Computers, Materials & Continua
DOI:10.32604/cmc.2021.013067
Article

Tech Science Press

Prediction of COVID-19 Cases Using Machine Learning for Effective Public Health Management

Fahad Ahmad[1,*], Saleh N. Almuayqil[2], Mamoona Humayun[2], Shahid Naseem[3], Wasim Ahmad Khan[4] and Kashaf Junaid[5]

[1]Department of Computer Sciences, Kinnaird College for Women, Lahore, 54000, Pakistan
[2]Department of Information Systems, College of Computer and Information Sciences, Jouf University, Sakaka, Aljouf, 72341, Saudi Arabia
[3]Division of Computer Science & Information Technology, University of Education, Lahore, 54000, Pakistan
[4]School of Computer Science, National College of Business Administration & Economics, Lahore, 54000, Pakistan
[5]Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, Jouf University, Sakaka, Aljouf, 72341, Saudi Arabia
*Corresponding Author: Fahad Ahmad. Email: fahad.ahmad@kinnaird.edu.pk; drfahadahmadmian@gmail.com

Received: 24 July 2020; Accepted: 30 September 2020

Abstract: COVID-19 is a pandemic that has affected nearly every country in the world. At present, sustainable development in the area of public health is considered vital to securing a promising and prosperous future for humans. However, widespread diseases, such as COVID-19, create numerous challenges to this goal, and some of those challenges are not yet defined. In this study, a Shallow Single-Layer Perceptron Neural Network (SSLPNN) and Gaussian Process Regression (GPR) model were used for the classification and prediction of confirmed COVID-19 cases in five geographically distributed regions of Asia with diverse settings and environmental conditions: namely, China, South Korea, Japan, Saudi Arabia, and Pakistan. Significant environmental and non-environmental features were taken as the input dataset, and confirmed COVID-19 cases were taken as the output dataset. A correlation analysis was done to identify patterns in the cases related to fluctuations in the associated variables. The results of this study established that the population and air quality index of a region had a statistically significant influence on the cases. However, age and the human development index had a negative influence on the cases. The proposed SSLPNN-based classification model performed well when predicting the classes of confirmed cases. During training, the binary classification model was highly accurate, with a Root Mean Square Error (RMSE) of 0.91. Likewise, the results of the regression analysis using the GPR technique with Matern 5/2 were highly accurate (RMSE = 0.95239) when predicting the number of confirmed COVID-19 cases in an area. However, dynamic management has occupied a core place in studies on the sustainable development of public health but dynamic management depends on proactive strategies based on statistically verified approaches, like Artificial Intelligence (AI). In this study, an SSLPNN model has been trained to fit public health associated data into an appropriate class, allowing GPR to predict the number of confirmed COVID-19 cases in an area based on the given values of selected parameters. Therefore, this tool can help authorities in different ecological settings effectively manage COVID-19.

Keywords: Public health; sustainable development; artificial intelligence; SARS-CoV-2; shallow single-layer perceptron neural network; binary classification; gaussian process regression

1 Introduction

In December 2019, an outbreak of pneumonia of unknown etiology was noticed in Wuhan City, China, which later spread across the globe. In January 2020, the cause of this pneumonia-like disease was confirmed to be a novel coronavirus known as SARS-CoV-2 [1]. This virus belongs to Coronaviridae, a large family of enveloped single-stranded RNA viruses [2]. Coronaviruses are well known to cause a variety of diseases, from the common cold to significant epidemics, like severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) [1–3].

In March 2020, the World Health Organization (WHO) classified COVID-19 as a pandemic that could threaten millions of people all over the world [4]. Ever since, the number of confirmed cases has increased, partially because this new viral disease is highly contagious during the incubation period. Asymptomatic individuals infected with COVID-19 can spread the disease throughout their communities [5]. Thus, asymptomatic carriers can play a significant role in related viral infections, such as rhinovirus and the influenza virus [6,7]. In addition, there is no antiviral drug or vaccine available against this virus. Therefore, molecular testing is the most reliable diagnostic test for COVID-19.

COVID-19 poses a significant challenge for governments. Though stakeholders have dedicated many resources to fight it, the epidemic has nevertheless caused a social and economic crisis in both developed and developing countries. During the present crisis, it is important to understand how to maintain sustainability practices with limited resources so that long-term public health outcomes can still be achieved. Sustainable development depends on the cooperation of stakeholders across social, ecological, cultural, and political domains. The current challenges of COVID-19 have caused mortality and morbidity on a massive scale, directly or indirectly influencing all these domains. After the emergency declaration from the WHO, all trade and travel were banned, which led to social unrest and devastating economic consequences. In the past, the Ebola, influenza, SARS, and HIN1 epidemics caused almost US $10 billion in losses. The current crisis is similar in nature to what occurred during the SARS epidemic and may have worse consequences; if the spread of the epidemic continues as it has, the worldwide losses are projected to exceed US $150 billion [8].

As the situation worsens, relevant tools based on artificial intelligence (AI) need to be studied; a machine learning process uses big data for pattern recognition, explanation, and prediction based on input data [9,10]. Therefore, AI has the potential to design tools to fight COVID-19. In this study, we utilized SSLPNN and GPR to predict the classes to which specific case studies belonged and the number of confirmed COVID-19 cases in specific geographical areas. Though, climatic and socio-economic conditions have a strong relationship with the incidence and spread of infectious diseases [11,12]. Nevertheless, this analysis will help to design public health policies to manage sustainable development policies.

2 Materials and Methods

This study was designed to predict the number of COVID-19 cases based on environmental and non-environmental factors. We used two different approaches. First, we analyzed the correlations between the confirmed cases (from February 1, 2020 to April 20, 2020) and several environmental factors (temperature, humidity, wind speed, ultraviolet (UV) index, elevation, air quality index and pollution level) and non-environmental factors (population, population density, gender ratio, and human development index). Second, we built a binary classification model to predict and classify COVID-19 cases using an SSLPNN algorithm based on critical factors related to sustainable development in the area of public health. These factors were divided into two significant modules: the first was the non-environmental module, and the second was the environmental module. Both modules were used as the inputs, with the number of confirmed COVID-19 cases designated as the outputs. The study design is presented in Fig. 1.

