Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 9953283, 9 pages
https://doi.org/10.1155/2021/9953283

Research Article

Correlation Determination between COVID-19 and Weather Parameters Using Time Series Forecasting: A Case Study in Pakistan

Humera Batool [1] **and Lixin Tian** [1,2,3]

[1]*School of Mathematical Sciences, Nanjing Normal University, Nanjing, Jiangsu 210023, China*
[2]*Center for Energy Development and Environmental Protection, Jiangsu University, Zhenjiang, Jiangsu 212013, China*
[3]*Research Centre of Energy-Interdependent Behavior and Strategy, Nanjing Normal University, Nanjing, Jiangsu 210023, China*

Correspondence should be addressed to Humera Batool; drhumerabatool@hotmail.com

Received 21 March 2021; Revised 24 April 2021; Accepted 29 May 2021; Published 15 June 2021

Academic Editor: Bhawani Shankar Chowdhry

Infectious diseases like COVID-19 spread rapidly and have led to substantial economic loss worldwide, including in Pakistan. The effect of weather on COVID-19 spreading needs more detailed examination, as some studies have claimed to mitigate its spread. COVID-19 was declared a pandemic by WHO and has been reported in about 210 countries worldwide, including Asia, Europe, the USA, and North America. Person-to-person contact and international air travel between the nations were the leading causes behind the spreading of SARS-CoV-2 from its point of origin, besides the natural forces. However, further spread and infection within the community or country can be aided by natural elements, such as the weather. Therefore, the correlation between COVID-19 and temperature can be better elucidated in countries like Pakistan, where SARS-CoV-2 has affected at least 0.37 million people. This study collected Pakistan's COVID-19 infection and mortality data for ten months (March–December 2020). Related weather parameters, temperature, and humidity were also obtained for the same course of time. The collected data were processed and used to compare the performance of various time series prediction models in terms of mean squared error (MSE), root-mean-squared error (RMSE), and mean absolute percentage error (MAPE). This paper, using the time series model, estimates the effect of humidity, temperature, and other weather parameters on COVID-19 transmission by obtaining the correlation among the total infected cases and the number of deaths and weather variables in a particular region. Results depict that weather parameters hold more influence in evaluating the sum number of cases and deaths than other factors like community, age, and the total population. Therefore, temperature and humidity are salient parameters for predicting COVID-19 affected instances. Moreover, it is concluded that the higher the temperature, the lesser the mortality due to COVID-19 infection.

1. Introduction

A viral infection named COVID-19 was initially discovered in mid-December 2019 in Wuhan city of China [1], which spread across the whole world, and eventually WHO declared it as a pandemic [2]. Figure 1 shows the map along with the total number of confirmed cases in the province. Up to November 22, 2020, there was 58,475,749 COVID-19 cases, 1,385,775 deaths, and 40,459,596 recoveries across the world, out of which 371,508 total cases, 7,603 deaths, and 328, 931 recoveries were in Pakistan [3]. Although SARS-CoV-2 originated from China, the world's biggest population, it was effectively controlled in China's epicenter and other regions since February 2020 [4]. Daily COVID-19 cases in Pakistan peaked at 6,825 on June 14, 2020; then, it declined to 331 on August 3, 2020; and from the first week of November 2020, again it showed ascending pattern. Albeit there is a cure, the main focus is to curb the spread through national blockades and quarantine measures [5]. Such a high daily number of cases warrants an immediate plan of action to control it effectively and its need to prepare for future outbreaks in Pakistan and other nations.

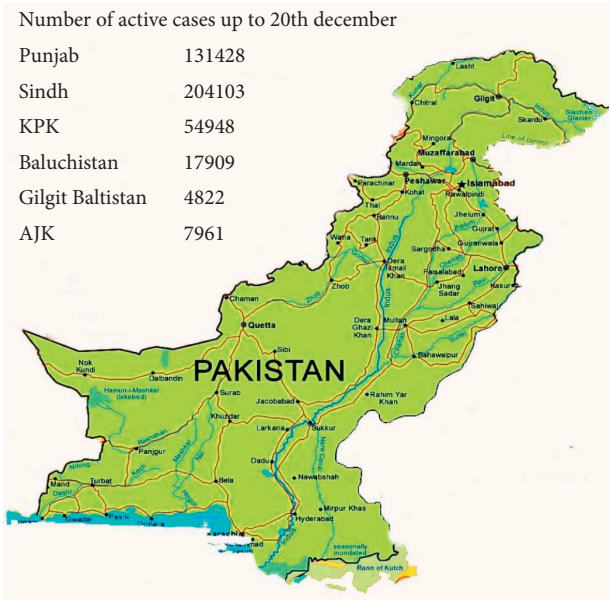

Figure 1: Map of Pakistan showing active cases in all regions of Pakistan, that is, Gilgit Baltistan (GB), Azad Jammu and Kashmir (AJK), Khyber Pakhtunkhwa (KPK), Baluchistan, Islamabad, Punjab, Sindh, as of December 20, 2020

Recently, scientists elucidated close affiliation between weather parameters and main COVID-19 epidemic areas. Moreover, these areas are located in a relatively temperate region in the northern partition [6]. Although pandemic is a global issue, the outbreak epicenters of the world have a mean temperature of 5°C–11°C with 47%–79% humidity in the first two months of the year 2020. Based on these facts, our primary hypothesis is that virus spread is curtailed in high-temperature and humidity areas rather than areas having average temperature and humidity.

