Heliyon 9 (2023) e13483

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Improved autoregressive integrated moving average model for COVID-19 prediction by using statistical significance and clustering techniques

Saratu Yusuf Ilu, Rajesh Prasad *

Department of Computer Science, African University of Science and Technology, Abuja, Nigeria

ARTICLE INFO

Keywords:
Coronavirus
ARIMA
Machine learning
Prediction
Feature selection
And clustering

ABSTRACT

Purpose: The COVID-19 pandemic has affected more than 192 countries. The condition results in a respiratory illness (e.g., influenza) with signs and symptoms such as cold, cough, fever, and breathing difficulties. Predicting new instances of COVID-19 is always a challenging task.
Methods: This study improved the autoregressive integrated moving average (ARIMA)-based time series prediction model by incorporating statistical significance for feature selection and k-means clustering for outlier detection. The accuracy of the improved model (ARIMAI) was examined using World Health Organization's official data on the COVID-19 pandemic worldwide and compared with that of many modern, cutting-edge algorithms.
Results: The ARIMAI model (RSS score = 0.279, accuracy = 97.75%) outperformed the current ARIMA model (RSS score = 0.659, accuracy = 93%).
Conclusions: The ARIMAI model is not only an efficient but also a rapid and simple technique to forecast COVID-19 trends. The usage of this model enables the prediction of any disease that will affect patients in the future pandemics.

1. Introduction

In December 2019, numerous individuals in Wuhan, Hubei Province, China, began experiencing severe health problems. This is the period when the COVID-19 pandemic started. According to the World Health Organization (WHO), as of June 22, 2022, approximately 545,891,254 cases and 6,343,938 fatalities for COVID-19 have been reported. During the early pandemic, to reduce incidence and mortality rates, the WHO promoted self-quarantine and isolation of people who were ill, which resulted in the largest lockdown in history [1–3]. On June 3, 2020, a total of 188 countries and regions reported having over six million COVID-19 cases [3,4]. The most common symptoms of COVID-19 include fever, respiratory problems, fatigue, and loss of taste and smell [5,6]. In advanced cases, multiorgan failure, septic syncope, acute respiratory distress syndrome, and blood clots are observed. Although adverse symptoms are often observed after approximately five days, they can intensify between two and fourteen days [7]. Older and young individuals and those with heart disease, obesity, or diabetes are the most vulnerable to COVID-19 [8,9].

COVID-19 has resulted in the loss of jobs worldwide. The gross domestic product is predicted to decrease by 3% globally, a situation that is significantly worse compared with the financial crisis of 2008 and 2009. According to the International Labor Organization, approximately half of the world's workforce is facing job loss and hundreds of millions of businesses are in the danger of running out of

* Corresponding author.
E-mail address: rprasad@aust.edu.ng (R. Prasad).

https://doi.org/10.1016/j.heliyon.2023.e13483
Received 7 July 2022; Received in revised form 28 January 2023; Accepted 31 January 2023
Available online 3 February 2023

business [10].

COVID-19 has killed millions of people and continues to incarcerate individuals globally. This disease can be prevented by regularly washing hands, wearing a mask on the mouth, isolating oneself, and avoiding contact with other individuals. However, these precautionary measures are insufficient. Preventing this illness is the only means to avoid it.

Contact tracing was a preventative and containment measure employed by various health systems and governments because of the nonavailability of COVID-19 vaccines until August 2020. On some occasions, the infected person either forgets or does not have their contact information. By performing a literature search in Google Scholar, ScienceDirect, PubMed, Web of Science, and IEEE and evaluating WHO COVID-19 reports and guidelines, a previous study [11] examined the benefits and drawbacks of using new technology for COVID-19 contact tracing. Although that study reported favorable results, the technologies used still had to overcome various obstacles, including technical limitations, dealing with patients with no symptoms, and lack of IT infrastructure support and electronic health policies.

The COVID-19 pandemic is a global public health problem [12,13]. Numerous epidemiological models have been used to determine the dynamics of transmission, techniques, and approaches under some presumptions. Because most of these models employ hypothetical criteria, the accuracy of forecasting future COVID-19 cases is low.

Diverse containment strategies were implemented in various countries, including social exclusion, travel restrictions on tourists from high-risk regions, quarantine for new arrivals from high-risk regions, and shutdown of offices and educational facilities. On March 12, 2020, the Turkish government declared that starting from March 16, all schools and colleges would be closed. Turkey implemented various measures to limit people's freedom of movement. People aged over 65 years; those with immune system deficiencies, asthma, chronic obstructive pulmonary disease, chronic cardiovascular disease, chronic renal diseases, hypertension, and chronic liver disease; and those receiving immune-suppressing medications were prohibited from leaving their homes or using public transportation [14].

On January 11, 2020, the genetic sequence of COVID-19 was made public, which immediately spurred a global response to speed up the creation of a vaccine and prepare for an outbreak. Since then, the unprecedented collaboration between governments and the global pharmaceutical industry has accelerated the development of vaccines. By June 2020, corporations, governments, nongovernmental organizations, and academic research teams had committed tens of billions of dollars to the development of numerous vaccine candidates as well as international immunization programs. According to the Coalition for Epidemic Preparedness Innovation, approximately 40% of the COVID-19 vaccine development effort was undertaken by North American organizations. In February 2020, the WHO stated that it did not anticipate that a COVID-19 vaccine would be available in less than 18 months [15].

A specific time period is frequently predicted using machine learning (ML) models, remote sensing techniques, and empirical models [16,17]. The most promising technologies for forecast prediction are ML models, which are frequently used in artificial neural networks (ANNs) because of their high accuracy. ARIMA is a well-known ML model that is particularly popular for time series data and has excellent accuracy for small datasets [17,18]. Because of its high accuracy in forecasting, researchers have used the ARIMA model to determine the pricing of natural gas, oil, and electricity [19].

COVID-19 data can be categorized as time series data. Time series analysis can be performed to identify data trends, clean data, and make future predictions [20,21]. Strategic decision making and future activity planning are dependent on seasonal time series forecasting [22]. In addition, time series analysis can be conducted to measure and analyse changes in datasets, including the past, present, and potential future changes.

A previous study [23] employed time series analysis to forecast COVID-19 validated instances. By using WHO time series data from January 22, 2020, to April 7, 2020, the study predicted the number of confirmed COVID-19 cases 3 months in advance by using the ARIMA model. One evaluation model, residual sum of squares (RSS) for the ARIMA model, which produced a result of 0.405828, served as the foundation for that study. However, the limitations of that were as follows: (i) the dataset was small, and the forecast was based on a pandemic with high levels of volatility; (ii) only RSS was to evaluate performance, and other metrics, including accuracy, precision, and recall, were not tested; and (iii) feature selection and outlier analysis were not performed for the dataset.

