Modeling and analysis of COVID-19 new deaths using tree-based ensemble

Preprint · September 2021

DOI: 10.36227/techrxiv.16566012

Ibrahim Abaker Targio Hashem
University of Sharjah
91 PUBLICATIONS **11,589** CITATIONS

Raja Sher Afgun Usmani
BlueBlaze.Earth
30 PUBLICATIONS **632** CITATIONS

Muhammad Bilal
National University of Computer and Emerging Sciences
55 PUBLICATIONS **1,120** CITATIONS

Modeling and analysis of COVID-19 new deaths using tree-based ensemble

SUBMISSION DATE / POSTED DATE

03-09-2021 / 09-09-2021

CITATION

Hashem, Ibrahim Abaker Targio; Usmani, Raja Sher Afgun; Shah, Asad Ali; Almazroi, Abdulwahab Ali; Bilal, Muhammad (2021): Modeling and analysis of COVID-19 new deaths using tree-based ensemble. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.16566012.v1

DOI

10.36227/techrxiv.16566012.v1

Modeling and analysis of COVID-19 new deaths using tree-based ensemble

[1]Ibrahim Abaker Targio Hashem, [2]Raja Sher Afgun Usmani, [3,4]Asad Ali Shah, [3]Abdulwahab Ali Almazroi,
[5]Muhammad Bilal

[1] College of Computing and Informatics, Department of Computer Science, University of Sharjah, 27272 Sharjah, UAE.
[2]Department of Software Engineering, Faculty of Computing, and Information Technology, University of Sialkot
[3,4]Department of Computing, School of Electrical Engineering & Computer Science, National University of Sciences and Technology, Pakistan
[4]Department of Information Technology, College of Computing and Information Technology at Khulais, University of Jeddah, Jeddah, Kingdom of Saudi Arabia
[5]Department of Computer Science, FAST National University of Computer and Emerging, Sciences, Islamabad 44000, Pakistan.

Email: ihashem@sharjah.ac.ae, rajasherafgun@gmail.com, asad.safdar@hotmail.com, maz-2000@hotmail.com
bilal.m@nu.edu.pk

Corresponding author: ihashem@sharjah.ac.ae,

Abstract— The COVID-19 pandemic has emerged as the world's most serious health crisis, affecting millions of people all over the world. The majority of nations have imposed nationwide curfews and reduced economic activity to combat the spread of this infectious disease. Governments are monitoring the situation and making critical decisions based on the daily number of new cases and deaths reported. Therefore, this study aims to predict the daily new deaths using four tree-based ensemble models i.e., Gradient Tree Boosting (GB), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Voting Regressor (VR) for the three most affected countries, which are the United States, Brazil, and India. The results showed that VR outperformed other models in predicting daily new deaths for all three countries. The predictions of daily new deaths made using VR for Brazil and India are very close to the actual new deaths, whereas the prediction of daily new deaths for the United States still needs to be improved.

Index Terms— COVID-19; Voting Regressor; Machine Learning; Analysis; Gradient Tree Boosting.

1.0 INTRODUCTION

In the last century, science has made significant contributions in the field of medicine. These advancements have allowed the world's medical systems to improve health care, lower the mortality rate and find cures for diseases such as smallpox and polio [1, 2]. Despite these In the past four decades, the world has witnessed many viral pandemics [3]. Starting with Human Immunodeficiency Virus in 1981 to Severe Acute Respiratory Syndrome Coronavirus-2 also known as COVID-19 the world is currently facing, most of these pandemics have affected world economies, country’s health systems, travel, education, and many other sectors. In short, one cannot rely upon medical systems along for combating pandemics alone.

Before scientists were able to develop vaccines for COVID-19, governments and states relied on taking cautionary measures to control the spread of the virus. These included imposing lockdowns, sanitization, wearing masks, temperature sensing, mobile tracing, telemedicine, robots [4]. Among these lockdowns have been successful in controlling the spread as it ensures people avoiding contact altogether. However, these lockdowns do come at the cost of huge economical losses. From experience, governments have learned to use smart lockdowns, where instead of shutting down the country, state, or city completely, only specific areas are locked down based on the data gathered. This is done based on data collected and correct predictions made by systems [5]. Therefore, researchers and experts are making great efforts in improving the accuracy of their systems to predict new, recovered, death cases accurately.

There are various ways through which predictions can be made for different types of COVID cases. Some scientists have made predictions using mathematical models while others have used artificial intelligence techniques. All techniques hold certain advantages and disadvantages making them suitable in specific scenarios and not suitable in others [6, 7]. Instead of relying on only one technique, researchers have suggested the use of ensemble techniques where multiple techniques are combined to produce better predictions.

This research will specifically investigate ensemble models for predicting covid cases. The research will test out different models and select the ones performing better than others in the system by evaluating a dataset using evaluation metrics used to evaluate such systems.

The rest of the paper is structured as follows. Section 2 covers the literature review where past studies and research gap is identified. Section 3 covers the methodology of the paper as well as the dataset used and the evaluation metrics for testing the system. Section 4 discusses the results and findings of the research and in section 5 we conclude the paper following up with future work.

2.0 LITERATURE REVIEW

Mankind has faced devastating pandemics in the past, whether it was Malaria that struct Ancient Egypt 5000 years ago, the plague of Justinian, or the Spanish Flu. However, over the years research has improved and models have been proposed in understanding how diseases reached pandemic levels better [8]. A great amount of effort is being an investment in by researchers in preparing models that can better project the COVID-19 pandemic deaths, to help us understand the spread better and fight it off. Before discussing different models that have been proposed, it is important to understand the different types of models present, which can range from simple to complex.

There are different models available for projecting pandemics. SIR is one of the popular models for project pandemics, due to its simplistic nature. However, variations of SIR can add more complexity such as SIS, SIRD, MSIR, SEIR, SEIS, MSEIR, MSEIRS, and Carrier state are also present that is useful as different diseases behave differently [8, 9]. Moreover, researchers have also used artificial intelligence techniques including supervised, and unsupervised machine learning techniques, and deep learning for making new death predictions [10]. Additionally, ensemble learning, which uses multiple machine learning models and classifiers, has also shown great success to solve such problems [6, 7]. In this research, the scope has been limited mainly towards ensemble techniques due to their exploitation of utilizing the advantages of multiple machine learning models to produce better results. However, papers making some use of ensemble techniques have also been looked into.