Initially, two cases in Pakistan appeared on February 26, 2020. In parallel, three more cases were recorded in subsequent hours from different cities and there was no affinity/contact among these COVID-19 victims. Unexpectedly, an increase in the number of affected persons on April 14, 2020, was witnessed with the highest number of cases in Punjab, that is, 2826. Sindh was the second with 1452 patients, KPK was the third, having 800 patients, Gilgit Baltistan was in fourth place, having 233 patients, and Baluchistan and Islamabad had 321 and 131 cases, respectively. In contrast, AJK had the least number of cases, i.e., 43 [7]. As there is no proper cure for this pandemic yet discovered and multiple forms of SARS-CoV-2 are also dependent on seasonality [8], these all factors make SARS-CoV-2 spread more alarming and lethal. Short-term forecasting is inevitable to maintain the balance between social, economic, and health aspects in subsequent months [9]. To illustrate the nature of SARS-CoV-2 and to forecast its transmission, there is a dire need to explore its effect on weather. In this regard, the systematic approach of our study includes the following:

(a) Using existing data to predict the number of actual COVID-19 affected cases and the total number of deaths in upcoming months with or without weather data in Pakistan.

(b) To determine the fragile range of climatic factors and verifying these factors at various periods through statistical analysis.

(c) Aiding Pakistan government institutions and policymakers to adopt new strategies to strengthen existing preventive measures to combat the COVID-19 pandemic.

Demongeot et al. [10] identified that the virulence of SARS-CoV-2 and their lethal strains get downregulated in hot and humid climate conditions. The presumed temperature-dependent virulence of COVID-19 also got an eminent interest in the medical field. Instead of the above, our study aims to determine critical factors relying on temperature and transmission kinetics of COVID-19, which increases with cold and dry weather.

Sajadi et al. [11] explained a simplified model describing a zone at high virulence of the COVID-19 outbreak. Bloom-Feshbach et al. [12] elaborated that COVID-19 prevails high in cold and temperate climates than warm and tropical climates, which acknowledges respiratory influenza viruses. For natural distancing calculation and estimation, Prem et al. [13] utilized an age-structured susceptible-exposed-infected-removed (SEIR) model. This study illustrated that if arrival to work initiated in April, physical distancing measures would be most efficient. Eikenberry et al. [14] stated that the SEIR model aimed to evaluate the potential colony impact of the adoption of masks by the public on the mobility and control of the COVID-19. The study recommended using masks nationwide and implementing their use strictly.

Research work related to applying machine learning tools to elucidate the impact of weather parameters on transmission and circulation of COVID-19 seemed lacking and needs more attention. In addition, ascending temperature may or may not lower SARS-CoV-2 spread, and likewise role of other weather factors is also still under debate.

Therefore, past studies are concise to various models, and findings are also not authentic. Hence, it is time to understand the relationship between weather variables and the epidemic spreading of COVID-19 in Pakistan.

2. Materials and Methods

2.1. Data Collection. The daily cumulative total number of confirmed cases and the total number of deaths were obtained from the official website of the National Institute of Health (NIH) in Islamabad, Pakistan. The National Institute of Health is an independent health research department under the Ministry of National Health Services of Pakistan. It is located in Islamabad and is engaged in various research activities and vaccine making. Daily COVID-19 diagnosed cases, recoveries, deaths, and COVID-19 diagnostic tests conducted across Pakistan were updated on the official website of NIH [15].

2.2. Examination. The data were collected from March 10 to December 20, 2020, both for COVID-19 and weather, and was further divided into training and testing datasets. The training dataset comprises the data from March 10 to November 15, 2020, and the testing dataset has data from November 15 to December 20, 2020. Test data was further analyzed for a cumulative number of cases and deaths with and without weather data. Figure 2 shows the division of complete data into training and testing datasets.

2.3. Methods. We have applied simple machine learning models, deep learning techniques, and statistical models to predict the total number of cases and total deaths with and without weather data for COVID-19. Time series prediction models such as ARIMA, linear regression, SVM, MLP, RNN, LSTM, and GRU were used. Statistical performance of time series prediction models was measured in terms of mean squared error (MSE), root-mean-squared error (RMSE), and mean absolute percentage error (MAPE). For all these experiments, we used Python version 3.8, Scikit-learn version 0.21.0, and deep learning library Keras v.2.2.5 using tensor flow at the backend.

2.4. Autoregressive Integrated Moving Average (ARIMA). There are three types of ARIMA, namely, autoregression, data-dependent integration, and parameter estimation. All these three types are implemented according to the issue that needs to be focused on [16, 17]. The time series form of the process is

$$x^t = \Theta_0 + \Theta_1 x^{t-1} + \Theta_2 x^{t-2} + \cdots + \Theta_p x^{t-p} + \varepsilon^t - \Theta_1 \varepsilon^{t-1} - \cdots \Theta_q \varepsilon^{t-q}. \quad (1)$$

In the previous equation, x^t and ε^t depict the original value and random error at time t. Model parameters are $\emptyset a$ ($a = 1, 2..., p$) and $\emptyset b$ ($a = 0, 1, 2,..., q$). An unexpected error is defined by ε^t and considered with zero mean and $\sigma 2$ of standard variance. Equation (1) represents the ARIMA model and is applied to various applications for problem-solving.

Taking value $q = 0$, in equation (1) works as an AR model with order p, and for $p = 0$, it becomes MA model with order q. Thus, (p, q) are both inevitable factors for ARIMA model determination.

2.5. Linear Regression. Linear regression can be defined as

$$Y = \alpha + bX + \varepsilon, \quad (2)$$

where Y = dependent variable, X = independent variable, α = intercept, b = regression parameter as slope, and ε = random error.

The disadvantage of linear regression is that it usually correlates among an average of input and input variables. Unfortunately, a simple average is not a full illustration of a single variable.

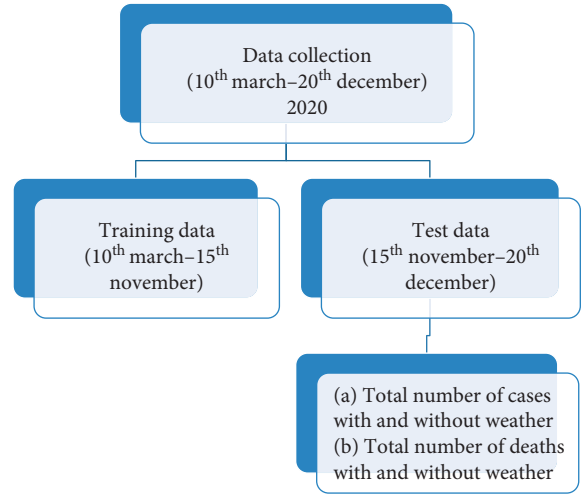

Figure 2: Diagram showing division of complete data into training and testing datasets.