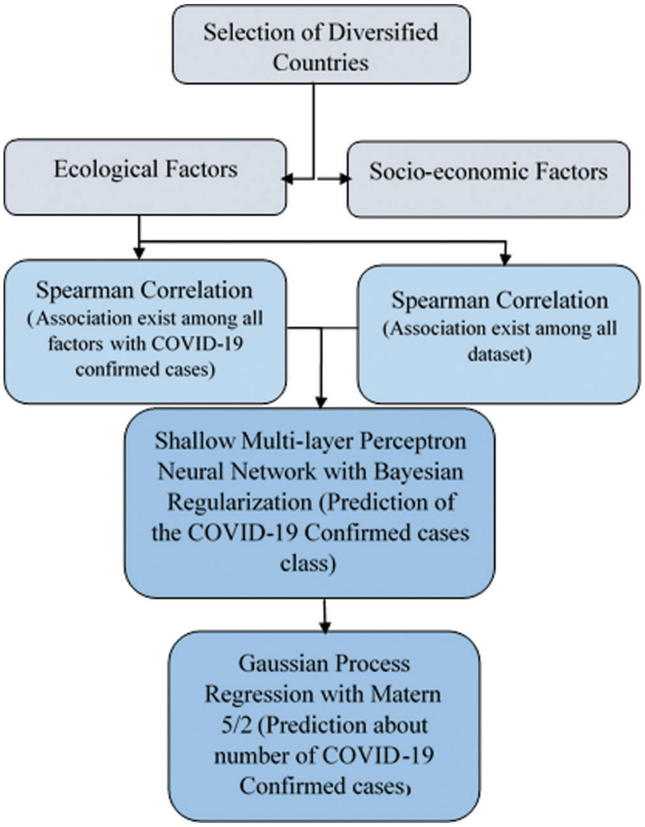

Selection of Diversified Countries
Ecological Factors
Socio-economic Factors
Spearman Correlation (Association exist among all factors with COVID-19 confirmed cases)
Spearman Correlation (Association exist among all dataset)
Shallow Multi-layer Perceptron Neural Network with Bayesian Regularization (Prediction of the COVID-19 Confirmed cases class)
Gaussian Process Regression with Matern 5/2 (Prediction about number of COVID-19 Confirmed cases)

Figure 1: Artificial intelligence–based framework for the prediction of confirmed COVID-19 cases

2.1 Conditions for Analysis

In the analysis, specific conditions were applied. These conditions included the following:

- In addition to the number of COVID-19 cases, 14 different environmental and non-environmental variables were used, including temperature (minimum, maximum, and average), humidity, wind speed, air quality index, UV index, pollution level, population, population density, gender ratio, average age, and human development index levels.
- To enhance the precision of the estimates and to reduce bias, different countries were considered due to their different topographical, monetary, and ecological situations.
- The environmental data used in this study was based on the capital cities of the selected regions, as these regions generally had larger populaces.
- The non-environmental data used in this study was also taken from the regions of the respective countries.

- The analysis period was from February 1, 2020 to April 20, 2020.
- Different countries of the Asian continent were selected for this study.

2.2 Data Collection

In this study, data was collected from the various official and independent websites of the selected countries, which were China, South Korea, Japan, Pakistan, and Saudi Arabia [13–23]. These countries were selected due to their diverse climatic conditions. The details of this dataset are provided in the supplementary file.

2.3 Correlation Analysis

To see the relationships between the total confirmed COVID-19 cases in the 54 provinces of the five countries included in this study and the 14 environmental and non-environmental variables, a correlation analysis was performed.

2.4 Spearman's Rank Correlation

Before building the model, it was necessary to evaluate the correlations between each independent dataset. For this purpose, a Spearman's correlation analysis was done on the non-parametric dataset. In a non-parametric dataset, the population data usually does not have a normal distribution and is randomly distributed vertically and horizontally. For this test, the selected parameters (environmental and non-environmental) and the total number of confirmed COVID-19 cases were included. This test was conducted to reveal the associations between two different variables without considering the distribution of the data, which is highly recommended for a dataset with at least ordinal scale. The relationships among the non-parametric variables are represented by parallel plot in Fig. 2. The mathematical formulation of Spearman's rank correlation can be represented by the following equation:

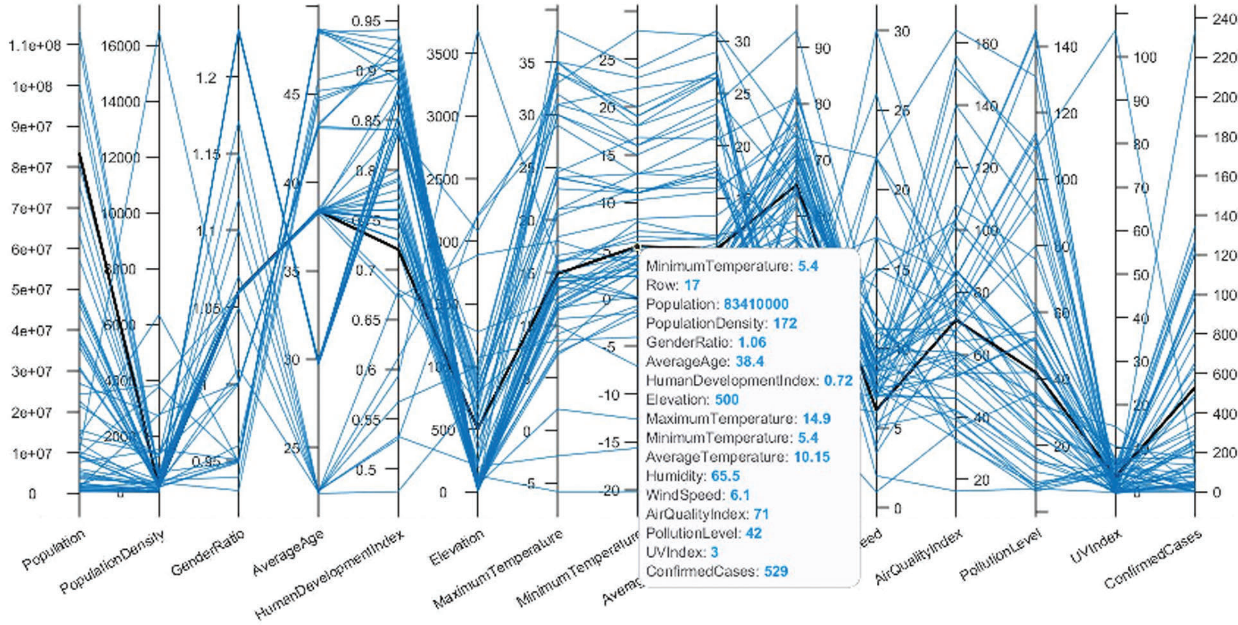


Figure 2: Parallel plot depicting the relationships among the non-parametric environmental & non-environmental variables

$$\delta = 1 - \frac{6\sum s_i^2}{m(m^2 - 1)} \tag{1}$$

δ = Spearman's rank correlation

s_i = the transformation between the ranks of corresponding parameters

m = the number of values

2.5 Shallow Single-Layer Perceptron Neural Networks (SSLPNN)

AI embraces a wide variety of approaches and algorithms based on machine intelligence. It has numerous applications in innumerable areas of science, encompassing fuzzy logic theory, machine learning techniques, risk valuations and hazard detection, meta-heuristic algorithms and classification, and clustering techniques [24]. SSLPNN is a type of neural network that has a smaller number of hidden layers and can be used for pattern recognition. While studies have shown that a shallow network can fit any function in the identification of patterns and prediction of problems, it is also considered a less complex artificial neural network. Though the use of deep learning is rapidly increasing in different fields of science, SSLPNN is still widely used in regression problems.

