Research Article/Araştırma Makalesi

GAZİ
JOURNAL OF ENGINEERING SCIENCES

Prediction of Tropospheric Ozone Concentration with Bagging-MLP Method

Pınar Cihan[*a], Husayin Kurtulus Ozcan[b], Atakan Ongen[c]

Submitted: 18.05.2023 Revised: 18.07.2023 Accepted: 31.08.2023 doi:10.30855/gmbd.0705087

ABSTRACT

Human activities are linked to atmospheric pollution and are affected by economic development. Ground-level ozone has become an important and harmful pollutant for many countries, adversely affecting public health. As there is a limited number of on-site measurements, alternative methods are required to accurately estimate ozone concentrations. In this study, a database containing annual average concentrations of CO_2, N_2O, CO, NO_x, SO_x, and O_3, covering the years 2008-2018 for ten countries in Europe, was created. Ten different artificial intelligence regression methods were developed to predict the O_3 concentration using these variables. The predictive performance of the developed artificial intelligence models was compared using the coefficient of determination, mean absolute error, root mean square error, and relative absolute error criteria. Experimental results show that the Bagging-MLP method has a better predictive performance than other models in ozone concentration estimation, with an R^2 value of 0.9994, mean absolute error of 24.67, root mean square error of 33.85, and relative absolute error of 2.9%. This study shows that the O_3 concentration can be estimated very close to the actual value by using the Bagging-MLP method, one of the artificial intelligence methods.

Keywords: Artificial intelligence, bagging, multilayer perceptron, urban environment, ozone

[a,*] Tekirdag Namık Kemal University,
Corlu Engineering Faculty,
Dept. of Computer Engineering
59860 - Tekirdağ, Türkiye
Orcid: 0000-0001-7958-7251
e mail: pkaya@nku.edu.tr

[b] Istanbul University-Cerrahpasa University,
Engineering Faculty,
Dept. of Environmental Engineering
34320 - İstanbul, Türkiye
Orcid: 0000-0002-5979-4197

[c] Istanbul University-Cerrahpasa University,
Engineering Faculty,
Dept. of Environmental Engineering
34320 - İstanbul, Türkiye
Orcid: 0000-0002-9043-7382

*Corresponding author: pkaya@nku.edu.tr

Bagging-MLP Yöntemiyle Troposferik Ozon Konsantrasyonunun Tahmini

ÖZ

İnsan faaliyetleri atmosfer kirliliği ile bağlantılıdır ve ekonomik gelişmelerden etkilenir. Yer seviyesindeki ozon birçok ülke için önemli ve zararlı bir kirletici haline gelmiş olup halk sağlığını olumsuz etkiler. Yerinde yapılan ölçümlerin sınırlı sayıda olmasından dolayı, ozon konsantrasyonlarını doğru bir şekilde tahmin etmek için alternatif yöntemlere ihtiyaç vardır. Bu çalışmada, Avrupa'da on ülkede 2008-2018 yıllarını kapsayan CO_2, N_2O, CO, NO_x, SO_x, ve O_3 yıllık ortalama konsantrasyonlarını içeren bir veritabanı oluşturuldu. Bu değişkenleri kullanarak O_3 konsantrasyonunu tahmin etmek için on farklı yapay zeka regresyon yöntemi geliştirildi. Geliştirilen yapay zeka modellerinin tahmin performansı, determinasyon katsayısı, ortalama mutlak hata, kök ortalama karesel hata ve göreceli mutlak hata ölçütleri kullanılarak karşılaştırıldı. Deneysel sonuçlar, Bagging-MLP yönteminin diğer modellere göre ozon konsantrasyonu tahmininde daha iyi bir performansa sahip olduğunu, R^2 değeri 0.9994, ortalama mutlak hata 24.67, kök ortalama karesel hata 33.85 ve göreceli mutlak hata ise %2.9 olarak ortaya koydu. Bu çalışma, yapay zeka yöntemlerinden olan Bagging-MLP yöntemi kullanılarak O_3 konsantrasyonunun gerçek değere oldukça yakın bir şekilde tahmin edilebileceğini göstermektedir.

Anahtar Kelimeler: Yapay zeka, torbalama yöntemi, çok katmanlı algılayıcı, kentsel çevre, ozon

To cite this article: P. Cihan, H. K. Ozcan and A. Ongen, "Prediction of Tropospheric Ozone Concentration with Bagging-MLP Method," *Gazi Journal of Engineering Sciences,* vol. 9, no. 3, pp. 557-573, 2023. doi:10.30855/gmbd.0705087

1. Introduction

Ozone (O_3), which was discovered in the mid-19th century, is a reactive oxidizing gas that occurs naturally in trace amounts in the Earth's atmosphere. It is a relatively unstable molecule made up of three atoms of oxygen (O), blue in color, and has a strong odor. Although ozone represents only a tiny fraction of the atmosphere, it is crucial for life on Earth and it plays a key role in atmospheric chemistry and the overall radiative balance of the atmosphere [1]. For example, most of the ozone in the stratospheric ozone layer (a layer 12–48 m above the Earth) acts as a shield to protect the Earth's surface from the Sun's harmful ultraviolet radiation [2]. Approximately 90% of atmospheric ozone is found between the top of the troposphere layer and within the stratospheric layer at an altitude of about 50 km. The remaining 10% of atmospheric Ozone is present in the lower parts of the atmosphere (the Troposphere), which is very close to the earth's surface.

The troposphere, which begins at the Earth's surface, is composed of multiple layers and stretches from 8 to 14.5 kilometers above the Earth's surface. When present in high concentrations, tropospheric ozone is a photochemical oxidizing gas that harms the environment and human health. Tropospheric ozone, which causes photochemical smog, is a secondary pollutant that forms when the concentration of primary pollutants like hydrocarbons and nitrogen oxides (NO_x) rises during peak hours. At a concentration of 0.15 ppm, it can cause burning in the eyes and at 0.25 ppm it is considered hazardous to human health [3]. In addition to the negative effects on human health; oxidizing substances in the atmosphere reduce visibility and it has been observed by many researchers that tropospheric O_3 has adverse effects on rubber, plastics, and paints [4]. Furthermore, tropospheric O_3 is also the third-largest greenhouse gas, contributing about 3%-7% of the greenhouse effect, and has a substantial impact on climate change [5].

In recent years human activities have caused a dramatic increase in ozone concentrations. In the atmosphere, ozone is formed only as a result of the reaction between atomic oxygen and molecular oxygen. However, the troposphere is an environment where many oxidation reactions occur. Under the influence of daylight, the oxidation of organic molecules in the presence of nitrogen oxides takes place, and the components in the troposphere tend to move towards a more oxidized state. Thus, the primary product in troposphere chemistry is thought to be ozone. Ground-level ozone is less concentrated compared to ozone in the upper atmosphere; however, it is considered more dangerous due to its hazardous nature and the risk it poses to public health and well-being. Ozone concentration varies between large cities and rural areas, as ozone formation is entirely related to other pollutants released into the atmosphere. Hence, the determination and modeling of the relationship between tropospheric ozone concentrations and other components in the atmosphere has been studied extensively [6-11].

The formation and distribution of ground-level ozone compounds depend on factors such as altitude, land use type, atmospheric components, and some meteorological factors such as temperature, wind, sunlight, humidity, and precipitation [12]. Statistical models are used to directly determine the relationships between the tropospheric ozone concentration and these variables. Based on the temporal and spatial variations in these factors the models have the potential to predict the ozone concentration when and where monitoring points are deficient [13].