2.6. Support Vector Regression (SVM). SVR involves evaluating the support vectors (points) near the hyperplane to upgrade boundary among two-point groups acquired by variation between objective value and threshold. SVR is employing kernel functions to elucidate nonlinear issues, which calculates the affinity between two values. We used the linear kernel function in our study. The main leverage of SVR is that it can capture the nonlinearity of the prediction and then use it to raise the prediction case. In the same scenario, it is beneficial to adopt this view in the case studies used because the sample is inadequate [18]. SVR for the complicated data is

$$y = f(x) = \sum_{i=1}^{M} w_i x_i + b, \quad (3)$$

where wi = input weights, y = actual values, b = bias, and M = total number of data samples. This comparison illustrates the purpose of use of SVR and $||W||$ = magnitude of vector:

$$\min_w \frac{1}{2} w^2. \quad (4)$$

Enabling SVM consists of two inadequate variables, that is, ε and ϵ^*. They are used to guard against anomalies, and $1/2||W||2$ is used for the precision of function. Both specifications rely on the C parameter. Then, equation (4) will transform into the following equation:

$$\min \frac{1}{2} w^2 + C \sum_{i=1}^{M} (\varepsilon_i + \varepsilon_i^*). \quad (5)$$

With the suppression,

$$\begin{aligned} &y_i - w^T x_i \le \varepsilon + \varepsilon_{i'}^*, \quad i = 1, 2, 3 \ldots M, \\ &W^T X_i y_i - y_i \le \varepsilon + \varepsilon_{i'}^*, \quad i = 1, 2, 3 \ldots M, \\ &\varepsilon_i, \varepsilon_i^* \ge 0. \end{aligned} \quad (6)$$

Finally, SVR task is accessed as

$$f(x) = \sum_{i=1}^{M} (\propto_i^* + \propto_i) k(X_i - X) + b. \tag{7}$$

2.7. Multilayered Perceptron (MLP). Multilayered perceptron (MLP) is the frequently used artificial neural network (ANN) for modeling and forecasting. For evaluating tasks in simple and semicomplicated datasets, this method provides considerable accuracy. It is a wholly joined feed-forward artificial neural network in which neurons are overlapped [19]. MLP has layers: an input, output, and hidden layer. The output layer in the presented research is the total number of cases and deaths. The MLP used in this study has three neurons in the input layer, and each neuron corresponds to an input data point (total cases, total deaths, and days since infection). MLP as the method has ease of implementation. In comparison to complex forms, MLP results in high-quality models while keeping robustness and accuracy in prediction.

Because MLP regressor can only revert an individual value, an adaptive model must be used if the issue hinders multiple output values. Although there may be resemblances among the models, training the whole model means that the dataset will be tested, so a better predictive model can be gained to address each issue. In the present study, three independent MLPs were employed.

2.8. Recurrent Neural Networks (RNNs). In deep learning, it is assumed that classified models are more prospering than flat models in regression tasks [20]. As RNN holds hidden states allocated across time, it favors them to accumulate previous information. Moreover, due to their capability of analyzing the variable length of data, they are abundantly used in forecasting [21]. Our research aims to analyze and evaluate the proposed prediction model, RNN, with different hyperparameters. The essential aspect of RNNs is to consider the impact of previous data on the generated output. Most importantly, RNN is effective for learning time information [22]. LSTM and GRU are two robust RNN models. These illustrations have depicted sublime outcomes in precision and accuracy compared to the classic time series models, and commonly used networks have identified that they can attain multiple outputs in various purposeful domains with time series [23, 24]. Figure 3 shows the conceptual framework of the applied proposed model depicting splitting of data into training and testing data. Further, testing data was evaluated using MSE, RMSE, and MAPE, while training data was validated through time series prediction models ARIMA, linear regression, SVR, MLP, RNN, GRU, and MAPE.

2.9. Gated Recurrent Units (GRU). GRU was presented by [25], which solves vanishing gradient with a standard RNN. GRU is reciprocal to LSTM, but it joins LSTM into one update gate. The GRU further combines cell and concealed form. It consists of a cell containing multiple operations which are duplicated and could be a neural network. When the neural network is applied through BPTT, it can prevent gradient vanishing [26]. The GRU layer, comprising reset gates and update gates, can learn long-term and short-term interdependence from the flow [25]. The mathematical interrelationship among different GRU factors is given by

$$\begin{aligned}
\text{updata gate } \mathscr{Z}_t &= o(\mathscr{X}_t \mathscr{W}_{xz} + \mathscr{H}_{t-1} \mathscr{W}_{hz} + b_z),\\
\text{reset gate } \mathscr{R}_t &= o(\mathscr{X}_t \mathscr{W}_{xr} + \mathscr{H}_{t-1} \mathscr{W}_{hr} + b_r),\\
\text{cell gate } \tilde{\mathscr{H}_t} &= tanh(\mathscr{X}_t \mathscr{W}_{xh} + (R_t o H_{t-1}) \mathscr{W}_{hh} + b_h),\\
\text{new state } \mathscr{H}_t &= \mathscr{Z}_t o \mathscr{H}_{t-1} (1 - \mathscr{Z}_t) o \tilde{\mathscr{H}_t},\\
\mathscr{W}_{xr}, \mathscr{W}_{xz}, \mathscr{W}_{hr} &= \text{weight parameters},\\
b_r, b_z &= \text{bias parameters}.
\end{aligned} \tag{8}$$

2.10. Long Short-Term Memory (LSTM). The common application of LSTM is in speech recognition and data prediction. Its robust performance in evaluating future predictions by modeling the issue as a series regression problem caught various scientists' attention due to its applications such as activity recognition, prediction, risk resolve, and fall detection [27, 28]. As a deep learning methodology, it leads to other forecasting methods [29]. LSTM is a type of RNN, and its original purpose is to eliminate errors in previous algorithms when back-propagating the information contained in the most recent input event [30]. There are two reasons for using LSTM. First, it returns the error to the machine to calibrate the model in the first training phase. At the same time, errors have been deliberately applied in mechanical gates. Second, the LSTM network is impartial to lag among events in the time series. Therefore, when we are trying to derive an unknown prediction model, the LSTM algorithm is more effective as compared to ANN's (such as hidden Markov and SVR) or other prediction methodologies (such as ARIMA) [29].