2.6 Mathematical Modeling of the Shallow Single-Layer Neural Network

In this study, we used a neural network architecture with 2,352 inputs for each selected parameter, one output neuron with a linear output function, and a single-layer grid. Through forward propagation, the network calculated the dot product between the nth sample x(n) and the weight vector w and then added the bias b. This calculation produced the weighted sum of the inputs with bias correction:

$$z^{(n)} = w^T x^n + b \tag{2}$$

$$\hat{y}^{(n)} \tag{3}$$

w = weight vector

b = bias

g = activation function

$\hat{y}^{(n)}$ = network output

2.7 Objective Function

The mean square error function assesses the credibility of the algorithm on a distinct trial:

$$L\left(\hat{y}^{(n)}, y^{(n)}\right) = \left(\hat{y}^{(n)}, y^{(n)}\right)^2 \tag{4}$$

where, $y^{(n)}$ is 2 if the nth trial fits category 2, 1 if the nth trial fits category 1, and 0 if the nth trial fits category 0. A cost function with L2 regularization of the weights is used to assess the global performance of the classifier. The term is affixed with the cost function to handle huge weights and to lessen the search space, reducing the inoperable weights toward zero, thus delivering more straightforward representations:

$$J(w, b) = \frac{1}{m}\sum_{1=1}^{m} L\left(\hat{y}^{(n)}, y^{(n)}\right) + \frac{\rho}{2m} \parallel w \parallel_w^2 \tag{5}$$

where:

ρ = regularization parameter

$\parallel w \parallel_w^2$ = L2 norm of the weight vector

For large values of ρ, the regularization is robust, enhancing the capacity associated with the weights. Consequently, the weights, which are not able to lessen the Mean Square Error (MSE), decrease to zero. However, for small values of ρ, the regularization outcome is weak. Here, the regression results are converted into class tags by using a Heaviside step function to deliver a numerical measure of the grid performance:

$$\hat{a}^{(n)} = \frac{d}{dx}\max\left\{0, \hat{y}^{(n)}\right\} \tag{6}$$

Of note, the accuracy is computed as if it were a classification part.

2.8 Parameter Optimization

The cost function is deputed to compute the errors in the recent forecasts. The learning process matter is comparable to the cost function reduction. While the training samples are fixed, the cost function depends only on the network parameters (the weights and bias). Thus, the cost function reduction is also comparable to the optimization of the grid parameters. The whole process is controlled by the following equations.

The objective function to be reduced is the cost function $K_n(\theta)$, where n denotes the n^{th} epoch, μ is a label for w and b, and g_n represents the gradient.

$$g_n = \nabla_\mu K_n(\mu_{n-1}) \tag{7}$$

This evaluation is then utilized to consider two exponential moving averages of the gradient m_n and the squared gradient v_n, respectively.

$$m_n = \beta_1.m_{n-1} + (1-\beta_1).\, g_n \tag{8}$$

$$v_n = \beta_2.m_{n-1} + (1-\beta_2).\, g_n \odot g_n \tag{9}$$

The two hyper-parameters $\beta_1, \beta_2 \in [0, 1)$ regulate the exponential decline rates of these moving averages.

Finally, the grid parameters are restructured by utilizing the classical method of gradient descent represented by $\hat{m}_n$ & $\hat{v}_n$, respectively.

$$\hat{m}_n = \frac{m_n}{1-\beta_1^n} \tag{10}$$

$$\hat{v}_n = \frac{v_n}{1-\beta_2^n} \tag{11}$$

The term ε certifies that the denominator is always non-zero and avoiding mathematical difficulties.

$$\mu = \mu - \alpha\frac{m_n}{\sqrt{v_n}+\varepsilon} \tag{12}$$

The SSLPNN algorithm can forecast the value $\hat{y}^{(n)}$ through an estimated function for every input vector x_n. Fig. 3 shows the flow of the input and output variables in the SSLPNN algorithm. The inputs are brought into the opening layer. Then, the valuation and optimization procedures are conducted, ending when the algorithm obtain better results The SSLPNN algorithm can be used as an influential instrument to deal with unexpected and indefinite problems. Thus, in the present study, a binary classification analysis was done by the SSLPNN algorithm.

To refine the precision of the model and to reduce the learning errors so as to obtain optimized outcomes, dissimilar models were created by hit-and-trial methods to find the appropriate number of layers and neurons for each layer. The input variables were the previously mentioned 14 notable factors: Namely, the population, population density, gender ratio, average age, human development index, elevation, temperatures (maximum, minimum, and average), relative humidity, wind speed, air quality index, pollution level, and UV index of each region. The number of confirmed COVID-19 cases was used as the output dataset. Two classes (labels) were assigned to the number of confirmed cases. Specifically, the number of confirmed cases under or equal to 800 were labeled as “0,” and the number of confirmed cases above 800 were labeled as “1.” The number of cases in five countries were included in the study. For modeling, 70% of the cases were used as training cases, 15% were used for validation, and 15% were reserved for testing.

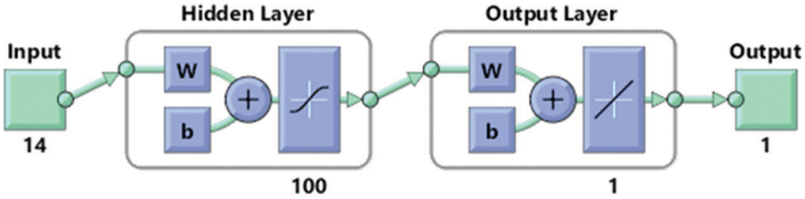


Figure 3: Structure of the Neural Network

2.9 Regression Learning through a Gaussian Process Regression (GPR) with the Matern 5/2 Preset

Determining the regulating parameters in an algorithm is important, as it aids in the quick convergence of the algorithm. There were no explicit associations among most of the parameters in this study. Thus, these parameters were considered independent and identified with the assistance of recent studies, experts, and trial-and-error methods. It is also important to identify the relationships between parameters through regression analysis, which helps with predictions based on the least learning error that are measured by the Root Mean Square Error (RMSE). Therefore, the process of selection was dimensionless and influenced the sensitivity of the modeling error. It is worth mentioning that the RMSE was used by the GPR algorithm with the Matern 5/2 GPR preset as a measurement of accuracy for the regression learner model.

When the dimensionality of the data is high, parameter identification typically turns out to be instinctive for the learning algorithms, as high-dimensional data tends to undesirably affect the efficacy of the majority of learning algorithms. Parameter identification is an effective dimensionality reduction procedure that chooses an ideal subclass of the unique parameters, delivering exceptional predictive control when modeling the data. These diverse structures can then be utilized to segregate trials into dissimilar modules. In this study, the Principal Component Analysis (PCA) procedure was used to select the optimal parameters.

