NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 2022; 11(3), 484-495

Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Derleme makalesi / Review article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

A survey of recent studies on COVID-19 outbreak prediction using statistical and machine learning methods

İstatistiksel ve makine öğrenme yöntemleri ile COVID-19 salgın tahmini üzerine yapılan güncel çalışmaların incelemesi

Umut Ahmet Çetin[1], Fatih Abut[2,*]

[1,2] *Çukurova University, Engineering Faculty, Computer Engineering Department, 01330, Adana, Turkey*

Abstract

COVID-19 is an infectious disease first discovered in Wuhan City, China, in December 2019. Ever since, COVID-19 has infected more than 70 million people and caused more than 1 million deaths worldwide. There is a need for models that predict the COVID-19 outbreak as accurately as possible to combat such an infectious and deadly disease. By using the results of the prediction models, governments can make better decisions and control measures about the disease, such as arranging budget and facility planning to combat the disease, deciding on how many medicines and medical equipment should be produced or imported, and how much medical staff is going to be needed. Consequently, various regression and classification models have been proposed for time series or supervised prediction of the COVID-19 outbreak in several countries and continents. This study aims to give an overview of recent studies on predicting the COVID-19 outbreak utilizing statistical and machine learning methods. Particularly, for each study, we outline the utilized ground-truth dataset characteristics, the type of the developed models, the predictor variables, the statistical and machine learning methods, the performance metrics, and finally, the major conclusion. The survey results reveal that machine learning methods are promising tools for making predictions for various needs, such as predicting whether a patient is infected with COVID-19 or not, predicting the trend of COVID-19 outbreaks, or predicting which age groups are most affected by COVID-19.

Keywords: Machine learning, Classification, Regression, COVID-19, Coronavirus

Öz

COVID-19, ilk olarak Aralık 2019'da Çin'in Wuhan şehrinde ortaya çıkan bulaşıcı bir hastalıktır. O zaman beri COVID-19 dünya çapında 70 milyondan fazla insanı enfekte etmiştir ve 1 milyondan fazla ölüme neden olmuştur. Bu denli bulaşıcı ve ölümcül bir hastalıkla mücadele etmek için COVID-19 salgınını mümkün olduğunca doğru tahmin eden modellere ihtiyaç duyulmaktadır. Hükümetler tahmin modellerinin sonuçlarını kullanarak hastalıkla mücadele için bütçe ve tesis planlaması, ne kadar ilaç ve tıbbi ekipmanın üretileceğine veya ithal edileceği ve ne kadar tıbbi personele ihtiyaç duyulacağı hakkında daha iyi kararlar ve kontrol önlemleri alabilir. Sonuç olarak, çeşitli ülke ve kıtalarda COVID-19 salgınının zaman serileri veya denetimli tahmini için çeşitli regresyon ve sınıflandırma modelleri önerilmiştir. Bu makale, istatistiksel ve makine öğrenimi yöntemlerini kullanarak COVID-19 salgınını tahmin etmeye yönelik son çalışmalara genel bir bakış sunmayı amaçlamaktadır. Özellikle, her çalışma için, kullanılan veri kümesi özelliklerini, geliştirilen modellerin türünü, tahmin değişkenlerini, istatistiksel ve makine öğrenimi yöntemlerini, performans ölçümlerini ve son olarak ana sonucu ana hatlarıyla özetlenmiştir. Araştırma sonucu, makine öğrenme yöntemlerinin, COVID-19 salgını eğilimini tahmin etmek veya bir hastanın COVID-19 ile enfekte olup olmadığını tespit etmek gibi çeşitli ihtiyaçlar için tahminler yapmak için umut verici araçlar olduğunu ortaya koymaktadır.

Anahtar kelimeler: Makine öğrenimi, Sınıflandırma, Regresyon, COVID-19, Koronavirüs

1 Introduction

COVID-19 is an infectious disease caused by the coronavirus, also known as SARS-CoV-2. COVID-19 is a respiratory pathogen that was first discovered in Wuhan City, China, in December 2019 [1], and ever since, COVID-19 has infected more than 70 million people and caused more than 1 million deaths worldwide [2], which was recognized as a pandemic by WHO since March 11, 2020 [3].

To combat such an infectious and deadly disease, there is a need for models that predict the COVID-19 outbreak as accurately as possible so that governments can make better decisions about the disease. Some of such governmental decisions include (a) arranging a budget to combat disease as effectively as possible, (b) checking the sufficiency of facilities before becoming overfilled, (c) deciding on how many medicines and medical equipment should be produced or imported, and finally (d) how much medical staff is going to be needed.

Studies for developing prediction models for other diseases have been conducted in the past decades, such as swine fever, dengue fever, H1N1 flu, and influenza. But due to different characteristics and COVID-19's spread being on

* Sorumlu yazar / Corresponding author, e-posta / e-mail: fabut@cu.edu.tr
Geliş / Recieved: 17.11.2021 Kabul / Accepted: 22.04.2022 Yayımlanma / Published: 18.07.2022
doi: 10.28948/ngmuh.1025095

an unprecedented scale, measures were taken that had almost never been done, such as nationwide curfews, nationwide closing of business and public places, being required to wear a mask outside homes, turning public places like stadiums into temporary hospitals, and many similar measures.

There are two types of prediction models currently used to produce COVID-19 outbreak predictions: regression and classification models. Regression models predict a continuous quantity output, for example. In contrast, classification is a process of assigning data to appropriate classes, having either two-class (i.e., binary) or multi-class labels to assign data. The main difference between regression and classification models is that the output variable in the regression is numerical (or continuous) while that for classification is categorical. For regression and classification models, the output variable can be predicted in two different ways. The first way is to find a mapping function f to map the input variables x_1, x_2, …, x_n to the continuous output variable y with $f(x_1, x_2, \ldots, x_n) = y$. To such types of predictions, we are referring to as the regular prediction of COVID-19 outbreaks. Alternatively, the output variable can also be predicted based on time series data. Particularly, time series prediction is a process of predicting future data values through a sequence of time based on past values of the data. The principle of the time series prediction method is predicting changes in the future, assuming that values will change similar to how they changed in the past. By combining the regular and time series prediction with the regression and classification models, we obtain four types of models: regular regression models, regular classification models, time series-based regression models, and time series-based classification models.

This study aims to outline recent studies on predicting the COVID-19 outbreak utilizing statistical and machine learning methods. Particularly, for each study, we outline the utilized ground-truth dataset characteristics, the type of the developed models, the predictor variables, the statistical and machine learning methods, the performance metrics, and finally, the major conclusion.

The rest of the paper is organized as follows. Section 2 introduces the studies examining Eurasian cases. Section 3 presents the studies examining African cases. Section 4 outlines the studies examining American cases. Section 5 presents the studies examining intercontinental cases. Finally, the paper is concluded along with research gaps.

2 Studies examining Eurasian cases

Satrio et al. [4] proposed models to forecast COVID-19 spread in Indonesia. To build the models, Autoregressive Integrated Moving Average (ARIMA) and PROPHET algorithms were used. ARIMA is usually written in the notation ARIMA(p, d, q) to indicate its parameters, where the "p" value means the number of time lags, "d" value means the degree of difference, and "q" means the size of the average window. The dataset used in this study was obtained from the Kaggle website and consists of the serial number, observation date, province/state, country/region, last update, confirmed, recovered, and death columns. It originally had 27.618 rows, but after choosing only Indonesian data, only 81 rows remained. Then, they split the confirmed, recovery, and death into separate data frames because ARIMA only uses univariate data. Then, they checked whether the data is stationary or not because ARIMA performs better when data has a stable or consistent pattern over time. Their finding indicated that the data was not stationary. Then, they used log-scale transformation and time-shifting transformation to convert data into stationary form. In this study, time-series regression models were used where the past confirmed, recovered, and death counts were used as input variables to predict their future values. Coefficient of determination (R^2), MSE (Mean Square Error), MAE (Mean Absolute Error), and MFE (Mean Forecast Error) have been used as performance metrics. They evaluated a 30-day prediction period for both models from April 22, 2020, to May 21, 2020. It has been concluded that PROPHET generally performed better than ARIMA, but both models become more inaccurate as more days are forecasted.

