An assessment of meteorological parameters effects on COVID-19 pandemic in Bangladesh using machine learning models

Jaionto Karmokar1 · Mohammad Aminul Islam1 · Machbah Uddin1 · Md. Rakib Hassan1 · Md. Sayeed Iftekhar Yousuf1

Abstract
Coronavirus (COVID-19) is a highly contagious virus (SARS-CoV-2) that has caused a global pandemic since January 2020. Scientists around the world are doing extensive research to control this disease. They are working tirelessly to find out the origin and causes of the disease. Several studies and experiments mentioned that there are some meteorological parameters which are highly correlated with COVID-19 transmission. In this work, we studied the effects of 11 meteorological parameters on the transmission of COVID-19 in Bangladesh. We first applied statistical analysis and observed that there is no significant effect of these parameters. Therefore, we proposed a novel technique to analyze the insight effects of these parameters by using a combination of Random Forest, CART, and Lasso feature selection techniques. We observed that 4 parameters are highly influential for COVID-19 where $T_{min}$ and Cloud have positive association whereas WS and AQ have negative impact. Among them, Cloud has the highest positive impact which is 0.063 and WS has the highest negative association which is $-0.021$. Moreover, we have validated our performance using DLNM technique. The result of this investigation can be used to develop an alert system that will assist the policymakers to know the characteristics of COVID-19 against meteorological parameters and can impose different policies based on the weather conditions.

Keywords Bangladesh · COVID-19 · Meteorological parameters · Random Forest · CART · Lasso · DLNM

Introduction
Coronavirus (COVID-19) was unknown to the world before December 2019 and this epidemic was first reported in the city of Wuhan in Hubei Province, China (Li et al. (2020)). Then, it rapidly spread to 222 countries in the world1. Due to the extreme infectious nature of COVID-19 and its quick spreadability, the World Health Organization (WHO) declared this disease as a pandemic on 11 March 20202. As of 31 January 2022, the total number of confirmed cases and deaths was 379,589,352 and 5,697,638 respectively.1 Besides, the current active cases (number of infected patients) are 74,198,207 all over the world1. In Bangladesh, the infection of COVID-19 was first introduced in 8 March 2020 and the number of new cases and deaths increased rapidly from April 2020. According to IEDCR (Institute of Epidemiology, Disease Control and Research) Bangladesh, the number of confirmed cases and deaths is 17,98,833 and 28,394 respectively (until 31 January 2022).

Different researchers (Liu et al. (2020); Ahmadi et al. (2020); Travaglio et al. (2021); Xie and Zhu (2020)) observed that meteorological parameters play a vital role in spreading or suppressing the COVID-19 disease. Prata et al. (2020); Shi et al. (2020); Ma et al. (2020); Abdelhafez

1 Department of Computer Science and Mathematics, Bangladesh Agricultural University, Mymensingh 2202, Bangladesh

1 https://www.worldometers.info/coronavirus/

2 https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf
et al. (2021) found strong association between meteorological parameters and COVID-19 infection. Bherwani et al. (2020) conducted experiments to analyze the impact of environmental parameters (temperature and relative humidity) on the COVID-19 outbreak in India. They found that city temperature has a moderate linkage with the number of daily cases in cities. Araujo and Naimi (2020) reported that the spread of the virus is favorable in cool and dry weather. Bu et al. (2020) reported that 13–24°C temperature, 50–80% humidity, and 30 mm/month rainfall are the prime factors for the COVID-19 outbreak in China. Moreover, variations of temperature and humidity may be the important factors affecting the COVID-19 mortality in Wuhan, China (Ma et al. (2020)). Pani et al. (2020) reported that mean temperature and relative humidity had a positive linear relationship with daily cases in Singapore but air pressure has an inverse relationship. Chan et al. (2011) suggested that at temperatures range (22–25°C) and relative humidity range (40–50%), the virus was retained on smooth surfaces over 5 days. They said that the sustainability of this virus was quickly vanished at high temperatures (>38°C) and high relative humidity (>95%). Gupta et al. (2020) suggested that temperature had a huge effect on faster transmission of COVID-19 under certain conditions in the USA. Rosario et al. (2020) found a negative relationship between climate factors (maximum temperature, average temperature, wind speed, and solar radiation) and spread of the virus in Brazil. Alongside, humidity and solar radiation had an inverse relationship with infection rate in Iran (Ahmadi et al. (2020)). However, Bhattacharjee (2020) found that there is no strong connection between the effectiveness of virus and several environmental factors in different areas of China and Italy. Adhikari and Yin (2020) found strong correlation between Cloud and COVID-19 in New York. Nasirpour et al. (2021) also tried to find the relationship of solar activity with COVID-19.

For the COVID-19 situation in Bangladesh, Islam et al. (2021) said that temperature has positive correlations with the COVID-19 confirmed cases. They also suggested that certain range of minimum and mean temperature, wind speed, relative humidity, and absolute humidity are the most favorable conditions for the spread of COVID-19. In another study, Mofijur et al. (2020) observed that minimum temperature and mean temperature had a significant association with new cases whereas air quality had a strong negative correlation with cumulative cases in Dhaka, Bangladesh. However, the observation periods of Islam et al. (2021) and Mofijur et al. (2020) are only 81 days and 30 days respectively.

