deterministic and non-deterministic data series into a Monte-Carlo (MC) simulation for fuzzy decision making to help with early decision making of a novel disease, as decision making for a novel disease can be critical in the initial stage an epidemic especially when available data considered scarce.

A Modified Auto-Encoder (MAE) method for real time forecasting of the new and cumulative COVID-19 cases based on WHO data under various interventions strategies in various countries presented in (Hu *et al*., 2020) and concluded that public health intervention was extremely necessary. A delay of one month in Italy increased the maximum number of cases from 29,475 to 1,493,498 and a delay of one month in Germany from increased the maximum number of cases from 8,795 to 144,542.

Unlike most existing related works which were based data from China or limited set of countries, our study is based on publicly available data related to most countries, when such data could be retrieved. In this study we attempt to consider a large number of macro-level factors such as GDP rather than considering micro-level factors such as repertory system complications or considering a limited number of factors.

Methodology

We start by analyzing potential factors that may be used to model the spread of the disease and group them into categories to simplify their analysis. Factors analyzed in this study were divided into Geographic, Demographic, Economic, Healthcare System, Transportation, Technological, Social, Cultural, and Religious and Political metrics categories as indicated in Table 1. For the purpose of this study, we rely on publicly available data only as processing privacy-protected data can be done in a separate study due to the time needed to obtain privacy-protected data. Following data extraction, we create a dataset for further analysis. To avoid noise and find the most appropriate set of features that can use to model the spread of the disease, we apply some well-known feature selection

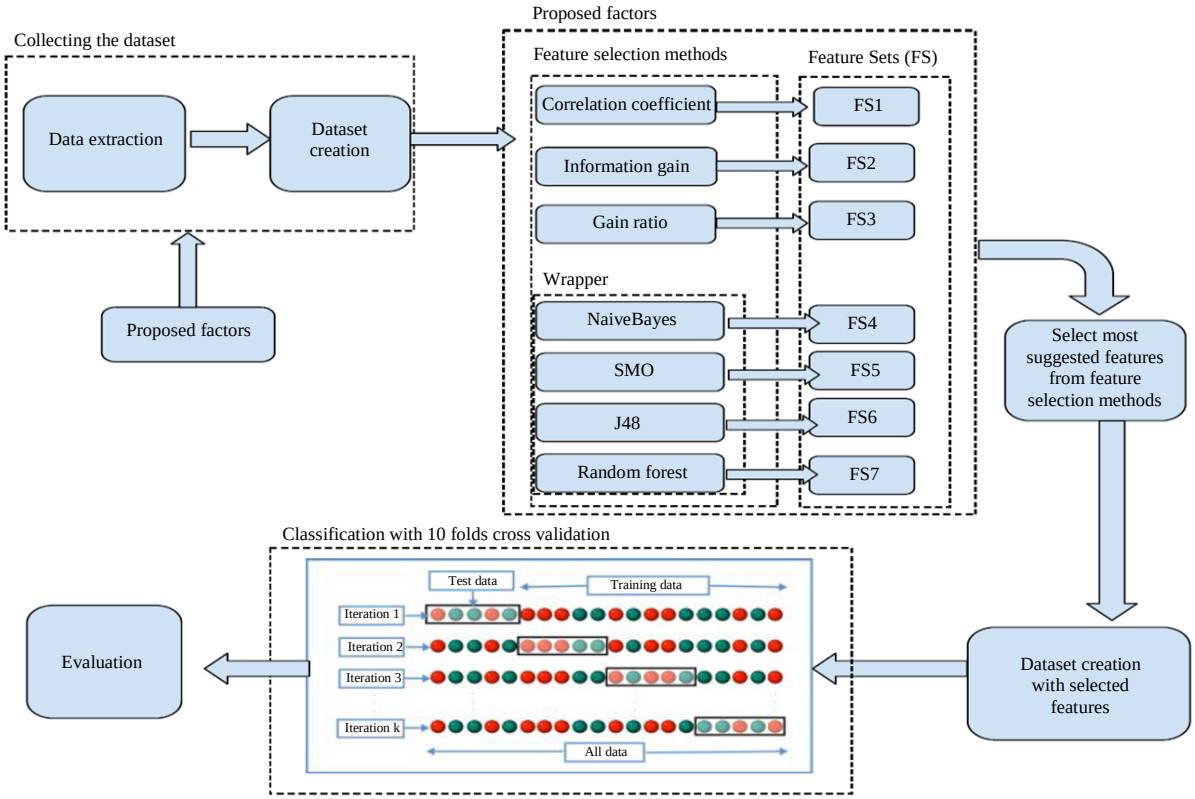

Fig. 1: COVID-19 spread radar methodology

schemes before building and evaluation a model. Figure 1 shows the flow layout of these stages.

Data Extraction and Dataset Creation

We tried to extract publicly available data related to all factors in Table 1. Unfortunately, for several factors, we could not find such data for most countries. Table 2 lists potential factors and the number of countries for which relevant data we could extract. All data extracted on 15/4/2020. Because of missing values in features for some countries; we limit ourselves to the 25 factors that, each of which covers at least 137 countries such as GDP, population density and air traffic out. The list of these factors is presented in Table 3 followed by the corresponding countries in Table 4.

Following data extraction, typical data normalization adapted to address differences in data ranges for various factors in the dataset.

Selection of Best Subset of Features

Feature selection schemes can be used for removing noisy, irrelevant, and redundant features which results in a smaller subset of relevant features from the original ones (Miao and Niu, 2016) aiming to get more accurate models. Information gain and wrapper selection among the well known feature selection schemes. A survey of feature selection schemes can be found in (Miao and Niu, 2016; Chandrashekar and Sahin, 2014; Molina *et al*., 2002).

The correlation coefficient indicates the strength and direction of a relationship between two random variables. The commonest use refers to a linear relationship. Two variables have strong dependency when their correlation coefficient value is close to 1 or -1. When the value is 0, it means that the two variables are not related at all (Hsu and Hsieh, 2010). Information Gain (IG) is an entropy-based feature evaluation method, widely used in the field of machine learning. IG measures the number of bits of information obtained for category prediction by knowing the presence or absence of a feature in an instance (Lei, 2012). IG evaluates features individually, scores each feature without considering the redundancy between them and selects the number of features which are predefined with the highest correlation rates (Quinlan, 1986).