The literature review was done system following a systematic literature review using the PRISMA framework [11]. Figure 1, shows the systematic literature review conducted for this research. These articles shortlisted for review are discussed in detail in their respective paras below:

Da Silva, Ribeiro [12] suggested a framework that uses different artificial intelligence models including Bayesian regression neural network, cubist regression, k-nearest neighbors, quantile random forest, and support vector regression, for time-series forecasting. They used different artificial intelligence models, which are used standalone. These models are coupled with pre-processing variational mode decomposition. This allows them to decompose the time-series into intrinsic mode functions and allows them to forecast with one, three, and six days ahead of new COVID-19 cases. The hybrid variational mode decomposition framework was able to higher accuracy in 70% of the cases. The results show that cubist regression couple with variational mode decomposition model is much more suitable than other models.

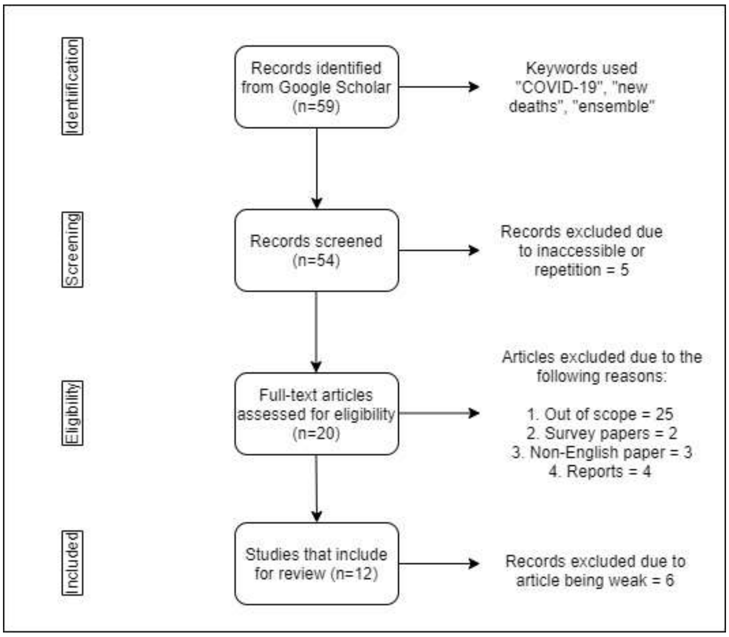
Identification
Records identified from Google Scholar (n=59)
Keywords used "COVID-19", "new deaths", "ensemble"
Screening
Records screened (n=54)
Records excluded due to inaccessible or repetition = 5
Eligibility
Full-text articles assessed for eligibility (n=20)
Articles excluded due to the following reasons:
1. Out of scope = 25
2. Survey papers = 2
3. Non-English paper = 3
4. Reports = 4
Included
Studies that include for review (n=12)
Records excluded due to article being weak = 6

Figure 1 – PRISMA systematic literature review

Friston, Parr [13] proposed the use of a dynamic casual model for modeling the spread of COVID-19 through the population. The model can predict new cases and deaths. This model uses state-of-the-art variational (Bayesian) model inversion and comparison procedures, that is can highlight the effects of interventions such as social distancing and herd immunity. This model is used to epidemiological populations to predict the results as per time series data.

Evensen, Amezcua [14] proposed using an iterative ensemble smoother for correcting predicting future deaths and people hospitalized. This was done by using the ensemble smoothers on parameters of a susceptible, exposed, infectious, recovered model by using age classes as well as classes for sick, hospitalized, and dead. The system was tested on data collected from 11 countries and regions. The results showed that the best results for the system were obtained using N=100 ensemble size with random seed value set to 5000-10000.

Altieri, Barter [15] suggested the use of Combined Linear and Exponential Predictors for forecasting new COVID-19 cases up to 14 days. The research involved various kinds of predictors that utilize features that are either country-specific, common across all counties, data for neighboring counties, and demographics-based data. The research used combined predictors involving 1) separate-county exponential predictors 2) separate-county linear predictor, 3) shared-county exponential predictor, 4) expanded shared exponential predictor, and 5) demographics shared exponential predictor. The results showed a mean absolute percentage error of 8.18%, 12.21%, 15.14%, and 26.45% for 3-, 5-, 7-, and 14-day-ahead prediction. This is one of the only predictors that can estimate deaths per country instead of country and has been adopted by non-profit organizations as well.

Elsheikh, Saba [16] proposed the use of a deep learning model using a long-short term memory model for predicting the number of total confirmed cases, recovered cases, and deaths due to COVID-19 in Saudi Arabia. The proposed model is also tested for other countries as well for verification purposes including Brazil, India, South Africa, Spain, and the USA. The system utilized the optimal hidden value and learning rate for achieving better results, which were 100 and 0.005 respectively. The system was able to predict results up to 1 week, which far better results with baseline systems tested against with including NARANN and ARIMA. The system also used several evaluation metrics for testing the results including Root mean square error, coefficient of determination, mean absolute error, efficiency coefficient, overall index, coefficient of variation, and coefficient of residual mass. In coefficient of determination, which highlights the correlation of predicted results vs actual results (with a score between 0 and 1), the system achieved 0.976 for total cases and 0.944 for total deaths.

Hao, Xu [17] suggested the use of Elman neural network, long short-term memory, and support vector machine for predicting new COVID-19 cases including confirmed cases, deaths, and cured cases. The system can predict trends and analyze cumulative confirmed, death, and cured cases. It is also able to predict the growth range of new, death, and cured cases. For the system, the research used Elman neural network, due to its effectiveness for working with historical data. This technique is also used once more data is added dynamically. Additionally, long short-term memory and Support vector machines are used for classification depending on the data being dealt with. The system evaluated the Wuhan dataset using a square correlation coefficient where cumulative confirmed was 84.51%, cumulative deaths was 99.06% and cumulative cured was 99.74%. For the USA dataset, cumulative confirmed was 99.03%, and cumulative deaths were 99.12%.

