DOI: 10.24996/ijs.2024.65.3.36

ISSN: 0067-2904

An Analytic Model for COVID-19 Cases in Iran and Its Neighbors Using Deep Learning and Time Series Methods

Razieh Abedi, Kheirolah Rahsepar Fard*

Department of Computer Engineering and Information Technology, University of Qom, Qom, Iran

Received: 29/4/2022 Accepted: 20/3/2023 Published: 30/3/2024

Abstract

Since the pandemic of the coronavirus (COVID-19) in 2019, it has rapidly become a major global health concern. Various mutations and rapid spread of the virus, a lack of specific treatment, and limited hospital facilities highlight the importance of anticipation, risk analysis, and timely treatment. The use of mathematical models, artificial intelligence, and simulation methods are effective tools in predicting the spread and providing effective solutions to prevent virus transmission. Analysis and forecasting require an integrated model to cover different aspects of the problem and use different methods to obtain appropriate results.

In this research, a proposed model for analysis and prediction of COVID-19 cases in Iran and neighboring countries is presented. The performance of mathematical and deep learning models in the proposed model has been evaluated using data from Johns Hopkins University from January 29, 2020, to April 30, 2021. Evaluation of the predictive outcomes of daily cases was performed using RMSE criteria. Then, the effect of the trend of cases in neighboring countries of Iran on the rate of new cases in this country has been studied. These models can help governments predict the number of infections to provide the necessary solutions and prevent a new wave of the virus.

Keywords: COVID-19, Risk Analysis, Deep Learning, Time Series

1. Introduction

On March 11, 2020, the World Health Organization announced the outbreak of the global coronavirus. The coronavirus was first identified in December 2019 in Wuhan, China, as COVID-19 and has spread around the world in a matter of weeks. More than 340 million definitive cases and more than 5.5 million deaths have been reported worldwide. However, researchers have not found a specific treatment for it [1].

Despite the development of various vaccines by some countries, including Iran, to prevent COVID-19, factors such as new mutations in the virus, access to the vaccine in only a few countries in the world, and the lack of complete immunity of vaccinated people against these mutations have raised many concerns about controlling the epidemic. Therefore, predicting new cases and identifying critical cases using clinical data has become an important and challenging mission. Such advanced models can provide immediate preventive measures and help physicians identify and intervene early to reduce pandemic mortality.

There are several approaches to analyzing the prevalence of the coronavirus, including creating predictive models and diagnosing the condition of patients. For this purpose, statistical

*Email: rahsepar@qom.ac.ir

models [2, 3] and methods based on machine learning [4] and deep learning [5, 6] can increase accuracy and speed in estimating the number of confirmed, deceased, and recovered cases. They can also be used to prioritize acute coronavirus patients, improve pandemic management and control, and reduce mortality.

The following are some of the research and achievements that have been made in this field to prevent and control the prevalence of COVID-19.

1.1 Statistical models

Using time series data from five countries: Canada, France, India, South Korea, and the United Kingdom from late January to early April 2020 and the Autoregressive Integrated Moving Average-Wavelet Based Forecasting (ARIMA-WBF) hybrid model, Chakraborty [7] made a short-term prediction of COVID-19 cases. The combination of ARIMA and WBF models can model complex autocorrelation structures in the COVID-19 time series dataset and reduce model prediction error.

Singh et al. [8] propose a hybrid method involving the use of wavelet analysis with an ARIMA model developed to predict the death of COVID-19 cases in the coming month. The computational results show that the hybrid model minimizes the prediction errors compared to the ARIMA model.

Toga et al. [9] used ARIMA models and an artificial neural network to predict the daily infection, death, and recovered cases of COVID-19 patients in Turkey. Both models had almost the same results, and the ARIMA model was able to predict the prevalence of the coronavirus cases with high accuracy.

In 16 countries with high incidence, Arun Kumar et al. [10] have implemented the ARIMA and Seasonal Autoregressive Integrated Moving Average (SARIMA) to forecast confirmed and recovered cases in the next 60 days. The COVID-19 trend in these countries is divided into three classes: exponential, linearly sloping, and gradually increasing linearly. The results show that in most countries, including the United States and India, there is a 7-day seasonal pattern. The predictions of the SARIMA model are more realistic than the predictions of the ARIMA model, which confirms the existence of seasonality in COVID-19 data.

Knopov et al. [11] made a switching regression model with unknown switching points to investigate the infection dynamics of COVID-19 cases in Ukraine. In a step-by-step solution, specific switching points and, for each input, being in a regime or creating a switching point and a new regime are specified.

1.2 Deep and machine learning models

In a study of 375 patients in Wuhan, China, Yan et al. [12] implemented a predictive model for identifying critical cases of COVID-19 patients using clinical data-based machine learning (Extreme Gradient Boost (XGBoost)). After extracting clinically important features, the model was tested on 29 other discharged patients. The proposed model extracts three important clinical features from more than 300 features and, based on them, is able to predict the risk of mortality and provide a clinical route to distinguish critical cases from severe cases.

Luo et al. [4] used Long Short-Term Memory (LSTM) and XGBoost models to study the trend of COVID-19 cases in the United States and identify the effective indicators. The dataset includes case information from April 1, 2020, to September 30, 2020, and forecast results for the next 30 days show that the LSTM model is able to better estimate cases with less error than XGBoost. According to the results obtained by the XGBoost model, the most important features are the average number of cases in the past seven days and the number of cases in the past seven days.

Suzuki et al. [13] implemented a hybrid model using XGBoost and multivariate regression to predict COVID-19 cases for the next 24 days in 17 South Korean provinces. The machine learning model, as a binary classifier, predicts the number of cases above 100 in the next 24 days in each province.

Using the Johns Hopkins University Time Series Database, Chimmula et al. [5] implemented a deep learning model (LSTM) in Canada to forecast COVID-19 occurrence and assess key features in predicting the trends and possible stopping time of the current COVID-19 outbreak.

Wang et al. [14] examined the COVID-19 epidemic trend in Russia, Peru, and Iran using the development of LSTM networks. The results show that the implemented LSTM network is able to model the infection process well in the next 150 days in these three countries.

Zeroual et al. [15] developed five deep learning algorithms (recurrent neural networks (RNN), LSTM, bidirectional LSTM (Bi-LSTM), gated recurrent units (GRU), and variational auto encoder (VAE)) to estimate the number of confirmed and recovered cases of COVID-19 over the next 17 days in six countries: Italy, Spain, France, China, the United States, and Australia. The results show that the VAE model can obtain almost all the diversity in the data and provide a more accurate forecast than other models.

Yudistira et al. [16] predicted new outbreaks in Italy, Sweden, Indonesia, and Norway and conducted a multivariate analysis consisting of environmental, mobility, demographic, COVID-19, health, health facilities, government, economic, and education categories.

In the study of Ayoobi et al. [6], prediction of new cases and mortality rates of COVID-19 patients in the next 100 days using six deep neural network prediction models (LSTM, GRU, Convolutional–LSTM (Conv_LSTM), Bi_LSTM, Bi_GRU, and Bi_Conv_LSTM) was done. The results show that bidirectional models have less error than other models.

Shastri et al. [17] evaluated three models, stacked-LSTM, Bi-LSTM and Conv-LSTM, to demonstrate forecasting of COVID-19 for India and the USA for the next one month. COVID-19 is confirmed, and death cases in both countries are taken into consideration. The results showed that Conv-LSTM performed best.

Arora et al. [18] implemented the stacked-LSTM, Conv-LSTM and Bi-LSTM models to predict COVID-19 cases in 32 Indian shires. Among the 3 implemented models, the Bi-LSTM model had the lowest error and highest accuracy, and the Conv-LSTM model had the poorest performance.

Therefore, predicting and estimating the number of cases or deaths of patients or the peak of the pandemic using random models and deep learning can be used in the field of management and prevention. In this research, considering the importance of forecasting and taking the necessary measures, a proposed model for using different methods and determining the project framework is presented. Deep learning and machine learning models, such as XGBoost, LSTM, BiLSTM, Conv_LSTM, ARIMA, and Switching Regression as statistical models, are implemented and analyzed to predict the incidence of COVID-19 in Iran and neighboring countries. This study, based on the proposed model, has not been done in this area so far.