The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values (Han *et al*., 2011). Gain Ratio (GR) is a modification of the information gain that reduces its bias. Gain ratio takes number and size of branches into account when choosing an attribute (Priyadarsini *et al*., 2011).

Table 1: List of all suggested features

Geographic	Economic	Social + Cultural + Religious
Total Area (Km^2)	GDP (2019) $	Lifestyle
Elevation (above sea leavel)	GDP percapita $	Agregation n gathering
Temperature (Jan - Apr)	Economic Model (communist, socialist, capitalist, free market..)	Schooling
Relative Humidity % (Jan-Apr)	Imports from China ($) 2019	Extended family
Wind speed (avg)	Top product	Emotional + intimcy (hsndshaking, kissing..)
Rural land %	CO_2 Emissions	Personal hygene
Agricultural land %	Digital Economy index (% of GDP)	Queing
Terrain (mountains Vs desert vs rivers)	eCommerce penetration	UNDP index (human develoment index)
Coastal line length (total)	ePayments Size/Total Transactions (internal)	Reserved (conservative, radical, open..)
Number of camels	Currency Exchange Rate (to $)	Worship attendance + relgious status
Numbers of cats	Unemployment %	Per capita expenditure on food
Number of bats	Energy cost ($ price per L gasoline)	smoking habit
Number of Pets	Expenditure on organic food	shopping centers + commercial properties
Distance from equator (north)	Wealth distribution	Personal Space
Distance from equator (south)		opennes about relationships (legalised prostitution)
Distance from China (Originating Point)		Single parents + no parents
Natural Resources		Sports events + sports club
		number of cinema + theutres
Healthcare System & General Health	Demographic	Entertainmnet (Night clubs + dancing clubs)
Previous epedimics	Population	Alcohol consumption
Radioactive materials	Population Density (/km^2)	Drugs (legalised, consumptions..)
Obesity %	Avg. age	Crime rate
Cancer %	Life expectancy	NGO + Charity
Diabetis %	Literacy % (basic education)	Urban green spaces - parks (open spaces Vs closed)
Beds/1000	Per capita from built up space	Hapiness index
Doctors/1000	Per capita water consumption (L/Year)	Social media + messaging penetration
Nurses/1000	Per capita cleaning materials consumption ($/Year)	Corona search on google by country
Avg. occupancy of hospitals (%)	HDI	
Digitization in Healthcare System	Avg. family size	
Vaccines %	Bachelors %	
Child mortality (per 1000)	Number of persons per household	
% of heart & coronery diseases	Avg. household area (m^2)	
public health expenditure	Higher education	
	Expats %	
Transportation	Technological	Political
Air Traffic out	Internet Penetration	Democracy index
Airport handling capacity (per year)	Mobile Penetration	Transparancy index
trains + buses (total)/local transportation	ICT development index	Moral freedom index
Maritime Traffic	% GDP expenditure on R&D	Freedom of press
Transportation behaviors (frequency + distances...etc)	Innovation index	Human freedom index
Avg commute per person	# of Stratups/per capita	
	ICT exports	
	Tech sector slice in GDP %	
	eGovt Development Index (EGDI)	
	Digital Competitiveness Ranking	

Wrappers are methods for feedback which incorporate the ML algorithm into the feature selection process, i.e., they depend on a specific classifier's performance to assess the quality of a set of features. Wrapper methods look through the space of feature subsets and calculate the accuracy of one classifier for each feature that can be added to or removed from the feature subset (Janecek *et al*., 2008).

Evaluation Measures

We compare the performance of all these methods based on a set of standard evaluation measurements (described next):

- Accuracy: Accuracy is a metric used to estimate how a classifier can correctly predict low, neutral, and high instances for each class. It can be calculated as the ratio of correctly classified instances to the total number of instances (Sokolova and Lapalme, 2009).
- Root Mean Squared Error (RMSE): Displays the error in both predicted and actual classes of the dataset instances. For more precise classification results, RMSE should have lower values (Witten and Frank, 2002).
- F-Measure: F-Measure is a composition of Precision and Recall. It is a consistent average of the two metrics which is used as an accumulated performance score. F-Measure of a class C, where C is low, neutral, or high can be calculated as in Equation 1 which is adapted from the general macro F-Measure equation (Sokolova and Lapalme, 2009):

$$F-Measure(C)=\frac{2*Precision(C)*Recall(C)}{Precision(C)+Recall(C)} \quad (1)$$

- Area Under Curve (AUC) Area Under the Curve (AUC) is commonly used as a summary measure of the Receiver Operating Characteristic (ROC) curve. Which measures the trade-off between sensitivity and specificity. The higher the area the better is the decision rule (Metz, 1978)

Table 2: List of features with number of countries

Feature	Num
Population[3]	189
ICT service exports[1]	210
crime rate[9]	108
Average elevation meter[4]	175
Population density[3]	181
Forest area[3]	227
Coastal line length total[7]	245
Life expectancy[3]	189
Happiness index[3]	120
Smoking habit[2]	159
cancer[3]	170
Transparency index[3]	157
Land area[1]	242
Diabetes[3]	181
Research development[1]	144
Agricultural land[1]	237
ICT development index[6]	156
Percapita food[3]	154
Rural land area[1]	211
Techsector GDP[1]	188
Democracy overall score[10]	151
GDP[1]	232
E-government index[8]	166
World index of moral freedom[2]	140
GDP per capita[1]	237
hygiene[1]	213
Freedom of press[2]	155
CO2Emissions[1]	233
UNDP index[3]	166
HUMAN FREEDOM[5]	146
Unemployment[1]	218
Air traffic out[1]	245
TotalCases[11]	173
Nurses and midwives[1]	209
Airport handling capacity[1]	184
Total deaths[11]	173
Individuals using internet[1]	236
International tourism[1]	137
Mobile cellular subscriptions[1]	235
Total alcohol consumption per capita[1]	216
Daily propagation speed[12]	186

