World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

Retrospective study of machine learning based Covid-19 prediction frameworks

R. John Martin [1, *]

[1] *School of Computer Science & Information Technology, Jazan University, Jazan, KSA.*

World Journal of Advanced Research and Reviews, 2023, 17(01), 890–903

Publication history: Received on 12 December 2022; revised on 21 January 2023; accepted on 23 January 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.17.1.0097

Abstract

During the deadly pandemic of COVID-19, several clinical and non-clinical mechanisms were adopted by the governments and World Health Organization to flatter the pandemic curve and succeed to a certain extent. Diversified research groups involved themselves in this mission to their utmost capacity. The contributions of artificial intelligence and data analytics cannot be forgotten. Several research outcomes are brought up with the use of machine learning (ML) and deep learning (DL) by analyzing worldwide COVID-19 related datasets. Various predictive analytics models have been proposed by statisticians, clinicians, and computer scientists. Now, this is the time to evaluate these models and to prove the validity of the proposed methods. This study aims to analyze the effectiveness of the predictive analytics models proposed during the pandemic by using ML and DL methods. In this review, the original research works put forth in the indexed journals during the pandemic period are analyzed and how they are relevant in today's context. Besides, the featured algorithms widely used in various frameworks and their importance in predictive analytics are presented.

Keywords: COVID-19; Predictive Analytics; Data Mining; Machine Learning; Deep Learning; Healthcare Analytics

1. Introduction

Covid-19 is caused by the SARS-CoV-2 virus, which is spreading at an unprecedented rate across the globe right now. Most people who are infected are asymptomatic or have mild pneumonia, so there is no need for treatment (Huang C, et al., 2020) but new data shows that people over 60 and those with comorbid conditions are particularly vulnerable to respiratory problems that can be fatal (Johns Hopkins, 2020). It is virtually impossible to prevent or treat an infection caused by the Corona virus, despite its high contagiousness (Sohrabi C, et al., 2020). There were 300 million infections worldwide by January 2022 (Johns hopkins, 2022) after it was first reported in Wuhan, China, in December 2019. There are still many unknowns about the pandemic's temporal nature, even though states have taken extreme measures to contain it. The virus has already mutated several times, and the world is on edge because of the new mutation, Omicron.

Covid-19-related data is collected from all over the world and is considered valuable if appropriate data analysis methods are used to gain insights. It is important to emphasize the value of Big Data Analytics, which includes Machine Learning, in order to meet the current and future challenges of combating this pandemic. When conducting this research, the team relied on already-existing information about the Covid-19 pandemic, including data on global infection rates, clinical data on the various virus strains and virus mutations, mitigation strategies, vaccines and drugs. During the Covid-19 pandemic, we have carefully selected frameworks for predictive analytics that have been published in reputable indexed journals. There were a few frameworks that used statistical techniques along with Machine Learning, but the majority of the frameworks used Machine Learning and Deep Learning. From January 2020 to January 2022, a selection of papers based on the most recent findings will be published. Covid-19 data sets from Kaggle and the Johns Hopkins Centre for Systems Science (Johns hopkins, 2022) were the most commonly used, but other sources, such

* Corresponding author: R. John Martin; Email:jmartin@jazanu.edu.sa

as those in the region or state, also used Covid-19 data. Some works are regionally specific, while others are global in scope.

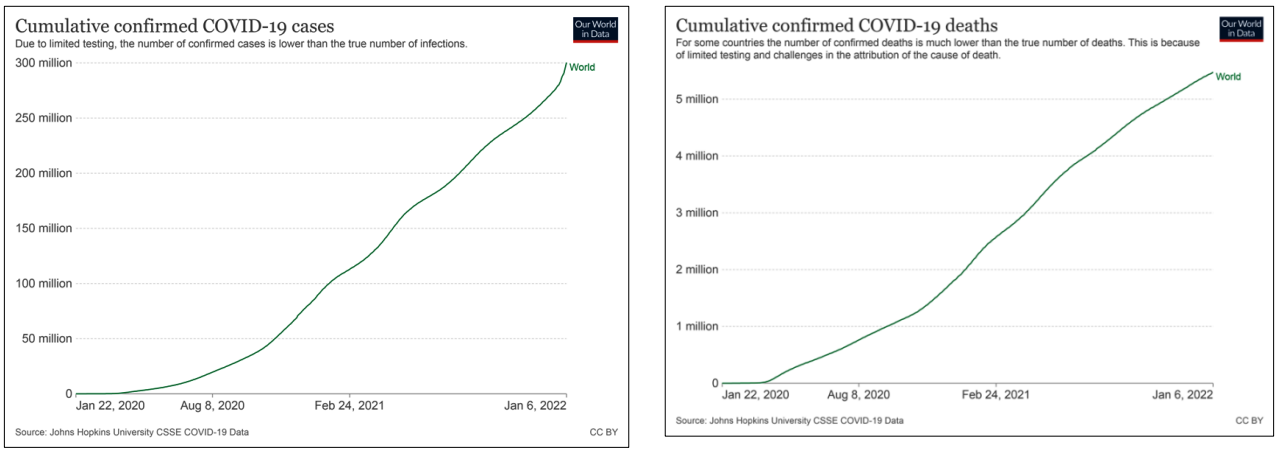

Figure 1 Cumulative conformed Covid-19 cases and deaths (Source: https://ourworldindata.org/covid-cases)

The observations and discussions on the findings of this retrospective analysis are presented in the sections that follow. Section 2 provides an overview and description of the individual research projects conducted in each domain, such as Machine Learning, Deep Learning, Big Data Analytical methods, and Statistical methods. Section 3 discusses potential Machine Learning and Deep Learning models for Predictive Analytics that can be used in future research projects of a similar nature. Sections 4 provide an overview of this study by comparing various machine learning and deep learning models based on a variety of criteria.

2. Literature Review

Machine learning algorithms are being widely used in a variety of contexts, from pandemic forecasting to patient critical illness prediction, mortality prediction, and outbreak forecasting (John Martin, et al., 2018). Predictive Analytics frameworks for the Covid-19 pandemic are listed in Table I using machine learning and deep learning. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) (Moher D, et al., 2009) guidelines were used to conduct this retrospective analysis and systematic review.