2. Proposal model

The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a six-stage process model to define the data science life cycle. CRISP-DM defines the steps of a project, tasks, and relations between tasks. The main steps in CRISP-DM are as follows: [19]

- Business understanding: understanding project goals and requirements and turning this knowledge into a data mining problem.
- Understanding data: collecting and checking data quality; identifying hypotheses regarding hidden information in data.
- Data preprocessing: including all the activities needed to create the final data set, such as cleaning and formatting the data.
- Modeling: including choosing the modeling method, designing the test, building, and evaluating the model.
- Evaluation: checking the results of the model to achieve business goals before the final deployment of the model.
- Development: project development and monitoring, and documentation of results.

One of the most important features of the CRISP-DM model is the possibility of reverse transfer between stages. When working with real data, any mistakes can be fixed without having to complete the entire cycle. In this research as well, inspired by the CRISP-DM model, work processes have been carried out (Figure 1).

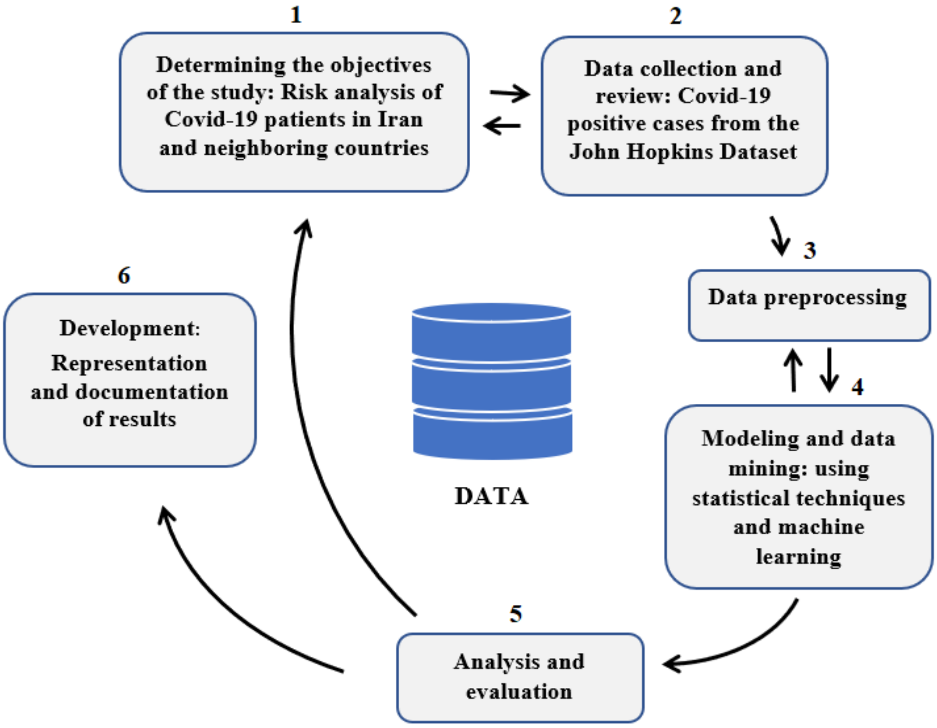
1
Determining the objectives of the study: Risk analysis of Covid-19 patients in Iran and neighboring countries
2
Data collection and review: Covid-19 positive cases from the John Hopkins Dataset
3
Data preprocessing
4
Modeling and data mining: using statistical techniques and machine learning
5
Analysis and evaluation
6
Development: Representation and documentation of results
DATA

Figure 1: Steps of proposal model

2.1 Types of models for prediction of COVID-19 cases

2.1.1 LSTM Model

The LSTM network has been developed to solve the problem of vanishing gradient in RNN [20]. the major change is the replacement of hidden layers of RNN with memory cells to remember long-term dependencies. In this network, there are three gates: input, forget, and output, which control the flow of information [18].

2.1.2 Bi_LSTM Model

Bidirectional LSTM is an extension of the LSTM network. In these networks, data is extracted at time t by considering both past and future information [18]. Forward and backward hidden neurons are not connected [18]. Bi-LSTM networks can be trained faster than LSTM [6].

2.1.3 CONV_LSTM Model

The convolutional LSTM network, unlike the LSTM networks used for one-dimensional data, can manage multidimensional data. Spatio-temporal data are encoded in the Conv_LSTM memory cell. In Conv-LSTM, there are convolutional layers instead of LSTM layers. Using convolution operators, it decides whether information should be remembered or forgotten [6].

2.1.4 ARIMA Model

The autoregressive integrated moving average (ARIMA(p, d, q)) is a statistical model to forecast non-stationary time series data. Three terms of an ARIMA model are auto-regression (AR), differentiation (D), and moving average (MA). p is the order of the AR term, q is the order of the MA term, and d is the order of differencing to make a series stationary. AR models consider the correlation between time sequences of variables. In other words, the value of a variable at a given time depends on the value of that variable at previous time steps. The MA model is a type of linear regression that uses prediction-related errors in the previous time step to predict a variable in the next time step [10].

The model of AR for variable Z_t is as follows:

$$Z_t = \sum_{i=1}^{p} (\Phi_i Z_{t-i} + \varepsilon_t) , \tag{1}$$

Where Φ_i is the j-th parameter of the model and ε_t is random variable.

The model of MA for variable Z_t is as follows:

$$Z_t = \varepsilon_t + \sum_{j=1}^{q} (\Theta_j \, \varepsilon_{t-j}), \tag{2}$$

Where Θ_j is the j-th parameter of the model and ε_t is random variable.

The autocorrelation function (ACF) plot is used to determine the stationarity of time series, and non-stationary series are converted to stationary using the differencing transformation. Also, ACF is used to specify the lag structure in the AR model. A partial autocorrelation function (PACF) plot is used to specify the MA lag structure [2].

After making timeseries stationary, all models are compared based on the Akaike information criterion (AIC) or Bayesian information criterion (BIC). The model with the lowest AIC or BIC is selected [21].

The approaches used in this article to analyze the risk of COVID-19 patients are shown in Figure 2.

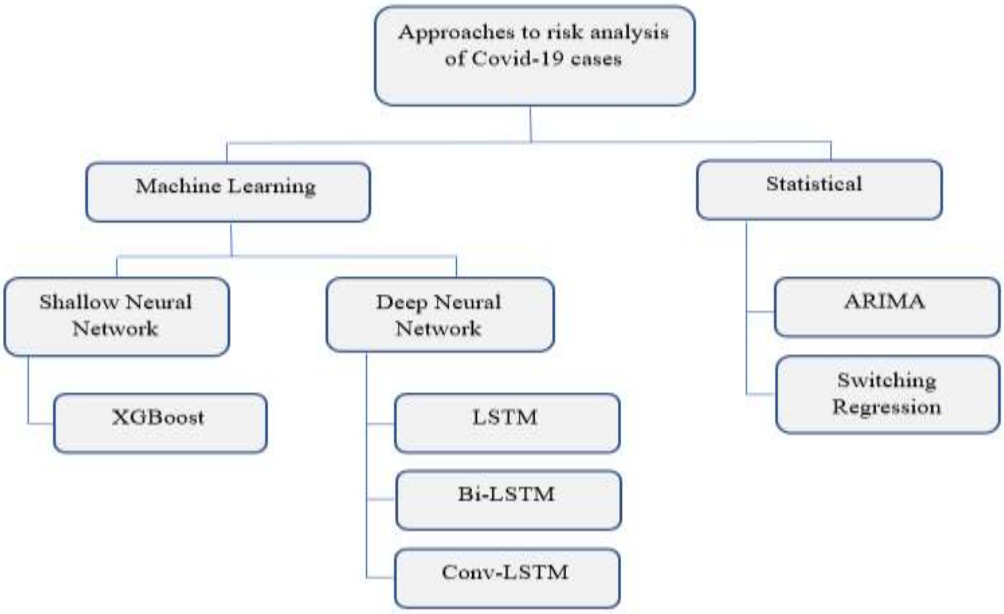

Figure 2: Approaches to risk analysis of Covid-19 cases

2.1.5 Modelling

The statistical approach (ARIMA) and three models of deep neural networks (LSTM, Bi_LSTM, and Conv_LSTM) have been used to predict the cases of COVID-19. Prediction is made in two steps: training and evaluation.

In the statistical approach, the stationary set of data is first examined, and after differential conversion (if required), the best parameters are selected for the model using an appropriate evaluation criterion (AIC). The ARIMA model framework for prediction is shown in Figure 3.

