Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1247962

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 6 - Issue 4: 421-425 / October 2023

FORECASTING COVID-19 CASES IN TÜRKİYE WITH THE HELP OF LSTM

Nurgül GÖKGÖZ KÜÇÜKSAKALLI[1]*

[1]*Çankaya University, Faculty of Arts and Sciences, Department of Mathematics, 06790, Ankara, Türkiye*

Abstract: Even though, it is thought that the pandemic has come to an end, the humanity is still under the danger of upcoming pandemics. In that sense, every effort to understand or predict the nature of an infectious disease is very precious since those efforts will provide experience for upcoming infectious disease epidemic/pandemic. Mathematical models provide a common way to analyze the nature of the pandemic. Apart from those mathematical models that mostly determine which variables should be used in the model to predict the nature of the epidemic and at which rate the disease will spread, deep learning models can also provide a fast and practical tool. Moreover, they can shed a light on which variables should be taken into account in the construction of a mathematical model. And also, deep learning methods give rapid results in the robust forecasting trends of the number of new patients that a country will deal with. In this work, a deep learning model that forecasts time series data using a long short-term memory (LSTM) network is used. The time series data used in this project is COVID-19 data taken from the Health Ministry of Republic of Türkiye. The weekend isolation and vaccination are not considered in the deep learning model. It is seen that even though the graph is consistent and similar to the graph of real number of patients, and LSTM is an effective tool to forecast new cases, those parameters, isolation and vaccination, must be taken into account in the construction of mathematical models and also in deep learning models as well.

Keywords: Long-Short Term Memory (LSTM), COVID-19, Forecasting, Modeling

***Corresponding author:** Çankaya University, Faculty of Arts and Sciences, Department of Mathematics, 06790, Ankara, Türkiye
E mail: ngokgoz@cankaya.edu.tr (N. GÖKGÖZ KÜÇÜKSAKALLI)
Nurgül GÖKGÖZ KÜÇÜKSAKALLI https://orcid.org/0000-0002-9640-4194

Received: February 27, 2023
Accepted: September 10, 2023
Published: October 15, 2023

Cite as: Gökgöz Küçüksakalli N. 2023. Forecasting COVID-19 cases in Türkiye with the help of LSTM. BSJ Eng Sci, 6(4): 421-425.

1. Introduction

The world is still fighting against the coronavirus disease 2019 (COVID-19) caused by the novel coronavirus, SARS-CoV-2, which is thought as a highly contagious virus that affects the human respiratory system as a first step. Many researchers from different fields started to work on this topic to find a treatment, to understand the nature of the disease or to forecast the trend in the case numbers or death numbers.

To provide a solid background to the biological data, mathematical models provide a powerful tool (Belen et al., 2011; Pedamallu et al., 2012; Brauer et al., 2019). In that sense, many mathematical models are investigated to estimate the trend in the case/death numbers or to decide how contagious is this disease. The mathematical approach used in the model aside (ordinary differential equations, partial differential equations, difference equations, etc.), how many variables are considered make a difference in the arrangement of the model. It is certain that some simplifications and therefore assumptions must be made in order to be able to analyze the model. From the mathematical modeling perspective, some of the works disregard the incubation period (Arino et al., 2020), some works include the social isolation period in their model (Vega, 2020), some consider both social isolation and vaccination (Demirci, 2023). With those models, basic reproduction number, i.e., the number of next generation cases produced by the present generation, stability of the model and numerical simulations are obtained (Demirci, 2023). However, which variables must be taken into account, which parameters should be included, basically which simplifications or assumptions we are allowed to make for a more realistic mathematical model must be determined. In that sense, apart from mathematical models, deep learning, machine learning, and artificial intelligence models are very popular in identification (Subramanian et al., 2022; Paul et al., 2023) and forecasting the trend of any real world problem (Livieris et al. 2020), and also COVID-19 cases (Xu et al. 2022). Apart from their effectiveness in the forecast of the trends, deep learning models can be used to overcome the difficulties we face listed above in the use of mathematical models. The model should be fast, reliable and practical as possible as to allow government to take action before the pandemic spreads quickly.

