PLOS GLOBAL PUBLIC HEALTH

RESEARCH ARTICLE

Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh

Md. Siddikur Rahman*☯, Arman Hossain Chowdhury☯, Miftahuzzannat Amrin

Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh

☯ These authors contributed equally to this work.
* siddikur@brur.ac.bd

Citation: Rahman M.S, Chowdhury AH, Amrin M (2022) Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh. PLOS Glob Public Health 2(5): e0000495. https://doi.org/10.1371/journal.pgph.0000495

Editor: Abraham D. Flaxman, University of Washington, UNITED STATES

Received: August 29, 2021

Accepted: April 27, 2022

Published: May 18, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgph.0000495

Data Availability Statement: All data are in the manuscript and/or supporting information files.

Funding: The authors received no specific funding for this work.

Abstract

Accurate predictive time series modelling is important in public health planning and response during the emergence of a novel pandemic. Therefore, the aims of the study are three-fold: (a) to model the overall trend of COVID-19 confirmed cases and deaths in Bangladesh; (b) to generate a short-term forecast of 8 weeks of COVID-19 cases and deaths; (c) to compare the predictive accuracy of the Autoregressive Integrated Moving Average (ARIMA) and eXtreme Gradient Boosting (XGBoost) for precise modelling of non-linear features and seasonal trends of the time series. The data were collected from the onset of the epidemic in Bangladesh from the Directorate General of Health Service (DGHS) and Institute of Epidemiology, Disease Control and Research (IEDCR). The daily confirmed cases and deaths of COVID-19 of 633 days in Bangladesh were divided into several training and test sets. The ARIMA and XGBoost models were established using those training data, and the test sets were used to evaluate each model's ability to forecast and finally averaged all the predictive performances to choose the best model. The predictive accuracy of the models was assessed using the mean absolute error (MAE), mean percentage error (MPE), root mean square error (RMSE) and mean absolute percentage error (MAPE). The findings reveal the existence of a nonlinear trend and weekly seasonality in the dataset. The average error measures of the ARIMA model for both COVID-19 confirmed cases and deaths were lower than XGBoost model. Hence, in our study, the ARIMA model performed better than the XGBoost model in predicting COVID-19 confirmed cases and deaths in Bangladesh. The suggested prediction model might play a critical role in estimating the spread of a novel pandemic in Bangladesh and similar countries.

Introduction

The coronavirus disease 2019 (COVID-19) is a major global public health threat. A group of pneumonia infections caused by a newly found β-coronavirus occurred in Wuhan, China in December 2019 [1]. On January 12, 2020, the World Health Organization (WHO) labelled this coronavirus the 2019-novel coronavirus (2019-nCoV) [2, 3]. More than 222 nations, including Bangladesh, have reported more than 263.1 million confirmed COVID-19 cases as of

Competing interests: The authors have declared that no competing interests exist.

November 30, 2021, resulting in 5.2 million fatalities worldwide [4]. On March 8, 2020, IEDCR detected the first COVID-19 case in Bangladesh. On March 9, 2020, the number of infected cases began to rise, and as of December 31, 2021, Bangladesh had 1.6 million infected cases and 28,072 fatalities [5].

In South Asia, especially Bangladesh, COVID-19 has portrayed a significant gap in public health preparedness and response to contagious disease risks and outbreaks [6]. The lack of a dependable public health surveillance system is noticeable [7]. Of the 222 countries, Bangladesh globally ranks 4th on the daily increase of COVID-19 deaths [8] and 3rd in fatality rate in South Asia [9]. Bangladesh is a densely populated country, with almost 161.4 million people living in overcrowded cities and villages, with a population density of over 1115 persons per square kilometre [10]. The healthcare system of Bangladesh is falling short of international standards due to a scarcity of competent workers and inadequate healthcare services, despite the Bangladesh government's efforts to address these challenges in the health service [11]. Furthermore, there are insufficient Intensive Care Unit (ICU) beds for the population. The government faces an uphill battle to control the COVID-19 spread. The Impact of COVID-19 in Bangladesh on education is also noticeable. Due to the lengthy university shutdown and home confinement caused by COVID-19, students' learning was severely disrupted [12]. Students had a higher psychological effect due to COVID-19 [13]. The spread of COVID-19 poses a tremendous challenge for any administration in terms of public health system capacity and management in the event of a catastrophic emergency [14]. As a result, knowing the exact prediction and usual pattern of this virus is crucial for Bangladesh. The prediction model can assist hospitals, healthcare administration and related stakeholders in public health planning and response during the emergence of the COVID-19 pandemic.

The autoregressive integrated moving average (ARIMA) model is commonly used in the modelling of contagious diseases [15], such as influenza viruses [16], malaria [17], and hemorrhagic fever [18]. Several studies regarding COVID-19 forecasting used ARIMA model for predicting the confirmed cases and examined it as the best model [19–22]. On the other hand, the eXtreme Gradient Boosting, a new approach, is an uptrend machine learning technique in time series modelling [23, 24]. The XGBoost model has performed admirably in many medical research sectors [25–28], but the application of XGBoost model in predicting COVID-19 incidence is scanty [29–32]. Time series forecasting methods play a critical role in estimating the spread of an epidemic. Therefore, this study aimed to (a) model the overall trend of COVID-19 confirmed cases and deaths in Bangladesh; (b) generate a short-term forecast of 8 weeks of confirmed COVID-19 cases and deaths; (c) compare the predictive accuracy of the Autoregressive Integrated Moving Average (ARIMA) and eXtreme Gradient Boosting (XGBoost) for precise modelling of non-linear features and seasonal trends of the time series (Fig 1). The findings of this study will help policymakers and government officials with effective public health interventions to control the spread of an epidemic.