In the deep neural network approach, the data is first pre-processed and standardized before being used to learn the model. Then the best hyperparameters are selected to optimize the model and reduce errors. Finally, the models created for evaluation are tested on the test data. The steps for implementing deep neural network models are shown in Figure 4. Finally, the performance of the models is evaluated using the root-mean-square error (RMSE).

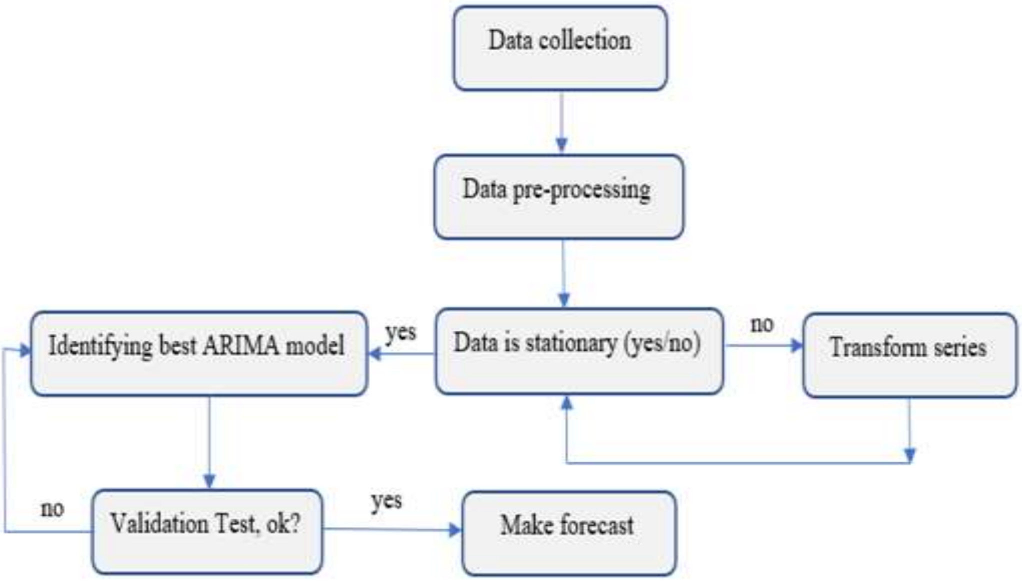

Figure 3: Steps of implementation for ARIMA model [2]

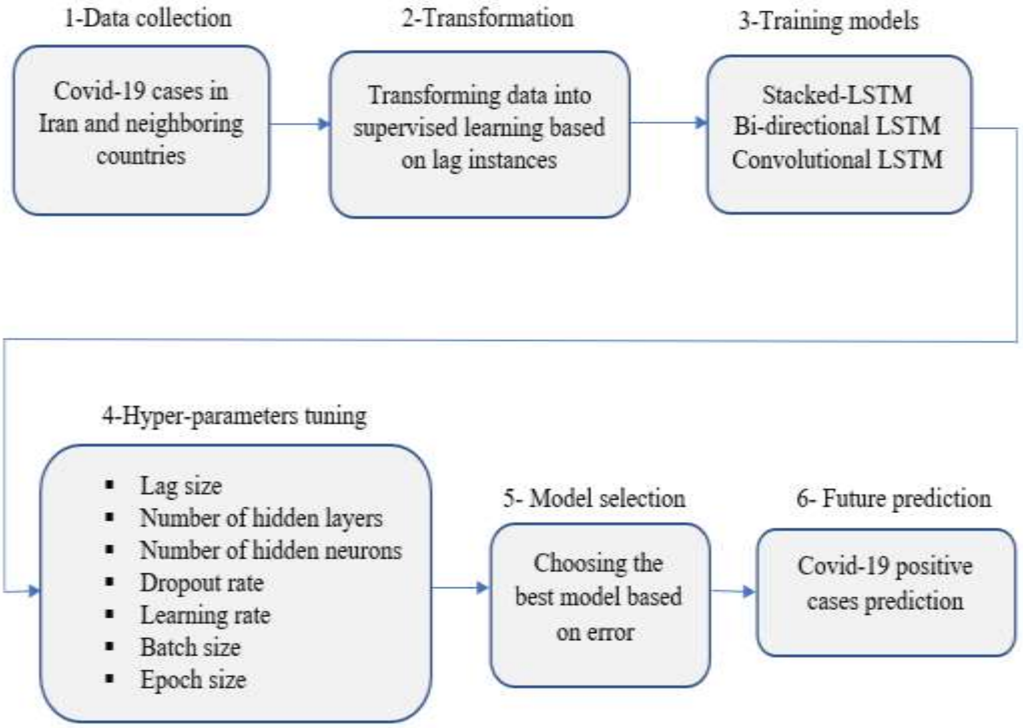

Figure 4: Steps of implementation for deep learning model [18]

2.2 The evaluation criterion

The RMSE criterion has been used to evaluate and compare the performance of models. The mean square error is a method for estimating the amount of error that calculates the difference between the predicted values and the actual values. To calculate the mean square error of a set with n data points, the following equation is used, where y_i's and y'_i's are equal to the actual and predicted values, respectively, and n is the measure of the data set.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2}. \quad (3)$$

2.3 Switching Regression

The Markov chain is a famous type of stochastic process that has many applications. The Markov switching approach is useful when a series is thought to undergo shifts from one type of behavior to another and back again, but where the forcing variable that causes the regime shifts is unobservable. The Markov process defines movements of the state variable among regimes [22]. Markov's property is:

$$P[a < y_t \leq b \mid y_1, y_2, \ldots, y_{t-1}] = P[a < y_t \leq b \mid y_{t-1}]. \quad (4)$$

In order to forecast the probability of a variable that follows a Markov process being in a specific regime at the end of the next period, the current period's probability and a transition probability matrix should be considered [22].

$$P = [p_{ij}]_{m\times m} \,. \quad (5)$$

Where m is the number of regimes and P_{ij} is the transmission probability from regime i to regime j. The vector probabilities in the present state are defined as: [22]

$$\pi_t = [\pi_1\ \pi_2 \ldots \pi_m]\,. \quad (6)$$

where π_i is a probability in state i. It is possible to predict the possibility of being on a certain diet in the next period using π_t and P: [22]

$$\pi_{t+1} = \pi_t . P\,. \quad (7)$$

The Python programming language is used to implement the switching regression models. Using Augmented Dickey Fuller (ADF), the stationarity of data is evaluated, and if necessary, one can use differentiations for stationarity. The ADF is the unit root test for stationary. Unit roots can cause unpredictable results in your time series analysis. A stationarity time series is a sequence of time-dependent values whose mean and variance are not time-dependent and are constant over time.

By analyzing the implemented model, we can approximate the length of periods of high, medium, and low variance. To implement the switching regression model in the studied countries, the number of switching points 1, 2, and 3 have been investigated; in all cases, three switching points have been selected using AIC and BIC. The AIC evaluation criterion measures the amount of information lost by the model. The AIC criterion is a balance between the model parameters and the model's fit to the data. The penalty for losing information in the BIC is greater than the AIC.

2.4 Spatial approach

2.4.1 The linear combination of prediction vectors

Suppose V is a vector space on the field F and α1,… αn are scalars from the field F. In this case, if v1, v2,… vn are vectors of the vector space V, then the linear combination of vi's is:

$$a_1v_1 + a_2v_2 + a_3v_3 + \cdots + a_nv_n. \tag{8}$$

Using the linear combination of the prediction vectors of the countries in the region, it could examine the trend of cases and countries based on the linear combination vector of the prediction.

In order to implement the linear combination model of forecast vectors, after implementing the forecast models for each country, the results of each forecast vector are considered inputs to the model, and the linear combination of the forecast vectors will be the output of the model. Finally, the performance of the model is evaluated and compared with the forecast models of each country using the RMSE.

2.4.2 XGBoost Model

The XGBoost algorithm is a scalable machine learning system for tree boosting. XGBoost is an optimized distributed gradient boosting algorithm that can bring weak classifiers together to make a strong classifier. It efficiently indicates the importance of input features. The advantages of XGBoost are handling missing values, overfitting prevention, and decreasing running time through parallel and distributed calculation. XGBoost uses gradient descent optimization to minimize the loss function [4].

