Pan *et al. Journal of Big Data* (2023) 10:63
https://doi.org/10.1186/s40537-023-00748-x

Journal of Big Data

RESEARCH Open Access

A comparison of machine learning methods for ozone pollution prediction

Qilong Pan[1†], Fouzi Harrou[1*†] and Ying Sun[1]

†Qilong Pan and Fouzi Harrou contributed equally to this work.

*Correspondence: fouzi.harrou@kaust.edu.sa

[1] Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), 23955-6900, Thuwal, Saudi Arabia

Abstract

Precise and efficient ozone (O_3) concentration prediction is crucial for weather monitoring and environmental policymaking due to the harmful effects of high O_3 pollution levels on human health and ecosystems. However, the complexity of O_3 formation mechanisms in the troposphere presents a significant challenge in modeling O_3 accurately and quickly, especially in the absence of a process model. Data-driven machine-learning techniques have demonstrated promising performance in modeling air pollution, mainly when a process model is unavailable. This study evaluates the predictive performance of nineteen machine learning models for ozone pollution prediction. Specifically, we assess how incorporating features using Random Forest affects O_3 concentration prediction and investigate using time-lagged measurements to improve prediction accuracy. Air pollution and meteorological data collected at King Abdullah University of Science and Technology are used. Results show that dynamic models using time-lagged data outperform static and reduced machine learning models. Incorporating time-lagged data improves the accuracy of machine learning models by 300% and 200%, respectively, compared to static and reduced models, under RMSE metrics. And importantly, the best dynamic model with time-lagged information only requires 0.01 s, indicating its practical use. The Diebold-Mariano Test, a statistical test used to compare the forecasting accuracy of models, is also conducted.

Keywords: Ozone pollution, Machine learning, Multivariate time-series data

Introduction

Background and motivation

Atmospheric pollution is becoming a global problem with a harmful influence on human health and ecoecosystems [1, 2]. Ground-level ozone pollution could cause substantial damage to crops, forests, and native plants. There were plenty of negative cases concerning the globally decreasing yields of crops caused by the gradually increasing concentration of ground-level ozone pollution, such as 2.2-—5.5% for maize to 3.9–15% and 8.5–14% for wheat and soybean, respectively [3]. It is considered one of the important greenhouse gases that would exacerbate global warming [2]. For instance, the heatwave that occurred in France during the summer of 2003 was associated with an atypical ozone pollution that had an impact on the whole European population [4]. In addition, ozone can damage the tissues of the respiratory

tract, causing inflammation and irritation and resulting in symptoms such as coughing, chest tightness, and worsening asthma symptoms [5]. Therefore, it would be meaningful to accurately predict the concentration of the ground level of the zone, which is beneficial for people's daily activities.

Accurately predicting Ground-level ozone levels is a key step for the safety of humans and ecosystems. Towards this end, different methods have been proposed in the literature to predict ozone concentrations in the past two decades [6, 7]. Two main approaches are used to predict ozone concentrations: physical-based and data-driven models [8–10]. Various physical-driven methods have been presented in the literature, including the Comprehensive Air Quality Model with Extensions model (CAMx) [11], Weather Research and Forecasting with Chemistry model (WRF-Chem) [12], and Weather Research and Forecasting coupled with Community Multi-scale Air Quality (WRF-CMAQ) [13]. However, due to the heavily intensive computation and the difficulty in accurately collecting the environmental and pollutant data, the use of physical methods is challenging to meet the requirements in the real world [8].