Table 1 Summary of prediction and forecasting frameworks using ML & DL during Covid-19 pandemic period

[Ref]	Targeted Prediction	Research Period (MMM YYYY)	Data Source/ Country	Methods Used
Assaf D, et al., 2020	Predicting Critical Illness / Mortality	Jun 2020	Israel	ANN | RF| DT
Batista AFM, et al., 2020	Predicting Critical Illness / Mortality	Mar 2020	Brazil Hospital Israelita Albert Einstein (HIAE) in São Paulo, Brazil,	SVM | ANN | RF
Li Yan, et al., 2020	Predicting Critical Illness / Mortality	Jan-Feb 2020	Wuhan, China Tongji Hospital	XGBoost - ML
M Pourhomayoun and M Shakibi, 2021	Predicting Critical Illness / Mortality	Unknown	[c] Worldwide	SVM | ANN | RF | DT | LR | kNN
Patterson Bruce K, et al., 2021	Predicting Critical Illness / Mortality	Jun 2021	US	RF

Sankaranarayanan S et al., 2021	Predicting Critical Illness / Mortality	Mar 2020	USA	Recurrent NN with other ML
Mohammad reza Nemati, et al., 2020	Predicting Critical Illness / Mortality	Jun- Aug 2020	China	GBoost | Statistical
Banoei M M, et al., 2021	Predicting Critical Illness / Mortality	Jun 2020 – Jun 2021	USA	PLS | LCA
An C, et al., 2020	Predicting Critical Illness / Mortality	Jan 2020 – Apr 2020	South Korea	LASSO | LSVM | SVM (RBF) | DT | RF | kNN
Malki Z, et al., 2020	Predicting Critical Illness / Mortality	Dec 2019 – Apr 2020	[a, b] Worldwide	LR | LASSO | etc
Farooq, et al., 2020)	Predicting Critical Illness / Mortality	Unknown	India	ANN
Sujata Dash, et al., 2021	Pandemic Prediction	Mar 2020 - Mar 2021	[a] US | Brazil | India | France | Russia | UK	ARIMA
Alomari E, et al., 2021	Pandemic Prediction	Nov 2020	Saudi Arabia	LDA
Pinter G, et al., 2020	Pandemic Prediction	Apr 2020	Hungary	ANFIS
Sujath R, et al., 2020	Pandemic Prediction	May 2020	[b] India	LR | ANN | VAR
Muhammad LJ, et al., 2021	Pandemic Prediction	Oct 2020	Mexico	LR | DT | SVM | NB | ANN
Shalini Ramanathan, et al., 2021	Pandemic Prediction	Dec 2020	[c] Worldwide	SVM | kNN | Big Data Mining
Zoabi Y, el al., 2021	Pandemic Prediction	Mar 2020	Israel	GBoost | ANN
Shreshth Tuli, et al., 2020	Pandemic Prediction	Till May 2020	[d] Worldwide	LM | BDA (FogBus)
Ardabili SF, et al.,	Pandemic Prediction	Aug-Sept 2020 30 days sample	[e] Italy | Germany | Iran | USA | China	MLP | ANFIS | GA | PSO
Da Silva RG et al., 2020	Pandemic Prediction	Mar-Apr 2020	Brazil/USA	NN | Quantile RF | SVM
Durga Prasad K, et al., 2020	Pandemic Prediction	Unknown	India	PDR-NML
Furqan Rustam et al., 2020	Pandemic Prediction	Till May 2020	[a] Worldwide	LR | LASSO | SVM | ES
Talha Burak Alakus, et al., 2020	Pandemic Prediction	Early months of 2020	Brazil	ANN | CNN | RNN | LSTM
Wang P et al., 2020.	Pandemic Prediction	Jan 2020 – Jun 2020	[a] Worldwide	Logistic Model and FBProphet Model
Cafer Mert, 2020	Pandemic Prediction	Jan 2020 – Jun 2020	[a] Worldwide	RF
Zivkovic M et al., 2021	Pandemic Prediction	Jan – Feb 2020	China (WHO)	ANFIS | Adaptive ML

Wadhwa P et al., 2021	Pandemic Prediction	Jun – Aug 2020	India	LR
Efthimios Kaxiras et al., 2020	Pandemic Prediction	Till Apr 2020	Worldwide	Kermack McKendrick Mathematical Model
Singhal A, et al., 2020	Pandemic Prediction	Till Jun 2020	India | Italy | USA	Mathematical & Non-Parametrical Models
Wang P, et al.,	Predicting Infection	Aug 2021	[b] Pakistan	CNN

Abbreviations: DT - Decision Tree | ANN – Artificial Neural Network | RF – Random Forest | DT – Decision Tree | SVM – Support Vector Machine | ARIMA - Autoregressive Integrated Moving Average | kNN – k Nearest Neighbour | LDA - Latent Dirichlet Allocation | LCA - Latent Class Analysis | ANFIS - Adaptive network-based fuzzy inference system | VAR – Vector Autoregression | NB – Naive Bayes | PLS – Partial Least Square Regression | LM - Levenberg-Marquardt | PDR-NML – Partial Derivative Regression – Nonlinear ML | CNN – Convolution Neural Network | LSTM - Long-Short Term Memory | RNN - Recurrent Neural Net- works | LR – Linear Regression | ES – Exponential Smothing | MLP – Multi Layer Perceptron | GA – Genetic Algorithm | PSO – Particle Swarm Optimizer; Covid-19 Data Sources: [a] JH - Johns Hopkins University Corona Virus Data Stream | [b] Kaggle dataset | [c] UCI ML Repository | [d] Our World in Data by Hannah Ritchie | [e] https://www.worldometers.info/coronavirus/#countries

2.1. Predicting Covid-19 Critical Illness/Mortality

Assaf D, et al. (2020) developed a predictive analytics model using ANN and Random Forest algorithms. Developed models are trained with Covid-19 patient data of the state of Israel and claimed that ML models were outperformed with 92% accuracy over other existing frameworks. The authors proclaim that the ML model work well for predicting critical illness among the patients. Hence, the model can be viable tool for patients' classification and will help the healthcare professional to effectively deal with the pandemic by reducing the death rate.

Another research carried out in Brazil by Batista AFM et al., (2020) for predicting critical patients from the covid positive patients with biological parameters. Different Machine learning algorithms are tested with the proposed model such as ANN, RF, GBT, LG and SVM. Among the five, SVM performed well with an AUC 0.85. It was concluded that the use of machine learning could be right choice for identifying critical care patients along with RT-PCR test results and physiological parameters of the patients.

A Machine learning based prognostic model proposed in Wuhan China at the early stage of pandemic by Li Yan et al., (2020) and used gradient boosted decision tree (XGBoost) algorithm for their prediction framework. The model used major clinical features of the patients for classification. Though the authors claim efficiency in their framework, it is noted that the model was trained with limited number of samples. Another experimental analysis conducted by Sankaranarayanan, S et al., (2021) with USA based covid-19 dataset using deep neural network based gated recurrent unit (GRU) binary classification model by using various different biomarkers as features. It was targeted individual patients to predict the mortality possibility and obtained AUC of 0.938 which is higher than other ML models.