De Figueiredo, Dos Santos [18] proposed the use of adjusted determination coefficient, Akaike information criterion, and residual mean square performance measures in the training process. For selecting the best model for the dataset, mean absolute percentage error and the relative error criterion were used. The results showed that the Bertalanffy model correlated closely with the deaths for China, whereas the Gompertz model worked best for Brazil, Germany, Italy, Spain, and the United States. The results generated achievements predictions with a confidence score of 95%.

Jo, Kim [19] suggested the use of deep learning using long short-term memory networks with a quantile output model. This system was able to project new death rates for up to two weeks. One of the uniqueness of this system was that it was also able to project data down to the county level. It also was able to show results on a fine geographical scale, making it useful to manage resources at the state level. The system was tested on 2,721 counties out of 3,114 counties in total. The results showed that the system performed poorly with lesser data to deal with but gradually improved once the model was trained with more data.

Silva, Barreira [20] evaluated how econometrics, machine learning models, and ensemble methods can be used to predict new COVID-19 cases. In econometrics, the study used ARIMA and SARIMA econometrics models. For machine learning models, AdaBoost and GBR models were evaluated. Moreover, ensemble methods were also evaluated. The study evaluated these models on Brazil, South Korea, China, and Italy datasets, using features such as the total number of cases, the total number of deaths, new cases, new deaths in the day, and recovered patients. The results showed that no single model gave better predicts in all datasets. However, the ensemble of machine learning and econometrics showed a lot of potentials. This is because machine learning models performing poorly with less data, but can be compensated by using ensemble methods.

Goic, Bozanic-Leal [21] proposed a framework for predicting up to two weeks forecast for utilization and availability of ICU beds during the COVID-19 pandemic. The framework uses an ensemble approach that combines autoregressive, artificial neural networks, and a compartment model. The system was tested on a Chile dataset and achieved a mean error of 4% for the first week and 9% for the second week. Results showed the ensemble approach performing better than individual models for handling different scenarios.

Ngie, Nderu [22] suggested the use of a Simultaneous Tree-based Regressor Interactive Model that uses an ensemble of Decision Tree and Logistic Regression models. The ensemble method allows the system to achieve better accuracy than individual models instead. The accuracy was increased further by introducing particle swarm optimization for parameter tuning. The results were very encouraging, where the ensemble approach was better than both individual models and improved further with the inclusion of particle swarm optimization.

Ngie, Nderu [22] also prosed the use of an ensemble of different models. Their system was able to forecast up to 30 to 60 days of new cases and deaths. The system utilized an ensemble of Monte Carlo simulations, wavelet analysis, and least-squares Optimization that is applied onto an SEIR compartmental model. This allows the system to produce stochastic epidemiological models that can produce better predictions. The results were tested on Greece, Germany, and Switzerland datasets, and was able to keep a ratio of tested to real cases between 0.12 and 0.18.

Shastri et al. [23] proposed a nested ensemble model based on long short-term memory (LSTM) for predicting confirmed and fatal Covid-19 cases in India. The results showed that the deep learning-based ensemble model can predict daily new cases with 97.59 percent accuracy and daily new deaths with 98.88 percent accuracy. Furthermore, the MAPE value of 2.40 for new confirmed cases and 1.11 for daily new deaths demonstrated the effectiveness of the proposed approach. Shastri et al. [24] compared the prediction performance of Bi-directional LSTM and Convolutional LSTM with the ensemble of Convolutional and Bi-directional LSTM to predict the number of daily new cases and deaths in Brazil, India, and the United States. It was noticed that the proposed ensemble model performed significantly better than other models.

In summary, table 1 shows the research gap of the research. Findings show that the ensemble methods perform better and are preferred over individual models. This is due to the fact they perform better as being able to adapt to different scenarios. This is a major problem for individual models that perform well in one dataset but poorly in another. Moreover, the research gap also shows a good mix of features that can be utilized, but the essential ones being daily reports of cases, deaths, and recovery. In our research, we hope to learn from this research gap and implement a methodology accordingl

Table 1: Summary of the related works

Ref	Framework	Dataset	Results	Models/techniques	Features

[12]	Hybrid variational mode decomposition framework	Brazil and USA	~70% accuracy	non-decomposed and decomposed models including BRNN, CUBIST, KNN, QRF, SVR, and VMD	climatic exogenous variables
[13]	Dynamic causal modeling	UK	-	variational (Bayesian) model inversion and comparison procedures	Number of initial cases, effective population size, herd immunity proportion, Location, infection, clinical and testing
[14]	Ensemble Data Assimilation	Argentina, Brazil, England, France, Norway, The Netherlands, four states (New York, California, Alabama, and North Carolina) of the USA, and the province Qu´ebec of Canada	The best results were achieved with N-100 ensemble member size and random seed value set between 5000-10000 realizations	Ensemble Kalman filter on SEIR model. This includes the function initial growth rate R(t), and the initial infected and exposed Ii and Ei	age-classes and compartments of sick, hospitalized, and dea
[15]	Combined Linear and Exponential Predictors	USA	8.18%, 12.21%, 15.14% and 26.45% for 3-, 5-, 7-, and 14-day-ahead prediction for mean absolute percentage error	Combined Linear and Exponential Predictor	Features that are county-specific, common across all countries, neighboring counties, and demographics. These include population density per square mile, population estimate, number of hospitals, number of ICU beds, median age, percentage of the population who are smokers, percentage of the population with diabetes, deaths due to heart diseases per 100,000 and cumulative death count
[16]	Deep learning-based forecasting model	Brazil, India, Saudi Arabia, South Africa, Spain, and USA	The correlation results (score between 0 to 1) between predicted results	long short-term memory	the total number of confirmed cases, recoveries, and deaths