In this study, we built a model called improved ARIMA (termed as ARIMAI), which uses time series data on daily COVID-19 cases worldwide (gathered from https://www.who.int/) to estimate impending COVID-19 cumulative cases by employing the ARIMA model. The feature selection method employed was statistically significant, and k-means clustering was used to identify outliers. The accuracy of the ARIMAI model was 97.75%, which is greater than the 93% accuracy of the ARIMA model.

The remainder of the paper is organized as follows. Section 2 presents a literature review. Section 3 discusses system architecture, processes, and methodologies. Section 4 describes the preferred methodology. Section 5 discusses the findings. Section 6 draws the conclusions and provides suggestions for the future work.

2. Literature review

In this section, we review some studies using ML algorithms to forecast time series data.

2.1. COVID-19 detection using ML algorithms

By using a support vector machine (SVM) model, a study generated a real COVID-19 forecast for confirmed, recovered, and dead patients [3]. For 4 months, from January to April 2020, the authors collected global data on various factors, including confirmed location, deaths, retrieved COVID-19 data, longitude, and latitude. By using the SVM model, the authors identified factors affecting recognition, mortality, and recovery. They determined a strong correlation between the number of COVID-19 deaths and the number

of confirmed cases. Furthermore, they indicated that the food patterns and immunological state of a population are influential factors.

Another study used a hybrid ML method to forecast COVID-19 data by utilizing the Hungary dataset [23]. They used the fuzzy inference system and multi-layered perceptron-imperialist competitive algorithm. The study indicated that ML is a promising approach for pandemic modelling. However, more studies are warranted to confirm the results and improve prediction quality.

A study examined COVID-19 data to determine the age group that was the most severely affected by the virus [24]. Numerous forecasting models were developed using ML, and the outcomes were collected and examined. The XGBoost classifier, SVM, Gaussian naive Bayesian classifier, logistic regression, KNN + NCA, decision tree classifier, and multilinear regression demonstrated poorer performance than did the random forest classifier and random forest regressor. The findings revealed that COVID-19 affects people aged from 20 to 30, 30–40, and 40–50 years. The random forest classifier and regressor outperformed other models in terms of accuracy.

Another study determined how ML models helped in combating the COVID-19 pandemic threat [25]. This study identified possible strategies for ML-based defence against the COVID-19 virus and the accompanying outbreak. In addition, this study discussed ML applications and methodologies as well as key ML strategies to combat the COVID-19 outbreak. The study discovered that ML is a useful technique for locating current drugs that may be used in diverse settings and can be helpful for treating patients with COVID-19. Instead of a conventional overtly calculation-based approach, ML algorithms offer precise and beneficial properties. Furthermore, ML algorithms are useful for predicting health-care risk during the COVID-19 crisis. Risk factors for ML include region, climate, age, and social behaviour.

Another study proposed a potential model to forecast COVID-19 transmission [26]. The authors evaluated the epidemiological evidence for the illness and the incidence of COVID-19 cases in India by using COVID-19 Kaggle data. They employed vector autoregression, multilayer perceptron, and linear regression methods. Historical data on confirmed and recovered cases and fatalities can be utilized to forecast the near future in India. To make the model fully informative, additional attributes might be added, including information on different hospitals, patients' immune system, age, sex, and techniques employed to stop viral reproduction.

Multilayer perceptron, naive Bayes, and J48 were used in Ref. [27] to categorize responses to questionnaires distributed to Basra city residents. The effectiveness of COVID-19 preventive tactics was examined, and the most precise method was determined. A framework using supervised ML algorithms was developed. The survey consisted of 25 questions, including those on demographics, cognitive ability, health management, and prevention. A total of 1017 persons participated in the survey. Weka 3.8 was used to create the model. The study revealed that quarantine was essential in halting the spread of the disease. J48 exhibited the highest precision.

2.2. Time series analysis using the ARIMA model

Wavelet transform and ARIMA models have been used to develop a unique method for predicting electricity expenditure in the coming days. The historical power price series is divided into several better-behaved constituent series by using the wavelet transform. The values of these constituent series are then projected into the future by using well-fitted ARIMA models. The use of the wavelet transform as a data pre-processor enhances the forecasting performance of any strategy, including neural networks and ARIMA [28]. The previous study compared their approach to an ARIMA model that directly forecasted the original price series.

Another study suggested a more accurate method to predict store product sales by using the ARIMA model that considered structural reform [29]. Depending on whether the primary product is being pushed, they assessed the forecasting performance of the model. They determined that the recommended technique performed the best for the promoted forecast and outperformed the non-promoted forecast. Using information from a well-known American shop, they demonstrated that their method outperformed conventional forecasting techniques that do not account for the likelihood of such changes.

Researchers in Ref. [30] used time series SARIMA and MARIMA with interventions to assess 10 arrival series for Hong Kong to forecast the demand for tourism. The used the upgraded Dickey–Fuller evaluation to illustrate the seasonal nonstationary of all series. Significant test findings and anticipated signals enabled the empirical identification of significant actions, such as the loosening of requirements for issuing outbound visitor permits and the Asian financial crisis.

A study classified whether the COVID-19 pandemic is a significant barrier to sustainable development [31]. This study utilized fbprophet for forecasting. An additive model can be used to anticipate time series data by using a Python library package called Prophet. This study successfully used both seasonal time series and data from previous seasons. Non-linear patterns were fitted using seasonality and holiday effects in addition to yearly, monthly, and daily seasonality. The model assists in analyzing public sentiment toward the broadcast of pertinent health information as well as the evaluation of the political and economic effects of COVID-19 infection.

2.3. COVID-19 detection using time series data

To evaluate the disease's potential future spread in human civilization, researchers in Ref. [4] conducted a COVID-19 survey over a 5-month period by using data on the global incidence of confirmed incidents, fatalities, and recovery. The date of observation, the state, the country, and the most recent updates were all included in the data they acquired from the WHO. They compared various forecasting techniques, including the naïve methodology, single exponential smoothing, simple average, moving average, Holt Winter method, and ARIMA. They discovered that the naive approach is the most favorable.

The authors of [32] added to the existing time series prediction systems by using ML and methodologies that were inspired by nature. Nearly every nation has been forced to implement strict laws and regulations because of the current COVID-19 epidemic in order to stop the virus from spreading. The number of COVID-19 cases was estimated using the most recent forecasting methodology.