The objective function of XGBoost is the sum of the loss functions for all predictions and the regularization functions for all predictors.

$$obj(\theta) = \sum\nolimits_i L(y_i', y_i) + \sum_j \Omega(f_j)\ , \tag{9}$$

Where f_j is prediction of the jth tree, L is the training loss function to determine the difference between the predicted value y'_i and actual value y_i. The Ω is regularization function to prevent overfitting [4].

In order to implement the XGBoost model, data must first be collected and pre-processed. Then the model parameters are set, and the model is trained. Key features are ranked based on their importance in making a model decision. Due to the fact that this model includes three time steps, the model is actually trained using 39 features (13 countries and, for each country, the information of the last 3 days). After ranking the features, the countries with the most impact on the results are selected for training. Starting features from one country's data will add another country to the model's features until the performance of the model decreases. Finally, the model will be trained again with new features, and evaluation will be performed on the test data. To evaluate the model, the RMSE criterion is used.

2.5 Dataset

In order to predict new confirmed cases of COVID-19 in Iran and neighboring countries including Afghanistan, Pakistan, Iraq, Azerbaijan, Turkey, Russia, Kuwait, Saudi Arabia, Bahrain, Qatar, the UAE, and Kazakhstan, data is used from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) [23]. The training data is from January 29, 2020, to March 31, 2021, and the test data is from April 1, 2021, to April 30, 2021.

3. Trend analysis of cases

3.1 Trend of Iran and neighboring countries

Figure 5 shows the diagrams of the cases in these countries. The test data is marked with a black box.

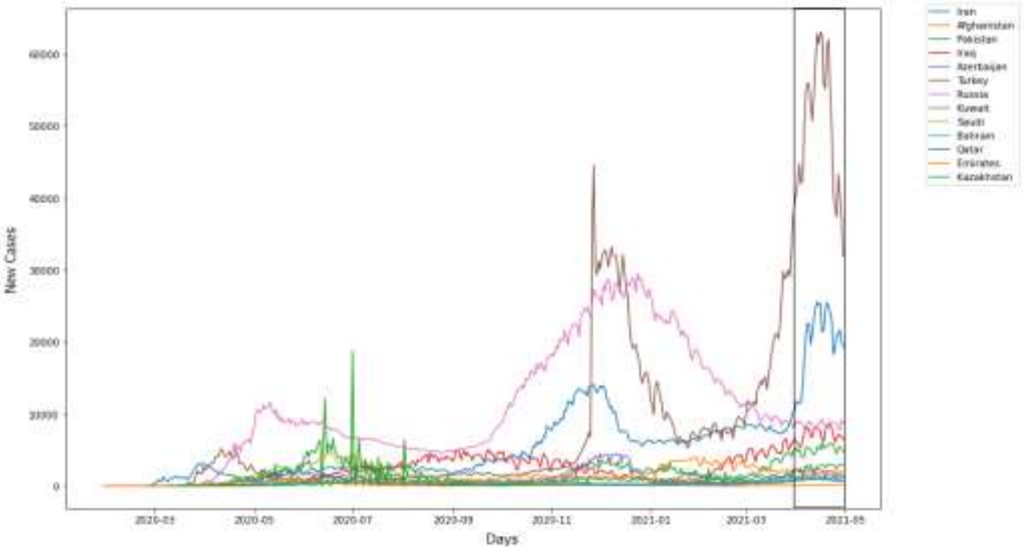

Figure 5: New cases of COVID-19 in Iran and its neighbors

In Iran, Azerbaijan, Russia, the United Arab Emirates, and Turkey, the trend of daily new cases and the peak of new cases are observed simultaneously or with a slight delay.

3.2 Results in Switching Regression Model

Figure 6 shows changes in variance for UEA. The results of implementing the switching regression model and estimating the length of each period for countries are shown in Table 2. According to the results, Bahrain, Iran, Iraq, Azerbaijan, and Russia have a long period of high variance, which indicates that after entering the new peak, more time will be spent at the peak. Therefore, before being in this situation, it is necessary to plan for the necessary measures, create social restrictions, and provide medical facilities to reduce the deaths from the coronavirus. Iran and then Azerbaijan have been in a period of medium variance and indeed unstable conditions for a longer period of time. The use of masks, social distance, and health tips are essential to get out of this situation and achieve a low risk of coronavirus infection.

The implementation results according to the trend in these countries show that in most countries in this region, the length of each corona wave has a repetitive and relatively predictable pattern. In the current situation, with the injection of the vaccine to a high percentage of the population in countries such as Iran, the duration of high and medium variance is expected to decrease.

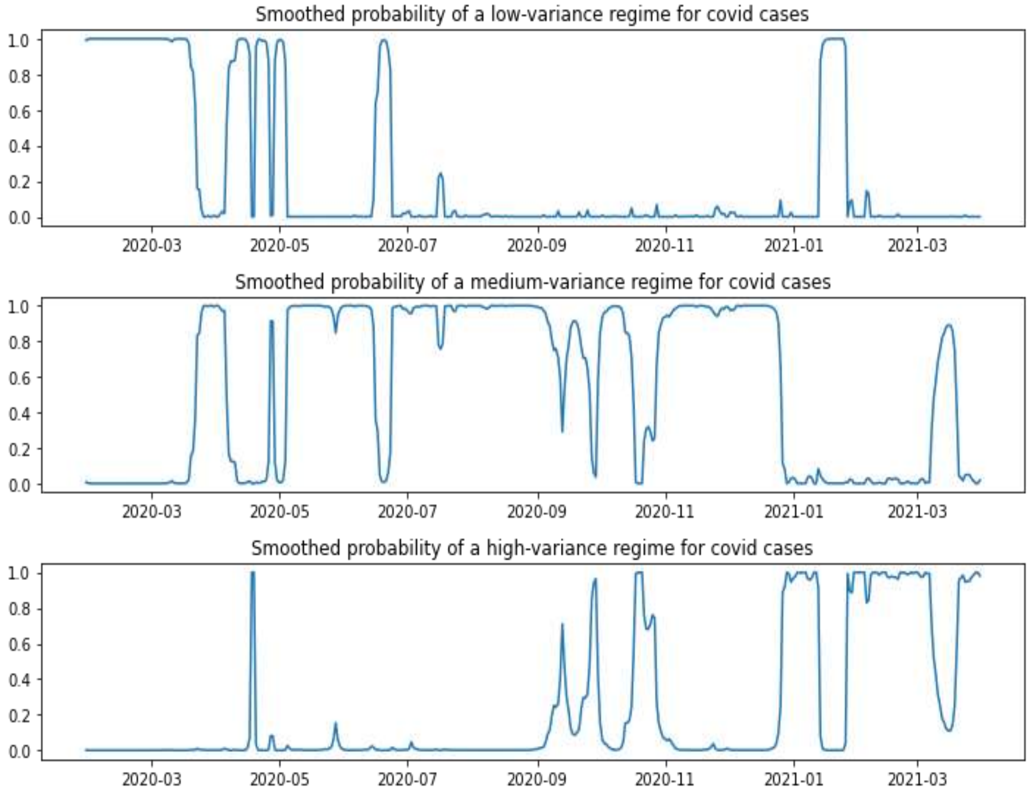

Figure 6: Diagram of switching regression implementation for UEA

Table 1: Switching regression results for Iran and neighboring countries

Country	Low Variance Period	Medium Variance Period	High Variance Period
Iran	24	117	76
Afghanistan	22	23	28
Pakistan	18	13	4
Iraq	85	31	64
Azerbaijan	23	48	62
Turkey	150	25	4
Russia	103	32	61
Kuwait	46	28	17
Saudi	79	27	27
Bahrain	70	24	129
Qatar	10	29	27
Emirates	16	23	14
Kazakhstan	8	4	5

4. Results of prediction

4.1 Results for deep learning and time series models

In this section, the performance results of the implemented LSTM, Bi_LSTM, Conv_LSTM, and ARIMA models to predict the daily confirmed cases of COVID-19 in Iran and neighboring countries are analyzed and compared. The Python programming language is used to implement the models. In deep neural network modeling, open source libraries such as NumPy, Pandas, and Keras have been used.