			and actual results are: 0.976 for total cases 0.944 for total deaths		
[17]	Prediction and Analysis using Elman neural network, long short-term memory, and support vector machine	China and USA	The system results for Wuhan dataset using square correlation coefficient are as follows: Cumulative confirmed: 84.51% Cumulative deaths: 99.06% Cumulative cured: 99.74% The system results for USA dataset using square correlation coefficient are as follows: Cumulative confirmed: 99.03% Cumulative deaths: 99.12%	It mainly uses Elman neural network and long short-term memory for prediction and analysis. Additionally, scalar vector machines are used on non-linear data for achieving higher prediction	Time series of confirmed cases, deaths, and cured cases and their commutative
[18]	An application to COVID-19 S-shaped models using Long-Term Time Prediction	Brazil, China, Germany, Italy, Spain, the United States	The confidence score of 95%	Techniques include adjusted determination coefficient, Akaike Information Criterion, and Residual Mean Square	Observation data such as new cases, recovered cases, and deaths
[19]	Deep Learning Model For Predicting using long short-term memory networks with quantile output model	USA	The model showed poor results at the beginning where there was not enough data to model, but with enough data, the model was able to correlate with the actual results.	Long Short-Term Memory networks with Quantile output	Seasonality features, 64 features of the demographics and local health data, 43 features including the population estimate and the mortality rate of various underlying diseases, and policy actions features
[20]	Use of econometrics and machine learning models to predict new cases	Brazil, China, Italy, and South Korea	In mean absolute error, the best performing models for different countries are	Econometrics, machine learning, and ensemble models	total number of cases, the total number of deaths, new cases, new deaths in the day, and recovered

			Econometrics models Brazil: SARIMA 163.53 China: Ensemble 21.50 Italy: ARIMA 682.37 South Korea: Ensemble 26.43 Machine Learning Models Brazil: AdaBoost 51.66 China: GBR 186.19 Italy: AdaBoost 1748.00 South Korea: GBR 41.98		patients
[21]	COVID-19: Short-term forecast of ICU beds	Chile	average forecasting errors 4% - one week horizon 9% - two week horizon	Ensemble method utilizing autoregressive, machine learning, and epidemiological models	number of new infections, the positivity rate and the number of fatality cases
[22]	Tree-Based Regressor Ensemble	Over a hundred countries	Root mean square error Decision Trees 88.02 Regression Tree 90.43 STRIM Ensemble 96.84 mean absolute percentage error Decision Trees 87.07 Regression Tree 92.04 STRIM Ensemble 97.42	tree-based regressor model christened Simultaneous Tree-based Regressor Interactive Model , which is a ensemble of Decision Tree and Logistic Regression models	identification variables (country short name, country/state), input variable in the dataset included (population, weight, date, cumulative cases, cumulative deaths, new cases and new deaths
[25]	Ensemble forecasting models with statistically calibrated parameters and	Greece, Switzerland and Germany	ratio of tested to real cases is estimated to be between 0.12 and 0.18	combination of Monte Carlo simulations, wavelet analysis and least squares optimization is applied to a known basis of	confirmed daily and cumulative cases and deaths

	stochastic noise			SEIR compartmental models	
[23]	Deep learning	India	98.88% accuracy, 1.11 MAPE	Deep learning-based ensemble model based on LSTM	Time series of confirmed cases and deaths.
[24]	Deep learning	India, Bazil and United States	98.10% accuracy (India), 98.30% accuracy (Brazil), and 98.40% accuracy (USA)	Bi-directional LSTM, Convolutional LSTM, and ensemble of Convolutional and Bi-directional LSTM	Time series of confirmed cases and deaths.

3.0 MATERIALS AND METHODS

Study Location

This study is conducted for the three most affected countries by COVID-19, i.e., the United States of America (USA), Brazil, and India. The details of the study location and selected statistics are provided in Table 2. As displayed in 2, these countries have three of the highest COVID-19 cases and deaths. 2 also presents the statistics related to populations older than 65 and 70. The USA has a big percentage of the population older than 65 in comparison with Brazil and India. Brazil has a much higher value of extreme poverty than India and the USA, as well as a fairly low number of hospital beds per thousand.

Table 2: Study area statistics

Parameter	**USA**	**Brazil**	**India**
Population	331,002,647	212,559,409	1,380,004,385
Population Density	35.608	25.04	450.419
Median Age	38.3	33.5	28.2
Aged 65 Older	15.413	8.552	5.989
Aged 70 Older	9.732	5.06	3.414
GDP Per Capita	54,225.45	14,103.45	6,426.67
Extreme Poverty	1.2	3.4	21.2
Cardiovasc Death Rate	151.089	177.961	282.28
Diabetes Prevalence	10.79	8.11	10.39
Female Smokers	19.1	10.1	1.9
Male Smokers	24.6	17.9	20.6
Hospital Beds Per Thousand	2.77	2.2	0.53
Life Expectancy	78.86	75.88	69.66
Human Development Index	0.926	0.765	0.645
Total COVID-19 cases	20,061,903	7,675,973	10,266,674
Total COVID-19 deaths	351,817	194,949	148,738

Note: All statistics are from OWID dataset – December 31, 2020 (OWID, 2021)

Dataset

This study is carried out using the *Our Data in World* (OWID, 2021) COVID-19 dataset The COVID-19 dataset by OWID is provided on daily basis, and it includes data on deaths, confirmed cases, and testing (Hasell et al., 2020; OWID, 2021). COVID-19 dataset by OWID has 59 parameters, but for this study, we have removed the parameters with constant data, and 22 parameters are utilized for the study. Table 3 presents the selected parameters from the OWID COVID dataset. The data provided in the dataset is of a time series nature, with a timestep of one day. The duration of the OWID dataset is from the first confirmed case in each country, i.e., 22nd January 2020 for USA, 26th February 2020 for Brazil, and 30th January 2020 for India, to 31st December 2020 for all three countries. Parameters 1-21 are our independent parameters, and Parameter 22 (new deaths) is our dependent parameter.