[1]https://data.worldbank.org [2]https://en.wikipedia.org
[3]https://ourworldindata.org
[4]https://www.atlasbig.com
[5]https://www.cato.org 6 https://www.itu.int
[7]http://world.bymap.org 8 https://knoema.com/
[9]https://worldpopulationreview.com
[10]https://www.eiu.com
[11]https://www.worldometers.info [12]Calculated data

Table 3: Final selected features

Feature	Feature
Population	GDPPerCapita
Average Elevation Meter	CO^2 emissions
Coastal Line Length Total	Unemployment
Smoking Habit	Nurses and midwives
Land area	Individuals using internet
Agricultural Land	Mobile cellular subscriptions
GDP	Population density
Life expectancy	Total alcohol consumption per capita
Cancer	Forest area
Diabetes	Transparency Index
E-Government Index	Freedom of press
Hygiene	Total cases
UNDP index	Total deaths
Air traffic out	Daily propagation speed

Table 4: Final selected countries

Country	Country	Country
Afghanistan	Ghana	Niger
Albania	Greece	Nigeria
Algeria	Guatemala	Norway
Argentina	Guinea	Oman
Armenia	Guinea-Bissau	Pakistan
Australia	Guyana	Panama
Austria	Haiti	Paraguay
Azerbaijan	Honduras	Peru
Bahrain	Hungary	Philippines
Bangladesh	India	Poland
Belarus	Indonesia	Portugal
Belgium	Iran Islamic Rep.	Qatar
Benin	Iraq	Romania
Bolivia	Ireland	Russian Federation
Bosnia and Herzegovina	Israel	Rwanda
Botswana	Italy	Saudi Arabia
Brazil	Jamaica	Senegal
Brunei Darussalam	Japan	Serbia
Bulgaria	Jordan	Sierra Leone
Burkina Faso	Kazakhstan	Singapore
Burundi	Kenya	Slovak Republic
Cambodia	Kuwait	Slovenia
Canada	Kyrgyz Republic	South Africa
Chad	Lao PDR	Spain
Chile	Latvia	Sri Lanka
China	Lebanon	Suriname
Colombia	Liberia	Sweden
Costa Rica	Libya	Switzerland
Croatia	Lithuania	Thailand
Cuba	Luxembourg	The Gambia
Cyprus	Madagascar	Togo
Czech Republic	Malawi	Trinidad and Tobago
Denmark	Malaysia	Tunisia
Djibouti	Maldives	Turkey
Dominican Republic	Mali	Uganda
Ecuador	Malta	Ukraine
Egypt Arab Rep.	Mauritania	United Arab Emirates
El Salvador	Mauritius	United Kingdom
Equatorial Guinea	Mongolia	United States
Estonia	Montenegro	Uruguay
Ethiopia	Morocco	Uzbekistan
Finland	Mozambique	Vietnam
France	Myanmar	Yemen Rep.
Gabon	Nepal Zambia	
Georgia	New Zealand	Zimbabwe
Germany	Nicaragua	

Experiments and Evaluation Results

In this section, we present the conducted experiments to test the performance of the proposed classification models and discuss their evaluation results.

Experiments Setup

All experiments were conducted using a personal computer with Intel^{R}core^{TM} i5-5500U CPU @ 2.53 GHz/4 GB RAM. We experimented with different algorithms, namely: (1) Sequential Minimal Optimization (SMO), (2) Random Forest, (3) J48 and (4) Naive Bayes. In all experiments, all algorithms were implemented using Weka. All classification algorithms are trained using 10-fold cross-validation. In 10-fold cross validation, the available data is randomly divided into 10 disjoint subsets of approximately the same size. Nine sets are used for building the classifier and the remaining subset is used as the test set. Then the test set is used to determine the accuracy. This is done ten times in order to use every subset as a test subset. The accuracy calculated as a mean of the accuracy value for each of the classifiers.

Creating Datasets

We illustrate the methods presented in this study using two datasets:

- Predicting the spread of COVID-19 with respect to the total cases (COVID-19STC[1])
- Predicting the spread of COVID-19 with respect to the total deaths (COVID-19STD[2])

Each dataset contains 25 features of 137 countries. In the COVID-19STC dataset, the target class is Total Cases, whereas, in the COVID-19STD dataset, the target class is Total Death.

The COVID-19STC dataset is sorted in ascending order according to the total cases feature. We assign ’1’ (low) as a label to the countries in the first one-third part in the sorted dataset; we assign ’2’ (intermediate) to the countries in the second third in the sorted dataset; finally, the remaining countries in the third part will have ’3’ (high) as a class label. The same process done for COVID-19STD dataset. The COVID-19STD dataset is sorted according to the total deaths feature in ascending order. The dataset is labelled similar to the COVID-19STC.

Building a Model based on all Factors

Initially, we try to use all extracted factors presented in Table 3. We tried several machine learning algorithms and reported best results in Table 5 and for the COVID-19STC dataset and Table 6 COVID-19STD dataset where random forest outperformed other algorithms for the two datasets. In terms of AUC, NaiveBayes came next for both datasets. Such results help setting a base line for later comparisons.

Selection of the Best Subset of Features

A reductionist view assumes that the prediction of virus speed relies on the sum of risk features, as is the case with most scoring systems. We believe that such a reductionist approach is limited in its ability to successfully predict the spread of virus. The majority of virus speed (e.g., low, intermediate, or high) do not arise from a linear interaction between isolated factors, but from non-linear interactions among a web of determinants (Geographic, Demographic, Economic, ...etc.). For each data set (COVID-19STC and COVID-19STD), we run three feature selection algorithms: Correlation coefficient, information gain and gain ratio. The values from these methods are recorded in Tables 7 and 8 for COVID-19STC and COVID-19STD datasets, respectively. For COVID-19STC, Individuals Using Internet and E-Government Index had the highest correlation coefficient, GDP has the highest information gain and CO_2 Emissions has the highest gain ratio. For COVID-19STD UNDP index has the highest correlation coefficient, CO_2 Emissions has the highest information gain and GDP has the highest gain ratio.