Artificial intelligence is the general name of computer algorithms that model a problem situation according to the data belonging to that problem [14]. Artificial intelligence uses the information obtained from previous experiences, examines the new information in this direction, and constantly tries to improve its performance. The main purpose of artificial intelligence is to make inferences using the information that already exists without any additional intervention from the outside and to make these inferences ready to be used in future estimations or when appropriate [15]. Artificial intelligence algorithms are widely used in most applications due to their unique nature of problem-solving. Such algorithms deal with the construction of machines that move automatically by gaining experience, the formation of these algorithms with low computational costs, the design of new algorithms and the usability of big data have made progress in recent years. Since artificial intelligence has a very wide usage area, studies on artificial intelligence can be found in nearly every subject when reviewing the literature [16-28]. Due to artificial intelligence, computers can be programmed to perform specific tasks or process [29], desired classifications can be made [30-33], models can be designed, and these models can make predictions about the future [21, 34], based on previous experiences or dataset presented as examples [35].
It is observed in the literature that ozone concentration is successfully predicted using artificial intelligence methods. Although studies may have different results based on the methods or data used, overall, they

demonstrate the successful application of artificial intelligence methods. Juarez and Petersen [36] used XGBoost, Random Forest (RF), K-Nearest Neighbor Regression (K-NNR), Support Vector Regression (SVR), Decision Trees (DT), AdaBoost, LSTM, and Linear Regression (LR) methods to predict the ozone level in Delhi, India. The study utilized a dataset containing 12 air pollutants and 5 weather variables recorded hourly throughout one year (2015). Each model was trained and tested ten times. The performance of the models was compared using the determination coefficient (R^2) statistical criterion. According to the findings of the study, the most successful method was XGBoost with an R^2 value of 0.614. Additionally, predictions were made based on seasons, and during the winter period, the XGBoost method exhibited an approximate prediction success rate of 97%

Jumin et al. [37] used LR, Neural Network (NN), and Boosted Decision Tree (BDT) methods to predict the tropospheric ozone concentration in Malaysia, Kuala Lumpur, and Selangor. The model prediction performances were compared using the R^2 statistical criterion. The dataset used in the study consists of variables such as humidity, wind speed, nitrogen oxide, nitrogen dioxide, sulphur dioxide, carbon monoxide and ozone. 75% of the dataset was used for training, and 25% for testing. According to the findings of the study, the most successful method was BDT. The R^2 values for the proposed method of ozone concentration in these three regions were determined as 0.87, 0.88, and 0.91, respectively.

Pan et al. [38] used 19 machine learning (ML) methods for predicting ozone pollution. To compare the prediction performances of the methods, they utilized R^2, RMSE, MAPE, MAE, and J^2 metrics. The study employed air pollution and meteorological data collected at King Abdullah University of Science and Technology in Saudi Arabia. The data was collected every 15 minutes from May 20 - Dec 20, 2020, and Jan 21-Oct 21, 2021. The findings of the study reported that the SVR method outperformed other ML models.

Wang et al. [39] proposed a random forest model for predicting ground-level ozone concentrations in California. The study utilized Troposphere Monitoring Instrument (TROPOMI) and High-Resolution Rapid Refresh (HRRR) data. According to the obtained results, it was reported that daily surface ozone concentration was predicted with an R^2 value of 84%. Three cross-validation (CV) strategies were applied to evaluate the model performance.

Yafouz et al. [40] aimed to predict ozone intensity using various ML models such as LR, Tree Regression (TR), Support Vector Regression (SVR), Gaussian Process Regression (GPR), Ensemble Regression (ER), and Artificial Neural Network (ANN). The data used in the study was hourly averaged from three different stations located in Putrajaya, Kelang, and KL on the Malay Peninsula. According to the findings of the study, the best prediction with an R^2 value of 0.89 was achieved using LR, SVR, GPR, and ANN methods with the data obtained from the KL station.

Aljanabi et al. [41] aimed to predict ozone concentration using a combination of meteorological and seasonal variable data for Amman City in Jordan. For this purpose, they compared MLP, SVR, DTR, and XGBoost methods. In the study, they reported that MLP outperformed other algorithms and the use of Savitzky-Golay improved the R^2 by 50% and the RMSE and MAE by 80%. Feature selection was applied to predict ozone concentration, and they obtained an approximate R^2 score of 98%.

To estimate ozone concentrations using artificial intelligence approaches this study created a database that contains the annual average concentrations of Carbon dioxide (CO_2), Carbon monoxide (CO), Nitrous oxide (N_2O), Nitrogen oxides (NO_x), Sulfur oxides (SO_x), and Ozone (O_3). Data from 2008 to 2018 was collected from ten European countries (Czechia, Germany, Greece, Spain, France, Italy, Romania, Switzerland, United Kingdom, and Turkey) and the accuracy of the predictions was determined. Furthermore, a performance comparison between the selected artificial intelligence models was conducted to determine the most successful method.

2. Material and Methods

The summary of the methodologies for data processing is illustrated in Figure 1.

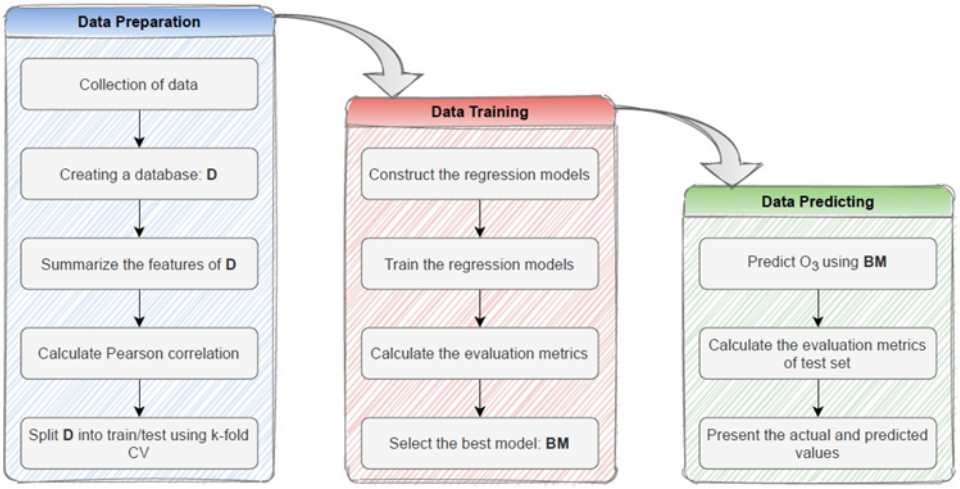

Figure 1. Flowchart for prediction of ozone

2.1. Dataset and study area

In this study, data collected from 10 different European countries was used to predict Ozone concentration with artificial intelligence methods. These countries are Czechia, Germany, Greece, Spain, France, Italy, Romania, Switzerland, United Kingdom, and Turkey as illustrated on the map in Figure 2.

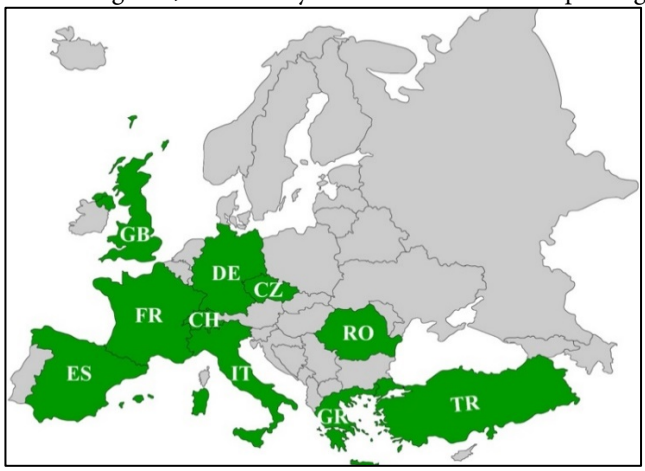

Figure 2. Countries involved in this study

This study's dataset includes input variables such as carbon dioxide (CO_2), Nitrous oxide (N_2O), Carbon monoxide (CO), Nitrogen oxides (NO_x), and Sulfur oxides (SO_x) concentrations. The Ozone (O_3) concentration is used as an output (target) variable. The dataset was collected from the Eurostat website, and it contains the annual average concentrations of selected countries from 2008 to 2018. Input and output variables are summarized in Table 1.