Sharifi et al. (2021) said that different artificial intelligence techniques can be used to find the relationship between meteorological parameters and COVID-19. Different studies used a distributed lag nonlinear model (DLNM) (Yuan et al. (2021)), generalized additive model (GAM) (Prata et al. (2020)), and artificial neural network (ANN) (Borghi et al. (2021)) for this purpose. Moreover, Artin et al. (2021) used neural architecture search (NAS) and linear regression to predict the traffic condition based on meteorological parameters. Inspired from the above literature, we aimed to develop a new machine learning technique based on Random Forest, classification and regression trees (CART), and Lasso feature selection technique. Therefore, in this research, we try to investigate the influence of meteorological parameters on COVID-19 transmission in Bangladesh for daily cases, deaths, and recovered over a duration from 8 March 2020 to 31 January 2022. We considered 11 meteorological parameters: minimum temperature ($T_{\text{min}}$), maximum temperature ($T_{\text{max}}$), mean temperature ($T_{\text{mean}}$), air pressure ($AP$), air quality ($AQ$), rainfall ($\text{rain}$), water vapor ($WV$), wind speed ($WS$), relative humidity ($RH$), absolute humidity ($AH$), and Cloud. We proposed a novel ensemble machine learning technique which gives us insight about the effects of meteorological parameters on COVID-19 transmission. The outputs of this experiment can be used with other non-meteorological parameters and human contact transmission factors to suppress the spreading of COVID-19 transmission.

The manuscript is organized as follows: Section 2 discusses the materials and methodology, Sect. 3 includes the results and discussion of the detail analysis, and Sect. 4 contains the conclusion of the research work.

Materials and methods

Study area

Bangladesh is a highly populated country which is situated near the Bay of Bengal. Geographically, the location of this country is between 20°34’ to 26°38’ north latitude and 88°01’ to 92°41’ east longitude. The total area of Bangladesh is 148,560 square km and the total population is 161,376,708. However, due to the hydrogeological and social-economic reasons, Bangladesh is considered one of the climatic vulnerability regions. According to BMD, there are four seasons: pre-monsoon (March to May months are hot and humid), monsoon (June to September are humid and rainy), post-monsoon (October and November are quite hot and dry), and winter (December to February are cool and dry). There are 64 districts in Bangladesh, but the regional climatic variations are negligible since all the regions of the country are horizontal. Therefore, we consider the average value of all the meteorological parameters in our dataset.

---

3 https://en.wikipedia.org/wiki/Bangladesh

4 Bangladesh Meteorological Department: http://bmd.gov.bd/
Methodology

Our proposed model is divided into three phases: data collection and dataset preparation, data normalization, and feature importance. Then, we selected the meteorological features. The workflow of our proposed model is depicted in Fig. 1. Each of the following phases is described below.

Data collection and dataset preparation

The data of this study of the COVID-19 pandemic (daily number of confirmed cases, deaths, and recovered) is collected from the official website of the Ministry of Health\(^5\), Bangladesh for the period of 695 days from 8 March 2020 to 31 January 2022. The data of meteorological parameters such as minimum temperature (\(T_{\text{min}}\)), maximum temperature (\(T_{\text{max}}\)), and rainfall (\(\text{rain}\)) are collected from BMD\(^4\), air quality (\(\text{AQ}\)) from the Air Pollution in World: Real-time Air Quality Index (AQI)\(^6\), cloud data are collected from the World Weather Online\(^7\), and wind speed (\(\text{WS}\)), relative humidity (\(\text{RH}\)), and air pressure (\(\text{AP}\)) are collected from Time and Date\(^8\). Many researchers (Islam et al. (2021); Mofijur et al. (2020)) used these sources as their dataset and they are popular and trustworthy. Moreover, the other parameters, such as absolute humidity (\(AH\)) (Herrmann and Bucksch (2014); Qi et al. (2020); Gupta et al. (2020); Pani et al. (2020)) and water vapor (\(WV\)) (Ou-Yang et al. (2014); Pani et al. (2020)) are calculated using mathematical Eqs. (1) and (2).

\begin{equation}
AH = 2.1674 \times RH \times \frac{6.112 \times e^{\left(\frac{17.67}{234.5+T}\right)}}{273.15 + T} \quad (1)
\end{equation}

\begin{equation}
WV = 6.22 \times RH \times \frac{6.112 \times e^{\left(\frac{17.67}{234.5+T}\right)}}{P} \quad (2)
\end{equation}

Here, \(AH\) is absolute humidity (\(\text{gm}^{-3}\)), \(RH\) is relative humidity (\(\%\)), \(T\) is average temperature (\(^\circ\text{C}\)), \(P\) is air pressure (\(\text{mbar}\)) respectively.

Data normalization

The data or features in our dataset are integrated from various sources and the value range of each data or feature differs from each other. To eliminate the bias of different numeric features (Singh and Singh (2020)), we need to transform the data in a common range so that the larger numeric feature values cannot dominate the smaller values. Therefore, we transform the feature range to a common scale using the data normalization technique without distorting the range differences of the features so that the relative importance of the features can vanish and each feature contributes equally to the feature selection process. In this study, we use Eq. (3) to perform the data normalization process.

\begin{equation}
x_{\text{new}} = \frac{x_{\text{old}}}{x_{\text{max}}} \quad (3)
\end{equation}

Here, \(x_{\text{new}}\) is the new value of a feature after the normalization, \(x_{\text{max}}\) is the maximum value of the feature in the dataset, and \(x_{\text{old}}\) is the present value of the feature.

Feature importance

In our dataset, we use 11 meteorological parameters and try to find the effectiveness of each parameter on the COVID-19 daily cases, deaths, and recovered. To calculate the important features, we used two different types of techniques: statistical analysis and machine learning techniques.

Statistical analysis In our experiment, two non-parametric tests (Spearman’s rank correlation and Kendall’s correlation)
are used to analyze the relationship between meteorological and COVID-19 parameters (number of confirmed cases, deaths, and recovered). Spearman’s rank correlation coefficient is used to measure the monotonic relationship between two variables which is denoted by \( r_s \) (Pani et al. (2020); Şahin (2020)). The coefficient can be estimated using Eq. (4). Negative value of \( r_s \) represents the negative correlation and positive value of \( r_s \) represents the positive correlation.

\[
r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}
\]

Here, \( r_s \) is Spearman’s rank correlation coefficient, \( d_i^2 \) is difference between the two ranks of each observation, and \( n \) is the number of observation.