The larger values correspond features for each method, indicate the importance of these features to the prediction. The wrapper model techniques evaluate the features using the learning algorithm that will ultimately be employed. Thus, they “wrap” the selection process around the learning algorithm. We apply Wrapper method on the original data sets by using four classifiers: NaiveBayes, SMO, J48 and Random Forest. A comparison of selected features based on wrapper feature selection presented in Table 9 for COVID-19STC dataset and in Table 10 for COVID-19STD dataset.

Table 5: The Classification results based on the all extracted features (COVID-19STC)

	NaiveBayes	SMO	J48	Random forest
Accuracy	0.71	0.61	0.67	0.76
F-Measure	0.72	0.60	0.67	0.76
Root MSE	0.42	0.42	0.45	0.36
AUC	0.84	0.74	0.77	0.86

Table 6: The Classification results based on the all extracted features (COVID-19STD)

	NaiveBayes	SMO	J48	Random forest
Accuracy	0.55	0.58	0.56	0.71
F-Measure	0.55	0.58	0.56	0.7
Root MSE	0.53	0.44	0.51	0.38
AUC	0.77	0.71	0.65	0.83

[1]https://drive.google.com/file/d/1EJKEDgR2beSTRwWT7I3v_zpm1n2Ll1Ik/view?usp=sharing
[2]https://drive.google.com/file/d/1VzmpdLQ6x8QUzY-PAn7zWySK6Ij5bJbz/view?usp=sharing

Table 7: Comparison of selected features based on feature selection methods for COVID-19STC dataset

Feature	Correlation coefficient	Information gain	Gain ratio
Population	0.17	0	0
AverageElevationMeter	0.03	0	0
CoastalLineLengthTotal	0.19	0.20	0.234
smokingHabit	0.31	0.28	0.28
Land area	0.20	0	0
AgriculturalLand	0.19	0	0
GDP	0.22	0.66	0.43
GDPPerCapita	0.37	0.39	0.307
CO2Emissions	0.19	0.45	0.45
Unemployment	0.088	0	0
NursesAndMidwives	0.32	0.22	0.28
IndividualsUsingInternet	0.41	0.31	0.33
MobileCellularSubscriptions	0.17	0.12	0.17
Population Density	0.067	0	0
Life expectancy	0.38	0.30	0.30
cancer	0.26	0.22	0.28
Diabetes	0.11	0.12	0.19
E-Government Index	0.41	0.39	0.39
hygiene	0.30	0.26	0.26
UNDPindex	0.40	0.45	0.3
AirTrafficout	0.21	0.38	0.38
Total alcohol consumption per capita	0.20	0	0
Forestarea	0.20	0.13	0.14
TransparencyIndex	0.35	0.30	0.24
freedomofpress	0.12	0	0

Table 8: Comparison of selected features based on feature selection methods for COVID-19STD dataset

Feature	Correlation coefficient	Information gain	Gain ratio
Population	0.19	0.18	0.205
AverageElevationMeter	0.09	0	0
CoastalLineLengthTotal	0.19	0.23	0.26
smokingHabit	0.24	0.17	0.171
Land	0.19	0	0
AgriculturalLand	0.15	0	0
GDP	0.22	0.52	0.53
GDPPerCapita	0.23	0.22	0.28
CO2Emissions	0.18	0.44	0.45
Unemployment	0.01	0	0
NursesAndMidwives	0.22	0.13	0.15
IndividualsUsingInternet	0.32	0.20	0.20
MobileCellularSubscriptions	0.15	0.12	0.18
Population	0.07	0	0
Life	0.34	0.25	0.25
cancer	0.21	0.17	0.247
Diabetes	0.14	0	0
E-Government	0.36	0.28	0.304
hygiene	0.25	0.22	0.30
UNDPindex	0.34	0.23	0.24
AirTrafficout	0.20	0.35	0.36
Total	0.16	0	0
Forestarea	0.18	0	0
TransparencyIndex	0.25	0.14	0.16
freedomofpress	0.11	0	0

Table 9: Comparison of selected features based on wrapper feature selection (COVID-19STC)

NaiveBayes	SMO	J48	Random Forest
smokingHabit	Population	smokingHabit	Population
GDP	AverageElevationMeter	GDP	smokingHabit
E-Government	smokingHabit		hygiene GDP
hygiene	Land		Unemployment
	GDP Diabetes		
	GDPPerCapita		
	CO2Emissions		
	E-Government		
	AirTrafficout		
	TransparencyIndex		

Table 10: Comparison of selected features based on wrapper feature selection (COVID-19STD)

NaiveBayes	SMO	J48	Random Forest
MobileCellular Subscriptions	Population	AgriculturalLand	GDP
E-Government	Population Density	GDP	NursesAndMidwives
	E-Government	GDPPerCapita	MobileCellular Subscriptions
		MobileCellular Subscriptions	E-Government Index
			UNDPindex
			Total alcohol consumption per capita
			freedomofpress

Table 11: Experiment result based on wrapper feature selection (COVID-19STC)

	NaiveBayes	SMO	J48	Random forest
Accuracy	74.45	0.68	0.73	0.72
F-Measure	0.75	0.69	0.73	0.71
Root MSE	0.37	0.40	0.39	0.37
AUC	0.87	0.79	0.83	0.84

Table 12: Experiment result based on wrapper feature selection (COVID-19STD)

	NaiveBayes	SMO	J48	Random forest
Accuracy	74.45	0.68	0.73	0.72
F-Measure	0.75	0.69	0.73	0.71
Root MSE	0.37	0.40	0.39	0.37
AUC	0.87	0.79	0.83	0.84

Each newly obtained data set contains only the selected features from each algorithm and calculate overall accuracy, F-Measure, RMSE and AUC by 10-fold cross-validation as presented in Tables 11 and 12 for COVID-19STC and COVID-19STD datasets respectively.

Let's take a closer look at how good were the feature selection methods in choosing the best subset of features for better prediction. Figure 2 summarizes the essential features that result from the various feature selection methods.