Precise air quality forecasting offers valuable insights that can assist individuals in taking necessary precautions to avert unfavorable outcomes. However, developing accurate models of ozone concentration is challenging due to multiple factors, including the complex mechanisms responsible for ozone formation in the troposphere [14], the intricacies of meteorological conditions in urban areas, and the uncertainty in the measurements of all the involved parameters. When fundamental process models are unavailable, data-driven techniques, such as machine learning, can reveal the linear relationships among the process variables. Data-driven approaches are not limited to describing the exact relationship between pollutants and environmental variables as physical models do [6]. In contrast to physical models, data-driven methods exploit available measurements to learn the underlying reference empirical model. With large available and easily collected data from sensors, the data-driven methods attempt to learn this relationship from the data with different algorithms, which enclose statistical models and machine learning-based methods. Statistical models have the advantage of being highly interpretable, and a measure of uncertainty in the results, and these advantages have made statistical models shine in environmental analysis [15, 16]. Several statistical techniques aimed at predicting and monitoring ozone pollution levels have been presented in the literature, including the autoregressive (AR) model and autoregressive integrated moving average (ARIMA) and its variants model [17, 18], multivariate linear regression [19, 20], the least absolute shrinkage, and selection operator (Lasso) [21], and principal component regression (PCR) [22, 23]. Unlike model-based methods, machine learning methods are within a nonparametric framework and do not require prior information about data distributions. In recent years, machine learning methods have become more appealing due to their flexibility and capability to explicitly learn relevant features from multivariate data. Thus. machine learning have been applied in wide range of application, in engineering applications, such as predicting the axial compressive strength of concrete-filled steel tubular (CFST) columns [24, 25], as well as in biomedical applications [26] and air quality monitoring [27, 28]. Also, numerous machine learning methods have been used to model and predict ozone prediction, such as artificial neural networks [29], Support Vector Regression (SVR) [30], Random Forests (RF) [31], and (XGBoost) eXtreme Gradient

Boosting [32]. Previous studies have demonstrated that machine-learning techniques can be employed for ozone pollution modeling and prediction.

Contributions

With the advantages of computing efficiency and simplified modeling, machine learning algorithms have achieved significant progress in the prediction, classification, outlier detection, and generation tasks. As presented above, several machine learning algorithms have been implemented in the prediction of the concentration of ozone. In this paper, we attempt to compare the performance of different machine learning algorithms, such as their accuracy and consumed time. Here we studied nineteen machine learning-based models for ozone pollution prediction, aiming to find the most efficient and accurate model. By involving more input variables, we anticipate the model will incorporate more information and be more accurate. By involving fewer water parameters, we anticipate the model is more efficient and requires less computational power. Somehow there should be a trade-off between accuracy and efficiency. The main contributions of this work are recapitulated as follows.

- This study's first contribution is to explore machine learning models' capabilities in predicting Ozone pollution levels at KAUST based on environmental factors as input, such as absolute humidity, air temperature, ambient pressure, global radiation, wind direction, and wind speed. Here, the performance of nineteen models was investigated to predict Ozone concentration, including linear models, support vector regression (SVR), Gaussian Processes Regression (GPR), Multi-layer perceptron, and ensemble methods. Specifically, in the linear model, we use linear regression, linear regression with l_2 norm regularization, Lasso and partial least square regression; in the GPR and SVR, we choose different kernels such as exponential function, radius basis function, Matern function; in the MLP, the combination of different depth and width is adopted; and in ensemble methods, we used bagging, boosting and random forest models. Five performance evaluation metrics are employed to assess the goodness of predictions. However, the results showed that using only weather conditions to predict ozone levels, the machine learning models did not provide satisfactory predictions. This suggests that other factors, such as atmospheric pollutants, also significantly affect ozone pollution.
- The second contribution of this study focused on improving the predictive capability of the machine learning models in predicting ozone pollution levels by investigating the use of weather and pollution data (such as NO_2, SO_2, and PM) as input. Additionally, feature selection techniques based on the random forest algorithm were used to identify the most important variables significantly influencing the models' predictive capability. Results revealed that using only a subset of the features, specifically NO_2, absolute humidity, air temperature, absolute pressure, wind direction, and SO_2, led to parsimonious prediction models with slightly improved accuracy. This highlights the importance of feature selection in reducing overfitting, improving model accuracy, and speeding up computation in machine learning model building.
- However, it is worth noting that the investigated machine learning-driven approaches do not consider information from past data in predicting ozone concentration lev-

els. This study's final contribution involves investigating the effect of incorporating lagged ozone data with other input variables and selecting important features to predict ozone pollution levels. We proposed dynamic machine learning models incorporating information from past data to predict ozone concentrations, providing useful insights into trends and patterns that can help predict future ozone pollution levels. The results demonstrated that incorporating lagged data considerably improved the accuracy of the machine learning models in predicting ozone concentrations. Overall, this study highlights the potential of dynamic machine learning models that incorporate past data and important features in predicting ozone pollution levels and underscores the importance of considering multiple factors when developing such models.