Methods

Data source

Daily confirmed cases and deaths of COVID-19 in Bangladesh from March 08, 2020, to November 30, 2021 were collected from the Directorate General of Health Service (DGHS) and Institute of Epidemiology, Disease Control and Research (IEDCR) [33, 34]. The daily confirmed cases and deaths of COVID-19 of 633 days in Bangladesh were divided into several training and test sets. The ARIMA and XGBoost models were established using those training data, and the test sets were used to evaluate each model's ability to forecast and finally averaged all the predictive performances to choose the best model.

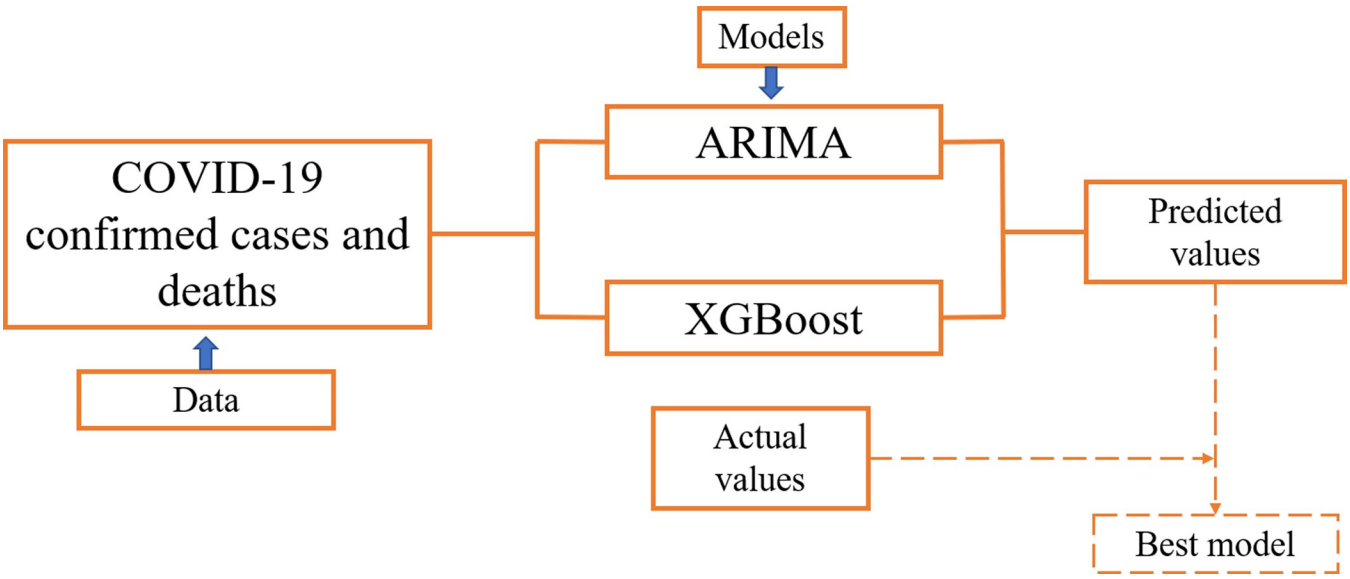

Fig 1. Proposed methodology.

https://doi.org/10.1371/journal.pgph.0000495.g001

ARIMA model

The ARIMA model is frequently used for time series modelling of contagious diseases [35]. It is one of the most often used time-series models in a variety of sectors of data analysis because it accounts for changing trends, periodic variations, and random disturbances in the data. It's utilized for forecasting and better interpreting the data [36]. ARIMA(p, d, q) is a combination of the Autoregressive (AR) and Moving Average (MA) models, with the 'I' standing for integration; where p denotes the autoregressive order, d for differencing order, and q for moving average order [37]. Stationary is a discardable property for a time series analysis. The difference order d is used to make a nonstationary time series to stationary. It is estimated by the Augmented Dickey-Fuller (ADF) test. An ARMA (p, q) model combines AR(p) and MA(q) models, which is best suited to univariate time series analysis. The AR(p) model assumes that a variable's future value is determined by a linear combination of p previous observations plus a random error term. The AR(p) model is represented mathematically as follows:

$$Y_t = C + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \emptyset_3 Y_{t-3} + \emptyset_4 Y_{t-4} \ldots\ldots \emptyset_p Y_{t-p} + \varepsilon_t \quad (1)$$

Y_t and ε_t denote the actual value and error terms at time t, $\emptyset_i$ (i = 1,2,3,4. . ..) denotes model parameters, and c denotes a constant. The order of the model is a positive integer p. Unlike the AR(p) model, the MA(q) model includes a dependent variable for previous errors. Following is the MA(q) model:

$$Y_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \theta_4 \varepsilon_{t-4} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (2)$$

Here, μ denotes the series' mean, θ_j (j = 1, 2, 3. . . q) denotes model parameters, and q is the model's order. A mathematical representation of an ARMA (p, q) model is as follows:

$$Y_t = C + \mu + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \emptyset_3 Y_{t-3} + \emptyset_4 Y_{t-4} \ldots\ldots + \emptyset_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \theta_4 \varepsilon_{t-4} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (3)$$

Seasonal ARIMA model

A seasonal ARIMA model collects information from seasonal components that the conventional ARIMA model cannot comprehend. The seasonal model may be split into two types based on its complexity: an additive model (simple seasonal model) and a product seasonal model. The mathematical expression of the simple seasonal model's is:

$$X_t = S_t + T_t + I_t \tag{4}$$

Where, S_t, T_t, and I_t denote seasonal information, trend information, and random fluctuation information in the data, respectively. To build a seasonal ARIMA model, the components of the nonseasonal part is identified first. After that, the seasonal part is identified. For the seasonal information, the time series data were plotted to see the seasonality pattern. Then the Box-Cox transformation was performed to reduce the variance of the original COVID-19 time series. At the same time, the long term trend and seasonal variations were fixed by performing first-order differencing and seasonal differencing. An Augmented Dickey-Fuller (ADF) test can be used to determine if the time series is stable. The potential values of the autoregressive order p, moving average order q, seasonal autoregressive order P, and the seasonal moving average order Q may be calculated using the graphs of the autocorrelation function (ACF) and partial ACF (PACF) determined by the Box- Jenkins order determination method [38]. The corrected Akaike information criterion (AICc) value was used to evaluate the benefits and drawbacks of the model fit, and the model with its least AICc value was deemed the best. The Ljung-Box test is thus used to determine the white noise of the residuals [18, 38].