For the flow of information, LSTM has input, output, and duplicate gates. These gates are composed of weighted sum logistic functions, and the weights can be gained through backpropagation, all along the process of training. The input gate manages the unit state and the forget gate. The output is accomplished from the output gate or hidden state, and it illustrates the memory used by the direction. This structure permits the network to remember for a long duration, while the traditional single RNN does not have such memory. The ideal feature of LSTM is its extended quality to capture long-term dependencies and powerful ability to process time-series data. For example, given the input time-series X_t and the number of hidden units as h, the gates have the following equation:

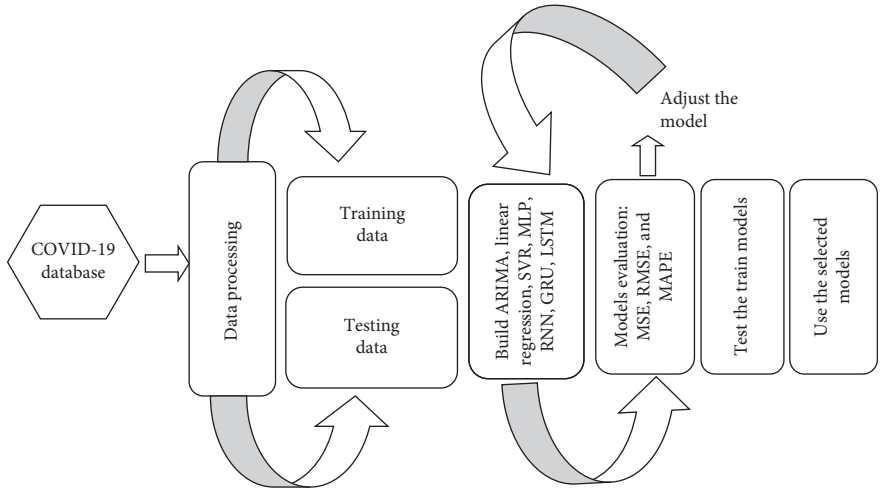

Figure 3: Conceptual framework of the proposed forecasting methods.

$$\begin{aligned}
\text{Input gate: } \mathrm{I_t} &= o\left(\mathscr{X}_t \mathscr{W}_{\mathrm{xi}} + \mathscr{H}_{\mathrm{t-1}} \mathscr{W}_{\mathrm{hi}} + b_{\mathrm{i}}\right),\\
\text{Forward gate: } F_{\mathrm{t=0}} &\left(\mathscr{X}_t \mathscr{W}_{\mathrm{xf}} + \mathscr{H}_{\mathrm{t-1}} \mathscr{W}_{\mathrm{hf}} + b_{\mathrm{f}}\right),\\
\text{Output gate: } O_{\mathrm{t=0}} &\left(\mathscr{X}_t \mathscr{W}_{\mathrm{xo}} + \mathscr{H}_{\mathrm{t-1}} \mathscr{W}_{\mathrm{ho}} + b_{\mathrm{o}}\right),\\
\text{Intermediate cell state: } \mathscr{C}_{\mathrm{t}}^{\sim} &= \tanh\left(\mathscr{X}_t \mathscr{W}_{\mathrm{xc}} + \mathscr{H}_{\mathrm{t-1}} \mathscr{W}_{\mathrm{hc}} + b_{\mathrm{c}}\right),\\
\text{Cell state (next memory input) } \mathscr{C}_{\mathrm{t}} &= \mathscr{F}_{\mathrm{t}} o \mathscr{C}_{t-1}\, \mathrm{o} \mathscr{C}_t^{\sim},\\
\text{New state: } \mathscr{H}_{\mathrm{t}} &= \mathscr{O}_{\mathrm{t}} \mathrm{o}\, \tanh\left(\mathscr{C}_t\right).
\end{aligned} \tag{9}$$

$\mathscr{W}_{\mathrm{xi}}$, $\mathscr{W}_{\mathrm{xf}}, \mathscr{W}_{\mathrm{xo}}$, and $\mathscr{W}_{\mathrm{hc}}, \mathscr{W}_{\mathrm{hf}}, \mathscr{W}_{\mathrm{ho}}$ are weight parameters and $\mathrm{b_i}, \mathrm{b_f}, \mathrm{b_o}$ denote bias parameters. $\mathscr{W}_{\mathrm{xc}}, \mathscr{W}_{\mathrm{hc}}$ = weight parameter, b_c is bias parameter, and o = elementwise multiplication. The estimation of $\mathscr{C}_{\mathrm{t}}$ depends on the output information's from memory cells ($\mathscr{C}_{\mathrm{t-1}}$) and the current time step $\mathscr{C}_{\mathrm{t}}^{\sim}$.

2.11. Performance Metrics. Measure the average of the squares of the errors. It is the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss:

$$\mathrm{MSE} = \sqrt{\frac{1}{x}\sum_{i-1}^{n}\left(Y_{i+} - Y_t^{\sim}\right)^2}. \tag{10}$$

3. Root-Mean-Squared Error

Root-mean-square error is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed:

$$\mathrm{RMSE} = \sqrt{\frac{1}{x}\sum_{t-1}^{n}\left(Y_t - Y_t^{\wedge}\right)^2}. \tag{11}$$

3.1. Mean Absolute Percentage Error. The mean absolute percentage error (MAPE) is a measure of prediction accuracy of a forecasting method in statistics, for example, in trend estimation, also used as a loss function for regression problems in machine learning. It usually expresses the accuracy as a ratio defined by the formula

$$\mathrm{MAPE} = \frac{100}{x}\sum_{t-1}^{n}\left|Y_t - Y_t^{\wedge} \div Y_t\right|_{\%}. \tag{12}$$

4. Results

Evaluation of COVID-19 transmission using mathematical models requires training on a large number of datasets. The size of the dataset affects the performance of the proposed algorithms and holds a considerable role in training. The dataset is classified into two parts, the training and the testing datasets. A training dataset is employed during model development, whereas testing datasets are used to validate datasets that are not previously used [31, 32].

