applied sciences

MDPI

Article

Predicting the Spread of a Pandemic Using Machine Learning: A Case Study of COVID-19 in the UAE

Donthi Sankalpa *, Salam Dhou, Michel Pasquier and Assim Sagahyroon

Computer Science and Engineering Department, American University of Sharjah, Sharjah P.O. Box 26666, United Arab Emirates; sdhou@aus.edu (S.D.); mpasquier@aus.edu (M.P.); asagahyroon@aus.edu (A.S.)
* Correspondence: dsankalpa@aus.edu

Abstract: Pandemics can result in large morbidity and mortality rates that can cause significant adverse effects on the social and economic situations of communities. Monitoring and predicting the spread of pandemics helps the concerned authorities manage the required resources, formulate preventive measures, and control the spread effectively. In the specific case of COVID-19, the UAE (United Arab Emirates) has undertaken many initiatives, such as surveillance and contact tracing by introducing mobile apps such as Al Hosn, containment of spread by limiting the gathering of people, online schooling and remote work, sanitation drives, and closure of public places. The aim of this paper is to predict the trends occurring in pandemic outbreak, with COVID-19 in the UAE being a specific case study to investigate. In this paper, a predictive modeling approach is proposed to predict the future number of cases based on the recorded history, taking into consideration the enforced policies and provided vaccinations. Machine learning models such as LASSO Regression and Exponential Smoothing, and deep learning models such as LSTM, LSTM-AE, and bi-directional LSTM-AE, are utilized. The dataset used is publicly available from the UAE government, Federal Competitiveness and Statistics Centre (FCSC) and consists of several attributes, such as the numbers of confirmed cases, recovered cases, deaths, tests, and vaccinations. An additional categorical attribute is manually added to the dataset describing whether an event has taken place, such as a national holiday or a sanitization drive, to study the effect of such events on the pandemic trends. Experimental results showed that the Univariate LSTM model with an input of a five-day history of Confirmed Cases achieved the best performance with an RMSE of 275.85, surpassing the current state of the art related to the UAE by over 30%. It was also found that the bi-directional LSTMs performed relatively well. The approach proposed in the paper can be applied to monitor similar infectious disease outbreaks and thus contribute to strengthening the authorities' preparedness for future pandemics.

Keywords: COVID-19; UAE; machine learning; deep learning; forecasting; trend analysis

Citation: Sankalpa, D.; Dhou, S.; Pasquier, M.; Sagahyroon, A. Predicting the Spread of a Pandemic Using Machine Learning: A Case Study of COVID-19 in the UAE. *Appl. Sci.* **2024**, *14*, 4022. https://doi.org/10.3390/app14104022

Academic Editor: Aleksander Mendyk

Received: 29 February 2024
Revised: 8 April 2024
Accepted: 24 April 2024
Published: 9 May 2024

1. Introduction

A pandemic is the rapid spread of newly emerging pathogens through human hosts on a global scale. Over the years, the world has witnessed many such pandemics and epidemics, namely SARS-CoV-2, HIV, and AIDS with the most recent being COVID-19 [1]. Coronavirus or COVID-19 is a severely infectious disease that was declared a global pandemic in March 2020 [2]. Since then, the World Health Organization (WHO) has been updating the world every day with the number of cases and has reported over 200 million cases and over 4 million deaths [2]. The virus is said to spread through the air via respiratory droplets and mainly spreads when a person has come in close contact (within six feet) of a COVID-19-positive person when they cough, sneeze, breathe, or talk. The virus can also spread when a person is exposed to small droplets in the air or on surfaces, as they remain for a few minutes to several hours [3].

A pandemic of this nature tends to vary in magnitude over time due to the inherent nature of the virus that is constantly adapting, in addition to the non-negligible effects of seasonal changes and various environmental factors. This general phenomenon is exemplified in the case of the evolution of the initial strain of the novel Coronavirus into the rapidly spreading and arguably more harmful Delta or Omicron variants [4]. As a consequence of the multiple aspects influencing the contagion susceptibility of the Coronavirus and its variant, the relationship between the contributing components and the virus itself tend to be non-linear [5]. This necessitates the utilization of machine learning/deep learning algorithms that can capture these underlying relationships effectively. Statistical techniques make generalized assumptions about the intrinsic behavior of the data (i.e., the daily instances of COVID-19 cases, vaccinations, etc.). These assumptions pertain, more specifically, to the measures of central tendency associated with the periodic, fixed interval measurements of the data samples of interest. For instance, linear correlations in this historical data appear to be incapable of detecting non-spurious patterns of COVID-19-specific factors due to its aforementioned non-linear and potentially complex nature [6]. Machine learning has been proven to work effectively with chaotic, noisy data even without a priori domain knowledge (although, this can impact its performance favorably if available) and hence can be a solution for recognizing patterns of interest in a continuum of equally spaced data measurements such as the daily case counts of COVID-19. To learn more about this phenomenon, this paper focuses on the United Arab Emirates (UAE) as a case study.

The first case in the UAE was reported on 23 January 2020 and case numbers have been rising ever since, with an average of 1000 cases daily during the pandemic days [6,7]. To combat COVID-19, the UAE has taken many initiatives, which are: surveillance and contact tracing, by introduction of apps such as Al Hosn; containment of spread, by limiting the gathering of people; online schooling and remote work; closure of public places; sanitation drives; more frequent testing and more accessibility to testing, by adding more government testing centers and adding extra PCR-test requirements, such as the need to have a negative test result 6 h before flight when traveling from high-risk countries; and fast and organized vaccine drives for all people living in the UAE. Despite the effectiveness of these intensive precautious measures, the country was far from having zero cases per day. To reduce the spread of the disease, the available data must be analyzed in order to be used in forming more accurate future policies and interventions. The UAE government provided such data through the Federal Competitiveness and Statistics Centre (FCSC) which reports daily the cumulative number of cases, deaths, recovered cases, tests, and vaccinations that were recorded on that specific day.

Hence, the goal of this work is to build a model that can forecast and analyze the trend of the spread of a pandemic or contagious disease. COVID-19 in the UAE was considered as a case study. Thus, the main objectives of this paper are defined as follows:

1. Determine whether there is a relationship between the vaccinations and the trend in number of positive cases.
2. Forecast the number of cases per day based on previous history of the number of cases, deaths, and recoveries.
3. Forecast the number of cases per day using the added feature regarding events that can have a major impact on the spread of the disease, such as national holidays, airport closures, etc., to see the event's effect on trends and compare with the actual case trend.
4. Apply the objectives (1–3) considering the COVID-19 pandemic outbreak in the UAE as a case study.

The contribution of this work lies in it being the only work found in the literature that looks into vaccination and events as suitable attributes for finding the best model for forecasting COVID-19.

The rest of the paper is organized as follows. Section 2 presents the literature review. Section 3 discusses the materials and methods utilized in this work. The experimental

results are presented in Section 4 and these results are discussed and interpreted in Section 5. Section 6 concludes the paper.

2. Literature Review

There are several works in the literature that focused on monitoring, analyzing, and forecasting the spread of the pandemic. These works use several methods such as statistical techniques [8,9], agent-based techniques [10,11], and machine learning techniques. In this section, we focus on previous work related to forecasting COVID-19 trends using machine learning. This section is divided into two subsections depending on the methods used. Section 2.1 discusses the works using the traditional machine learning methods and Section 2.2 discusses the works using deep learning methods.

2.1. Forecasting of COVID-19 Cases Using Machine Learning

Several studies in the literature utilized machine learning techniques to forecast COVID-19 trends. Rustam et al. [12] compared four ML models for forecasting of COVID-19. The authors used these models to predict three numerical outputs, namely the number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days. The authors used the Johns Hopkins University dataset. The four ML algorithms used were Linear Regression, LASSO Regression, Support Vector Machine (SVM), and Exponential Smoothing. It was found that Exponential Smoothing performed the best given that the size of the dataset was very small. LASSO worked well only for predicting death rates and confirmed cases. SVM proved to perform the worst of the four techniques.

Similarly, the authors Bhadana et al. [13] compared six ML models for forecasting COVID-19 in India. The authors selected the following features for their dataset: number of cases, announced date of detection, age, detected city, detected district, regional state of detection, gender, and current status. The models, Random Forest (RF), Decision Tree Regression, and Polynomial Linear Regression (LR), were tested. The best performing models were found to be Polynomial LASSO and Polynomial LR.

Gupta et al. [14] selected three features from [15], which are observation date, time, and state/union territory for forecasting. They dropped two features, which were confirmed Indian national and confirmed foreign nations, as these only mattered in the beginning of the pandemic when people were travelling from abroad.

The authors Romadhon et al. [16] performed a comparison of Naïve Bayes method, Logistic Regression, and K-Nearest Neighbors (KNN) model to predict the recovery of COVID-19 in Indonesia. The authors also proposed future work by adding more variables such as travel history and diet to their data for more accurate prediction.

Kumari et al. [17] used Multiple Regression Analysis to predict the number of deaths due to COVID-19 in India. The authors used a dataset from [18] that consisted of state wise number of cases recorded weekly. The input features for their model were the active cases and the recovered cases. Decision Tree and Auto Regression were used. Their results were in agreement with the ground truth.

Another use case of utilizing an exponential smoothing is of Petropoulos et al. [19]. The authors used the Johns Hopkins University dataset to perform a 10-step prediction using a non-seasonal Multiplicative Error and Multiplicative Trend Exponential Smoothing model (ETS(MMN)). It was observed that a large forecast error was found in certain areas due to the global measures taken to control the spread and declines in certain countries, which the model was unable to pick up due to the nature of the data. The signed error does show whether a certain policy worked or not in the past. The authors also conducted experiments country wise and found that they had higher percentage errors.

The authors Leon et al. [20] also evaluated multiple machine learning models to find the best model for predicting COVID-19 infections and deaths in Bangladesh. The authors compared Linear Regression, Polynomial Regression (PR), Support Vector Regression, Auto Regressive Model, Moving Average Model, Holt's Winter Additive model, Auto Regressive Integrated Moving Average Model, and Facebook Prophet's Model (FP).

Table 1 provides a summary of some of the studies that used classical machine learning approaches to predict COVID-19 and gives details on the best results.

Table 1. Summary of Literature Review for Classic Machine Learning approaches.