Table 1. Statistics the of variables in the dataset

	CO_2	N_2O	CO	NO_X	SO_X	O_3
Minimum	25959.1	9.3	64.9	71.8	5.9	154.9
1st Quarter	75395.0	17.9	217.6	201.9	137.6	412.7
Median	234325.2	56.2	766.1	680.8	267.9	1578.1
Mean	237126.4	61.7	763.4	914.2	427.7	1409.6
3rd Quarter	315613.1	110.1	1151.0	914.2	437.1	2003.7
Maximum	725664.4	150.3	2190.4	1831.9	2296.7	3563.1
Standard dev.	187421.8	45.7	584.2	493.5	551.8	973.2

2.2. Methods

In this study, nine different regression models were employed to predict ozone concentration with accurately, and the performance of these methods was compared. The regression models used in the study are presented in Table 2.

Table 2. Artificial intelligence regression methods used in this study

Method	Abbreviation
Linear Regression	LR
Multilayer Perceptron	MLP
Support Vector Regression	SVR
Fuzzy k-nearest neighbor	FKNN
K-Nearest Neighbors	KNN
Weighted K-Nearest Neighbors	WKNN
Random Forest Regression	RFR
Bagging MLP	Bagging-MLP
Bagging SVR	Bagging-SVR

Used regression methods are briefly described below.

2.2.1. Linear Regression (LR)

LR is describing the linear relationship between a dependent variable and one or more independent variables. LR involves utilizing weighted samples to construct a prediction model, and it employs least-squares regression to ascertain linear relationships. The following steps are traced in the LR method:

Weights are calculated from the training dataset (Eq. (1)). Weights should be chosen to minimize errors (actual output value - predicted output value).

$$x = w_0 + w_1 a_1 + \cdots + w_k a_k \quad (1)$$

Where x is the output value, a is the input value, and w is the weight of each input attribute (a0 is considered as 1 and w1 is the weight of a1).
The Predicted value for the first training instance a(1) is calculated as shown in Eq. (2).

$$\sum_{j=0}^{k} w_j a_j^{(1)} = w_0 + w_1 a_1^{(1)} + \cdots + w_k a_k^{(1)} \quad (2)$$

Lastly, Weights are updated to minimize the squared error between actual output and predicted output as shown in Eq. (3).

$$\sum_{i=1}^{n} \left(x^{(i)} - \sum_{j=0}^{k} w_j a_j^{(i)} \right) \quad (3)$$

2.2.2. Fuzzy K-Nearest Neighbor (FKNN)

In this method, the concept of fuzzy logic is combined with the k-nearest neighbor technique (KNN). Here, different degrees of membership values are assigned considering the distance of the KNNs. FKNN consists of two steps. In the first step, the KNN's are determined for the training dataset and the fuzzy membership values are estimated for the feature vector. In the second step, the fuzzy membership value is calculated and assigned to the unknown test sample [42].

2.2.3. Multilayer Perceptron (MLP)

MLP method is a feedforward neural network. In a multilayer neural network, the neurons are fully connected, that is, there are connections from the neuron cell in one layer to all the neuron cells in the other layer. Neurons are mapped from input data to a series of outputs with hidden layers, as shown in Figure 3. The most popular learning method in a multilayer neural network is backpropagation. To minimize the error in the output layer, the weights of the neurons between the layers behind are updated with each backpropagation iteration. So the weights on the connections change over time during learning. After a certain repetition, the change of weights decreases, from that moment the system has completed the learning process.

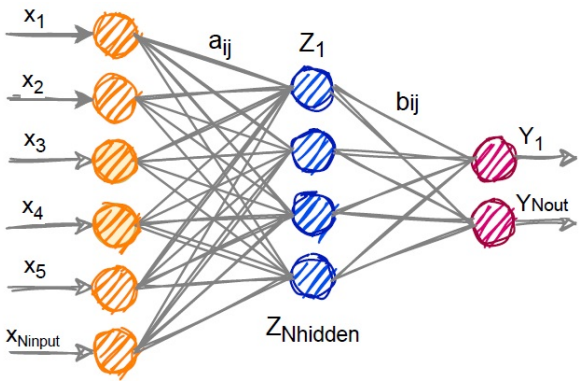

Figure 3. A schematic diagram of the MLP neural network

An MLP with a hidden layer can be mathematically described by the following equations. Weighted sums of inputs are calculated using Eq. (4).

$$u_j = \sum_{i=1}^{N_{inp}} X_i a_{ij} + a_{0j} \quad (4)$$

Where N is the number of input nodes, Xi is the i^{th} input, a_{ij} shows the weight vectors and a_{0j} is the bias of the hidden node.

In Eq. (5), by transforming this sum defined in Equation 4, the outputs of the Z_j hidden layer are obtained. For this g activation function (transfer function) is used.

$$Z_j = g(u_j) \quad (5)$$

The output of each hidden node is based on the sigmoid function and is defined in Eq. (6).

$$g(x) = sigmoid(x) = \frac{1}{(1+e^{-x})} \quad (6)$$

The sum-product of the Z_j hidden layer's outputs and the b_{jk} weight vectors and the b_{0k} bias term of the output layer are calculated using Eq. (7).

$$u_k = \sum_{j=1}^{N_{hid}} Z_j b_{jk} + b_{0k} \quad (7)$$

Using Eq. (6) and Eq. (7), the final outputs are defined as shown in Eq. (8).

$$Y_k = g(u_k) \quad (8)$$

2.2.4. Support Vector Regression (SVR)

SVR is a kernel-based learning algorithm. The basic idea in SVR is to minimize the error by individualizing the hyperplane where the error is maximized. The kernel function is used to map the input data to a higher dimensional feature space through nonlinear mapping. Thus, SVR solves a linear regression problem in this feature space. Although there are different kernel functions, the most commonly used are polynomial, linear, sigmoid (MLP) and Gaussian (RBF) kernels [43]. In this study, a sequential minimal optimization algorithm is applied to train the SVR model.

2.2.5. K-Nearest Neighbor (KNN)

KNN is the most popular of the nearest neighbor approaches [44]. When calculating the output value of a test

sample; first, the distance of the test sample to all training samples is calculated, then the nearest k neighbors are determined, and finally, the output is determined by averaging the value of these k neighbors [45]. In addition, distances can be weighted in the KNN algorithm. Euclidian distance is usually used for distance measurement. In this study, both weighted and unweighted KNN methods were used. Euclidean distance was used for distance measurement and k value 3 was chosen. Eq. (9) is used to estimate the output value of a test sample (q_i).

$$\mathrm{Sim}(q_i;s_j) = \frac{\sqrt{\sum_{f-1}^{N} \delta(q,s,f)^2}}{\sqrt{N}} \quad (9)$$

2.2.6. Random Forest Regression (RFR)

RFR is a tree-based regression method. It is an ensemble learning algorithm developed by Leo Breiman [46] and consists of a combination of many regression trees. Each tree in the forest is trained using a bootstrap sample extracted from the training set. The output values predicted by the models that complete the learning process are combined. In the regression, decisions are aggregated by taking the average of the predicted values.

2.2.7. Bagging (Bootstrap Aggregation)

Bagging is also an ensemble learning algorithm developed by Breiman [47]. The purpose of the bagging algorithm is to generate a large number of similar training sets by taking a random bootstrap sample from the training dataset. These subsets are used for training the base learners. To predict the test set, models that have learned from these subsets are used collectively. Bagging uses averaging to aggregate the outputs of the base learners [48]. Bagging can be formalized in Eq. (10) and the workflow of the bagging technique is illustrated in Figure 4.

$$\hat{y}_{BAG} = \frac{1}{n}\sum_{T=1}^{n} \emptyset(x;T_i) \quad (10)$$

In the equation, x is the input and n is the number of bootstrap samples of training set T.