Althnian et al. [5] conducted a study to forecast the susceptibility of individuals based on demographic data, including age, gender, nationality, and location. The dataset was obtained from the Saudi Ministry of Health. The dataset contains data from 229224 patients in Saudi Arabia between 02.03.2020 to 25.04.2020. The predictor variables are the city, confirmatory result, date of birth, gender, nationality, patient's subject id, result date, screening result, and test date. Date of birth was substituted by one of the following age groups: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80+. The city was replaced by 15 different Saudi regions. Because negative test results were overwhelmingly high, random oversampling was used for the positive class, and as a result, the resulting database included 50% positive and 50% negative test results. In this study, regular classification models were used, and the variables age, gender, nationality, and location were used as predictors of COVID-10 outbreak. In this study, MLP (Multi-Layer Perceptron), SVM (Support Vector Machine), DT (Decision Tree), and RF (Random Forest) methods were preferred. 5-fold cross-validation was used to train and test the models. Accuracy, precision, recall, F1 score, and AUC (Area under the Curve) metrics were used for performance evaluation. It has been concluded that the DT-based model performs best in terms of all performance metrics. It was also concluded that susceptibility could be predicted using only the mentioned predictor variables without any medical variables.

Ceylan [6] conducted a study to estimate the prevalence of COVID-19 in Italy, Spain, and France. The dataset was taken from the WHO website and consisted of data collected between 21.02.2020 and 15.04.2020 from Italy, Spain, and France. In this study, 45 samples were used to create a stable and effective time series ARIMA regression model. Various architectures of the ARIMA model are used, and confirmed cases were used as the predictor variable, whereas the number of total cases was predicted. Root Mean Square Error (RMSE), MAE, and Mean Absolute Percentage Error (MAPE) were used as performance metrics. It has been concluded that ARIMA (0, 2, 1), ARIMA (1, 2, 0), and ARIMA (0, 2, 1) are the most suitable models for Italy, Spain, and France, respectively.

Fang et al. [7] conducted a study to create a COVID-19 forecasting model to prevent and control COVID-19 outbreaks. The dataset is taken from the John Hopkins University Coronavirus resource center. The dataset contains data between 31.01.2020 and 30.05.2020. In this study, the number of confirmed recovered and death cases in Russia were used as variables. Time series regression models were developed based on the ARIMA forecasting method. To verify the fitting effect of the developed ARIMA model, the MAPE metric is preferred. The forecast model for confirmed cases yielded a MAPE value of 0.60, and it has been concluded that the model has high prediction accuracy and is robust. Forecast models for death and recovery yield MAPE values of 3.90 and 2.40, respectively, and it has also been reported that these models were relatively robust.

Fayyoumi et al. [8] proposed regular classification models to predict possible patients of COVID-19 using machine learning and statistical models. The utilized dataset was obtained by an online survey in which various eligible participants from Jordan participated, and 105 of those participants were selected for the study. For developing the statistical learning models, Logistic Regression (LogReg) is used, and for building the machine learning models, SVM and MLP were preferred. The dataset includes various predictor variables, including positive PCR test, age, gender, smoking, positive X-Ray, fever, breathing, diarrhea or vomiting, lack of smell, nasal congestion, dry cough, aches and pain, and sore throat. All models were tested employing 10-fold cross-validation. This study predicts whether the patient is infected with COVID-19 or not. The performance of the models was evaluated using several metrics, including accuracy, sensitivity, specificity, Geometric Mean (G_Mean), and precision. As a result, MLP has been chosen to be the best performing model in terms of the used performance metrics.

Gupta et al. [9] proposed a model for predicting the active, death, and cured rate of COVID-19 in India. The dataset contains the active, death, and cured rate for India between 20/01/2020 to 07/05/2020. They used the time series regression model in this study. To build the model, SVM, LinReg (Linear Regression), and prophet forecasting were used. The predictor variables used are active cases, death cases, and cured cases. Active cases, death cases, and cured cases were predicted as the output variables. MAE, MSE, and RMSE have been used as performance metrics. It has been concluded that the prophet forecasting model performed better than SVM and LinReg thanks to its features that are not available on the other two models, such as the seasonal component set. Furthermore, it has been reported that SVM has become the second-best prediction model among those three.

Önder [10] proposed time series regression models that aim to predict the COVID-19 trend. The utilized dataset was collected from the Turkey Ministry of Health as daily confirmed cases. The dataset contains information between March 16, 2020, and March 28, 2020. The models used are Richards Model, Generalized Logistic Growth Model, and the Sub-Exponential growth model. The confirmed case count was used as the predictor variable, whereas the reproductive number was predicted as the output variable. Adjusted R^2, RMSE, and AIC (Akaike information criterion) have been used as performance metrics. As a result of this study, the sub-exponential method with a scaling of growth parameter of 0.91 showed the best results among the three.

Pinter et al. [11] proposed a COVID-19 prediction model that utilizes hybrid machine learning methods. The dataset includes COVID-19 cases and death rates in Hungary, which consists of data collected between March 4, 2020, and April 28, 2020, taken from worldometers. For training the model, data collected between March 4, 2020, and April 19, 2020, was utilized, whereas the remaining data was used for validating the predicted results. They used the time series regression model in this study. Total cases and mortality rate were predicted as the target variables. The utilized methods included the Hybrid Multi-Layered Perceptron-Imperialist Competitive Algorithm (MLP-ICA) and Adaptive Network-Based Fuzzy Inference System (ANFIS). Training is done by using two scenarios, the first one (i.e., scenario 1) is trained using odd-numbered days, and the second (i.e., scenario 2) is trained using even-numbered days. The training of ANFIS is done by using three Membership Functions (MF) types: Triangular, Trapezoidal, and Gaussian. RMSE and MAPE were used for comparing these MF results, and it has been concluded that Gaussian MF has the highest performance. The first scenario provided the lowest RMSE results for ANFIS compared to scenario 2 using selected MF because scenario one is more suitable for COVID-19 prediction. Still, it has also been concluded that scenario 2 gives the highest performance for predicting mortality rate using the selected MF type. Also, according to their results, ten neuron MLP architecture yielded the best results for MLP-ICA using both scenarios. But for MLP-ICA, unlike ANFIS, scenario 2 produces the highest accuracy for predicting COVID-19 cases compared to scenario 1. For the mortality rates, neuron number 14 for scenario one and neuron number 18 for scenario 2 yielded the highest accuracy in integrating by ICA method compared to other architectures. Because of it, scenario 1 concluded as the most suitable scenario for mortality rates for MLP-ICA. Evaluations of models were performed by calculating R^2, MAPE, RMSE values. As a result, it has been concluded that different scenarios used for training models do not create any results that are too different from each other. Also, though both methods showed promising results, MLP-ICA surpassed ANFIS after 122.608 km^2. Also, Fars province is split into 36 counties, 93 districts, and 112 cities. The dataset was collected from the Iranian Ministry of Health and Medical Education (IMHME) on April 10, 2020. In this study utilizing data collected from WHO and IMHME, between February 25, 2020, and June 10, 2020, for active cases and from March 2, 2020, to June 10, 2020, for death cases, the growth rate of activity and death cases around the world, Iran, and Fars comparing their results with actual cases.