Table 3: OWID COVID-19 dataset parameters

#	Parameter Name	#	Parameter Name
1	date	12	tests per case
2	total COVID-19 cases	13	stringency index
3	total COVID-19 deaths	14	total COVID-19 tests
4	new COVID-19 deaths	15	total COVID-19 tests *pt*
5	total COVID-19 cases *pm*	16	new cases
6	total COVID-19 deaths *pm*	17	new cases
7	new COVID-19 deaths *pm*	18	ICU patients
8	reproduction rate	19	ICU patients *pm*
9	new COVID-19 tests	20	hospital patients
10	new COVID-19 tests *pt*	21	hospital patients *pm*
11	positive rate	22	new deaths

Note: pm= per million, pt=per thousand, ph=per hundred

Methods

Machine Learning algorithms, such as Decision Tree and Artificial Neural Network are considered to be inherently unstable because these algorithms lead to significantly different predictions if the training dataset utilized in the algorithms has any perturbation (Hassan et al., 2017). These machine learning prediction algorithms are known to have low bias and high variance. Researchers have recommended tree-based ensemble methods to reduce the bias and/or variance. In these tree-based methods, various base predictor models are created and joined as an ensemble to form a single predictor (Hochkirchen, 2010). In this study, we are using four tree-based ensemble models, i.e., Gradient Tree Boosting (GB), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Voting Regressor (VR).

Gradient Tree Boosting

Gradient Tree Boosting or Gradient Boosted Decision Trees (GB) is a machine learning algorithm based on a decision tree-based ensemble model. GB is considered one of the most effective and versatile machines learning predictive models. GB is considered an accurate and effective method that can be applied for regression as well as classification problems. The graphical illustration of GB is presented in Figure 2. In GB, numerous sequential regression trees are iteratively chained together, making sure that each tree is trained using the residuals of the precious tree in the iteration. A new learner is included to reduce the loss function optimally, at each step. Afterward, an additive model is utilized to combine these trees, hence, creating a robust and effective ensemble model.

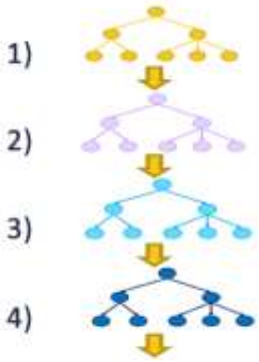

Figure 2: Gradient Tree Boosting

Extreme Gradient Boosting

XGBoost, formally known as Extreme Gradient Boosting, is an extension and implementation of GB. XGBoost is designed to prevent the phenomenon of overfitting, and enhance the speed and performance of the prediction (Chen et al., 2016). It is a scalable end-to-end method and during the training phase, it can adapt easily and make the best use of the resources available. Data scientists use XGBoost to tackle many machine learning challenges, often producing state-of-the-art results (Brownlee, 2013).

Random Forest

Random Forest or Random decision forests (RF) are one of the most commonly used decision tree-based learning algorithms. It is very popular in both classification and regression problems in machine learning. RF was developed in 2001 by Leo Breiman (Breiman, 2001). Leo Breiman formulated an algorithm for building a forest of uncorrelated trees using a method similar to regression and classification trees and the method included bagging and randomized node optimization. The overall working of RF is presented in Figure . Multiple trees are trained on slightly different training data and are combined into a robust and stronger model, whose prediction by committee is more precise than any individual decision tree in the RF.

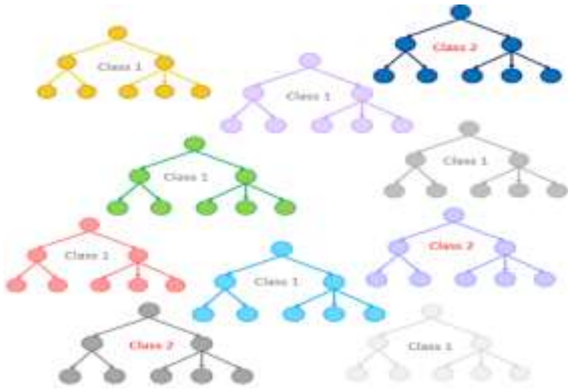

Figure 3: Random Forests

Voting Regressor

The Voting Regressor (VR) is created on an intuitive and simple concept. The concept is to combine multiple machine learning models and a final predicted value is calculated by using either their average predicted value or a value predicted by the majority of the machine learning algorithms in the ensemble. The working of VR is presented in Figure . VR is considered very useful in the machine learning models, which are equally well-performing. It will help to predict more accurately by balancing out their weaknesses.

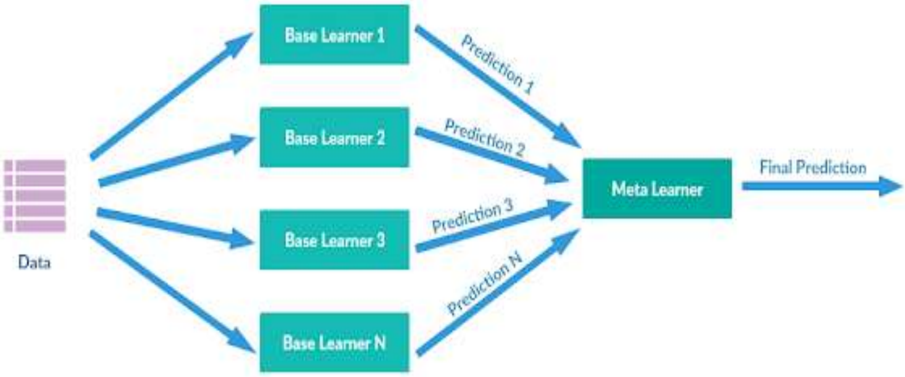

Figure 4: Voting Regressor

4.0 RESULTS & DISCUSSIONS

In this study, we are focusing on three of the most affected countries by COVID-19 in the world, i.e., Brazil, India, and USA. These countries have one of the highest death counts in the world due to COVID-19. As of 22 March 2021, the USA is leading the world in death counts with 555,314 deaths, and Brazil and India have 294,115 and 160,003 deaths due to COVID-19 respectively. In this study, we aimed to predict the new death count due to COVID-19 using tree-based ensemble methods, so the countries and the world can plan accordingly. The results show that the tree-based ensemble methods show good prediction power with all three countries. The models can detect the upward and downward trends of death counts relatively well, with the Voting regressor showing the best prediction in all three countries. In the coming sections, we will discuss the results of the countries and provide a comparative discussion at the end of the section.