The training data is divided into training and validation sections with a ratio of 0.7:0.3 and is normalized using the MinMax Scaler. Time step 3 is included in the models; that is, the number of cases on the next day is predicted according to the cases in the last three days. To increase the efficiency of all 3 models, a two-layer stack model has been used. The number of epochs is set to 100, and the number of features and batch size are both set to 1. After training the model and implementing the model architecture, an initial evaluation was performed on the validation data. After modifying the model and obtaining the appropriate result, the final evaluation of the test data has been done.

In the ARIMA model, for each country, the data were examined for stationarity, and (if necessary) differentiation was performed. Figures 7 and 8 show the ACF and PACF charts related to Iran before and after differentiation. The best parameters are selected using the AIC. At the end, the evaluation is done on the test set. Because the daily cases in Iran have a trend, the existing trend has been eliminated by differentiating, and the existing data has become stationary data. The performance of ARIMA models depends on the choice of p, q, and d values. According to the ACF and PACF diagrams, the best values of autocorrelation and moving average are interpreted and determined. The ARIMA models (3, 1, and 3) have been selected for Iran using the AIC.

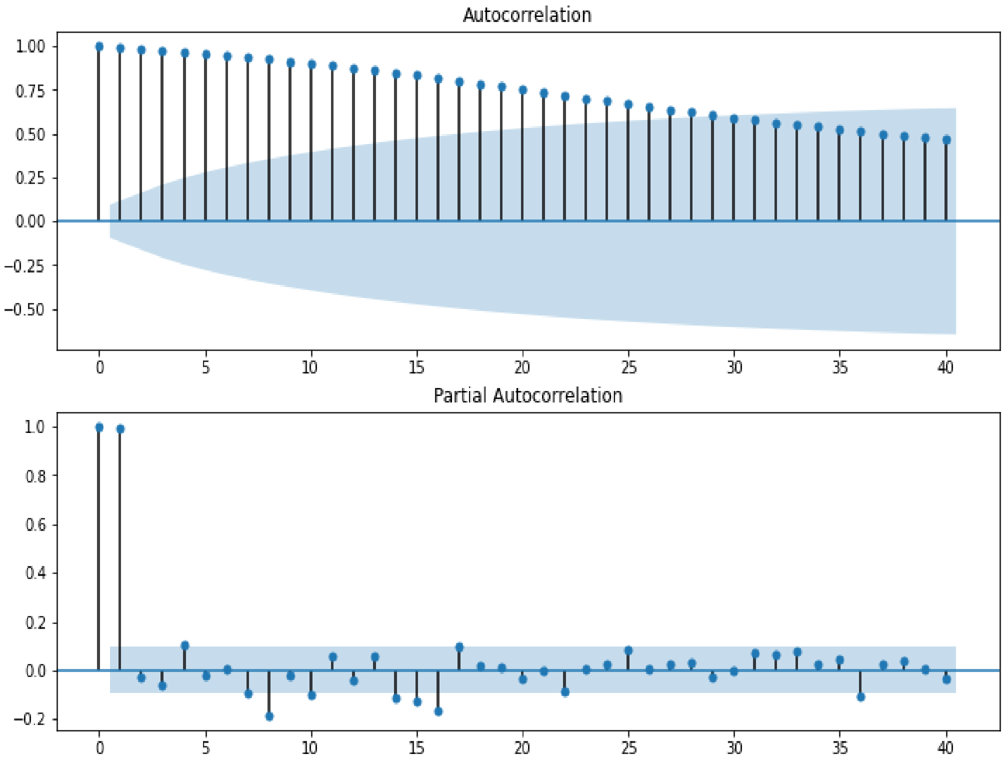

Figure 7: ACF and PACF for real data in Iran

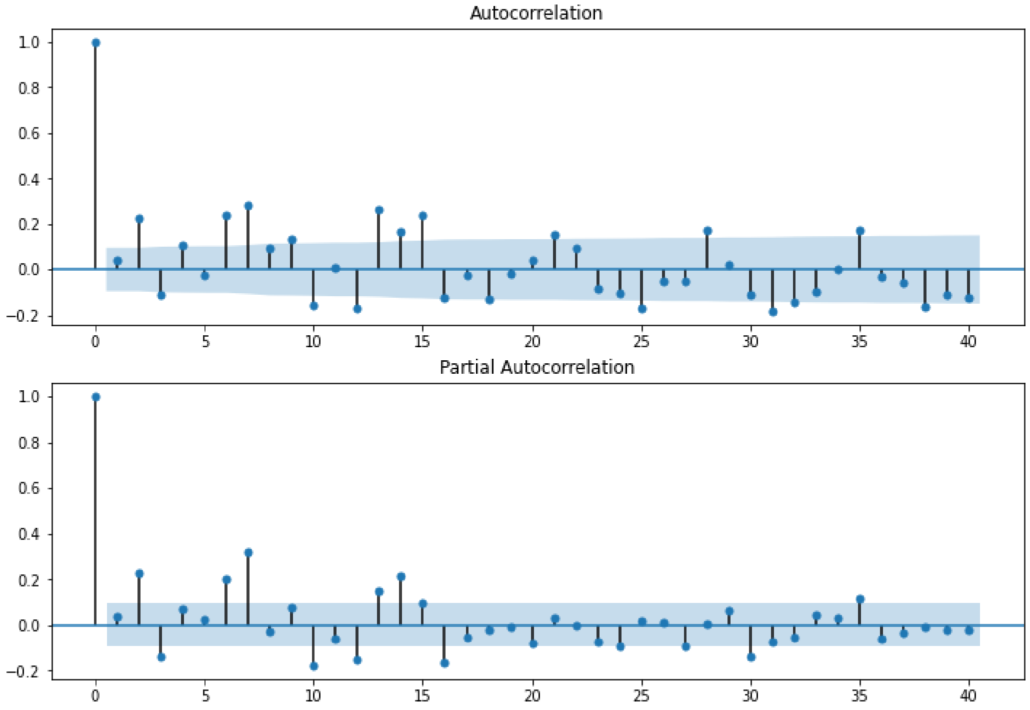

Figure 8: ACF and PACF for differentiation data in Iran

In time-series modeling, the accuracy of the prediction decreases over time. This can be done more in the case of COVID-19 because, due to the influence of various factors, the trend of cases is changing. Therefore, a suitable approach for evaluating test data is to retrain the model using the available real data that has been added to the data set and used in predicting the model. Step-by-step validation is the most desirable solution to obtain more accurate results and has been used in the implementation of all models.

The forecast chart of the LSTM, Bi_LSTM, Conv_LSTM, and ARIMA models for Iran is shown in Figure 9.

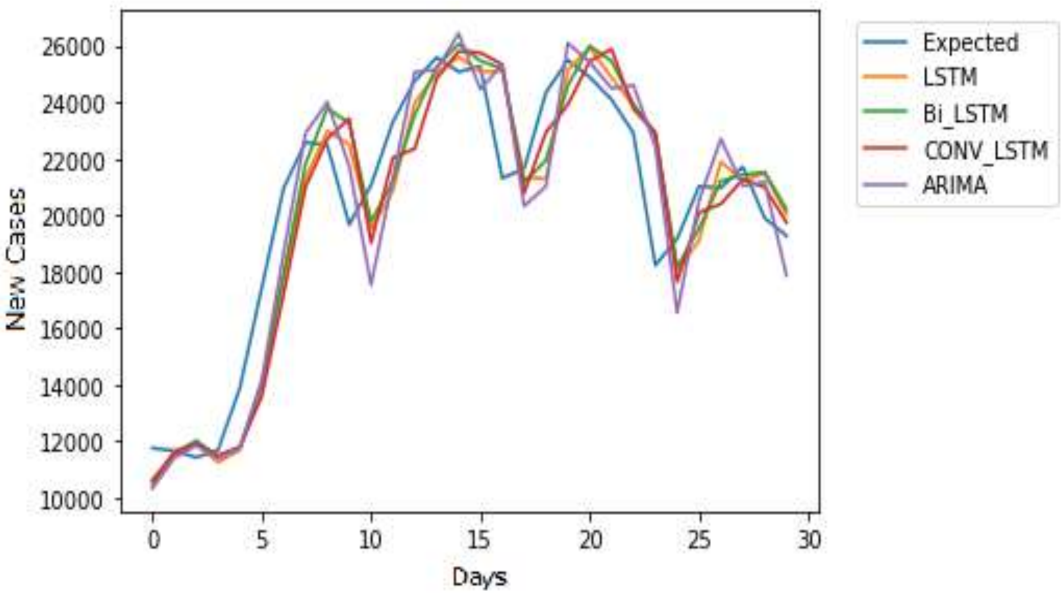

Figure 9: LSTM, Bi_LSTM, Conv_LSTM, ARIMA results to predict new cases in Iran

The results of implementing ARIMA, LSTM, Bi-LSTM, and Conv-LSTM models for the studied countries are as follows (Table 3).

