NTU JOURNAL OF PURE SCIENCES
EISSN: 2789-1097

RESEARCH ARTICLE
Open Access

Predict the Risk Level in Iraqi Governorates According to the Spread of COVID-19 Using Data Mining

Ibtisam Abbas Othman
ibtisamabbas81@gmail.com

Ban Sharief Mustafa
Banmustafa66@uomosul.edu.com

Received: 06-03-2022, **Accepted:** 22-03-2022, **Published online:** 24-05-2022

Abstract. The massive spread of COVID-19 made it one of the biggest current pandemics in the world. Predicting the extent of the virus' spread is critical to containing the threat, because it helps to take appropriate measures and decisions at the state level as well as at the personal level, where it is possible to avoid travel to the places of spread and take the necessary measures to limit the spread of the virus. In this research, an intelligent model has been built to predict the extent of the spread of the Covid-19 disease in the Iraqi governorates. COVID-19 data for Iraq's governorates was obtained from a website affiliated with the Iraqi Ministry of Health. The data was reconstructed according to a certain structure to be used in training the prediction model. The LSTM deep learning algorithm was adopted for its effective performance in predicting the direction of the recorded cases in the future. The results showed high accuracy in the performance of the proposed model. The model performance was measured using mean logarithmic square error (RMSLE) loss function. The model loss values are 0.0833 for predicting number of cases, 0.0616 for predicting number of deaths and 0.6096 for predicting the incidence value for cases. It can be said that this is the first model built to predict COVID-19 cases based on IRAQ's governorates dataset.

Keywords: Covid-19, Covid-19 prediction tool, Iraq's covid-19 dataset, LSTM.

Introduction

Recently, and due to its efficient and successful performance, Deep Learning (DL) algorithms have become widely used in many computing applications. In December 2019, the world witnessed a widespread spread of a new strain of the SARS virus, which was named COVID-19. Its rapid transmission from one person to another led to the spread of the virus on a large scale [1]. On 30 January, 2020, the World Health Organization declared that Covid-19 represents a health emergency of concern to the international community. Therefore, on basis of this rapid spread of the virus in many countries of the world, the Director-General of the World Health Organization announced that Coronavirus has become an epidemic that threatens the lives of millions of people in the world [2]. Non-clinical approaches such as data mining and machine and deep learning techniques can help reduce the spread of the virus. Data mining is used to find similar patterns and extract knowledge from large amounts of data. Machine and deep learning algorithms are data processing techniques that help in building analytical models [3]. In this research, the aim is to predict the number of infected persons or deaths in each Iraqi governorate and for a subsequent period determined by a range of days. According to the information available to us, Covid-19 dataset for Iraq's governorates has not been use till now in a research paper. The data available on the World Health Organization's HDX website has been used, processed and restructured to be suitable as inputs for training models. A deep learning model based on the LSTM network has been built to predict the path of disease spread, such as the number of new cases or the number of deaths in the coming days according to the Iraqi- governorates, and it has achieved very good performance rates. Predicting the current situation in Iraq is critical to containing the threat because it helps to take appropriate measures and decisions at the governorates level as well as at the personal level where travel to places of deployment can be avoided. It is also possible to study the impact of some social and religious events on increasing the natural expectation rate of the model in some governorates.

Related work

Several research papers have been published in the field of contributing to the response to the Corona pandemic. There has been a lot of research that focused on building predictive models to predict the spread of disease in specific geographical locations. Ramchandani and others [4], proposed a deep learning model to predict an increase in (Covid-19) cases in the coming days. The model is used to identify the features most influential in predicting infection growth. Experiments show that the proposed model obtains satisfactory predictive performance. The proposed model can complement the national standard epidemiological models for epidemiological surveillance, such as (Covid-19). Wendi [5] introduced a model that implements the Boosted Random Forest algorithm, and the model provides accurate predictions even on unbalanced data sets. The data analyzed revealed that death rates were higher among Wuhan citizens than among non-citizens. As well as the male mortality rate was more aware of it than the female. The ages of the infected people ranged between 20 and 70 years. Roy [6] study relied on data obtained from January 30 to April 26, 2020, and from April 27, 2020 to May 11, 2020 as samples for modeling and forecasting, respectively, in which an empirical study was conducted in predicting the pattern of the Covid-19 epidemic, and they also compared the actual and expected differences of values in principle and practical aspects. The area has been classified into a high-risk, medium-risk, and low-risk area for COVID-19. Ramadan [7] presents a potential spatial attempt to model the prediction of COVID-19 risk areas in Cairo Governorate, Egypt. The developed approach builds on previous studies and the guidelines for the Sustainable Development Goals of the United Nations taking into account good health and implementation. The proposed model provides a systematic framework for forecasting areas highly vulnerable to COVID-19, as its propagation is carried out through four urban indicator models (experimental, residential, environmental, and topographic). The study concluded that urban planners should consider the environment and health as aspects of sustainable urban planning for cities and neighborhoods.