For better beetle antennae search, the proposed prediction model combines ML, adaptive neuro-fuzzy inference, and metaheuristics. Tuning the extended beetle antennae search and the adaptive neurofuzzy inference system improves the performance of the prediction model. When used to test against a larger collection of benchmark functions, the new technique outperformed the prior implementation significantly. The suggested hybrid method outperformed more intricate algorithms using the same datasets and can forecast time series.

A study evaluated the tendency of advanced deep learning architecture to predict time series data for COVID-19 by using short datasets. This study employed six data-driven models to reproduce the number of confirmed and recovered daily cases from seven of the most affected nations: Saudi Arabia, Mexico, Brazil, Russia, France, India, and the United States. The analysis considered the total number of COVID-19 occurrences reported between January 22 and September 6, 2020. Six efficacy metrics were employed to evaluate and compare the predictive power of the models: explained variance (EV), mean absolute error (MAE), mean absolute percentage error (MAPE), R2 (squared), mean squared logarithmic error (MLSE), and root mean square error (RMSE). They discovered that the long short-term memory (LSTM) convolutional neural network performed better, with an averaged MAE of 3.718% [33]. This is because LSTM may accumulate higher-level information that enhances the precision of prediction.

To create a prediction model for daily confirmed COVID-19 cases, a study used many covariates. The ideal prediction model was discovered using a portion of these variables [34]. The Malaysian Ministry of Health and John Hopkins University's websites were used to collect data for the study on daily confirmed COVID-19 cases. An ARIMA model was built using data from instances observed between January 22 and March 31, 2020, and the model was validated using cases observed between April 1 and April 17, 2020. The ARIMA model successfully forecasted daily COVID-19 cases between April 18 and May 1, 2020. They discovered that the ARIMA model, with a BIC and MAPE of 4.17 and 16.01, respectively, best fits the data. The incidence of COVID-19 decreased to the predicted level between May 1, 2020, and May 1, 2021. Projected cases were included in the fitted model's prediction intervals during the forecast period. This study demonstrated how carefully selected factors can be utilized to track and forecast COVID-19 case trends in Malaysia.

The Mann–Kendall test was used to determine the COVID-19 pandemic trend, and the recurrent forecasting singular spectrum analysis (RF-SSA) model was used to forecast future COVID-19 cases in Malaysia [35]. The RF-SSA model was created utilizing validated instances to analyse and estimate daily COVID-19 cases over the next 10 days. To extract noise in a time series trend, a forecasting technique based on SSA was suggested. On the basis of official COVID-19 data from the WHO, the RF-SSA model forecasted daily confirmed cases from April 29 to May 9, 2020. The findings indicated that the RF-SSA model significantly underestimated COVID-19 case counts by 0.36%, demonstrating the capacity of the RF-SSA model to forecast case numbers in the future. By the beginning of June 2020, the number of confirmed COVID-19 cases in Malaysia was in single digits according to their forecast. These results imply that by identifying trends, the RF-SSA model may reliably estimate COVID-19 occurrences. A more efficient RF-SSA strategy should be developed to determine any crucial data changes.

Our Contribution.

Several studies on time series data for COVID-19 prediction have been conducted. However, none of the studies used efficient feature selection techniques with statistical significance for evaluating time-series COVID-19 data. The proposed research uses statistical significance for feature selection and k-means clustering with the Mahala Nobis distance [36] for outlier detection and removal. The proposed research can improve the accuracy of the existing model.

3. System architecture

This section explains several methods and tools used to create the COVID-19 prediction model (ARIMAI) as well as metrics employed to investigate the effectiveness of the suggested system.

3.1. Dataset description

From the website (https://www.worldometers.info/coronavirus/#main table), we obtained daily information on confirmed cases

Table 1
Sample of the dataset before pre-processing.

Date	ID	Cases	Country
01/03/2020	1	31709	USA
01/04/2020	2	1685	India
01/05/2020	3	37690	Brazil
01/06/2020	4	148635	France
01/07/2020	5	51762	UK
01/08/2020	6	305592	Germany
01/09/2020	7	25387	Russia
01/10/2020	8	16894	Turkey
01/11/2020	9	8811	Italy
01/12/2020	10	0	Spain
1/13/2020	11	395589	South Korea
1/14/2020	12	1580	Argentina

globally. Instances of COVID-19 were discovered in 198 nations worldwide. Data updates were gathered between January 3, 2020, and July 18, 2020. The dataset has four attributes: Date, ID, new cases, and country. The dataset was divided into training and testing datasets by using the k-fold cross validation method. A sample of a raw dataset is presented in Table 1.

The COVID-19 dataset exhibits non-linear characteristics. If a dataset cannot be linearly separated, it is non-linear. According to the COVID-19 dataset, when the number of days or weeks increases, no linear separation of cases would be observed. The numbers of cases did not follow a linear trend. The proposed work is chosen due to its characteristics of handling a nonlinear dataset.

3.2. Data pre-processing

Data pre-processing, which involves the preparation and translation of data into an appropriate mining approach, is among the most crucial data mining techniques [37]. The objectives of this method include shrinking the quantity of data, creating data links, normalizing data, removing outliers, and gathering data features. Data cleansing, transformation, integration, and reduction are some of the strategies used [35,38].

Inadequate data efficiency and quality can produce substandard prediction results. Numerous pre-processing approaches were used with various tools available inside the Python programming language to maximize the usefulness and applicability of our initial dataset for COVID-19 forecasting. Table 2 displays a sample of the dataset after pre-processing.

4. Materials and methods

The flowchart and pseudocode for the system are described in this section. Fig. 1 displays the flowchart for the proposed ARIMAI model. The proposed solution is designed and put into practice by using two ML algorithms, namely k-means and ARIMA, as well as a feature selection method called statistical significance. Statistical significance removes any feature that is not correlated with the model on the basis of the correlation coefficient, thus improving performance while reducing training time and expense. This is an excellent feature selection strategy for time series data. The k-means clustering technique uses the statistical significance result to filter out any outlier. The proposed COVID-19 prediction model is then created using the k-means result and ARIMA approach.

