Environmental Science and Pollution Research (2022) 29:68269–68279
https://doi.org/10.1007/s11356-022-20231-z

RESEARCH ARTICLE

Estimation of COVID-19 patient numbers using artificial neural networks based on air pollutant concentration levels

Gülşen Aydın Keskin[1] · Şenay Çetin Doğruparmak[2] · Kadriye Ergün[1]

Received: 30 December 2021 / Accepted: 9 April 2022 / Published online: 10 May 2022

Abstract

The dilemma between health concerns and the economy is apparent in the context of strategic decision making during the pandemic. In particular, estimating the patient numbers and achieving an informed management of the dilemma are crucial in terms of the strategic decisions to be taken. The Covid-19 pandemic presents an important case in this context. Sustaining the efforts to cope with and to put an end to this pandemic requires investigation of the spread and infection mechanisms of the disease, and the factors which facilitate its spread. Covid-19 symptoms culminating in respiratory failure are known to cause death. Since air quality is one of the most significant factors in the progression of lung and respiratory diseases, it is aimed to estimate the number of Covid-19 patients corresponding to the pollutant parameters (PM10, PM2.5, SO_2, NO_X, NO_2, CO, O_3) after determining the relationship between air pollutant parameters and Covid-19 patient numbers in Turkey. For this purpose, artificial neural network was used to estimate the number of Covid-19 patients corresponding to air pollutant parameters in Turkey. To obtain highest accuracy levels in terms of network architecture structure, various network structures were tested. The optimal performance level was developed with 15 neurons combined with one hidden layer, which achieved a network performance level as high as 0.97342. It was concluded that Covid-19 disease is affected from air pollutant parameters and the number of patients can be estimated depending on these parameters by this study. Since it is known that the struggle against the pandemic should be handled in all aspects, the result of the study will contribute to the establishment of environmental decisions and precautions.

Keywords Covid-19 pandemic · Air pollutants · Monitoring · Artificial neural network · Prediction · Turkey

Highlights

- The Covid-19 pandemic is one of the most critical problems affecting almost all countries today. The success in controlling pandemics caused by infectious diseases depends mostly on significant information being provided as early as possible in the life cycle of the pandemic, based often on a very limited set of data.
- In this study, it was aimed to estimate the number of patients based on air pollutant parameters. To do so, Covid-19 data map of Turkey was subjected to Multilayer Perceptron Backpropagation (MLP-BP), to estimate the number of Covid-19 patients based on various air pollutant parameters (PM10, PM2.5, SO_2, NO_X, NO_2, CO, O_3).
- In conclusion, one can forcefully argue that the MLP-BP model proposed in this study can be used for analyzing and predicting Covid-19 patient numbers based on air pollution levels.

Responsible Editor: Lotfi Aleya

✉ Şenay Çetin Doğruparmak
senayc@kocaeli.edu.tr

Extended author information available on the last page of the article

Introduction

Today, Covid-19 ranks at the top of the agenda for virtually all countries. The Covid-19 (CO: corona, VI: Virus, D: Disease, 2019) has been qualified as a novel coronavirus (SARS-CoV-2) (Hamit et al. 2021). It is a disease caused by a pathogen which affects human respiratory system, leading to significant symptoms including but not limited to fever, dry cough, dyspnea, and pneumonia. The disease was detected for the first time on December 1, 2019, in Wuhan, province of China. Due to its high speed of spread, it soon reached other parts of China, followed by the rest of the globe. As the case counts and associated deaths increased quickly in the countries affected, the World Health Organization declared it as a pandemic on March 11, 2020 (Balcı and Çetin 2019; Shaban et al. 2021).

The governments often face a dilemma between health concerns and the economy in formulating their responses to a pandemic. And they often prefer strategic decision making through diligent observation and assessment of the developments, instead of going with sudden and dramatic decisions. In particular, estimating the patient numbers and achieving an informed management of the dilemma are crucial in terms of the governments' strategic decision-making processes. Predicting pandemic cases well in advance can provide the governments an important guide in better-informed and improved management of the dilemma between public health and the economy. Predictions about the future, as well as the use of methods and models involving machine learning for coming up with such predictions, has become essential. The literature is fulfilling in terms of prediction models regarding pandemics (Eroğlu 2020; Ergül et al. 2020). The success in controlling pandemics caused by infectious diseases depends mostly on significant information being provided as early as possible in the life cycle of the pandemic, based often on a very limited set of data. To do so, cases should be monitored accurately using various variables, in order to achieve increased reliability levels for forecasts.

It has emphasized that improving air quality will play an important role in overcoming the epidemic and mitigating the disease (Conticini et al. 2020). In this context, during the Covid-19 epidemic disease, it has great importance to clarify how much the air pollutant parameters, PM10 (particulate matter 10), PM2.5 (particulate matter 2.5), SO_2 (sulfur dioxide), NO_X (nitrogen oxide), NO_2 (nitrogen dioxide), CO (carbon monoxide), and O_3 (ozone) that increase the severity of the disease in humans, affect the number of patients. Because of this information, the present study focused on air pollutants (PM10, PM2.5, SO_2, NO_X, NO_2, CO, O_3), which can be considered a significant set of factors to accelerate the spread of the disease and to increase case numbers. To achieve a versatile approach to the fight against Covid-19 pandemic in particular, one should take into account the correlation between air pollutants at a regional level on the one hand, and the patient number and distribution on the other. The main reason that air pollution increases the severity of Covid-19 disease is its damage to the respiratory system and other organ systems caused by long-term exposure to the pollutants. It has seen that the severity of Covid-19 disease was higher in regions dominated by air pollution compared to areas exposed to cleaner air. The studies have shown that there is a positive correlation between current high levels of air pollution and the severity of Covid-19 in some regions such as parts of Spain, Italy, USA, Germany, and Saudi Arabia (Paital and Agrawal 2021; THHP 2021). Therefore, in this study, air pollutants, which can be considered as one of the factors that increase the Covid-19 patient numbers and the rate of transmission of the disease, are discussed.