XGBoost model

Extreme Gradient boosting (XGBoost) technique is an optimized distributed Gradient boosting library that can rapidly assess the importance of all input features and is a scalable machine learning system for tree boosting. It has proven to be a qualified and competent problem solver for machine learning [39, 40]. Gradient boosting is a popular method for building a forecasting model and a quantifiable boosting algorithm [38]. It was initially developed by Chen Tianqi and Carlos Gestrin in 2011 and has since been improved and polished by numerous scientists in the follow-up study [41]. The core concept of boosting (enhancing machine learning models) is to merge hundreds of low-accuracy prediction models into a single high-accuracy model. Several models must frequently be integrated to obtain good prediction accuracy under tolerable parameter values. The model may need to be iterated or repeated multiple times or more to attain sufficient accuracy if the data collection is vast or complicated; the XGBoost model could better handle this problem [18]. XGBoost is a robust and effective gradient boosting machine algorithm [42, 43]. The objective function can be written as follows:

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant \tag{5}$$

Where y_i is the observed values, $\hat{y}_i^{(t-1)}$ is the predicted value of the last iteration, x_i is the feature vector, n is the sample size, f_t is a new function which model learns, $\Omega(f_t)$ is the regularization term which saves the model from complexity. l denotes the loss function, which calculates the difference between the label and the prediction in the previous phase, the new tree's output [38, 44].

Evaluation parameter of models

A model's real accuracy can be measured by comparing predicted and actual values. A variety of performance metrics can be performed to calculate accuracy [45]. We used four prominent

forecasting parameters to assess the predictive efficacy of our model: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), Mean Percentage Error (MPE), as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{7}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{\hat{y}_i - y_i}{y_i}| \times 100\% \tag{8}$$

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{y}_i - y_i}{y_i}\right) \times 100\% \tag{9}$$

Where n denotes the number of observations, $\hat{y}_i - y_i$ denotes the error between the forecasted and actual value. The mean of the actual forecasting error is calculated by taking the arithmetic average of the absolute errors between the prediction and the actual value. The root mean square error (RMSE) is a commonly used metric for comparing the values forecasted by a model or estimate to the values observed, and it's the average squared error squared. The MAPE measure calculates accuracy as a percentage, computed as the actual values minus the forecasted values divided by the actual values for each time period [46].

Data analysis

Statistical analyses were performed using RStudio (Version 4.1.0) [47]. The 'tseries', 'TSstudio' and stats packages were used to process the time series. ARIMA models were built with the 'forecast' package using auto.arima function for choosing the best model based on the AICc values [48]. The 'forecastxgb' package was used for fitting XGBoost model. The necessary codes are available at https://github.com/ [49].

Results

In Bangladesh, 1.6 million cases and 27,983 deaths of COVID-19 of 633 days (91 weeks) were recorded from March 08, 2020 to November 30, 2021. The highest COVID-19 confirmed cases were recorded 16,230 and deaths 264 in Bangladesh (Table 1). The data vary considerably and show weekly seasonality and nonlinearity pattern in both cases and deaths. Although the number of confirmed cases and deaths fluctuated in different weeks, there was a highly upward trend between 70 and 80 weeks. After that, it began to alleviate (Fig 2). The ADF test confirms that the data are not smooth. The entire data set (COVID-19 confirmed cases and deaths) was split into several training and test sets (S1 Text).

Table 1. Summary of COVID-19 confirmed cases and deaths count during March 08- November 30, 2021.

Variables	Minimum	Maximum	Mean±SD	Total
Confirmed cases	0	16,230	2490±2938.7	1,576,284
Deaths	0	264	45.2±54.5	27,981

SD: Standard deviation.