BRAZIL

The first case for COVID-19 was reported in Brazil on 27th February 2020. The first casualty associated with COVID-19 was reported on 17th March 2020. The COVID-19 cases and deaths in Brazil increased considerably in the year 2020, with the highest death count of 1703 occurring on 24th September 2020.

Figure presents the predictions of death count in Brazil. The predictions were completed on the COVID-19 dataset, 80% (248 days) for training and 20% for testing (62 days). Figure presents the comparison of the test data with the predictions of models utilized in the study. The comparison shows that all four models perform well in comparison with real data. The models can detect the trend of the death count. The Voting regressor performs the best in comparison with other models in predicting the upward and downward curve of death count in Brazil.

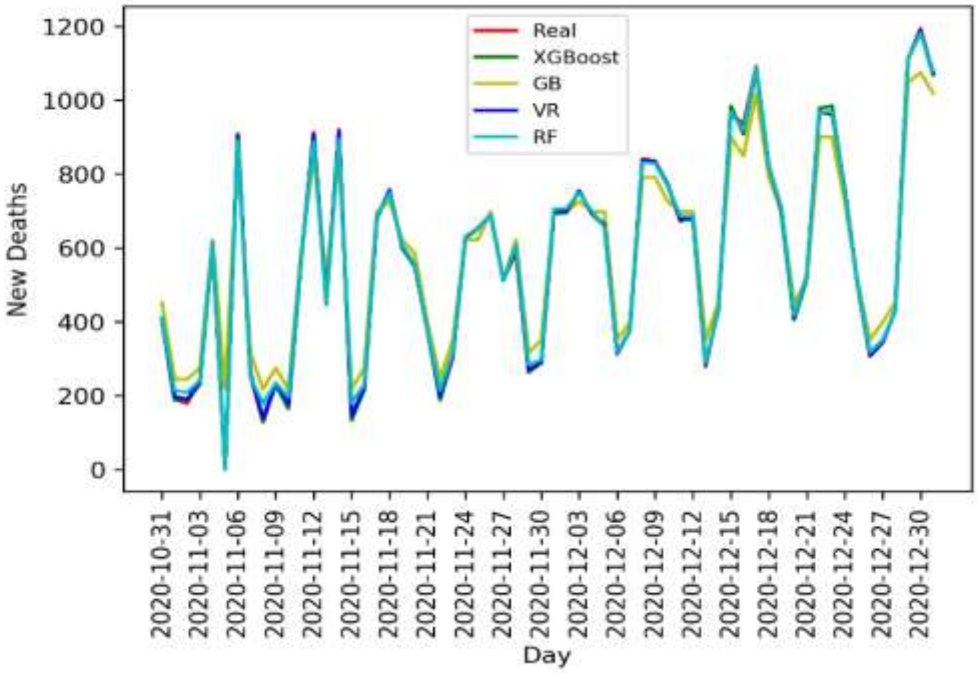

Figure 5: Comparison of new deaths count prediction in Brazil

The values presented in Table 4 show the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for each model. RMSE is considered the primary criterion for prediction models. The values of RMSE indicate that the Voting regressor performed better than the other models in the study. Among other models, XGBoost and GB also performed well. The RF performed well if considered separately but it scored the lowest RMSE in comparison with other models in the study.

Table 4 Evaluation metrics - Brazil

Location	Model	MAE	RMSE
Brazil	Voting	3.10	4.26
	RF	9.13	13.43
	XGBoost	5.89	8.09
	GB	4.42	6.10

Table 4 also presents the MAE values for the models. It is well known that RMSE will always be equal or greater to MAE values as RMSE gives more importance to the biggest errors. Hence, it is useful to compare these two matrices when large errors are undesirable, which is the case in the current study.

The comparison of RMSE and MAE for the models show that the models do not have large residual errors and they can follow the trend of death count and predict accordingly. The Voting regressor, GB and XGBoost show the models have small residual errors but the RF model has the highest difference with its MAE values, hence; the largest residual errors comparatively.

The results clearly show that the tree-based ensemble models, especially Voting regressor can be used to accurately predict the death count in Brazil.

INDIA

The first case for COVID-19 was reported in India on 30th January 2020. The first casualty associated with COVID-19 was reported on 11th March 2020. Similar to Brazil, the COVID-19 cases and deaths in India increased considerably in the year 2020, with the highest death count of 2,003 occurring on 06th June 2020.

Figure presents the predictions of death count in India. The predictions were completed on the COVID-19 dataset, 80% (270 days) for training and 20% for testing (67 days). Figure presents the comparison of the test data with the predictions of models utilized in the study. Similar to Brazil, the comparison shows that all four models perform well in comparison with real data and all four models can detect the trend of the death count. The Voting regressor performs the best in comparison with other models in predicting the upward and downward curve of death count in India as well.