Basically, deep learning refers to an artificial neural network with feature learning. It uses multiple layers in the architecture of the network and that is the reason why it is called as deep learning. When we talk about the deep learning methods, it has to be mentioned that one of the extensive types of artificial networks is Recurrent Neural Networks (RNN) which is capable of using

arbitrary data from their internal states (Tealab 2018). However, they can capture data only from a few steps earlier. A long-short term memory (LSTM) networks which is a kind of RNN has been proved to be successful in many applications (Graves et al., 2008). Throughout the years many different applications of LSTM have been conducted to various fields (Wang et al. 2020; Xu et al., 2022) and proved to be successful and therefore, they are still extensively used.

In this project, an LSTM network is used where the LSTM network learns to predict the value of the next time step. The data is COVID-19 data from 27th April 2021 to 21st June 2021 taken from the Health Ministry of Republic of Türkiye (one may check https://covid19.saglik.gov.tr/EN-69532/general-coronavirus-table.html).

2.1. Related Work

When the use of machine learning is searched throughout the literature, it can be seen that they are both used in the identification of a pattern and in the forecasting of some cases. When COVID-19 cases are considered, deep learning or machine learning networks are used both in detection and prediction of it (Paul et al., 2023; Subramanian et al., 2022; Xu et al., 2022; Jin et al., 2021). To be more precise, machine learning models that predict antibody response using antibody sequences (Magar et al., 2021), using chest X-rays to determine if the patient is infected or not (Toğaçar et al., 2020), forecasting cases of COVID-19 (Xu et al., 2022; Wang et al., 2020).

Making a prediction using time series data play an important role in many areas, and especially multi-step time-series forecast is challenging and essential in many real world problems e.g. forecasting stock-price, river flow, disease cases, etc. The most effective multi-step model is Long Short Term Memory (LSTM) because of its structure that allows capturing the long-term dependencies. There are works on the literature that apply LSTM on different data patterns (Yunpeng et al., 2017). To name a few of them, there are works on financial data (Siami-Namini et al., 2018), petroleum production, gold price (Sagheer et al., 2019; Livieris et al., 2020). Forecasting COVID-19 cases using LSTM networks plays an important role in the literature as well. It has been shown that it is more effective than convolutional neural networks (CNN), and a combination of LSTM-CNN (Xu et al. 2022). Therefore, in this work, we use LSTM networks to forecast COVID-19 cases in Türkiye.

2. Materials and Methods

In this section, the data used in the experiment and the steps used to construct the model are explained in detailed. The model is constructed and implemented on MATLAB R2019b. For a better explanation, figures of the data, forecasted data, graphs of RMSE are given.

2.1. Dataset

The data includes a single time series where time steps are days and the values are the number of COVID-19 cases in Türkiye. The data is a cell array, where each element is a time step (Figure 1). Then the data is divided into two groups as the training and the test data. The training is done on the first 90% of the sequence and the test is done on the remaining 10%. At this point, a standardization of the data has been applied. It is crucial because, it allows data exchange in different models or computers. Moreover, it improves quality of the data and therefore leads to a better decision-making. Therefore, standardization on the training data is done in order to obtain a better fit and to get away from the divergence of the training. By this way, the training data have zero mean and unit variance. Later on, the test data is standardized using the same parameters as the training data. During the overall process, CPU is used since the data is rather small collection of data.

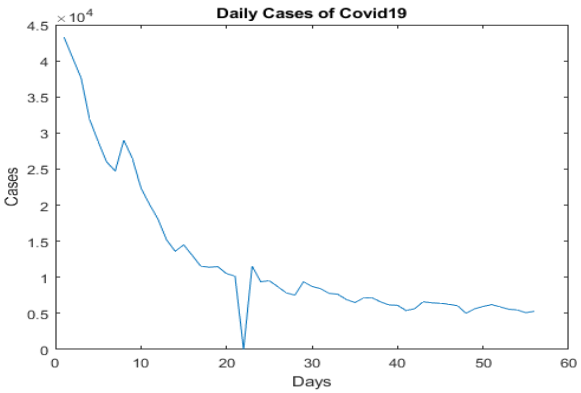

Figure 1. COVID-19 daily dataset of Türkiye (starts on 27th April 2021, ends on 21st June 2021).