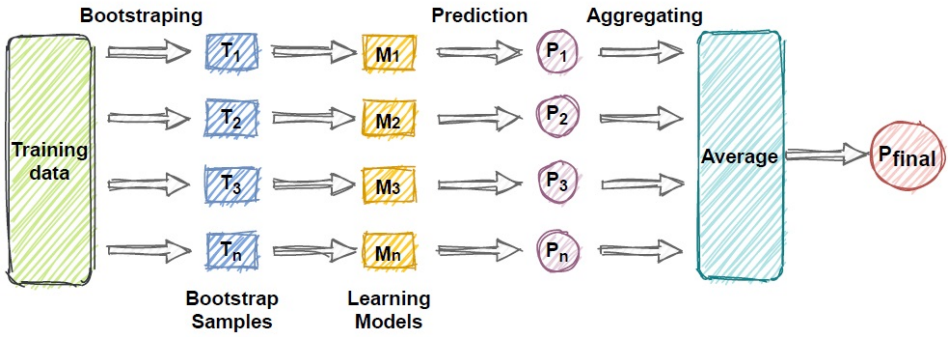

Figure 4. Workflow of the Bagging technique

2.2. Comparison of Models' Performance

In the modeling phase, the dataset is split into two parts apart for training and the other part for testing to perform model training and to test the prediction performance of the model. K-fold cross-validation is a technique used to divide the dataset into training and test set. In this method, the data set is divided into train/test according to the determined number of k. In this study, the k value of 10 was chosen. In other words, our dataset consisting of 110 samples will be divided into 10 folds. In each iteration, the model will test the training process with 99 samples (90%) and the prediction performance of the model with the remaining 11 samples (10%). After 10 iterations are completed, the error of the model is calculated by taking the average. Hence, CV helps to estimate the error of the model and to select the best model. In the CV technique, the entire dataset is used for both training and testing, thus eliminating bias. This method is useful for small datasets as the training and testing process of models is time-consuming. The process of CV is illustrated in Figure 5.

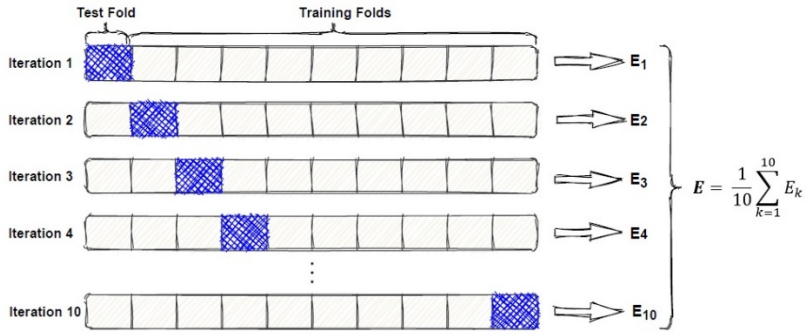

Figure 5. Diagram of k-fold CV with k=10

The performances of the methods used in the study in predicting the Ozone concentration were compared according to the evaluation criteria of R^2, root mean square error (RMSE), mean absolute error (MAE), and relative absolute error (RAE). These metrics are expressed mathematically as in Eqs. (11-14), respectively:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(a_i - p_i)^2}{\sum_{i=1}^{n}(a_i - \bar{a})^2} \tag{11}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{\text{n}}} \tag{12}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|p_i - a_i|}{\text{n}} \tag{13}$$

$$\text{RAE} = \frac{\sum_{i=1}^{n}|p_i - a_i|}{\sum_{i=1}^{n}|\bar{a} - a_i|} \tag{14}$$

In the above equations, p is the predicted value, a is the actual value and $\bar{a}$ is the mean of the actual values. The above three error measurement criteria (RMSE, MAE, RAE) should be lower. Error is zero indicates that it is a statistically perfect model. The R^2 measures how well the predicted values match the actual values. In other words, this value should be high since it shows the predictive accuracy of the model.

3. Experimental Results

In this study, the prediction performances of various regression techniques were evaluated and compared to determine the most successful artificial intelligence regression method in estimating O_3 concentration. Firstly, the cross-correlation of the variables in the dataset, density plots, and 2D density charts are presented in Figure 6.

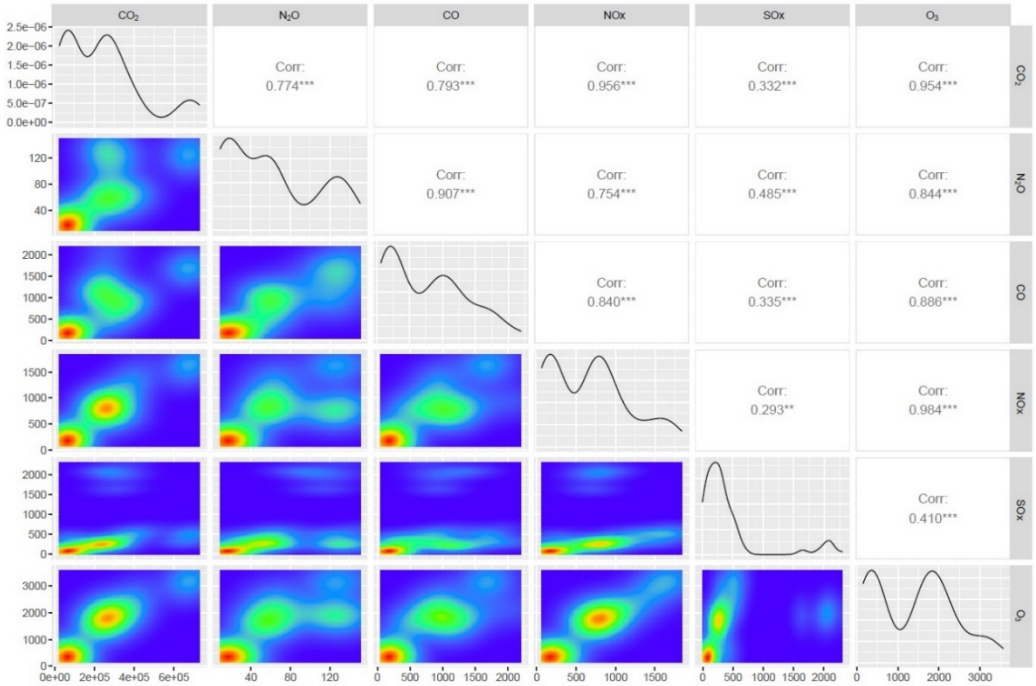

Figure 6. Correlation and density plot of variables in dataset (thousand tonnes)

In Figure 6, it is seen from the density plots of the input variables that they do not have a normal distribution. The CO_2 variable has very high values compared to other variables (see Table 1). The fact that the value ranges of the input variables are different, especially the features with high values such as that of CO_2, affects the success of the methods that are based on distance measurement. This is because variables with high values tend to overshadow the impact of variables with lower values. Box-plot graphs of the input variables are given in Figure 7.

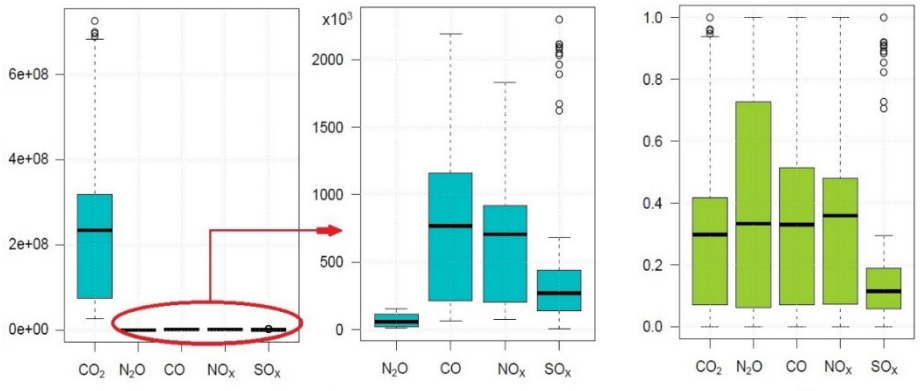

Figure 7. Boxplot of (A) original dataset and (B) normalized dataset

Below, the impacts of each input variable on the target variable are presented.