https://doi.org/10.1371/journal.pgph.0000495.t001

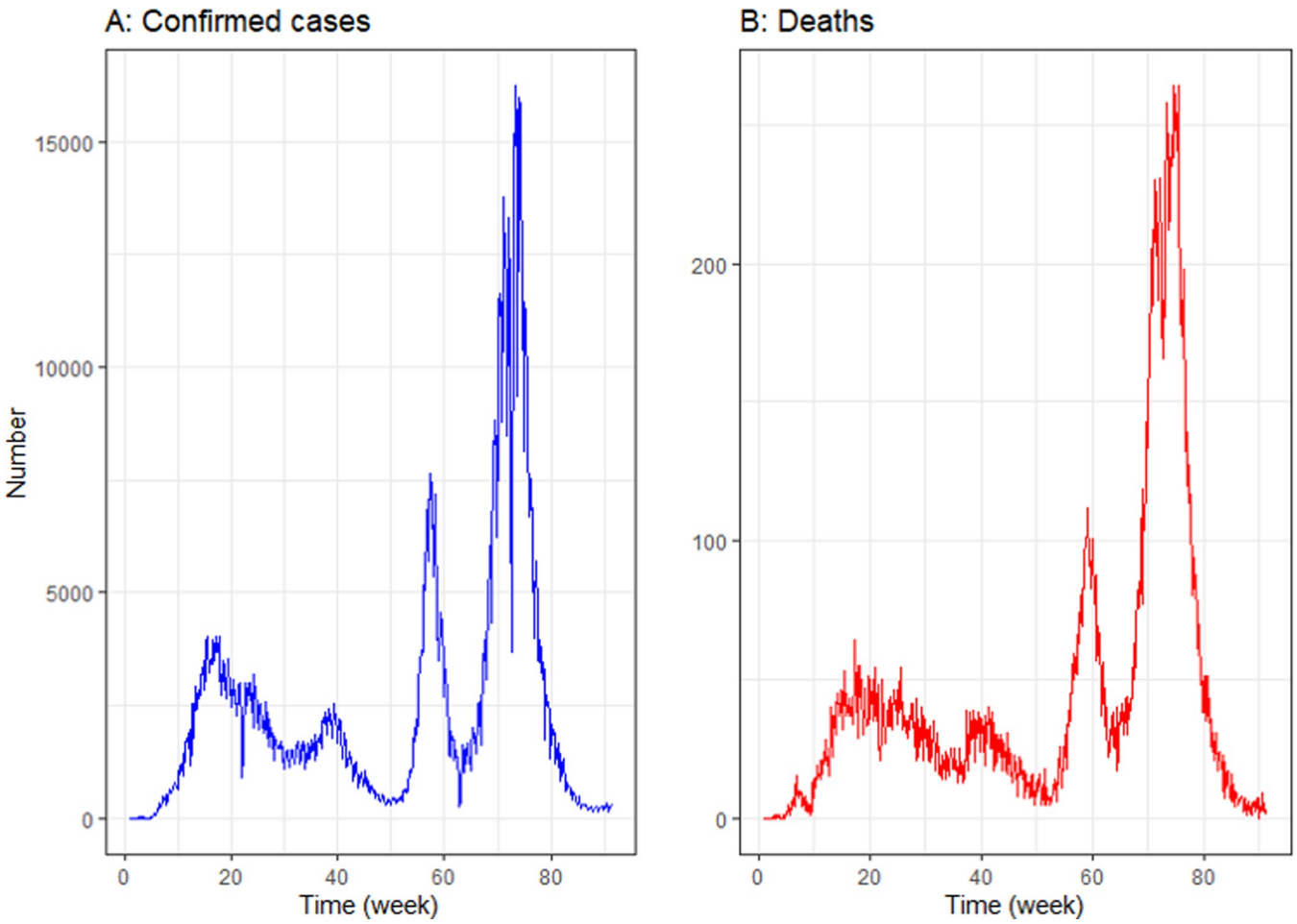

Fig 2. A 633 day (91 weeks) time series plot for confirmed COVID-19 cases and deaths in Bangladesh from March 08, 2020 to November 30, 2021.

https://doi.org/10.1371/journal.pgph.0000495.g002

To decrease anomalies such as non-normality and heteroscedasticity that the variances are not constant, Box & Cox (1964) presented a parametric power transformation technique [50]. The Box-Cox transformation was applied to each training data set to remove the non-normality and exhibit less variation [51]. The decomposed data shows a weekly seasonal pattern in both cases and deaths [52]. Table 2 illustrates the predictive performance of different ARIMA models built from seven different training sets and their average values for COVID-19 confirmed cases.

The XGBoost model for COVID-19 confirmed cases were built by adjusting different parameters like seas_method= 'dummies', trend_method= 'none', power transformation parameter 'lambda' for each training set. Table 3 illustrates the predictive performances of different training and test sets of the XGBoost model and their average values for confirmed cases.

For COVID-19 death data, we built five different ARIMA models for five different training and test sets. The appropriate model for each training data set was selected based on the AICc value. The predictive performances of ARIMA models of five different training and test data sets and their average values were shown in Table 4.

We also built the XGBoost model for COVID-19 deaths by adjusting the parameters seas_method= 'dummies', trend_method= 'none', power transformation parameter 'lambda' for

Table 2. Evaluation of parameters for the ARIMA model of different training and test sets for COVID-19 confirmed cases.

ARIMA model	Train				Test			
	RMSE	MAE	MPE	MAPE	RMSE	MAE	MPE	MAPE
Sample 1	258.50	166.16	-2.53	15.00	421.92	359.07	-2.63	22.46
Sample 2	241.82	154.73	-2.72	13.39	244.33	216.94	-47.26	48.85
Sample 3	224.11	139.44	-2.67	12.80	3844.89	2988.57	75.95	75.95
Sample 4	280.81	165.97	-1.72	12.50	2160.95	2031.37	-163.43	163.44
Sample 5	275.28	173.14	-1.96	12.82	6325.54	5481.82	49.68	50.74
Sample 6	560.19	276.18	-2.15	13.11	8984.24	8217.15	-549.35	549.35
Sample 7	558.91	280.01	-2.47	12.82	67.31	55.24	1.02	21.36
Average error measures	342.80	193.66	-2.32	13.21	3149.88	2764.31	-90.86	133.16

ARIMA: Autoregressive Integrated Moving Average; RMSE: Root Mean Square Error; MAE: Mean Absolute Error; MPE: Mean Percentage Error; MAPE: Mean Absolute Percentage Error.

https://doi.org/10.1371/journal.pgph.0000495.t002

each training set. The predictive measures of different training and test set for XGBoost model and their average values for COVID-19 deaths were shown in Table 5.

The average MAPE values of the ARIMA model for COVID-19 confirmed cases is comparatively lower than the XGBoost model indicating that ARIMA performs better than XGBoost in predicting COVID-19 confirmed cases in Bangladesh. On the other hand, the average MAPE value of the ARIMA model for COVID-19 deaths is smaller than XGBoost which also indicates that ARIMA performs better than XGBoost in predicting COVID-19 deaths in Bangladesh.