Kendall’s correlation coefficient which is denoted by tau (\( \tau \)) is used to assess the ordinal association between the two variables (Pani et al. (2020)). This coefficient can be estimated by using Eq. (5). The coefficient value is ranging from \(-1\) to \(+1\).

\[
\tau = \frac{concor - discor}{0.5 \times n(n-1)}
\]

Here, \( concor \) is the number of concordant pairs, \( discor \) is the number of discordant pairs, and \( n \) is the number of pairs.

**Algorithm 1** Algorithm for finding the most influenced meteorological parameters on COVID-19

**Input:** Meteorological Parameters: \( AQ, AP, AH, RH, WV, WS, Tmax, Tmin, Tmean, Rainfall \)

**COVID-19:** Daily cases, Daily deaths and Daily recovered

**Output:** A set of important meteorological parameters

1. \( RF_{Feature} \) ← Selected features using Random Forest where feature importance score \( \geq 0.05 \)
2. \( CART_{Feature} \) ← Selected features using CART where feature importance score \( \geq 0.05 \)
3. \( Lasso_{Feature} \) ← Selected feature using Lasso from \( (RF_{Feature} \cup CART_{Feature}) \)

Machine learning techniques In this step, we propose an algorithm by using a combination of Random Forest, CART, and Lasso feature selection techniques to find the effectiveness of meteorological parameters on COVID-19 parameters. The proposed algorithm is described in Algorithm 1. The descriptions of these techniques are described in the following section.

**Random Forest feature selection technique:** Random Forest is a machine learning algorithm that is widely used for regression and classification tasks. Basically, it ensembles many decision trees on given data and generates an optimal model (Speiser et al. (2019); Chen et al. (2020)). Again, a decision tree is an algorithmic technique in which the dataset is split into a huge number of smaller subsets so that it can efficiently predict the target value. Here, each branch represents an output and nodes represent the conditions. The algorithm continues the splitting process until a specific criterion is met (Shaikhina et al. (2019)). In our proposed method, we use this technique for identifying the influence of each meteorological parameter on COVID-19 parameters. In a Random Forest algorithm, it is known as feature importance which is calculated based on the metric of decrease in node impurity (mean square error impurity) in Eq. (6) and the node probability. Again, node probability is calculated from the number of incoming edges and total sample nodes.

\[
I = \frac{1}{s} \sum_{i=1}^{s} (P_i - \mu)^2
\]

where \( s \) is the size of \( P \) and \( \mu \) is the mean value of \( P \).

Hence, the feature importance among two nodes \( L \) and \( R \) is calculated in Eq. (7).

\[
f_j = w_j \times I_j - w_{Lj} \times I_{Lj} - w_{Rj} \times I_{Rj}
\]

Here, \( w_j \) is the number of edges incoming to node \( j \), \( I_j \) is the impurity of node \( j \) calculated in Eq. (6), and \( L_j \) and \( R_j \) are the left and right child node of \( j \) node. Then, the feature importance of each feature with respect to all features is calculated in Eq. (8).

\[
s_j = \frac{\sum_{j=1}^{k} f_{ij}}{\sum_{j=1}^{k} f_{ij}}
\]

Here, \( k \) is the number of outgoing edges in node \( i \), \( x \) is the number of total incoming nodes and outgoing edges in node \( i \), and \( f_{ij} \) is the importance of \( j^{th} \) child level nodes for
\(i^{th}\) node. Later, normalized feature importance (0 to 1) is generated using Eq. (9).

\[
    n_i = \frac{s_i}{\sum_{j=1}^{n} s_j} \tag{9}
\]

where \(x\) is the number of initial features and \(s_i\) is the \(i^{th}\) feature’s importance. Thus, the final Random Forest importance is defined in Eq. (10).

\[
    r_i = \frac{\sum_{j=1}^{n} n_{ij}}{T} \tag{10}
\]

Here, \(x\) is the number of any depth sub tree generated from node \(i\), \(T\) is the number of total subtree generated by the Random Forest algorithm, and \(n_{ij}\) is the normalized feature importance for feature \(i\). Therefore, the Random Forest algorithm finds the importance matrix \(r\) for the given number of features. We can apply threshold values to select the number of features from \(r\).

**CART feature selection technique:**

Classification and Regression Trees (CART) algorithm is a popular technique of machine learning which works based on a decision tree. It is gaining more popularity for its transparent nature. In each branching of a decision tree, the CART algorithm uses the gini index score for homogeneity checking of subnodes (Zimmerman et al. 2021). It generates a large tree and then minimizes it by tree pruning technique. The CART algorithm works by calculating the gini index or gini impurity of any data which represents the degree of probabilities of being the wrong classification of a variable (Saba et al. 2021). The maximum value of gini index is 1 and the minimum is 0. Here, 0 means that all values are classified into a single class, 1 means randomly distributed where no specific class is found, and 0.5 indicates uniform distribution into several classes. The gini index is calculated using Eq. (11).

\[
g_i = 1 - \sum_{j=1}^{n} (p_j)^2 \tag{11}
\]

Here, \(n\) is the number of classes and \(p_j\) is the probability of \(j^{th}\) class.

In the case of the given dataset, the algorithm applies the greedy approach to generate a decision tree. It selects a node and splits it into different spaces, then computes the gini index for each space. The branching procedure goes until a condition is met using a recursive algorithm. Then, a cost function is used to choose the optimal path, where the cost function is defined in Eq. (12).

\[
z = \sum_{j=1}^{n} (x - y)^2 \tag{12}
\]

Here, \(n\) is the number of nodes in the decision tree, \(x\) is the given data, and \(y\) is the predicted data.

Thus, it generates a large decision tree and then it is optimized. Therefore, in our case, we provide a matrix of daily meteorological conditions as input variables and COVID-19 parameters as output variables. The CART algorithm generates a lot of subdecision trees and makes a regression for each COVID-19 parameter and ranks the impact of meteorological parameters. Thus, we find the important parameters that have huge influence on spreading the pandemic in Bangladesh.