Reference	Purpose	Dataset	Methods	Best Results
Rustam et al. [12]	Predict number of newly infected cases, number of deaths, number of recoveries for a span of 10 days	Global totals from Johns Hopkins University dataset, 56 days were used	Exponential Smoothing, Linear Regression, LASSO Regression, and SVM	Exponential Smoothing with an R^2 score of 0.98 and RMSE of 16,828.58.
Bhadana et al. [13]	Predict number of newly infected cases, number of deaths, number of recoveries	COVID-19 tracking website of India	LASSO, Random Forest, Decision Tree Regression, Linear Regression, SVM, and polynomial LASSO	LASSO Regression with and average R^2 score of 87.0 for forecasting the future values.
Gupta et al. [14]	Predict number of newly infected cases, number of deaths, number of recoveries	Indian dataset [15]	Decision Tree, Random Forest, Multinomial Logistic Regression, Neural network, and SVM	Confirmed cases had an accuracy of 83.54, death cases had an accuracy of 72.79, and cured cases had an accuracy of 81.27%.
Romadhon et al. [16]	Predict number of newly infected cases	Indonesian local websites	Naïve Bayes, Logistic Regression, and KNN	KNN had the highest accuracy of 75%.
Kumari et al. [17]	Predict number of newly infected cases	Indian dataset [18]	Decision Tree and Auto Regression	Was graphically represented, very similar to real trend.
Petropoulos et al. [19]	Predict number of newly infected cases and deaths	Johns Hopkins University dataset	(ETS(MMN))	Was not able to predict properly as could not pick up on effects of policies implemented.
Leon et al. [20]	Predict number of newly infected cases	Data from Bangladesh Center For Systems Science and Engineering, this included holiday events	Linear Regression, Polynomial Regression (PR), Support Vector Regression, Auto Regressive Model, Moving Average Model, Holt's Winter Additive model, Auto Regressive Integrated Moving Average Model, and Facebook Prophet's Model (FP)	The FP model had the lowest RMSE error of 518.0 for confirmed cases and Holt's had the lowest RMSE error of 13.0 for predicting death cases.

2.2. Forecasting of COVID-19 Cases Using Deep Learning

Several works that use deep learning in forecasting COVID-19 were found in the literature. The authors Zheng et al. [21] proposed an improved susceptible-infected (ISI) model to analyze the development law of COVID-19 in China. Their method used the ratio of the number of newly confirmed cases at a certain time to the cumulative number of new cases over different time scales to calculate the infection rate for their epidemic model. A LSTM network was used to estimate the rate deviation of the epidemic and combine with ISI was used to estimate the number of infected cases. To include the effect of events and government control measures, a pre-trained natural language processing (NLP) model named BERT was used to extract features from relevant news articles of various cities and combined with the LSTM network to correct the deviation of infection rate and further predict a more accurate number of infected cases. The Centre of Disease Control and Prevention (CDC) data was used for this paper. It was found that NLP features provided extra information and guidance for more accurate prediction.

Kumar et al. [22] also proposed the use of an LSTM to predict the spread of COVID-19 in New Zealand using data extracted over a period of 11 months from the Worldometer website. The authors aimed to predict when the virus could be contained. To achieve that aim, the authors calculated a range with an error rate of 1% from either side of the date when it was first predicted as zero cases. To validate their model, they utilized the data for New Zealand, which has already controlled the spread of the pandemic. The model predicted that the curve depicting the number of new cases daily would flatten to zero around 25 April 2020 before rising again and eventually flatlining on 15 May 2020, whereas New Zealand got its first zero cases on 13 May although there were one or two more cases a few days afterwards. For countries where the number of new cases was decreasing, the authors used a time step of one day and predicted the number of cases for the next day; where the number of cases was increasing, the authors used a two-phase procedure where in the first phase, the log of the number of daily cases used a seven-point moving-average filter to determine the peak, and after the peak, the same method of the decreasing number of cases was applied. To avoid overfitting and improve generalization, Bayesian optimization was applied to tune the hyper-parameters of the LSTM network.

Similarly, Chimmula et al. [23] used a LSTM network to forecast COVID-19 transmission in Canada. The authors used the data from Johns Hopkins University and the Canadian Health Authorities that included several features, namely the number of confirmed cases to 31 March 2020, the number of deaths, and the number of recoveries, along with the dates. The authors also inversely trained their network and found that the start of the outbreak in Canada was around early January 2020 but was not actually reported until the last week of January.

Another useful case is from Helli et al. [24], who also proposed using a LSTM for short-term forecasting of COVID-19 cases in Turkey. The authors used data from the Turkish Ministry of Health that consisted of number of confirmed cases from 11 March 2020 to 8 May 2020. It was also found that using an exponential linear unit for activation function worked better as compared to the common choice of hyperbolic tangent.

Ramchandani et al. [25] proposed a deep learning time series model to forecast the range of increase in COVID-19 data in the future days using large numbers of heterogeneous features. The features were grouped into three categories, those being constant feature groups, time-dependent feature groups and cross-county time-dependent feature groups. Each of these feature's groups were then embedded so that it can be fed as input the deep learning model. The authors used the DeepFM model with slight modifications. Their model deemed to be effective in forecasting. It was hard to do a comparison with other models due to the nature of their data. The authors also conducted feature importance evaluation and feature interactions to get a clearer idea on what can be used to make COVID-19 related policies.

The authors Kafieh et al. [26] compared five different models to find the most optimum model for forecasting. The authors first trained their model using countrywide data from Johns Hopkins hospital and then applied the best model to Iran. The models were evaluated based on the recovered cases as an output. The five models, Random Forest (RF), Multilayer Perception (MLP), LSTM, LSTM with regular features, LSTM with extended features, and multivariate LSTM (M-LSTM) were judged and evaluated based on MAPE, RMSE, NRMSE, and R^2 metrics. The input to the models was the number of confirmed, deaths and recovered cases. The authors also varied the lag to find the most optimal lag parameter, which was found to be six days. To consider the effect of actions taken by the government, the authors stopped the training in three different situations; for example, they stopped the training for dates between 27 March and 4 April 2020, due to road closures between cities. This was to demonstrate what would have happened if the above decisions were not taken by the government.

The authors Zain et al. [27] proposed the use of a hybrid CNN-LSTM model to forecast the COVID-19 pandemic. The thought was to combine the advantages of Convolutional Neural Network (CNN) in filtering out noise from the input data and learning the time

series representation using the LSTM model which is effective in identifying and modeling short- and long-term temporal dependencies embedded within a data sequence. The authors used the WHO COVID-19 dashboard data that contained countrywide information such as number of confirmed dead, cumulative confirmed, and cumulative deaths. Their hybrid model was compared against 17 other baseline models, which were divided into two deep learning models, two statistical models, three linear models, five ensemble models, and five machine learning models.

Ghany et al. [28] used a LSTM to predict COVID-19 in the Gulf Cooperation Council (GCC) countries, which includes the UAE. The data were downloaded from the Johns Hopkins dataset for seven months. The inputs were sampled into one-week windows for a seven-time step input and a one-time step output. The data were then fed into a LSTM to generate the output. From their experimentation, they were able to forecast that the number of deaths in the UAE would be under control from the second half of March 2021. For the UAE, the LSTM had a MARE for confirmed cases of 58.50 and a MARE for death cases of 0.63.

Table 2 summarizes the studies that used the deep learning approaches to predict COVID-19 and gives details on the best results.

Table 2. Summary of Literature Review for Deep Learning approaches.

Reference	Purpose	Dataset	Methods	Best Results
Zheng et al. [21]	Predict number of newly infected cases using additional event attributes	Centre of Disease Control and Prevention (CDC)	LSTM + NLP	The MAPE was 0.52% for the predictions in Wuhan and 0.38% in the predictions for Beijing.
Kumar et al. [22]	Predict containment of the virus	11 months of data from the Worldometer website	LSTM	Their model achieved an RMSE of 2 for the New Zealand data.
Chimmula et al. [23]	Predict COVID transmissions in Canada	Johns Hopkins University and the Canadian Health Authorities	LSTM	RMSE error of 34.83 with and accuracy of 93.4%.
Helli et al. [24]	Predict number of newly infected cases	Turkish Ministry of Health	LSTM	MAPE of 0.70.
Ramchandani et al. [25]	Predict number of newly infected cases	Created 3 groups of new features	DeepFM	Accuracy of 63.7%.
Kafieh et al. [26]	Predict number of newly infected cases	Johns Hopkins, applied the best model to Iran	Random Forest (RF), Multilayer Perception (MLP), LSTM, LSTM with regular features, LSTM with extended features and Multivariate LSTM (M-LSTM)	M-LSTM performed best with an MAPE of 0.509%, RMSE of 458.12, NRMSE of 0.001624, and R^2 score of 0.99997.
Zain et al. [27]	Predict number of newly infected cases	WHO COVID-19 dashboard	hybrid CNN-LSTM	MAPE of 01.9, RMSE of 13,275.00, and RRMSE of 5.30.
Ghany et al. [28]	Predict number of newly infected cases and deaths	UAE data from Johns Hopkins dataset	LSTM	MARE for confirmed cases of 58.50 and MARE for death cases of 0.63.

From the literature review, it has been found that LSTM is the most common deep learning model that is used for forecasting of COVID-19 cases and that it performs better than the other machine learning models. Amongst the machine learning models, LASSO Regression and Exponential Smoothing seem to have performed the best. Most of the papers have extracted data for a specific country from the Johns Hopkins database [29],

which is a public dataset that has accumulated all the data from over the world and uses three common features, which are Confirmed Cases, Recovered Cases, and Deaths, and future forecasts for the same.

We also see a lack of studies carried out on the consideration of the effect of the number of vaccinated people on the trends in the spread of COVID-19. Furthermore, only Zhang et al. [21] have used events as an extra feature in training the model for forecasting.

3. Methods and Datasets

This section describes the methodology used for forecasting COVID-19 cases. Section 3.1 presents the machine learning models used for pandemic forecasting. Section 3.2 presents the evaluation metrics. Section 3.3 discusses the model selection. Sections 3.4 and 3.5 present a thorough description of the data used as a part of the case study on the UAE and some preliminary data analysis, respectively. Figure 1 describes the workflow of the study.

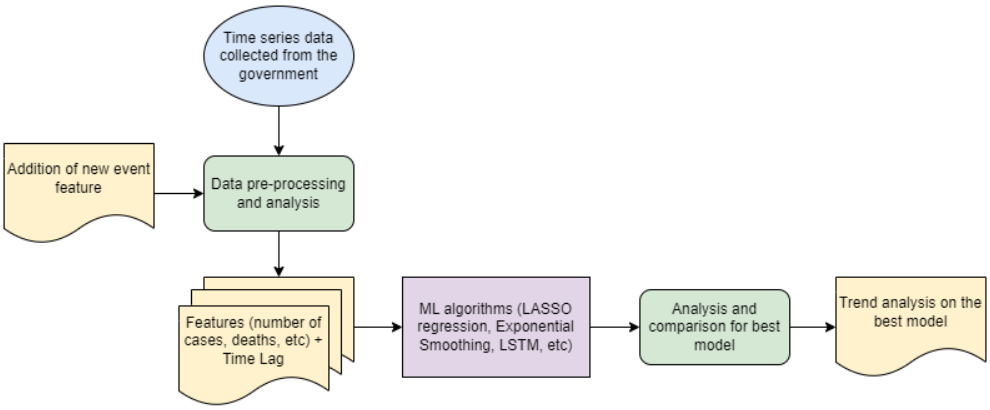


Figure 1. Proposed methodology that demonstrates the flow of getting the data until forecasting and results analysis.

3.1. Machine Learning Models

Several models were trained to predict the future values in this time series problem based on different input conditions. These models were selected from the best performing models in the literature review and were used in addition to the LSTM-AE and bidirectional AE models that were not used by any of the works in the literature review.