In the study, it was aimed to estimate the number of Covid-19 patients corresponding to the pollutant parameters (PM10, PM2.5, SO_2, NO_X, NO_2, CO, O_3) after determining the relationship between air pollutant parameters and Covid-19 patient numbers in Turkey. Artificial neural networks (ANN) were used for prediction. ANN, which are widely used in machine learning and are inspired by the information processing technique of the human brain, enable the understanding of the relationships between dependent and independent variables affecting the system when the mathematical model of the system cannot be created (Borghi et al. 2021; Kuvvetli et al. 2021; Khojasteh et al. 2021).

The data used in the study include the "Daily Status Report" for Covid-19, published by the Ministry of Health of the Republic of Turkey, for the period June 29, 2020–November 23, 2020, and the air pollutant concentration levels published by the Ministry of Environment and Urbanization of the Republic of Turkey. The numbers provided in the report published by the Ministry of Health are given for 12 regions (Istanbul, Western Marmara, Aegean, Eastern Marmara, Western Anatolia, Mediterranean, Central Anatolia, Western Blacksea, Eastern Blacksea, North-Eastern Anatolia, Central-Eastern Anatolia, and South-Eastern Anatolia) according to the regional classification of TUIK (Turkish Statistical Institute) IBBS (Statistical Regional Units Classification). However, as the data for air pollutants other than PM10 and PM2.5 were not available for the Central-Eastern Anatolia and South-Eastern Anatolia regions, these regions were removed from the study scope, to reduce the number to 10 regions. The daily air pollutant parameters were downloaded from the website www.havaizleme.gov.tr for 10 regions for which the Covid-19 patient numbers were available. The station numbers as well as measured values vary from region to region. As the model required a single value as the daily pollutant parameter, the median of numbers received from all stations covered was used as the value pertaining to the region. The arithmetic mean was not used, as some data points were lost or unavailable.

Forecasting studies are significantly important to efficiently manage Covid-19 pandemic and the affected areas such as economy, education, and manufacturing. It is also obvious that the relationship of air pollutants with the patient numbers and distribution of patients should be considered in order to versatile struggle against Covid-19 and similar pandemics and to take regional restrictions or precautions. The developed model will provide policy makers general information to pursue the spread of the disease and support decision making. Besides, the estimation results will have an impact on people having a chronic disease in the society to review their individual precautions and comprise environmental awareness indirectly. For better control of the pandemic, it is crucial to consider taking proactive decisions based on regional assessments such as regulating the manufacturing processes contributing to air pollution, determining the traffic time schedule for different types of vehicles. Thereby, related studies on these subjects would have significant contribution to the literature.

Related literature

There are several studies in the literature, focusing on the estimation and assessment of patient number, infection rate, death rate, or case number, based on a number of variables known to have an effect on Covid-19 involving machine learning and forecasting methods. Wieczorek et al. (2020) also worked on developing a neural network model for predicting Covid-19 spread. The predictor that study used is based on a classical approach to deep architecture. Using a deep learning method with rolling update mechanism, Wang et al. (2020) developed a forecasting model for Covid-19 based on epidemical data. Kumar and Kumar (2020), in turn, looked at the correlation between various meteorological parameters and the Covid-19 pandemic in Mumbai, India, using an artificial neural network technique and Spearman rank correlation test to predict the associations of Covid-19 with the meteorological parameters covered. Hasan (2020) carried out a study where a hybrid model incorporating ensemble empirical mode decomposition and ANN was proposed for the purposes of predicting the spread of the Covid-19 pandemic. To do so, real-time Covid-19 time series data was used. Using ANN experiments, Magazzino et al. (2020) found that the concentration levels of PM2.5 and PM10 were linked to Covid-19-related deaths. The study focused on the potential effects of particulate matter in terms of spreading the epidemic. Silva et al. (2020) explored and compared the predictive capacity of Bayesian regression neural network, cubist regression, *k*-nearest neighbors, quantile random forest, and support vector regression used on a stand-alone basis, and a hybrid framework comprised of variational mode decomposition coupled with one of the above-mentioned models on each case. All forecasting models used Covid-19 case data accumulated to date, as well as exogenous variables as daily temperature and precipitation as inputs. Ahangar et al. (2020) investigated the relationship between the Covid-19 pandemic's growth and various weather factors and proceeded to propose an optimal mathematical model to forecast daily Covid-19 case counts. In order to unearth and assess the potential links between Covid-19 and absolute humidity and temperature in 5 European countries, they conducted a Poisson analysis with general linear neural network model, to predict the trends regarding and the number of daily Covid-19 cases. Rasjid et al. (2021) predicted the death and infection rates associated with Covid-19 in Indonesia by Savitzky Golay smoothing and long short-term memory neural network model. The study was based on a dataset containing daily death and infection numbers caused by Covid-19. Mohammadi et al. (2021) carried out a comparative study on Covid-19 patients from six provinces of Iran and applied multilayer perceptron neural network and logistic regression models to contribute to the efforts to diagnose and assess the spread of Covid-19. Borghi et al. (2021) used a machine learning model based on MLP ANN structure, to achieve effective and efficient prediction of the behavior of the series analyzed, in time frames extending up to 6 days. In another study, Magazzino et al. (2021) looked at the relationship between Covid-19-related deaths, economic growth, PM10, PM2.5, and NO_2 concentration levels in New York state, running daily city-level data through two distinct machine learning models. In a study which involved the first empirical analysis of the correlation between the threshold values of NO_2 concentrations and Covid-19 related deaths in France and using ANN experiments and a Causal Direction from Dependency algorithm, Mele et al. (2021) found that the concentration of NO_2 was linked to Covid-19-related deaths in three major French cities. Using a multilayer perceptron, Abdelhafez et al. (2021) analyzed the correlation between the daily confirmed Covid-19 cases in Jordan and various meteorological parameters including the maximum ambient temperature, average daily temperature, wind speed, pressure, relative humidity, and average daily solar radiation. Behnam and Jahanmahin (2021) compiled a dataset comprised of the number of confirmed cases, the daily number of deaths associated with Covid-19, and the number of recovered cases. Furthermore, by combining case number-related variables with measures taken and policies introduced over time along with any changes thereof, with the help of machine learning algorithms such as the logistic function using inflection point, the authors came up with new rates such as weekly death rate, life rate, and new approaches to mortality rate and recovery rate which would have occurred with specific policies and measures. Guo and He (2021), in turn, developed an artificial neural network to help with the modeling of confirmed cases and deaths associated with Covid-19. The model allowed the prediction of the next day's infected cases and deaths associated with Covid-19 worldwide on a cumulative basis, using the infected case and death counts of the preceding days. Long et al. (2021) adapted a variant of physics-informed neural network for an effort to identify the time-varying parameters of the Susceptible-Infectious-Recovered-Deceased model pertaining to the spread of Covid-19. To do so, they used daily reported case numbers. Malki et al. (2021) evaluated various machine learning methods in the context of predicting the spread of the Covid-19. The overall endeavor tried to come up with daily total confirmed Covid-19 positive cases, daily and total deaths, and total and daily recoveries. Gupta et al. (2021a) employed a support vector machine to analyze Covid-19 data from India, using the prophet prediction and the linear regression models. Gupta et al. (2021b) proposed a prediction model designed primarily to deal with the small size of the dataset available for the purpose of estimating the pandemic curve in Europe. In conclusion, they adopted for the generalized regression neural network, given the fact that it can be trained accurately with a rather small dataset. Kuvvetli et al. (2021), have designed an artificial neural network model that will predict the daily number of cases and deaths caused by the Covid-19 pandemic in a generalized manner. Reddy et al. (in press), in turn, proposed a neural network-based prediction model for the number of Covid-19 cases in India. The recurrent neural network-based LSTM they developed was applied on a dataset for India, to train it for prediction purposes.