In regression analysis, a GPR algorithm with variable models can adapt to numerous types of pattern recognition data for prediction through classification. The excellent experimental results demonstrate that GPR models provide a very promising feature selection solution to numerous pattern recognition problems through PCA. The algorithm can acquire patterns from the global distribution, therefore improving the precision of its pattern recognition capabilities.

GPR models with a finite-dimensional group of arbitrary variables and multivariate distribution are non-parametric kernel-based probabilistic models. Therefore, each linear combination is consistently distributed and the notion of Gaussian procedures is named after Carl Friedrich Gauss, as it emerges from Gaussian distribution to be an infinite-dimensional generalization of multivariate normal distributions. In this study, Gaussian process was used in the statistical modeling, regression to multiple target values, and analyses of mapping in higher dimensions. In addition, a GPR model with the Matern 5/2 GPR preset was used to plot the behavior of the algorithm; calculate the RMSE, R-Squared Value, MSE, Mean Absolute Error (MAE), prediction speed, and training time; and analyze the results of the GPR to see the similarities and differences in the data. The Matern 5/2 kernel does not have competence for measure problems in high dimensional spaces. The mathematical model of the Matern 5/2 GPR is illustrated as follows:

$$k\left(x_i, x_j|\theta\right) = \sigma_f^2\left(1 + \frac{\sqrt{3r}}{\sigma_l}\right) \tag{13}$$

where:

$$r = \sqrt{\left(x_i - x_j\right)^T\left(x_i - x_j\right)} \tag{14}$$

3 Results

3.1 Relationships between Environmental and Non-Environmental Parameters and COVID-19 Cases

The number of cases showed a significant correlation with the population and air quality index of a region. A statistically significant inverse relationship was observed between the number of cases and the average age and human development index levels. The results of the correlation analysis are presented in Tab. 1.

Table 1: Correlation analysis of COVID-19 cases with selected environmental and non-environmental variables

Variable	Correlation coefficient	*P*-value
Population	0.597*	<0.001
Population Density	−0.01	0.93
Gender Ratio	0.23	0.079
Average Age	−0.26*	0.04
Human Development Index (HDI)	−0.52*	<0.001
Elevation	−0.19	0.14
Maximum Temperature	0.19	0.14
Minimum Temperature	0.06	0.63
Average Temperature	0.15	0.23
Humidity	0.11	0.37
Wind Speed	0.03	0.76
Air Quality Index	0.37*	0.004
Pollution Level	0.23	0.07
UV Index	−0.04	0.74

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

3.2 Independent Association of Environmental and Non-Environmental Parameters

Furthermore, all included variables were analyzed to assess their correlations. An independent association was observed between each of the parameters. The results of this analysis are presented in Tab. 2. The correlation coefficient (R) indicated that the relationships were either negatively correlated or positively correlated among the independent parameters. Thus, the obtained results showed that SSLPNN could be used to further develop a pattern recognition model based on the selected parameters.

Table 2: Spearman's correlation for all mutually independent variables

Variables	Population	Gender ratio	Age	HDI	Elevation	Max temp	Min temp	Avg temp	Humidity	Wind speed	Air quality	Pollution level	UV index	Covid-19 cases
Population	1.0	−0.19	0.14	−0.48*	−0.27*	−0.25	−0.37*	−0.30*	0.39*	−0.16	0.35*	0.04	0.20	0.59*
Gender Ratio	−0.19	1.0	−0.84*	−0.30*	0.48*	0.64*	0.56*	0.63*	−0.49*	0.13	0.32*	0.59*	−0.24	0.23
Age	0.14	−0.84*	1.0	0.55*	−0.62*	−0.73*	−0.64*	−0.72*	0.43*	−0.22	−0.33*	−0.70*	0.34*	−0.26*
HDI	−0.48*	−0.30*	0.55*	1.00	−0.36*	−0-.26*	−0.14	−0.24	−0.17	0.01	−0.43*	−0.45*	0.00	−0.52*
Elevation	−0.27*	0.48*	−0.62*	−0.36*	1.0	0.42*	0.39**	0.42*	−0.41**	0.11	0.01	0.34*	−0.20	−0.19
Max Temp	−0.25	0.64*	−0.73*	−0.26*	0.42*	1.0	0.91**	0.97*	−0.31*	0.30*	0.10	0.42*	−0.41*	0.19
Min Temp	−0.37*	0.56*	−0.64*	−0.14	0.39*	0.91*	1.0	0.97*	−0.31*	0.36*	0.00	0.35*	−0.36*	0.06
Avg Temp	−0.30*	0.63*	−0.72*	−0.24	0.42*	0.97*	0.97*	1.0	−0.31*	0.31*	0.06	0.40*	−0.41*	0.15
Humidity	0.39*	−0.49*	0.43*	−0.17	−0.41*	−0.31*	−0.31*	−0.31*	1.0	−0.27*	−0.19	−0.44*	0.03	0.11
Wind speed	−0.16	0.13	−0.22	0.01	0.11	0.30*	0.36*	0.31*	−0.27*	1.00	0.10	0.24	0.15	0.03
Air Quality	0.35*	0.32*	−0.33*	−0.43*	0.01	0.10	0.08	0.06	−0.19	0.10	1.0	0.80*	0.12	0.37*
Pollution Level	0.04	0.59*	−0.70*	−0.45*	0.34*	0.42*	0.3*	0.40*	−0.44*	0.24	0.80*	1.00	−0.12	0.235
UV Index	0.20	−0.24	0.34*	0.06	−0.20	−0.41*	−0.36*	−0.41*	0.03	0.15	0.12	−0.12	1.0	−0.04
Covid-19 Cases	0.59*	0.23	−0.26*	−0.52*	−0.19	0.19	0.06	0.15	0.11	0.039	0.37*	0.23	−0.04	1.00

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

3.3 Results of the Pattern Recognition Model for Binary Classification Using SSLPNN

Before applying the binary classification through the pattern recognition model using the SSLPNN algorithm, a correlation analysis was conducted for the 54 case studies in the five countries, which included China, South Korea, Japan, Saudi Arabia, and Pakistan. This analysis showed a reasonable correlation coefficient (R) among the non-parametric variables (Tab. 3). Thus, it was decided that the 54 case studies could be evaluated in a cluster with the binary classification through the pattern recognition model. The results of our analysis indicate that the SSLPNN algorithm performed excellently, predicting the classes of the number of COVID-19 cases with an accuracy of 99.09% during training and an accuracy of 99.04% during testing, as shown in Tab. 3. These results demonstrated the high accuracy of the system, as presented in Fig. 4. The MSE for testing was almost 0 (MSE testing $9.11804e^{-01}$).

3.4 Prediction of COVID-19 Cases by Regression Analysis

The results of the regression analysis using the GPR technique with Matern 5/2 were reasonably accurate, with an RMSE of 0.95239 in the prediction of confirmed COVID-19 cases. The PCA technique was used for the removal of noise and redundant parameters in order to reduce the dimensionality of the dataset. The information and results for the models are presented in Tab. 4. In addition, the response plot and predicted values *vs.* actual values plot for the whole scenario are provided in Fig. 5. These results were promising, with the lowest RMSE value (0.952) and an R-value of 1 that could predict the values more accurately than all other competing models, as presented in Tab. 4. The overall training time required for this model was 134.04 sec, and the prediction speed was 11000 obs/sec.