**Lasso feature selection technique:**

The Least Absolute Shrinkage and Selection Operator technique is known as the Lasso method which is widely used for multivariate feature selection. The basic principle of Lasso is the linear regression model but it has the capability to extend with generalized linear models. At first, it shrinks the regression coefficients obtained from the regression model and uses the L1 regularization technique to regularize the model (Thanh et al. 2020; Zhang et al. 2019)). Later, it selects the non-zero elements and discards others. Lasso is very much suitable for minimizing prediction errors in comparison to other prediction methods (Cui et al. 2021)). It is calculated using Eqs. (13) to (16) (Ueno et al. 2021; Manhrawy et al. 2021)).

\[
    R(\theta_j) = \frac{1}{s} \sum_{i=1}^{s} \text{Cost}(h_y(p^{(i)},q^{(i)}) + \lambda \sum_{j=1}^{s} \text{abs}(d_j) \tag{13}
\]

where \(s\) is the size of vector \(p\) and \(q\), \(\lambda\) is the L1 regularization parameter, \(p^{(i)}\) represents a vector contains given observations of an object, \(q^{(i)}\) holds the information of predicted values, gradient descent is represented by \(d_j\) in Eq. (14), and the cost function is defined in Eq. (15).

\[
    d_j = \frac{1}{s} \sum_{i=1}^{s} 2(q_i - (m \times p_i + c))(-p_i) \tag{14}
\]

where \(m\) is the slope of \(p\) and \(q\) and \(c\) is the constant.

\[
    R(\theta_{a,b}) = \frac{1}{2s} \sum_{i=1}^{s} (w_a(p^{(i)} - q^{(i)})^2 \tag{15}
\]

where

\[
    w_a(r) = \theta_0 + \theta_1r_1 + \theta_2r_2 + \theta_3r_3 + \ldots + \theta_s r_s \tag{16}
\]

where \(s\) is the size of given observations \(p\).

To analyze the impact of meteorological parameters of COVID-19 in Bangladesh, we provide matrix \(P\) of \(X \times Y\) dimensions where \(X\) is the number of days we consider and \(Y\) is the number of meteorological parameters and \(Q\) is a dimension of \(X \times R\), the actual value of COVID-19 parameters. Therefore, the Lasso technique selects \(Z\) number
of features based on the values of $P$ and $Q$ where $Z \leq Y$. Thus, $Z$ represents the important meteorological features of COVID-19 in Bangladesh.

### Results and discussion

#### Descriptive statistics

The descriptive statistics of the meteorological and the COVID-19 parameters are shown in Table 1. The total number of confirmed cases, deaths, and recovered for 695 days is 1798793, 28399, and 1551211, respectively. Among the meteorological parameters, we observed the highest SD for AQ which is 65.66 and the minimum is 3.66 for $T_{max}$. The second highest variance was found for cloud which is 920.19. Among the temperature parameters, the highest variation was found for $T_{min}$. Also, we observed the highest range (difference between maximum and minimum) for AQ and the second highest is AP. Moreover, we found the positive skewness for Rain, AQ, WS, and Cloud.

| Parameters | Mean | SD  | Variance   | Skewness | Min   | Max   |
|------------|------|-----|------------|----------|-------|-------|
| Daily cases | 2588.19 | 3207.30 | 10286796.30 | 2.31 | 0.00 | 16230.00 |
| Daily deaths | 40.86 | 53.14 | 2823.96 | 2.50 | 0.00 | 264.00 |
| Daily recovered | 2235.17 | 2923.45 | 8546554.88 | 2.69 | 0.00 | 16627.00 |
| Tmax (°C) | 30.78 | 3.66 | 13.42 | -0.79 | 18.00 | 38.00 |
| Tmin (°C) | 23.82 | 4.36 | 19.02 | -1.00 | 10.00 | 31.00 |
| Tmean (°C) | 27.30 | 3.78 | 14.30 | -1.04 | 15.00 | 34.00 |
| Rain (mm) | 6.17 | 13.23 | 175.10 | 5.24 | 0.00 | 142.00 |
| RH (%) | 73.93 | 17.97 | 322.87 | -0.53 | 20.00 | 100.00 |
| AH (gm$^{-3}$) | 19.50 | 5.60 | 31.37 | -0.11 | 5.91 | 31.78 |
| WV (gkg$^{-1}$) | 16.73 | 4.94 | 24.38 | -0.09 | 5.08 | 27.66 |
| AQ (μgm$^{-3}$) | 148.22 | 65.66 | 4311.22 | 0.61 | 21.00 | 365.00 |
| AP (mbar) | 1007.84 | 10.38 | 107.83 | -11.87 | 814.00 | 1021.00 |
| WS (kmh$^{-1}$) | 9.39 | 6.68 | 44.62 | 0.53 | 0.00 | 34.00 |
| Cloud (%) | 38.13 | 30.33 | 920.19 | 0.07 | 0.00 | 100.00 |

#### Non-parametric test

The non-parametric test i.e. Spearman and Kendall rank correlation test between meteorological parameters and COVID-19 is shown in Table 3. From this table, we observed that the relationship between temperature parameters ($T_{max}$, $T_{min}$, and $T_{mean}$) and COVID-19 daily cases, deaths, and recovered are positive where $T_{min}$ relationship is higher than the $T_{max}$ and $T_{mean}$ which implies that $T_{min}$ favors the spread of COVID-19 confirmed cases, deaths, and also people recovered from COVID-19 and $T_{max}$ has less impact.