We select features that have value greater than 0.3 for correlation coefficient, information gain and gain ratio, as shown in Table 7, in addition to the features selected using wrappers: NaiveBayes, SMO, J48 and Random Forest, which are presented in Table 11, with respect to COVID- 19STC dataset.

There are four factors: GDP, GDP Per Capital, E-Government Index and Smoking Habit that are highly selected by the selection methods. This gives an indication of the factors that effect Covid19 spread in terms of the total cases.

We have evaluated the performance of our feature selection processes using accuracy, F-Measure, RMSE and AUC. These metrics help us to examine whether the methods can correctly and efficiently recognize the optimized features and show us the effect of feature selection in the classification stage.

Table 13 presents the classification results depending on these four final selected features for COVID-19STC dataset.

It was observed from Table 13, that J48 gives highest accuracy of (73%), NaiveBayes and Random Forest were in the second place with (72%).

Figure 3 summarizes the essential features that result from the various feature selection methods for COVID-19STD dataset. For correlation coefficient, information gain and gain ratio, features that have value greater than 0.3 for correlation coefficient, information gain and gain ratio, as shown in Table 8, in addition to the features selected using wrappers: NaiveBayes, SMO, J48 and Random Forest, which are presented in Table 12.

The are two factors (Vis. GDP and E-Government Index) that are highly selected by the selection methods. That gives an indication of factors that effect Covid19 number of deaths. Table 14 present classification result depending on these four final selected features.

Table 14 presents the classification results using these two final selected features for COVID-19STD dataset. One can notice that J48 gives higher accuracy with (71%), Random Forest is in the second place with (68 %).

Table 13: The Classification results based on the final selected feature (COVID-19STC)

	NaiveBayes	SMO	J48	Random forest
Accuracy	0.72	0.62	0.73	0.72
F-Measure	0.73	0.62	0.73	0.72
Root MSE	0.40	0.42	0.40	0.37
AUC	0.88	0.75	0.78	0.86

Table 14: The Classification results based on the final selected features (COVID-19STD)

	NaiveBayes	SMO	J48	Random forest
Accuracy	0.58	0.58	0.71	0.68
F-Measure	0.56	0.55	0.69	0.674
Root MSE	0.46	0.44	0.40	0.40
AUC	0.79	0.69	0.74	0.81

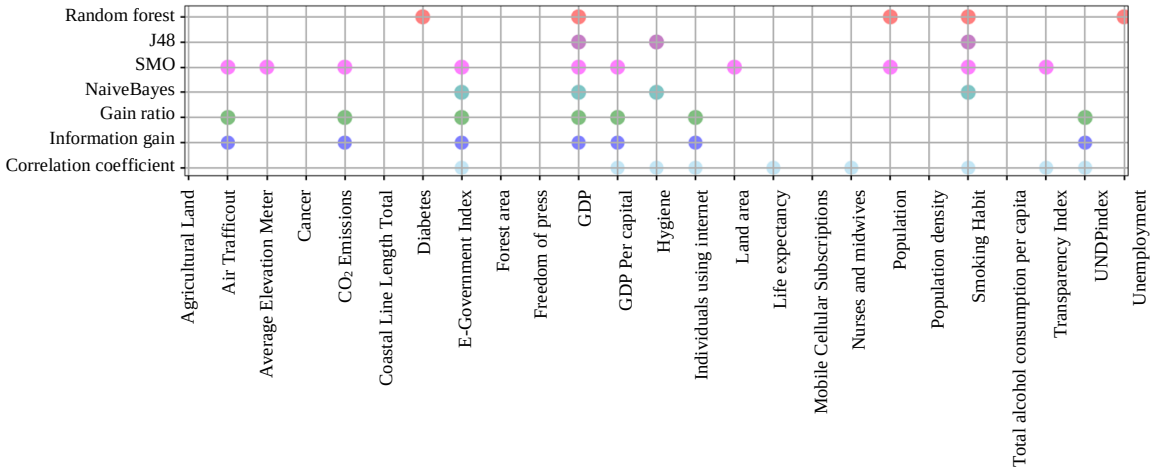

Fig. 2: Selected features from all methods that are related to COVID-19STC

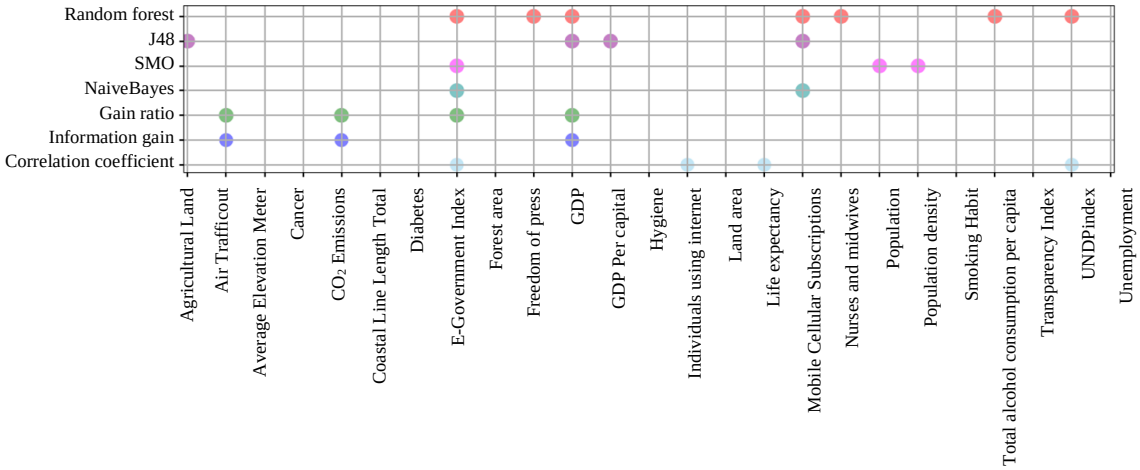

Fig. 3: Selected features from all methods that are related to COVID-19STD

Discussion

Results indicate there are four factors that gained sufficient weight to be considered of strong correlation with Covid19 spread. The weight of each factor is stipulated from observing that factor's appearing within the top result group of several approach angles used. Each one of these four factors appeared among the top results of at least 4 methods used (representing approach angles).