Deep Learning

Machine and deep learning algorithms, can be considered as a branch of artificial intelligence. It is a field that is based on learning and updating depending on the analysis of mathematical and statistical algorithms. Deep learning algorithms differ from machine learning algorithms in that they require greater computing power and complexity than the latter. Despite this, progress in the field of big data has led to the emergence of larger and deeper networks, enabling algorithms to learn, monitor, and deal with complex data and cases more efficiently and faster than humans. In general, deep learning has been effective in applications of image classification, speech recognition, bioinformatics, etc. [8]. Deep learning is one of the most important tools for data scientists who specialize in collecting, analyzing and interpreting large amounts of data. This is due to its enormous efficiency and speed in the process of processing large data in a way that simulates and even surpasses the human mind. [9]. DL allows computational models consisting of multiple processing layers to learn data representations with multiple levels of abstraction. These methods have greatly improved the latest technologies in speech recognition, visual object recognition, object detection and many other fields such as drug discovery and genetics. DL discovers the complex structure in large data sets using a reverse propagation algorithm to indicate how the machine works [10]. DL methods do not rely on human intervention; it consists of many layers of algorithms that provide a different interpretation of the data they feed on [9].

Long Short-Term Memory

Currently, recurrent neural networks for long-term memory (LSTM) are one of the most important types of deep learning networks. It has been used in challenging areas such as language translation, image annotation and text production. LSTMS dispute significantly from other intricate learning techniques, such as multi-layer cognition (MLPS) and convolutional neural networks (CNN), in that it is specifically designed for sequence prediction problems [11]. LSTM is a neural network. In each LSTM module, there are three gates: the forget gate, the input gate and the output gate. An input gate is an interested to direct the memory of input data and prevent worthless information from incoming the storage unit. The forgetting gate is used to selectively discard the last coin information. The output gate is used to control the output of information each time [12].

Methods

1. Data Set

The data set of (Covid-19) virus cases for the State of Iraq has been used for training the model. It is located on the HDX website. This website provides statistics for Covid-19 cases in Iraq [13]. It is a part of the COVID-19 data of the World Health Organization (WHO), which is the source of this data. The WHO provides Covid-19 data for all parts of the world and is updated daily, and supported by maps and statistics of the number of confirmed and deaths for each country, which are updated weekly according to this data. HDX is partnering with WHO to provide this data. The first update of the data set was on April 29, 2020, and it is updated daily and is available and visible to the public. The Iraq data set consists of six columns (the governorate, the number of confirmed cases, the number of deaths, the number of recovered, active cases, and the date). This research has approved the data file until October 30, 2021.

2. Preprocessing step

It has been assumed that the infection rate in any region follows the pattern of infections spread in the last days in that region. Therefore, the model was trained using the data values recorded for a specific period and considering the value that follows it as the goal or output of this model. The outputs of this model, which represent the values of registered cases for the coming days, are taken as an indicator of the extent of the epidemic in that period. However, after determining the period, which was determined for seven days, the information related to the cases in each governorate was added to the data set. Initially, the data was divided according to the value of a pivot date that can be chosen according to the size of the data and the time period for updating it. In the approved data file, the date of data registration extends from 10/3/2020 to 10/30/2021, so the value of the pivot date that was adopted is 7/8/2021. All data, recorded before this one, are training and validation data are used to train the model. All data recorded after this date are test data to observe the model's performance in determining the locations of the disease's spread in the coming period. The structure of the training and validation dataset was reconstructed through a function that transforms the data for the specified target (number of cases, number of deaths, and number of new cases) into a dataset in which each row is the target registration values for seven days, and the eighth day is the output of the model. When creating this data, the governorates in which the data were recorded are taken into consideration. The data entry was also reconstructed in the form of time series by generating the new deaths and New confirmed columns based on the Deaths and Confirmed columns. The new values are the difference in values between the current days and the previous day. These data were used to train the model to predict the number of new cases.To show predictability of disease prevalence over a specific period, the test data are restructured based on this period and governorates. A new column is generated for each day in this period, meaning that there are 20 columns for the period of 20 days and for each particular governorate. Each column contains the values of registered cases (the number of cases, the number of deaths, and the number of new cases) in that governorate, and the governorates column is added to the file.