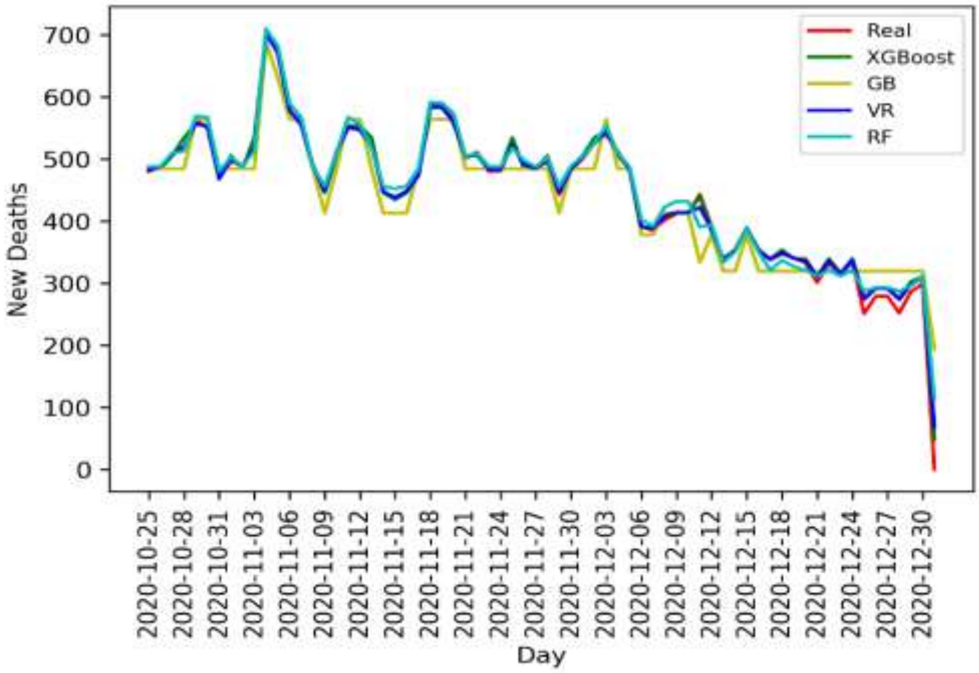

Figure 6: Comparison of new deaths count prediction in India

The values presented in Table 5 show the RMSE and MAE for each model. As discussed, RMSE is considered the primary criterion for prediction models. Similar to Brazil, the values of RMSE indicate that the Voting regressor performed better than the other models in the study. Similarly, among other models, XGBoost and GB also performed well. The RF performed well if considered separately but it scored the lowest RMSE in comparison with other models in the study.

Table 5 Evaluation metrics - India

Location	Model	MAE	RMSE
India	Voting	4.58	10.21
	RF	11.11	18.88
	XGBoost	6.47	9.81
	GB	4.60	7.23

Table 5 also presents the MAE values for the models. The comparison of RMSE and MAE for the models show that the models do not have large residual errors. Similar to Brazil, The models can follow the trend of death count and predict accordingly. The Voting regressor, GB and XGBoost show the models have small residual errors but the RF model has the highest difference with its MAE values, hence; the largest residual errors comparatively.

Similar to Brazil, the results clearly show that the models in the study, especially the Voting regressor can be used to accurately predict the death count in India.

USA

The first case for COVID-19 was reported in the USA on 24th January 2020. The first casuality associated with COVID-19 was reported on 26th February, 2020. Similar to Brazil and India, the COVID-19 cases and deaths in USA increased considerably in the year 2020, with the highest death count of 3729 occurring on 27th December, 2020. Figure presents the predictions of death count in USA. The predictions were completed on the COVID-19 dataset, 80% (276 days) for training and 20% for testing (69 days).

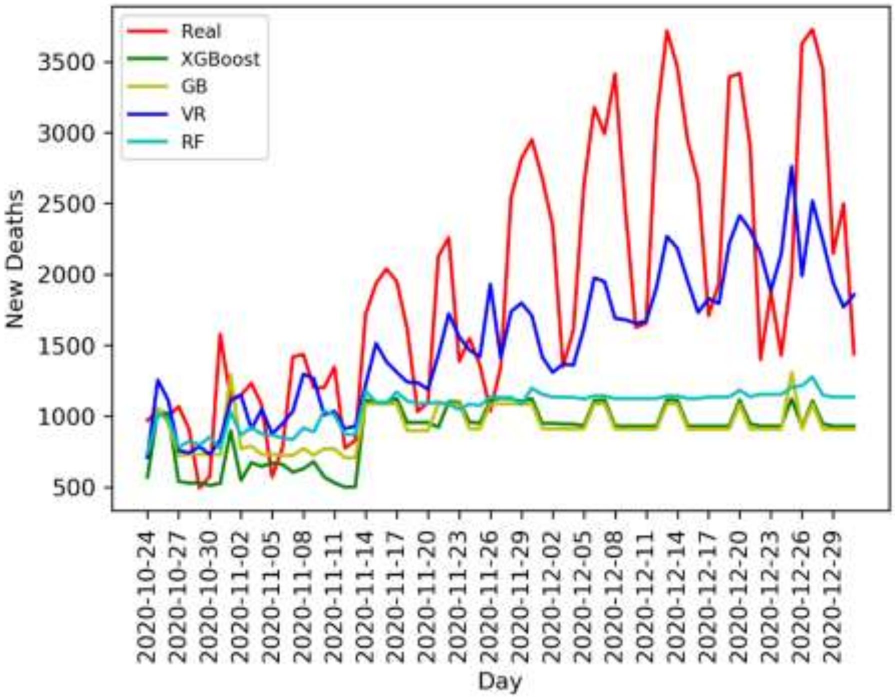

Figure 7 Comparison of new deaths count prediction in USA

Figure presents the comparison of the test data with the predictions of all four models used in the study. Unlike Brazil and India, the comparison shows that all models in the study perform do not perform relatively well in comparison with real data. The models are able to detect the trend of death count, but most RF, GB and XGBoost are unable to follow the peaks in death counts. The Voting regressor performs the good in comparison with other models in predicting the upward and downward curve of death count in USA. The Voting regressor can follow the rise and fall of the death count relatively well.

The values presented in Table 6 show the RMSE and MAE for each model. As discussed, RMSE is considered as the primary criterion for prediction models. Similar to Brazil and India, the values of RMSE indicate that the Voting regressor performed better than the other models in the study. RF, GB, and XGBoost models have large RMSE values which indicate that these models do not provide the best prediction power.