The first model used in this work is Least Absolute Shrinkage and Selection Operator (LASSO) Regression. It is a variation of the general linear regression algorithm which tries to express the output as a linear combination of attributes with pre-determined weights [30]. It generates a linear equation by summing the product between attributes and their weights, as seen in Equation (1).

$$y = w_0 + w_1 a_1 + w_2 a_2 + \cdots + w_k a_k, \tag{1}$$

where w_k stands for the weight to be applied on attribute a_k and y stands for the output variable to be predicted using the linear equation. The advantage of this model is that it is suitable for multicollinear data [31]. Unlike multivariate regression, where all the attributes are used in the regression, the LASSO algorithm adds the attributes one at a time, and if the new feature does not improve the fit enough to outweigh the penalty term by including the feature, then it cannot be added. This helps make the model sparse with few coefficients [12].

The second model trained is the Exponential Smoothing model. This model is commonly used in all the work related to the forecasting of COVID-19 using classical machine learning techniques as it is well known to work well with lesser amount of data [32]. Exponential Smoothing is a powerful forecasting model, especially for univariate data.

Univariate data is a type of data that consists of observations only on a single attribute. In this model, the influence of past data observations decays exponentially as it becomes older. Hence, the weight assigned to different lag values is reduced [12]. Equation (2) is used to forecast for current time y_t.

$$y_t = \alpha x_{t-1} + (1 - \alpha) y_{t-1}, \tag{2}$$

where α is the smoothing cost that is $0 <= \alpha <= 1$, x_{t-1} is the actual value of the previous history in the time series and y_{t-1} is the predicted value of the previous forecast [8]. The smoothing cost is similar to the weights used in deep learning models.

Another model used is a classic LSTM. LSTM is a deep learning model that is commonly used for forecasting in time series. LSTM networks works with memory blocks that were created to solve vanishing gradients by memorizing network parameters for long durations [23]. A vanishing gradient simply means that the model is unable to change the weights applied to the nodes. The memory blocks are like the differential storage systems of a digital systems gate. The information is processed within these gates with the help of an activation sigmoid function and gives an output of 0 or 1. Sigmoid is used, as we require that only positive values to be passed on [23]. They are generally built with three gates being input gate, forget gate, and output gate. The input gate gives information that needs to be stored in the cell state. The forget gate throws information based on the forget-gate activation equation, and finally, the output gate combines information from the cell state and forget state at a certain time step t to generate the output. The gradients and weights are shared throughout for long durations and, by adjusting them, we can adjust the time scale to detect the dynamically changing parameters [33] and hence avoid the vanishing gradient factor. The following equations [33] illustrate the three gates.

$$J_t = \text{sigmoid}\big(w_J[h_{t-1}, k_t] + b_J\big), \tag{3}$$

$$G_t = \text{sigmoid}(w_G[h_{t-1}, k_t] + b_G), \tag{4}$$

$$P_t = \text{sigmoid}(w_P[h_{t-1}, k_t] + b_P). \tag{5}$$

Here, J_t is the function for the input gate, G_t is the function for the forget gate, and P_t is the function for the output gate. w_x is the coefficient of neurons at gate (x) and b_x is the bias of the neurons at gate (x). Finally, k_t is the input to the current function at time step t and h_{t-1} the result from the previous time step.

LSTMs are advantageous for the COVID-19 dataset as they can capture the non-linearity nature of the data and can result in state-of-the-art results on time-series-related data [13]. The actual architecture of the network used for this work follows the architecture of [28]. This is illustrated in the following Figure 2.

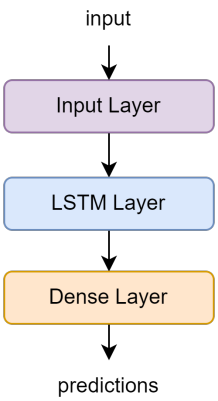


Figure 2. Architecture of the simple LSTM.

Finally, different variations of the LSTM Autoencoder are used. An LSTM Autoencoder, or LSTM-AE, merges the use of LSTM layers with a traditional Autoencoder architecture. This type of architecture is trained to copy the input to the output. The architecture is divided into two sections, an encoder and decoder. The encoder encodes the data into a lower dimension latent encoding while the decoder decodes the latent representation back into the original context [34].

A bi-directional LSTM adds to this concept by changing the LSTM layers to bi-directional LSTM layers. Bi-directional LSTM layers address the limitation of the standard unidirectional LSTM architectures in terms of being restricted by the previous context only [35]. This architecture is able to process data in both forward and backward directions, consequently focusing on both the sequence as well as the breakdown of the sequence. The forward layer consists of T cells that are denoted by h_t^f while the backward layer consists of the same T cells that are denoted by h_t^b. The forward layer processes the inputs in the manner of $[t_0, t_1, \ldots t_T]$, while the backward layer processes in the opposite direction $[t_T, t_{T-1}, \ldots t_0]$. The result of combining the outputs of both layers results in a vector which is computed by the following Equation (6):

$$\begin{aligned} h_t^f &= \tan h\left(W_{xh}^f x_t + W_{hh}^f h_{t-1} + b_h^f\right), \\ h_t^b &= \tan h\left(W_{xh}^b x_t + W_{hh}^b h_{t+1} + b_h^b\right), \\ y_t &= \left(W_{hy}^f h_t^f + W_{hy}^b h_t^b + b_y\right). \end{aligned} \tag{6}$$

This theory is applied onto the architecture for this work and is illustrated in the following architecture diagrams seen in Figure 3.

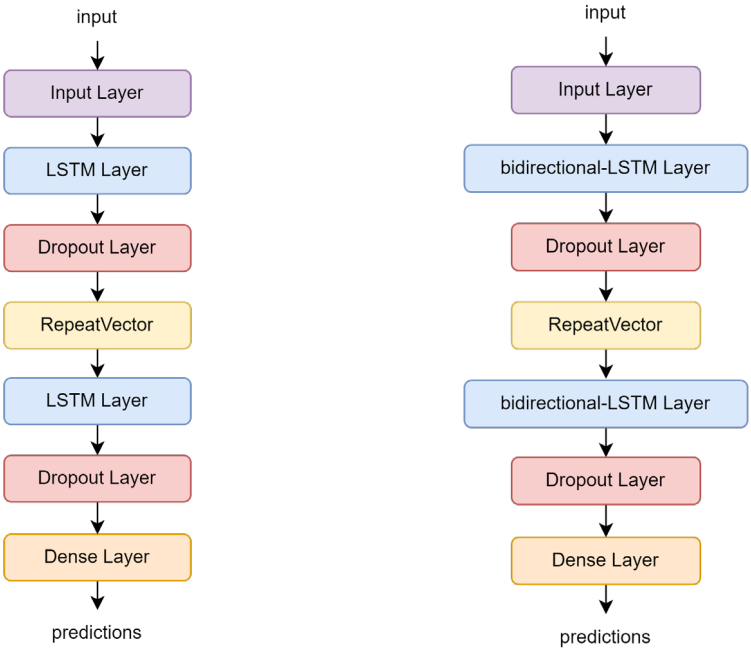


Figure 3. (**a**) architecture of the unidirectional LSTM-AE (**b**) architecture of the bidirectional LSTM-AE.

3.2. Evaluation Metrics

Different evaluation metrics were used to evaluate the performance of the machine learning models applied in this work. The metrics used in this work are the coefficient of determination (R^2) score, mean squared error (MSE), and root mean squared error (RMSE).

R^2 score is a statistical performance metric that can be used to determine the proportion of variance in an output variable that can be explained or predicted by an input variable in

a regression model. It can be used to measure how well a regression model predicts the output variable. The following equation formulates the R^2 score [36].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \quad (7)$$

where $\hat{y}_i$ represents the predicted value, y_i is the actual value, and $\bar{y}_i$ is the mean of all the actual values. An R^2 value ranges between 0 and 1; where 0 means that the model explains or predicts 0% of the relationship between the output and input variables, a value of 1 indicates that the model predicts 100% of the relationship, a value of 0.5 indicates that the model predicts 50% of the relationship, and so on.

MSE is a commonly used metric to measure the error of a model. The objective of the training process in numerical predictions is to minimize the *MSE*, as the smaller the *MSE*, the lower the error; therefore, the better the model. It is calculated using the following equation [37].

$$MSE = \frac{1}{n} * \sum(y_i - \hat{y}_i), \quad (8)$$

where $\hat{y}_i$ represents the predicted values while y_i are the actual values.

RMSE is calculated using the following equation, where $\hat{y}_i$ represents the predicted values and y_i are actual values. It is a good measure to estimate the standard deviation of an observed value from the actual value, thereby allowing insights into the efficacy of the model being evaluated.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}. \quad (9)$$

3.3. Model Tuning and Selection

This section provides details on how the data used for training the models was split, the usage of hyperparameter tuning to get the most performant LSTM model, and the use of feature normalization.

For the purpose of training and testing the models, the data was split into 80% training and 20% testing. For the LSTM models, out of the 80% kept aside for training, 20% was used for validation to tune weights via the loss functions. Another important parameter is the time window over past history, also known as the lag time, which is used for both training and testing the model. The optimal time window was found by doing trial and error experimentation. Four different values of lag were tested with window sizes of 1, 5, 7, and 14. The results of this analysis are shown and elaborated on in Section 4. The reason 5 was selected is because the number of working days in the week is 5 and, even with an overlap of the lag window, at least more than 1 day of the working week will be covered. Next, 7 and 14 were chosen to cover the full week including the weekend days and 2 consecutive weeks.

To find the most optimal LSTM model, hyperparameter tuning was used to select the values of certain parameters such as learning rate, number of layers within the model, and loss functions. Table 3 summarizes the specifications that were found to result in the most optimal model.

Table 3. Hyperparameter specifications.

Hyperparameter	Value
Learning Rate	1×10^{-3}
Optimizer	Adam
Loss Function	MAE
No. of Epochs	1×10^{3}

Normalization is generally a vital step, for almost all algorithms, as it can lead to faster convergence towards the minima, because the feature values are on a similar scale. In this work, Min-Max scaling is applied to the values such that the resulting range of values lies between 0 and 1 as follows:

$$x' = (x - x_{min})/(x_{max} - x_{min}), \quad (10)$$

where x' is the normalized value, x is the original value of the attribute, x_{max} is the biggest value in the column of the attribute, and x_{min} is the smallest.

3.4. The UAE COVID-19 Dataset

The UAE is home to approximately 9.28 million people and is divided into 7 Emirates, namely, Abu Dhabi, Ajman, Dubai, Fujairah, Ras Al Khaimah, Sharjah and Umm Al Quwain. Abu Dhabi is the largest emirate while Dubai is the most populated. The dataset used in this paper consists of data in time series format. The dataset is updated every day by the FCSC and provided through the official UAE COVID-19 updates website [7]. The data reported represent all values across the UAE and are not divided by Emirate. The dataset was taken over a duration of round 27 months, between 1 of February 2020 and 29 April 2022, resulting in 791 records (1 record per day). Each record consists of 10 attributes as follows:

- Day: an index of number of days since the first record of the virus.
- Date: the date of when the data were recorded and added to the dataset.
- Confirmed Cases: the number of positive cases recorded on the specified date.
- Recovered Cases: the number of recovered cases recorded on the specified date.
- Confirmed Deaths: the number of deaths caused by the virus, recorded on the specified date.
- Tests: the number of PCR tests conducted on the specified date.
- Active Cases: a case is considered active if there is a possibility of the virus being active in the person's body. For instance, an active case can happen when a person shows symptoms but gets a negative PCR test, or when a person comes into contact with another person who has been diagnosed with COVID-19.
- Vaccine Doses: the number of vaccine doses administered on the specific date. This number does not differentiate between people who are having their first, second, or third dose.
- One Dose: the percentage of the population of the UAE that has had at least one single dose of any of the available vaccines. This record is cumulative, as each day adds to the previous day.
- Fully Vaccinated: the percentage of the population of the UAE that has had both doses of the vaccine. This record is cumulative, as each day adds to the previous day.