A clinical decision making model for covid-19 mortality risk evolved by M Pourhomayoun and M Shakibi (2021) used various machine Learning algorithms including ANN. The empirical study conducted in US with world wide data. On evaluation, the ANN based model outperformed for predicting critical care patients with the accuracy of 89.9%.

In order to predict the covid severity, a machine learning based framework proposed by Bruce K. Patterson et al., (2021) during the second wave in Unites States of America. They used systematic diagnostic parameters including plasma and isolated peripheral blood mononuclear cells (PBMCs) and T-cells activation levels in patients for classification with Random Forest as one of the classifier.

According to Mohammad reza Nemati et al., (2020), the analysis of infected patients with healthcare support was effective with a machine learning framework using Gradient Boost algorithm. The work was aimed at predicting critical care patients and the patient's length of stay in the hospital. The research was conducted in China during early months of the pandemic.

Banoei M M et al., (2021) devised a Mortality model in USA by using one year covid-19 data from June 2020 to June 2021. Partial Least Square (PLS) Regression model was used for classification along with Latent Class Analysis (LCA) for clustering the patients. The model endow with considerably high accuracy of AUC>0.85.

A South Korea based research study conducted by An C et al., (2020) by taking into account of covid-19 patient data collected from January 2020 to April 2020. It was aimed at mortality rate prediction among the patients by adopting various machine learning algorithms such as the linear support vector machine (SVM), SVM with RBF kernel, Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), and k-Nearest Neighbours (kNN). The proposed ML model performed well using LASSO and Linear SVM algorithms and both achieved high accuracy of AUC>0.96. The most significant predictors in this framework were age and cancer comorbidity.

Malki Z et al., (2020) brought out an important study by associating weather factor for predicting mortality rate. The authors reported that they were used worldwide data from different sources during the period of five months since December 2019. Various regression based machine learning algorithms are employed to set the predictive model. Weather variables such as temperature and humidity are used to train the model along with patient's clinical data. It is noted that weather is also a factor in the pandemic and one of the reason for mortality rate. Farooq J et al., (2020) reported their covid-19 strategic model for reducing mortality using neural network based deep learning algorithm during June 2020. The model simulated strategy of controlled natural immunization for the patients with comorbidities. Though the model is technically valid, its use is still questionable.

2.2. Pandemic Prediction

Large number of experimental studies came out during the covid-19 pandemic to predict and forecast the pandemic in order to deal with the consequences. A Big Data Analytics (BDA) based forecasting model put forth by Sujata Dash et al. (2021) with an objective of predicting further outbreak in upcoming 90 days since June 2021 for the six worst hit countries and six worst hit states in India. The autoregressive integrated moving average (ARIMA) model was used in the machine learning framework. The study confirms the efficiency of ARIMA Model and recommends the model for epidemiological analysis in future.

In June 2020, Alomari E et al., (2021) brought out an analytical model by focussing Saudi Arabia to understand the public opinion and assess the government's pandemic measures using twitter data. The unsupervised machine learning model used Latent Dirichlet Allocation (LDA) for simulation. Pinter G et al., (2020) developed a pandemic prediction framework in Hungary by adopting hybrid machine learning approach. They used adaptive network-based fuzzy inference system (ANFIS) and multi-layered perceptron - imperialist competitive algorithm (MLP-ICA) in order to predict wave of infections and death rate. It is claimed that the developed hybrid machine learning model is an alternative to traditional epidemiological models and has the potential for predicting COVID-19 outbreak accurately.

Sujath R et al., (2020) developed a forecasting model for India during April 2020 to predict the potential of anticipated covid-19 spread. The model used various machine learning algorithms namely linear regression, Multilayer perception and Vector Auto Regression. Though the predicted numbers are different from reality, the models fit for predicting the anticipated waves of pandemic.

A trained model of a covid-19 decease detection system was developed by Muhammad L, J et al., (2021) using the positive and negative cases of Mexican based dataset. Different supervised machine learning models were employed for this framework namely LR, DT, SVM, NB, and ANN. On evaluation, DT emerged as an efficient algorithm with the accuracy of 94.99%. It is noted that all the ML algorithms performed well in detecting the covid-19 cases with accuracy rate greater than 93%.

Shalini Ramanathan et al., (2021) developed a covid-19 patient classification framework using the worldwide dataset obtained from UCI ML repository. Dataset are obtained from the repository using big data mining concepts. They used RT-PCR test results for classification using SVM and kNN classifiers. Three stages of COVID-19 cases are classified. The study put forth the alternative method of detecting the category of COVID-19-positive patients using blood tests and machine learning combined instead of rRT-PCR. This discovery is considered as the most significant and applicable one during the stages of pandemic with shortage of RT-PCR test kits. In an Israel based empirical study carried out by Zoabi Y el al., (2021), a covid-19 symptom based diagnostic framework is brought out with the use of Gradient Boost and ANN algorithms. They obtained data from local health administration and applied to train the models using covid-19 positive and negative data along with patients' physiological parameters. It is noted that their framework is working well with high accuracy and shall be used to identify the COVID-19 cases whenever testing resources are limited.

A pandemic prediction machine learning model Covid-19 was proposed by Shreshth Tuli et al., (2020) during the mid of 2020 by considering worldwide data till May 2020. The model is build in a cloud based platform and used a machine learning technique Levenberg-Marquardt (LM) for curve fitting. The researchers observed that their framework

achieved accurate and real-time prediction of pandemic curve. This is considered as an important forecasting model which gives governments clear message to be proactive in dealing with pandemic.

With 30 days covid-19 data samples of five countries (Italy, Germany, Iran, China and USA), Ardabili S F et al., (2020) developed predictive analytics model for covid-19 pandemic by using Machine Learning and few soft computing methods (GA & PSO). Among the various algorithms tested, MLP and ANFIS given promising results. The study also recommends machine learning as a potential tool for future researches on pandemic predictions.

USA and Brazil based pandemic prediction system was proposed by Da Silva RG et al., (2020) in the early stages of the pandemic with the use of diversified machine learning algorithms with climate variables. The model was tested with five highly reported states one, three, and six-days-ahead the cumulative COVID-19 cases with a high number of cases up to April 28th, 2020 and achieved better accuracy of 70%.

Durga Prasad K, et al., (2020) brought out a pandemic outbreak prediction model using partial derivative regression and nonlinear machine learning (PDR-NML) method. The work was targeted for Indian state of Tamilnadu during early stages of the wave. It is reported that the proposed model outperformed most of the recent frameworks in pandemic prediction in India.