Finally, the predictive number of COVID-19 cases was compared with the actual observed cases; the results were close. The overall observed cases were 1,271.00, and our model predicted 1,118.2 with an 87.96% accuracy. The results are presented in Tab. 5.

4 Discussion

This paper examined the relationship between COVID-19 cases and different environmental, ecological, and socio-economic factors and established a model system based on these variables to classify and predict rates of infection. COVID-19 has created a panic among the public. Scientific approaches must be identified and developed to predict the impact of these factors and to help policymakers take appropriate actions in the future.

Weather conditions, such as temperature, humidity, wind speed, and air quality, can affect the viability of viruses. Studies suggest that temperature and humidity have a strong influence on the transmission of COVID-19 [25]; researchers have also found that temperature and humidity may affect COVID-19 mortality [26]. In this study, we reported a statistically significant positive correlation between pollution, air quality index, and the number of positive COVID-19 cases in an area. Poor air quality is associated with the incidence of many diseases, such as asthma, bronchitis, lung and heart diseases, and many respiratory allergies [27]. China, where the epidemic started, is also severely affected by air pollution [28], indicating a relationship between poor air quality and COVID-19 [29].

The results of this study indicate that population density and human development index levels can also be associated with the number of COVID-19 cases in an area. Socio-economic factors like population size and low human development index levels are a significant driver for emerging infectious diseases and their subsequent effects on public health [30,31]. According to the Spearman's correlation coefficient, a direct and inverse relationship exists among the independent parameters of different case studies, due to the policies and restrictions in different countries for this issue. This finding supports the predictive power of our study, indicating we may be able to generalize it for other countries and extend its scope [32].

Table 3: Pattern recognition model for confirmed COVID-19 cases using SSLPNN

Neural Network Architecture	Neural Network Performance										
	Number of Hidden Nodes	Epocs	Gradient	Validation Checks	Performance	Accuracy (%)	RMSE (%)	Cross-Entropy	Training (Classified/ Misclassified) out of 1646	Testing (Classified/ Misclassified) out of 353	Validation (Classified/ Misclassified) out of 353	Overall (Classified/ Misclassified) out of 2352
Shallow Neural Network	5	40	0.003	6	0.06	98.67	1.33	1.31	1624/22	347/6	353/0	2324/28
	10	32	0.009	6	0.05	98.66	1.34	1.23	1624/22	351/2	349/4	2324/28
	50	21	0.02	6	0.05	98.97	1.03	1.26	1629/17	347/6	348/5	2324/28
	100	33	0.03	6	0.06	99.09	0.91	1.33	1631/15	347/6	346/7	2324/28
	200	54	0.04	6	0.08	98.67	1.33	1.29	1624/22	351/2	349/4	2324/28

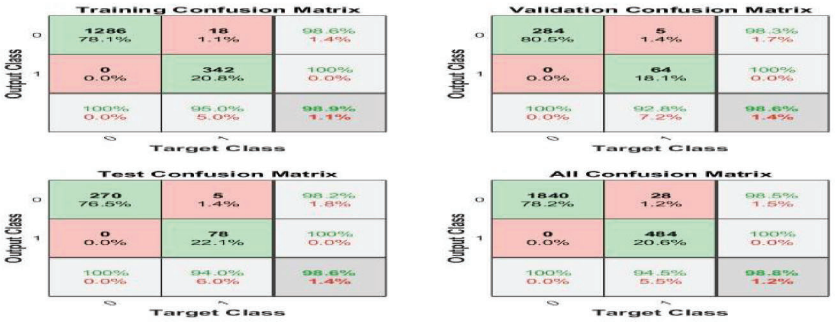
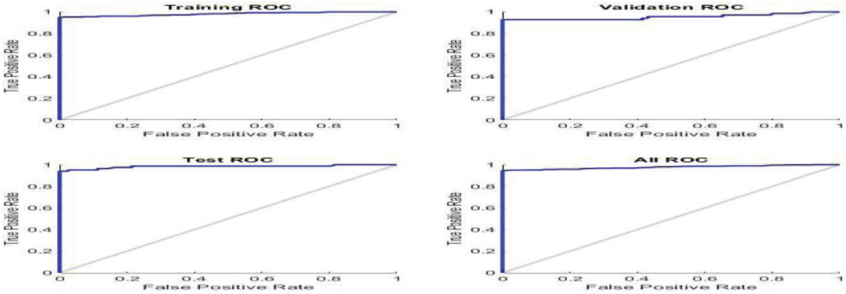
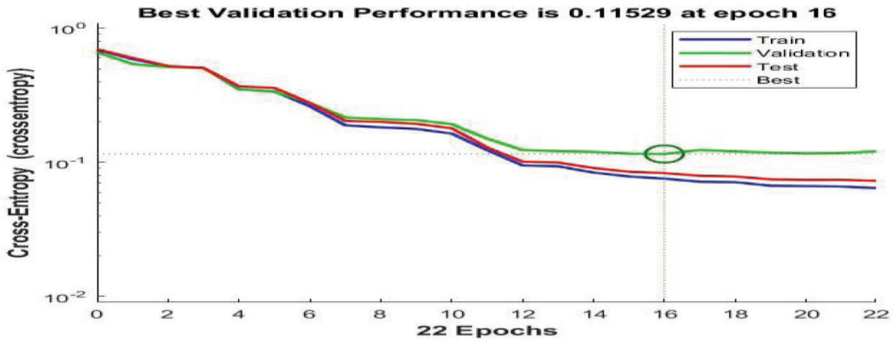
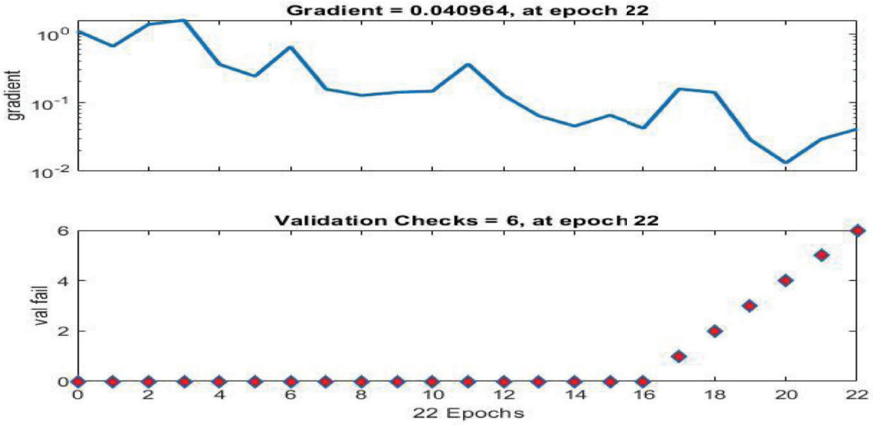
Results Neurons = 100	Graphs
Confusion Matrix	
Receiver Operating Characteristic (ROC) Curve	
Performance	
Gradient & Validation Checks	