4.1. Feature selection using statistical significance

The fundamental goal of feature selection is selecting a feature subset that enables the highest classification between groups. Use of all characteristics in a classifier does not yield the optimal result in many circumstances. In addition, feature selection aids individuals in gaining a better grasp of aspects that are crucial for diagnosing relevant data [39,40]. In this study, correlation analysis was performed to determine how strongly two variables are linked to one another. Using Pearson's correlation analysis, the link between feature and target variables was determined; this analysis examined only the strength of the correlation between the number of confirmed instances and dates. The Pearson correlation coefficient ranges from -1 to $+1$, with -1 signifying a weakly negative relationship, 0 signifying no association, and 1 signifying a strong positive correlation. Equation (1) theoretically expresses this metric. This phase can assist in identifying only the feature variable that significantly affects the incidence of COVID-19 positive cases.

$$p(x, y) = \sigma xy \,/\, \sigma x \sigma y \tag{1}$$

$$where : \sigma x = standard\ deviation\ of\ x$$

$$\sigma y = standard\ deviation\ of\ y$$

$$\sigma \mathrm{xy} = \text{population covariance}$$

4.1.1. Collinearity test

When predictors are related to one another linearly, collinearity occurs. To determine predictors with a high level of collinearity, the variance inflation factor (VIF) is used to assess the strength of the correlation between variables. After the inquiry, if the VIF is > 1,

Table 2
Sample of data set after pre-processing.

Date	ID	Cases in last 7 days
01/03/2020	1	3
01/04/2020	2	4
01/05/2020	3	4
01/06/2020	4	4
01/07/2020	5	4
01/08/2020	6	5
01/09/2020	7	5
01/10/2020	8	5
01/11/2020	9	6
01/12/2020	10	7

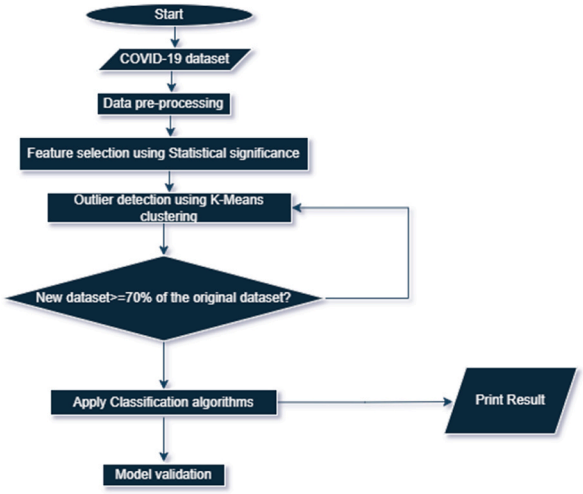

Fig. 1. Flowchart of the ARIMAI prediction model.

collinearity is present; otherwise, it indicates no collinearity [41]. Equations (2) and (3) are used to numerically express VIF as follows:

$$\mathrm{VIF} = \frac{1}{1 - R_i^2} \tag{2}$$

$$Where: i = The\ predictors\ \left(x_1, x_{2,\ldots\ldots\ldots x_n}\right)$$

and

$$R_{adj}^2 = \left[\frac{\left(1 - R^2\right)(n-1)}{n - 1 - K}\right] \tag{3}$$

R_{adj}^2 = adjusted R squared
n = total number of data samples
k = number of feature variables.

4.2. k-means clustering

A method of data partitioning called clustering divides a large dataset into smaller, easier-to-manage groups. In k-means clustering, each input in a dataset is assigned to one of the groups [42,43]. The goal is to create k clusters with observations that are remarkably similar within each cluster but significantly dissimilar between clusters [44]. Similarity was checked by using the Mahalanobis distance. The covariance between the variables in feature vectors that are being compared is the foundation for this distance. The benefits of using group averages and variances for each variable outweigh the drawbacks of scale and correlation that the Euclidean distance has. The data in this study were clustered using the aforementioned methods.

1. k = 6 as the initial value. By rounding each input data point to the nearest mean and comparing similarities by using the Mahalanobis distance, we divided data into six groups.
2. We calculated the mean value for input data for each cluster. We repeated steps 1 and 2 until the mean cluster value merged as necessary.
3. We removed outliers by deleting incorrectly classified data. This method can be used to generate a new dataset with the desired size. We moved to the next phase of classification if the size is greater than 70%; otherwise, the k-means clustering procedure was repeated until the data size was appropriate. Approximately 0.8% of outliers were found and deleted at the end of the clustering procedure.

4.3. ARIMA model

The most popular and effective statistical time series forecasting models are the ARIMA models, which were created by George Box and Gwilym Jenkins. Using various historical facts and random errors, the ARIMA model predicted the future value of an assumed variable [45,46]. In the ARIMA (p, q, d) model, p denotes the order of the autoregressive component, d denotes the extent of initial differentiation, and q denotes the order of the first moving component [21,47,48]. The mathematical model for ARIMA is as follows (see Equation (4)):

$$\mathscr{W}_t = \mu + \frac{\beta(\upsilon)}{\varepsilon(\upsilon)}\mathrm{a}_t \tag{4}$$

where:

t represents time, $\mathscr{W}_t$ signifies the response series itself or a variation in the response series, μ *signifies the mean term*, υ indicates the back shift operator, $\varepsilon(\upsilon)$signifies t he autoregressive operator, $\beta(\upsilon)$ signifies the moving-average operator, and a is the random error, which is also known as the independent disturbance.

Some components of the model building process are essential to ensure the accuracy of predictions, such as assumptions on error terms [28,49]. The steps include model identification, model parameter estimation, model hypothesis validation, and actual prediction.

4.4. Performance metrics

Several evaluation metrics were used to compare the ARIMAI model and other classification models. These metrics included accuracy, mean absolute percentage error (MAPE), mean error (ME), mean absolute error (MAE), R-Squared (R^2), root mean square error (RMSE), and residual sum of squares (RSS) [17,50–52]. These metrics were calculated as shown below. Equations (5)–(12) give the following mathematical formulation of the evaluation values:

$$Accuracy = \frac{No.\ of\ accurate\ predictions}{Total\ no.\ of\ predictions} \tag{5}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|A_i - F_i|}{A_i}\right) \times 100 \tag{6}$$

$$ME = sum\ of\ all\ errors/number\ in\ the\ set \tag{7}$$

$$MPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{a_t - f_t}{a_t} \tag{8}$$

$$MAE = \frac{1}{n}\left\{\sum_{i=1}^{n}|A_i - F_i|\right\} \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}}\left\{\sum_{i=1}^{n}(A_i - F_i)^2\right\} \tag{10}$$

$$R^2 = 1 - \left\{\frac{\sum_{i}^{n}(A_i - F_i)^2}{\sum_{i}^{n}(A_i - \overline{A}_i)^2}\right\} \tag{11}$$

$$RSS = \sum_{i=1}^{n}\left(y^i - F_i\right)^2 \tag{12}$$

Ai is the actual value, *Fi* is the forecast value, $\overline{A}_i$ is the mean actual value, p is the number of predictors, and n is the number of datasets.