The interrelationship between COVID-19 and weather factors in the case of Pakistan is examined in this study. The number of confirmed COVID-19 cases (dependent variable) was log-transformed to make it work as normal distribution as the original data is highly skewed in the selected area. For a specified period up to November 15, 2020, training data evaluates the statistics of cases by considering Pakistan's humidity and temperature data. The hypothesis is that high humidity and temperature (weather variables) shall coincide with a lowered count of SARS-CoV-2 cases. Figures 4(a) and 4(b) illustrate a scatter plot among the number of proved infections compared to thermal readings and humidity in Pakistan.

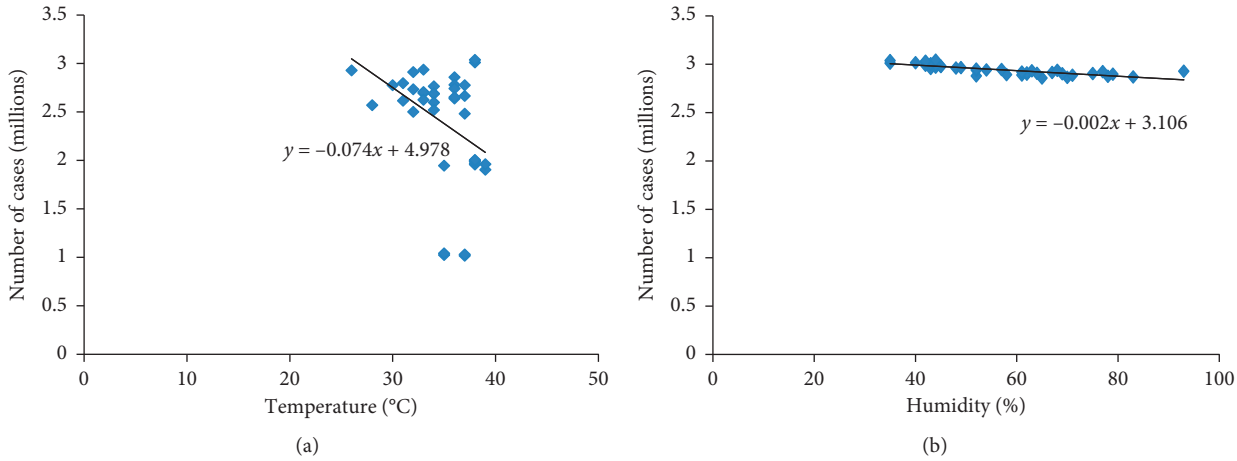

Figure 4: Scatter plot showing the number of cases in Pakistan based on temperature (a) and humidity (b).

Figures 5(a) and 5(b) depict a scatter plot among the number of deaths compared to temperature and humidity in Pakistan. From these findings, it is understood that, as atmospheric moisture and thermal reading decline (increase in temperature), the numbers of infected cases and death rates also decline.

When temperature and humidity showed ascending pattern, the infection rate descends. However, this fact is unavoidable that when sunlight hours increase, interaction among people increases. As a result, the infection rate may elevate. The people residing in urban areas are also strongly influenced because it means a higher population density, making COVID-19 inferior. Several parameters which can affect COVID-19 spread could be considered as a potential carrier. Population density also matters in epidemic spreading. Older people are more susceptible to the epidemic. Figures 6(a) and 6(b) depict the total number of cases without weather, while Figures 6(c) and 6(d) denote deaths. In both cases, we can observe in Figures 6(a) and 6(c) that the difference between actual and predicted graph lines is more significant than Figures 6(b) and 6(d). Predictions evaluated with weather data showed that the addition of weather parameters improved predicted results.

In order to understand whether the weather parameters, that is, temperature and humidity inclusion, affect the results or not, we created more comprehensive time series prediction models using ARIMA, linear regression, SVR, MLP, GRU, and LSTM. The current time series prediction model gives better facilitation to elucidate the impact of weather parameters on epidemic spreading. In addition, these time series models aid in illustrating the authentic interrelationship among the number of proved cases, deaths, and weather factors. Tables 1 and 2 predict the total number of cases (actual vs. predicted) and Table 3 and 4 elaborate a total number of fatalities (actual vs. predicted) with and without including weather variables, where the performance of these models are indicated in terms of MSE, RMSE, and MAPE.

5. Discussion

This study aims to figure out an output of seven-time series prediction models with and without weather data on the total number of COVID-19 cases and their mortality. In Table 1, it is clear that LSTM achieved better results with lower MSE, RMSE, and MAPE values. For illustration, the LSTM model achieved MAPE values 0.022, 0.0217, 0.0208, 0.0198, 0.0176, 0.0164, and 0.0155 for the total number of cases with weather data in Pakistan. Thus, the results depict that prediction of new COVID-19 confirmed cases by LSTM has sublime performance. The efficiency of the actual versus predicted total number of cases with weather data for COVID-19 is promising and evident. LSTM's outperforming ability to handle fewer datasets than the other models (linear regression, SVR, MLP, RNN, and GRU) which possibly require lengthier data to evaluate correlated fluctuation in time series data has made it a better choice. Conversely, RNN and its updated version GRU accommodate comparatively balanced forecasting performance due to the evaluation metrics (RMSE and MAPE), and explained variance is executed ambiguously.