Figure 4: (continued)

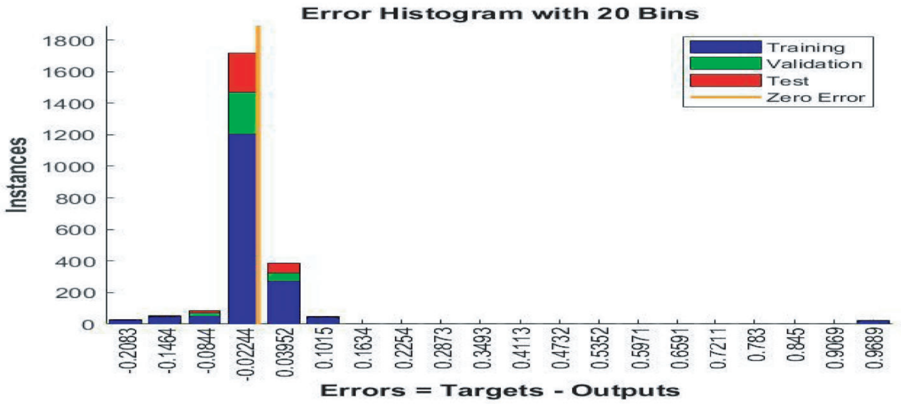
Error Histogram	

Figure 4: Best results of the shallow neural network (neurons = 100)

Despite significant advancements in medical science, infectious diseases are a leading cause of mortality. For a novel disease like COVID-19 that does not have any standard guidelines for treatment and vaccination, the short-term response from medical science will be limited. However, we can utilize mathematical tools to better understand and forecast the impacts of such diseases. In the last few years, AI has been widely adopted to better understand infectious diseases and to predict epidemics [33].

Studies have reported on the use of neural networks to predict the outbreaks of many diseases, such as foot and mouth disease, influenza, epidemic diarrhea, Ebola virus, Rift Valley fever virus, Nipah virus, and SARS [34–36]. A recent report utilized neural network models to identify the risk of COVID-19 cases in a specific country based on weather conditions, and promising results were reported [37].

In this paper, we used the SSLPNN algorithm, which performed excellently, predicting the classes of COVID-19 cases for both the training and testing datasets with an accuracy of 99.09% and 99.04%, respectively.

The results of the binary classification modeling using SSLPNN with Scaled Conjugate Gradient Backpropagation (SCGB) showed high accuracy, with an MSE of 0.0114858 in five selected countries. Moreover, the results of the regression analysis using the GPR technique with Matern 5/2 for 54 case studies in five countries also showed high accuracy in the prediction of COVID-19 confirmed cases, with an RMSE of 0.952. This study established some previously unexplored patterns in the relationships between COVID-19 infections and the environmental and non-environmental conditions of select countries. Based on this analysis, we propose that both SCGB and GPR may be applicable to classifying and predicting patterns of COVID-19 cases. The results show that AI techniques can provide reasonable estimates about upcoming events based on specific inputs by learning the hidden structures of a scenario [38,39]. These rational outcomes can support governments in policy-making decisions, particularly those regarding public health, to ensure a sustainable development process. Our comparative analysis of daily weather parameters and trends of confirmed cases also demonstrate the role of these variables in the rate of COVID-19 cases.

Our findings are consistent with previous studies into the effects of climatic conditions on epidemic diseases and public health [40,41]. A recent analysis of confirmed COVID-19 cases through a binary classification using artificial intelligence and regression analysis also showed the impact of weather conditions in the COVID-19 epidemic [42]. Overall, machine learning is an innovative technique that is helpful when predicting upcoming trends in COVID-19 cases in relation to specific ecological and socio-economic factors.

Table 4: Gaussian process regression model with the highest prediction performance

Serial. No	Model	Preset	RMSE	R-Squared	MSE	MAE	Prediction Speed	Training Time	Terms	Upper bound on Terms	Robust	PCA	Max. no. of Steps
1	Linear Regression	Linear	317.4	0.58	1.01e +05	243.96	~26000 obs/s	3.1733 sec	Linear		Off	After training 7 components were kept	
2	Linear Regression	Interactions Linear	231.2	0.78	53454	169.98	~24000 obs/s	0.5699 sec	Interactions		Off		
3	Linear Regression	Robust Linear	342.71	0.51	1.18e +05	223.13	~34000 obs/s	0.6687 sec	Linear		On		
4	Stepwise Linear Regression	Stepwise Linear	231.94	0.78	53794	172.57	~29000 obs/s	21.744 sec	Linear	Interactions			1000
5	Tree	Medium Tree	18,37	1.00	337.31	1.1309	~35000 obs/s	0.385 Sec					
6	Tree	Fine Medium Tree	18,05	1.00	325.91	1.0737	~37000 obs/s	2.334 Sec					
7	Tree	Fine Coarse Tree	131.97	0.93	17417	1.0737	~39000 obs/s	0.383 Sec					
8	SVM	Linear SVM	334.53	0.53	1.1191e +05	221.5	~35000 obs/s	8.31 Sec					
9	SVM	Quadratic SVM	219.4	0.80	48137	134.58	~33000 obs/s	73.372 Sec					
10	SVM	Cubic	39.41	0.99	1553	36.47	~35000 obs/s	95.735 sec					
11	SVM	Fine Gaussian SVM	44.15	0.99	1949.6	43.352	~31000 obs/s	754625.73 sec					
12	SVM	Fine Gaussian SVM	77.54	0.97	6013	52.411	~36000 obs/s	1.016 sec					
13	SVM	Coarse Gaussian SVM	279.48	0.67	78108	184.23	~32000 obs/s	1.280 sec					
14	Ensemble	Boosted Trees	57.12	0.99	3262.3	42.551	~15000 obs/s	4.051 sec					
15	Ensemble	Bagged Trees	12.82	1.00	164.29	1.162	~19000 obs/s	2.34 sec					
16	Gaussian Process Regression	Squared Exponential GPR	0.995	1.00	0.98963	0.224	~14000 obs/s	117.9 sec					
17	Gaussian Process Regression	Matern 5/2 GPR	0.952	1.00	0.90705	0.208	~11000 obs/s	134.04 sec					
18	Gaussian Process Regression	Exponential GPR	1.579	1.00	2.4923	0.172	~16000 obs/s	120.04 sec					
19	Gaussian Process Regression	Rational Quadratic GPR	0.9973	1.00	0.99457	0.206	~7800 obs/s	171.41 sec					