Yet another machine learning framework of pandemic forecasting presented by Furqan Rustam et al., (2020) used most promising algorithms such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES). Worldwide covid-19 data repository (Johns Hopkins University Corona Virus Data Stream) was used for this framework. It was aimed at identifying new infections, possible deaths and recoveries in the next 10 days. Notably all the above algorithms performed well in the framework for forecasting and the winner was ES.

A deep learning based covid-19 prediction model was introduced by Talha Burak Alakus et al., (2020) during the early months of the pandemic with an objective to identify who are likely to receive a COVID-19 disease. The most promising deep learning algorithms (ANN | CNN | RNN | LSTM) are adopted for the model. According to the reported results, the predictive models identify patients that have COVID- 19 disease at an average accuracy of 86.66%. The CNN-LSTM combination obtained high accuracy of 92.3% in prediction. These types of predictive models could be used to effectively handle the pandemic and to prioritize the resources appropriately.

An AI based pandemic trend simulation model was evolved by adopting worldwide covid-19 pandemic data received from January – June 2020 by Wang P et al., (2020). The framework was developed using Logistic Model and FBProphet Model coupled with machine learning techniques. As a result of this simulation model, it was estimated that the global pandemic outbreak will peak in late October 2020 with the cumulative infection of 14.12 million. In reality there is a big gap between the real data and the projected simulation data.

Cafer Mert (2020) developed a pandemic wave estimation model by considering covid-19 data from 170 countries and taking into the account of the infection rate from January 2020 to June 2020. Random Forest machine learning algorithm was used for the model, and claimed that the model performed well in estimating the infection rates of the nearby dates.

An Artificial Intelligence based model was proposed in a covid-19 prediction framework by Zivkovic M et al., (2021) by inducting adaptive neuro-fuzzy inference system (AFNIS) and enhanced beetle antennae search swarm intelligence metaheuristics. The study aimed at predicting the pandemic wave using the data obtained from China during the early days of the pandemic from January 2020 to February 2020. It is concluded that the hybrid machine learning model managed well to outscore other benchmarking models proposed at that time and recommended for time series prediction frameworks.

A Linear Regression based predictive analytics model was put forth by Wadhwa P et al., (2021) with an objective of predicting lock down extension period in India amidst sharp increase in covid-19 cases during June-August 2020. The model predicted expected covid positive cases, critical illness and deaths. Based on the projections the study recommends appropriate measures to the government on lockdown.

An analytical model was simulated by Efthimios Kaxiras et al., (2020) to study the evolution of the pandemic by considering first 100 days covid-19 pandemic data. They used Kermack McKendrick Mathematical Model for simulation with the worldwide data till April 2020. According to the authors, the model predicted expected number of infections in each country as well as the date at which the total number of covid cases reach maximum. When we check the current data (https://ourworldindata.org/coronavirus), the projections of this mathematical model is completely wrong.

Similarly, Singhal A, et al. (2020) developed a pandemic prediction model using mathematical and non-parametric model which was based on the pandemic data from India, Italy and USA. Despite the fact that the model not used any Machine Learning techniques, due to its significance we include in this retrospective study. The model records 95% confidence intervals as June 2020. This Gaussian mixture model estimates that the total number of expected cases and deaths are 12.7 ×10 6 and 5.27 ×10 5, respectively. But as on today (16/01/2023), the number of cases and deaths are much more than numbers projected by the model. Hence, we report that these types of statistical or mathematical models are less significant in predicting pandemic waves and the importance of Machine Learning is highly significant for pandemic prediction.

3. Potential Algorithms/Methods used in Prediction Frameworks

Several machine learning and deep learning frameworks put forth for the prediction of pandemic and its related incidents as elaborated in section 2. In this section we present the most promising algorithms and their impacts in accurate prediction of covid-19.

3.1. Linear Regression

A well-known and widely used regression algorithm is a linear regression model that can able to learn with a linear combination of input features (Awan MJ, et al., 2021) Let's have the inputs: $\{(x_i, y_i)\}_{i=1}^{N}$, where N is the size of the input collection, $\boldsymbol{x_i}$ is the D-dimensional feature vector of example *1 to N*, y_i is a real-valued target and every feature $x_i^{(j)}$, $j = 1, \ldots, D$, is also a real number.

To build a model $f_{w,b}(x)$ as a linear combination of features of **x**:

$$f_{w,b}(x) = wx + b \tag{1}$$

where **w** is a D-dimensional parameter vector and $\boldsymbol{b}$ is a real number. The entry $f_{w,b}$ represents that the model $\boldsymbol{f}$ is parameterized by the values **w** and $\boldsymbol{b}$.

The model being used to predict the unknown factor $\boldsymbol{y}$ for a given **x:** $\leftarrow f_{w,b}(x)$. It is obvious that the two models parameterized by two different training pairs ($\boldsymbol{w}, \boldsymbol{b}$) will possibly give two dissimilar predictions with the same example. It is important to optimize ($\boldsymbol{w}, \boldsymbol{b}$) in such a way to fit with the model to make most accurate predictions. The linear model as stated in Eq.1 is almost similar to the SVM model. Yet, the decision boundary is decided only based on the hyperplane in the SVM so as to separate two classes of samples from one another.

3.2. Random Forest

Random forest (RF) and Gradient Boosting (GB) are the two widely used ensemble learning algorithms for the forecasting models. The concept of learning for classification in Random forest (Breiman, L, 2001) is straightforward, wherein weak data samples pooled together to form strong learners with a set of disassociated decision trees. Random forest normally aggregates the output from a number of shallow trees, forming an additional layer to *bagging*. *Bagging* forms *n* predictors with independent subsequent trees, by bootstrapping samples of the dataset. The combination of n predictors will solve a classification or estimation problem through averaging. It is quite natural that there will be weak learners, but the combination will form a strong learner. The working principle of bagging algorithm is as follows.

For the given training set, choose N random samples S_b (for each $b = 1, \ldots, B$) from the training set and make a decision tree model d_b using each sample S_b as the training set. In order to sample S_b for some b, sampling with replacement principle is applied. Meaning that, begin with an empty set, and then pick and choose at random a sample from the training set and put it into to S_b by keeping the original training set as it is until the $|S_b| = N$.