Effect of Nitrogen oxides (NO_x):

NO_x is a powerful greenhouse gas that is produced during fossil fuel combustion and biomass burning. In the troposphere, NO_2 is the main source that provides the oxygen atoms necessary for O_3 formation. NO_2 is broken down into NO and oxygen atoms by sunlight. Then the oxygen atom combines with the oxygen molecule to form O_3. Therefore, it is expected that there will be a strong correlation between NO_x and O_3. Figure 6 shows that NO_x exhibits a strong influence on O_3, with a correlation of 0.984. Figure 8 illustrates the scatter diagram of NO_x versus O_3 and includes the fitted linear regression model. The result of the predictive model for the dataset is $O_3 = 1.9405 * NO_x + 88.652$. The regression function's slope indicates that a unit increase in NO_x is associated with a rise of 1.9405 thousand tonnes in O_3. It is seen that when NO_x increases,

O_3 concentration also increases.

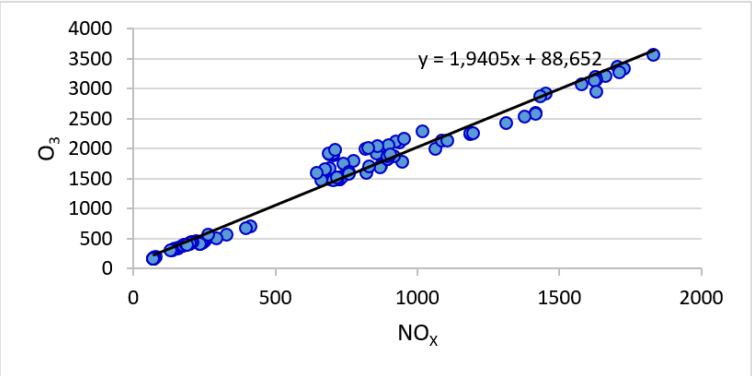

Figure 8. Scatter diagram of NO_x vs. O_3 in thousand tones

Effect of Carbon dioxide (CO_2):

CO_2, a major greenhouse gas, is emitted by both human activities like deforestation and burning fossil fuels, and natural processes including respiration and volcanic eruptions. The second most effective factor for the formation of O_3 is CO_2 with a correlation coefficient of 0.954 (Figure 6). Figure 9 shows the scatter diagram and the linear regression model that fits the data of CO_2 vs. O_3. The result of the predictive model for the dataset is $O_3 = 0.005 * CO_2 + 234.4$. The slope of the function shows that the unit increase in CO_2 corresponds to an increase in O_3 of 0.0005 thousand tones. When all other variables are held constant, the performance is improved with the presence of CO_2.

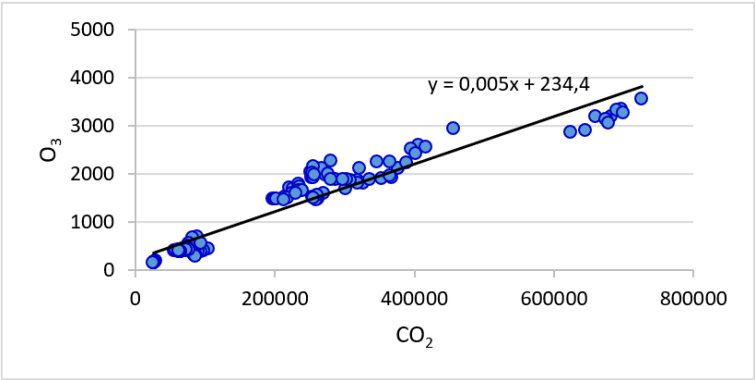

Figure 9. Scatter diagram of CO_2 vs. O_3 in thousand tones

Effect of Nitrous Oxide (N_2O):

N_2O is a substantial contributor to global warming as a greenhouse gas. When considered per molecule over 100 years, nitrous oxide has approximately 265 times the heat-trapping capacity of CO_2 in the atmosphere. However, due to its lower concentration, its overall contribution to the greenhouse effect is less than one-third that of CO_2. Nitrous oxide is emitted as a by-product of burning fossil fuels, though the quantity released varies depending on the type of fuel used. The linear relationship between N_2O and O_3, shows another high correlation of 0.844 with ozone, as seen in Figure 10. The result of the predictive model for the dataset is $O_3 = 17.972 * N_2O + 300.16$.

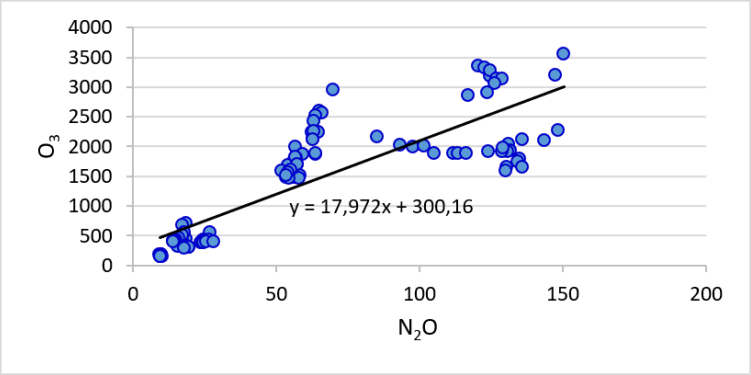

Figure 10. Scatter diagram of N_2O vs. O_3 in thousand tones

Effect of Carbon monoxide (CO):

CO is a colorless, odorless, tasteless, flammable gas that results from the incomplete combustion of carbon. CO could play roles with potential impacts on climate change. It indirectly influences radiative forcing by increasing the concentrations of direct greenhouse gases like methane and tropospheric ozone. Natural atmospheric processes lead to the oxidation of CO to carbon dioxide and ozone [49]. This variable shows another high correlation of 0.886 with O_3. The result of the predictive model for the dataset is O_3 = 1.4762 * CO + 282.65. The linear relationship between CO and O_3 is shown in Figure 11.

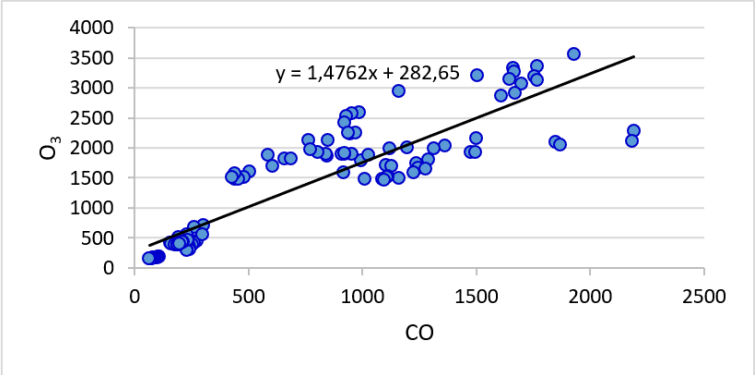

Figure 11. Scatter diagram of CO vs. O_3 in thousand tones

Effect of Sulfur Oxides (SO_x):

SO_x, which stands for compounds composed of sulfur and oxygen molecules. The main form found in the lower atmosphere is sulfur dioxide (SO2). It is a colorless gas that can be detected at concentrations ranging from 1,000 to 3,000 µg/m3 due to its distinct odor and taste. The majority of sulfur dioxide is generated by burning fuels containing sulfur or by roasting metal sulfide ores, while natural sources like volcanoes also contribute to sulfur dioxide emissions, accounting for 35-65% of the total. In comparison to other variables in the dataset, SO_x show the weakest correlation with O_3 (0.41). Figure 12 shows the scatter diagram and the linear regression model that fits the data of SO_x vs. O_3. The result of the predictive model for the dataset is O_3 = 0.7231 * SO_x + 1100.3. The slope of the regression function shows that the unit increase in SO_x corresponds to an increase of 0.7231 thousand tones in O_3. From the correlation value and as seen in Figure 12, SO_x is not self-sufficient for estimation.