3.5. Dataset Preprocessing

The dataset was inspected and several data-preprocessing techniques were applied prior to the analysis. For example, it was noticed that the Date and Day attributes both indicate the same thing, with the only difference being that the former is an absolute value while the latter is a relative value. To avoid redundant information in the dataset, the Date attribute was excluded from the analysis. Additionally, the attributes One Dose and Fully Vaccinated were dropped because they are just cumulative percentages that show an overall picture of the vaccination drive conducted in the UAE but do not show daily information that could affect the trends directly.

For the event-based prediction, three additional features are manually added to the dataset, namely Outdoor Events, Policies, and National Holidays. These three attributes are of type categorical Boolean, with a value of 1 when there was an event e.g., when it was a national holiday, when a new policy was put in place, when a major outdoor event was

held involving crowds, etc., and a value of 0 when nothing happened on the specified date. This data was manually retrieved by traversing through the nation calendar of the UAE.

It was found that the attributes Confirmed Cases, Recovered Cases, and Confirmed Deaths had a Minimum of 0. The zero value recorded as the number of cases is irrelevant as it was recorded in the beginning of February in 2020, when there were only 1 or 2 cases per day and the virus was still only contained within China. It was also noticed that the attributes related to the vaccinations had some missing values in the beginning of the pandemic because there was no vaccination present at that time. Thus, these missing values were replaced with 0.

Figure 4 shows a graphical representation of the time series dataset used in this paper. The graphs are generated for the period from the 1 of February 2020 to the 29 of April 2022. From these graphs we can see that the Recoveries and Confirmed Cases are almost similar in magnitude but vary in the number of peaks and falls. The similarity in magnitude is expected as within the UAE, the recovery rate is extremely higher than the death rate, implying that majority of the people have recovered. We can see in the initial stages of the virus, although the number of Confirmed Cases was low, the number of Recovered Cases was high. This can signify that the Recovered Cases were most likely the initial cases where the effect of the virus was stronger on people due to lack of knowledge and precautionary measures.

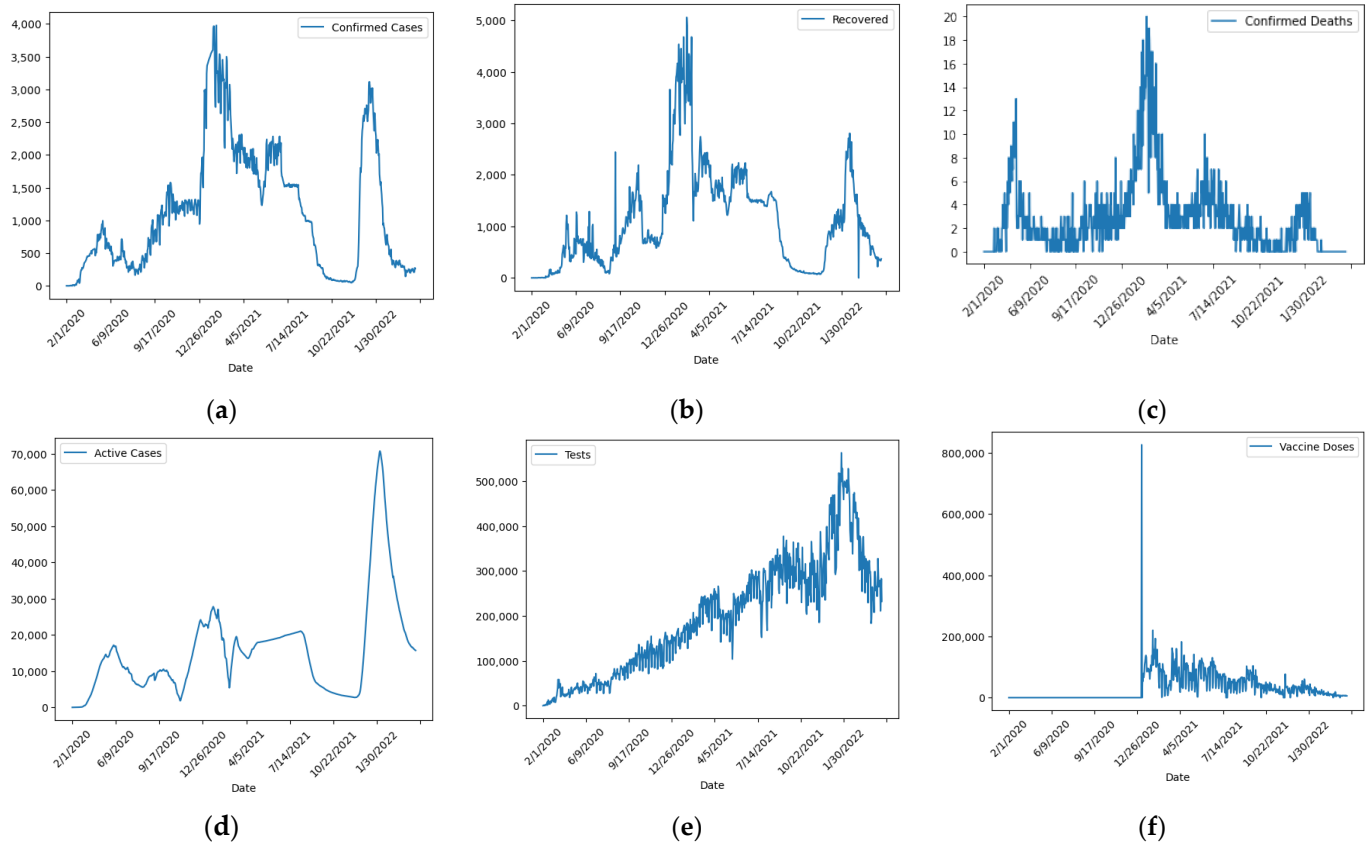


Figure 4. Graphical representation of the time series. The outbreak in the UAE started in February 2020 as can be seen in Figures (**a**–**e**), while in (**f**), the vaccination stated in December 2020.

Additionally, we can see that the number of tests conducted are gradually increasing except for the sudden spike in between of October 2021 and January 2022. This can be accounted for by the introduction of the Al Hosn green-pass mobile application, which required people to take a PCR test every 14 days to maintain their green status. There is also an interesting phenomenon where the number of vaccine doses were initially very high but then started to downfall and have swindled down to zero doses given per day. This

can be seen as the UAE has a very high vaccination rate where over 99% of the population are already fully vaccinated [38].

3.6. Data Exploration

Before the data can be fed into the machine learning models, certain analysis needs to be done so as to understand the behavior of the time series. Time series data can have trends only if a certain pattern in the data repeats itself on regular intervals of time due to any external factors and is not random; therefore, we need to test the quality of the given dataset such that it consists of a trend in which we can analyze. For this, autocorrelation plots are derived to determine whether the elements of the time series are random or not. Autocorrelation determines the similarity between a certain data point and its past within a time interval called lag. As seen in Figure 5, the horizontal axis shows the lag between two elements, so, for example, a lag of 2 means values observed two time periods earlier than the current observation. The vertical axis shows the results of the autocorrelation function and ranges from -1 to 1, where a value from -1 to 0 is a negative correlation and a value from 0 to 1 is a positive relation [39]. Any value closer to $1/-1$ means a higher level of correlation. Figure 5 shows the autocorrelation plots for each of the attributes. The autocorrelation with lag 0 will always be 1 as it is the autocorrelation between itself and can be ignored. As all the spikes are above zero, they are statistically significant which means that all the values are highly correlated and signifies that when the values of the trend are rising, it will continue to rise with time and when the values of the trend are falling, they will continue to fall. Hence, the time series can be considered as not random but following a significant trend.

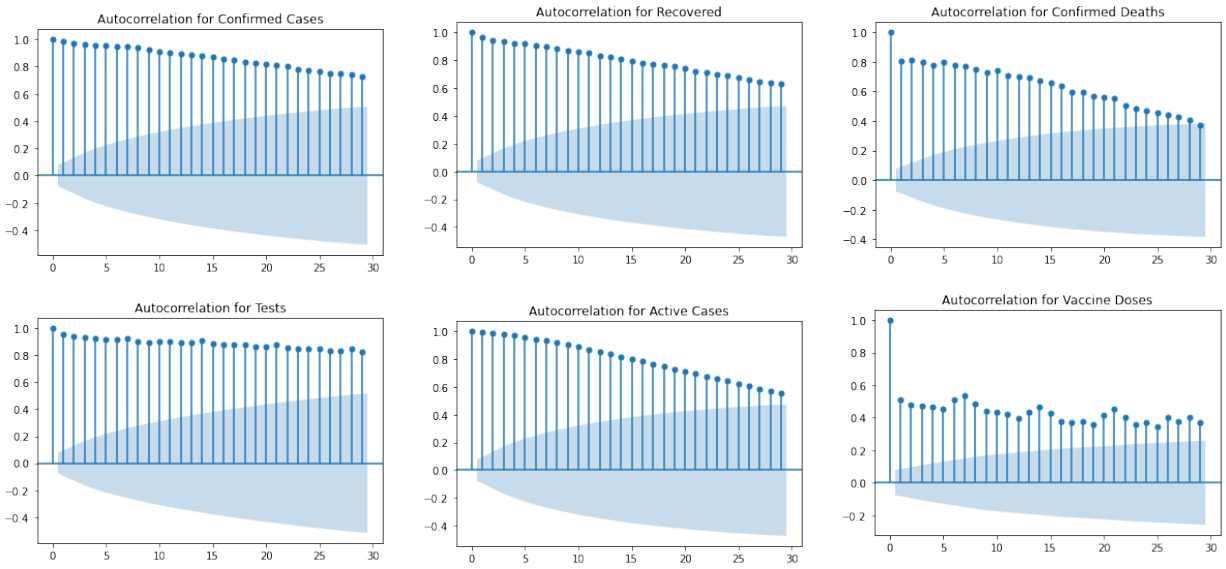


Figure 5. Autocorrelation graph for each of the attributes. The X-axis represents the different lag periods while the Y-axis represents the autocorrelation.