3 . Constructing the prediction model

The study methodology mainly consists of five main steps: data preparation, data pre-processing, model constructing, model evaluation, and visualization. A diagram of the proposed model is shown in Fig. 1.

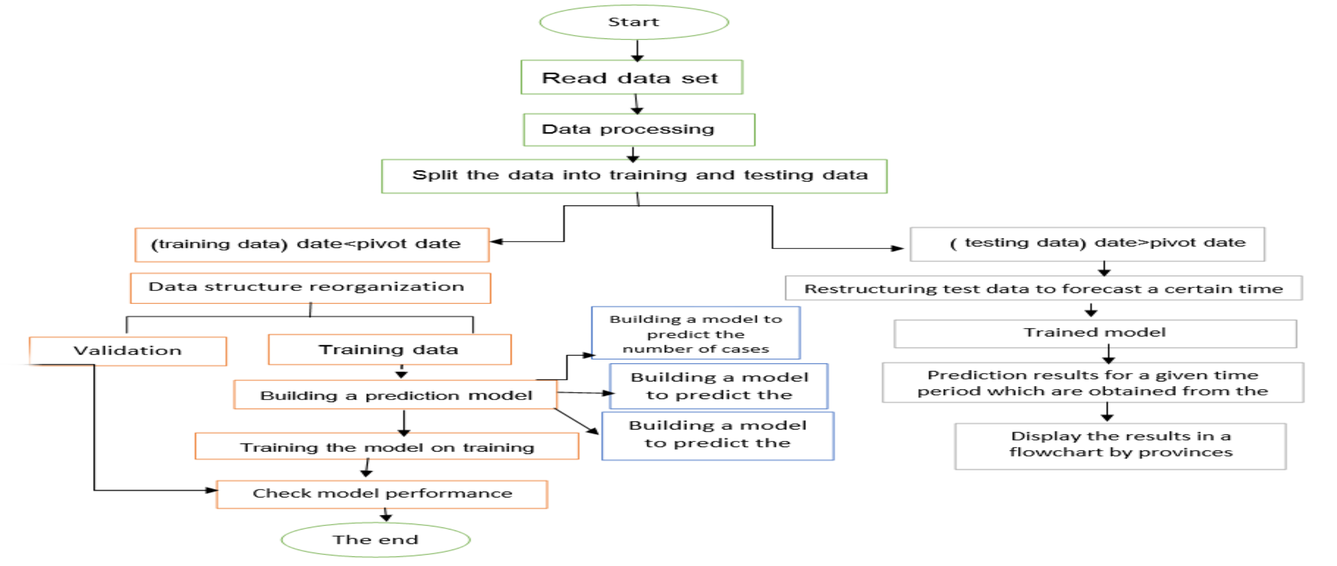

Start
Read data set
Data processing
Split the data into training and testing data
(training data) date<pivot date
(testing data) date>pivot date
Data structure reorganization
Restructuring test data to forecast a certain time
Validation
Training data
Building a model to predict the number of cases
Trained model
Building a prediction model
Building a model to predict the
Prediction results for a given time period which are obtained from the
Building a model to predict the
Display the results in a flowchart by provinces
Training the model on training
Check model performance
The end

Fig. (1) The structure of the Proposed mode

The model was built using more than one layer of the LSTM network. After the data is divided into training and validation data, it is entered into the LSTM model. Models were built for the number of cases, deaths and new cases. Where this model takes the data of the specified days as input and gives a forecast for the next day and continues in the same format and creates the forecast data for the specified period to predict the extent of the disease. For example, two governorates were selected and their data has been applied to the case prediction model for 20 days. The result has been visualized as shown in Fig. 2 for the number of confirmed cases.