While the above-mentioned studies utilize machine learning methods for the prediction and evaluation of Covid-19 cases, the following literature consists of papers using time series analysis. Khan and Gupta (2020) employed an autoregressive integrated moving average (ARIMA) model with the data collected during the earlier months of the pandemic, and then proceeded to verify the model using the data from a shorter time frame to follow. To compare the accuracy of the models, a nonlinear autoregressive neural network was developed. Singh et al. (2021) interested in the development of ARIMA model to predict the number of Covid-19 cases and the effect of a rise in temperature in most affected states of India. Yudistira et al. (2021) presented a multivariate analysis to emerge comprehensive explanations regarding several factors contributing to the dynamics of the pandemic. The explanations thus developed were then used to point out critical characteristics of the pandemic, which were in turn helpful in predicting daily case numbers over the time period analyzed. What the authors proposed for this purpose was gradient-based visual attribution for generating a saliency map. The aim was to explain which specific variables contributed to daily case numbers during a given time frame. Alaraj et al. (2021) worked on an improvement on the classical SEIR model, this time using the SEIRD and ARIMA models, with a view to compensating for the differences between actual and predicted data. Overall, the study aimed to come up with a reliable approach for predicting the evolution of the epidemic, so that policymakers could take the correct steps for reducing contagion and implement selective measures with reference to the specific characteristics of individual regions. Chaurasia and Pal (2021) used the data for the number of cases, deaths, and recovery cases worldwide through a period of 5 months to implement a number of forecasting techniques including the naive method, moving average, simple average, Holt linear trend method, Holt-Winters method, single exponential smoothing, and ARIMA for comparison, and to understand if and how these methods stand as improvements over the root mean square error score. Toğa et al. (2021) analyzed the prevalence of Covid-19 in Turkey, and proceeded to predict infected case numbers, the number of deaths, and the recovered cases in Turkey, using ARIMA and ANN. The study then compared the techniques with reference to MSE and correlation coefficient. Aabed and Lashin (2021) designed a fuzzy logic system to predict the effects of the variables covered, on the rate of Covid-19 spread. Apart from these studies, Göreke et al. (2021) designed a new hybrid classifier architecture using deep learning techniques and used it to detect Covid-19 spread. Another innovative study for Covid-19 detection was designed by Noshad et al. (2021). The authors presented a deep residual architecture for identification of Covid-19 using raw chest X-ray images in their study.

When the literature was examined in depth, the studies investigating the relationship between the number of Covid-19 patients and air pollutants were also encountered (Zhu et al. 2020; Magazzino et al. 2020; Velasquez and Lara 2020). For a more detailed review, see Hu et al. (2021). These studies revealed that although there have differences between regions, many researchers have found an association between air pollutants and the number of Covid-19 patients. On the other hand, when the studies on Covid-19 using artificial neural networks are examined, it has been seen that almost all of them use the number of patients or deaths from previous periods as the input of artificial neural networks. Ahangar et al. (2020) has been distinguished as an exception using the parameters of temperature, and relative and absolute humidity as inputs in their study. Still, the fewness of studies using air pollutants as input data to estimate the number of Covid-19 patients for a given period in the future is quite obvious. When the estimation studies on Covid-19 are examined, it is seen that outputs such as patient numbers, mortality rate, number of cases, number of infected cases per day, and recovered patient numbers per day are tried to be predicted by different methods by independent variables such as precipitation the number of past patients, mortality rate, susceptible cases, days, curfews, laboratory tests, number of daily cases, cumulative cases, cumulative number of deaths, cumulative number of recovered patients, number of previous cases, daily temperature, and rainfall as independent variables. In this study, in spite of other studies in the literature, Covid-19 patient numbers are predicted in response to the pollutant parameters. The reason to use these parameters is their effects on the formation and progressing of the disease. At the same time, this study will take a base role on helping to understand how environmental factors are affecting the tendency of the pandemic.

Methodology

The present study begins with an assessment of the air pollutants' effects on Covid-19 patient numbers, followed by determining to what extent air pollutant concentrations can explain Covid-19 patient numbers, using IBM SPSS 22 software. The rate thus found is important in explaining patient numbers, as it affects the infection rates. After the determination of this rate, the Covid-19 patient numbers based on air pollution concentrations were estimated using multilayer perceptron (MLP), which is the most known Artificial Neural Networks model by MATLAB R2021a.

Study area, air pollutant, and Covid-19 data

The study area is on Turkey, which is divided — for the purposes of Covid-19 data —into 12 regions as shown in Fig. 1: Istanbul, Western Marmara, Aegean, Eastern Marmara, Western Anatolia, Mediterranean, Central Anatolia, Western Blacksea, Eastern Blacksea, North-Eastern Anatolia, Central-Eastern Anatolia, and South-Eastern Anatolia.