Min. Leaf Size	Surrogate Decision Splits	Kernel Function	Kernel Scale	Box Constraint	Epsilon	Standardize Data	Learning Rate	No. of Leaves	Basis Function	Use of Isotropic Kernel	Optimized Numeric Parameter	Kernel Sigma
4	Off											
12	Off											
36	Off											
		Linear	Automatic	Automatic	Automatic	True						
		Quadratic	Automatic	Automatic	Automatic	True						
		Cubic	Automatic	Automatic	Automatic	True						
		Gaussian	0.66	Automatic	Automatic	True						
		Gaussian	2.6	Automatic	Automatic	True						
		Gaussian	11	Automatic	Automatic	True						
8							0.1	30				
8								30				
		Squared Exponential	Automatic		True			Constant		True	Automatic	
			5/2 GPR	Automatic		True			Constant	True	True	Automatic
		Exponential			True			Constant	True	True	Automatic	
		Rational Quadratic	Automatic		True			Constant	True	True	Automatic	

Table 5: Comparison between the observed and predicted COVID-19 cases

Observed Value														
96,050,000.00	575.00	1.06	38.40	0.73	104.00	13.70	−0.10	6.80	61.90	6.80	123.00	58.00	6.00	1,271.00
Predicted Value														
96,050,000.00	575.00	1.06	38.40	0.73	104.00	13.70	−0.10	6.80	61.90	6.80	123.00	58.00	6.00	1,118.20

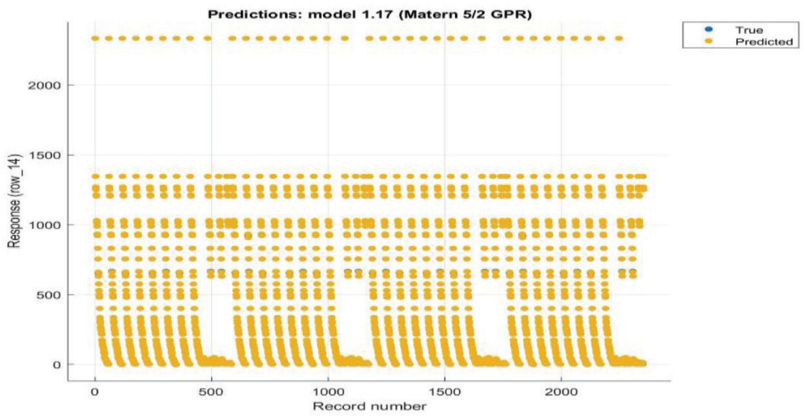
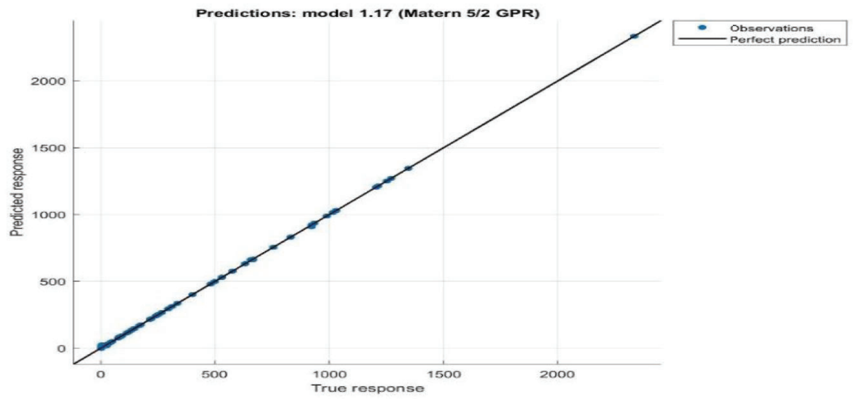
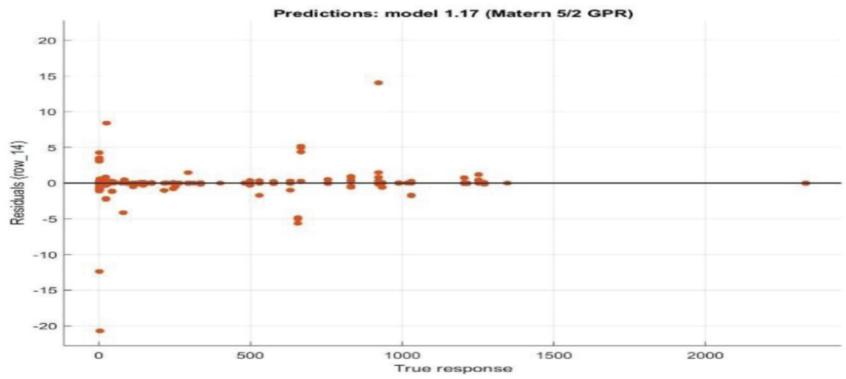
Response Plot	
Predicted *Vs*. Actual Plot	
Residuals Plot	

Figure 5: Best results of the regression analysis with Gaussian process regression kernel function

Funding Statement: The authors received no specific funding for this study.

Conflict of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] V. Surveillances, "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)—China, 2020," *China CDC Weekly*, vol. 2, no. 8, pp. 113–122, 2020.

[2] S. Kannan, S. S. P. Ali, A. Sheeza and K. Hemalatha, "COVID-19 (novel coronavirus 2019)-recent trends," *European Review for Medical and Pharmacological Sciences*, vol. 24, no. 4, pp. 2006–2011, 2020.

[3] T. Phan, "Novel coronavirus: From discovery to clinical diagnostics," *Infection, Genetics and Evolution*, vol. 79, 104211, 2020.

[4] World Health Organization, "WHO director-general's opening remarks at the media briefing on COVID-19 March 2020," 2020.

[5] S. Lei, F. Jiang, W. Su, C. Chen, J. Chen *et al.,* "Clinical characteristics and outcomes of patients undergoing surgeries during the incubation period of COVID-19 infection," *EClinicalMedicine*, vol. 21, pp. 100331, 2020.

[6] A. Granados, E. C. Goodall, K. Luinstra, M. Smieja and J. Mahony, "Comparison of asymptomatic and symptomatic rhinovirus infections in university students: Incidence, species diversity, and viral load," *Diagnostic Microbiology and Infectious Disease*, vol. 82, no. 4, pp. 292–296, 2015.

[7] L. Furuya-Kanamori, M. Cox, G. J. Milinovich, R. J. S. Magalhaes, I. M. Mackay *et al.,* "Heterogeneous and dynamic prevalence of asymptomatic influenza virus infections," *Emerging Infectious Diseases*, vol. 22, no. 6, pp. 1052, 2016.

[8] M. Di Marco, M. L. Baker, P. Daszak, P. De Barro, E. A. Eskew *et al.,* "Opinion: Sustainable development must account for pandemic risk," in *Proc. of the National Academy of Sciences of the United States of America*, vol. 117, no. 8, pp. 3888–3892, 2020.

[9] C. Garcia-Vidal, G. Sanjuan, P. Puerta-Alcalde, E. Moreno-García and A. Soriano, "Artificial intelligence to support clinical decision-making processes," *EBioMedicine*, vol. 46, pp. 27–29, 2019.