2.2. Procedure

After data preparation step, an LSTM network is constructed. A flowchart of the general workflow of this structure can be seen from Figure 2.

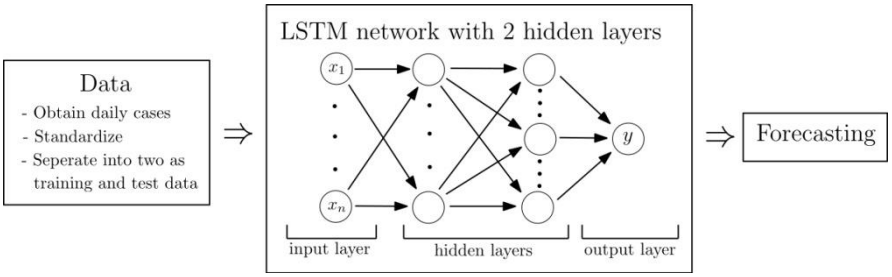

Figure 2. Flowchart that shows the workflow of the LSTM network to forecast cases of COVID-19 according to daily dataset of Türkiye.

At every time step of the input sequence, the network learns to forecast the next time step's value. In this work, the LSTM layer has 200 hidden units. The training is executed for 250 epochs with the solver 'adam'. The get away from the problem of gradient exploding, gradient threshold is chosen as 1. Moreover, the learning rate is initially is chosen as 0.005 and dropped to 0.001 after the halfway of the training. A standardization is applied on the test data using the same parameters as the training data. A detailed list of hyper parameter optimization and model architecture is given in Table 1.

Table 1. Parameters for the architecture for the LSTM network

Model	LSTM
Hidden Layer	2
Hidden Units	200
Learning Rate	0.005
Optimizer	Adam
Epoch	250

Using the training progress plot, the root-mean-square error (RMSE) may be calculated from the standardized data. In this work, for the error calculation root-mean-square error (RMSE) is used. Even though, it is not scale invariant, using standardized data overcomes this disadvantage. Moreover, it is mostly used to check the model performance. In the literature, (RMSE) formula (equation 1) is given by

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

where y_i values are observed values, $\hat{y}_i$ are predicted values and n is the number of observations. In our model, the MATLAB output for the RMSE measure (equation 2) is

RMSE= 528.0957. (2)

One may see the forecasted values (Figure 3). We may compare the forecasted values with the test data (Figure 4). Since we have the actual values of time steps between predictions as well, then we update the network state with the observed values (Figure 5).

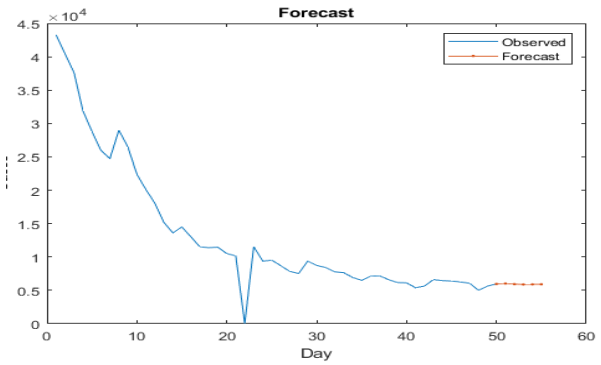

Figure 3. Forecasted data. The red curve indicates the forecasted data in the graph.