Table2: Implementation results of LSTM, Bi_LSTM, Conv_LSTM, ARIMA for Iran and its neighbors

Country	ARIMA Model	ARIMA RMSE	LSTM RMSE	Bi_LSTM RMSE	CONV_LSTM RMSE
Iran	ARIMA(3,1,3)	1944.5	1876.8	1790.2	2005.7
Afghanistan	ARIMA(0,1,1)	42.2	40.9	42.4	42.1
Pakistan	ARIMA(0,1,3)	582.3	557.6	576.7	557.5
Iraq	ARIMA(2,1,3)	536.9	598.3	581.6	574.3
Azerbaijan	ARIMA(3,2,3)	507.01	491.4	483.0	468.3
Turkey	ARIMA(2,1,0)	4669.2	3950.6	4059.8	3892.0
Russia	ARIMA(0,1,1)	464.1	471.9	499.02	516.6
Kuwait	ARIMA(3,1,3)	123.1	131.7	133.3	134.9
Saudi	ARIMA(0,1,1)	61.4	61.9	58.8	58.4
Bahrain	ARIMA(0,1,1)	108.6	109.19	109.15	110.5
Qatar	ARIMA(1,1,2)	54.1	44.3	40.4	40.1
Emirates	ARIMA(3,1,3)	147.2	127.5	172.2	179.8
Kazakhstan	ARIMA(3,2,2)	338.5	353.4	356.8	343.2

The parameters of the ARIMA model have been determined for all countries, and the model has been predicted using these parameters. According to the RMSE values of the four implemented models, in most countries, the highest and lowest RMSE rates are observed in the ARIMA and Conv_LSTM models. In fact, the best performance for predicting the trend in these countries is achieved with two different approaches: either using simple mathematical models (ARIMA) or using more complex deep neural network models (Conv_LSTM).

The error values are obtained from four models that are very close to each other. In 5 countries (Iraq, Russia, Kuwait, Bahrain, and Kazakhstan), the ARIMA model has the lowest error among other models, and in 5 countries (Pakistan, Azerbaijan, Turkey, Saudi Arabia, and Qatar), the ARIMA model has the lowest error among other models. And also, the Conv_LSTM model had the least error.

In Iran, the Bi-LSTM model had the lowest error, while the LSTM model had the lowest error in Afghanistan and the UAE. These results show that the ARIMA model has been able to predict the daily new cases of COVID-19 in comparison with the models of deep neural networks. Examining the results, in general, among the neural network models, the Bi_LSTM model has not significantly improved compared to the LSTM model in predicting COVID-19 cases, and the performance of both models is almost the same. However, the complexity of the Conv_LSTM model has in some cases improved the results and in others reduced the performance and increased the prediction error.

4.2 Results of prediction for spatial approach

4.2.1 Results for the linear combination of prediction vectors

To implement the linear combination vector of prediction vectors, we have created the linear combination of prediction vectors from LSTM models in the previous section and compared them with the results of single vectors. For this purpose, we review the implementation results over two different time periods. In the first period (training data from January 29, 2020, to March 31, 2021, and test data from April 1, 2021, to April 30, 2021), Iran and Turkey are at the corona peak. In the second period (training data from January 29, 2020, to May 13, 2021, and test data from May 14, 2021, to June 12, 2021), almost no country is in the corona peak. These two intervals are shown in Figure 10. The black box is related to the test data.

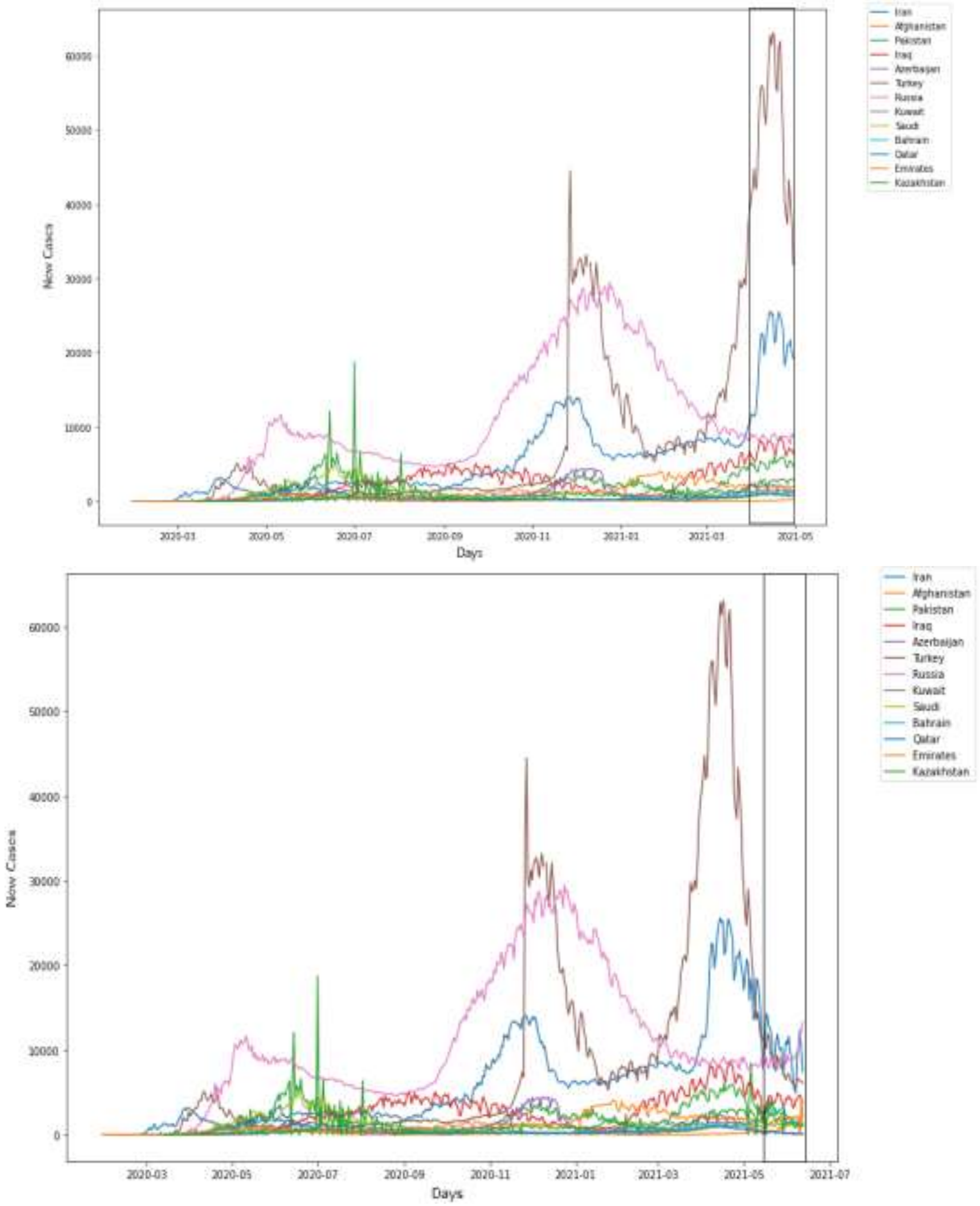

Figure 10: Trend of new cases of COVID-19 for Iran and its neighbors in two intervals

The results of implementing the linear combination model versus single LSTM models for both time periods are shown in Figure 11. (In the error diagram, the orange line is the prediction error of the general model resulting from the linear combination of the vectors, and the blue line is the prediction error of each LSTM model for each country.)

Russia, Turkey, and Iran, due to their high populations compared to other countries, have a higher overall trend (although Pakistan has a very low overall incidence despite having a population of more than 220 million). These countries have a higher trend in cases than the general cases model. In the first case, where Turkey and Iran are at the peak of the corona and there has been a very significant increase in cases, the general model prediction has a high error due to the high difference between cases. In the second case, where no country is in the corona peak, the overall model error is greatly reduced compared to the first case. Therefore, this model has a more accurate interpretation and less error when there is no exposure to the corona peak in these countries.