Pourghasemi et al. [12] examined risk factors of COVID-19 and tested the SVM model for mapping areas that have a high risk of human infection in Fars Province, Iran. Also, the growth trend of COVID-19 in Fars Province was analyzed and compared with the growth rate of Iran and various other

countries. They developed time series classification models in this study. The methods included SVM and ARIMA. SVM was used to create a risk map of the COVID-19 outbreak, and ARIMA models were used to examine the patterns of COVID-19 spread in the province. The predictor variables include MTCM (Minimum Temperature of Coldest Month), MTWM (Maximum Temperature of Warmest Month), PWM (Precipitation of Wettest Month), PDM (Precipitation of Driest Month), distance from roads, distance from mosques, distance from hospitals, distance from fuel stations, distance from bus stations, distance from banks, distance from bakeries, distance from attraction sites, distance from ATM's, human footprints, the density of cities, density of villages. For SVM, accuracy and AUC metric were used, and for ARIMA coefficient, standard error, t-statistics, and probability metrics were calculated. It has been concluded that climate factors have the least influence on COVID-19 risk mapping and, ATMs, distance from; attraction sites, fuel stations, mosques, and roads, MTCM and, the density of villages and cities have moderate influence, but distances from bus stations, distance from bakeries, and distance from hospitals had the strongest influence. As for ARIMA, it showed that COVID-19 is not expected to show any explosive process, but it seems infection will keep its general trend for several months.

Tamhane et al. [13] published a paper about creating a time series regression model that predicts the case number for the next 20 days. The dataset was obtained from John Hopkins University, consisting of dates between January 1, 2020, and May 25, 2020. SVM and Polynomial Regression (PR) machine learning methods are used to build the models, whereas the number of cases has been utilized as the predictor variable. First, the data is divided into train and test sets. MAE and MSE are used as performance metrics. This study demonstrated that machine learning methods are very promising and essential tools to deal with the crisis

3 Studies examining African cases

Gebretensae et al. [14] predicted the future trend of COVID-19 in Ethiopia using ARIMA. The dataset used in this study was obtained from the official website of the Ethiopian Public Health Institute and consisted of confirmed, recovered, and death cases in Ethiopia. The dataset contains data from March 13, 2020, to August 31, 2020. Models were built based on ARIMA using the time-series regression. After determining the suitable ARIMA order, the authors tried to find precise estimates of model parameters utilizing least squares as described by Box and Jenkins. Autocorrection Function 8ACF) and Partial Autocorrection (PACF) were used in confirmed cases to check if the data is stationary. To predict confirmed and recovered cases, ARIMA (0, 1, 5) and ARIMA (2, 1, 3) were selected, respectively, based on ACF and PACF graphs. RMSE, MAPE, and Bayesian information criteria (BIC) have been used as performance metrics. It has been concluded that both ARIMA (0, 1, 5) and ARIMA (2, 1, 3) are the best models for confirmed and recovered cases, respectively, and the confirmed and recovered cases will increase in Ethiopia in the next 60 days on a daily basis.

Marzouk et al. [15] proposed a study to forecast the COVID-19 outbreak prevalence in Egypt using deep learning algorithms. The dataset was recognized by the Egyptian Ministry of Health and Population and consists of daily confirmed, recovered, and death cases from Egypt. The dataset consists of data collected between February 14, 2020, and June 30, 2021, and spans over 503 days. The data were split into 90% training data and 10% testing data. In this study, time-series regression models were used. Long Short-Term Memory (LSTM), Convolutional neural network (CNN), and MLP were used to build models. The past values of confirmed, recovered, and death counts were used to predict their future values. RMSE and R^2 have been used as performance metrics. It has been concluded that LSTM models performed better than CNN and MLP models for 1 week ahead and 1 month ahead predictions because it can capture nonlinear patterns in the input data over time, and feedback connections characterize the LSTM network to propagate the data in the backward pass.

Ahmed [16] published a study examining the prediction of Ethiopian cases to predict the spread of COVID-19. The dataset is taken from the official GitHub repository of the John Hopkins University Center of System Science and Engineering. Only data concerning Ethiopia was used from this dataset. The dataset contains data recorded since 22.01.2020. The dataset was divided into training and testing sets. For training, 3/4 of the dataset was used, whereas the rest was used for testing. The time series regression model was developed based on SVM and PR. The predictor and output variables are confirmed, recovered, and death cases. MAE and MSE metrics were used for performance evaluation. As a result, it has been concluded that SVM performs better than PR in both confirmed, recovered, and death cases.

Djeddou et al. [17] predicted new COVID-19 cases in Algeria. The utilized dataset was obtained from the public health database of the Algeria health ministry. They build a time series regression model based on the Extreme Learning Machine (ELM) method. Cumulative confirmed COVID-19 cases, calculated COVID-19 new cases, and index day were used as input variables. The model gives the output of COVID-19 new cases as output. MSE, RMSE, MAE, Nash–Sutcliffe Coefficient of efficiency (NSE), the overall index of model performance (OI), and R^2 were preferred as performance metrics. It has been concluded that the ELM architecture is suitable for predicting new COVID-19 cases.

Saba et al. [18] created a model for short-term prediction of COVID-19 to help authorities make better decisions. The dataset was obtained from the Egyptian Ministry of Health and Population and consisted of data collected between March 1, 2020, and May 10, 2020. Time series regression models were developed based on ARIMA and Nonlinear Autoregressive Artificial Neural Networks (NARANN) algorithms. Reported and new cases were used as predictor and target variables, respectively. The model's performance was evaluated using the MAE, RMSE, R^2, coefficient of residual mass (CRM), and deviation ratio (RD) metrics. They concluded that NARANN performs better than ARIMA to predict the new COVID-19 case number.

Takele [19] proposed a time series regression model based on the ARIMA method to predict the prevalence of COVID-19 in East African Countries. The motivation of the study is to provide the public health institutions with reliable day-to-day predictions so that proper intervention strategies can be produced and an immediate and long-term prevalence trajectory of COVID-19 can be provided. The dataset was taken from the official GitHub repository of John Hopkins University and included data gathered between March 13, 2020, and June 30, 2020. The countries examined in this study are Ethiopia, Sudan, Djibouti, and Somalia. Confirmed cases and total confirmed cases are used both as the predictor and target variables, respectively. AIC, BIC (Bayesian Information Criterion), MAPE, Coefficient, and P-value were used as the model evaluation metrics. ARIMA (1, 2, 1) was chosen as the best model for predicting COVID-19 cases in Ethiopia, ARIMA (2, 1, 1) with drift was chosen as the best model for predicting COVID-19 cases in Djibouti, ARIMA (0, 2, 2) was chosen as the best model for predicting COVID-19 cases in Somalia, and finally, ARIMA (2, 2, 1) was chosen as the best model for predicting COVID-19 cases in Sudan.

Table 1. Overview of studies examining Eurasian cases.