The rest of the paper consists of five sections. "Related works" section gives the related works in the prediction task of ozone concentration. "Materials and methods" section briefly describes the machine learning methods used in this study and the evaluation metric. "Results and discussion" section presents the exploratory data analysis. "Ozone prediction results" section is the results and discussion of the machine learning algorithms within our datasets. Lastly, "Conclusion" section summarizes the paper and provides future directions for possible improvements.

Related works

Accurate prediction of ozone concentrations is needed to enhance ozone control, plan for implementing preventive measures, and manage public health. Thus, ground-level ozone prediction in different conditions was the subject of several research studies in the last two decades. Machine learning-based techniques have recently risen in popularity in numerous applications, including ozone pollution modeling, prediction, and monitoring, owing to their flexibility and feature extraction capacity without the need to understand the underlying mechanisms for constructing empirical models. Machine learning methods, such as RF, SVR, Decision Tree (DT), and XGBoost, have been used to predict ozone concentration levels [33–35]. For example, Jiang et al. adopted the RF with a large amount of feature engineering in the task of ozone prediction [36]. However, the naive RF algorithm is complicated to interpret and takes much time to build trees, with the time complexity being $O(v \times n \log(n))$ where n is the number of samples, and v is the number of attributes [36]. In [37], Allu et al. applied multiple linear regression (MLR) models to predict surface ozone levels based on air pollutants and meteorological parameters in Hyderabad, India, in 2016. Results showed that the adjusted R^2 for the MLR models ranged from 0.6 to 0.9 for precursor gases and 0.9 for meteorological variables. However, MLR assumes a linear relationship between the dependent variable and independent variables, and as a result, it may not be able to capture nonlinear relationships between variables. Chelani et al. conducted an empirical study on predicting ozone levels using the standard SVM based solely on environmental variables. The results show that the SVM outperformed the MLP and linear regression in terms of three evaluation metrics [38]. The study in [39] focused on predicting ground-level ozone concentration in the air near Zrenjanin, Serbia, using MLR and artificial neural networks (ANNs). Results revealed that the ANNs model outperformed MLRA with

coefficient of determination of 0.919 during training and 0.873 during testing, compared to MLRA's 0.663 and 0.672 for training and testing, respectively. In [40], Hosh et al. compared two models, deterministic (WRF-Chem) and ANN, for predicting ground-level ozone concentration in São Paulo, Brazil. WRF-Chem simulations satisfactorily predicted CO concentrations and correlated with O3 and NOx measurements at air quality monitoring stations. The FS-ANN model achieved correlation coefficients of 0.84 and 0.75 for the daily mean and 0.64 and 0.67 for the daily peak ozone during testing. While WRF-Chem performed better in predicting mean and peak ozone concentrations, FS-ANN was advantageous due to its lower computational costs and ease of development and implementation.

Braik et al. considered three models, recurrent multilayer perceptron (RMLP), recurrent fuzzy neural network (RFNN), and hybridization of RFNN and grey wolf optimizer (GWO), to forecast daily ozone, particulate matter (PM10 and PM2.5) concentrations in a highly polluted city in the Republic of China [41]. They showed that the hybrid RFNN-GWO model achieved the best results in the modeling of ozone, PM10, and PM2.5 compared with the RMLP-ANN and RFNN models. In [42], Ren et al. compared thirteen machine learning algorithms with linear land-use regression (LUR) for modeling ozone concentrations across the contiguous United States. The nonlinear machine learning methods achieved higher prediction accuracy than LUR, with the improvements being more significant for spatiotemporal modeling (nearly 10%-40% decrease of predicted RMSE). By tuning the sample weights, spatiotemporal models can predict concentrations used to calculate ozone design values that are comparable or even better than spatial models (nearly 30% decrease of cross-validated RMSE). Random Forest and Extreme Gradient Boosting were found to be the two best-performing machine learning algorithms. In [43], Oufdou et al. explored the advantages and disadvantages of parametric and non-parametric statistical models, such as Sparse Partial Least Squares (SPLS), Lasso, RF, bagging, and Classification and Regression Tree (CART), in forecasting daily ozone concentration. Results indicate that the parametric models have more accurate predictions than that non-parametric approaches [43]. Juarez et al. applied several machine learning methods, including XGBoost, Random Forest, K-Nearest Neighbor Regression, Support Vector Regression, Decision Trees, AdaBoost, and linear regression, for ozone prediction using data comprising twelve air pollutants and five weather variables over one year in Delhi. They showed the importance of training machine learning methods with season-specific data sets [44]. To build on the interpretability of decision trees on ozone prediction, Jumin et al. further considered the Boosted DT and compared its performance with a deep neural network (DNN) model and the linear regression model [33]. However, the problem of large-time complexity was not well addressed, even though the accuracy is improved when predicting ground-level ozone concentration. In [34], Yilmaz et al. applied the XGBoost and used a reweighted ensembled model to combine SVR, RF, and DNN to improve the model's stability and performance [34]. Moreover, Marvin et al. considered comparing machine learning methods and conducted experiments to verify the performance of LR, LASSO, ARIMA, RF, XGboost, and Natural Gradient Boosting (NGBoost) in predicting ozone concentration levels [45]. Results showed that least-squares boosting and NGBoost dominate the other tested forecasting models.