On training the model, B decision trees are obtained. In regression, the prediction for a new sample x is the average of B forecasts using Eq.2 or the majority of votes in classification.

$$y \leftarrow \hat{f}(x) \stackrel{\text{def}}{=} \frac{1}{B}\sum_{b=1}^{B} d_b(x) \tag{2}$$

While the individual decision trees experience high variance and high bias, random forest aggregates multiple decision trees in order to improve overall performance. This means, though a decision tree showcase a weak performance in the forest, the collective performances of decision trees in the forest will make a prediction. The variance of final estimation is reduced by averaging across the ensemble of decision trees. Random forest provides good accuracy and efficient in

using with large datasets. It is also a suitable estimation method for using with incomplete or inconsistent datasets. In addition, random forest can able to estimate the relative importance of a variable for classification. Hence, random forest was widely used for covid-19 prediction frameworks at the early stages of the pandemic.

3.3. Gradient Boost

Gradient boosting is considered as an effective ensemble learning method and is widely seen in Covid-19 forecasting frameworks. In order to construct a strong predictor, we start with a constant model $f = f_0$.

$$f = f_0(x) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} y_i \tag{3}$$

Then we modify labels of each example $i = 1, \ldots, N$ in our training set like follows:

$$\widehat{y_i} \leftarrow y_i - f(x_i) \tag{4}$$

where $\hat{y}_i$ called the *residual*, is the new label for example $\mathbf{x}_i$.

Use the modified training set with residuals in order to build a new decision tree model, f_1. The boosting model is now stated as $f \stackrel{\text{def}}{=} f_0 + \alpha f_1$, where α is the learning rate coefficient.

Residuals are being computed using eq. 4 and again replace the labels in the training data, train the new model f_2, redefine the boosting model as $f \stackrel{\text{def}}{=} f_0 + \alpha f_1 + \alpha f_2$ and continue till the maximum of M trees aggregated.

Naturally, by calculating the residuals, it could be found how best the target of each training example is identified by the current model f. Then chose another tree to train and fix the error gaps of the current model and subsequently add the new tree to the existing model with some weight α. As a result, each tree added to the model in some extent fixes the errors resulted by the previous trees until reaching the optimum number of trees are to be combined.

The concept of gradient computing is different from linear regression models as we don't calculate any gradient here as compared to linear regression. In general, the gradient is showing the direction, but it is unclear that how far it should go further in this direction. Hence, used an iterative step and then re-assessed its direction. This is what happening in gradient boosting; however, rather than getting the gradient directly, a proxy is deployed in the form of residuals to show how the error gap (the residual) is reduced by tuning the model. There are three principal coefficients also called hyper parameters which may affect model accuracy includes number of trees, learning rate coefficient and the tree depth. The classification algorithm of gradient boosting is fairly similar to regression, but with slightly different steps.

It is evident from the referred bench marking literatures that the gradient boosting is one of the most influential machines learning algorithms particularly for the predictive frameworks of Covid-19. It is not just because of its accuracy; rather, it is capable of handling large datasets with huge number of distinctive features.

3.4. kNN

k-Nearest Neighbors (kNN) is a class of non-parametric learning. In contrast to other learning algorithms of similar nature that permit discarding the training data once the model is built, kNN holds all training examples in memory. On detecting a new and unseen example, the kNN algorithm traverse k training examples closest to the new occurrence and awards the most dominant class label (in case of classification) or the average label (in case of regression). The three major elements of kNN includes: i) existing labelled set of points ii) a distance function to estimate the closeness of points and iii) the number of nearest neighbors (k).

The proximity of two points is given by a distance function. Negative cosine similarity function is the popular choices for the distance function in kNN. The cosine similarity is defined as,

$$S(x_i, x_k) \stackrel{\text{def}}{=} \cos\big(\angle(x_i, x_k)\big) = \frac{\sum_{j=1}^{D} x_i^{(j)} x_k^{(j)}}{\sqrt{\sum_{j=1}^{D}\left(x_i^{(j)}\right)^2}\sqrt{\sum_{j=1}^{D}\left(x_k^{(j)}\right)^2}} \tag{5}$$

The Eq.5 is the similarity measure of the orientation of two vectors. When the angle is 0 degrees, the two vectors are in the same direction and cosine similarity is 1. If the vectors are orthogonal, the cosine similarity will be 0. Whenever the vectors are pointing in opposite directions, the cosine similarity is ≠1. On using cosine similarity as a distance metric, it should be multiplied by ≠1. Chebychev distance, Mahalanobis distance, and Hamming distance are the well-known

distance metrics. The selection of distance metric and the value for the parameter *k*, is done by the analyst with trade-off between different values of *k*.

3.5. Lasso Regression

The Eq.1 above shows the linear model as a linear combination of features. With only a single feature, it could understand that *w* is the slope and *b* represents intercept. Optimizing *w* and *b* in linear regression is obtained by minimizing the cost function. The cost function for the simple linear model with *N* instances and *p* features is expressed as:

$$\sum_{i=1}^{N}(y_i - \hat{y_i})^2 = \sum_{i=1}^{N}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 \quad (6)$$

The problem of over-fitting and under-fitting is common in linear regression models. On using linear regression with the data-set divided into training and test set, calculating the scores on training and test set can give us an estimation about whether the model is suffering from over-fitting or under-fitting. The under-fitting problem comes usually with less number of features and scores poor with the proposed models. Conversely, using large number of features and getting relatively poor test score than the training score called over-fitting. With the use of some simple techniques, the under-fitting and over-fitting problems can be addressed. Lasso regression is a regularization technique to reduce model complexity in a simple linear regression model so as to prevent over-fitting. The cost function of a Lasso (least absolute shrinkage and selection operator) regression model can be expressed as:

$$\sum_{i=1}^{N}(y_i - \hat{y_i})^2 = \sum_{i=1}^{N}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p}|w_j| \quad (7)$$

Here in the above equation, if λ =0 then it reduces to Eq.6. Lasso regression does L1 regularization by adding a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can lead to zero coefficients which mean some of the features are completely eliminated from the model for the evaluation of output. Hence Lasso regression helps in solving over-fitting problem and produce simpler models. Alternatively, L2 regularization (e.g. Ridge regression) may not lead to exclusion of coefficients.

3.6. Adaptive Network-based Fuzzy Inference System (ANFIS)

ANFIS is considered as the most influential AI based prediction system with a combined framework of neural networks and fuzzy based inference mechanism (JSR Jang, 1993). Feed-forward neural network such as backpropagation is generally used with ANFIS. It builds a fuzzy inference system whose membership function coefficients are adjusted using a feed forward NN algorithm. This is to enable the fuzzy system to learn from the data.