Lower values for MAE, MAPE, and RMSE indicate a more accurate regression model. By contrast, the greater the RSS value is, the worse the model fits the data. Percent errors indicate how big experimental errors are when measuring a variable in an analysis. The lower the MPE is, the better is the model. The more closely the model matches the data, the lower the RSS is. In contrast to a number of 1, this shows that the model does not fit the data. A larger R-squared value is better. In the case of ME, a value of 0 implies no error [53,54].

Model evaluation is the process of using several evaluation metrics to assess the effectiveness, advantages, and disadvantages of a ML model. This method is critical to assess a model's efficacy in the early stages of a study. Monitoring a model is aided by model evaluation. We can use various evaluation criteria to determine whether our model(s) performs well with new data.

4.5. System implementation

The proposed model was implemented using Anaconda 3, which is compatible with Python 3.8 with an Intel Core i5-1135G7 processor, HP computer. It is an open-source software project with several tools to help with ML and data science application development. The COVID-19 prediction tool's design flow consists of the following steps:

Step 1. Data cleaning and preprocessing
Obtain the dataset and combine the datasets
For each of the dataset's records, if the record is not found, it is considered NULL
Record should be discarded

Step 2. Apply statistical significance (ref. Section 4.1) on the preprocessed dataset

Step3. Remove outliers (if any) from the dataset using K-means clustering.

Step 4. Compare the effectiveness of the XGboost algorithm, LSTM, naive Bayes, logistic regression, and ARIMAI.

5. Results and discussion

As listed in Table 3, the letters ARIMAI stand for SS/KMEANS/ARIMA, LSTMI for SS/KMEANS/LSTM, NAIVEBAYESI for SS/KMEANS/NAIVE-BAYES, LOGISTICREGRESSIONI for SS/KMEANS/LOGISTICREGRESSION, and EXGBOOSTI for SS/KMEANS/XGBOOST. The suggested method included three stages. In the first stage, the statistically significant and most crucial features are selected from processed and cleaned data. In the second stage, the output from SS is sent to the k-means clustering algorithm to find and eliminate outliers. In the last step, k-means clustering results are used with ML techniques (ARIMA, LSTM, NAVE BAYES, LOGISTIC REGRESSION, and XGBOOST) to predict confirmed cases.

We compared the performance of numerous models, namely ARIMAI, LSTMI, NAIVEBAYESI, LOGISTICREGRESSIONI, and EXGBOOSTI, by using various accuracy metrics, namely MAPE, ME, MPE, MAE, RMSE, R-Squared, and RSS.

To select the best daily data prediction model, we examined the least MAPE, ME, MPE, MAE, RMSE, and RSS values and the maximum R-squared and accuracy values. The maximum accuracy of the ARIMAI model was 97.75%. This had the smallest MAPE and RSS, measuring at 0.023 and 0.279 respectively. The findings indicated that the suggested ARIMAI model is the best for making predictions.

Compared with the other models, the ARIMAI model had a higher accuracy (Table 3), demonstrating its high performance. Compared with the ARIMAI, LSTMI, NAIVEBAYESI, LOGISTIC REGRESSIONI, and XG-BOOSTI models, the NAIVEBAYESI model had a lower ME value (the best). The proposed strategy outperformed the other approaches in terms of performance because the ARIMAI model had the lowest RMSE and MAE values among all the algorithms. The ARIMAI model had a higher MAE value than did the other algorithms. MPE data indicated that the ARIMAI performed better compared with other methods. The chosen model is inside the realm of the ideal model according to the MAPE value, which also revealed that the ARIMAI value is < 1 [56]. Apart from LSTMI, which has the same R-square value as does the ARIMAI model, no other approaches had a higher R-squared value than did the ARIMAI model. Thus, the ARIMAI model was superior to other approaches. We emphasized the RSS value of ARIMAI because this model had the lowest value, demonstrating how better the suggested model fits the dataset compared with other models. The ARIMAI model often outperformed the other approaches by a wide margin.

As listed in Table 4, the outcomes of the dataset utilized for the ARIMA (existing work) and the proposed work (ARIMAI) were compared. The existing work was implemented under the same condition as the proposed work to enable comparison. The proposed ARIMAI model had greater accuracy than did the current ARIMA model, demonstrating that the proposed ARIMAI model outperformed the ARIMA model. Moreover, the MAPE value of the ARIMAI model was smaller than that of the existing ARIMA model. The findings indicate that the ARIMAI model is better than the current approach.

The MAPE value for the ARIMAI model was <1, indicating that the chosen model was within the range of the effective model [57]. The values of ME, MAE, and RMSE for the ARIMAI model were lower than those of the existing ARIMA model, indicating that the proposed model outperformed the existing model. The RSS value of the ARIMAI model was lower than that of the ARIMA model, indicating that the proposed model fitted the dataset more than the other existing ARIMA.

Table 5 compares the results of previous studies on COVID-19 with those obtained using the dataset used for the ARIMAI (proposed work). To enable comparison, previous studies on COVID-19 that employed various prediction models had their results implemented

Table 3
Performance analysis with different classifiers.

	Accuracy (%)	ME	RMSE	MAE	MPE	MAPE	R-SQUARE	RSS
ARIMAI	**97.75**	**0.315**	**2.815**	**0.793**	**3.254**	**0.022**	**0.993**	**0.279**
LSTMI	91.71	0.553	5.138	0.808	1.002	0.083	0.993	0.593
NAÏVEBAYESI	88.83	0.016	5.138	1.011	1.002	0.112	0.796	1.190
LOGISTIC REGRESSIONI	93.91	0.052	5.138	1.825	1.002	0.060	0.796	0.551
XG-BOOSTI	96.75	0.019	5.137	0.972	1.002	0.032	0.796	0.401

Table 4
Comparison with the existing work.

S/N	Algorithm	Accuracy (%)	ME	RMSE	MAE	MPE	MAPE	RSS
1	ARIMA (Existing work)	93	0.489	18.669	1.065	0.043	0.065	0.659
2	ARIMAI (Proposed Work)	**97.75**	0.315	2.815	0.793	3.254	0.022	**0.279**

Table 5
Comparison with the existing work on COVID-19 prediction using different models.

S/N	Algorithm	Accuracy	ME	RMSE	MAE	MPE	MAPE
1	ARIMAI (Proposed Work)	97.75	0.315	2.815	0.793	3.254	0.022
2	CNN [55]	83.4	0.991	3.975	2.008	0.505	0.165
3	Linear regression [56]	82.7	0.997	6.925	1.499	0.993	0.172
4	Polynomial regression [56]	69.1	0.759	3.983	2.851	0.939	0.309

Fig. 2. ARIMAI evaluation model.