Time series are divided into two types, stationary and non-stationary. Stationary series are those series that do not depend on time components such as seasonality effects and trend, while non-stationary series are series that change trend with respect to time and seasonality effects. Although, from domain knowledge, it is well-known that a pandemic such as COVID-19 changes with time and external environment, and hence will be a non-stationary series, it is still important to verify this hypothesis. For this purpose, the Augmented Dickey Fuller test [40] was conducted. It is also important to check as statistical models can only use stationary series and, hence, if the series is non-stationary, it is better to use machine learning models. If the *p*-value obtained is between 5 and 1%, the null hypothesis is rejected and does not have a unit root and is hence a stationary series. If the

p-value obtained is greater than 5% or 0.05, the input data has a unit root and is regarded as a non-stationary series [23].

Table 4 shows the *p*-values for each of the time series. It is noticed that all attributes except Confirmed Deaths show much higher *p*-values than the threshold. The statistical significance for Confirmed Deaths is on the edge of the hypothesis threshold, which is why it is still considered as a potential confounding variable and introduced into the studies. Although there is a change in the values of the attribute based on external factors, the magnitude of the relation change is less. This explains why the death rate in the UAE with regard to COVID-19 is very low.

Table 4. *p*-values of each of the time series and Pearson Correlation between Confirmed Cases and other attributes.

Attributes	*p*-Value	Pearson Correlation Coefficient(r) with Attribute Confirmed Cases
Confirmed Cases	4.20×10^{-1}	1.00×10^{0}
Recovered	3.60×10^{-1}	9.20×10^{-1}
Confirmed Deaths	5.00×10^{-2}	6.70×10^{-1}
Tests	6.70×10^{-1}	7.50×10^{-1}
Active Cases	1.10×10^{-1}	6.90×10^{-1}
Vaccine Doses	1.20×10^{-1}	6.20×10^{-1}

Moreover, Pearson correlation [41] was conducted to explore correlations between the dataset attributes. The Pearson product-moment correlation coefficient (r) measures the strength of a linear association between two variables. The value of r ranges from -1 to +1, with values towards +1 having stronger correlation, values of 0 having no correlation at all, and values towards -1 showing a strong negative correlation, i.e., if one variable increases the other decreases [42]. Table 4 explores the correlations between the attribute Confirmed Cases and other attributes in the dataset.

Figure 6 shows the heat map derived using the correlations between all the attributes in the dataset. A heat map is a two-dimensional representation of data where the values are represented in color. As seen in the color scale, as the values go towards darker green, that means that r is moving towards 1. The diagonal in the heat map can be ignored as the correlation between a variable against itself will always be 1. We can additionally see that most of the attributes are correlated with the Confirmed Cases and Recovered Cases having the highest correlation. A priori, we know that the greater number of tests conducted means that there is a higher chance of finding positive cases, and this is proved by the high correlation between number of Confirmed Cases and Tests. The coefficient values between 0 and 0.5 or 0 and –0.5 indicate weak correlation, diminishing to mean almost no direct correlation at 0. Table 4 also shows the actual Pearson value for correlation between the number of Confirmed Cases and the other attributes.

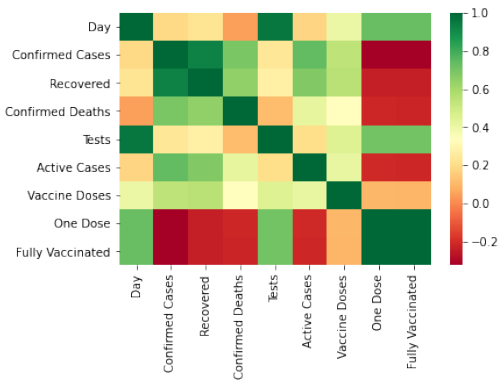


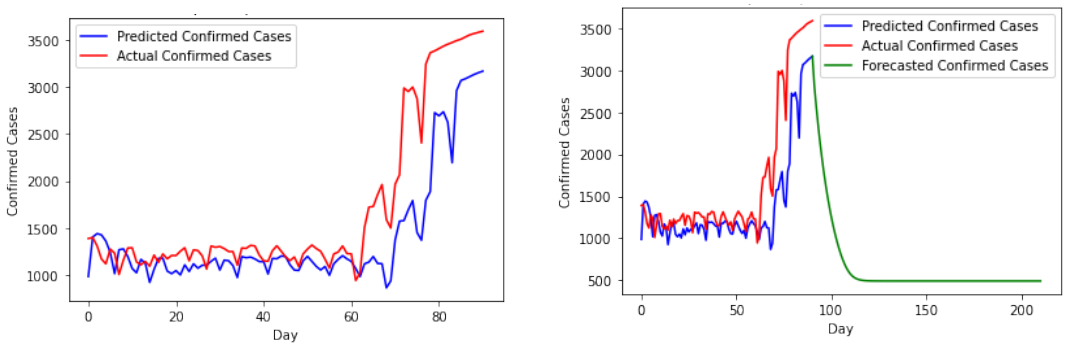
Figure 6. Heat map derived from Pearson correlation.

4. Results

First, to understand the data and the nature of the series, the LSTM model from the work by [28] was recreated, since the authors used the same dataset. It is also known a priori that all the countries were affected at different levels at different times with the virus, and hence would have different training methods and results. At the time of publishing, the authors were limited to barely one year of data, from 14 January 2020 to 9 February 2021. To keep all the characteristics the same and enable the same grounds of comparison, the data was limited to this time period. Following the paper, a single attribute Confirmed Cases was used, keeping the consistent lag period of seven, after which the value of the same variable is forecasted/predicted. This baseline is initially considered for this period only, with the primary intention of ensuring that the findings are indeed similar. The subsequent experiments extend this baseline by means of additional predictor variables (the earlier works uses only Confirmed Cases), in addition to the consideration of a much longer time frame, i.e., 791 days (ours) versus 333 (theirs). Table 5 reports the results, followed by Figure 7, which shows a visual representation of the testing.

Table 5. Results for the Baseline Model.

Metric	[28]	Training	Testing
RMSE	4.46×10^2	1.50×10^2	4.95×10^2
MRAE	5.85×10^1	12.39×10^7	1.38×10^1
MSE	Not reported	22.57×10^3	24.47×10^4
R^2	Not reported	7.80×10^{-1}	6.60×10^{-1}



(**a**) (**b**)

Figure 7. (**a**) Prediction using the original LSTM on the testing set (**b**) Predicting beyond the test set using the original LSTM.

Regarding the terminology in Figure 8, it is worth mentioning that we do not use the expressions "predicted" and "forecast" interchangeably, despite their very similar semantic connotations. The Keyword "predicted" is used for new number of Confirmed Cases predicted for windows of time from the testing set while the expression "forecast" is used for new number of Confirmed Cases using the "predicted" number of Confirmed Cases.

Observing further, it appears that the model tends to systematically underestimate the actual values, and a potential reason may be the confounding effect of trend, i.e., the average level of a series. This trend is likely being caused by the presence of a variable that we have not factored in or unknown to us, such as public policy, or the most obvious reason: insufficiency of data. With the relatively high parameterization of deep learning networks with low amounts of data, the true patterns are not detected effectively. As seen in (c), the model forecasts a decline in COVID-19 cases and then starts to flatline and continue at this single value (500). This green line is produced using only newly forecasted data, in simpler terms, it is data that the model has never seen before. Hence, we can say that the model is unable to account for any patterns of seasonality and hence converges the values

and predicts a flatline value, which is just the previous value as it stops considering all the values before the immediately last time step due to insufficient data. Here, seasonality refers to the periodic (cyclical) patterns that occur in the curve as described in Section 1 previously. This "last " time-step, somewhere ~110th day outputs a prediction of 500, and this is propagated to the next step, and so on such that only 500 is predicted by the collapsed learning behavior of the model.

Figure 8. Prediction using Exponential Smoothing model.

Table 6 shows the extensive results of the experiments conducted with various combinations of attributes. The combinations are mainly grouped into four cases, them being as follows:

- Case 1: The input is the previous history of Confirmed Cases while the output is the prediction of the number of Confirmed Cases for the next day.
- Case 2: Attributes Recovered Cases and Death are additionally added to the input list while the output remains the same.
- Case 3: The attribute Vaccinations is added to the input list of Case 2 while the output remains the same.
- Case 4: The attributes related to the Event feature as explained in Section 3 are added to the input list of Case 2 and the output remains the same.

Exponential Smoothing has a major limitation, which is that the model can only be used for univariate series. This means that the series should depend only on a single variable. Although it was known that the model would not work well, it was selected to establish a baseline performance, and while we expect LSTMs to fare better based on previous research, a point of comparison is needed with standard models. To test this model, we followed the authors of [28] and selected Confirmed Cases as the only input for the model. The output is the number of Confirmed Cases predicted for the next day. Alpha was varied with values of 0.8, 0.2, and 0.5. Alpha, in Exponential Smoothing, defines the weighting level of the smoothness factor. In Figure 8, we can see that the smoothing does not greatly affect the predictions and that all three models overfit at first before tending to flatline at a singular value. This result is expected as Exponential Smoothing is mostly used for data that do not depend highly on trends and seasonality, whereas COVID-19 high depends on the environment and surroundings and, hence, depends on trends and seasonality.

LASSO Regression was tested on all the cases, as it works on both univariate and multivariate series. For this model, the time lag was set to seven days, following the logic of the previous works discussed in the literature review that seven days would cover the data accurately as it would most likely include weekdays as well as weekends.

Table 6. Results of all the experimentation conducted.

Model			Exponential Smoothing	LASSO	LSTM	LSTM-AE	Bi-Directional LSTM
Univariate (Confirmed Cases as input)	Train	R^2	-	-	9.10×10^{-1}	5.00×10^{-2}	7.10×10^{-1}
		RMSE	-	-	2.68×10^{2}	8.82×10^{2}	4.84×10^{2}
	Test	R^2	7.00×10^{-2}	-	9.10×10^{-1}	4.50×10^{-1}	8.60×10^{-1}
		RMSE	-	-	2.76×10^{2}	6.80×10^{2}	3.33×10^{2}
Multivariate (Added Recovered Cases and Death to input)	Train	R^2	-	-	9.10×10^{-1}	2.00×10^{-2}	6.50×10^{-1}
		RMSE	-	-	2.65×10^{2}	9.00×10^{2}	5.31×10^{2}
	Test	R^2	-	9.80×10^{-1}	8.80×10^{-1}	3.50×10^{-1}	5.70×10^{-1}
		RMSE	-	-	3.15×10^{2}	7.40×10^{2}	5.99×10^{2}
Multivariate (Added Vaccination Attribute to input)	Train	R^2	-	-	9.10×10^{-1}	3.00×10^{-2}	6.80×10^{-1}
		RMSE	-	-	2.61×10^{2}	8.92×10^{2}	5.11×10^{2}
	Test	R^2	-	9.80×10^{-1}	8.60×10^{-1}	6.00×10^{-2}	1.73×10^{0}
		RMSE	-	-	3.43×10^{2}	9.44×10^{2}	1.52×10^{3}
Multivariate (Added Events to the input)	Train	R^2	-	-	9.10×10^{-1}	2.00×10^{-2}	7.10×10^{-1}
		RMSE	-	-	2.70×10^{2}	9.01×10^{2}	4.89×10^{2}
	Test	R^2	-	9.80×10^{-1}	8.50×10^{-1}	1.73×10^{0}	6.10×10^{-1}
		RMSE	-	-	3.47×10^{2}	1.52×10^{3}	5.71×10^{2}

In Figure 9, similar to the previous model, the model has a tendency to overfit. We also see that there is not a discernible improvement with the addition of the extra variables. LASSO Regression works by adding an attribute only if it is useful to the model. More specifically, the L1 regularization component of LASSO confers a coefficient of zero to features that were completely neglected by the model to make predictions. From this, we can deduce that the addition of the vaccination, policies, and outdoor events might not be providing any additional information to the trends, at least in the case of the UAE, and hence may likely be unnecessary for the forecasting of COVID-19.