3.7. Autoregressive Integrated Moving Average (ARIMA)

An ARIMA is a statistical regression analysis model that uses time series data to either better understand the data set or to predict potential trends by measuring the strength of one dependent variable relative to other changing variables. For instance, an ARIMA model has been used to predict projected covid-19 cases in the near future based on the covid-19 cases reported in the previous months. It has been used widely in covid-19 pandemic prediction models along with machine learning frameworks. ARIMA is blended in nature with a mixture of differenced autoregressive model and moving average model. Mathematically, ARIMA is expressed as:

$$y'_t = I + \alpha_1 y'_{t-1} + \alpha_2 y'_{t-2} + \cdots + \alpha_p y'_{t-p} + e_t + \theta_1 e_{t-1} + \theta_{12} e_{t-2} + \cdots + \theta_{1q} e_{t-q} \quad (8)$$

The above Eq.8 demonstrates that the predictors are the lagged *p* data samples for the differenced auto regression and the lagged *q* errors are for the differenced moving average. The prediction is the differenced y_t in the d^{th} order. The ARIMA(p,d,q) model is the estimation of the coefficients α and θ for a given *p, d, q* as the model learns from the training data samples in a time series.

3.8. Deep Learning

One of the most reliable regression models for prediction frameworks is ANN-based regression (Abdulaal, A et al., 2020). ANN generates an approximate output by estimating the weights and constant values for each neuron in the hidden layer and reducing the inaccuracy in the target values. A matching weight w for every input x indicates how much influence it has on the outcome. Eq. 9 is used to calculate neuron output (Y):

$$Y = f(b + W X) \quad (9)$$

Y is the output, X is the input to the model from the preceding layer, b is the proportional constant and W is the weight attributed to each X variable. An activation function calculates Y. Common activation functions include Sigmoid, ReLU, and Tan h. The ANN model employs learning methods to estimate weights and correlation coefficient. Backpropagation is the most commonly used learning algorithm for regression. It starts by determining the weights and coefficients. After that, the errors for out-of-sight values in the output layer are calculated. Finally, the weights and constants are revised. Convolutional Neural Network [29] was used for prediction in a few of the deep learning works reported here.

4. Results and discussion

The purpose of this retrospective study was to conduct a systematic review of the covid-19 pandemic prediction frameworks using machine learning and deep learning and show to the prospective researchers how these models could be effective in upcoming frameworks. The researchers used various machine learning models while taking into account population metrics, health facilities, cured, deaths, and confirmed positive cases in order to build the model and to achieve accurate prediction. The prediction frameworks were considered variety of other factors in addition to covid-19 data, such as social media data, meteorological data, and medical imaging data. Despite the fact that no predictive models proposed by the researchers give exact accurate results when all features related to COVID-19 are considered, thanks to the researchers for their works that predict the waves of pandemic and disease severity and are well informed to decision makers for timely management of the situation.

Patient data will be a valuable resource for researchers working on machine learning-based predictive and forecasting frameworks to develop automatic diagnostic mechanisms and therapeutic strategies for Covid-19. The majority of the works presented here made use of global Covid-19 data repositories such as Johns Hopkins University Corona Virus Data Stream, Kaggle dataset, UCI ML Repository, and Hannah Ritchie's Our World in Data. A few models were developed using regional health data gathered from the respective countries.

Table 2 Prediction performance of potential ML Algorithms in the reported works in terms of ROAUC

Sl.No	ML/DL Algorithms		ROAUC (Area Under the Receiver Operating Curve)							
		[5]	[3]	[7]	[8]	[21]	[30]	[35]	[39]	Avg.
1	SVM	-	0.94	-	0.85	0.88	-		-	0.89
2	Decision Tree	0.90	0.87	-	-	0.93	-	-	-	0.90
3	Random Forest	0.93	0.93	-	0.85	0.94	-	-	-	0.91
4	K Nearest Neighbour	-	0.77	-	-	0.90	-	-	-	0.84
5	Gradient Boost/XBoost	-	-	-	0.84	-	-		0.94	0.89
6	LASSO	-	0.94	-	-	-	-	-	-	0.94
7	Logistic Regression	-	-	-	0.84	0.92	-	-	-	0.88
8	ANN/DL	0.92	-	-	0.84	0.93	0.94	0.90	0.97	0.92
9	SIMPLS	-	-	0.85	-	-	-	-	-	0.85

Table 1 summarizes the machine learning and deep learning frameworks for pandemic and critical illness prediction as a result of this in-depth analysis. It should be noted that not all ML/DL-based prediction works used the same evaluation metrics. Table 2 shows the results of the prediction frameworks in terms of ROAUC, while Table 3 shows the results of the work in terms of Accuracy.

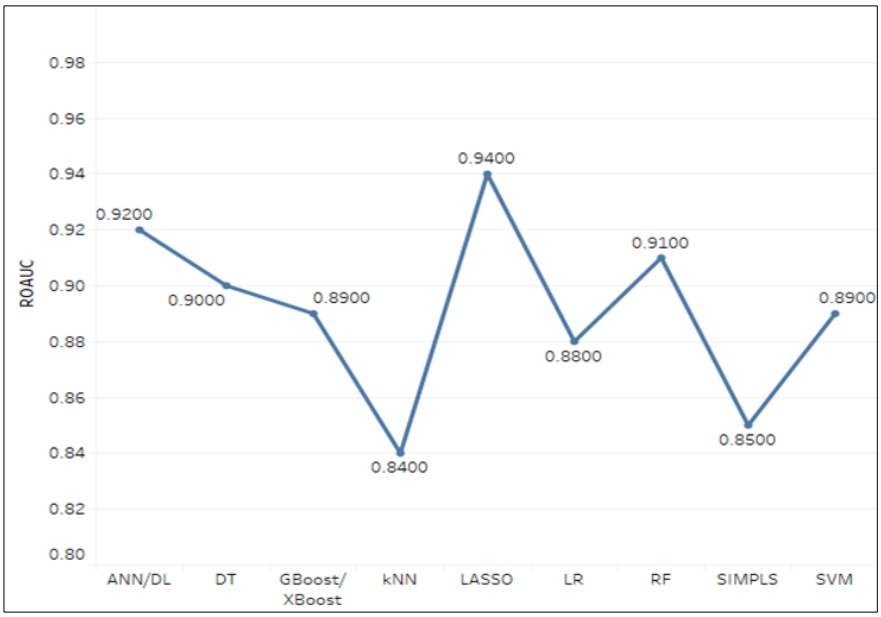

Figure 2 ROAUC metric of Covid-19 Prediction frameworks in ML algorithms

Though parts of the results are expressed in terms of ROAUC (Fig.2) and the remaining in terms of accuracy (Fig.3), most other studies used alternative measures to assess prediction performance. The metrics used to predict the performance include ROAUC, f1-score, accuracy, and statistical measures.