The performance of time series models ARIMA, linear regression, SVR, MLP, RNN, GRU, and LSTM in MSE, RMSE, and MAPE predicting the total number of cases without considering weather data parameters (temperature and humidity) is shown in Table 2. It is clear that values of MSE, RMSE, and MAPE for all-time series prediction models were enhanced without the addition of weather data; for example, GRU showed values of 180.8178718, 13.4468536, and 0.018989281 for MSE, RMSE, and MAPE without weather data. In contrast, it was 140.0163399, 11.83285003, and 0.01641411, respectively, for the number of cases with weather parameters.

Similarly, Table 3 shows the application of time series prediction models on the number of deaths in Pakistan, considering the weather parameters, temperature, and humidity. LSTM shows the best MSE, RMSE, and MAPE values, that is, 1711, 41.36423576, and 0.492211157, respectively. But in Table 4, it is predicted that if we see performance of time series models ARIMA, linear regression, SVR, MLP, RNN, GRU, and LSTM in terms of MSE, RMSE, and MAPE without temperature and humidity, the accuracy of models descends. In both Tables 3 and 4, the ARIMA model shows the least accuracy and high error values, and LSTM predicts the least values of MSE, RMSE, and MAPE.

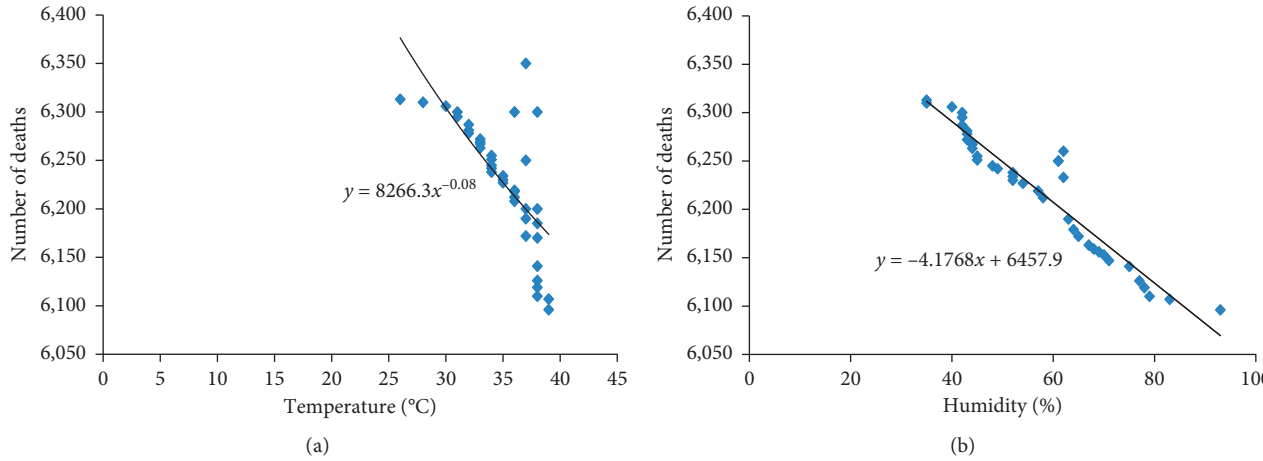

Figure 5: Scatter plot showing the number of deaths in Pakistan based on temperature (a) and humidity (b).