In our study, it was found that ARIMA model performs better than XGBoost in predicting COVID-19 confirmed cases and deaths in Bangladesh. The detailed procedure of ARIMA and XGBoost model fitting for COVID-19 confirmed cases and deaths were shown in S1 Text.

Discussion

In our study, we found a weekly seasonality for daily COVID-19 confirmed cases and deaths in Bangladesh. Because of the weekend, fewer health care staffs were available to report new cases or fewer people are tested, which causes weekly seasonality [11, 53]. It was simpler to assess the seasonality and pattern of this disease using seasonal decomposition, which offered a reference for us to analyze, process, and stabilize data, laying the groundwork for building a

Table 3. Evaluation of parameters for the XGBoost model of different training and test sets for COVID-19 confirmed cases.

XGBoost model	Train				Test			
	RMSE	MAE	MPE	MAPE	RMSE	MAE	MPE	MAPE
Sample 1	47.19	31.66	-0.11	2.30	520.76	436.20	11.44	23.97
Sample 2	31.39	22.05	-0.16	1.72	925.60	865.12	-190.79	190.81
Sample 3	64.91	42.74	0.36	3.10	3727.73	2874.31	71.05	71.12
Sample 4	76.87	51.28	-0.11	3.64	1989.87	1814.20	-156.36	156.56
Sample 5	53.71	35.71	-0.24	2.53	7374.24	6413.23	59.46	59.46
Sample 6	130.08	80.66	-0.17	4.06	9683.20	9183.94	-561.51	561.55
Sample 7	168.82	105.89	-0.14	4.64	196.60	185.76	-76.20	76.20
Average error measures	81.85	52.86	-0.08	3.14	3488.29	3110.39	-120.42	162.81

XGBoost: eXtreme Gradient Boosting; RMSE: Root Mean Square Error; MAE: Mean Absolute Error; MPE: Mean Percentage Error; MAPE: Mean Absolute Percentage Error.

https://doi.org/10.1371/journal.pgph.0000495.t003

Table 4. Evaluation of parameters for the ARIMA models of different training and test sets for COVID-19 deaths.

ARIMA model	Train				Test			
	RMSE	MAE	MPE	MAPE	RMSE	MAE	MPE	MAPE
Sample 1	5.84	4.35	-6.69	23.46	46.32	33.48	49.72	69.15
Sample 2	6.14	4.59	-4.71	23.09	110.31	101.42	-287.79	286.79
Sample 3	7.23	5.37	-4.64	24.34	100.32	90.70	47.81	47.81
Sample 4	9.98	6.67	-4.35	22.51	246.09	228.36	-620.60	620.60
Sample 5	9.64	6.60	-4.98	21.22	6.00	5.40	-152.45	154.40
Average error measures	7.77	5.52	-5.07	22.92	101.81	91.87	-192.66	235.75

ARIMA: Autoregressive Integrated Moving Average; RMSE: Root Mean Square Error; MAE: Mean Absolute Error; MPE: Mean Percentage Error; MAPE: Mean Absolute Percentage Error.

https://doi.org/10.1371/journal.pgph.0000495.t004

mathematical model. The ARIMA models for cases and deaths were created using a linear regression model to expose the data's dynamic rules and forecast future data values. The ARIMA model combines the trend components, cyclical factors, and random errors originally included in the time series. This model combines the benefits of autoregressive and moving average models, is unconstrained by data sources, has high adaptability, and has good short-term predictions [18]. Instead of requiring particular influencing elements, the ARIMA model uses merely historical data to comprehend the illness pattern and achieve a more accurate forecast impact. As a result, the ARIMA approach is simple to learn and frequently employed [38]. In this study, the ARIMA method is compared to the XGBoost model for its fairly mature time series prediction approach and widespread application. The ARIMA model performs well on the nonstationary time series after applying Box-Cox transformation and differencing adjustments, demonstrating the model's capacity to forecast diseases. In general, the greater the number of differences utilized, the more data is lost. We built different ARIMA models for different training sets and selected the best for each training set based on the AICc value for both confirmed cases and deaths [53, 54]. Finally, we averaged all the error measures from all models. The average MAPE value of the training data sets for confirmed cases was 13.21%, whereas it was 133.16% for the test data sets. On the other hand, the average MAPE value of the training sets for death data was 22.92%, whereas the test sets was 235.95%. On the other hand, we used the most popular machine learning model to fit the nonlinear data [55]. The XGBoost model, a relatively new approach, is a gradient boosting-based ensemble machine learning technique that utilizes decision trees. The XGBoost technique offers several benefits in terms of model prediction, including the lack of data preprocessing, a quick operation speed, complete feature

Table 5. Evaluation of parameters for the XGBoost models of different training and test sets for COVID-19 deaths.

XGBoost model	Train				Test			
	RMSE	MAE	MPE	MAPE	RMSE	MAE	MPE	MAPE
Sample 1	2.19	1.45	0.32	6.34	40.32	28.11	20.68	63.18
Sample 2	1.95	1.39	-1.03	6.65	49.17	45.01	-131.30	132.18
Sample 3	3.80	2.66	1.69	10.05	150.98	136.59	71.82	72.82
Sample 4	2.70	1.92	-1.49	7.37	179.88	169.85	-444.68	445.68
Sample 5	3.20	2.27	-1.18	7.50	16.27	15.47	-470.79	471.27
Average error measures	2.77	1.94	-0.34	7.58	87.32	79.01	-190.85	237.03

XGBoost: eXtreme Gradient Boosting; RMSE: Root Mean Square Error; MAE: Mean Absolute Error; MPE: Mean Percentage Error; MAPE: Mean Absolute Percentage Error.