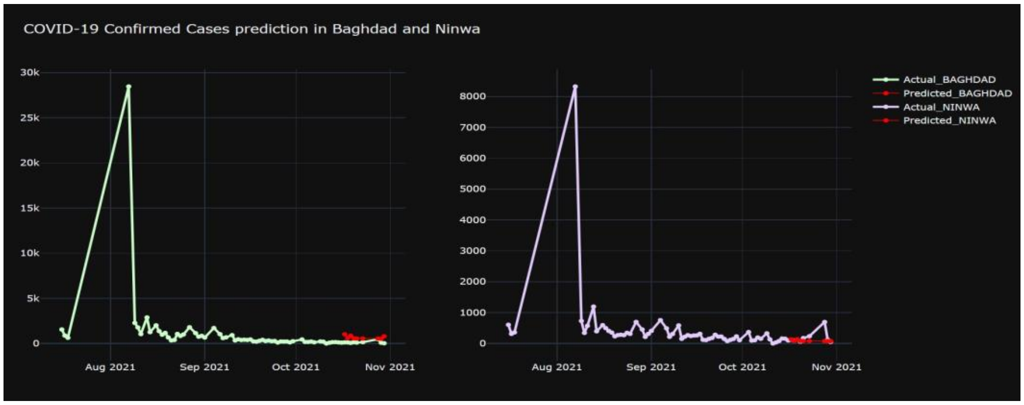


Fig. (2) COVID-19 Confirmed cases prediction in Baghdad and Ninawa

Another example, data for two governorates has been applied to the Death prediction model for 20 days. The result has been visualized as shown in Fig. 3 for the number of deaths cases.

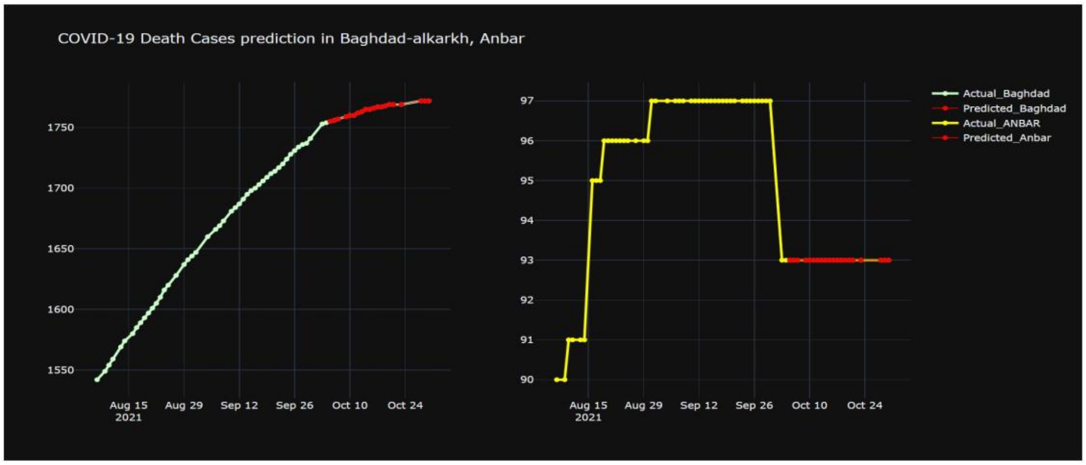


Fig. (3) COVID-19 Death cases prediction in Baghdad and Anbar

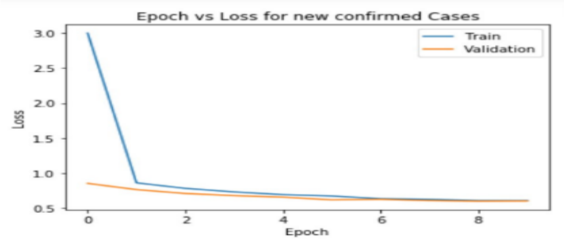
Model Evolution

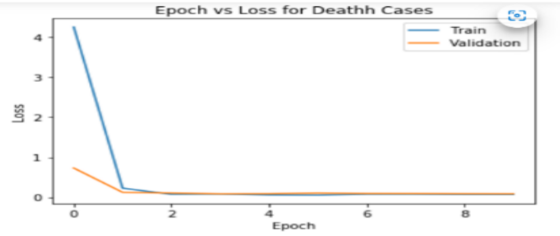
The logarithmic mean squared error (RMSLE) loss function has been used to measure the model's prediction accuracy. The RMSLE function can be defined as a function to calculate the mean squared error between Y_true and Y_pred.

Table 1 shows the value of the loss function for models trained on confirmed cases, deaths, new confirmed cases as targets.