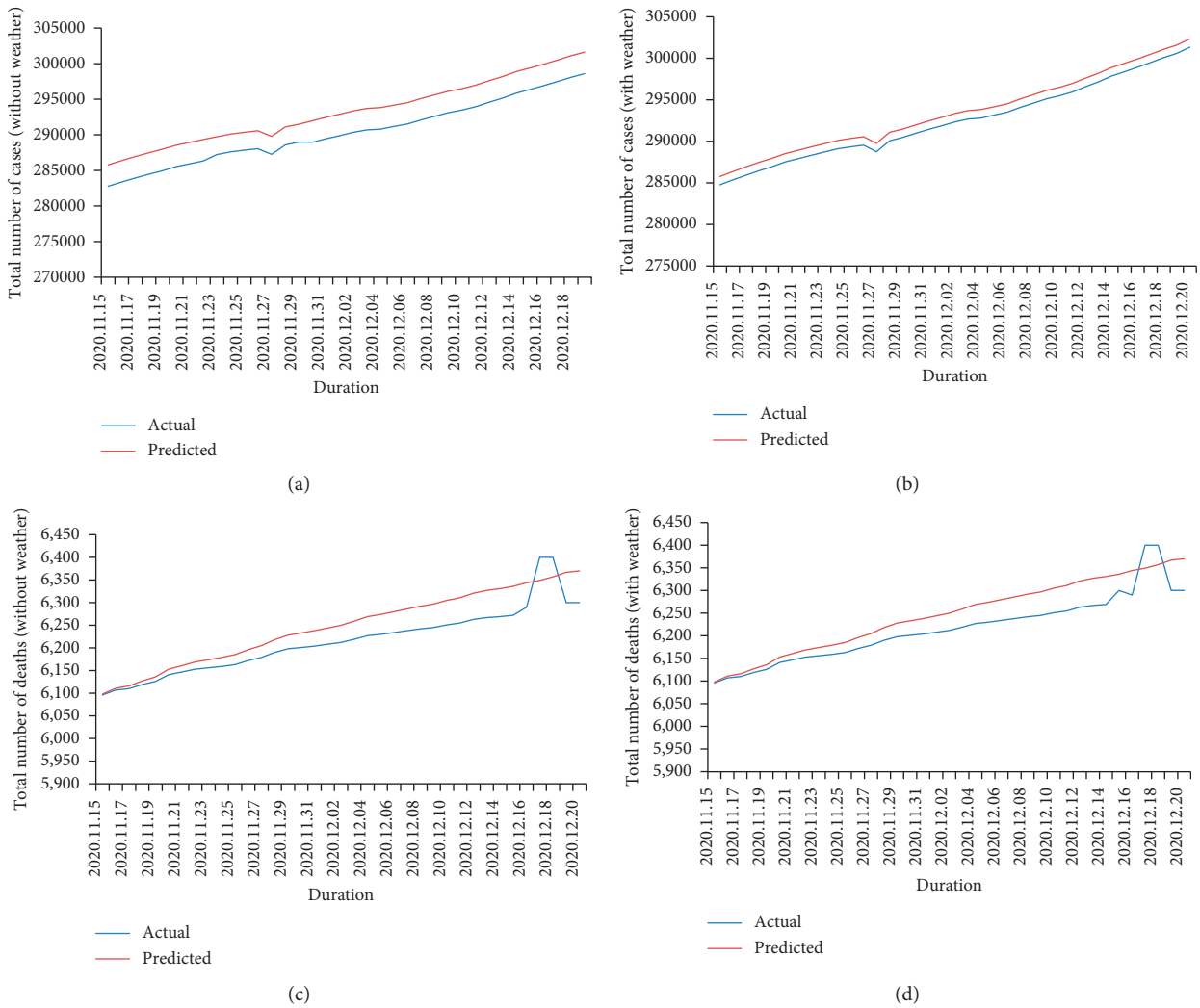

Figure 6: Total number of COVID-19 cases (actual vs. predicted) from November 15 to December 20, 2020, in Pakistan: (a) without weather and (b) with the weather. Total number of deaths due to COVID-19 (actual vs. predicted) from November 15 to December 20, 2020: (c) without weather and (d) with the weather.

Table 1: Validation metric for total number of cases (actual vs. predicted) with weather.

Methods	MSE	RMSE	MAPE
ARIMA	**374.0833333**	**19.34123402**	**0.022592539**
Linear regression	348.6187913	18.67133609	0.021763819
SVR	310.8357811	17.63053547	0.020891829
MLP	274.3420311	274.3420311	0.019862762
RNN	190.9222324	13.81746114	0.017662732
GRU	140.0163399	11.83285003	0.01641411
LSTM	103.5873714	10.17778814	0.015592019

Table 2: Validation metric for total number of cases (actual vs. predicted) without weather.

Methods	MSE	RMSE	MAPE
ARIMA	407.1029011	20.17679115	0.025818721
Linear regression	380.0182181	19.49405597	0.023817828
SVR	330.1928918	18.17121052	0.022661666
MLP	301.18281	17.35461927	0.020109929
RNN	210.4517527	14.50695532	0.019827818
GRU	180.8178718	13.4468536	0.018989281
LSTM	**135.9179218**	**11.65838418**	**0.017928821**

Table 3: Validation metric for total number of deaths (actual vs. predicted) with weather.

Methods	MSE	RMSE	MAPE
ARIMA	2214.672	47.05316142	0.65601851
Linear regression	2167.525	46.55668588	0.625752876
SVR	2112.525	45.96221274	0.614681416
MLP	1953.075	44.20124433	0.600669098
RNN	1891.225	43.48821679	0.590657314
GRU	1797.575	42.39781834	0.544018402
LSTM	**1711.188**	**41.36423576**	**0.492211157**

Table 4: Validation metric for total number of deaths (actual vs. predicted).

Methods	MSE	RMSE	MAPE
ARIMA	2270.907	47.644517	0.666156352
Linear regression	2190.502	46.8027777	0.647128252
SVR	2135.556	46.20606021	0.639812675
MLP	1970.124	44.38468204	0.628978829
RNN	1910.327	43.70354677	0.61928292
GRU	1821.201	42.67551991	0.585261726
LSTM	**1745.017**	**41.77319715**	**0.541672819**

Based on our study results shown in Tables 1–4, it can be illustrated that weather parameters like moisture and temperature can pervade SARS-CoV-2. From the results, we can conclude that there can be elevated epidemic spreading when atmospheric temperature and humidity descend. While on the other hand, when both temperature and humidity are high, the infection rate of SARS-CoV-2 declines. In forecasting the total number of cases and total deaths with and without weather data in Pakistan, the current research illustrated comparability among deep learning models using time series models ARIMA, linear regression, SVR, MLP, RNN, GRU, and LSTM to training datasets. The present study findings elucidate the sublime performance of LSTM over other models by showing high accuracy and precision compared to other time series prediction models.

Table 5: Parameter settings of the studied approaches.

Methods	Parameters	Values
ARIMA	(p, d, q)	(1, 1, 14)
Linear regression		
SVR	C	3
	Epsilon	0.0000001
	Degree	3
	Tolerance	0.000001
MLP	Learning rate	0.004
	Time step	4
	Invisible units	64
	Training epochs	1000
RNN	Learning rate	0.005
	Time step	5
	Invisible units	16
	Training epochs	1000
GRU	Learning rate	0.0005
	Time step	5
	Invisible units	16
	Training epochs	1000
LSTM	Learning rate	0.0005
	Time step	5
	Invisible units	16
	Training epochs	1000

In this study, we focused on the number of cases and death cases from Pakistan. First, each model is trained. Then, we forecast each variable. Parameters of the constructed ARIMA, linear regression, SVR, MLP, RNN, GRU, and LSTM models based on training datasets are presented in Table 5.

6. Conclusion

The present investigation analyzed the effect of prime weather factors (temperature and humidity) on the number of reported cases and deaths due to COVID-19 in Pakistan. Different time series prediction models such as ARIMA, linear regression, SVR, MLP, RNN, GRU, and LSTM were used, and the execution of each model was analyzed in terms of MSE, RMSE, and MAPE. Results illustrated that the LSTM could better predict the COVID-19 spread as compared to other models. From the present results, we can deliberately conclude that weather holds significance in COVID-19 prediction. Thus, it is advised to wear masks and personal protective wears, keep social distancing, and continue isolation (on infection/suspect) until the temperature rises or the vaccine is fully deployed. Further, predicting the COVID-19 spread/incidence by considering other weather parameters like rainfall, wind speed, and so forth shall provide additional clues to mitigate the epidemic.

Data Availability

The data used to elaborate the results and findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

Humera Batool conceptualized the study, developed the methodology, performed formal analysis, reviewed and edited the article, performed validation, and performed visualization. Lixin Tian reviewed the article, performed supervision, and performed project administration.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants nos. 71690242, 91546118, and 11731014).