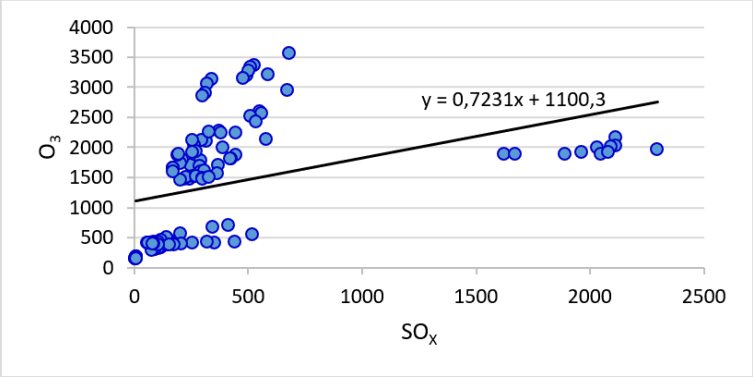

Figure 12. Scatter diagram of SO_x vs. O_3 in thousand tones

Ozone concentration was estimated with different artificial intelligence regression methods and the prediction performances of these models were compared to each other in Table 3. Since the dataset does not have a normal distribution, the data were normalized with the min-max normalization technique. The original dataset and the normalized dataset estimations were made separately, and the results obtained are presented in comparison in Table 3.

Table 3. Comparison of the prediction results of the models for the original and normalized dataset

Method	Original Data				Normalized Data			
	R^2	MAE	RMSE	RAE (%)	R^2	MAE	RMSE	RAE (%)
LR	0.9984	42.266	54.586	4.962	0.9984	0.0124	0.0160	4.962
MLP	0.9990	34.618	44.356	4.064	0.9990	0.0102	0.0130	4.064
SVR	0.9983	42.241	56.889	4.959	0.9983	0.0124	0.0167	4.946
FKNN	0.9967	46.576	80.735	5.467	0.9967	0.0137	0.0237	5.467
KNN	0.9957	54.663	90.653	6.417	0.9957	0.0160	0.0266	6.417
WKNN	0.9962	48.906	85.676	5.741	0.9967	0.0137	0.0237	5.467
RFR	0.9972	46.620	72.764	5.473	0.9973	0.0135	0.0212	5.385
Bagging-MLP	**0.9994**	**24.668**	**33.846**	**2.896**	**0.9994**	**0.0072**	**0.0099**	**2.896**
Bagging-SVR	0.9984	43.016	55.544	5.049	0.9983	0.0125	0.0169	5.007

Table 4 presents the predicted and actual ozone values for the Bagging-MLP method, which outperforms other methods. The study employed the CV technique, and it provides separate data for actual and estimated ozone values (thousand tons) as well as errors for each iteration. When Table 4 is examined, it is seen that the Ozone values estimated by the Bagging with MLP method are quite close to the actual and therefore the estimation errors are quite low.

Table 4. Actual and predicted values for each CV iteration of Bagging-MLP

Iteration 1			Iteration 2			Iteration 3			Iteration 4			Iteration 5		
Actual	Predicted	Error	Actual	Predicted	Error	Actual	Predicted	Error	Actual	Predicted	Error	Actual	Predicted	Error
561.8	598.2	36.3	412.7	411.5	-1.3	670.9	678.0	7.0	2095.4	2105.0	9.6	2862.0	2930.0	68.1
1589.0	1614.9	25.9	2120.0	2086.5	-33.5	3563.1	3446.2	-116.9	1479.6	1439.3	-40.3	422.6	433.2	10.6
1592.6	1603.7	11.1	1699.4	1691.3	-8.2	2046.3	2014.2	-32.1	3133.6	3156.5	22.9	405.8	415.4	9.6
3359.9	3252.6	-107.2	2158.7	2149.8	-8.9	175.2	155.9	-19.3	2595.7	2597.5	1.9	2118.0	2126.0	8.0
1462.6	1396.2	-66.5	295.0	323.0	28.0	380.2	380.7	0.5	323.4	336.5	13.2	2254.9	2249.0	-5.9
2276.7	2280.2	3.5	186.5	156.6	-29.8	169.7	155.2	-14.5	1817.8	1815.6	-2.2	452.9	471.5	18.6
1989.7	1982.3	-7.4	413.7	414.5	0.8	172.9	156.7	-16.2	1915.7	1968.1	52.4	3189.1	3209.1	19.9
557.6	561.8	4.2	2028.5	2030.4	2.0	1515.5	1481.2	-34.3	177.8	152.6	-25.2	1482.6	1465.0	-17.6
394.1	394.3	0.2	163.9	143.4	-20.5	388.6	410.4	21.9	3141.0	3178.0	37.0	702.0	724.8	22.8
378.6	384.1	5.5	158.9	140.9	-18.0	334.3	357.5	23.2	503.5	525.4	21.9	2908.6	2984.2	75.5
1739.6	1755.2	15.6	1889.2	1851.0	-38.2	1884.8	1831.1	-53.8	173.9	154.2	-19.6	156.6	143.0	-13.6
Iteration 6			**Iteration 7**			**Iteration 8**			**Iteration 9**			**Iteration 10**		
Actual	Predicted	Error	Actual	Predicted	Error	Actual	Predicted	Error	Actual	Predicted	Error	Actual	Predicted	Error
1815.1	1811.3	-3.8	464.0	466.2	2.2	1508.9	1468.8	-40.2	1476.6	1484.0	7.4	1891.6	1880.2	-11.4
1471.8	1441.6	-30.1	1508.4	1454.8	-53.5	1487.7	1465.7	-22.0	1872.5	1850.7	-21.8	1796.6	1795.3	-1.3
3272.1	3261.1	-11.0	2008.4	2032.7	24.3	340.9	361.7	20.8	2240.8	2315.7	74.9	395.6	400.2	4.6
424.4	414.2	-10.2	389.2	378.2	-11.0	1651.4	1670.3	18.9	1567.1	1534.2	-32.9	1707.0	1702.9	-4.1
1659.6	1668.6	9.1	155.2	142.6	-12.6	2424.7	2468.4	43.7	393.0	391.9	-1.1	1892.0	1818.2	-73.9
1918.2	1932.3	14.2	434.0	440.7	6.7	384.1	380.9	-3.2	1521.5	1531.7	10.2	1781.0	1803.8	22.9
416.6	393.4	-23.3	1929.9	1949.8	20.0	1892.7	1798.2	-94.4	304.5	333.4	28.9	1687.5	1703.1	15.7
2521.3	2520.4	-0.8	2126.4	2164.2	37.8	2247.5	2269.1	21.5	2570.9	2627.9	57.0	1918.6	1983.7	65.1
1871.1	1802.4	-68.7	154.9	140.3	-14.6	412.6	418.3	5.7	1603.9	1553.9	-50.0	3062.8	3099.6	36.8
3204.7	3241.3	36.7	434.3	419.0	-15.3	311.6	342.6	31.0	1987.8	2039.1	51.3	431.2	426.6	-4.6
1971.8	1986.9	15.1	415.8	420.7	4.8	411.0	421.4	10.4	2949.2	2931.0	-18.2	3325.3	3272.1	-53.3

3. Conclusion and Discussion

In this work, the temporal variation of the most important air pollutant, O_3, was examined, and the relationship of O_3 components with other air pollutants was investigated, to model these pollutants using various artificial intelligence methods. The findings obtained in the study are summarized below.
A strong correlation of 0.956 is observed between NO_x and CO_2, which is highest correlation among the input variables. Also, the highest correlation with the target variable (O_3) is observed with NO_x (0.984). A high and significant correlation is also observed with CO_2 (0.954), N_2O (0.844), and CO (0.886) with the target variable (O_3). From Figure 6, it can be observed that SO_x has a relatively weak correlation (0.41) with O3. However, this level of correlation is still adequate for individual estimation. The order of correlations of the data set with O_3 was NO_x > CO2 > CO > N2O > SO_x.

When reviewing the literature on the subject, it becomes apparent that the application of machine learning algorithms in air pollution studies largely centers around the temporal estimation of air pollutant gas concentrations. In a study by Gao et al. [50], the R^2 value was found to be 0.80 in ozone estimation. Jia et al. [51], tried to predict ozone with artificial neural networks using different model structures in their work. The R^2 values obtained in the study vary between 0.89 and 0.92. They found the correlation coefficients in values ranging from 0.40 to 0.60. Liu et al. [13], tried to predict long-term ozone concentrations using ML algorithms in their work. The R^2 values of the ML model results used in the study ranged from 0.60 to 0.87. Considering these values, it is seen that the statistical results obtained in this study are compatible with the literature (Table 5).