Study	Year	Model	Method	Predictor Variables	Metrics	Target Variable
Satrio et al. [4]	2021	Time Series Regression	Prophet Forecasting Model	confirmed, recovered, death	R^2, MSE, MAE, MFE	confirmed, recovered death
Althnian et al. [5]	2020	Regular Classification	DT, MLP	age, gender, nationality, and location	accuracy, precision, recall, F1 score, AUC	susceptibility
Ceylan [6]	2020	Time Series Regression	ARIMA	confirmed cases	RMSE, MAE, MAPE	total cases
Fang et al. [7]	2020	Time Series Regression	ARIMA	confirmed, recovered, death	MAPE	confirmed, recovered, death
Fayyoumi et al. [8]	2020	Regular Classification	MLP	positive PCR test, age, gender, smoking, positive X-Ray, fever, breathing, diarrhea or vomiting, lack of smell, nasal congestion, dry cough, aches, and pain, sore throat	accuracy, sensitivity, specificity, G_Mean, Precision	positive or negative COVID-19
Gupta et al. [9]	2020	Time Series Regression	Prophet Forecasting Model	confirmed cases, death cases, recovered cases	MAE, MSE, RMSE	confirmed cases, death cases, recovered cases
Önder [10]	2020	Time Series Regression	Sub-Exponential Growth	confirmed cases	Adj. R^2, RMSE, AIC	reproductive number
Pinter et al. [11]	2020	Time Series Regression	MLP-ICA	daily cases, number of deaths	RMSE, MAPE, R^2	number of cases, mortality rate
Pourghasemi et.al. [12]	2020	Time Series Classification	SVM, ARIMA	MTCM, MTWM, PWM, PDM, distance from roads. distance from mosques, distance from hospitals, distance from fuel stations, distance from bus stations, distance from banks, distance from bakeries, distance from attraction sites, distance from ATM's, human footprints, the density of cities, the density of villages	accuracy, AUC, coefficient, standard error, t-statistics, probability	high or low-risk area
Tamhane et al. [13]	2020	Time Series Regression	PR, SVM	number of cases	MAE, MSE	number of cases

LSTM: Long Short-Term Memory, MSE: Mean Square Error, MAE: Mean Absolute Error, MFE: Mean Forecast Error, DT: Decision Tree MLP: Multi-Layer Perceptron, AUC: Area under the Curve, ARIMA: Autoregressive Integrated Moving Average, RMSE: Root Mean Square Error, MAPE: Mean Absolute Percentage Error, G_Mean: Geometric Mean, Adj. R^2: Adjusted Coefficient of determination, AIC: Akaike information criterion, MLP-ICA: Multi-Layered Perceptron-Imperialist Competitive Algorithm, R^2: Coefficient of determination, SVM: Support Vector Machines, MTCM: Minimum Temperature of Coldest Month, MTWM: Maximum Temperature of Warmest Month, PWM: Precipitation of Wettest Month, PDM: Precipitation of Driest Month, PR: Polynomial Regression

4 Studies examining American cases

Luo et al. [20] proposed a study to predict the future trend of COVID-19 in the US. The dataset was obtained from the WHO website, and the dataset was in time-series format. Due to the US did not initiate isolation and treatment measures between January and March, new confirmed cases from April 1, 2020, to September 30, 2020, were utilized as 19 or not. Precision, accuracy, recall, and AUC were used as performance metrics. It has been concluded that MLP exhibited the best performance compared to other methods.

Jojoa et al. [22] predicted COVID-19 spread in different countries of America. The dataset was downloaded from the open data repository of the European Union. Time series regression-based prediction models using MLP and SVM were used for building the models. This study uses confirmed COVID-19 cases as the predictor and output variable. As for performance measures, CP (Pearson's Correlation Coefficient), MAE, and Mean Percentage Error (MPE) were used. It has been concluded that MLP performs better when an optimization algorithm determines the hyperparameters, but when MLP does not perform well, it is necessary to use SVM instead. MLP showed better performance than SVM for Chile, Mexico, and the USA. However, SVM performs better than MLP for Brazil, Colombia, and Peru using the same performance metrics.

Moreau [23] conducted a study to predict the evolution of the COVID-19 pandemic. The used dataset stems from "Our World in Data Project". The dataset contains the daily modeling objects. The data were split into 90% training data and 10% testing data. To build models, LSTM and XGBoost were used using the time-series regression. The past values of confirmed cases were used as an input variable to predict the confirmed cases for the next 30 days. MAE, MSE, RMSE, and MAPE have been used as performance metrics. It has been concluded that LSTM performed better than XGBoost as LSTM has a lower MAPE value.

Santana et al. [21] conducted a study to assist the detection of COVID-19 based on early symptoms using machine learning. The dataset was created in Brazilian and includes data from 55.676 patients. This study developed regular classification models based on RF, SVM, MLP, K-Nearest Neighbors (KNN), DT, Gradient Boosting Machine (GBM), and XGBoost. Gender, sore throat, dyspnea, fever, cough, headache, taste disorder, olfactory disorder, and health professional were used as the predictor variable. This study predicts whether the patient is infected with COVID-19 or not. Precision, accuracy, recall, and AUC were used as performance metrics. It has been concluded that MLP exhibited the best performance compared to other methods.

Jojoa et al. [22] predicted COVID-19 spread in different countries of America. The dataset was downloaded from the open data repository of the European Union. Time series regression-based prediction models using MLP and SVM were used for building the models. This study uses confirmed COVID-19 cases as the predictor and output variable. As for performance measures, CP (Pearson's Correlation Coefficient), MAE, and Mean Percentage Error (MPE) were used. It has been concluded that MLP performs better when an optimization algorithm determines the hyperparameters, but when MLP does not perform well, it is necessary to use SVM instead. MLP showed better performance than SVM for Chile, Mexico, and the USA. However, SVM performs better than MLP for Brazil, Colombia, and Peru using the same performance metrics.

Moreau [23] conducted a study to predict the evolution of the COVID-19 pandemic. The used dataset stems from "Our World in Data Project". The dataset contains the daily and the total number of confirmed cases and deaths. This dataset includes data starting from February 26, 2020, which is the date of the first confirmed case in Brazil. The Weibull Distribution model using time series regression is proposed in this study. Daily confirmed cases and daily deaths were used as predictor variables, whereas the number of daily new cases and daily new deaths were predicted as output variables. A parameter called diagnostic-death lag is used because it has been previously demonstrated that there is a correlation between lethality rate and diagnostic-death lag. This parameter is obtained by the temporal distance between the peak of daily new cases and daily new deaths. In this study, four different scenarios were conducted. These scenarios were based on the daily number of new deaths. The first, second, third, and fourth scenarios consider 1250, 1500, 1750, and 200 daily new deaths, respectively, at the maximum turning point of the curve. R^2 was preferred as the performance evaluation metric. This study concluded that the first scenario produces the most optimistic results, and the fourth scenario gives the most pessimistic results among those four scenarios. Also, the second and third scenarios yielded results similar to each other and fitted the actual data for daily new cases. Still, because of their similar precision, it's impossible to choose any of these scenarios as the most probable. However, it's mentioned that the diagnostic-death lag value points to a prospective scenario that lies between 1750 and 2000 daily new deaths.

Silva et al. [24] conducted a study incorporating the exogenous climatic variables in forecasting models. The datasets for confirmed cases contain cumulative cases in 5 states of the USA and Brazil. The datasets contain data until the date of 28.04.2020. The Brazilian dataset was obtained from an API that collects information about COVID-19 from all Brazilian states. The USA dataset was retrieved from the official repository of Johns Hopkins University, whereas the USA climate database was obtained from the National Centers for Environmental Information (NCEI) from the National Oceanic and Atmospheric Administration. The Brazilian climate database was obtained from Instituto Nacional de Meteorologia. Minimum and maximum temperature and precipitation were used as exogenous climatic inputs for each model. Time series regression models based on Bayesian regression neural network (BRNN), Cubist Regression (CUBIST), KNN, QRF (Quantile Random Forest), SVR, and variational mode decomposition (VMD)-based models were proposed. COVID-19 cases, precipitation, maximum temperature, and minimum temperature were used as predictor variables. The output variable is the number of confirmed cumulative cases.