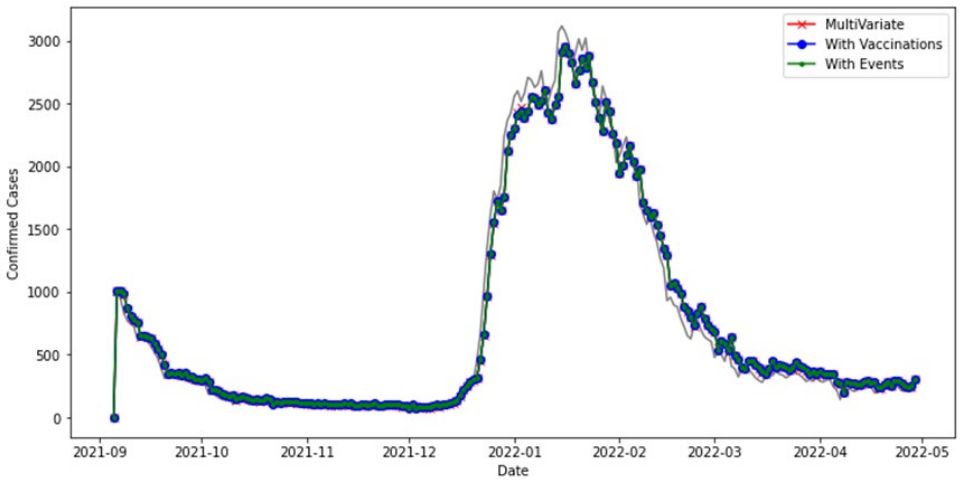


Figure 9. Prediction using the LASSO Regression model on the testing set.

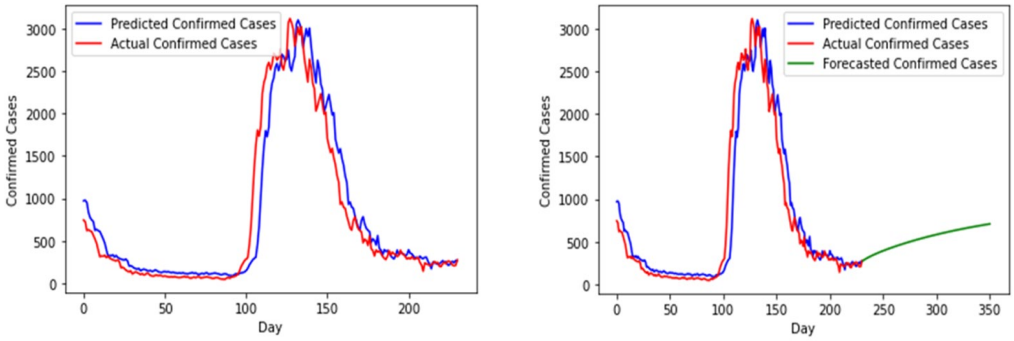
Table 7 shows the MSE and R^2 Score resulting from testing the LSTM using the different lag periods, as mentioned in Section 3. From this, a lag of five days was chosen. Although

Day 1 has a better score, with just one day of history, it can easily overfit the model and, hence, the next best value of a lag of five days was selected.

Table 7. Results for the Baseline Model.

Lag Period	MSE	R^2 Score
1 day	1.53×10^4	9.80×10^{-1}
5 days	7.81×10^4	9.00×10^{-1}
7 days	1.30×10^5	8.40×10^{-1}
14 days	4.27×10^5	5.10×10^{-1}

We can already see a huge improvement on the model using the univariate trend as compared to [28]. This is illustrated in Figure 10. We can also see that the prediction line has started to not flatline and to forecast some values. This shows that the original model was not trained with enough data. Although this model is not perfect either, we can see that more data do help improve it drastically, and in the future, the model may be perfected. This also proves that LSTMs highly depend on trends and seasonality and enough of these factors should be captured in the training set.



(**a**) (**b**)

Figure 10. Prediction of cases based on the univariate trend using the LSTM on (**a**) testing set (**b**) Predicting beyond the test set.

The following Figure 11 illustrates the way the LSTM worked on the other variations of the dataset. From this, we could infer that the model works moderately well with the addition of the Confirmed Cases, Deaths, and Recoveries attributes, except for a few areas marked on the Figure where the model gets the values slightly wrong (b), although one would think that the addition of the Recovered Cases and Deaths attributes should help improve the model as they correlated to each other, as seen in Figure 3. This could be because in the UAE dataset, deaths are almost constant, and this might cause a confusion to the trend. With the addition of the Vaccination attribute, the effect of the Recovered Cases and Deaths attributes on the final prediction is reduced but is still not as good as the univariate case (d). This can show that there might be a slight relation between the Vaccination attribute and the others, but the actual values may not be truly helpful. Finally, the case with the additional event-related variables is tested. Unlike the previous Vaccinations case, the model is slightly more stable with an R^2 score of 0.85, as seen in Table 7. This proves that the additional variables help the model learn the trend. Although it does work well, there is still room for improvement. This could be because the variables are simply 1 or 0 for whether an event was there or not on the day. Not enough information about the effect of the events is transferable through these simple variables.

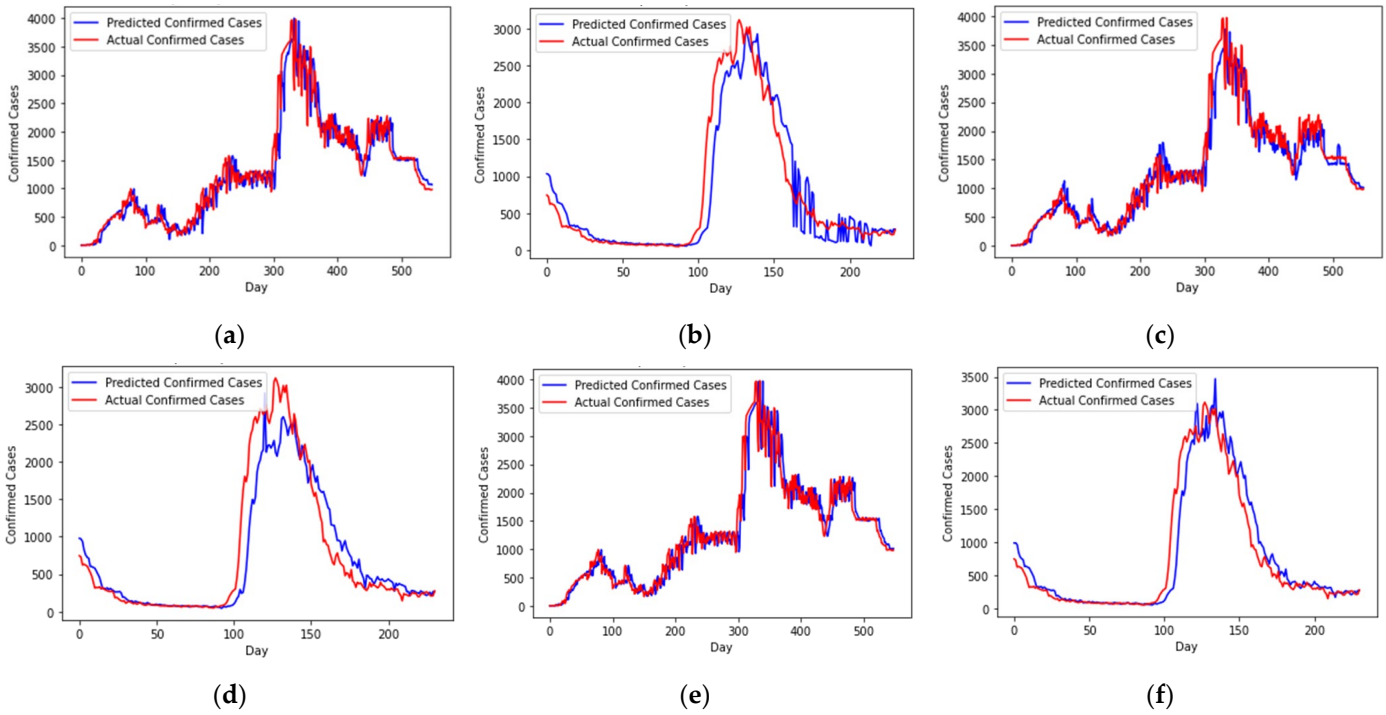


Figure 11. (**a**) Prediction using the LSTM on the training set using multivariate data (**b**) Prediction using the LSTM on the testing set using multivariate data (**c**) Prediction using the LSTM on the training set using vaccination data (**d**) Prediction using the LSTM on the testing set using vaccination data (**e**) Prediction using the LSTM on the training set using event data (**f**) Prediction using the LSTM on the testing set using event data.

The final models tested were the LSTM-AE and the Bidirectional LSTM-AE. Like with the LSTM, first the lag period was tested, and the lag window of 5 performed the best. In the case of the Univariate trends, we can see that the bidirectional LSTM-AE outperforms the classic LSTM-AE, yet the classic LSTM performs the best. This could be because the data is very simple and small and hence does not require a heavy and complex architecture for the prediction. It can be further observed through the training data predictions that the model gets stuck in a local optimum after a while and does not learn very well. However, we see that the bidirectional model does perform better than the unidirectional, as expected, because of the focus on both sequence and breakdown of the sequence.

Overall, we notice that the bidirectional LSTM-AE works better than the classic LSTM-AE. Figure 12 shows the training curve and testing curve for the remaining cases. From this figure we deduce that, unlike with classic LSTMs, there is a huge drop in performance with the addition of the Death and Recovered Cases attributes. When the Vaccination attribute is added, we notice a major problem; the model can predict the training set, but when it comes to an un unseen range, it starts to flatline, as observed previously. Here, too, it led to the model stopping predictions at around the value of 2000 cases when predicting the testing cases. With the complexity of the models, the trends were not being captured.

One of the hypotheses we had was that, as the Vaccine values were very high as compared to the other attributes, we thought that could be overpowering the model and creating a bias. However, the simple LSTM did not have this issue at all. However, to test this assertion, we employed normalization for LSTM-AE, as discussed in Section 3, but we observed early overfitting and learning termination due to the zero-mean feature values and subpar prediction/forecasting behavior. Upon training, we found the model to be extremely overfit. This can also be seen in the case of the LASSO Regression caused by the inbuilt normalization of the algorithm.