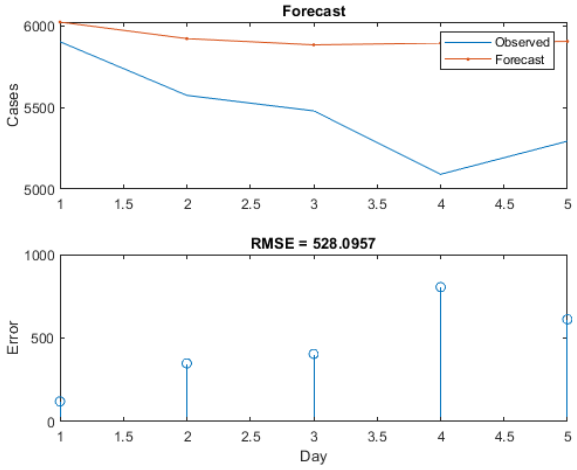

Figure 4. Forecasted values with the test data (with predicted values).

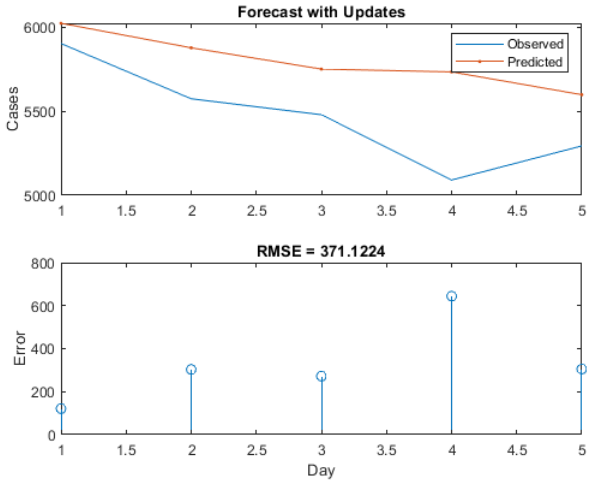

Figure 5. Forecasted values with the test data (with actual values).

In that case, the MATLAB output for the RMSE measure (equation 3) is

RMSE= 371.1224. (3)

As expected, in that case, we have a more accurate prediction. Moreover, it can be concluded that, by investigating Figure 4 and Figure 5, observed values and the forecasted values show a similar trend. Therefore, even though the error value, since the general behavior of the system is captured, it can be stated that the model should include the disregarded variables such as isolation and vaccination.

3. Results and Discussion

In this work, forecasting of the number of cases for COVID-19 which is a time series data is considered. Obtaining a future prediction in the number of cases is very important from different aspects as explained earlier. In the literature, LSTM networks are proved to be efficient among different deep learning networks among

other deep learning models (Xu et al., 2022; Wang et al. 2020). In this work, we use LSTM networks to determine which parameters will be used in a mathematical model. We conclude that, even though the RMSE value is high, this study shows that predicting of the number of cases is available by using LSTM. That is because the tendency of the observed and the forecasted values are similar to each other in the graphs (Figure 4 and Figure 5). High RMSE value is depending on the vaccination of different age groups and occupation groups started before and during this period of time and the quarantine applied during weekends. Moreover, high RMSE value is known to be causing from the outliers in the dataset. In this case, we may conclude that a mathematical model should include those parameters, as well. This result is also compatible with the mathematical models considered in the literature (Demirci, 2023). Therefore, we conclude that even though LSTM is an effective tool to predict number of cases, other factors (like vaccination, quarantine) must be considered in the future models. In the literature, there are works that consider different deep learning networks and compare them according to different error measures (Xu et al., 2022). As a future work, a comparison of the LSTM model with different deep learning networks with different error measures may be considered. Moreover, using LSTM networks to determine which parameters to include in a mathematical model may be extended to different real world problems. To name a few of them, it may be used in tumor-immune system (Gokgoz et al., 2021), HIV/AIDS transmission (Pedamallu et al., 2012), rumour model (Belen et al. 2011) or fishery model (Çifdalöz, 2022). By this way, by adding more variables in the mathematical models better reproduction number, more realistic stability results, etc. will be obtained.

Author Contributions

The percentage of the author contributions is presented below. The author reviewed and approved the final version of the manuscript.

	N.G.K.
C	100
D	100
S	100
DCP	100
DAI	100
L	100
W	100
CR	100
SR	100
PM	100
FA	100

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

Conflict of Interest

The author declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans. The authors confirm that the ethical policies of the journal, as noted on the journal's author guidelines page, have been adhered to.