Table 5. Summary and comparison of studies based on predicting ozone concentration

Reference	Study Area	Best Model	R^2
[36]	Delhi	XGBoost	0.614
[37]	Malaysia, Kuala Lumpur, Selangor	BDT	0.87, 0.88, 0.91
[38]	Saudi Arabia	SVR	0.924
[39]	California	RF	0.84
[40]	Malaya	LR, SVR, GPR, ANN	0.89
[41]	Amman	MLP	0.98
[50]	Hebei	ANN	0.80
[51]	Lanzhou	CANN	0.89 - 0.92
[13]	Beijing-Tianjin-Hebei, Yangtze River Delta, Sichuan Basin, Pearl River Delta, Jianghan Plain, Northeast Plain	XGBoost	0.60 - 0.87
Our study	Czechia, Germany, Greece, Spain, France, Italy, Romania, Switzerland, United Kingdom, Turkey	Bagging-MLP	0.9994

A review of studies in the literature reveals that ozone concentrations have been predicted using different regions, various methods, or distinct features. Consequently, while the findings from these studies may vary, there is a general consensus that artificial intelligence methods have been successful in predicting ozone concentration.

Upon examining the experimental results in this study (Table 3), it becomes evident that the Bagging-MLP method was the most successful in estimating O_3 levels. A comparison between the O_3 values predicted by the proposed Bagging-MLP method and the actual values (Table 4) demonstrates a close alignment between the two. These findings indicate that estimation systems employing the Bagging-MLP method can predict O_3 levels with minimal error.

Estimation performances of the normalized dataset and the original dataset were compared, the error of the SVR, FKNN, RFR, and Bagging-SVR methods for the normalized dataset decreased. However, this decrease is less than 1%. In the Bagging-MLP prediction model, there is no difference in the success of the estimations for O_3 concentration in the original dataset and the normalized dataset, however, the results obtained show that when compared to other regression methods it is the most successful method.

The use of a limited sample size in this study is acknowledged as a limitation. To address this constraint, we employed 10 cross-validation techniques to ensure the reliability of results and mitigate issues such as overfitting. Cross-validation is a validation method that involves dividing the dataset into smaller subsets and training and testing the model on these subsets. This helps improve the model's generalization and reduces misleading results arising from the restricted sample size.
There are many studies in the literature on the distribution of air pollutants in the atmosphere and their relations with each other. For this reason, modeling studies are important in the follow-up of the long-term relationships of air pollutants with each other. The results obtained in this study show that the relationship of ozone with other air pollutants can be successfully predicted by artificial intelligence methods.

Conflict of Interest Statement

The authors declare that there is no conflict of interest.

References

[1] M. O. Andreae and P. J. Crutzen, "Atmospheric aerosols: Biogeochemical sources and role in atmospheric chemistry," *Science,* vol. 276, no. 5315, pp. 1052-1058, 1997.

[2] H. K. Ozcan, E. Bilgili, U. Sahin, O. N. Ucan, and C. Bayat, "Modeling of trophospheric ozone concentrations using genetically trained multi-level cellular neural networks," *Advances in Atmospheric Sciences,* vol. 24, no. 5, pp. 907-914, 2007. doi:10.1007/s00376-007-0907-y

[3] O. P. Tripathi *et al.*, "An assessment of the surface ozone trend in Ireland relevant to air pollution and environmental protection," *Atmospheric Pollution Research,* vol. 3, no. 3, pp. 341-351, 2012. doi:10.5094/APR.2012.038

[4] Z. Feng, E. Hu, X. Wang, L. Jiang, and X. Liu, "Ground-level O3 pollution and its impacts on food crops in China: a review," *Environmental Pollution,* vol. 199, pp. 42-48, 2015. doi:10.1016/j.envpol.2015.01.016

[5] R. Tang, X. Huang, D. Zhou, H. Wang, J. Xu, and A. Ding, "Global air quality change during the COVID-19 pandemic: Regionally different ozone pollution responses COVID-19: **疫情期**间全球空气质量变化: **臭氧响**应的区域间差异," *Atmospheric and Oceanic Science Letters,* vol. 14, no. 4, p. 100015, 2021. doi:10.1016/j.aosl.2020.100015

[6] Y. Ma, B. Ma, H. Jiao, Y. Zhang, J. Xin, and Z. Yu, "An analysis of the effects of weather and air pollution on tropospheric ozone using a generalized additive model in Western China: Lanzhou, Gansu," *Atmospheric Environment,* vol. 224, p. 117342, 2020. doi:10.1016/j.atmosenv.2020.117342

[7] X. Ren, Z. Mi, and P. G. Georgopoulos, "Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States," *Environment International,* vol. 142, p. 105827, 2020. doi:10.1016/j.envint.2020.105827

[8] X. Yang *et al.*, "Summertime ozone pollution in Sichuan Basin, China: Meteorological conditions, sources and process analysis," *Atmospheric Environment,* vol. 226, p. 117392, 2020. doi:10.1016/j.atmosenv.2020.117392

[9] A. Yerramilli *et al.*, "Simulation of surface ozone pollution in the Central Gulf Coast region during summer synoptic condition using WRF/Chem air quality model," *Atmospheric Pollution Research,* vol. 3, no. 1, pp. 55-71, 2012. doi:10.5094/APR.2012.005

[10] T. Zhang *et al.*, "Modeling the joint impacts of ozone and aerosols on crop yields in China: An air pollution policy scenario analysis," *Atmospheric Environment,* vol. 247, p. 118216, 2021. doi:10.1016/j.atmosenv.2021.118216

[11] M. Zunckel *et al.*, "Modelled surface ozone over southern Africa during the cross border air pollution impact assessment project," *Environmental Modelling & Software,* vol. 21, no. 7, pp. 911-924, 2006. doi:10.1016/j.envsoft.2005.04.004

[12] Y. Fu and A. Tai, "Impact of climate and land cover changes on tropospheric ozone air quality and public health in East Asia between 1980 and 2010," *Atmospheric Chemistry and Physics,* vol. 15, no. 17, pp. 10093-10106, 2015. doi:10.5194/acp-15-10093-2015

[13] R. Liu, Z. Ma, Y. Liu, Y. Shao, W. Zhao, and J. Bi, "Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach," *Environment international,* vol. 142, p. 105823, 2020. doi:10.1016/j.envint.2020.105823

[14] S. Amini and S. Mohaghegh, "Application of machine learning and artificial intelligence in proxy modeling for fluid flow in porous media," *Fluids,* vol. 4, no. 3, p. 126, 2019. doi:10.1016/j.envint.2020.105823

[15] M. Meroni, F. Waldner, L. Seguini, H. Kerdiles, and F. Rembold, "Yield forecasting with machine learning and small data: what gains for grains?," *Agricultural and Forest Meteorology,* vol. 308, 2021. doi:10.1016/j.agrformet.2021.108555

[16] B. Dietrich, J. Walther, M. Weigold, and E. Abele, "Machine learning based very short term load forecasting of machine tools," *Applied Energy,* vol. 276, p. 115440, 2020. doi:10.1016/j.apenergy.2020.115440

[17] M. Cheng, F. Fang, T. Kinouchi, I. Navon, and C. Pain, "Long lead-time daily and monthly streamflow forecasting using machine learning methods," *Journal of Hydrology,* vol. 590, p. 125376, 2020. doi:10.1016/j.jhydrol.2020.125376

[18] V. Chandran *et al.*, "Wind power forecasting based on time series model using deep machine learning algorithms," *Materials Today: Proceedings,* vol. 47, 2021. doi:10.1016/j.matpr.2021.03.728

[19] S. George and A. Dixit, "A machine learning approach for prioritizing groundwater testing for per-and polyfluoroalkyl substances (PFAS)," *Journal of Environmental Management,* vol. 295, p. 113359, 2021. doi:10.1016/j.jenvman.2021.113359