Table 6 Evaluation metrics - USA

Location	Model	MAE	RMSE
USA	Voting	547.51	715.21
	RF	880.00	1169.93
	XGBoost	1017.80	1283.05
	GB	613.79	809.20

Table 6 also presents the MAE values for the models. The comparison of RMSE and MAE for the models shows that the models have large residual errors. Unlike Brazil and India, the RF, GB, and XGBoost models are not able to follow the trend of the death count. The

Voting regressor has a relatively small residual error but RF, GB, and XGBoost models have the highest difference with its MAE values, hence; the largest residual errors comparatively.

The results show that the Voting regressor model can be used to predict the death count in the USA, but there is still room for improvement in the models in USA COVID-19 death count prediction.

Our findings suggest that COVID-19 death count can be predicted accurately using tree-based ensemble models, especially Voting regressors. An interesting aspect of this study is that four different models are used, and they were able to produce good prediction results using the little amount of data available for COVID-19 cases.

5.0 CONCLUSION

Covid-19 has presented unique challenges to all economies around the world by disrupting healthcare systems and limiting economic activity. Governments are closely monitoring the situation and attempting to reduce the number of daily new cases and deaths so that economic recovery can begin. This study attempts to predict the daily number of new deaths per day in the United States, Brazil, and India. The predictions in this study are made using four tree-based ensemble models including GB, RF, XGBoost, and VR. The datasets used in this study for Brazil, India, and the United States span 310, 337, and 345 days, respectively. The datasets for all three countries are divided into 80% train and 20% test splits. The results showed that VR achieved the lowest MAE and RMSE values for all three countries compared to MAE and RMSE values achieved by GB, RF, and XGBoost. VR achieves an MAE of 3.10 for Brazil, 4.58 for India, and 547.51 for the United States. All tree-based ensemble models performed well for Brazil and India, but the prediction results for the United States could be improved. Therefore, future research will employ deep learning-based models to predict daily new deaths using recent datasets.

Funding

Not available

Conflicts of interest/Competing interests

The authors have no conflicts of interest to declare.

Availability of data and material

Upon request

Code availability

Upon request

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent for publication

Not applicable

REFERENCES

1. Greenwood, B., *The contribution of vaccination to global health: past, present and future.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 2014. **369**(1645): p. 20130433-20130433.

2. Braithwaite, J., et al., *Health systems improvement across the globe: success stories from 60 countries*. 2017: CRC Press.

3. Roychoudhury, S., et al., *Viral Pandemics of the Last Four Decades: Pathophysiology, Health Impacts and Perspectives.* International Journal of Environmental Research and Public Health, 2020. **17**(24): p. 9411.

4. Waheed, A. and J. Shafi. *Successful Role of Smart Technology to Combat COVID-19*. in *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. 2020.

5. Tsallis, C. and U. Tirnakli, *Predicting COVID-19 Peaks Around the World.* Frontiers in Physics, 2020. **8**(217).

6. Bathwal, R., et al., *Ensemble Machine Learning Methods for Modeling COVID19 Deaths.* arXiv preprint arXiv:2010.04052, 2020.

7. Rokach, L., *Ensemble Learning: Pattern Classification Using Ensemble Methods*. 2019: World Scientific Publishing Co Pte Ltd.

8. Yates, C. *How to model a Pandemic*. 2020.

9. Menon, A., et al., *Modelling and simulation of COVID-19 propagation in a large population with specific reference to India.* medRxiv, 2020.

10. Vaid, S., C. Cakan, and M. Bhandari, *Using Machine Learning to Estimate Unobserved COVID-19 Infections in North America.* The Journal of Bone and Joint Surgery. American Volume, 2020: p. 10.2106/JBJS.20.00715.

11. Moher, D., et al., *Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement.* Systematic reviews, 2015. **4**(1): p. 1.

12. Da Silva, R.G., et al., *Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables.* Chaos, solitons, and fractals, 2020. **139**: p. 110027-110027.

13. Friston, K.J., et al., *Dynamic causal modelling of COVID-19.* arXiv preprint arXiv:2004.04463, 2020.

14. Evensen, G., et al., *An international assessment of the COVID-19 pandemic using ensemble data assimilation.* medRxiv, 2020.

15. Altieri, N., et al., *Curating a COVID-19 data repository and forecasting county-level death counts in the United States.* arXiv preprint arXiv:2005.07882, 2020.

16. Elsheikh, A.H., et al., *Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia.* Process safety and environmental protection : transactions of the Institution of Chemical Engineers, Part B, 2021. **149**: p. 223-233.

17. Hao, Y., et al., *Prediction and analysis of corona virus disease 2019.* PloS one, 2020. **15**(10): p. e0239960.

18. De Figueiredo, M.P.S., et al., *Long-Term Time Prediction of Cumulative Number of Deaths in Brazil, China, Germany, Italy, Spain, the United States: an application to COVID-19 S-shaped models.* Research, Society and Development, 2020. **9**(8): p. e749986565-e749986565.

19. Jo, H., et al., *condLSTM-Q: A novel deep learning model for predicting Covid-19 mortality in fine geographical Scale.* arXiv preprint arXiv:2011.11507, 2020.

20. Silva, R., et al. *Use of econometrics and machine learning models to predict the number of new cases per day of COVID-19*. in *Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde*. 2020. SBC.

21. Goic, M., et al., *COVID-19: Short-term forecast of ICU beds in times of crisis.* Available at SSRN 3693447, 2020.

22. Ngie, H.M., L. Nderu, and D.G. Mwigereri, *Tree-Based Regressor Ensemble for Viral Infectious Diseases Spread Prediction.*

23. Shastri, S., et al., *Deep-LSTM ensemble framework to forecast Covid-19: an insight to the global pandemic.* International Journal of Information Technology, 2021: p. 1-11.

24. Shastri, S., et al., *CoBiD-net: a tailored deep learning ensemble model for time series forecasting of covid-19.* Spatial Information Research, 2021: p. 1-14.

25. Politis, G. and L. Hadjileontiadis, *Covid19 infection spread in Greece: Ensemble forecasting models with statistically calibrated parameters and stochastic noise.* medRxiv, 2020.