References

[1] C. Paules, H. Marston, and A. Fauci, "Infeksi coronavirus—lebih dari sekedar pilekCoronavirus Infections-More Than jold," *JAMA*, vol. 323, pp. 707-708, 2020.

[2] Q. Bukhari, J. M. Massaro, R. B. D'agostino, and S. Khan, "Effects of weather on coronavirus pandemic," *International journal of environmental research and public health*, vol. 17, p. 5399, 2020.

[3] Worldometer, *Covid-19 coronavirus pandemic*, Worldometer, Yorkville, IL, USA, 2020.

[4] K. Kupferschmidt and J. Cohen, "Can china's covid-19 strategy work elsewhere?" *Science*, vol. 367, no. 6482, pp. 1061-1062, 2020.

[5] J. Hamzelou, "World in lockdown," *New Science*, vol. 245, pp. 30611–30614, 2020.

[6] M. M. Sajadi, P. Habibzadeh, A. Vintzileos, S. Shokouhi, F. Miralles-Wilhelm, and A. Amoroso, "Temperature, humidity and latitude analysis to predict potential spread and seasonality for covid-19," *SSRN*, vol. 9, p. 3550308, 2020.

[7] M. Yousaf, S. Zahir, M. Riaz, S. M. Hussain, and K. Shah, "Statistical analysis of forecasting covid-19 for the upcoming month in Pakistan," *Chaos Solitons Fractals*, vol. 138, p. 25, 2020.

[8] L. M. Casanova, S. Jeon, W. A. Rutala, D. J. Weber, and M. D. Sobsey, "Effects of air temperature and relative humidity on coronavirus survival on surfaces," *Applied Environment Microbiology*, vol. 76, pp. 2712–2717, 2010.

[9] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus covid-19," *PloS One*, vol. 15, 2020.

[10] J. Demongeot, Y. Fleet-Berliac, H. Seligmann et al., "Temperature decreases spread parameters of the new covid-19 case dynamics," *Biology*, vol. 9, 2020.

[11] M. Sajadi, P. Habibzadeh, A. Vintzileos, S. Shokouhi, F. Miralles-Wilhelm, and A. Amoroso, "Temperature, humidity, and latitude analysis to estimate potential spread and seasonality of coronavirus disease 2019 (covid-19)," *JAMA Netw OpenJAMA network open*, vol. 3, p. 11834, 2020.

[12] K. Bloom-Feshbach, W.J. Alonso, and V. Charu, "Latitudinal variations in seasonal activity of influenza and respiratory syncytial virus (RSV): A global comparative review," *PloS One*, vol. 8, p. 14, 2013.

[13] K. Prem, Y. Liu, T. W. Russell et al., "The effect of control strategies to reduce social mixing on outcomes of the covid-19 epidemic in Wuhan, China: A modelling study," *Lancet Public Health*, vol. 5, pp. e261–e270, 2020.

[14] S.E. Eikenberry, M. Mancuso, and E. Iboi, "To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the covid-19 pandemic," *Infection Disease Model*, vol. 5, pp. 293–308, 2020.

[15] National Institute of Health, Islamabad, Pakistan, https://www.nih.org.pk/, 2020.

[16] J. Contreras, R. Espinola, F.J. Nogales, and A.J. Conejo, "Arima models to predict next-day electricity prices," *IEEE transactions on power systems*, vol. 18, pp. 1014–1020, 2003.

[17] R. Adhikari and R.K. Agrawal, "An introductory study on time series modelling and forecasting," 2013, http://arxiv.org/abs/1302.6613.

[18] H. Drucker, C.J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, pp. 155–161, 1996.

[19] D.S. Hui, I.A. E, T.A. Madani et al., "The continuing 2019-ncov epidemic threat of novel coronaviruses to global health - the latest 2019 novel coronavirus outbreak in Wuhan, China," *International Journal of Infectious Diseases*, vol. 91, pp. 264–266, 2020.

[20] Y. Bengio, *Learning Deep Architectures for AI*, Now Publishers Inc, The Netherlands, 2009.

[21] A. Graves, "Generating sequences with recurrent neural networks," 2013, http://arxiv.org/abs/1308.0850.

[22] A. Zeroual, F. Harrow, A. Dairi, and Y. Sun, "Deep learning methods for forecasting covid-19 time-series data: A comparative study," *Chaos Solitons FractalsChaos, solitons, and fractals*, vol. 140, p. 15, 2020.

[23] F. Harrow, F. Kadri, and Y. Sun, "Forecasting of photovoltaic solar power production using lstm approach," *Advanced Statistical Modeling, Forecasting, and Fault Detection in Renewable Energy Systems*, Springer, Berlin, Germany, 2020.

[24] A.S. Ashour, A. Attar, N. Dey, H. Abd Elkader, and M. Elnaby, "Long short term memory based patient- dependent model for fog detection in Parkinson's disease," *Pattern recognition letters*, vol. 131, 2019.

[25] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, http://arxiv.org/abs/1406.1078.

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modelling," 2014, http://arxiv.org/abs/1412.3555.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computing*, vol. 9, pp. 1735–1780, 1997.

[28] G. Forbes, S. Massie, and S. Craw, "Fall prediction using behavioural modelling from sensor data in smart homes," *Artificial Intelligence Review*, vol. 53, pp. 1071–1091, 2020.

[29] R. Law, G. Li, D. Fong, and X. Han, "Tourism demand forecasting: A deep learning approach," *Annals of Tourism Research*, vol. 75, pp. 410–423, 2019.

[30] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transition Neural Network*, vol. 5, pp. 157–166, 1994.

[31] T. Trappenberg, *Machine learning with sklearn*, pp. 38–65, Oxford University Press, Oxford, UK, 2019.

[32] W. M. Lee, *Getting Started with Scikit-learn for Machine Learning*, pp. 93–117, John Wiley & Sons, Inc. Hoboken, NJ, USA, 2019.