References

Arino J, Protet S. 2020. A simple model for COVID-19. Infectious Disease Modelling, 5: 309-315.

Belen S, Kropat, E, Weber, GW. 2011. On the classical Maki–Thompson rumour model in continuous time. Cent Eur J Oper Res, 19: 1–17.

Brauer F, Castillo-Chavez C, Feng Z. 2019. Mathematical models in epidemiology. Springer-Verlag, New York, USA, First Edition, pp: 254.

Çifdalöz, O. 2022. Sustainable Management of a Renewable Fishery Resource with Depensation Dynamics from a Control Systems Perspective. Gazi University J Sci, 35 (3): 936-955.

Demirci E. 2023. A Novel Mathematical Model of the Dynamics of COVID-19. GU J Sci, 36(3): 1302-1309.

Gokgoz N, Oktem H. 2021. Modeling of tumor-immune system interaction with stochastic hybrid systems with memory: a piecewise linear approach. Advances in the Theory of Nonlinear Analysis and its Application, 5(1): 25-38.

Graves A, Schmidhuber J. 2008. Offline handwriting recognition with multidimensional recurrent neural networks. Advances in neural information processing systems, 21, 545-552.

Jin W, Stokes JM, Eastman RT, Itkin Z, Zakharov AV, Collins JJ, Jaakkola TS, Barzilay R. 2021. Deep learning identifies synergistic drug combinations for treating COVID-19. In: Proceedings of the National Academy of Sciences of the United States of America, 118(39): e2105070118.

Livieris IE, Pintelas E, Pintelas, P A. 2020. CNN–LSTM model for gold price time-series forecasting. Neural Comput Applic, 32: 17351–17360.

Magar R, Yadav P, Farimani AB. 2021. Potential neutralizing antibodies discovered for novel corona virus using machine learning. Sci Rep, 11: 5261.

Paul SG, Saha A, Biswas A, Zulfiker S, Arefin MS, Rahman M, Reza AW. 2023. Combating Covid-19 using machine learning and deep learning: Applications, challenges, and future perspectives. Array, 2023: 100271.

Pedamallu CS, Özdamar L, Kropat E, Weber GW. 2012. A system dynamics model for intentional transmission of HIV/AIDS using cross impact analysis. CEJOR, 20(2): 319-336.

Sagheer A, Kotb M. 2019. Time series forecasting of petroleum production using deep LSTM recurrent networks. Neurocomput, 323: 203-213.

Siami-Namini S, Tavakoli N, Siami Namin A. 2018. A Comparison of ARIMA and LSTM in Forecasting Time Series. In: 17th IEEE International Conference on Machine Learning and Applications (ICMLA): pp: 1394-1401. doi: 10.1109/ICMLA.2018.

Subramanian N, Elharrouss O, Al-Maadeed S, Chowdhury M. 2022. A review of deep learning-based detection methods for COVID-19. Computers Biol Med, 2022: 105233

Tealab A. 2018. Time series forecasting using artificial neural networks methodologies: A systematic review. Future Computing Informatics J, 3(2): 334-340.

Toğaçar M, Ergen B, Cömert Z. 2020. COVID-19 detection using deep learning models to exploit social mimic optimization and structured chest X-ray images using fuzzy color and stacking approaches. Comput Biol Med, 121: 103805.

Vega DI. 2020. Lockdown, one, two, none, or smart. Modeling containing COVID-19 infection. A conceptual model. Sci Total Environ, 730: 138917.

Wang P, Zheng X, Ai G, Liu D, Zhu B. 2020. 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 Solit Fractals, 140.

Xu L, Magar R, Farimani AB. 2022. Forecasting COVID-19 new cases using deep learning methods. Computers Biol Med, 2022: 105342.

Yunpeng L, Di H, Junpeng B, Yong Q. 2017. Multi-step ahead time series forecasting for different data patterns based on lstm recurrent neural network. In: 14th Web Information Systems and Applications Conference (WISA): pp: 305-310. doi: 10.1109/WISA.2017.25.