#### Normality test

In this step, we apply both the Kolmogorov-Smirnov test and Shapiro-Wilk test to find the normality of the dependent variables of COVID-19 parameters: daily cases, deaths, and recovered. This test is performed at 95% confidence interval level. The test result is shown in Table 2. We observed that the significance values of all the cases are less than 0.01 which implies the parameters are not normally distributed. Therefore, to find the association between meteorological parameters and COVID-19 parameters, we applied a non-parametric test (Spearman’s rank correlation and Kendall’s correlation).

| Parameters | Kolmogorov-Smirnov Statistic | df | Sig. | Shapiro-Wilk Statistic | df | Sig. |
|------------|-----------------------------|----|------|------------------------|----|------|
| Daily cases | 0.21 | 694 | 0.00 | 0.70 | 694 | 0.00 |
| Daily deaths | 0.26 | 694 | 0.00 | 0.66 | 694 | 0.00 |
| Daily recovered | 0.22 | 694 | 0.00 | 0.68 | 694 | 0.00 |
on daily recovered. The *Rain*, *AH*, and *WV* also have a positive relationship with COVID-19 and *RH* has less effect on COVID-19 where the correlation for daily cases and deaths is close to zero. On the other hand, *AQ* and *AP* have a negative correlation with COVID-19 transmission, where *AP* helps significantly to reduce the spread of COVID-19. However, for all the cases, the *p* value is less than the significant value of 0.01 and 0.05 except *WS* in daily recovered. Therefore, to find the important meteorological parameters which are highly correlated with the COVID-19 transmission, we apply a modified version of the machine learning technique by using the combination of Random Forest, CART, and Lasso feature selection techniques.

**Table 3** Empirical results of nonlinear correlation between COVID-19 and meteorological parameters

| Parameters | Kendall’s tau_b | Spearman’s rho |
|------------|-----------------|----------------|
|            | Daily cases     | Daily deaths   | Daily recovered |
|            | *τ*   | *p*  | *τ*   | *p*  | *τ*   | *p*  |
| *T* max    | .124** <0.01   | .217** <0.01  | .141** <0.01  |
| *T* min    | .348** <0.01   | .439** <0.01  | .375** <0.01  |
| *T* mean   | .258** <0.01   | .356** <0.01  | .280** <0.01  |
| *Rain*     | .284** <0.01   | .354** <0.01  | .283** <0.01  |
| *RH*       | .137** <0.01   | .116** <0.01  | .183** <0.01  |
| *AH*       | .291** <0.01   | .343** <0.01  | .335** <0.01  |
| *WV*       | .296** <0.01   | .351** <0.01  | .338** <0.01  |
| *AQ*       | -.281** <0.01  | -.337** <0.01 | -.245** <0.01 |
| *AP*       | -.328** <0.01  | -.406** <0.01 | -.320** <0.01 |
| *WS*       | .127** <0.01   | .170** <0.01  | .082** .002   |
| *Cloud*    | .338** <0.01   | .371** <0.01  | .287** <0.01  |

*Correlation is significant at the 0.01 level (2-tailed)*

*Correlation is significant at the 0.05 level (2-tailed)*

**Fig. 2** Analysis of relationship between COVID-19 and meteorological parameters using Random Forest

**Table 4** Parameters selection using Random Forest and CART feature selection techniques

| (a) *RF feature* | (b) *CART feature* | (c) *RF feature ∪ CART feature* | (d) *Lasso feature* from Column (c) |
|------------------|--------------------|---------------------------------|-----------------------------------|
| Cases            | Deaths             | Recovered                       | Cases                            | Deaths             | Recovered                       | Cases | Deaths | Recovered |
| AQ               | AQ                 | AQ                              | AQ                               | AQ                 | AQ                              | AQ    | AQ     | AQ       |
| AP               | AP                 | WV                              | AP                               | AP                 | AH                              | AP    | AQ     | AQ       |
| RH               | WS                 | WS                              | RH                               | WS                 | RH                              | RH    | WS     | WS       |
| WV               | Tmin               | Tmin                            | WV                               | Tmin               | WV                              | WV    | Tmin   | Tmin     |
| WS               | Rain               | Rain                            | WS                               | Rain               | WS                              | WS    | Tmin   | Tmin     |
| *Tmax*           | Cloud              | Cloud                           | *Tmax*                           | Cloud              | *Tmax*                          | *Tmax* | Cloud  | Cloud    |
| *Tmin*           | Rain               | Cloud                           | *Tmin*                           | Rain               | *Tmin*                          | *Tmin* | Rain   | Cloud    |
| *Rain*           | Cloud              |                                 | *Rain*                           | Cloud              |                                 | *Rain* | Cloud  | Cloud    |
| *Cloud*          |                    |                                 | *Cloud*                          |                    |                                 | *Cloud* |        |          |

![Environmental Science and Pollution Research](https://example.com/environmental-science-and-pollution-research)
In this study, we proposed a new technique to identify the non-linear relationship between the meteorological parameters and COVID-19 daily cases, daily death, and daily recovered. We observed that Cloud has a strong positive association with all COVID-19 parameters, $T_{\text{min}}$ has a positive association with daily death and recovered, and $WV$ has a positive association with daily recovered. On the other hand, $WS$ has a strong negative association with all COVID-19 parameters, $AQ$ has a negative association with daily death and recovered, $AP$ has a negative weak association with daily cases and death, and $Rain$ has a weak negative association with daily case. The results are discussed in the following three parts.

Association analysis using DLNM: To validate our results, we applied a distributed lag nonlinear model (DLNM) between meteorological parameters and COVID-19 daily cases and deaths. For the simplicity of the model, we excluded daily recovered from this experiment. This model has been used in previous studies to analyze the exposure lag response curve. In our validation, we used a Quasi-Poisson lag-based DLNM model by selecting the parameters from Fig. 4. We used a 14-day lag response due

**Discussion**

In this step, we used Algorithm 1 to find the effectiveness of the meteorological parameters on COVID-19 transmission (daily confirmed cases, deaths and recovered). At first, we analyzed the effects of meteorological parameters and COVID-19 transmission using the Random Forest technique. The result of this experiment is shown in Fig. 2. From this figure, we observed that Cloud helps to spread COVID-19 whereas Rain, $WS$, and $AP$ have negative association with COVID-19, i.e., these parameters suppress the spreading of COVID-19. For daily cases, the negative relationship of $WS$ is the highest, whereas for daily deaths and recovered, we find a strong negative relationship with $WS$. On the other hand, Cloud and $T_{\text{min}}$ have a considerable positive relationship with daily death and recovered.