The four factors are: GDP, GDP Per Capita (ppp), eGovt. Development Index and Smoking:

- Finding of GDP as a Factor of Strong Correlation with Covid19 Spread
  Although the result may not sound intuitive at first glance, but it makes sense from different perspectives:
  - GDP is an economic indicator that represents a broad measure of overall domestic production. It functions as a comprehensive scorecard of the country's economic health. Thus, the higher the GDP the higher the national production activity is. Production activities would expectedly require a massive level of interactions at all tiers and throughout the entire value chain; from sourcing raw materials to finished products that involve processing, manufacturing and/or exchanging goods or services. The more human interactions there are, the higher the opportunity for Covid19 to spread
  - On another dimension, GDP is directly connected to exports, foreign trade (in industrial economies) and to tourism (in service economies) - both tie GDP to the influx of air traffic into and out of the country carrying visitors, labour and tourists, which are a major factor in spreading the disease by potential carries from abroad continuously mixing with population and increasing the likelihood of an outbreak
  - On a different level, GDP symbolizes the economic health of a country, which directly connects to that country's availability of sizable expenditure on vital sectors like the healthcare system. The better and more prepared the healthcare system, the higher Covid19 testing activities are going around. The higher the testing, the higher the numbers of positive cases registered
- Finding of GDP Per Capita as a Factor of Strong Correlation with Covid19 Spread GDP per capita is an economic indicator that represents a good measurement of a country's standard of living. It tells you how prosperous a country feels to each of its citizens as it reflects individual prosperity. Thus, result makes sense from two perspectives:
  - The higher GDP per capita the higher the prosperity on an individual level, which leaves the individual with a higher disposable income that forms a good motive to spend more on traveling according to this list[3] 3 that ranks the countries who travel most. The higher the likelihood of travel, the higher the likelihood of contracting Coronavirus
  - On the other hand, the higher the GDP per capita, the higher availability of Govt. expenditure on each individual's healthcare - leading to better availability of healthcare resources, among which is heavy testing of Coronavirus, that directly connects to the number of cases being discovered and announced (spread)
- Finding of EGovt. Development Index (EGDI) as a Factor of Strong Correlation with Covid19 Spread EGovt. Development Index (EGDI) is a measure of 3 elements: Telecommunication Infrastructure, Availability of Online Services, Human Capital. The higher the EGDI value the higher Covid19 Spread. This makes sense from two perspectives:
  - The higher the availability of online Govt. services, the higher the time saved that would've been otherwise spent obtaining the services offline, which leaves more leisure time for people to use on social activities, which increases the overall susceptibility to infection spread through human interactions
  - The better Telecommunication Infrastructure the higher the internet penetration in the society, leading to greater and faster exposure to misinformation and fake news that is associated with (and further intensifies) disease spread
- Finding of Smoking as a Factor of Strong Correlation with Covid19 Spread This result makes sense despite some unproven claims otherwise. WHO[4] officially announced that "There is currently insufficient information to confirm any link between tobacco or nicotine in the prevention or treatment of COVID-19" and stressed that "there are no peer-reviewed studies that have evaluated the risk of SARS-CoV-2 infection associated with smoking"
- However, WHO envisioned that "Tobacco smokers (cigarettes, waterpipes, bidis, cigars, heated tobacco products) may be more vulnerable to contracting COVID-19, as the act of smoking involves contact of fingers (and possibly contaminated cigarettes) with the lips, which increases the possibility of

[3] https://www.worldatlas.com/articles/countries-whose-citizenstravel-the-most.html
[4]https://www.who.int/news-room/detail/11-05-2020-who-statementtobacco-use-and-covid-19

transmission of viruses from hand to mouth. Smoking waterpipes, also known as shisha or hookah, often involves the sharing of mouth pieces and hoses, which could facilitate the transmission of the COVID-19 virus in communal and social settings"

Conclusion

In this study, we introduce two datasets, each of which consists of 25 country-level factors and covers 137 countries summarizing Geographic, Demographic, Economic, Healthcare System, Transportation, Technological, Social, Cultural, Religious and Political metrics. One of theme (COVID-19STC) aims to detect the increase of the total cases, whereas the other (COVID-19STD) aimed for total death detection. Then we analyzed the two datasets using different machine learning algorithms with various feature selection sachems. Four of these factors found to be able to create models comparable to those models created based of the twenty-five potential factors analyzed.

In the COVID-19STC dataset, the main features that are highly selected by the selection methods are GDP, GDP Per Capital, E-Government Index and Smoking Habit. This gives an indication of the factors that effect Covid19 spread in terms of the total cases. In the COVID-19STD dataset, GDP and E-Government Index were highly selected by the selection methods, which gives an indication of the factors that effect Covid19 number of deaths. GDP and GDP Per Capita are economic indicators that represent a good measurement of a country's standard of living, in addition to the higher production or export activities. Production and export activities would expectedly require a massive level of interactions. The more human interactions there are, the higher the opportunity for Covid19 to spread. Smoking Habit increases the possibility of transmission of viruses from hand to mouth. It often involves the sharing of mouth pieces and hoses, which could facilitate the transmission of the COVID-19 virus in communal and social settings. A natural future step would involve analyzing further factors, both based on proprietary data or privacy-protected data, such as patients' medical data, geographical location(s) and travel habits.

Acknowledgement

The authors acknowledge and thank Ameen Taha-Intern student-for his dedication and valuable contribution in Data gathering and data analysis.

Author's Contributions

Rana Husni Al Mahmoud: Participated in the creation of the model, carried out the experimental studies and participated in drafting the script.

Eman Omar: Participated in reviewing relevant litterateur the creation of the model and participated in drafting the script.

Khaled Taha: Participated in the creation of the model and revised the script.

Mahmoud Al-Sharif: Participated in the creation of the model and revised the script.

Abdullah Aref: Participated in the creation of the model and drafting the script.

All authors read and approved the final manuscript.