The regions used in the study were defined on the basis of the Covid-19 Status Reports published by Turkish Ministry of Health on the website https://covid19.saglik.gov.tr. The Covid-19 patient numbers for the period June 29, 2020–November 23, 2020 as provided in the report were used (TMH 2020a). The reports are public and cover only the stated time frame. Figure 2 shows the daily reported new case and cumulative patient numbers. In parallel, air quality data from 355 stations in 81 provinces of Turkey were

Fig. 1 Study area

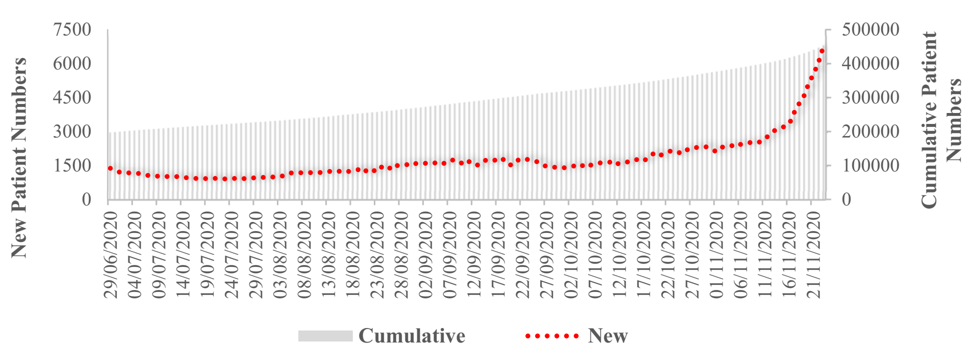

Fig. 2 Daily reported Covid-19 patient numbers for Turkey

also analyzed for the same time frame (EUM 2020). Daily air quality data were downloaded from the website (https://www.havaizleme.gov.tr) of Environment and Urban Ministry and were matched with the regions for which Covid-19 patient numbers were provided. As air quality parameters other than PM10 and PM2.5 were not available for Central-Eastern Anatolia and South-Eastern Anatolia regions, the analysis was carried out with reference to 10 regions. As the number of stations to supply data varied from region to region, median values were used for individual parameters.

The period June 29, 2020–November 23, 2020 is comprised of 148 days. Even though the infectious period for the disease is not established definitely, it is usually considered to be in the 2- to 14-day range (TMH 2020b). Considering the time frame required for the disease to present symptoms after infection, as well as the reports provided in the specified time frame, the present study is based on an infectious period of 3 days. Therefore, the air pollutant parameters are contrasted against the patient numbers to arise on the 3rd day to follow. Therefore, the data for patient numbers pertain to the time frame July 2, 2020–November 23, 2020. Air pollutant parameter and patient numbers data for the 145-day period is available for all regions covered. The patient numbers in specific regions in the relevant time frame are presented in Fig. 3. One should note with reference to this time frame, as the temperatures fall with the onset of fall and the subsequent winter, the resources used for heating are known to increase air pollution levels. The increase trend in patient numbers as shown in the graph is therefore associated with the fall in temperatures.

Artificial neural network

The term artificial neural network (ANN) refers to a machine learning method imitating the learning, remembering, and generalization abilities of the brain, by building a model of biological nervous systems (Cicek and Ozturk 2021). Figure 4 presents ANN typology.

As a concept, ANN is based on the workings of the biological neural network of the human brain. What it tries to achieve is to reproduce the behavior of the biological neuron (Kumar and Kumar 2020; Vakili et al. 2015). As an information processing method inspired by the human brain, it is one of the most effective and successful machine