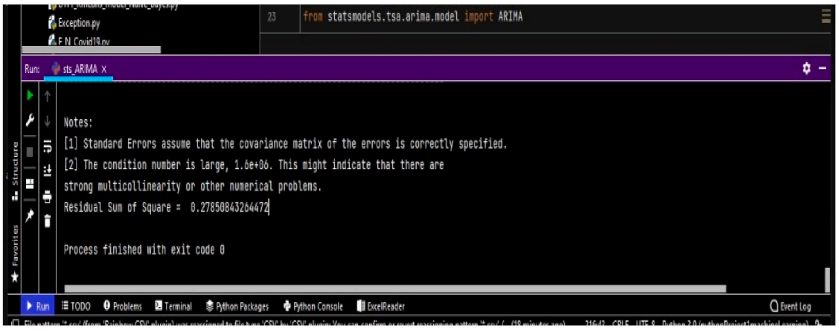

Fig. 3. Rss score for ARIMAI

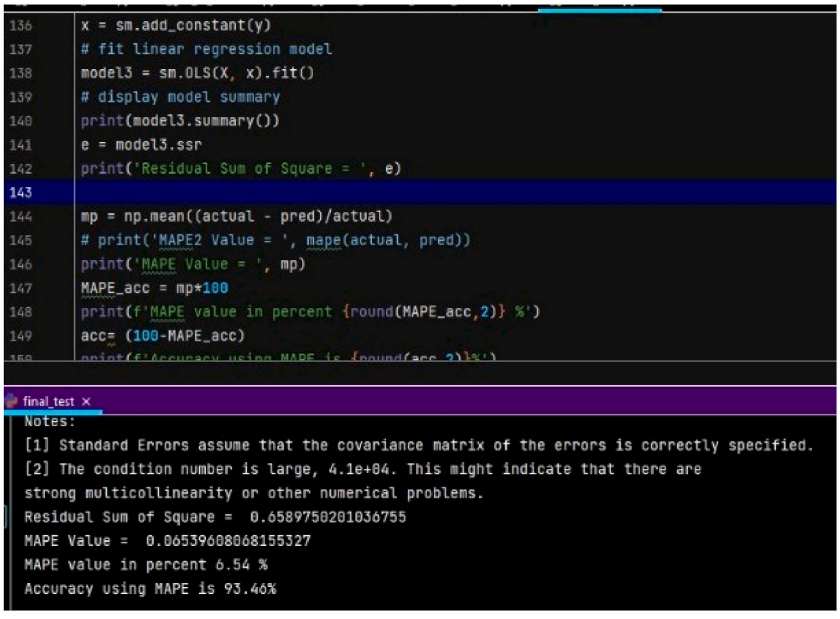

Fig. 4. RSS Score and evaluation model for ARIMA.

and assessed using the same evaluation models as did the proposed work. The suggested ARIMAI model surpassed all other models because its accuracy was higher than that of the current COVID-19 prediction model. In addition, the MAPE value of the ARIMAI model was lower than that of the current COVID-19 prediction models. This finding demonstrates that the ARIMAI model is superior to the present methods. The chosen model fits within the range of the effective model, as shown by the MAPE value for ARIMAI being <1. The suggested model outperformed the existing models, as evidenced by the fact that the ARIMAI's ME, MAE, and RMSE values are lower than those of the current COVID-19 prediction models.

Figs. 2, 3 and 4 presents the snapshots of the ARIMAI and ARIMA model when coded in Python. Fig. 2 presents all the evaluation models of ARIMAI, and Figs. 3 and 4 illustrate the RSS results for ARIMAI and ARIMA, respectively.

6. Conclusion and future work

The effectiveness of the feature selection method and the accuracy of COVID-19 predictions obtained utilizing a given dataset were determined in this study employing statistical significance and outlier detection using k-means clustering with Mahalanobis distance. The improved version is known as the ARIMAI model. With an accuracy of 97.75% and an RSS of 0.279, the suggested technique outperformed all other ML algorithms tested using the dataset for COVID-19 time series prediction. Using open-access data, the ARIMAI model was not only efficient but also a rapid and simple technique to forecast COVID-19 trends.

This study demonstrated that the ARIMA model for COVID-19 prediction delivered the most precise forecast for the short term due to the ability of the ARIMA model to produce a good forecast for information that pertains to a short timeframe [58]. Therefore, this method could be used by policymakers to advance updated recommendations for the short-term strategy.

7. Limitations of the study

The small dataset and the fact that the forecast was based on a pandemic with high volatility in the dataset are the two main limitations of this study. If the dataset was larger and less volatile, the results would be more precise. The usage of this method enables the prediction of any disease that will affect patients in the future pandemics.

Author contribution statement

Saratu Yusuf Ilu, Masters: Conceived and designed the experiments; Performed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Rajesh Prasad, PhD: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by Canada's International Development Research Centre (IDRC) and Swedish International Development Cooperation Agency (SIDA) [109559-001].

Data availability statement

Data associated with this study has been deposited at https://www.who.int/

Declaration of interest's statement

The authors declare no competing interests.

Additional information

No additional information is available for this paper.

References

[1] M. Mahdikhani, International Journal of Information Management Data Insights Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic, Int. J. Inf. Manag. Data Insights 2 (1) (2022), 100053, https://doi.org/10.1016/j.jjimei.2021.100053.

[2] L. Song, Y. Zhou, The COVID-19 pandemic and its impact on the global economy, What Does It Take to Turn Crisis into Opportunity 28 (4) (2020) 1–25, https://doi.org/10.1111/cwe.12349.

[3] V. Singh, et al., Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine, J. Discrete Math. Sci. Cryptogr. 23 (8) (2020) 1583–1597, https://doi.org/10.1080/09720529.2020.1784535.

[4] V. Chaurasia, S. Pal, Application of machine learning time series analysis for prediction COVID-19 pandemic, Res. Biomed. Eng. (2020) 1, https://doi.org/10.1007/s42600-020-00105-4. –16.

[5] K.M. Ridhwan, C.A. Hargreaves, International journal of information management data insights leveraging twitter data to understand public sentiment for the COVID-19 outbreak in Singapore, Int. J. Inf. Manag. Data Insights 1 (2) (2021), 100021, https://doi.org/10.1016/j.jjimei.2021.100021.

[6] M. Alazab, A. Awajan, A. Mesleh, A. Abraham, V. Jatana, S. Alhyari, COVID-19 Prediction and Detection Using Deep Learning 12 (April) (2020) 168–181.

[7] P. Srilatha Reddy, COVID-19 prevention and management : overview, Int. J. Sci. Res. Sci. Technol. 7 (6) (2020) 23–32.