https://doi.org/10.1371/journal.pgph.0000495.t005

extraction, a strong fitting effect, and high prediction accuracy. This study applied this new technique to predict COVID-19 confirmed cases and deaths in Bangladesh. We selected the most often used ARIMA time series model as the baseline of this study. But the XGBoost model did not perform well on the nonlinear data. The XGBoost model has a considerably worse influence on forecasting than the ARIMA model in this COVID-19 research in Bangladesh because the number of confirmed cases and deaths has increased significantly between 70 and 80 weeks. The number of confirmed cases in the country has also altered dramatically due to changes in government policies. In addition, there might have other climatic and environmental factors that impact the COVID-19 incidence observed from some previous studies which didn't incorporate in our study [46, 56–58]. As a result, the proposed model was no longer produced accurate predictions for this change. In this study, we compared the models' predictive performances to provide a reference for the country's policymakers to take effective steps and strategies to control the outbreak of the deadly disease. The study findings are useful to all other endemic countries similar to Bangladesh.

Conclusion

For controlling the spread of the COVID-19 pandemic in Bangladesh and similar settings elsewhere, we developed a seasonal ARIMA model and XGBoost model. These models were used to create short-term forecasts in this study. The ARIMA model performed better than the XGBoost model in predicting COVID-19 confirmed cases and deaths in Bangladesh.

Limitations

We compared the predictive performance of XGBoost and ARIMA models in this study, and the results help choose the best model for COVID-19 prediction in Bangladesh. There are many different prediction models, and we need to keep experimenting with them to find the best one for predicting confirmed COVID-19 cases and deaths. We focused on the impact of time on both cases and deaths in our research, which allows our model easier to build and forecast. Therefore, a limitation of our study is that, for example, meteorological data such as temperature, humidity, and wind speed variables were not incorporated but which are known to impact COVID-19. As mentioned above, this will be explored progressively with increasing data.

Supporting information

S1 Data. Time series COVID-19 data of Bangladesh from March 08, 2020 to November 30, 2021.
(XLSX)

S1 Text. Figs A–J, Tables A–B.
(DOCX)

Acknowledgments

The researchers are very grateful to the Directorate General of Health Service (DGHS) and the Institute of Epidemiology, Disease Control and Research (IEDCR) for providing COVID-19 data.

Author Contributions

Conceptualization: Md. Siddikur Rahman.

Data curation: Md. Siddikur Rahman, Arman Hossain Chowdhury, Miftahuzzannat Amrin.

Formal analysis: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Investigation: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Methodology: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Project administration: Md. Siddikur Rahman.

Resources: Md. Siddikur Rahman, Arman Hossain Chowdhury, Miftahuzzannat Amrin.

Software: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Supervision: Md. Siddikur Rahman.

Visualization: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Writing – original draft: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Writing – review & editing: Md. Siddikur Rahman, Arman Hossain Chowdhury, Miftahuzzannat Amrin.

References

1. Peeri NC, Shrestha N, Siddikur Rahman M, Zaki R, Tan Z, Bibi S, et al. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? Int J Epidemiol. 2020; 49: 717–726. https://doi.org/10.1093/ije/dyaa033 PMID: 32086938
2. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020; 395: 565–574. https://doi.org/10.1016/S0140-6736(20)30251-8 PMID: 32007145
3. Ruan S. Likelihood of survival of coronavirus disease 2019. Lancet Infect Dis. 2020; 20: 630–631. https://doi.org/10.1016/S1473-3099(20)30257-7 PMID: 32240633
4. Coronavirus Disease (COVID-19) Situation Reports. [cited 30 Nov 2021]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
5. COVID-19 Situation Updates | IEDCR. [cited 30 Nov 2021]. Available: https://iedcr.gov.bd/covid-19/covid-19-situation-updates
6. Rahman MS, Karamehic-Muratovic A, Amrin M, Chowdhury AH, Selim Mondol M, Haque U, et al. COVID-19 Epidemic in Bangladesh among Rural and Urban Residents: An Online Cross-Sectional Survey of Knowledge, Attitudes, and Practices. 2020. https://doi.org/10.3390/epidemiologia201
7. Bhutta ZA, Basnyat B, Saha S, Laxminarayan R. Covid-19 risks and response in South Asia. BMJ. 2020; 368: 1–2. https://doi.org/10.1136/bmj.m1190 PMID: 32213488
8. Fatalities hit yet another high in Bangladesh as 258 die of Covid in a day | Dhaka Tribune. [cited 28 Aug 2021]. Available: https://archive.dhakatribune.com/bangladesh/2021/07/27/fatalities-hit-yet-another-high-in-bangladesh-as-258-die-of-covid-in-a-day
9. Bangladesh Covid case fatality rate third in South Asia. [cited 28 Aug 2021]. Available: https://www.newagebd.net/article/144760/bangladesh-covid-case-fatality-rate-third-in-south-asia
10. Satu MS, Howlader KC, Mahmud M, Shamim Kaiser M, Islam SMS, Quinn JMW, et al. Short-term prediction of covid-19 cases using machine learning models. Appl Sci. 2021;11. https://doi.org/10.3390/app11094266
11. Ahmed SM, Hossain MA, RajaChowdhury AM, Bhuiya AU. The health workforce crisis in Bangladesh: Shortage, inappropriate skill-mix and inequitable distribution. Hum Resour Health. 2011; 9: 1–7. https://doi.org/10.1186/1478-4491-9-1 PMID: 21223546
12. Bari R, Sultana F. Second Wave of COVID-19 in Bangladesh: An Integrated and Coordinated Set of Actions Is Crucial to Tackle Current Upsurge of Cases and Deaths. Front Public Heal. 2021; 9: 1275. https://doi.org/10.3389/FPUBH.2021.699918/BIBTEX
13. Faisal RA, Jobe MC, Ahmed O, Sharker T. Mental Health Status, Anxiety, and Depression Levels of Bangladeshi University Students During the COVID-19 Pandemic. Int J Ment Health Addict. 2021. https://doi.org/10.1007/s11469-020-00458-y PMID: 33424514