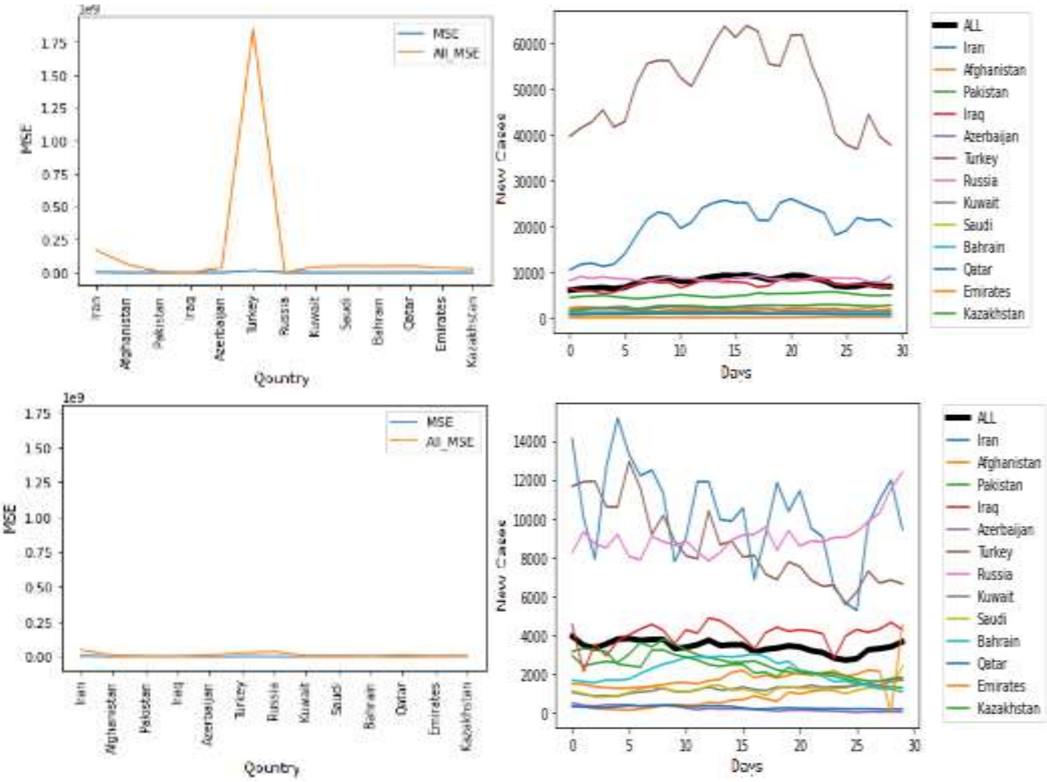

Figure 11: Linear combination versus LSTM in two intervals

4.2.2 Results in XGBoost Model

In this section, the results of the XGBoost model to predict daily cases of Iran using the cases of this country and neighboring countries are reviewed. For implementation, the training data is first divided into two parts, training and validation, and the final evaluation is performed on the test data. The training data is from January 29, 2020, to March 31, 2021, and the test data is from April 1, 2021, to April 30, 2021. These 13 countries are included as model features, and the number of time steps is 3. Training data is divided into two parts with a ratio of 7:3. The first part is used to train the model and select the most important features, and the evaluation is performed on the second part of the data to select the parameters of the final model. The model is set with the parameters of maximum depth (5), learning rate (0.06), estimator 100, and gamma equal to 6.77. After evaluating the model on the preprocessed data, the main features to predict new cases in Iran have been extracted.

According to the results, the most important features in predicting new cases of Iran are data related to Russia, Iran, Pakistan, and Turkey (Figure 12). The six main characteristics for predicting the new cases of one day in Iran are the new cases of two days before in Russia, two days before in Iran, one day before in Russia, one day before in Iran, three days before in Russia, and three days before in Iran.

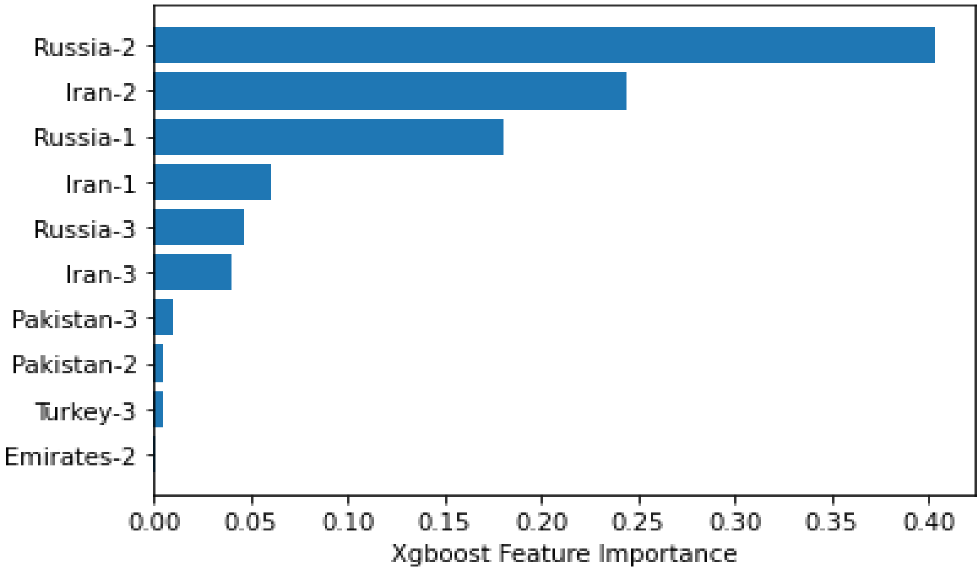

Figure 12: The importance of features in forecasting new cases in Iran

Considering the data of Iran and Russia as the main features, the model was retrained, and with increasing the number of features, the performance of the model has not improved. Therefore, the data of Iran and Russia have been selected as the main feature. The model has been retrained using data from Iran and Russia. To avoid over-fitting, some model parameters have been omitted, and the estimator number remains at 100. Then, using the main features, the test data is predicted (Figure 13). The model error using the XGBoost model using Covid cases in Russia, which had a similar trend to Iran, versus the XGBoost prediction model using only Iranian Covid cases, is decreased (Table 4). Although this error is higher than the error calculated by deep neural network models and statistical models, using this model, trending and effective countries in predicting the new cases of Iran have been identified. Russia's data with a high impact (even more than Iran’s new cases) in predicting Iran's cases has been able to significantly reduce the model error.

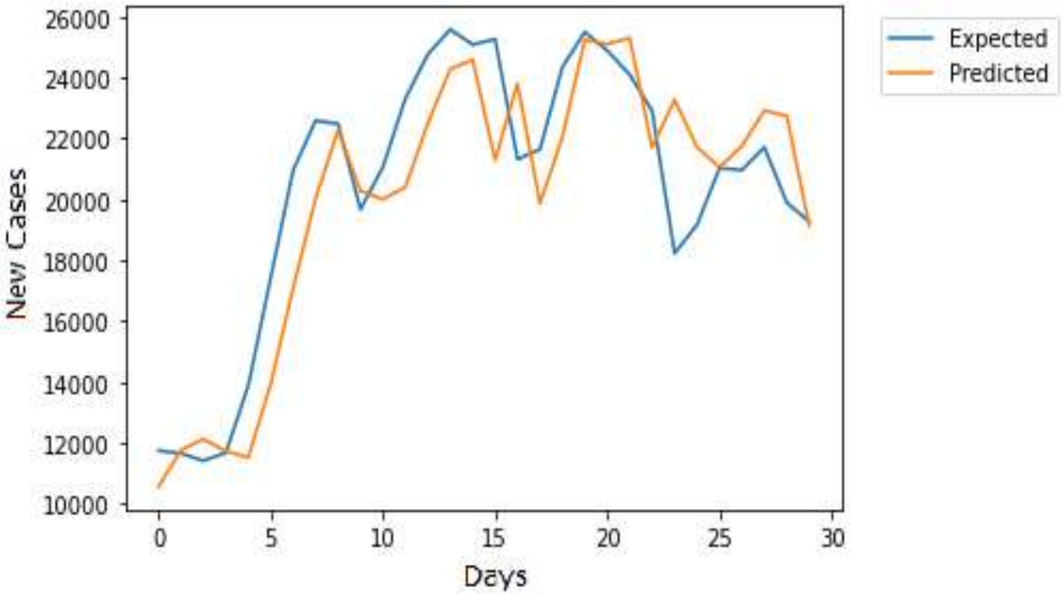

Figure 13: XGBoost implementation for new cases of COVID-19 in Iran

Table 3: Comparative RSME for implementation models to Iranian cases

country	ARIMA RMSE	LSTM RMSE	Bi_LSTM RMSE	CONV_LSTM RMSE	XGBoost_ALL RMSE	XGBoost RMSE
Iran	1944.5	1876.8	1790.2	2005.7	2111.2	2279.6

5. Conclusion

In this paper, a comprehensive prediction and analysis of COVID-19 new cases in Iran and neighboring countries has been done using deep learning models and time series. Using John Hopkins University's dataset, which includes data related to the daily cases of COVID-19 in different countries of the world, predictions of new cases in these countries have been implemented, and the results have been reviewed.