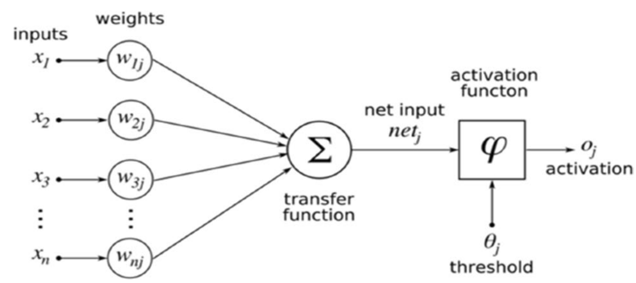

Fig. 4 The typology of ANN

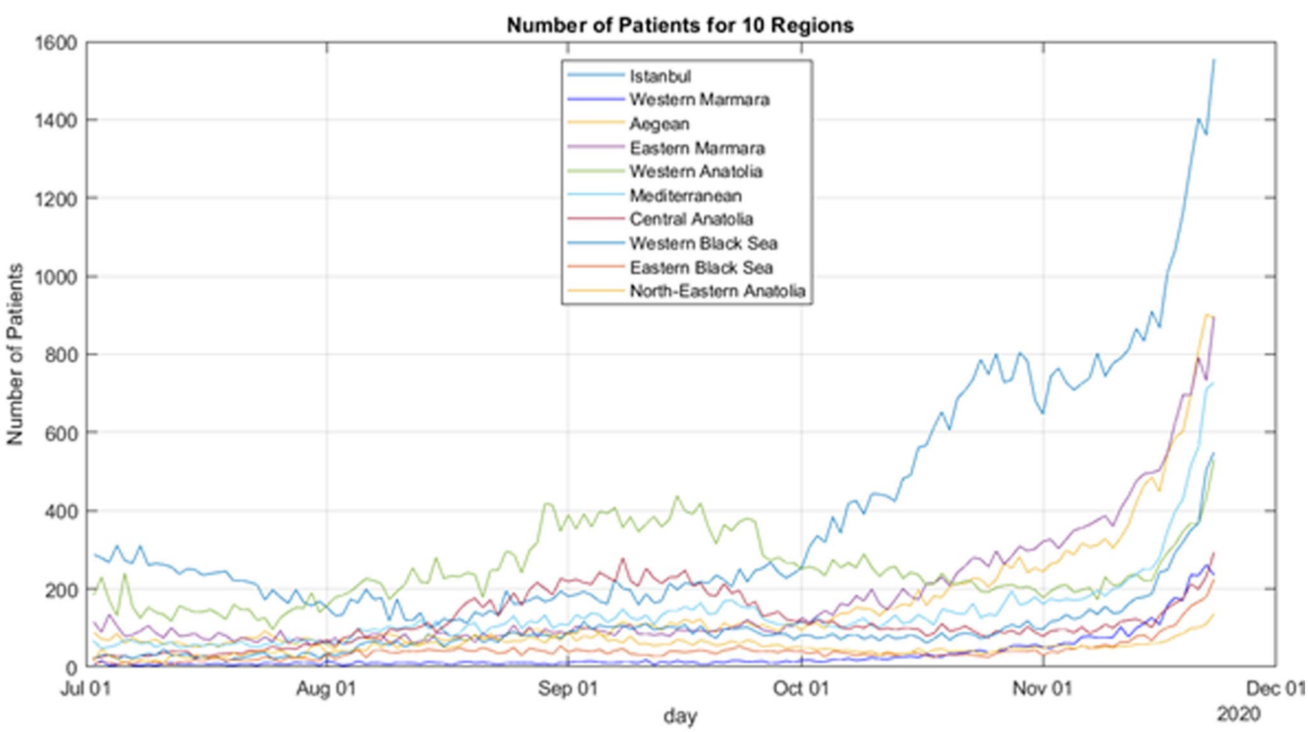

Fig. 3 Patient numbers observed in the period July 2, 2020–November 23, 2020 in 10 regions

learning techniques developed so far. While the human brain learns from human experiences, ANN mimics the brain in its endeavor to process data. The training of ANNs occur as either supervised or unsupervised learning systems, based on whether the output variable values are known in advance. ANNs can be used particularly in cases where no theoretical information about the functional form or the nonlinear structure of the model is available. For training, testing, and validation purposes, data are usually split into specific categories. As the name suggests, the training step involves the network's training — learning — based on the data provided. The validation step allows one to complete the training step by validating the structure of the network, whereas the testing step enables the assessment of the trained ANN's prediction ability (Toğa et al. 2021).

In the present study, a special type of feedforward backpropagation neural network called multilayer perceptron (MLP) is used. The MLP is the most widely used ANN model and usually involves seven inputs, one output, and one or more hidden layers. MATLAB R2021a software was used to try various network topologies, leading to the conclusion that Levenberg–Marquardt BackPropagation (LM-BP) algorithm produced the optimal performance levels. LM is essentially an iteration on the Newton algorithm and is used in forward feeding back propagating algorithms along with ANN. Thanks to its rapid convergence, and the fairly limited number of internal parameters involved, not to mention its dependence on nothing more than first order partial derivatives, LM is considered an advantageous option (Yang et al. 2020).

In the study, the model was run with 1450 data points (10 regions and 145 days of data). A 70% of the dataset was used for training, 15% was used for testing, and 15% was used for validation. Mean square error (MSE) is used as the performance criteria, so as to ensure that the model successfully predicts figures based on pollutant values.

Results of the methods

Variance analysis and correlation

Descriptive statistics for air pollutant and patient numbers data for 10 regions and 145 days were derived using IBM SPSS 22 software. The results are presented in Table 1.

The results of the analysis carried out to assess the covered pollutants' effects on Covid-19 patient numbers are summarized in Table 2. As shown in table, the multiple correlation value is 0.602 while the determination coefficient (R^2) is 0.362. The adjusted R^2 value is 0.359, suggesting that the proposed model is adequate at a rate of 36% for forecasting the Covid-19 patient number with reference to pollution parameters. The last column of the table, on the other hand, reflects the correlation between the actual patient numbers and the predicted numbers.

Table 1 Descriptive statistics of air pollutants and patient numbers

	N	Minimum	Maximum	Mean	Std. deviation
PM10	1450	13.29	106.84	39.8259	15.52068
PM2.5	1450	3.93	61.46	16.0803	8.20807
SO_2	1450	1.64	19.70	5.9096	2.47634
NO_2	1450	5.16	88.92	25.8195	12.38392
NO_X	1450	7.83	211.63	40.7433	27.31094
CO	1450	147.95	1660.66	525.3830	199.13676
O_3	1450	2.56	95.79	45.7669	20.12911
NEW_PATIENT	1450	4.00	1557.00	144.6614	168.25173
Valid N (listwise)	1450				

Table 2 Model's results

Model	*R*	R^2	R^2 (adj)	Std. error of the estimate	R^2 (predicted)
1	0.602[a]	0.362	0.359	134.71655	0.352

[a]Predictors: (constant), (PM10, PM2.5, SO_2, NO_X, NO_2, CO, O_3)

The analysis based on the pollutant and patient count data was followed by the running of the ANN method, leading to the results presented below.

Results of the artificial neural network

As the values included in the dataset used had different ranges, Min–Max normalization was applied to bring all values to the [0,1] range prior to use. To estimate patient numbers, the present study used a network topology involving 7 inputs, 1 output, and 1 hidden layer. Figure 5 presents the network topology used. It has been observed that the combination of Hyperbolic-Tangent (tansig) in the hidden layer and Logarithmic-Sigmoid (logsig) in the output layer gives the best performance as the activation function. The optimum neuron number for the hidden layer was found to be 15 through a comparison of the mean square error values chosen as performance criteria of training and validation set. The performance levels measured are presented in Table 3. As seen in table, it is possible to predict the patient numbers after 3 days with potential daily pollutant parameter values accurately by this model. In Fig. 6, it is seen that the actual and the predicted values overlap. Moreover, as Fig. 6 shows, this model is arguably capable of producing patient number predictions with reference to air pollutant parameters, at a rate of 97%.

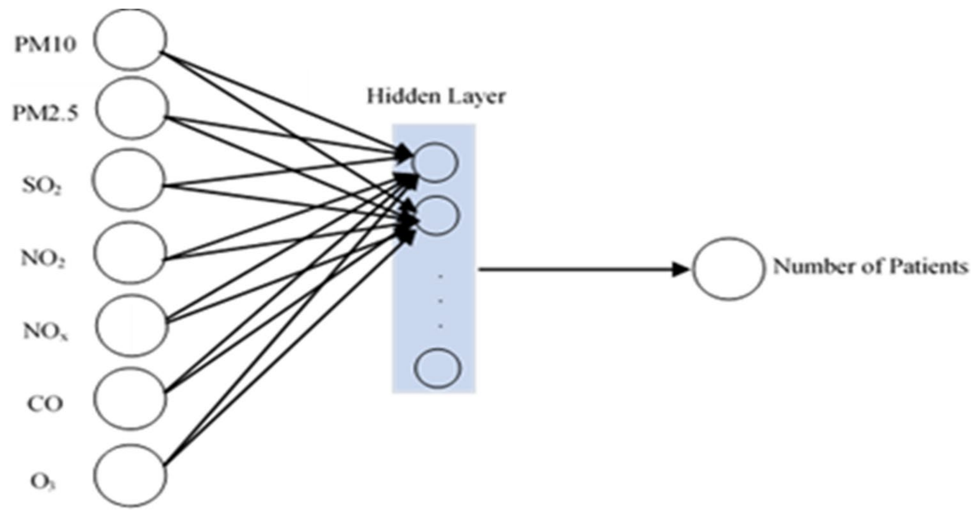

Fig. 5 MLP network topology used in the study

Table 3 Best network architecture and network's performance

	Training perf	Testing perf	Validation perf	Training algorithm	Error function	R^2	Optimal neuron number
MLP-BP (7–15-1)	0.98193	0.95933	0.89427	LM	MSE	0.97342	15