Machine learning techniques

In this study, we proposed a new technique to identify the non-linear relationship between the meteorological parameters and COVID-19 daily cases, daily death, and daily recovered. We observed that Cloud has a strong positive association with all COVID-19 parameters, $T_{\text{min}}$ has a positive association with daily death and recovered, and $WV$ has a positive association with daily recovered. On the other hand, $WS$ has a strong negative association with all COVID-19 parameters, $AQ$ has a negative association with daily death and recovered, $AP$ has a negative weak association with daily cases and death, and $Rain$ has a weak negative association with daily case. The results are discussed in the following three parts.

Association analysis using DLNM: To validate our results, we applied a distributed lag nonlinear model (DLNM) between meteorological parameters and COVID-19 daily cases and deaths. For the simplicity of the model, we excluded daily recovered from this experiment. This model has been used in previous studies to analyze the exposure lag response curve. In our validation, we used a Quasi-Poisson lag-based DLNM model by selecting the parameters from Fig. 4. We used a 14-day lag response due

![Fig. 3 Analysis of relationship between COVID-19 and meteorological parameters using CART](image)

![Fig. 4 Feature importance using Lasso feature selection technique](image)
to the 2–14 days incubation period of COVID-19 (Linton et al. (2020); Runkle et al. (2020)). The result of this experiment is shown in Fig. 5. From this figure, for daily case, we observed that the highest risk 7.995 for $T_{min}$ was obtained at lag 0 when temperature was 10°C. For daily death, the highest risk 1.361 was observed at lag 0 when the temperature is 10°C. The $Cloud$ parameter has a strong correlation with COVID-19 case where for all lags, the risk factor is around 1.0. In the case of daily death, we found the highest risk 1.49 for 80% $cloud$ at lag 0. As for $rain$, the highest risk for daily case is 1.1 at lag 0 when $rain$ is 35 mm and for death, the relative risk is 1.01 at lag 0 when $rain$ is 22 mm. When $AQ$ was 350 $\mu g/m^3$ at lag 14, the relative risk for daily case was highest at 3.6. For daily death, the highest relative risk was 1.8 when $AQ$ was 350 $\mu g/m^3$ at lag 14. As for $AP$, for both daily case and death, we found the relative risk is almost 0.

In comparison with machine learning selected parameters (Fig. 4) and DLNM (Fig. 5), we can see that the results are almost similar for both the experiments. Therefore, we can easily say that the impact analysis results from our machine learning technique are robust and effective.

**Relevance of analysis results with existing studies:**
Our proposed model not only gives a novel insight into the association of meteorological parameters with COVID-19 but also provides the important meteorological parameters which have a major influence on the transmission of COVID-19 in Bangladesh. We found the meteorological parameters such as minimum temperature, wind speed, air quality, and air pressure play a vital role in the transmission of COVID-19. Tobias and Molina (2020) said that temperature plays a vital role in the contamination and controlling the virus as the temperature is highly correlated with the human living system.

In our experiment, we found that minimum temperature ($T_{min}$) has a positive relationship with the daily death and recover. It was also found as a potential risk factor by Graudenz et al. (2006). Because rapid changes in minimum temperature affect the respiratory epithelium and increase the risk of respiratory coronavirus (Graudenz et al. (2006)). Similar to our study, Abdelhafez et al. (2021); Liu et al. (2020); Pani et al. (2020); Islam et al. (2021); Mofijur et al. (2020); Tobías and Molina (2020) also found that minimum temperature is more favorable to the spread of COVID-19. Meanwhile, Méndez-Arriaga (2020) found the negative association between temperature and daily cases in the Mexico capital city, because during the experimental period, the minimum temperature of study area was 0.87 to 17.79°C with an average of 9.97°C which is completely different from Bangladesh.

In addition, we also found that wind speed (WS) is a vital factor in COVID-19 virus transmission and has a negative association with it. Similarly, negative association between WS and COVID-19 transmission was found in Singapore (Pani et al. (2020)) and Iran (Ahmadi et al. (2020)). On the other hand, in Norway (Menebo (2020)), positive correlation was found between wind speed and COVID-19 case, because
in their study period the wind speed was 1.3 to 6.6 with an average of 3.32.

Cloud has a strong positive relationship with COVID-19 transmission. We know that cloud always has a negative relationship with WS (Adhikari and Yin (2020)). Similar to our study, Adhikari and Yin (2020) also found a positive relationship of Cloud with COVID-19.

Moreover, we found that air quality ($AQ$) has a negative correlation with the COVID-19 daily death and recovered. Similar to our study, Mofijur et al. (2020) also found a negative correlation with AQ. Because their study area is also Bangladesh. We also found that there is a weak negative correlation between AP with daily cases and death. Evidence from Islam et al. (2021) also suggested that AP has an inverse association with COVID-19 confirmed cases which is similar to our results because we analyzed the impact of meteorological parameters on COVID-19 in Bangladesh.

**Impact of lockdown and government policies:** We analyzed the lockdown impact on COVID-19 daily case and death. The result is shown in Figs. 6 and 7. We observed that the Government of Bangladesh imposed 4 lockdown periods in different intervals. After the lockdown period, the number of daily deaths and cases reduced significantly. Therefore, we can easily say that the policy of the Government of Bangladesh has a big impact in controlling the spread of COVID-19. Along with the meteorological parameters, we have to maintain the WHO guidelines such as keeping at least 3 feet distance (social distance) during face to face conversations; wearing a mask by covering the nose, mouth, and chin; avoid mass gathering; avoid touching the eyes, nose, and mouth; and cleaning hands regularly using an alcohol-based hand sanitizer, etc.

The findings of this research can be used to develop a meteorological-based warning or alert system. For example, we found that air quality has a significant reverse effect on raising the COVID-19 transmission in Bangladesh. However, if the air quality is forecasted to be going down, the government can warn the people, especially two vulnerable groups of people (children and elderly group) about taking the necessary precautions of COVID-19.