14. Li J, Guo K, Viedma EH, Lee H, Liu J, Zhong N, et al. Culture versus Policy: More Global Collaboration to Effectively Combat COVID-19. Innovation(China). 2020; 1: 100023. https://doi.org/10.1016/j.xinn.2020.100023 PMID: 32914139
15. Zhang X, Hou F, Qiao Z, Li X, Zhou L, Liu Y, et al. Temporal and long-term trend analysis of class C notifiable diseases in China from 2009 to 2014. BMJ Open. 2016; 6: 11038. https://doi.org/10.1136/bmjopen-2016-011038 PMID: 27797981
16. He Z, Tao H. Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study. Int J Infect Dis. 2018; 74: 61–70. https://doi.org/10.1016/j.ijid.2018.07.003 PMID: 29990540
17. Anwar MY, Lewnard JA, Parikh S, Pitzer VE. Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence. Malar J. 2016; 15: 1–10. https://doi.org/10.1186/s12936-015-1044-1 PMID: 26729363
18. Wang T, Liu J, Zhou Y, Cui F, Huang Z, Wang L, et al. Prevalence of hemorrhagic fever with renal syndrome in Yiyuan County, China, 2005-2014. BMC Infect Dis. 2016; 16: 69. https://doi.org/10.1186/s12879-016-1404-7 PMID: 26852019
19. Alzahrani SI, Aljamaan IA, Al-Fakih EA. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. J Infect Public Health. 2020; 13: 914–919. https://doi.org/10.1016/j.jiph.2020.06.001 PMID: 32546438
20. Khan FM, Gupta R. ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. J Saf Sci Resil. 2020; 1: 12–18. https://doi.org/10.1016/j.jnlssr.2020.06.007
21. Singh S, Parmar KS, Makkhan SJS, Kaur J, Peshoria S, Kumar J. Study of ARIMA and least square support vector machine (LS-SVM) models for the prediction of SARS-CoV-2 confirmed cases in the most affected countries. Chaos, Solitons & Fractals. 2020; 139: 110086. https://doi.org/10.1016/j.chaos.2020.110086 PMID: 32834622
22. Hernandez-Matamoros A, Fujita H, Hayashi T, Perez-Meana H. Forecasting of COVID19 per regions using ARIMA models and polynomial functions. Appl Soft Comput. 2020; 96: 106610. https://doi.org/10.1016/j.asoc.2020.106610 PMID: 32834798
23. Rahman MS, Pientong C, Zafar S, Ekalaksananan T, Paul RE, Haque U, et al. Mapping the spatial distribution of the dengue vector Aedes aegypti and predicting its abundance in northeastern Thailand using machine-learning approach. One Heal. 2021; 13: 100358. https://doi.org/10.1016/j.onehlt.2021.100358 PMID: 34934797
24. Li Z, Wang Z, Song H, Liu Q, He B, Shi P, et al. Application of a hybrid model in predicting the incidence of tuberculosis in a Chinese population. Infect Drug Resist. 2019; 12: 1011–1020. https://doi.org/10.2147/IDR.S190418 PMID: 31118707
25. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Informatics Decis Mak 2019 191. 2019; 19: 1–15. https://doi.org/10.1186/s12911-019-0918-5 PMID: 31694707
26. Liu L, Yu Y, Fei Z, Li M, Wu F-X, Li H-D, et al. An interpretable boosting model to predict side effects of analgesics for osteoarthritis. BMC Syst Biol 2018 126. 2018; 12: 29–38. https://doi.org/10.1186/s12918-018-0624-4 PMID: 30463545
27. Liu Z, Zhou T, Han X, Lang T, Liu S, Zhang P, et al. Mathematical models of amino acid panel for assisting diagnosis of children acute leukemia. J Transl Med 2019 171. 2019; 17: 1–11. https://doi.org/10.1186/s12967-019-1783-9 PMID: 30674317
28. Zou LS, Erdos MR, Taylor DL, Chines PS, Varshney A, Parker SCJ, et al. BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. BMC Genomics 2018 191. 2018; 19: 1–15. https://doi.org/10.1186/s12864-018-4766-y PMID: 29792182
29. Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell. 2020; 2: 283–288. https://doi.org/10.1038/s42256-020-0180-7
30. Li WT, Ma J, Shende N, Castaneda G, Chakladar J, Tsai J, et al. Using Machine Learning of Clinical Data to Diagnose COVID-19. BMC Med Informatics Decis Mak. 2020; 20: 247. https://doi.org/10.1101/2020.06.24.20138859
31. Chowdhury MEH, Rahman T, Khandakar A, Al-Madeed S, Zughaier SM, Doi SAR, et al. An Early Warning Tool for Predicting Mortality Risk of COVID-19 Patients Using Machine Learning. Cognit Comput. 2021. https://doi.org/10.1007/s12559-020-09812-7 PMID: 33897907
32. Romeo L, Frontoni E. A Unified Hierarchical XGBoost Model for Classifying Priorities for COVID-19 Vaccination Campaign. Pattern Recognit. 2021; 121: 108197. https://doi.org/10.1016/j.patcog.2021.108197 PMID: 34312570
33. COVID-19. [cited 30 Nov 2021]. Available: http://dashboard.dghs.gov.bd/webportal/pages/covid19.php