In the study, they first decomposed output variables into five IMFs by performing VMD. They chose a lag equal to two by the grid search and used it on the IMF, creating four input lags and exogenous inputs. Also, new data is split into training and test sets. In training, they adopted one-out-cross-validation with a time slice. Second, they trained each IMF with mentioned models, generated five prediction outputs, and named them VMD-BRNN, VMD-CUBIST, VMD-KNN, VMD-QRF, and VMD-SVR. They used a recursive strategy to develop multi-days ahead of COVID-19 case prediction. The model is fitted to do one day ahead prediction; the model uses this forecasting result as input and continues until the desired forecasting horizon is reached. Fourth, they used Improvement Percentage (IP), symmetric MAPE (sMAPE), and relative RMSE (RRMSE) as performance evaluation metrics. They ranked each of these models for Brazilian states and USA states. For Brazilian states, ranking was VMD-CUBIST, VMD-BRNN, SVR, CUBIST, VMD-SVR, BRNN, VMD-QRF, QRF, VMD-KNN, and KNN. For USA states ranking were VMD-CUBIST, BRNN, CUBIST, SVR, VMD-BRNN, VMD-SVR, VMD-QRF, QRF, KNN, and VMD-KNN. They also mentioned that six-day-ahead hybrid models were more suitable tools than non-decomposed models and also noted that climatic variables indeed influence the accuracy when predicting COVID-19 cases.

Souza et al. [25] conducted a study to make a prognosis or early identification of COVID-19 patients at increased risk of developing severe symptoms. The dataset is obtained from the Espírito Santo state portal. The dataset was made up of 13690 patients who were tested positive for COVID-19. Only closed cases due to death and recovery were used in this study because the objective of their study was to predict the outcome of the disease. In total, 4826 patient data were used for the training model, and 3617 patient data were used for validation. In this study, regular classification models were developed based on LogReg, Linear Discriminant Analysis, Naïve Bayes, K-Nearest Neighbors, Decision Trees, XGBoost, and SVM. Fever, respiratory distress, cough, runny nose, sore throat, diarrhea, headache, pulmonary disease, cardiac disease, kidney, diabetes, smoking, obesity, hospitalization, and "is patient older than 60" were used as predictor variables. The target variable being predicted indicates whether a patient is died or recovered from COVID-19. Receiver Operating Characteristic (ROC) AUC, Precision-Recall Curve (PR) AUC, precision, recall, and F1 score were used as performance evaluation metrics. Two different experiments were done to evaluate models. As a result of this study, it has been concluded that the best performing model regarding ROC AUC and PR AUC metrics was LogReg, Linear Discriminant Analysis, XGBOOST, and SVM.

Wollenstein-Betech et al. [26] proposed regular classification models to predict the following events: Hospitalization, mortality, need for Intensive Care Unit (ICU), and need for a ventilator. The Mexican government has provided the utilized dataset to the general public. This dataset contains demographic information such as age, nationality, location, the use of an indigenous language, and pre-existing conditions of patients such as diabetes, chronic obstructive pulmonary disease, immunosuppression, pregnancy, asthma, hypertension, obesity, chronic renal failure, "is using a tobacco" and other prior diseases. The dataset contains data of 91179 patients. 20737 of those patients have positive COVID-19 tests, whereas 15445 of them are waiting for the result of the test. In this study, negative COVID-19 test results were not used for training the models. This study used sparse SVMs, sparse LogReg, Random Forests, and XGBoost models. Accuracy, weighted F1 score, and AUC were used as performance evaluation metrics. It was concluded that according to SVMs and LogReg models, the features that contribute the most to predicting the hospitalization of the patient were age, gender, chronic renal insufficiency, diabetes, immunosuppression, and pregnancy. The other variables have a smaller impact on the prediction model. The most important features for predicting mortality were age, test status, immunosuppression, and pregnancy. Similarly, the most important features for predicting ICU need were the development of pneumonia (if available), cardiovascular disease, asthma, and test results. Finally, the most important features for predicting the need for a ventilator were: ICU and pneumonia (if available), age, gender, cardiovascular disease, obesity, pregnancy, and the test results.

5 Studies examining intercontinental cases

Ayoobi et al. [27] proposed models to forecast the COVID-19 outbreak and perform an in-depth comparison of Gated Recurrent Unit (GRU), LSTM, Convolutional Long Short-Term Memory (Conv-LSTM) with their bidirectional extensions. The dataset was obtained from the WHO website and contained columns such as date reported, country code, country, region, cumulative cases, and cumulative deaths. The data from Australia and Iran has been used to develop the models. The data for Australia ranges from January 25, 2020, to August 19, 2020, and the data for Iran is between January 3, 2020, to October 6, 2020. The data were split into approximately 70% training data and %30 testing data. To build models, GRU, LSTM, Conv-LSTM, and bidirectional extensions of these three methods have been applied using the time-series regression. In this study, new cases, cumulative cases, new deaths, and cumulative deaths have been used as predictors, and new cases and new deaths are predicted in one, three, and seven days ahead during the next hundred days. Mean Squared Log Error (MSLE), MAPE, Root Mean Squared Log Error (RMSLE), and Explained Variance (EV) have been used as performance metrics. It has been concluded that most of the time, bidirectional models outperformed their non-bidirectional counterparts. Also, it has been reported that no technique always produces the best predictions, i.e., the best- performing prediction method changes according to one, three, and seven days ahead prediction scenarios.

Bala [28] developed a model that predicts COVID-19 outbreaks in any particular country. The dataset is obtained from the "All our World in Data COVID-19 (OWID)" dataset on confirmed cases and confirmed deaths.

Table 2. Overview of studies examining African cases

Study	Year	Model	Method	Predictor Variables	Metrics	Target Variable
Gebretensae et al. [14]	2021	Time Series Regression	ARIMA	confirmed, recovered	RMSE, MAPE, BIC	confirmed, recovered
Marzouk et al. [15]	2021	Time Series Regression	LSTM	confirmed, recovered, death	RMSE, R^2	confirmed, recovered, death
Ahmed [16]	2020	Time Series Regression	SVM	confirmed, recovered, death	MAE, MSE	confirmed, recovered, death
Djeddou et al. [17]	2020	Time Series Regression	ELM	new COVID-19 cases	MSE, RMSE, MAE, NSE, OI, R^2	new cases
Saba et al. [18]	2020	Time Series Regression	NARANN	reported cases	MAE, RMSE, R^2, CRM, RD	new cases
Takele [19]	2020	Time Series Regression	ARIMA	confirmed cases	AIC, BIC, MAPE, coefficient, P-value	total confirmed cases

ARIMA: Autoregressive Integrated Moving Average LSTM: Long Short-Term Memory, SVM: Support Vector Machines, RMSE: Root Mean Square Error, MAPE: Mean Absolute Percentage Error, BIC: Bayesian information criterion, R^2: Coefficient of determination, MAE: Mean Absolute Error, MSE: Mean Square Error, ELM: Extreme Learning Machines, NSE: Nash–Sutcliffe Coefficient of efficiency, OI: The overall index of model performance, NARANN: Nonlinear Autoregressive Artificial Neural Networks, CRM: coefficient of residual mass, RD: deviation ratio, AIC: Akaike information criterion