Table 3 Prediction performance of potential ML Algorithms in the reported works in terms of Accuracy

Sl.No	ML/DL Algorithms		Accuracy %							
		[3]	[5]	[19]	[20]	[21]	[22]	[25]	[26]	Avg.
1	SVM	87.7	-	-	-	89.0	90.7	92.4	-	90.0
2	Decision Tree	-	92.0	-	-	86.8	-	95.0	-	91.3
3	Random Forest	90.7	92.9	-	-	87.9	-	-	95.0	91.6
4	K Nearest Neighbour	97.0	-	-	-	89.8	-	-	-	93.4
5	Gradient Boost/ XBoost	-	-	90.0	-	-	91.5	-	-	90.8
6	LASSO	96.8	-	-	93.2	-	-	-	-	95.0
9	Logistic Regression	-	-	-	-	87.9	-	94.4	-	91.2
10	ANN/DL	-	92.8	-	-	90.0	-	98.2	-	93.6

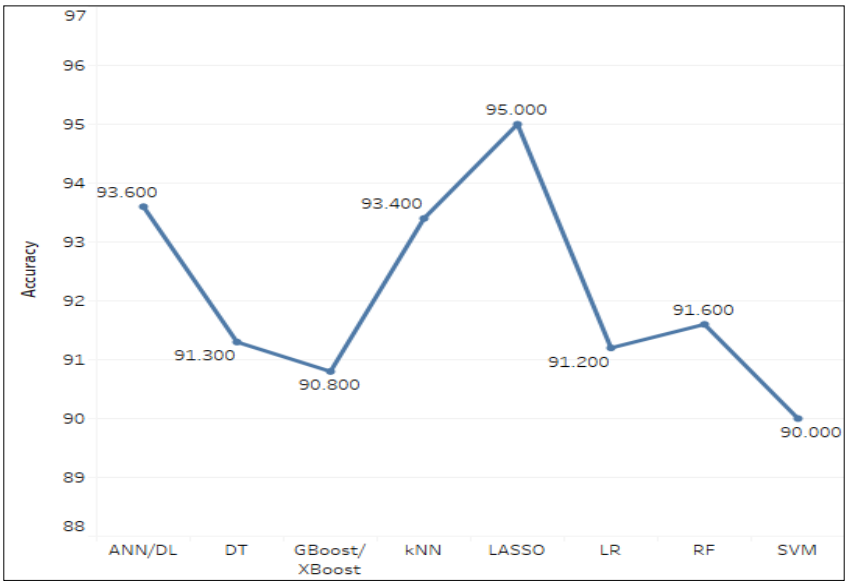

Figure 3 Accuracy of Covid-19 Prediction frameworks in ML algorithms

5. Conclusion

By considering all the results of the empirical studies reviewed, we found the following significant observations: i) Deep learning-based models won the race in predicting critical illness among COVID-19 patients, and they had a significant impact on the automatic detection and extraction of essential features from CT-Scan images. ii) LASSO emerged as a powerful predictor among the bench marking models by addressing the over fitting issues of linear regression models. iii) An ensemble learning-based XBoost, a more regularized version of Gradient Boost, emerged as a powerful predictor of the COVID-19 pandemic. On comparing the real-time data with projected data, it was noted that the AI-based models were outperforming traditional statistical models. The facts derived from benchmarking research works of pandemic prediction frameworks are important sources of knowledge and inspiration for future AI-based prediction frameworks.

Compliance with ethical standards

Acknowledgments

I would like to thank everyone who contributed to the benchmark research works that were published during the COVID-19 pandemic.

Disclosure of conflict of interest

There is no conflict of interest to declare.

References

[1] Abdulaal, A., Patel, A., Charani, E. et al. Comparison of deep learning with regression analysis in creating predictive models for SARS-CoV-2 outcomes. BMC Med Inform Decis Mak , 20, 299 (2020). https://doi.org/10.1186/s12911-020-01316-6

[2] Alomari E, Katib I, Albeshri A, Mehmood R. COVID-19: Detecting Government Pandemic Measures and Public Concerns from Twitter Arabic Data Using Distributed Machine Learning. Int J Environ Res Public Health. 2021;18(1):282. Published 2021 Jan 1. https://doi.org/10.3390/ijerph18010282

[3] An, C., Lim, H., Kim, DW. et al. Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study. Sci Rep 10, 18716 (2020).

[4] Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM. COVID-19 Outbreak Prediction with Machine Learning. Algorithms. 2020; 13(10):249. https://doi.org/10.3390/a13100249

[5] Assaf D, Gutman Y, Neuman Y, Segal G, Amit S, Gefen-Halevi S, Shilo N, Epstein A, Mor-Cohen R, Biber A, Rahav G, Levy I, Tirosh A. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern Emerg Med. 2020 Nov;15(8):1435-1443. https://doi.org/10.1007/s11739-020-02475-0

[6] Awan MJ, Bilal MH, Yasin A, Nobanee H, Khan NS, Zain AM. Detection of COVID-19 in Chest X-ray Images: A Big Data Enabled Deep Learning Approach. International Journal of Environmental Research and Public Health. 2021; 18(19):10147. https://doi.org/10.3390/ijerph181910147

[7] Banoei, M.M., Dinparastisaleh, R., Zadeh, A.V. et al. Machine-learning-based COVID-19 mortality prediction model and identification of patients at low and high risk of dying. Crit Care. 25, 328 (2021).

[8] Batista AFM, Miraglia JL, Donato THR, Chiavegatto Filho ADP. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. medRxiv, 2020.

[9] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[10] Cafer Mert Yeşilkanat. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos, Solitons & Fractals. Volume 140, 2020, 110210, https://doi.org/10.1016/j.chaos.2020.110210.

[11] da Silva RG, Ribeiro MHDM, Mariani VC, Coelho LDS. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. Chaos Solitons Fractals. 2020;139:110027. https://doi.org/10.1016/j.chaos.2020.110027

[12] Durga Prasad Kavadi, Rizwan Patan, Manikandan Ramachandran, Amir H. Gandomi. Partial derivative Nonlinear Global Pandemic Machine Learning prediction of COVID 19. Chaos, Solitons & Fractals. Volume 139, 2020, 110056, https://doi.org/10.1016/j.chaos.2020.110056.

[13] Efthimios Kaxiras, Georgios Neofotistos, Eleni Angelaki. The first 100 days: Modeling the evolution of the COVID-19 pandemic. Chaos, Solitons & Fractals. Volume 138, 2020, 110114. https://doi.org/10.1016/j.chaos.2020.110114.