34. IEDCR. [cited 30 Nov 2021]. Available: http://old.iedcr.gov.bd/
35. Sahai AK, Rath N, Sood V, Singh MP. ARIMA modelling & forecasting of COVID-19 in top five affected countries. Diabetes Metab Syndr Clin Res Rev. 2020; 14: 1419–1427. https://doi.org/10.1016/J.DSX.2020.07.042 PMID: 32755845
36. Brockwell PJ, Davis RA. Introduction to Time Series and Forecasting. 2016. https://doi.org/10.1007/978-3-319-29854-2
37. Kumar N, Susan S. COVID-19 Pandemic Prediction using Time Series Forecasting Models. 2020 11th Int Conf Comput Commun Netw Technol ICCCNT 2020. 2020. https://doi.org/10.1109/ICCCNT49239.2020.9225319
38. Lv CX, An SY, Qiao BJ, Wu W. Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. BMC Infect Dis. 2021; 21: 1–13. https://doi.org/10.1186/s12879-020-05706-z PMID: 33390160
39. Zheng Y, Zhu Y, Ji M, Wang R, Liu X, Zhang M, et al. A Learning-Based Model to Evaluate Hospitalization Priority in COVID-19 Pandemics. Patterns. 2020; 1: 100092. https://doi.org/10.1016/j.patter.2020.100092 PMID: 32838344
40. Hu CA, Chen CM, Fang YC, Liang SJ, Wang HC, Fang WF, et al. Using a machine learning approach to predict mortality in critically ill influenza patients: a cross-sectional retrospective multicentre study in Taiwan. BMJ Open. 2020; 10: e033898. https://doi.org/10.1136/bmjopen-2019-033898 PMID: 32102816
41. Li W, Yin Y, Quan X, Zhang H. Gene Expression Value Prediction Based on XGBoost Algorithm. Front Genet. 2019; 10: 1–7. https://doi.org/10.3389/fgene.2019.00001 PMID: 30804975
42. Shrivastav LK, Jha SK. A gradient boosting machine learning approach in modeling the impact of temperature and humidity on the transmission rate of COVID-19 in India. Appl Intell. 2021; 51: 2727–2739. https://doi.org/10.1007/s10489-020-01997-6 PMID: 34764559
43. Babajide Mustapha I, Saeed F. Bioactive Molecule Prediction Using Extreme Gradient Boosting. Molecules. 2016; 21: 1–11. https://doi.org/10.3390/molecules21080983 PMID: 27483216
44. Nishio M, Nishizawa M, Sugiyama O, Kojima R, Yakami M, Kuroda T, et al. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. PLoS One. 2018; 13: 1–13. https://doi.org/10.1371/journal.pone.0195875 PMID: 29672639
45. Prajapati S, Swaraj A, Lalwani R, Narwal A, Verma K, Singh G. Comparison of Traditional and Hybrid Time Series Models for Forecasting COVID-19 Cases. 2019;8.
46. Luo J, Zhang Z, Fu Y, Rao F. Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. Results Phys. 2021; 27: 104462. https://doi.org/10.1016/j.rinp.2021.104462 PMID: 34178594
47. RStudio: Integrated Development Environment for R RStudio Team. In: RStudio, PBC, Boston, MA (2021).
48. Hyndman RJ, Khandakar Y. Automatic Time Series Forecasting: The forecast Package for R. J Stat Softw. 2008; 27: 1–22. https://doi.org/10.18637/JSS.V027.I03
49. Arman-Hossain-Chowdhury/Time-series. [cited 29 Dec 2021]. Available: https://github.com/Arman-Hossain-Chowdhury/Time-series
50. Sakia RM. The Box-Cox Transformation Technique: A Review. Stat. 1992; 41: 169. https://doi.org/10.2307/2348250
51. Curran-Everett D. Explorations in statistics: The log transformation. Adv Physiol Educ. 2018; 42: 343–347. https://doi.org/10.1152/advan.00018.2018 PMID: 29761718
52. Rosselló J, Sansó A. Yearly, monthly and weekly seasonality of tourism demand: A decomposition analysis. Tour Manag. 2017; 60: 379–389. https://doi.org/10.1016/j.tourman.2016.12.019
53. Dehning J, Zierenberg J, Spitzner FP, Wibral M, Neto JP, Wilczek M, et al. Inferring change points in the COVID-19 spreading reveals the effectiveness of interventions. Science (80-). 2020;369. https://doi.org/10.1126/science.abb9789 PMID: 32414780
54. Zeng Q, Li D, Huang G, Xia J, Wang X, Zhang Y, et al. Time series analysis of temporal trends in the pertussis incidence in Mainland China from 2005 to 2016. Sci Rep. 2016; 6: 1–8. https://doi.org/10.1038/s41598-016-0001-8 PMID: 28442746
55. Wu W, Guo J, An S, Guan P, Ren Y, Xia L, et al. Comparison of two hybrid models for forecasting the incidence of hemorrhagic fever with renal syndrome in Jiangsu Province, China. PLoS One. 2015; 10: 1–13. https://doi.org/10.1371/journal.pone.0135492 PMID: 26270814
56. Pal SK, Masum MH. Effects of meteorological parameters on COVID-19 transmission trends in Bangladesh. Environ Sustain 2021 43. 2021; 4: 559–568. https://doi.org/10.1007/S42398-021-00195-5

57. Menebo MM. Temperature and precipitation associate with Covid-19 new daily cases: A correlation study between weather and Covid-19 pandemic in Oslo, Norway. Sci Total Environ. 2020; 737: 139659. https://doi.org/10.1016/j.scitotenv.2020.139659 PMID: 32492607

58. Hossain MS, Ahmed S, Uddin MJ. Impact of weather on COVID-19 transmission in south Asian countries: An application of the ARIMAX model. Sci Total Environ. 2021; 761: 143315. https://doi.org/10.1016/j.scitotenv.2020.143315 PMID: 33162141