Table 3. Overview of studies examining American cases

Study	Year	Model	Method	Predictor Variables	Metrics	Target Variable
Luo et al. [20]	2021	Time Series Regression	LSTM	confirmed cases	MAE, MSE, RMSE, MAPE	confirmed cases
Santana et al. [21]	2021	Regular Classification	MLP	gender, sore throat, dyspnea, fever, cough, headache, taste disorder, olfactory disorder, health professional	Precision, Accuracy, Recall, AUC	positive or negative COVID-19
Jojoa et al. [22]	2020	Time Series Regression	MLP, SVM	confirmed cases	CP, MAE, MPE	confirmed cases
Moreau [23]	2020	Time Series Regression	Weibull distribution	confirmed new cases, confirmed new deaths	R^2	daily new cases, daily new deaths
Silva et al. [24]	2020	Time Series Regression	BRNN, CUBIST, KNN, QRF, SVR, VMD-based models	COVID-19 cases, precipitation, maximum temperature, minimum temperature	IP, sMAPE, RRMSE	cumulative confirmed cases
Souza et al. [25]	2020	Regular Classification	LogReg, LDA, XGBoost, SVM	fever, respiratory distress, cough, runny nose, sore throat, diarrhea, headache, pulmonary disease, cardiac disease, kidney, diabetes, smoking, obesity, hospitalization, is patient older than 60 age, pregnant, chronic renal insufficiency,	ROC AUC, PR AUC, precision, recall, F1 score	is patient recovered or dead
Wollenstein-Betech et al. [26]	2020	Regular Classification	SVM, LogReg, RF, XGBoost	diabetes, immunosuppression, chronic obstructive pulmonary disease, obesity, other, hypertension, tobacco use, cardiovascular disease, asthma, gender, ventilator, ICU, test result	Accuracy, weighted F1 score, AUC	hospitalization, mortality, ICU need, ventilator need

LSTM: Long Short-Term Memory, MAE: Mean Absolute Error, MSE: Mean Square Error, RMSE: Root Mean Square Error, MAPE: Mean Absolute Percentage Error, MLP: Multi-Layer Perceptron, AUC: Area under Curve, SVM: Support Vector Machines, CP: Pearson's correlation coefficient, MPE: Mean Percentage Error, R^2: Coefficient of determination, BRNN: Bayesian regression neural network, CUBIST: Cubist Regression, KNN: k-nearest neighbors, QRF: Quantile Random Forest, SVR: Support Vector Regression, VMD: Variational Mode Decomposition, IP: Improvement Percentage, sMAPE: Symmetric Mean Absolute Percentage Error, RRMSE: Relative Root Mean Square Error, LogReg: Logistic Regression, RF: Random Forest, LDA: Linear Discriminant Analysis, ROC: Receiver Operating Characteristic, PR AUC: Precision-Recall Curve AUC, ICU: Intensive Care Unit

This dataset is updated daily and published by European CDC under Oxford University. Time series regression models were developed to predict the total number of cases using the date and total cases as predictor variables. This database was split into a 70:30 ratio for training and testing. LinReg, SVM, RF, and XGBoost Regression were used to build the models. MAE and RMSE were used as performance evaluation metrics. It has been concluded that the best performing model is based on XGBoost followed by RF, SVM, and LinReg.

Hassan et al. [29] proposed a study to develop a COVID-19 prediction model that predicts COVID-19 outbreak to contribute to humanity getting rid of the COVID-19 pandemic. The dataset was obtained from the GitHub repository of John Hopkins University. It consisted of worldwide daily confirmed, recovered, and death cases data collected between the start of the pandemic and October 4, 2020. The dataset has been divided into two subsets: a training set (220 days) to train models and a testing set (39 days). MLP, SVM, Bayesian Network (BN), PR, and Linear Regression (LinReg) were applied to build models using the time-series regression. The past values of confirmed, recovered, and death counts were used as input variables to predict their future variables. RMSE, MAE, MSE, EV, and R^2 have been used as performance metrics. It has been concluded that, in general, MLP yielded better results in the prediction of confirmed, recovered, and death cases.

Yu et al. [30] created an online artificial intelligence system to analyze the dynamic trend of COVID-19 called COVID-19 Pandemic AI System (CPAIS). Two datasets were utilized: the first one is from Oxford COVID-19 Government Response Tracker (OxCGRT), and the other one is from John Hopkins University Center for Systems Science and Engineering. Oxford dataset contains government responses to COVID-19 based on a variety of parameters ever since January 1, 2020, and includes 20 types of items. Their items can be grouped as containment and closure policies, economic policies, health system policies, and miscellaneous policies. The John Hopkins University dataset contains confirmed, recovered, and death cases containing data for 192 countries ever since January 21, 2020. ARIMA, Feedforward Neural Network (FNN), MLP, and LSTM were used to build models using the time-series regression. The policies used from the OxCGRT dataset with confirmed, recovered, and death cases from John Hopkins were used as predictors, and they used 1-year records from the dataset and used the last 14 days for validation. ME (Mean Error), RMSE, MAE, MPE, and MAPE have been used as performance metrics. It has been concluded that LSTM yielded better results than other methods in most countries with better accuracy. ARIMA, FNN, and MLP were not stable enough and only yielded competitive results in some specific countries.

Ardabili et al. [31] published another study for COVID-19 outbreak prediction using machine learning. They compared several machine learning models using various models for mathematical forecasting to present a comparative analysis of soft computing and machine learning models to predict the COVID-19 outbreak. The data consisted of population counts collected from Italy, China, Iran, Germany, and the USA over 30 days. They used the time series regression models. Their comparison of ML optimizer GWO (Gray Wolf Optimizer) outperformed other ML optimizer methods with the best accuracy because it had the smallest RMSE and the most significant correlation coefficient. In their comparison of equations, the Logistic model outperformed other models in general in terms of accuracy, which also had the smallest RMSE and largest r-square value. They predict the total number of cases. The training was done using MLP and ANFIS using different neuron numbers for MLP and other MF (Membership Functions) for ANFIS. After comparison, MLP showed the best accuracy.

Prakash et al. [32] conducted a study to find which age groups were most affected by COVID-19 using machine learning algorithms. Two datasets were used, and both were obtained from the Kaggle. The first dataset contained data obtained from India, and the second one includes global COVID-19 data. The datasets contain age, state, confirmed Indian national cases, confirmed foreign national cases, cured, deaths, confirmed, date, and month. Regular classification models were developed based on DT, MLR (Multi-LinReg), SVM, XGBoost Classifier, RF Classifier, RF Regressor, KNN+NCA, Gaussian Naïve Bayes (GNB) Classifier, and LogReg. The test dataset ratio is 7:3. The performance evaluation metrics are R^2 and accuracy. It has been reported that 20-30, 30-40, and 40-50 age groups are affected most by COVID-19, and the most successful machine learning methods are the RF Regressor and RF Classifier.

Punn et al. [33] predicted the behavior and reachability of COVID-19 by developing time series regression models using various machine learning models. The dataset was retrieved from the official repository of Johns Hopkins University, which includes data collected between January 22, 2020, and April 1, 2020. They used various predictor variables, including province/state, country/region, last update, confirmed, death, and recovered cases. Training and testing are done using real-time data, which uses the number of confirmed, recovered, and death cases as the label for the matching date. They predict the confirmed, recovered, and death cases. To predict the trend of COVID-19, SVR, PR, DNN (Deep Neural Network), and LTSM with worldwide data have been used. The RMSE metric has been preferred to evaluate the performance of the models. As a result, the PR approach has been reported as the best fit to follow the growing trend.

Rustam et al. [34] conducted a study on the future prediction of COVID-19, focusing on the number of new cases, deaths, and recoveries to demonstrate the capability of ML models to predict the upcoming patients affected by COVID-19. The dataset was obtained from the GitHub repository provided by the Center for Systems Science and Engineering, John Hopkins University.