[8] R. Dard, N. Janel, F. Vialard, COVID-19 and Down ' s syndrome : are we heading for a disaster, Eur. J. Hum. Genet. 28 (2020) 1477–1478, https://doi.org/10.1038/s41431-020-0696-7.

[9] D. Petrakis, D. Margină, K. Tsarouhas, D. Kouretas, D.A. Spandidos, A. Tsatsakis, Obesity-a risk factor for increased COVID-19 prevalence , severity and lethality (Review), Mol. Med. Rep. 22 (2020) 9–19, https://doi.org/10.3892/mmr.2020.11127.

[10] J. Koch, R. Plattfaut, I. Kregel, International journal of information management data insights looking for talent in times of crisis – the impact of the covid-19 pandemic on public sector job openings, Int. J. Inf. Manag. Data Insights 1 (2) (2021), 100014, https://doi.org/10.1016/j.jjimei.2021.100014.

[11] E. Mbunge, Diabetes & metabolic syndrome : clinical research & reviews integrating emerging technologies into COVID-19 contact tracing : opportunities , challenges and pitfalls, Diabetes Metab. Syndr. Clin. Res. Rev. 14 (6) (2020) 1631–1636, https://doi.org/10.1016/j.dsx.2020.08.029.

[12] H. Lee, G. Jang, G. Cho, Forecasting COVID-19 cases by assessing control- intervention effects in Republic of Korea : a statistical modeling approach, Alex. Eng. J. 61 (11) (2022) 9203–9217, https://doi.org/10.1016/j.aej.2022.02.037.

[13] G. Pinter, I. Felde, A. Mosavi, P. Ghamisi, R. Gloaguen, COVID-19 pandemic prediction for Hungary; A hybrid machine learning approach, SSRN Electron. J. (2020), https://doi.org/10.2139/ssrn.3590821.

[14] F. İ. R. A. Ş, COVID-19 : prevention and control measures in community, Turk. J. Med. Sci. 50 (9) (2020) 571–577, https://doi.org/10.3906/sag-2004-146.

[15] T.T. Le, et al., The COVID-19 vaccine development landscape, Nat. Rev. Drug Discov. (2020), https://doi.org/10.1038/d41573-020-00073-5. May.

[16] A. Malki, E. Atlam, I. Gad, Machine learning approach of detecting anomalies and forecasting time-series of IoT devices, Alex. Eng. J. 61 (11) (2022) 8973–8986, https://doi.org/10.1016/j.aej.2022.02.038.

[17] K.E. Arunkumar, D. V Kalaga, C. Mohan, S. Kumar, T.M. Brenza, Comparative analysis of Gated Recurrent Units (GRU), long Short-Term memory (LSTM) cells , autoregressive Integrated moving average (ARIMA), seasonal autoregressive Integrated moving average (SARIMA) for forecasting COVID-19 trends, Alex. Eng. J. 61 (10) (2022) 7585–7603, https://doi.org/10.1016/j.aej.2022.01.011.

[18] R.S. Society, Review author (s): M . G . Kendall review by : M . G . Kendall source : journal of the royal statistical society . Series A (general), J. Roy. Stat. Soc. 134 (3) (2016) 450–453, 134 , No . 3 (1971), Published by : Wiley for the Royal Statistical Society Stable URL : http://www.jstor.or.

[19] Y. Feng, W. Hao, H. Li, N. Cui, D. Gong, L. Gao, Machine learning models to quantify and map daily global solar radiation and photovoltaic power, Renew. Sustain. Energy Rev. 118 (August 2019) (2020), 109393, https://doi.org/10.1016/j.rser.2019.109393.

[20] S. Athiyarath, M. Paul, S. Krishnaswamy, A comparative study and analysis of time series forecasting techniques, SN Comput. Sci. (2020) 1–7, https://doi.org/10.1007/s42979-020-00180-5.

[21] I. Khandelwal, R. Adhikari, G. Verma, Time series forecasting using hybrid arima and ann models based on DWT Decomposition, Procedia Comput. Sci. 48 (C) (2015) 173–179, https://doi.org/10.1016/j.procs.2015.04.167.

[22] Y. Ensafi, S. Hassanzadeh, G. Zhang, B. Shah, International Journal of Information Management Data Insights Time-series forecasting of seasonal items sales using machine learning – a comparative analysis, Int. J. Inf. Manag. Data Insights 2 (1) (2022), 100058, https://doi.org/10.1016/j.jjimei.2022.100058.

[23] F.A. Chyon, M.N.H. Suman, M.R.I. Fahim, M.S. Ahmmed, Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning, J. Virol. Methods 301 (December 2021) (2022), 114433, https://doi.org/10.1016/j.jviromet.2021.114433.

[24] K.B. Prakash, Analysis, prediction and evaluation of COVID-19 datasets using machine learning algorithms, Int. J. Emerg. Trends Eng. Res. 8 (5) (2020) 2199–2204, https://doi.org/10.30534/ijeter/2020/117852020.

[25] S. Kushwaha, et al., Significant applications of machine learning for covid-19 pandemic, J. Ind. Integr. Manag. 5 (4) (2020) 453–479, https://doi.org/10.1142/S2424862220500268.

[26] R. Sujath, J.M. Chatterjee, A.E. Hassanien, A machine learning forecasting model for COVID-19 pandemic in India, Stoch. Environ. Res. Risk Assess. 34 (7) (2020) 959–972, https://doi.org/10.1007/s00477-020-01827-8.

[27] W.A. Awadh, A.S. Alasady, H.I. Mustafa, Predictions of COVID-19 spread by using supervised data mining techniques, J. Phys. Conf. Ser. 1879 (2) (2021), https://doi.org/10.1088/1742-6596/1879/2/022081.

[28] A.J. Conejo, M.A. Plazas, R. Espínola, A.B. Molina, Day-ahead electricity price forecasting using the wavelet transform and ARIMA models, IEEE Trans. Power Syst. 20 (2) (2005) 1035–1042, https://doi.org/10.1109/TPWRS.2005.846054.

[29] T. Huang, R. Fildes, D. Soopramanien, C.R. Pt Us, Eur. J. Oper. Res. 279 (6) (2019) 459–470, https://doi.org/10.1016/j.ejor.2019.06.011.

[30] C. Goh, R. Law, Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention, J. Tour. Manag. 23 (2002) 499–510.

[31] T.T. Mengistie, COVID-19 outbreak data analysis and prediction modeling using data mining technique, Int. J. Comput. 38 (1) (2020) 37–60.

[32] M. Zivkovic, et al., COVID-19 cases prediction by using hybrid machine learning and beetle antennae search approach, Sustain. Cities Soc. 66 (December 2020) (2021), 102669, https://doi.org/10.1016/j.scs.2020.102669.