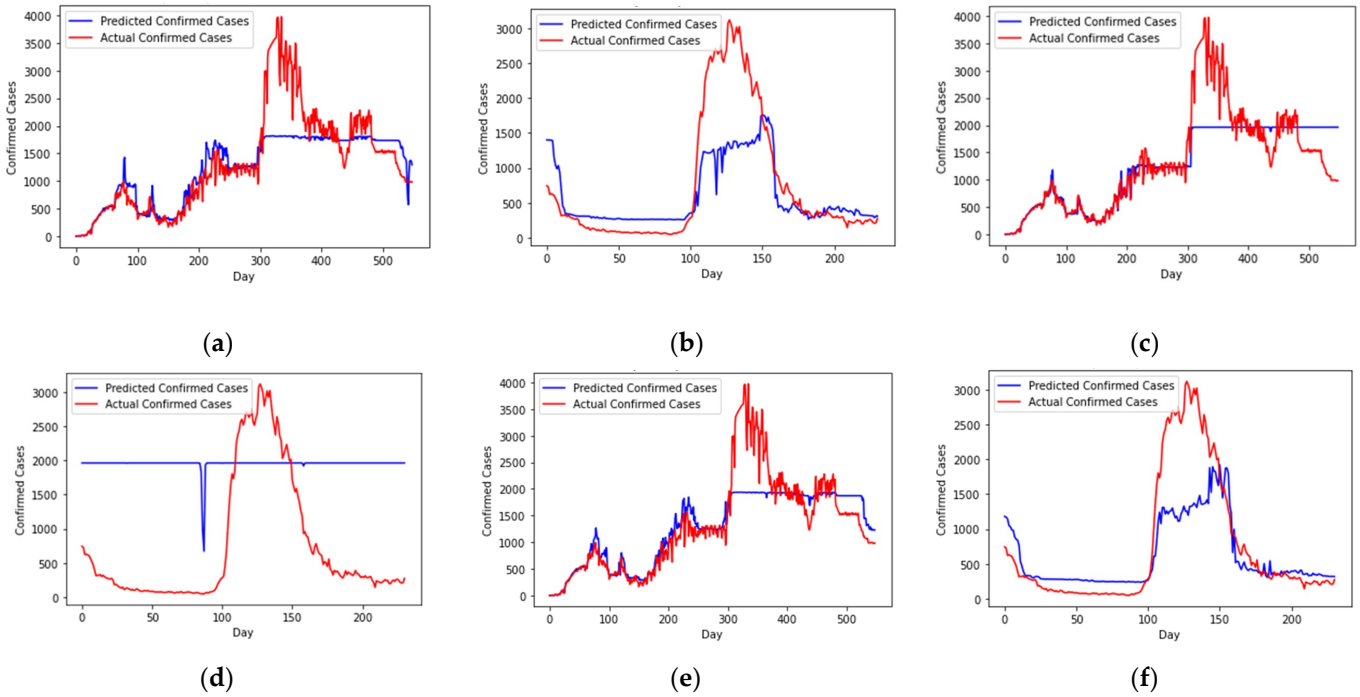


Figure 12. (**a**) Prediction using the bidirectional LSTM-AE on the training set using multivariate data (**b**) Prediction using the bidirectional LSTM-AE on the testing set using multivariate data (**c**) Prediction using the bidirectional LSTM-AE on the training set using vaccination data (**d**) Prediction using the bidirectional LSTM-AE on the testing set using vaccination data (**e**) Prediction using the bidirectional LSTM-AE on the training set using event data (**f**) Prediction using the bidirectional LSTM-AE on the testing using event data.

Similar to with the previous models, we see that, with the inclusion of the Events attributes, there is a slight improvement, but the new variables do not provide sufficient information to the seasonality for the model. Here, too, the model gets stuck in a local optimum after a while and does not learn very well. We can also observe that, due to the model stopping itself at around the value of 2000 cases, when predicting the testing cases, the model does not peak beyond this value.

5. Discussion

As shown in the previous section, multiple models with various combinations of attributes were studied to future forecast COVID-19 in the UAE. Table 8 summarizes the results of the models developed in this paper. It can be seen that four of the models have competitively outperformed the original UAE paper [28], with an improvement of over 30%. A more recent work published in 2024 [43] also tests multiple deep learning models for forecasting COVID-19 in the UAE and Malaysia. The authors reported an MAE of 0.046 and an R^2 score of 0.004 for their Univariate LSTM model. Our proposed model achieved an R^2 score of 0.91, which shows that the model learned the behavior better.

Table 8. Comparison of recent works for forecasting COVID-19 cases in UAE.

Source	Data Size	Model	RMSE
Proposed work	791	Univariate LSTM	2.76×10^2
Proposed work	791	Multivariate LSTM	3.15×10^2
Proposed work	791	Multivariate LSTM + Vaccinations	3.43×10^2
Proposed work	791	Multivariate LSTM + Events	3.47×10^2
K. K. A. Ghany et al. [28]	333	Univariate LSTM	4.46×10^2

The models developed in this work were compared to other UAE works only due to the fact that different countries were affected by different magnitudes of COVID-19 outbreak, as can be noticed from the literature review. Therefore, it is not fair to compare the results of this work to the works of other countries. Furthermore, the proposed work is the only one that used vaccination data as an attribute.

From these results, we can infer that LSTM works better with more data. It can also be gleaned that, due to the simplicity of the nature of the data, a relatively simple LSTM architecture is sufficient to forecast the number of cases. From prior knowledge, one would think that vaccinations, policies, and events would play a part in the spread of COVID-19. This can be seen partially here, with the improvement of predictions when those attributes were added, but the performance is still not the best due to the limited information provided to the model due to the nature of the data. It is suspected that, after the initial drives and efforts to curb the rapid contagion of COVID-19, public health measures and other similar policies remained roughly the same for many months during these periods without major changes. Thus, it could be said that concordance with public health measures such as mass vaccinations, booster drives, and quarantines helped flatten the curve only during the most infectious incubation period. Another factor is that the breakdown of cases by COVID-19 variant (original, Omicron, Delta, etc.) is not available, which means there is no "impurity" in the data, and no additional information that can be leveraged. This falls under the virus characteristics, where the mutation rate changes based on the different variants of the virus that are simultaneously rampant among the population. Since the active cases, the rate of infection, and even the mortality rate were different across different strata of the population (labor camps, social workers, essential staff, students, medical professionals, etc.), it would have been useful to have the average susceptibility of each demographic per day or per month to offer more modelling power.

Stricter protocols, such as fines imposed by the governments [44] to dissuade public gatherings, were not unanimously followed by all individuals in the UAE. This is also stated in a recent publication in the year 2022 by Chandra et al. [45], where the authors used LSTMs to predict the number of COVID-19 cases in India. The authors stated that, although the models work well, there is still uncertainty present due to limitations of COVID-19 related datasets. This uncertainty arises from the inherent flaw in the consolidation of information. Not everyone with the disease was detected and accounted for, their contacts were not logged, and the infection rate at gatherings organized at private residences could not be catalogued. Despite these general limitations, the UAE authorities were able to contain the spread and achieve high levels of immunization among the residents and citizens. Hence, we can assume that, since the data used are reported by the UAE Government, they provide the best estimate of the condition and any inherited/reporting error induced can be ignored [46].

Through this study, we see that there is a future for deep learning models to help forecast the number of cases based on different situations in the country, but there is still room for improvement in terms of data collection and its application on the models.

6. Conclusions

In this paper, we have applied machine learning and deep learning to analyze the trends in forecasting of COVID-19 cases in the UAE. We have used different combinations of the input attributes such as Vaccinations in addition to event-related attributes, with univariate and multivariate analysis. The dataset was initially analyzed and preprocessed, then the models were trained and compared. The dataset used was taken from the Federal Competitiveness and Statistics Centre website and is updated daily. Five different models were experimented on, including Exponential Smoothing, LASSO Regression, and LSTM, which were selected as they were the most commonly used models, along LSTM-AE and Bidirectional LSTM-AE. Upon evaluation, it was found that univariate LSTM performed the best with an RMSE of 275.85, which was an improvement of more than 30% from the current state of the art related to the UAE [28]. This model had the lowest RMSE,

which proves its feasibility as a model for predicting the COVID-19 trends in the UAE. The model was able to learn the trend and the seasonality of the data. It was also seen that the bidirectional LSTM was also very promising.

Although, the Vaccination attribute and the Event attributes did not perform as we expected, which was that the attributes would show a direct effect on the trend. However, there was only approximately a 10% decrease in RMSE as compared to the best model. This can be attributed to the amount of information these attributes are providing to the model in terms of trend and seasonality. From this we learned that the attributes are useful but require some additional attributes/biases to help them.

Through further research and the incorporation of granular informative attributes, more efficient models can be implemented. This can facilitate the simulation of various situations by the government and help policy making in terms of reducing the burden of pandemics on the nation.

Limitations and Future Work

Overall, we see that the simple univariate LSTM performs the best, with some additional information learned through the new added event related variables. We can also see that some variations of the bidirectional LSTM are very promising.

The major limitation to this work is the nature of the data. The data are purely numerical and statistical and lack depth in regard to the contextual severity of the effects of COVID-19 in the UAE. Another limitation is that the data are represented as a whole for the UAE and are not divided into districts. By dividing the data into districts, the model could be more accurate, as Emirates such as Abu Dhabi had more restrictions which would have slowed the spread there, whereas Emirates such as Sharjah that have a higher population would have caused an increase in spread. As these data were not provided by the government, studying the spread of COVID-19 by city of district was not addressed in this work; however, this observation can be taken into consideration during data reporting for future pandemics.

There is an additional limitation in regard to the Vaccination attribute, as it is just a number representing the number of doses administered per day and it does not consider which dose it is or whether a person who was vaccinated contracted COVID or not. Similarly, there is a challenge in regard to the number of cases, as we do not have any information on whether all are new cases or if a person has contracted the virus repeatedly. For future work, further variable selection and analysis can be applied, especially on the weakly trained LSTM-AE, to see if the performance of the model increases or decreases by using sets of meaningful and uncorrelated regressors.

Another major challenge is the difficulty of manually creating and updating the event variables. The information is difficult to find online and this is a very tedious job. There is also room for human error. A solution for this would be to follow [21] and build and train a Natural Language Processing (NLP) model that can scrap English and Arabic news sites and tweets to record important information related to the happening of a particular day.

Another idea to try in the future is inspired by a recent survey published by authors Elsheikh et al. [45], who concluded that a simple singular model might not be efficient for predicting the trends in COVID-19 and that we might have to venture into hybrid and ensemble models to play with the advantages of various different models and accurately learn the seasonality of the trends.

However, we consider this work to be of some importance to the research community, as it quantitatively shows the performance of the base LSTM model for modelling a novel and rapidly mutating contagion like the COVID-19 virus within the United Arab Emirates.

Author Contributions: Conceptualization, S.D., M.P. and A.S.; methodology, D.S.; software, D.S.; validation, D.S.; formal analysis, D.S., S.D., M.P. and A.S.; investigation, D.S.; resources, D.S.; data curation, D.S.; writing—original draft preparation, D.S.; writing—review and editing, S.D., M.P. and A.S.; visualization, D.S.; supervision, S.D., M.P. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this work is publicly available from the UAE government, Federal Competitiveness and Statistics Centre (FCSC) @: https://fcsc.gov.ae/en-us (accessed on 30 April 2024).