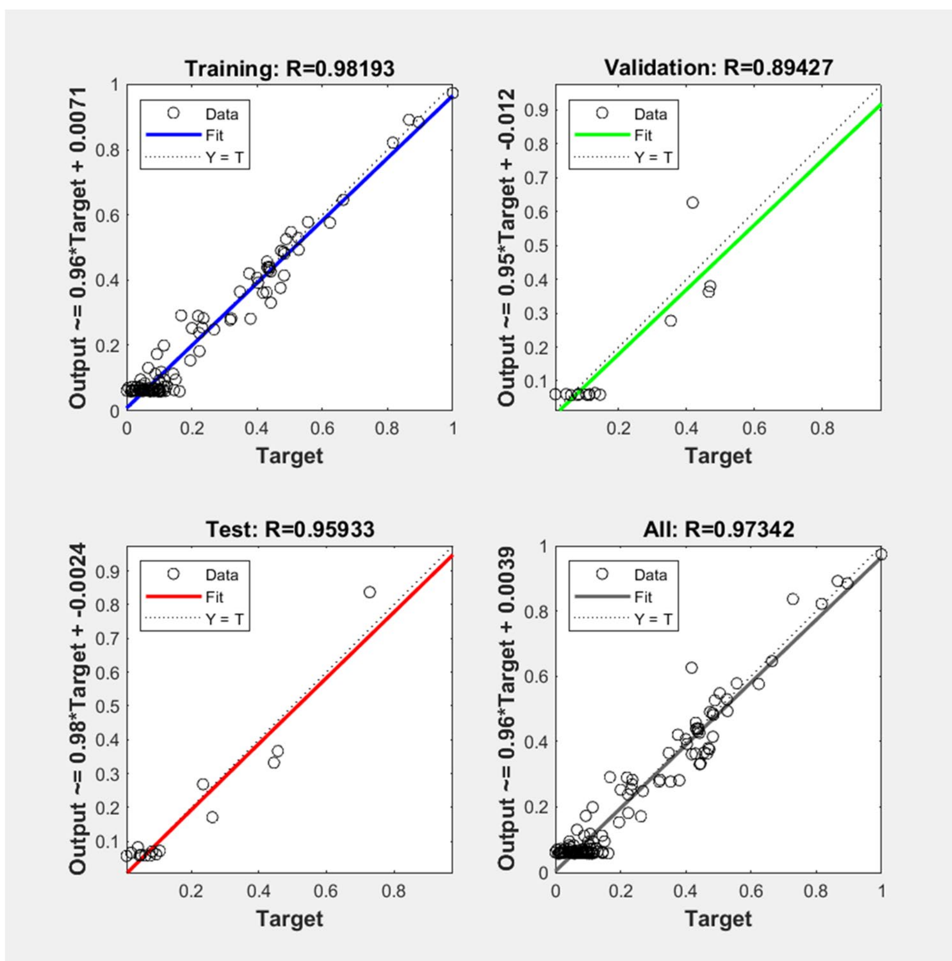

Fig. 6 Neural network training, validation and testing regression plot

Based on the results obtained, a simulation of the MLP model was ran to assess the output performance versus target values for predicting Covid-19 patient number for training, validation, and test stages, using all data available. The results are presented in Fig. 6. The training and testing phase performance obtained by R^2 values are 0.98193 and 0.95933, respectively. The R^2 of validation is 0.89427, showing that the training data is well prepared, and that training accuracy is strong. As shown in Fig. 6, the coefficient (R^2) value exhibits the relatively high correlation between the output and target values for forecasting Covid-19-related patient numbers in Turkey.

Conclusion and discussion

Air pollution poses an important and current environment issue, which should be tackled on all fronts, as studies show that people who are exposed to long-term air pollution have a higher risk of catching and being adversely affected by viruses such as Covid-19 due to emerging chronic diseases. Each 10 μg/m^3 increase in particulate matter (PM10) in the air can cause an increase of 0.7% in problems originating from the cardiovascular system and 1.4% in health problems originating from the respiratory tract (Ayta 2020). When the studies in the literature are examined, it has been seen that patient numbers, mortality numbers, and the spread rate are estimated against different input parameter values concerning Covid-19. Yet, none attempted to estimate patient numbers based on air pollutant parameters. Therefore, in this study, air pollutant parameters (PM10, PM2.5, SO_2, NO_X, NO_2, CO, O_3), which are one of the factors that affect the Covid-19 disease and increase the spread rate of the disease, were examined and patient numbers were estimated using these parameters.

The analysis was performed for the purpose of defining how input parameters are affecting the output parameters, and as a result, multiple correlation value of 0.602 and determination coefficient (R^2) value of 0.362 were found. Adjusted R^2 value is 0.359. Respiratory system diseases are affected by many factors such as personal factors, allergens, and environmental factors. It was found that air pollutant parameters, which is one of the factors affecting the Covid-19 respiratory tract disease, affect the number of patients by 36% by the established regression model. The model was evaluated with analysis of variance, and it was decided that the model was significant since Sig (0.000) < 0.005.

After this relationship, a multilayer perceptron model consisting of 7 input, 1 output, and 1 hidden layer was developed in the study to estimate the number of patients depending on the air pollutant parameters. In the model, Levenberg–Marquardt method was used as the training algorithm and mean square error was used as the performance criteria. It was seen that the number of patients after 3 days was predicted accurately at the rate of 97% against the air pollutant values by this model. In the study, the daily values of the input parameters were used as the median value due to the data obtaining from different monitoring stations. It is expected that the developed model will present a highly accurate estimation result in case of arranging the observation values of the data set by different measures of central tendency.