Ethics

All information provided in this study is confidential and unique. This paper has neither been published nor is under review elsewhere. There are no ethical issues associated with this research.

References

Abdulkareem, S. A., Augustijn, E. W., Filatova, T., Musial, K., & Mustafa, Y. T. (2020). Risk perception and behavioral change during epidemics: Comparing models of individual and collective learning. PloS one, 15(1), e0226483.

Anghel, M., Werley, K. A., & Motter, A. E. (2007, January). Stochastic model for power grid dynamics. In 2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07) (pp. 113-113). IEEE.

Bauch, C. T., Lloyd-Smith, J. O., Coffee, M. P., & Galvani, A. P. (2005). Dynamically modeling SARS and other newly emerging respiratory illnesses: past, present and future. Epidemiology, 791-801.

Boulos, M. N. K., & Geraghty, E. M. (2020). Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics.

Buizza, R. (2020). Probabilistic prediction of COVID-19 infections for China and Italy, using an ensemble of stochastically-perturbed logistic curves. arXiv preprint arXiv:2003.06418.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.

Chen, B., Liang, H., Yuan, X., Hu, Y., Xu, M., Zhao, Y., ... & Zhu, X. (2020a). Roles of meteorological conditions in COVID-19 transmission on a worldwide scale. MedRxiv.

Chen, B., Shi, M., Ni, X., Ruan, L., Jiang, H., Yao, H., ... & Ge, T. (2020b). Data visualization analysis and simulation prediction for covid-19. arXiv preprint arXiv:2002.07096.

Costa, L. D. F., Oliveira Jr, O. N., Travieso, G., Rodrigues, F. A., Villas Boas, P. R., Antiqueira, L., ... & Correa Rocha, L. E. (2011). Analyzing and modeling real-world phenomena with complex networks: a survey of applications. Advances in Physics, 60(3), 329-412.

DeCaprio, D., Gartner, J., Burgess, T., Kothari, S., & Sayed, S. (2020). Building a COVID-19 vulnerability index. arXiv preprint arXiv:2003.07347.

Diekmann, O., & Heesterbeek, J. A. P. (2000). Mathematical epidemiology of infectious diseases: model building, analysis and interpretation. John Wiley & Sons.

Fanelli, D., & Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons & Fractals, 134, 109761.

Fong, S. J., Li, G., Dey, N., Crespo, R. G., & Herrera-Viedma, E. (2020a). Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. Applied Soft Computing, 106282.

Fong, S. J., Li, G., Dey, N., Crespo, R. G., & Herrera-Viedma, E. (2020b). Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak. arXiv preprint arXiv:2003.10776.

Forna, A., Nouvellet, P., Dorigatti, I., & Donnelly, C. (2019). Case fatality ratio estimates for the 2013–2016 west african ebola epidemic: application of boosted regression trees for imputation. International Journal of Infectious Diseases, 79, 128.

Gao, Y., Zhang, Z., Yao, W., Ying, Q., Long, C., & Fu, X. (2020). Forecasting the cumulative number of COVID-19 deaths in China: a Boltzmann function-based modeling study. Infection Control & Hospital Epidemiology, 1-3.

Hamer, W. B., Birr, T., Verreet, J. A., Duttmann, R., & Klink, H. (2020). Spatio-Temporal Prediction of the Epidemic Spread of Dangerous Pathogens Using Machine Learning Methods. ISPRS International Journal of Geo-Information, 9(1), 44.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Hays, J. N. (2005). Epidemics and pandemics: their impacts on human history. Abc-clio.

Hilton, J., & Keeling, M. J. (2020). Estimation of country-level basic reproductive ratios for novel Coronavirus (COVID-19) using synthetic contact matrices. medRxiv.

Hsu, H. H., & Hsieh, C. W. (2010). Feature Selection via Correlation Coefficient Clustering. JSW, 5(12), 1371-1377.

Hu, Z., Ge, Q., Li, S., Jin, L., & Xiong, M. (2020). Evaluating the effect of public health intervention on the global-wide spread trajectory of Covid-19. medRxiv.

Imai, N., Dorigatti, I., Cori, A., Donnelly, C., Riley, S., & Ferguson, N. (2020). Report 2: Estimating the potential total number of novel Coronavirus cases in Wuhan City, China.

Ivanov, D. (2020). Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. Transportation Research Part E: Logistics and Transportation Review, 136, 101922.

Janecek, A., Gansterer, W., Demel, M., & Ecker, G. (2008, September). On the relationship between feature selection and classification accuracy. In New challenges for feature selection in data mining and knowledge discovery (pp. 90-105).

Jhuo, S. L., Hsieh, M. T., Weng, T. C., Chen, M. J., Yang, C. M., & Yeh, C. H. (2019, December). Trend prediction of influenza and the associated pneumonia in taiwan using machine learning. In 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) (pp. 1-2). IEEE.

Jia, L., Li, K., Jiang, Y., & Guo, X. (2020). Prediction and analysis of Coronavirus Disease 2019. arXiv preprint arXiv:2003.05447.

Jung, S. M., Akhmetzhanov, A. R., Hayashi, K., Linton, N. M., Yang, Y., Yuan, B., ... & Nishiura, H. (2020). Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: inference using exported cases. Journal of clinical medicine, 9(2), 523.

Kastner, J., Wei, H., & Samet, H. (2020). Viewing the progression of the novel corona virus (covid-19) with newsstand. arXiv preprint arXiv:2003.00107.

Kephart, J. O., & White, S. R. (1992). Directed-graph epidemiological models of computer viruses. In Computation: the micro and the macro view (pp. 71-102).

Khelil, A., Becker, C., Tian, J., & Rothermel, K. (2002). Directed-graph epidemiological models of computer viruses. In Proc. 5th ACM Int'l workshop on Modeling analysis and simulation of wireless and mobile systems (pp. 54-60).

Killeen, B. D., Wu, J. Y., Shah, K., Zapaishchykova, A., Nikutta, P., Tamhane, A., ... & Unberath, M. (2020). A County-level Dataset for Informing the United States' Response to COVID-19. arXiv preprint arXiv:2004.00756.