Table 1 Models RMSLE values

Model type	RMSLE values for validation data
LSTM_Confirmed	0.0944
LSTM_Death	0.1569
LSTM_new confirmed	0.6096





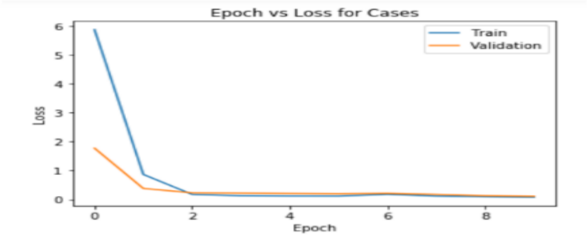


Fig. 4 shows Prediction Models accuracies for train and validation data set

These results show the high accuracy of the proposed models in prediction as shows in Fig. 4

For training and testing models, the Python language version (3.9 and (3.7) was used under the (anaconda) environment.

Conclusion and Recommendation

In this research, an intelligent model has been proposed to predict the risks of spreading Covid-19 disease in a specific Iraqi governorate. Data Set of Covid-19 cases registered in Iraq

was taken from a reliable website affiliated with the WHO. It is updated permanently based on the Iraqi governorates. The data has been processed and restructured to be suitable as input and output values for the model to be trained on. The LSTM deep learning algorithm is adopted for its effective performance in prediction. This is because the data can be treated as a time series to depend on the values recorded on the date field. Since the LSTM network can remember pre and post-values, it was the most suitable for model building Several variables can be changed to obtain different results, including the pivotal date value, which divides the model into training data and test data, as well as the number of days those were adopted to predict the value of the next day. Also, you can change the target field for one of the columns in the data file (number of cases, number of deaths, number of active cases ... and so on). The RMSLE scale was used, and the results of the proposed model had high prediction accuracy, and the network could be trained to obtain higher accuracy After obtaining the results from applying the proposed models to the COVID-19 data for Iraq, several ideas and improvements have emerged that will make the model more reliable and accurate. The following are some of these suggestions that can be worked on in the future to improve work performance:

- The possibility of adding the data of the neighboring governorates for each governorate in the training data set, given that these values recorded for the neighboring governorates have an impact on future values, which may help in obtaining more accurate prediction results.
- It is possible to benefit from the model to predict other diseases if the data are available, also, after making the necessary adjustments according to the data used.

References:

[1] Leppard, Gary G. ”Evaluation of electron microscope techniques for the description of aquatic colloids.” In Environmental particles, pp. 231-289. CRC Press, 2019.

[2] Ahouz, Fatemeh, and Amin Golabpour. "Predicting the incidence of COVID-19 using data mining." BMC public health 21.1 (2021): 1-12.

[3] Widyanto, R. Arri, et al. "Data Mining Predicts the Need for Immunization Vaccines Using the Naive Bayes Method." Journal of Applied Data Sciences 2.3 (2021): 93-101.

[4] Ramchandani, Ankit, Chao Fan, and Ali Mostafavi. "Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions." IEEE Access 8 (2020): 159915-159930.

[5] Roy, Santanu, Gouri Sankar Bhunia, and Pravat Kumar Shit. "Spatial prediction of COVID-19 epidemic using ARIMA techniques in India." Modeling earth systems and environment 7.2 (2021): 1385-1391

.

[6] Ramchandani, Ankit, Chao Fan, and Ali Mostafavi. "Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions." IEEE Access 8 (2020): 159915-159930.

[7] Ramadan, Rasha H., and Mona S. Ramadan. "Prediction of highly vulnerable areas to COVID-19 outbreaks using spatial model: A case study of Cairo Governorate, Egypt." The Egyptian Journal of Remote Sensing and Space Science (2021).

[8] Alakus, Talha Burak, and Ibrahim Turkoglu. "Comparison of deep learning approaches to predict COVID-19 infection." Chaos, Solitons & Fractals 140 (2020): 110120

[9] Alansari, Husain, Oksana Gerwe, and Anjum Razzaque. "Role of Artificial Intelligence During the Covid-19 Era." The Big Data-Driven Digital Economy: Artificial and Computational Intelligence. Springer, Cham, 2021. 157-173

.

[10] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 106-444

.

[11] Brownlee, Jason. Long short-term memory networks with python: develop sequence prediction models with deep learning. Machine Learning Mastery, 2017.

[12] Ghany, Kareem Kamal A., Hossam M. Zawbaa, and Heba M. Sabri. "COVID-19 prediction using LSTM algorithm: GCC case study." Informatics in Medicine Unlocked 23 (2021): 100566.

[13] “Iraq: Coronavirus (COVID-19) Subnational Cases” variety, data.humdata , variety, .org/dataset/iraq-coronavirus-covid-19-subnational-cases.