Acknowledgments: The work in this paper was supported, in part, by the Open Access Program from the American University of Sharjah (Award #: OAPCEN-1410-E00282). This paper represents the opinions of the authors and does not mean to represent the position or opinions of the American University of Sharjah.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Miranda, M.N.S.; Pingarilho, M.; Pimentel, V.; Torneri, A.; Seabra, S.G.; Libin, P.J.K.; Abecasis, A.B. A Tale of Three Recent Pandemics: Influenza, HIV and SARS-CoV-2. *Front. Microbiol.* **2022**, *13*, 889643. [CrossRef] [PubMed]
2. World Health Organization. *Statement on the Second Meeting of the International Health Regulations (2005) Emergency Committee Regarding the Outbreak of Novel Coronavirus (2019-nCoV)*; World Health Organization: Geneva, Switzerland, 2020; Available online: https://covid19.who.int/ (accessed on 10 October 2023).
3. Coronavirus Disease 2019 (COVID-19)—Symptoms and Causes—Mayo Clinic. Available online: https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963 (accessed on 20 August 2021).
4. Duong, B.V.; Larpruenrudee, P.; Fang, T.; Hossain, S.I.; Saha, S.C.; Gu, Y.; Islam, M.S. Is the SARS CoV-2 Omicron Variant Deadlier and More Transmissible Than Delta Variant? *Int. J. Environ. Res. Public Health* **2022**, *19*, 4586. [CrossRef] [PubMed]
5. Machado, J.A.T.; Ma, J. Nonlinear dynamics of COVID-19 pandemic: Modeling, control, and future perspectives. *Nonlinear Dyn.* **2020**, *101*, 1525–1526. [CrossRef] [PubMed]
6. Turak, N. First Middle East Cases of Coronavirus Confirmed in the UAE. *CNBC*, 29 January 2020. Available online: https://www.cnbc.com/2020/01/29/first-middle-east-cases-of-coronavirus-confirmed-in-the-uae.html (accessed on 20 August 2021).
7. UAE COVID-19 Updates. Available online: https://fcsc.gov.ae/en-us/Pages/Covid19/UAE-Covid-19-Updates.aspx (accessed on 20 August 2021).
8. Cooper, I.; Mondal, A.; Antonopoulos, C.G. A SIR model assumption for the spread of COVID-19 in different communities. *Chaos Solitons Fractals* **2020**, *139*, 110057. [CrossRef] [PubMed]
9. Alkhateeb, N.; Sallabi, F.; Harous, S.; Awad, M. A Study on Predicting the Outbreak of COVID-19 in the United Arab Emirates: A Monte Carlo Simulation Approach. *Mathematics* **2022**, *10*, 4434. [CrossRef]
10. Staffini, A.; Svensson, A.K.; Chung, U.-I.; Svensson, T. An Agent-Based Model of the Local Spread of SARS-CoV-2: Modeling Study. *JMIR Public Health Surveill.* **2021**, *9*, e24192. [CrossRef] [PubMed]
11. Kerr, C.C.; Stuart, R.M.; Mistry, D.; Abeysuriya, R.G.; Rosenfeld, K.; Hart, G.R.; Núñez, R.C.; Cohen, J.A.; Selvaraj, P.; Hagedorn, B.; et al. Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLoS Comput. Biol.* **2021**, *17*, e1009149. [CrossRef] [PubMed]
12. Rustam, F.; Reshi, A.A.; Mehmood, A.; Ullah, S.; On, B.-W.; Aslam, W.; Choi, G.S. COVID-19 Future Forecasting Using Supervised Machine Learning Models. *IEEE Access* **2020**, *8*, 101489–101499. [CrossRef]
13. Bhadana, V.; Jalal, A.S.; Pathak, P. A Comparative Study of Machine Learning Models for COVID-19 prediction in India. In Proceedings of the 2020 IEEE 4th Conference on Information & Communication Technology (CICT), Chennai, India, 3–5 December 2020; pp. 1–7.
14. Gupta, V.K.; Gupta, A.; Kumar, D.; Sardana, A. Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. *Big Data Min. Anal.* **2021**, *4*, 116–123. [CrossRef]
15. COVID-19 in India. Available online: https://kaggle.com/sudalairajkumar/covid19-in-india (accessed on 25 November 2021).
16. Romadhon, M.R.; Kurniawan, F. A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of COVID-19 Patients in Indonesia. In Proceedings of the 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), Surabaya, Indonesia, 9–11 April 2021; pp. 41–44.
17. Kumari, R.; Kumar, S.; Poonia, R.C.; Singh, V.; Raja, L.; Bhatnagar, V.; Agarwal, P. Analysis and predictions of spread, recovery, and death caused by COVID-19 in India. *Big Data Min. Anal.* **2021**, *4*, 65–75. [CrossRef]
18. COVID-19 Open Research Dataset Challenge (CORD-19). Available online: https://kaggle.com/allen-institute-for-ai/CORD-19-research-challenge (accessed on 25 November 2021).

19. Petropoulos, F.; Makridakis, S.; Stylianou, N. COVID-19: Forecasting confirmed cases and deaths with a simple time series model. *Int. J. Forecast.* **2020**, *38*, 439–452. [CrossRef] [PubMed]
20. Leon, M.I.; Iqbal, I.; Azim, S.M.; Al Mamun, K.A. Predicting COVID-19 infections and deaths in Bangladesh using Machine Learning Algorithms. In Proceedings of the 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 27–28 February 2021; pp. 70–75.
21. Zheng, N.; Du, S.; Wang, J.; Zhang, H.; Cui, W.; Kang, Z.; Yang, T.; Lou, B.; Chi, Y.; Long, H.; et al. Predicting COVID-19 in China Using Hybrid AI Model. *IEEE Trans. Cybern.* **2020**, *50*, 2891–2904. [CrossRef] [PubMed]
22. Kumar, S.; Sharma, R.; Tsunoda, T.; Kumarevel, T.; Sharma, A. Forecasting the spread of COVID-19 using LSTM network. *BMC Bioinform.* **2021**, *22*, 316. [CrossRef]
23. Chimmula, V.K.R.; Zhang, L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos Solitons Fractals* **2020**, *135*, 109864. [CrossRef] [PubMed]
24. Helli, S.S.; Demirci, O.; Coban, O.; Hamamci, A. Short-Term Forecasting COVID-19 Cases in Turkey Using Long Short-Term Memory Network. In Proceedings of the 2020 Medical Technologies Congress (TIPTEKNO), Antalya, Turkey, 19–20 November 2020; pp. 1–4.
25. Ramchandani, A.; Fan, C.; Mostafavi, A. DeepCOVIDNet: An Interpretable Deep Learning Model for Predictive Surveillance of COVID-19 Using Heterogeneous Features and Their Interactions. *IEEE Access* **2020**, *8*, 159915–159930. [CrossRef] [PubMed]
26. Kafieh, R.; Arian, R.; Saeedizadeh, N.; Amini, Z.; Serej, N.D.; Minaee, S.; Yadav, S.K.; Vaezi, A.; Rezaei, N.; Javanmard, S.H. COVID-19 in Iran: Forecasting Pandemic Using Deep Learning. *Comput. Math. Methods Med.* **2021**, *2021*, 1–16. [CrossRef] [PubMed]
27. Zain, Z.M.; Alturki, N.M. COVID-19 Pandemic Forecasting Using CNN-LSTM: A Hybrid Approach. *J. Control. Sci. Eng.* **2021**, *2021*, 8785636. [CrossRef]
28. Ghany, K.K.A.; Zawbaa, H.M.; Sabri, H.M. COVID-19 prediction using LSTM Algorithm: GCC Case Study. *Inform. Med. Unlocked* **2021**, *23*, 100566. [CrossRef]
29. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). Available online: https://coronavirus.jhu.edu/map.html (accessed on 8 June 2020).
30. Witten, I.H.; Frank, E.; Hall, M.A. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed.; Morgan Kaufmann/Elsevier: Burlington, NJ, USA, 2011; ISBN 978-0-12-374856-0.
31. Gomes, D.C.d.S.; Serra, G.L.d.O. Machine Learning Model for Computational Tracking and Forecasting the COVID-19 Dynamic Propagation. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 615–622. [CrossRef]
32. Petropoulos, F.; Makridakis, S. Forecasting the novel coronavirus COVID-19. *PLoS ONE* **2020**, *15*, e0231236. [CrossRef]
33. Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 420. [CrossRef] [PubMed]
34. Intro to Autoencoders | TensorFlow Core. Available online: https://www.tensorflow.org/tutorials/generative/autoencoder (accessed on 26 August 2022).
35. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
36. Fernando, J. R-Squared. *Investopedia*, 12 September 2021. Available online: https://www.investopedia.com/terms/r/r-squared.asp (accessed on 22 August 2021).
37. Mean Squared Error: Definition and Example. Statistics How To. Available online: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-squared-error/ (accessed on 27 September 2021).
38. Ritchie, H.E.; Mathieu, L.; Rodés-Guirao, C.; Appel, C.; Giattino, E.; Ortiz-Ospina, J.; Hasell, B.; Macdonald, D.; Roser, B.a.M. Coronavirus (COVID-19) Vaccinations. 2020. Available online: https://ourworldindata.org/covid-vaccinations (accessed on 28 December 2021).
39. Anderson, A.; Semmelroth, D. Autocorrelation Plots: Graphical Technique for Statistical Data—Dummies. Statistics for Big Data for Dummies. 2016. Available online: https://www.dummies.com/programming/big-data/data-science/autocorrelation-plots-graphical-technique-for-statistical-data/ (accessed on 22 August 2021).
40. Augmented Dickey-Fuller (ADF) Test—Must Read Guide. *ML+ Machine Learning Plus*, 2 November 2019. Available online: https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/ (accessed on 22 August 2021).
41. Mukaka, M.M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi. Med. J.* **2012**, *24*, 69–71. [PubMed]
42. Pearson Product-Moment Correlation—When you Should Run this Test, the Range of Values the Coefficient Can Take and How to Measure Strength of Association. Available online: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php (accessed on 22 August 2021).
43. Tariq, M.U.; Ismail, S.B. Deep learning in public health: Comparative predictive models for COVID-19 case forecasting. *PLoS ONE* **2024**, *19*, e0294289. [CrossRef] [PubMed]
44. News Details | UAE Coronavirus (COVID-19) Updates. Available online: https://covid19.ncema.gov.ae/en/News/Details/2316 (accessed on 27 August 2022).

45. Chandra, R.; Jain, A.; Chauhan, D.S. Deep learning via LSTM models for COVID-19 infection forecasting in India. *PLoS ONE* **2022**, *17*, e0262708. [CrossRef]
46. Elsheikh, A.H.; Saba, A.I.; Panchal, H.; Shanmugan, S.; Alsaleh, N.A.; Ahmadein, M. Artificial Intelligence for Forecasting the Prevalence of COVID-19 Pandemic: An Overview. *Healthcare* **2021**, *9*, 1614. [CrossRef]