[33] A. Dairi, F. Harrou, A. Zeroual, M.M. Hittawe, Y. Sun, Comparative study of machine learning methods for COVID-19 transmission forecasting, J. Biomed. Inf. 118 (April) (2021), 103791, https://doi.org/10.1016/j.jbi.2021.103791.

[34] S. Singh, et al., Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models, J. Infect. Dev. Ctries. 14 (9) (2020) 971–976, https://doi.org/10.3855/JIDC.13116.

[35] S.A. Alasadi, W.S. Bhaya, Review of data preprocessing techniques in data mining, J. Eng. Appl. Sci. 12 (16) (2017) 4102–4107, https://doi.org/10.3923/jeasci.2017.4102.4107.

[36] G.J. Mclachlan, Mahalanobis Distance," No, June, 1999.

[37] S.S. Baskar, L. Arockiam, S. Charles, Related papers A Systematic approach on data pre-processing in data mining, Intern. ional J. Sci. ific Res. Scence Technol. IJSRST 2 (11) (2013) 335–339.

[38] S.B. Kotsiantis, D. Kanellopoulos, Data preprocessing for supervised leaning, Int. J. Comput. Sci. 1 (2) (2006) 1–7, https://doi.org/10.1080/02331931003692557.

[39] C.A. Bejan, F. Xia, L. Vanderwende, M.M. Wurfel, M. Yetisgen-yildiz, Pneumonia identification using statistical feature selection, J. Am. Med. Inf. Assoc. 19 (2011) 817–823, https://doi.org/10.1136/amiajnl-2011-000752.

[40] A. Narin, Y. Isler, M. Ozer, Investigating the performance improvement of HRV Indices in CHF using feature selection methods based on backward elimination and statistical signi fi cance, Comput. Biol. Med. 45 (2014) 72–79, https://doi.org/10.1016/j.compbiomed.2013.11.016.

[41] O. Nkiruka, R. Prasad, O. Clement, Prediction of malaria incidence using climate variability and machine learning, Inform. Med. Unlocked 22 (2021), 100508, https://doi.org/10.1016/j.imu.2020.100508.

[42] L. Morissette, S. Chartier, The k -means clustering technique : general considerations and implementation in Mathematica, Tutor. Quant. Methods Psychol. 9 (1) (2013) 15–24.

[43] V. Faber, in: Clustering and the Continuous k -Means Algorithm, 22, 1994, pp. 138–144.

[44] D. Lei, Q. Zhu, J. Chen, H. Lin, P. Yang, Automatic K-means clustering algorithm for outlier detection, J. Inf. Eng. Appl. 154 (2012) 363–364, https://doi.org/10.1007/978-1-4471-2386-6.

[45] Â. Faruk, O. Durdu, Engineering Applications of Artificial Intelligence A hybrid neural network and ARIMA model for water quality time series prediction, Int. J. Intell. Real-Time Autom. 23 (2010) 586–594, https://doi.org/10.1016/j.engappai.2009.09.015.

[46] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, M. Ciccozzi, Data in brief Application of the ARIMA model on the COVID- 2019 epidemic dataset, Data Brief 29 (2020), 105340, https://doi.org/10.1016/j.dib.2020.105340.

[47] S. Al Wadi, M.T. Ismail, M.H. Alkhahazaleh, S.A.A. Addul Karim, Selecting wavelet transforms model in forecasting financial time series data based on ARIMA model, Appl. Math. Sci. 5 (5–8) (2011) 315–326.

[48] H. Li, et al., Prediction of gold price with ARIMA and SVM, J. Phys. Conf. Ser. 1767 (2021), 012022, https://doi.org/10.1088/1742-6596/1767/1/012022.

[49] F. Abdulla, Z. Hossain, S. Quality, Comparison of ARIMA and neural networks to forecast the jute production in comparison of ARIMA and neural network model to forecast the jute production in Bangladesh, Jahangirnagar Univ. J. Sci. 40 (July) (2017) 11–18.

[50] P. Xu, M. Aamir, A. Shabri, M. Ishaq, A. Aslam, L. Li, in: A New Approach for Reconstruction of IMFs of Decomposition and Ensemble Model for Forecasting Crude Oil Prices, 2020, 2020.
[51] K.E. Arunkumar, D. V Kalaga, C. Mohan, S. Kumar, G. Chilkoor, Forecasting the dynamics of cumulative COVID-19 cases (confirmed , recovered and deaths) for top-16 countries using statistical machine learning models : auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving, Appl. Soft Comput. J. 103 (December 2019) (2021), 107161, https://doi.org/10.1016/j.asoc.2021.107161.
[52] Q. Mao, K. Zhang, W. Yan, C. Cheng, Journal of Infection and Public Health Forecasting the incidence of tuberculosis in China using the seasonal auto-regressive integrated moving average (SARIMA) model, J. Infect. Public Health 11 (5) (2018) 707–712, https://doi.org/10.1016/j.jiph.2018.04.009.
[53] H. Liu, et al., Journal of infection and public health forecast of the trend in incidence of acute hemorrhagic conjunctivitis in China from 2011 – 2019 using the seasonal autoregressive integrated moving average (SARIMA) and exponential smoothing (ETS) models, J. Infect. Public Health 13 (2) (2020) 287–294, https://doi.org/10.1016/j.jiph.2019.12.008.
[54] L. Mamudu, A. Yahaya, S. Dan, Application of seasonal autoregressive integrated moving average (SARIMA) for flows of river kaduna, Niger. J. Eng. 28 (2) (2021).
[55] A. Srinivasulu, T. Barua, S. Nowduri, M. Subramanyam, COVID-19 Virus Prediction Using CNN and Logistic Regression Classification Strategies, 2022, pp. 78–89, https://doi.org/10.4236/jdaip.2022.101005.
[56] R.M. Shaikh Saud, Gala Jaini, Aishita Jain, Advani sunny, jaidhari sagar, "analysis and prediction of covid-19 using regression models and time series forecasting, in: International Conference on Cloud Computing, Data Science and Engineering, 2021, pp. 989–995.
[57] M. Naeem, J. Yu, M. Aamir, S. Ahmad, Comparative analysis of machine learning approaches to analyze and predict the COVID-19 outbreak, PeerJ. Comput. Sci. (2019) 2021, https://doi.org/10.7717/peerj-cs.746.
[58] X. Qiang, M. Aamir, M. Naeem, S. Ali, A. Aslam, Z. Shao, Analysis and Forecasting COVID-19 Outbreak in Pakistan Using Decomposition and Ensemble Model, July 2020, p. 2021, https://doi.org/10.32604/cmc.2021.012540.