Table 4. Overview of studies examining intercontinental cases

Study	Year	Model	Method	Predictor Variables	Metrics	Target Variable
Ayoobi et al. [27]	2021	Time Series Regression	GRU, Bi-GRU, LSTM, Bi-LSTM, Conv-LSTM, Bi-Conv-LSTM	new cases, cumulative cases, new deaths, cumulative deaths	MSLE, MAPE, RMSLE, EV	new cases, new deaths
Bala [28]	2021	Time Series Regression	XGBoost	date, total cases	MAE, RMSE	total cases
Hassan et al. [29]	2021	Time Series Regression	MLP	confirmed, recovered, death	RMSE, MAE, MSE	confirmed, recovered, death
Yu et al. [30]	2021	Time Series Regression	LSTM	government policies, confirmed, recovered, death	ME, RMSE, MAE, MPE, MAPE	number of cases
Ardabili et al. [31]	2020	Time Series Regression	MLP	number of cases	RMSE, R	number of cases
Prakash et al. [32]	2020	Regular Classification	RFC, RFR	Age, cured, date, month, confirmed foreign national, confirmed Indian national, deaths, state	R^2, Accuracy	most affected age group
Punn et al. [33]	2020	Time Series Regression	PR	confirmed, death, and recovered cases	RMSE	confirmed cases, recovered cases, death cases
Rustam et al. [34]	2020	Time Series Regression	ES	recovery, death, newly confirmed cases	MSE, MAE, RMSE, R^2 Score, R^2 Adjusted	confirmed cases, recovered cases, death cases
Sahin [35]	2020	Time Series Regression	LinReg, GPR, SVM, DT, EL	mobility index, population	MAPE	daily cases
Tuli et al. [36]	2020	Time Series Regression	Robust Weibull Fit	number of cases, mortality rate, new cases	MSE, R^2, MAPE	number of cases, mortality rate, new cases

GRU: Gated Recurrent Unit, LSTM: Long Short-Term Memory, Conv-LSTM: Convolutional Long Short-Term Memory, MSLE: Mean Squared Log Error, MAPE: Mean Absolute Percentage Error, RMSLE: Root Mean Squared Log Error, EV: Explained Variance MAE. Mean Absolute Error, RMSE: Root Mean Square Error, MSE: Mean Square Error, ME: Mean Error, MPE: Mean Percentage Error, MLP: Multi-Layer Perceptron, R=Coefficient of Correlation, RFC: Random Forest Classifier, RFR. Random Forest Regressor, R^2: Coefficient of Determination, PR: Polynomial Regression, ES: Exponential Smoothing, SVR: Support Vector Regression, GPR: Gaussian Process Regression, SVM: Support Vector Machines, DT. Decision Tree, EL: Ensemble Learning

The dataset contains confirmed cases, deaths, and recoveries in the past number of days since the pandemic started. After the initial data preprocessing step, the dataset has been divided into two subsets: a training set (56 days) to train models and the testing set (10 days). Several time series regression models were built based on LinReg, Least Absolute Shrinkage and Selection Operator (LASSO), SVM, and Exponential Smoothing (ES). To evaluate the performance of the models, R^2, R^2 Adjusted, MAE, MSE, and RMSE metrics were used. Future predictions of new cases, deaths, and recoveries were made, and as a result, ES has been reported to outperform all other models in both new cases, deaths, and recoveries. Even after bigger datasets were used, ES still outperformed other methods in accuracy, but other methods exhibited significantly improved performance after using bigger datasets.

Sahin [35] created a COVID-19 forecasting model for predicting daily cases of COVID-19 based on mobility data. The dataset of mobility data was taken from Apple Mobility Trends Reports. The training model is based on data collected between 13.01.2020 and 10.05.2020. The testing models are based on data collected between 13.05.2020 and 20.05.2020. Seven countries were analyzed, including Brazil, France, Germany, Italy, Spain, the UK, and the USA. The mobility indices and populations of the countries were used as inputs. Time series regression models were developed based on LinReg, Gaussian Process Regression, SVM, Decision Tree, and Ensemble Learning. MAPE was preferred to evaluate the model's performance. It has been concluded that it is impossible to predict daily cases in all countries based on mobility index through machine learning methods.

Tuli et al. [36] proposed a real-time, more realistic prediction model that works on a cloud-based computer. The dataset is obtained from "Our World in Data by Hannah Ritchie, "updated daily from WHO situation reports. For this study, three instance single-core Azure B1's virtual machines with 1 GB RAM, SSD Storage, and 64-bit Windows Server 2016 have been used. To do multiple analysis tasks for predicting various metrics, the HealthFog framework leveraging the FogBus was employed. Robust Weibull Fit and Gaussian Fit were employed to develop time series regression models. The number of cases, new cases, and mortality rate are both input and output variables. MSE, R^2, and MAPE are used as performance evaluation metrics. It has been concluded that the Robust Weibull model works better than the Gaussian Fit model as performance evaluation metrics. It has been concluded that the Robust Weibull model works better than the Gaussian Fit model.

6 Conclusion and research gaps

This study outlined recent studies on predicting the COVID-19 outbreak utilizing statistical and machine learning methods. The survey results reveal that many studies have been conducted for various needs in recent years, such as predicting whether a patient is infected with COVID-19 or not, predicting the trend of COVID-19 outbreaks, or predicting which age groups are most affected by COVID-19. A variety of prediction models with different types have been developed that can be categorized into four groups: regular regression models, regular classification models, time series-based regression models, and time series-based classification models.

Representative examples of variables for the prediction of COVID-19 outbreak range from demographical variables, such as age, gender, nationality, and location, to pre-existing condition variables, such as asthma, obesity, hypertension, tobacco smoking, chronic renal insufficiency, diabetes, pregnancy, to case count variables, such as the number of confirmed, recovered, death cases. Also, current health variables, such as fever, breathing, diarrhea or vomiting, lack of smell, nasal congestion, dry cough, headache, or sore throat, play an important role in predicting the COVID-19 infection. The most frequently used performance metrics for regression models are R^2, MAE, MSE, MAPE, and RMSE; whereas the performance of classification models is evaluated using accuracy, precision, recall, and F1 score. In general, it's not possible to have a ranking between these methods, but, in most models, the MLP method has pretty good results and could be used in new studies. The most used variables are case counts, recovered, and deaths.

A possible future research area is to investigate the effects of COVID-19 mutations on the degree of the outbreak. Furthermore, the survey results show that most studies use a very small timeframe to develop the forecasting models. New studies can be started that consider larger timeframes such as 6- or 9-month long periods to test whether more accurate models can be built. Also, new studies can consider the effects of new vaccines. Finally, much more feature selection-based studies can be undertaken to identify the relevant indicators of the COVID-19 disease.

Acknowledgment

The authors would like to thank Çukurova University Scientific Research Projects Center for supporting this work under grant no FYL-2021-14257.

Conflict of interest

The authors declare that there is no conflict of interest.

Similarity rate (iThenticate): 29%

References

[1] What is COVID-19, https://www.who.int/emergencies /diseases/novel-coronavirus-2019/ques-tion-and-ans wers-hub/q-a-detail/coronavirus-disease-COVID-19, Accessed December 14, 2020.

[2] Worldometer COVID-19 Count, https://www.worldo-meters.info/coronavirus/?, Accessed December 14, 2020.

[3] WHO Director-General's opening remarks at the media briefing on COVID-19 - March 11, 2020, https://www. who.int/ director-general/speeches /detail/ -who-director-general-s-opening-remarks-at-the-medi a-brie fing-on-COVID-19---11-march-2020, Accessed Dece mber 14, 2020.