[20] L. Nevasalmi, "Forecasting multinomial stock returns using machine learning methods," *The Journal of Finance and Data Science,* vol. 6, pp. 86-106, 2020. doi:10.1016/j.jfds.2020.09.001

[21] P. Cihan, "Forecasting fully vaccinated people against COVID-19 and examining future vaccination rate for herd immunity in the US, Asia, Europe, Africa, South America, and the World," *Applied Soft Computing,* vol. 111, p. 107708, 2021. doi:10.1016/j.asoc.2021.107708

[22] P. Cihan, "Fuzzy rule-based system for predicting daily case in covid-19 outbreak," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2020: IEEE, pp. 1-4. doi:10.1109/ISMSIT50672.2020.9254714

[23] P. Cihan, H. Ozel, and H. K. Ozcan, "Modeling of atmospheric particulate matters via artificial intelligence methods," *Environmental Monitoring and Assessment,* vol. 193, pp. 1-15, 2021. doi:10.1007/s10661-021-09091-1

[24] P. Cihan, O. Kalıpsız, and E. Gökçe, "Hayvan Hastalığı Teşhisinde Normalizasyon Tekniklerinin Yapay Sinir Ağı ve Özellik Seçim Performansına Etkisi," *Electronic Turkish Studies,* vol. 12, no. 11, 2017. doi:10.7827/TurkishStudies.11902

[25] P. Cihan and Z. B. Ozger, "A new heuristic approach for treating missing value: ABCIMP," *Elektronika ir Elektrotechnika,* vol. 25, no. 6, pp. 48-54, 2019. doi:10.5755/j01.eie.25.6.24826

[26] P. Cihan and Z. B. Ozger, "A new approach for determining SARS-CoV-2 epitopes using machine learning-based in silico methods," *Computational Biology and Chemistry,* vol. 98, p. 107688, 2022. doi:10.1016/j.compbiolchem.2022.107688

[27] P. Cihan, E. Gokce, and O. Kalipsiz, "A review of machine learning applications in veterinary field," *Kafkas Universitesi Veteriner Fakultesi Dergisi,* vol. 23, no. 4, 2017. doi:10.9775/kvfd.2016.17281

[28] P. Cihan, "The machine learning approach for predicting the number of intensive care, intubated patients and death: The COVID-19 pandemic in Turkey," *Sigma Journal of Engineering and Natural Sciences,* vol. 40, no. 1, pp. 85-94, 2022. doi:10.14744/sigma.2022.00007

[29] T. F. Cova and A. A. Pais, "Deep learning for deep chemistry: optimizing the prediction of chemical patterns," *Frontiers in chemistry,* vol. 7, p. 809, 2019. doi:10.3389/fchem.2019.00809

[30] M. Ani, G. Oluyemi, A. Petrovski, and S. Rezaei-Gomari, "Reservoir uncertainty analysis: The trends from probability to algorithms and machine learning," in *SPE Intelligent Energy International Conference and Exhibition*, 2016: OnePetro. doi:10.2118/181049-MS

[31] P. Cıhan and H. Coşkun, "Performance comparison of machine learning models for diabetes prediction," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*, 2021: IEEE, pp. 1-4. doi:10.1109/SIU53274.2021.9477824

[32] P. Cihan, O. Kalipsiz, and E. Gökçe, "Yenidoğan kuzularda bilgisayar destekli tanı," *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi,* vol. 26, no. 2, pp. 385-391, 2020. doi:10.5505/pajes.2019.51447

[33] Z. B. Ozger and P. Cihan, "A novel ensemble fuzzy classification model in SARS-CoV-2 B-cell epitope identification for development of protein-based vaccine," *Applied soft computing,* vol. 116, p. 108280, 2022. doi:10.1016/j.asoc.2021.108280

[34] P. Cihan, "Impact of the COVID-19 lockdowns on electricity and natural gas consumption in the different industrial zones and forecasting consumption amounts: Turkey case study," *International Journal of Electrical Power & Energy Systems,* vol. 134, p. 107369, 2022. doi:10.1016/j.ijepes.2021.107369

[35] E. E. Ozbas, D. Aksu, A. Ongen, M. A. Aydin, and H. K. Ozcan, "Hydrogen production via biomass gasification, and modeling by supervised machine learning algorithms," *International Journal of Hydrogen Energy,* vol. 44, no. 32, pp. 17260-17268, 2019. doi:10.1016/j.ijhydene.2019.02.108

[36] E. K. Juarez and M. R. Petersen, "A comparison of machine learning methods to forecast tropospheric ozone levels in Delhi," *Atmosphere,* vol. 13, no. 1, p. 46, 2021. doi:10.3390/atmos13010046

[37] E. Jumin *et al.*, "Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction," *Engineering Applications of Computational Fluid Mechanics,* vol. 14, no. 1, pp. 713-725, 2020. doi:10.1080/19942060.2020.1758792

[38] Q. Pan, F. Harrou, and Y. Sun, "A comparison of machine learning methods for ozone pollution prediction," *Journal of Big Data,* vol. 10, no. 1, p. 63, 2023. doi:10.1186/s40537-023-00748-x

[39] W. Wang, X. Liu, J. Bi, and Y. Liu, "A machine learning model to estimate ground-level ozone concentrations in California using TROPOMI data and high-resolution meteorology," *Environment International,* vol. 158, p. 106917, 2022. doi:10.1016/j.envint.2021.106917

[40] A. Yafouz *et al.*, "Comprehensive comparison of various machine learning algorithms for short-term ozone concentration prediction," *Alexandria Engineering Journal,* vol. 61, no. 6, pp. 4607-4622, 2022. doi:10.1016/j.aej.2021.10.021

[41] M. Aljanabi, M. Shkoukani, and M. Hijjawi, "Ground-level ozone prediction using machine learning techniques: A case study in Amman, Jordan," *International Journal of Automation and Computing,* vol. 17, pp. 667-677, 2020. doi:10.1007/s11633-020-1233-4

[42] D. T. Bui, Q. P. Nguyen, N.-D. Hoang, and H. Klempe, "A novel fuzzy K-nearest neighbor inference model with differential evolution for spatial prediction of rainfall-induced shallow landslides in a tropical hilly area using GIS," *Landslides,* vol. 14, no. 1, pp. 1-17, 2017. doi:10.1007/s10346-016-0708-4

[43] M. O. Elish, "A comparative study of fault density prediction in aspect-oriented systems using MLP, RBF, KNN, RT, DENFIS and SVR models," *Artificial Intelligence Review,* vol. 42, no. 4, pp. 695-703, 2014. doi:10.1007/s10462-012-9348-9

[44] P. Kumar, M. Folk, M. Markus, and J. C. Alameda, *Hydroinformatics: data integrative approaches in computation, analysis, and modeling*. CRC Press, 2005.

[45] J. Han and M. Kamber, "Data mining concepts and techniques, Morgan Kaufmann Publishers," *San Francisco, CA,* pp. 335-391, 2001.

[46] L. Breima, "Random Forests. Machine Learning," 2010.

[47] L. Breiman, "Bagging predictors," *Machine learning,* vol. 24, no. 2, pp. 123-140, 1996.

[48] J. D'Haen, D. Van den Poel, and D. Thorleuchter, "Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique," *Expert systems with applications,* vol. 40, no. 6, pp. 2007-2012, 2013.

[49] J. C. W. W. W. Carole and N. Beale, *Global climate change linkages: acid rain, air quality, and stratospheric ozone*. Springer Science & Business Media, 1989.

[50] S. Gao *et al.*, "Simulation of surface ozone over Hebei province, China using Kolmogorov-Zurbenko and artificial neural network (KZ-ANN) combined model," *Atmospheric Environment,* vol. 261, p. 118599, 2021. doi:10.1016/j.atmosenv.2021.118599

[51] B. Jia, R. Dong, and J. Du, "Ozone concentrations prediction in Lanzhou, China, using chaotic artificial neural network," *Chemometrics and Intelligent Laboratory Systems,* vol. 204, p. 104098, 2020. doi:10.1016/j.chemolab.2020.104098