Our study has the following limitations. Firstly, we do not count the population density and health care facility which are also an important factor in COVID-19 transmission. Secondly, we do not count whether the people maintained social distancing and human isolation (Bodrud-Doza et al. (2020)). Thirdly, we do not know whether those who have been affected have received adequate health care facilities, which is responsible for the increase in daily death records (Shammi et al. (2021)). It would be interesting to consider the daily testing rate, other non-environmental factors, COVID-19 vaccination, and human-serum antibody levels to find the important parameters which might be influential in the COVID-19 transmission.

**Conclusion**

In this study, we proposed a new technique to identify the important meteorological parameters which have major impacts on COVID-19 transmission. We first analyzed the effects of these parameters using statistical analysis. We found that the impacts of these parameters are not very significant using statistical analysis. Therefore, we proposed a new technique that gives a novel insight into the correlation between the meteorological parameters with COVID-19 transmission. We observed that cloud has a positive association with COVID-19, $T_{\text{min}}$ has a positive association with daily death and recovered, WS has a negative association with COVID-19, $AQ$ has a negative association with daily death and recovered, AP has a weak negative relation with daily case and death, and Rain has a negative relation with daily case. A decrease in the minimum temperature value and an increase in the air pressure, air quality, and wind
speed value can significantly suppress the spreading of the COVID-19 transmission. We also observed that for daily cases, the significance of Cloud is larger than the other meteorological parameters. However, the significance of minimum temperature is larger in daily death compared to the daily recovered. This research can be useful to develop an alert system for the public health care decision-makers for the general people in Bangladesh. However, the research is limited to not including data from vaccination programs, population density, human immunity system, and zone-based meteorological parameters. However, effective measures of these meteorological parameters and maintaining the necessary precautions of COVID-19 protocols such as maintaining the social distancing, taking the vaccine, individual hygiene, hand washing, and utilization of hand sanitizer should be taken to suppress the COVID-19 outbreaks and also the transmission.

The contribution of this manuscript are as follows:

1. We consider 11 meteorological parameters that may influence the COVID-19 explosion whereas other studies considered 2–6 parameters.
2. We apply a brand new technique based on combined machine learning algorithms for identifying the individual impact of meteorological parameters.
3. We develop a new and efficient algorithm (Algorithm 1) for combining three feature selection techniques that can measure the insight effects of parameters on COVID-19.
4. We conduct rigorous analysis in two folds: statistical and machine learning approaches. Among the two, we find that the statistical approach is not suitable for this targeted problem.
5. We also validated our model’s performance using DLNM.
6. By utilizing the selected parameters in Table 4, any prediction algorithm can provide the best estimation of the daily COVID-19 scenario in Bangladesh. That indicates the precision level of our proposed system.

Author Contributions Mr. Jaiont Karmokar collected the dataset from different sources as well as conducted experimental analysis and manuscript writing. Mr. Mohammad Aminul Islam developed the machine learning algorithms. He also conducted the experiments, analyzed the data, and drafted the manuscript. Mr. Machbub Uddin helped in organizing the manuscript as well as experimental analysis. Professor Dr. Md. Rakib Hassan and Mr. Md. Sayeed Iftekhar Yousuf provided the extensive review and produced the final draft of the article.

Data availability On request.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Abdelhafez E, Dabbour L, Hamdan M (2021) The effect of weather data on the spread of COVID-19 in Jordan. Environ Sci Pollut Res Int 28(30):40416–40423. https://doi.org/10.1007/s11356-020-12338-y

Adhikari A, Yin J (2020) Short-term effects of ambient ozone, pm2.5, and meteorological factors on COVID-19 confirmed cases and deaths in Queens, New York. Int J Environ Res Public Health 17(11):4047

Ahmadi M, Sharifi A, Dorosti S et al (2020) Investigation of effective climatology parameters on COVID-19 outbreak in Iran. Sci Total Environ 729(138):705

Araujo MB, Naimi B (2020) Spread of SARS-COV-2 coronavirus likely to be constrained by climate. MedRxiv

Arin J, Valizadeh A, Ahmadi M, Kumar SAP, Sharifi A (2021) Presentation of a Novel Method for Prediction of Traffic with Climate Condition Based on Ensemble Learning of Neural Architecture Search (NAS) and Linear Regression. Complexity. https://doi.org/10.1155/2021/8500572

Bhattarcharjee S (2020) Statistical investigation of relationship between spread of coronavirus disease (COVID-19) and environmental factors based on study of four mostly affected places of China and five mostly affected places of Italy. arXiv preprint arXiv:200311277

Bherwani H, Gupta A, Anjum S et al (2020) Exploring dependence of COVID-19 on environmental factors and spread prediction in India. npj Climate Atmos Sci 3(1):1–13

Bodrud-Doza M, Shammi M, Bahlman L et al (2020) Psychosocial and socio-economic crisis in Bangladesh due to COVID-19 pandemic: a perception-based assessment. Front Public Health 8:341

Borghi PH, Zakordonets O, Teixeira JP (2021) A COVID-19 time series forecasting model based on MLP ANN. Proc Comput Sci 181:940–947

Bu J, Peng DD, Xiao H et al (2020) Analysis of meteorological conditions and prediction of epidemic trend of 2019-NCOV infection in 2020. MedRxiv

Chan KH, Peiris JS, Lam SY, Poon LL, Yuen KY, Seto WH (2011) The Effects of Temperature and Relative Humidity on the Viability of the SARS Coronavirus. Adv Virol. https://doi.org/10.1155/2011/734690

Chen W, Li Y, Xue W et al (2020) Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and random forest methods. Sci Total Environ 701(134):979

Cui L., Bai L., Wang Y., Yu PS., Hancock. ER (2021) Fused lasso for feature selection using structural information. Pattern Recognition. https://doi.org/10.1016/j.patcog.2021.108058