[4] C. B. A. Satrio, W. Darmawan, B. U. Nadia and N. Hanafiah, Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET, Procedia Computer Science, 179, 524-532, 2021. https://doi.org/10.1016/j.procs.2021.01.036

[5] A. Althnian, A. A. Elwafa, N. Aloboud, H. Alrasheed and H. Kurdi, Prediction of COVID-19 Individual Susceptibility using Demographic Data: A Case Study on Saudi Arabia, Procedia Computer Science, 177, 379-386, 2020. https://doi.org/-10.1016/j.procs.2020.1 0.051

[6] Z. Ceylan, Estimation of COVID-19 prevalence in Italy, Spain, and France, Science of The Total Environment, 729, 138817, 2020. https://doi.org/10.10 16/j.scitotenv.2020.138817

[7] L. Fang, D. Wang and G. Pan, Analysis and Estimation of COVID-19 Spreading in Russia Based on ARIMA Model. SN Compr. Clin. Med. 2, 2521–2527, 2020. https://doi.org/10.1007/s42399-020-00555-y

[8] E. Fayyoumi, S. Idwan and H. AboShindi, Machine Learning and Statistical Modelling for Prediction of Novel COVID-19 Patients Case Study: Jordan. International Journal of Advanced Computer Science and Applications. 11(5), 122-126, 2020. https://doi.org. /10.14569/IJACSA.2020.0110518

[9] A. K. Gupta, V. Singh, P. Mathur and M. C. Travieso-Gonzalez, Prediction of COVID-19 pandemic measuring criteria using support vector machine, prophet and LinReg models in Indian scenario, Journal of Interdisciplinary Mathematics, 24(1), 89-108, 2020. doi:10.1080/09720502.2020.1833458

[10] H. Önder, Short-term forecasts of the COVID-19 epidemic in Turkey: March 16–28, Black Sea Journal of Health Science, 3(2), 27-30, 2020.

[11] G. Pinter, I. Felde, A. Mosavi, P. Ghamisi and R. Gloaguen, COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach. Mathematics, 8, 890, 2020.

[12] H. R. Pourghasemi, S. Pouyan, Z. F. N. Sadhasivam, B. Heidari, S. Babaei and J. P. Tiefenbacher, 2020. https://doi.org/10.1371/journal.pone.0236238

[13] R. Tamhane and S. Mulge, "Prediction of COVID-19 Outbreak using Machine Learning". In International Research Journal of Engineering and Technology (IRJET), 7:5, 2020.

[14] Y. A. Gebretensae and D. Asmelash, Trend Analysis and Forecasting the Spread of COVID-19 Pandemic in Ethiopia Using Box–Jenkins Modeling Procedure. Int J Gen Med. 14, 1485-1498, 2021. https://doi.org/10.2147 /IJGM.S306250

[15] M. Marzouk, N. Elshaboury, A. Abdel-Latif and S. Azab, Deep learning model for forecasting COVID-19 outbreak in Egypt, Process Safety and Environmental

Protection, 153, 363-375, 2021. https://doi.org/10.1016/j.psep.2021.07.034
[16] S. Z. Ahmed "Analysis and forecasting the outbreak of COVID-19 in Ethiopia using Machine learning". European Journal of Computer Science and Information Technology, 8(4), 1-13, 2020.
[17] M. Djeddou, I. A. Hameed, A. Nejatian and I. Loukam, Predictive Modelling of COVID-19 New Cases in Algeria using An Extreme Learning Machines (ELM) medRxiv 2020.09.28.20203299. doi: https://doi.org/10.1101/2020.09.28.20203299
[18] A. I. Saba and A. H. Elsheikh, Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks, Process Safety and Environmental Protection, 141, 1-8, 2020. https://doi.org/10.1016/j.psep.2020.05.029
[19] R. Takele, Stochastic modelling for predicting COVID-19 prevalence in East Africa Countries, Infectious Disease Modelling, 5, 598-607, 2020. https://doi.org/10.1016/j.idm.2020.08.005
[20] J. Luo, Z. Zhang, Y. Fu, F. Rao, Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms, Results in Physics, 27, 104462, 2021. https://doi.org/10.1016/j.rinp.2021.104462
[21] V. dos S. Santana et al., A Machine Learning Models for COVID-19 Detection in Brazil Based on Symptoms JMIR Preprints 25/01/2021:27293. doi: 10.2196/preprints.27293
[22] M. Jojoa and B. Garcia-Zapirain, Forecasting COVID 19 Confirmed Cases Using Machine Learning: the Case of America. Preprints 2020, 2020090228. doi: 10.20944/preprints202009.0228.v1
[23] V. H. Moreau, Forecast predictions for the COVID-19 pandemic in Brazil by statistical modeling using the Weibull distribution for daily new cases and deaths. Braz J Microbiol, 51, 1109–1115, 2020. https://doi.org/10.1007/s42770-020-00331-z
[24] R. G. da Silva, M. H. D. M. Ribeiro, V. C. Mariani, L. dos S. Coelho, Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables, Chaos, Solitons & Fractals, 139, 110027, 2020. https://doi.org/10.1016/j.chaos.2020.110027
[25] F. S. H. de Souza, N. S. Hojo-Souza, E. B. dos Santos, C. M. da Silva and D. L. Guidoni, Predicting the disease outcome in COVID-19 positive patients through machine learning: a retrospective cohort study with Brazilian data. medRxiv 2020.06.26.20140764. doi: https://doi.org/10.1101/2020.06.26.20140764
[26] S. Wollenstein-Betech, C. G. Cassandras, I. C. Paschalidis, Personalized Predictive Models for Symptomatic COVID-19 Patients Using Basic Preconditions: Hospitalizations, Mortality, and the Need for an ICU or Ventilator. medRxiv 2020.05.03.20089813. doi: https://doi.org/10.1101/-2020.05.03.20089813
[27] N. Ayoobi et al., Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods, Results in Physics, 27, 104495, 2021. https://doi.org/10.1016/j.rinp.2021.104495
[28] Bala, Sagar, COVID-19 Outbreak Prediction Analysis using Machine Learning. International Journal for Research in Applied Science and Engineering Technology, 9, 1-7, 2021. doi:10.22214/ijraset.2021.32690
[29] A. Hassan, A. Qasem, W. Abdalla and O. Elhassan. Visualization, Prediction of COVID-19 Future Outbreak by Using Machine Learning. International Journal of Information Technology and Computer Science. 13, 16-32, 2021. doi: 10.5815/ijitcs.2021.03.02
[30] C. Yu, S. Chang, T. Chang, J. Wu, Y. Lin, H. Chien and R. A. Chen, COVID-19 Pandemic Artificial Intelligence–Based System with Deep Learning Forecasting and Automatic Statistical Data Acquisition: Development and Implementation Study J Med Internet Res, 23(5), e27806, 2021. https://www.jmir.org/2021/5/e27806 doi:10.2196/278 06
[31] S. F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. R. Varkonyi-Koczy, U. Reuter, T. Rabczuk and P. M. Atkinson, "COVID-19 Outbreak Prediction with Machine Learning". medRxiv 2020.04.17.20070094. doi: https://doi.org/10.1101/2020.04.17.20070094
[32] K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar, Y. N. Pawan, "Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms" in International Journal of Emerging Trends in Engineering Research, 8(5), 2020. doi: https://doi.org/10.30534/ijeter/2020/117852020
[33] N. S. Punn, S. K. Sonbhadra and S. Agarwal. "COVID-19 Epidemic Analysis using Machine Learning. and Deep Learning Algorithms". medRxiv 2020.04.08. 20057679. doi: https://doi.org/10.1101/2020 .04.08 .20057679
[34] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B.-W. On, W. Aslam and G. S. Choi, "COVID-19 Future Forecasting Using Supervised Machine Learning Models". in IEEE Access, 8, 101489-101499, 2020. doi: 10.1109/ACCESS.2020.2997311
[35] M. Şahin, Forecasting COVID-19 cases based on mobility. MANAS Journal of Engineering, 8(2), 144-150, 2020. doi: 10.51354/mjen.769763
[36] S. Tuli, S. Tuli, R. Tuli, and S. S. Gill, "Predicting the Growth and Trend of COVID-19 Pandemic Using Machine Learning and Cloud Computing". Internet of Things, 11, 100222, 2020. https://doi.org/10.1016/j.iot.2020.100222