In the first section, a switching regression model is implemented to determine time intervals of change in COVID-19 new cases. Depending on the length of the intervals, Bahrain, Iran, Iraq, Azerbaijan, and Russia have long periods of high variance, and these countries will spend more time at the peak after entering the new corona peak. Therefore, it is necessary to take urgent decisions to manage the pandemic and reduce patient mortality.

In the second section, ARIMA, LSTM, Bi_LSTM, and Conv_LSTM models are developed to predict new cases of COVID-19 in each country, and the results are compared. According to the results, ARIMA and Conv_LSTM models have better results than others. These results show that time series models can have good results compared to deep neural network models. In Iraq, Russia, Kuwait, Bahrain, and Kazakhstan, the ARIMA model had the lowest error. In Pakistan, Azerbaijan, Turkey, Saudi Arabia, and Qatar, the Conv_LSTM model performed better. The bi-LSTM model performed best in Iran, while the LSTM model performed best in Afghanistan and the UAE. The best performance of predictions in these countries has been achieved using the ARIMA mathematical model and the Conv_LSTM deep neural network model.

In the third part of the research, the linear combination model of prediction vectors and the XGBoost model are implemented to predict the daily new cases of COVID-19 in Iran according to the trends of this country and neighboring countries. The linear combination model of predictive vectors was influenced by data from Iran, Russia, and Turkey, which have a higher new case trend than other countries in the region. Therefore, when one of these countries is at its peak, the model error increases dramatically. But in another period, when these countries were not at the peak of the corona, the error was greatly reduced. Using the XGBoost model, the most important features to predict new cases of Iran using the trends of countries in the region are Russia, Iran, Pakistan, and Turkey. Using the data of Russia and Iran, the error of predicting cases in Iran has been significantly reduced compared to the implemented model using only the data of Iran. Also, influential countries have been identified in predicting the trend of new cases in Iran.

Prediction models and analyzing new cases can be used as an effective solution by policymakers to enforce the necessary laws, allocate the required resources, improve pandemic management, and reduce mortality. Managers in different countries can plan to improve current performance and prevent a new COVID-19 wave.

References

[1] "World Health Organization website,." https://covid19.who.int/.

[2] F. M. Khan and R. Gupta, "ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India," *J. Saf. Sci. Resil.*, vol. 1, no. 1, pp. 12–18, 2020, doi: 10.1016/j.jnlssr.2020.06.007.

[3] Y. A. Shiferaw, “Regime shifts in the COVID-19 case fatality rate dynamics: A Markov-switching autoregressive model analysis,” *Chaos, Solitons Fractals X*, vol. 6, p. 100059, 2021, doi: 10.1016/j.csfx.2021.100059.

[4] J. Luo, Z. Zhang, Y. Fu, and F. Rao, “Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms,” *Results Phys.*, vol. 27, p. 104462, 2021, doi: 10.1016/j.rinp.2021.104462.

[5] V. K. R. Chimmula and L. Zhang, “Time series forecasting of COVID-19 transmission in Canada using LSTM networks,” *Chaos, Solitons and Fractals*, vol. 135, 2020, doi: 10.1016/j.chaos.2020.109864.

[6] N. Ayoobi *et al.*, “Time Series Forecasting of New Cases and New Deaths Rate for COVID-19 using Deep Learning Methods,” *Results Phys.*, vol. 27, no. June, p. 104495, 2021, doi: 10.1016/j.rinp.2021.104495.

[7] T. Chakraborty and I. Ghosh, “Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis,” *Chaos, Solitons and Fractals*, vol. 135, 2020, doi: 10.1016/j.chaos.2020.109850.

[8] S. Singh, K. S. Parmar, J. Kumar, and S. J. S. Makkhan, “Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19,” *Chaos, Solitons and Fractals*, vol. 135, pp. 1–8, 2020, doi: 10.1016/j.chaos.2020.109866.

[9] G. Toğa, B. Atalay, and M. D. Toksari, “COVID-19 prevalence forecasting using Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Networks (ANN): Case of Turkey,” *J. Infect. Public Health*, vol. 14, no. 7, pp. 811–816, 2021, doi: 10.1016/j.jiph.2021.04.015.

[10] K. E. ArunKumar, D. V. Kalaga, C. M. Sai Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, “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 Average,” *Appl. Soft Comput.*, vol. 103, 2021. doi: 10.1016/j.asoc.2021.107161.

[11] P. S. Knopov and A. S. Korkhin, “Statistical Analysis of the Dynamics of Coronavirus Cases using Stepwise Switching Regression,” *Cybern. Syst. Anal.*, vol. 56, no. 6, pp. 943–952, 2020, doi: 10.1007/s10559-020-00314-w.

[12] L. Yan *et al.*, “Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan,” *medRxiv*, p. 2020.02.27.20028027, 2020, [Online]. Available: http://medrxiv.org/content/early /2020/03/03/2020.02.27.20028027.abstract.

[13] Y. Suzuki and A. Suzuki, “Machine learning model estimating number of COVID-19 infection cases over coming 24 days in every province of South Korea (XGBoost and MultiOutputRegressor),” *medRxiv*, no. Ml, pp. 1–11, 2020, doi: 10.1101/2020.05.10.20097527.

[14] P. Wang, X. Zheng, G. Ai, D. Liu, and B. Zhu, “Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran,” *Chaos, Solitons and Fractals*, Vol. 140, January, 2020.

[15] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun, “Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study,” *Chaos, Solitons and Fractals*, vol. 140, p. 110121, 2020, doi: 10.1016/j.chaos.2020.110121.

[16] N. Yudistira, S. B. Sumitro, A. Nahas, and N. F. Riama, “Learning where to look for COVID-19 growth: Multivariate analysis of COVID-19 cases over time using explainable convolution–LSTM,” *Appl. Soft Comput.*, vol. 109, p. 107469, 2021, doi: 10.1016/j.asoc.2021.107469.

[17] S. Shastri, K. Singh, S. Kumar, P. Kour, and V. Mansotra, “Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study,” *Chaos, Solitons and Fractals*, vol. 140, p. 110227, 2020, doi: 10.1016/j.chaos.2020.110227.

[18] P. Arora, H. Kumar, and B. Ketan, “Prediction and analysis of COVID-19 positive cases using deep learning models : A descriptive case study of India,” *Chaos, Solitons Fractals Interdiscip. J. Nonlinear Sci. Nonequilibrium Complex Phenom.*, vol. 139, p. 110017, 2020, doi: 10.1016/j.chaos.2020.110017.

[19] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. “CRISP-DM 1.0: Step-by-step data mining guide.” (2000).

[20] Sepp Hochreiter, Jürgen Schmidhuber, "Long Short-Term Memory," Neural Comput, vol. 9, no. 8, pp. 1735–1780, 1997. doi: https://doi.org/10.1162/neco.1997.9.8.1735

[21] T. Dehesh, H. A. Mardani-Fard, and P. Dehesh, "Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models," *medRxiv*, pp. 1–12, 2020, doi: 10.1101/2020.03.13.20035345.

[22] C. Brooks, Introductory Econometrics for Finance, 4th ed. Cambridge: Cambridge University Press, 2019.

[23] "John Hopkins University repository," https://github.com/CSSEGISandData/COVID-19.