[10] T. Hassan and F. Ahmad, "Transaction and identity authentication security model for e-banking: Confluence of quantum cryptography and AI, in *Int. Conf. on Intelligent Technologies and Applications*, pp. 338–347, 2018.

[11] E. Rees, V. Ng, P. Gachon, A. Mawudeku, D. Mckenney *et al.,* "Risk assessment strategies for early detection and prediction of infectious disease outbreaks associated with climate change," *Canada Communicable Disease Reports*, vol. 45, no. 5, pp. 119–126, 2019.

[12] F. R. Chowdhury, Q. S. U. Ibrahim, M. S. Bari, M. J. Alam, S. J. Dunachie *et al.,* "The association between temperature, rainfall and humidity with common climate-sensitive infectious diseases in Bangladesh," *PLoS One*, vol. 13, no. 6, pp. 1–17, 2018.

[13] "List of countries by sex ratio," 2020. [Online]. Available: http://statisticstimes.com/demographics/countries-by-sex-ratio.php.

[14] "Coronavirus pandemic in Pakistan," 2020. [Online]. Available: https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_Pakistan.

[15] "List of Japanese prefectures by human development index," 2020. [Online]. Available: https://en.wikipedia.org/wiki/List_of_Japanese_prefectures_by_Human_Development_Index.

[16] "List of regions of South Korea by human development index," 2020. [Online]. Available: https://en.wikipedia.org/wiki/List_of_regions_of_South_Korea_by_Human_Development_Index.

[17] "List of administrative divisions of greater China by human development index," 2020. [Online]. Available: https://en.wikipedia.org/wiki/List_of_administrative_divisions_of_Greater_China_by_Human_Development_Index.

[18] "Lahore US embassy, Pakistan air pollution: real-time air quality index," 2020. [Online]. Available: https://aqicn.org/city/pakistan/lahore/us-embassy/.

[19] "Saudi Arabia population," 2020. [Online]. Available: https://countrymeters.info/en/Saudi_Arabia.

[20] "Saudi Arabia—median age of the population 1950–2050," 2020. [Online]. Available: https://www.statista.com/statistics/262482/median-age-of-the-population-in-saudi-arabia/.

[21] "List of countries and dependencies by population density," 2020. [Online]. Available: https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population_density.

[22] "List of administrative units of Pakistan by human development index," 2020. [Online]. Available: https://en.wikipedia.org/wiki/List_of_administrative_units_of_Pakistan_by_Human_Development_Index.

[23] "Coronavirus disease (COVID-19) situation reports," 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/.

[24] S. Shahzadi, B. Khaliq, M. Rizwan and F. Ahmad, "Security of cloud computing using adaptive neural fuzzy inference system," *Security and Communication Networks*, vol. 2020, no. 8, pp. 1–15, 2020.

[25] B. Chen, H. Liang, X. Yuan, Y. Hu, M. Xu *et al.,* "Roles of meteorological conditions in COVID-19 transmission on a worldwide scale," *medRxiv*, 2020.

[26] Y. Ma, Y. Zhao, J. Liu, X. He, B. Wang *et al.,* "Effects of temperature variation and humidity on the mortality of COVID-19 in Wuhan," *medRxiv*, 2020.

[27] M. Brauer, "How much, how long, what, and where: Air pollution exposure assessment for epidemiologic studies of respiratory disease," in *Proc. of the American Thoracic Society*, vol. 7, no. 2, pp. 111–115, 2010.

[28] M. Z. He, P. L. Kinney, T. Li, C. Chen, Q. Sun *et al.,* "Short-and intermediate-term exposure to NO2 and mortality: A multi-county analysis in China," *Environmental Pollution*, vol. 261, pp. 114165, 2020.

[29] F. Dutheil, J. S. Baker and V. Navel, "COVID-19 as a factor influencing air pollution?," *Environmental Pollution*, vol. 263, pp. 114466, 2020.

[30] K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk *et al.,* "Global trends in emerging infectious diseases," *Nature*, vol. 451, no. 7181, pp. 990–993, 2008.

[31] A. R. Parra, L. Echezuria and A. Rodriguez-Morales, "Epidemiological transition in Venezuela: Relationships between infectious diarrheas, ischemic heart diseases and transportation accidents mortalities and the human development index (HDI) in Venezuela, 2005–2007," *International Journal of Infectious Diseases*, vol. 14, pp. e426–e427, 2010.

[32] T. Cleff, "Applied statistics and multivariate data analysis for business and economics," Switzerland: Springer, 2019. [Online]. Available: https://link.springer.com/book/10.1007%2F978-3-030-17767-6.

[33] S. Agrebi and A. Larbi, "Use of artificial intelligence in infectious diseases," *Artificial Intelligence in Precision Health*, pp. 415–438, 2020.

[34] M. D. Philemon, Z. Ismail and J. Dare, "A review of epidemic forecasting using artificial neural networks," *International Journal of Epidemiology*, vol. 6, no. 3, pp. 132–143, 2019.

[35] W. Jia, X. Li, K. Tan and G. Xie, "Predicting the outbreak of the hand-foot-mouth diseases in China using recurrent neural network," in *IEEE Int. Conf. on Healthcare Informatics*, Xi'an: IEEE, pp. 1–4, 2019.

[36] A. Forna, P. Nouvellet, I. Dorigatti and C. A. Donnelly, "Case fatality ratio estimates for the 2013–2016 West African Ebola epidemic: Application of boosted regression trees for imputation," *International Journal of Infectious Diseases*, vol. 79, no. 12, pp. 128, 2019.

[37] R. Pal, A. A. Sekh, S. Kar and D. K. Prasad, "Neural network based country wise risk prediction of COVID-19," *arXiv preprint*, *arXiv:2004.00959*, 2020.

[38] A. P. Piotrowski, J. J. Napiorkowski and A. E. Piotrowska, "Impact of deep learning-based dropout on shallow neural networks applied to stream temperature modelling," *Earth-Science Reviews*, vol. 201, 103076, 2019.

[39] N. Zhang, J. Xiong, J. Zhong and K. Leatham, "Gaussian process regression method for classification for high-dimensional data with limited samples," in *Eighth Int. Conf. on Information Science and Technology*, Cordoba: IEEE, pp. 358–363, 2018.

[40] V. Martin, V. Chevalier, P. Ceccato, A. Anyamba, L. De Simone *et al.,* "The impact of climate change on the epidemiology and control of rift valley fever," *Revue Scientifique et Technique*, vol. 27, no. 2, pp. 413–426, 2008.

[41] B. Pirouz, S. Shaffiee Haghshenas, S. Shaffiee Haghshenas and P. Piro, "Investigating a serious challenge in the sustainable development process: Analysis of confirmed cases of COVID-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis," *Sustainability*, vol. 12, no. 6, pp. 2427, 2020.

[42] C. Bezirtzoglou, K. Dekas and E. Charvalos, "Climate changes, environment and infection: Facts, scenarios and growing awareness from the public health community within Europe," *Anaerobe*, vol. 17, no. 6, pp. 337–340, 2011.