The Covid-19 outbreak has revealed the necessity of reconsidering the importance of air quality in order to protect human health during and after the pandemic. Strategic decision making can immensely benefit from not only monitoring air quality using an adequate monitoring network and providing required warnings along with efforts to ameliorate air quality using emission control mechanisms, but also understanding of the relationship between pollutant concentration levels in a region and atmospheric events or health followed by forecasts based on that understanding. Making estimations about the future by the help of predictive models such as neural networks should be considered as a useful tool for versatile struggle against pandemics. The results obtained by this method will inform how to struggle the spread of the pandemic with appropriate data integration and analysis. The relationship revealed by the analysis should be evaluated as the determination of the effect amounts of air pollution among environmental factors.

For future studies, it is planned to use different techniques such as other central tendency and distribution measures, together with time series, which will best represent the daily pollutant parameter values of the regions. Thus, it will be possible to predict the number of patients for the future periods by accurate pollutant parameters.

Author contribution Gülşen Aydın Keskin: investigation, methodology, IBM SPSS 22 software usage, writing original draft, writing-review & editing.

Şenay Çetin Doğruparmak: investigation, data preprocessing, writing original draft, writing-review & editing.

Kadriye Ergün: investigation, data preprocessing, methodology, multilayer perceptron (MLP) model usage, MATLAB R2021a software usage, writing original draft, writing-review & editing.

Declarations

Ethics approval The regions used in the study were defined on the basis of the Covid-19 Status Reports published by Turkish Ministry of Health on the website https://covid19.saglik.gov.tr. Covid-19 patient numbers were used mentioned in the report.

Daily air quality data were downloaded from the website https://www.havaizleme.gov.tr of Environment and Urban Ministry. Everyone can access to the data.

Consent to participate Not applicable.

Consent to publish Not applicable.

Competing interests The authors declare no competing interests.

References

Aabed K, Lashin MMA (2021) An analytical study of the factors that influence COVID-19 spread. Saudi J Biol Sci 28:1177–1195

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

Ahangar RG, Pavur R, Fathi M, Shaik A (2020) Estimation and demographic analysis of COVID-19 infections with respect to weather factors in Europe. J Bus Anal 3(2):93–106

Alaraj M, Majdalawieh M, Nizamuddin N (2021) Modeling and forecasting of COVID-19 using a hybrid dynamic model based on SEIRD with ARIMA corrections. Infect Dis Model 6:98–111

Ayta S (2020) The effect of air pollution on our genes (based on epigenetic changes), 4th International 22nd National Public Health Congress. ISBN:978–605–69842–3–5. Çankaya, Ankara. http://www.hasuder.org.tr. Accessed in April 2020

Balcı, Y and Çetin, G (2019) The effects of the COVID-19 pandemic process on employment in Turkey and the measures to be taken in terms of the public. Istanbul Commerce University Journal of Social Sciences Covid-19 Social Sciences Special Issue, 37 Spring (Special Supplement): 40–58

Behnam A, Jahanmahin R (2021) A data analytics approach for COVID-19 spread and end prediction (with a case study in Iran). Model Earth Syst Environ. https://doi.org/10.1007/s40808-021-01086-8

Biswaranjan P, Pawan KA (2021) Air pollution by NO2 and PM2.5 explains COVID-19 infection severity by overexpression of angiotensin-converting enzyme 2 in respiratory cells: a review. Environ Chem Lett 19:25–42

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

Conticini E, Frediani B, Caro D (2020) Can atmospheric pollution be considered a co-factor in extremely high. Environ Pollut 261:11446. https://doi.org/10.1016/j.envpol.2020.114465

Chaurasia V, Pal S (2021) Application of machine learning time series analysis for prediction COVID-19 pandemic. Res Biomed Eng. https://doi.org/10.1007/s42600-020-00105-4

Cicek ZIE, Ozturk ZK (2021) Optimizing the artificial neural network parameters using a biased random key genetic algorithm for time series forecasting. Appl Soft Comput 102:107091

Google Earth (2020) https://earth.google.com/web/. Accessed in Jan 2020

Ergül B, Yavuz AA, Aşık EG, Kalay B (2020) Dünya'da ve Türkiye'de Nisan Ayı İtibariyle COVID-19 Salgın Verilerinin İstatistiksel Değerlendirilmesi. Anad Clin J Med Sci 25(1):130–141

Eroğlu Y (2020) Forecasting models for Covid-19 cases of turkey using artificial neural networks and deep learning. J Ind Eng 31(3):354–372

EUM (Environment and Urban Ministry) (2020) Air quality monitoring. Available at: https://www.havaizleme.gov.tr. accessed in Jan 2021

Göreke V, Sarı V, Kockanat S (2021) A novel classifier architecture based on deep neural network for COVID-19 detection using laboratory findings. Appl Soft Comput J 106:107329

Guo Q, He Z (2021) Prediction of the confirmed cases and deaths of global COVID-19 using artificial intelligence. Environ Sci Pollut Res 28:11672–11682

Gupta AK, Singh V, Mathur P, Carlos M, Gonzalez T (2021a) Prediction of COVID-19 pandemic measuring criteria using support vector machine, prophet and linear regression models in Indian scenario. J Interdiscipl Math 24(1):89–108

Gupta KD, Dwivedi R, Sharma DK (2021b) Prediction of Covid-19 trends in Europe using generalized regression neural network optimized by flower pollination algorithm. J Interdiscip Math 24(1):33–51

Hamit C, Nazmiye Y, Samettin G (2021) The spread of COVID-19 virus through population density and wind in Turkey cities. Sci Total Environ 751:141663

Hasan N (2020) A methodological approach for predicting COVID-19 epidemic using EEMD-ANN hybrid model. Internet of Things 11:100228

Hu L, Deng WJ, Ying GG, Hong H (2021) Environmental perspective of COVID-19: atmospheric and wastewater environment in relation to pandemic. Ecotoxicol Environ Saf 219:112297

Khan FM, Gupta R (2020) ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. J Saf Sci Resil 1:12–18