Kumar, J., & Hembram, K. P. S. S. (2020). Epidemiological study of novel coronavirus (COVID-19). arXiv preprint arXiv:2003.11376.

Legrand, J., Grais, R. F., Boelle, P. Y., Valleron, A. J., & Flahault, A. (2007). Understanding the dynamics of Ebola epidemics. Epidemiology & Infection, 135(4), 610-621.

Lei, S. (2012, March). A feature selection method based on information gain and genetic algorithm. In 2012 International Conference on Computer Science and Electronics Engineering (Vol. 2, pp. 355-358). IEEE.

Li, L., Yang, Z., Dang, Z., Meng, C., Huang, J., Meng, H., ... & Shao, Y. (2020a). Propagation analysis and prediction of the COVID-19. Infectious Disease Modelling, 5, 282-292.

Li, M., Chen, J., & Deng, Y. (2020b). Scaling features in the spreading of COVID-19. arXiv preprint arXiv:2002.09199.

Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., ... & Luo, B. (2020). Effects of temperature variation and humidity on the mortality of COVID-19 in Wuhan. medRxiv.

Machado, G., Vilalta, C., Recamonde-Mendoza, M., Corzo, C., Torremorell, M., Perez, A., & VanderWaal, K. (2019). Identifying outbreaks of Porcine Epidemic Diarrhea virus through animal movements and spatial neighborhoods. Scientific reports, 9(1), 1-12.

Marathe, M. V., & Ramakrishnan, N. (2013). Recent advances in computational epidemiology. IEEE intelligent systems, 28(4), 96-101.

Metz, C. E. (1978, October). Basic principles of ROC analysis. In Seminars in nuclear medicine (Vol. 8, No. 4, pp. 283-298). WB Saunders.

Mezzatesta, S., Torino, C., De Meo, P., Fiumara, G., & Vilasi, A. (2019). A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. Computer methods and programs in biomedicine, 177, 9-15.

Miao, J., & Niu, L. (2016). A survey on feature selection. Procedia Computer Science, 91, 919-926.

Molina, L. C., Belanche, L., & Nebot, À. (2002, December). Feature selection algorithms: A survey and experimental evaluation. In 2002 IEEE International Conference on Data Mining, 2002. Proceedings. (pp. 306-313). IEEE.

Pal, R., Sekh, A. A., Kar, S., & Prasad, D. K. (2020). Neural network based country wise risk prediction of COVID-19. arXiv preprint arXiv:2004.00959.

Pandey, M. K., & Karthikeyan, S. (2011). Performance analysis of time series forecasting of ebola casualties using machine learning algorithm. Proceedings ITISE, 201.

Peng, L., Yang, W., Zhang, D., Zhuge, C., & Hong, L. (2020). Epidemic analysis of COVID-19 in China by dynamical modeling. arXiv preprint arXiv:2002.06563.

Philemon, M. D., Ismail, Z., & Dare, J. (2019). A review of epidemic forecasting using artificial neural networks. International Journal of Epidemiologic Research, 6(3), 132-143.

Pigott, D. M., Golding, N., Mylne, A., Huang, Z., Henry, A. J., Weiss, D. J., ... & Bhatt, S. (2014). Mapping the zoonotic niche of Ebola virus disease in Africa. Elife, 3, e04395.

Poole, L. (2020). Seasonal Influences on the Spread of SARS-CoV-2 (COVID19), Causality and Forecastabililty (3-15-2020). Causality and Forecastabililty (3-15-2020)(March 15, 2020).

Priyadarsini, R. P., Valarmathi, M. L., & Sivakumari, S. (2011). Gain ratio based feature selection method for privacy preservation. ICTACT J Soft Comput, 1(04), 2229-6956.

Pyne, S., Vullikanti, A. K. S., & Marathe, M. V. (2015). Big data applications in health sciences and epidemiology. In Handbook of statistics (Vol. 33, pp. 171-202). Elsevier.

Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

Rao, A. S. S., & Vazquez, J. A. (2020). Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone–based survey when cities and towns are under quarantine. Infection Control & Hospital Epidemiology, 41(7), 826-830.

Sadilek, A., Kautz, H. A., & Silenzio, V. (2012, June). Modeling Spread of Disease from Social Interactions. In ICWSM (pp. 322-329).

Sajadi, M. M., Habibzadeh, P., Vintzileos, A., Shokouhi, S., Miralles-Wilhelm, F., & Amoroso, A. (2020). Temperature and latitude analysis to predict potential spread and seasonality for COVID-19. Available at SSRN 3550308.

Sau, A. (2017). A simulation study on hypothetical Ebola virus transmission in India using Spatiotemporal Epidemiological Modeler (STEM): a way towards precision public health. Journal of environmental and public health, 2017.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information processing & management, 45(4), 427-437.

Verdasca, J., Da Gama, M. T., Nunes, A., Bernardino, N. R., Pacheco, J. M., & Gomes, M. C. (2005). Recurrent epidemics in small world networks. Journal of Theoretical Biology, 233(4), 553-561.

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record, 31(1), 76-77.

Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet, 395(10225), 689-697.

Yan, L., Zhang, H. T., Xiao, Y., Wang, M., Sun, C., Liang, J., ... & Tang, X. (2020). Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. MedRxiv.

Yang, Z., Zeng, Z., Wang, K., Wong, S. S., Liang, W., Zanin, M., ... & Liang, J. (2020). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. Journal of Thoracic Disease, 12(3), 165.

Zeng, T., Zhang, Y., Li, Z., Liu, X., & Qiu, B. (2020). Predictions of 2019-ncov transmission ending via comprehensive methods. arXiv preprint arXiv:2002.04945.

Zhang, X., Ma, R., & Wang, L. (2020). Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. Chaos, Solitons & Fractals, 109829.

Zhao, S., Lin, Q., Ran, J., Musa, S. S., Yang, G., Wang, W., ... & Wang, M. H. (2020a). Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. International journal of infectious diseases, 92, 214-217.

Zhao, X., Liu, X., & Li, X. (2020b). Tracking the spread of novel coronavirus (2019-nCoV) based on big data. medRxiv.