Graudenz GS, Landgraf RF, Janzar S et al (2006) The role of allergic rhinitis in nasal responses to sudden temperature changes. J Allergy Clin Immunol 118(5):1126–1132

Gupta S, Raghuwanshi GS, Chanda A (2020) Effect of weather on COVID-19 spread in the US: a prediction model for India in 2020. Sci Total Environ 728(138):860

Herrmann H, Buchsch H (2014) Clausius-Clapeyron equation. Dictionary of Geotechnical Engineering/Wörterbuch GeoTechnik. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41714-6
Islam ARMT, Hasanuzzaman M, Azad MAK et al (2021) Effect of meteorological factors on COVID-19 cases in Bangladesh. Environ Dev Sustain 23(6):9139–9162

Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q, Zhou S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi G, Lam TTY, Wu JT, Gao GF, Cowling BJ, Yang B, Leung GM, Feng Z (2020) Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med 382(13):1199–1207. https://doi.org/10.1056/NEJMoa2001316

Linton NM, Kobayashi T, Yang Y et al (2020) Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. J Clin Med 9(2):538

Liu J, Zhou J, Yao J et al (2020) Impact of meteorological factors on the COVID-19 transmission: a multi-city study in China. Sci Total Environ 726(138):513

Ma Y, Zhao Y, Liu J et al (2020) Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. Sci Total Environ 724(138):226

Manhrawi Il, Qaraad M, El-Kafrawy P (2021) Hybrid feature selection model based on relief-based algorithms and regularizer algorithms for cancer classification.Concurrency and Computation: Practice and Experience 33(17):e6200

Méndez-Arriaga F (2020) The temperature and regional climate effects oncommunitarian COVID-19 contagion in Mexico throughout phase 1. Sci Total Environ 735(139):560

Menebo MM (2020) Temperature and precipitation associate with COVID-19 new daily cases: a correlation study between weather and COVID-19 pandemic in Oslo, Norway. Sci Total Environ 737(139):659

Mohijur M, Rizwanul Fattah I, Saiful Islam A et al (2020) Relationship between weather variables and new daily COVID-19 cases in Dhaka, Bangladesh. Sustainability 12(20):8319

Nasirpour MH, Sharifi A, Ahmadi M et al (2021) Revealing the relationship between solar activity and COVID-19 and forecasting of possible future viruses using multi-step autoregression (MSAR). Environ Sci Pollut Res 28(28):38,074–38,084

Ou-Yang CF, Lin NH, Lin CC et al (2014) Characteristics of atmospheric carbon monoxide at a high-mountain background station in East Asia. Atmos Env 89:613–622

Pani SK, Lin NH, RavindraBabu S (2020) Association of COVID-19 pandemic with meteorological parameters over Singapore. Sci Total Environ 740(140):112

Prata DN, Rodrigues W, Bermejo PH (2020) Temperature significantly changes COVID-19 transmission in (SUB) tropical cities of Brazil. Sci Total Environ 729(138):862

Qi L, Gao Y, Yang J et al (2020) The burden of influenza and pneumonia mortality attributable to absolute humidity among elderly people in Chongqing, China, 2012–2018. Sci Total Environ 716(136):682

Rosario DK, Mutz YS, Bernardes PC et al (2020) Relationship between COVID-19 and weather: case study in a tropical country. Int J Hyg Environ Health 229(113):587

Runkle JD, Sugg MM, Leeper RD et al (2020) Short-term effects of specific humidity and temperature on COVID-19 morbidity in select US cities. Sci Total Environ 740(140):693

Saba T, Abunadi I, Shahzad MN, Khan AR (2021) Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types. Micros Research Tech 84(7):1462–1474. https://doi.org/10.1002/mret.23702

Şahin M (2020) Impact of weather on COVID-19 pandemic in Turkey. Sci Total Environ 728(138):810

Shaikhina T, Lowe D, Daga S et al (2019) Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomed Signal Process Control 52:456–462

Shammi M, Bodrud-Doza M, Islam ARMT et al (2021) Strategic assessment of COVID-19 pandemic in Bangladesh: comparative lockdown scenario analysis, public perception, and management for sustainability. Environ Dev Sustain 23(4):6148–6191

Sharifi A, Ahmadi M, Ala A (2021) The impact of artificial intelligence and digital style on industry and energy post-COVID-19 pandemic. Environ Sci Pollut Res 28(34):46964–46984

Shi P, Dong Y, Yan H et al (2020) Impact of temperature on the dynamics of the COVID-19 outbreak in China. Sci Total Environ 728(138):890

Singh D, Singh B (2020) Investigating the impact of data normalization on classification performance. Appl Soft Comput 97(105):524

Speiser JL, Miller ME, Tooze J et al (2019) A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl 134:93–101

Thanh DNH, Hien NN, Prasath S et al (2020) Adaptive total variation II regularization for salt and pepper image denoising. Optik 208(163):677

Tobías A, Molina T (2020) Is temperature reducing the transmission of COVID-19? Environ Res 186(109):553

Travaglio M, Yu Y, Popovic R et al (2021) Links between air pollution and COVID-19 in England. Environ Pollut 268(115):859

Ueno D, Kawabe H, Yamasaki S et al (2021) Feature selection for RNA cleavage efficiency at specific sites using the lasso regression model in Arabidopsis thaliana. BMC Bioinf 22(1):1–17

Xie J, Zhu Y (2020) Association between ambient temperature and COVID-19 infection in 122 cities from China. Sci Total Environ 724(138):201

Yuan J, Wu Y, Jing W et al (2021) Association between meteorological factors and daily new cases of COVID-19 in 188 countries: a time series analysis. Sci Total Environ 780(146):538

Zhang H, Wang J, Sun Z et al (2019) Feature selection for neural networks using group lasso regularization. IEEE Trans Knowl Data Eng 32(4):659–673

Zimmerman RK, Nowalk MP, Bear T et al (2021) Proposed clinical indicators for efficient screening and testing for COVID-19 infection using classification and regression trees (CART) analysis. Hum Vaccin Immunother 17(4):1109–1112