[14] Farooq J, Bazaz MA. A novel adaptive deep learning model of Covid-19 with focus on mortality reduction strategies. Chaos Solitons Fractals. 2020 Sep;138:110148.

[15] Furqan Rustam et al., "COVID-19 Future Forecasting Using Supervised Machine Learning Models," in IEEE Access, vol. 8, pp. 101489-101499, 2020, https://doi.org/10.1109/ACCESS.2020.2997311

[16] Huang C , Wang Y , Li X , Ren L , Zhao J , Hu Y , et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395(10223):497–506.

[17] J.-S.R. Jang, "ANFIS: adaptive-network-based fuzzy inference system," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, no. 3, pp. 665-685, May-June 1993, https://doi.org/10.1109/21.256541

[18] Johns hopkins coronavirus resource center. 2022. https://coronavirus.jhu.edu/

[19] Li Yan et al. A machine learning-based model for survival prediction in patients with severe COVID-19 infection. medRxiv, 2020.

[20] Malki Z, Atlam ES, Hassanien AE, Dagnew G, Elhosseini MA, Gad I. Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Chaos Solitons Fractals. 2020 Sep;138:110137.

[21] Mohammad Pourhomayoun, Mahdi Shakibi. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart Health, Volume 20, 2021, 100178.

[22] Mohammadreza Nemati, Jamal Ansary, Nazafarin Nemati. Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data. Patterns. Volume 1, Issue 5,2020,100074.

[23] Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, PLoS Med 6(7):e1000097. https://doi:10.1371/journal.pmed.1000097

[24] Mortality Analyses - Johns Hopkins coronavirus resource center. 2020. https://coronavirus.jhu.edu/data/mortality

[25] Muhammad LJ, Algehyne EA, Usman SS, Ahmad A, Chakraborty C, Mohammed IA. Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. SN Comput Sci. 2021;2(1):11. https://doi.org/10.1007/s42979-020-00394-7

[26] Patterson Bruce K., Guevara-Coto Jose, Yogendra Ram, Francisco Edgar B., Long Emily, Pise Amruta, Rodrigues Hallison, Parikh Purvi, Mora Javier, Mora-Rodríguez Rodrigo A. Immune-Based Prediction of COVID-19 Severity and Chronicity Decoded Using Machine Learning . Frontiers in Immunology. 12. 2021. 2520 .

[27] Pinter G, Felde I, Mosavi A, Ghamisi P, Gloaguen R. COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach. Mathematics. 2020; 8(6):890. https://doi.org/10.3390/math8060890

[28] Ramanathan S, Ramasundaram M. Accurate computation: COVID-19 rRT-PCR positive test dataset using stages classification through textual big data mining with machine learning. J Supercomput. 2021 Jan 4:1-15. https://doi.org/10.1007/s11227-020-03586-3

[29] S. Dash, C. Chakraborty, S. K. Giri, S. K. Pani and J. Frnda, "BIFM: Big-Data Driven Intelligent Forecasting Model for COVID-19," in IEEE Access, vol. 9, pp. 97505-97517, 2021, https://doi.org/10.1109/ACCESS.2021.3094658

[30] Sankaranarayanan, S., Balan J. et al. COVID-19 Mortality Prediction From Deep Learning in a Large Multistate Electronic Health Record and Laboratory Information System Data Set: Algorithm Development and Validation. J Med Internet Res. 2021 Sep 28;23(9):e30157. https://doi.org/10.2196/30157

[31] Shreshth Tuli, Shikhar Tuli, Rakesh Tuli, Sukhpal Singh Gill. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. Internet of Things. Volume 11, 2020, 100222, https://doi.org/10.1016/j.iot.2020.100222.

[32] Singhal A, Singh P, Lall B, Joshi SD. Modeling and prediction of COVID-19 pandemic using Gaussian mixture model. Chaos Solitons Fractals. 2020;138:110023. https://doi.org/10.1016/j.chaos.2020.110023

[33] Sohrabi C , AlsafiZ , OfiNeill N , Khan M , Kerwan A , Al-Jabir A , et al. World health organization declares global emergency: a review of the 2019 novel coron- avirus (COVID-19). Int J Surg 2020 .

[34] Sujath, R., Chatterjee, J.M. & Hassanien, A.E. A machine learning forecasting model for COVID-19 pandemic in India. Stoch Environ Res Risk Assess 34, 959–972 (2020). https://doi.org/10.1007/s00477-020-01827-8

[35] Talha Burak Alakus, Ibrahim Turkoglu. Comparison of deep learning approaches to predict COVID-19 infection. Chaos, Solitons & Fractals. Volume 140, 2020, 110120, https://doi.org/10.1016/j.chaos.2020.110120.

[36] Wadhwa P, Aishwarya, Tripathi A, Singh P, Diwakar M, Kumar N. Predicting the time period of extension of lockdown due to increase in rate of COVID-19 cases in India using machine learning. Mater Today Proc. 2021;37:2617-2622. https://doi.org/10.1016/j.matpr.2020.08.509

[37] Wang P, Zheng X, Li J, Zhu B. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos Solitons Fractals. 2020;139:110058. doi:10.1016/j.chaos.2020.110058

[38] Zivkovic M, Bacanin N, Venkatachalam K, Nayyar A, Djordjevic A, Strumberger I, Al-Turjman F. COVID-19 cases prediction by using hybrid machine learning and beetle antennae search approach. Sustain Cities Soc. 2021 Mar;66:102669. https://doi.org/10.1016/j.scs.2020.102669

[39] Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digit. Med. 4, 3 (2021). https://doi.org/10.1038/s41746-020-00372-6

[40] John Martin R,, Swapna S.L, Sujatha S. Adopting Machine Learning Models for Data Analytics - A Technical Note. International Journal of Computer Sciences and Engineering, Vol.6, Issue.10, pp.359-364, 2018. https://doi.org/10.26438/ijcse/v6i10.359364

Author's short biography

Dr. R. John Martin has over 25 years of experience as an academic in the field of Computer Science. Dr. Martin earned his Ph.D in Computer Science from Bharathiar University in India and specialized in Machine Learning with an application to biomedical data analytics. He has vast experience in higher education as an educator and administrator in India and the Middle East. Currently, he is working in the School of Computer Science and Information Technology at *Jazan University* (Ministry of Education), KSA. He has published extensively in the fields of machine intelligence and biomedical data analytics and has served as editor and reviewer for refereed journals. His research was patented both nationally and internationally. His accomplishments as an educator, mentor, author, researcher, adjudicator and consultant are acknowledged by the global community. His research interests include Machine Intelligence, Signal Processing and Healthcare Data Analytics.