Khojasteh DN, Goudarzi G, Taghizadeh-Mehrjardi R, Asumadu-Sakyi AB, Fehresti-Sani M (2021) Long-term effects of outdoor air pollution on mortality and morbidity–prediction using nonlinear autoregressive and artificial neural networks models. Atmos Pollut Res 12:46–56

Kumar G, Kumar RR (2020) A correlation study between meteorological parameters and COVID-19 pandemic in Mumbai, India. Diabetes Metab Syndr 14:1735–1742

Kuvvetli Y, Deveci M, Paksoy T, Garg H (2021) A predictive analytics model for COVID-19 pandemic using artificial neural networks. Decis Anal J 1:100007

Long J, Khaliq AQM, Furati KM (2021) Identification and prediction of time-varying parameters of COVID-19 model: a data-driven deep learning approach, International Journal of Computer Mathematics. Int J Comput Math 98(8):1617–1632

Magazzino C, Mele M, Schneider N (2020) The relationship between air pollution and COVID-19-related deaths: an application to three French cities. Appl Energy 279:115835

Magazzino C, Mele M, Sarkodie SA (2021) The nexus between COVID-19 deaths, air pollution and economic growth in New York state: evidence from deep machine learning. J Environ Manag 286:112241

Malki Z, Atlam ES, Ewis A, Dagnew G, Ghoneim OA, Mohamed AA, Abdel-Daim MM, Gad I (2021) The COVID-19 pandemic: prediction study based on machine learning models. Environ Sci Pollut Res 10:1–11

Mele M, Magazzino C, Schneider N, Strezov V (2021) NO2 levels as a contributing factor to COVID-19 deaths: the first empirical estimate of threshold values. Environ Res 194:110663

Mohammadi F, Pourzamani H, Karimi H, Mohammadi M, Mohammadi M, Ardalan N, Khoshravesh R, Pooresmaeil H, Shahabi S, Sabahi M, Miryonesi FS, Najafi M, Yavari Z, Mohammadi F, Teiri H, Jannati M (2021) Artificial neural network and logistic regression modelling to characterize COVID-19 infected patients in local areas of Iran. Biomed J 44(3):304–316

Noshad A, Arjomand P, Khonaksar A, Iranpour P (2021) A novel method for detection of COVID-19 cases using deep residual neural network. Comput Methods Biomech Biomed Eng Imaging Vis 9:555–564

Paital B, Agrawal PK, (2021) Air pollution by NO2 and PM2.5 explains COVID-19 infection severity by overexpression of angiotensin-converting enzyme 2 in respiratory cells: a review. Environ Chem Lett 19:25–42

Rasjid ZE, Setiawan R, Effendi A (2021) A comparison: prediction of death and infected COVID-19 cases in Indonesia using time series smoothing and LSTM neural network. Proc Comput Sci 179:982–988

Reddy KSS, Reddy YCAP, Rao ChM, (InPress) Recurrent neural network based prediction of number of COVID-19 cases in India. Materials Today: Proceedings. https://doi.org/10.1016/j.matpr.2020.11.117.

Shaban WM, Rabie AH, Saleh AI, Abo-Elsoud MA (2021) Detecting COVID-19 patients based on fuzzy inference engine and Deep Neural Network. Appl Soft Comput J 99:106906

Shatnawi N, Abu-Qdais H (2021) Assessing and predicting air quality in northern Jordan during the lockdown due to the COVID-19 virus pandemic using artificial neural network. Air Qual Atmos Health 14:643–652

Silva RG, Ribeiro MHDM, Mariani VC, Coelho LDS (2020) Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. Chaos, Solitons Fractals 139:110027

Singh S, Singh Parmar K, Kaur J, Kumar J, Jitendra S, Makkhan S (2021) Prediction of COVID 19 pervasiveness in six major afected states of India and two stage variation with temperature. Air Qual Atmos Health 14(12):2079–2090

THHP (2021) Clean Air Right Platform, Black Report 2021. https://www.temizhavahakki.com Accessed in Oct 2021

TMH (Turkish Ministry of Health) (2020a) COVID-19 information platform. https://covid19.saglik.gov.tr. Accessed in Jan 2020a

TMH (Turkish Ministry of Health) (2020b) Covid-19 (Sars-Cov-2 infection) general information, epidemiology and diagnosis, Scientific Advisory Board Study. https://covid19.saglik.gov.tr/TR-66337/genel-bilgiler-epidemiyoloji-ve-tani.html. Accessed in Jan 2021

Toğa G, Atalay B, Toksari MD (2021) COVID-19 prevalence forecasting using autoregressive integrated moving average (ARIMA) and artificial neural networks (ANN): Case of Turkey. J Infect Public Health 14(7):811–816

Vakili M, Sabbagh-Yazdi SR, Kalhor K, Khosrojerdi S (2015) Using artificial neural networks for prediction of global solar radiation in Tehran considering particulate matter air pollution. Energy Proc 74:1205–1212

Velasquez RMA, Lara JVM (2020) Gaussian approach for probability and correlation between the number of COVID-19 cases and the air pollution in Lima. Urban Clim 33:100664

Wang P, Zheng X, Ai G, Liu D, Zhu B (2020) Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran. Chaos, Solitons Fractals 140:110214

Wieczorek M, Siłka J, Wozniak M (2020) Neural network powered COVID-19 spread forecasting model. Chaos, Solitons Fractals 140:110203

Yang J, Shen Z, Dong X, Shang X, Li W, Xiong G (2020) The prediction of the epidemic trend of COVID-19 using neural networks. IFAC-PapersOnLine 53:857–862

Yudistira N, Sumitro SB, Nahas A, Riama NF (2021) Learning where to look for COVID-19 growth: multivariate analysis of COVID-19 cases over time using explainable convolutional-LSTM. Appl Soft Comput 109:107469

Zhu Y, Xie J, Huang F, Cao L (2020) Association between short-term exposure to air pollution and COVID-19 infection: evidence from China. Sci Total Environ 727:138704

Authors and Affiliations

Gülşen Aydın Keskin[1] · Şenay Çetin Doğruparmak[2] · Kadriye Ergün[1]

[1] Faculty of Engineering, Department of Industrial Engineering, Balikesir University, Balikesir, Turkey

[2] Faculty of Engineering, Department of Environmental Engineering, Kocaeli University, Kocaeli, Turkey