Most of pthe revious studies on machine learning methods for ozone pollution prediction have neglected to utilize information from past data and have included both relevant and irrelevant features in their models. As a result, this study aims to investigate the impact of incorporating lagged data and feature selection on the performance of different machine learning methods in predicting ozone pollution. By considering the information from past data and selecting only important features, this study aims to develop more parsimonious models with improved prediction accuracy.

Materials and methods

This section presents the models investigated in this study for ozone concentration prediction. Machine learning methods compared in this paper are summarized into five categories: linear models, GPR, SVR, MLP, and ensemble methods. This section will cover each model's setting and its variants. In total, nineteen methods are investigated in this study.

Linear models

Linear regression is the classic statistical and machine learning model and has developed various variants nowadays. The variants we will implement in our experiment are Least absolute shrinkage and selection operator (LASSO) [46] and Partial Least Squares Regression (PLSR) [47, 48]. The Lasso is introduced for the feature selection, and PLSR is for multicollinearity between different features. Letting $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ be the covariate matrix and $Y = (y_1, y_2, \ldots, y_n)$ be the outcome of $\mathbf{X}$, then the expression of lasso can be written as:

$$\min_{\beta_0,\beta} \left\{ \left\| y - \beta_0 - X\beta \right\|_2^2 \right\} s.t.\ \|\beta\|_1 \leq t, \tag{1}$$

where β_0 is the constant coefficient, $\beta := (\beta_1, \beta_2, \ldots, \beta_p)$ is the coefficient vector, and $t > 0$ is a prespecified hyperparameter that determines the degree of regularization, $\|u\|_p$ is the standard ℓ^p norm. The complexity of linear regression is only concerned with the dimension of features; mathematically, $O(p^3)$ and p is the dimension of features. The limitation of linear regression models is the difficulty of modeling the nonlinearity in the data, where it is necessary for manual manipulation of selecting degrees of variables or transformation functions.

GPR models

GPR models are nonparametric kernel-driven learning models; this approach has shown extended modeling ability for handling nonlinear prediction problems because of its nonlinear approximation capabilities [49, 50]. Importantly, a Gaussian process is fully specified by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$. The Gaussian distribution is over vectors, whereas the Gaussian process is over functions [51], which is written as,

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \tag{2}$$

where $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n) \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I})$, and $f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{X}), k(\mathbf{X}, \mathbf{X}))$, for abbreviation, $\mu := m(\mathbf{X})$ and $\Sigma := k(\mathbf{X}, \mathbf{X})$). Without loss of generality, the $m(\mathbf{X})$ is set 0 or constant,

and $k(\mathbf{X}, \mathbf{X})$ could be any kernel that fits the data, such as Matern kernel or (non-)stationary kernels (see more kernels in [52]). In addition, the complexity of GPR increases cubically with the number of samples, that is $O(n^3)$, which is an important limitation of the GRP. In the big data case, the computation is heavily intensive, whereas parallel computing cannot help when computing the inverse of a large matrix.

SVR models

This subsection briefly presents another widely used approach, the SVR model, a flexible kernel-based method with good learning ability through kernel tricks. The SVR model is widely used for nonlinear regression problems. It maps the training data into a higher-dimensional space and accomplishes linear regression, enabling efficient handling of nonlinear data through the kernel trick [53, 54]. Moreover, structural risk minimization is the relevant concept used in designing the SVR model. It is demonstrated that SVR performs satisfactorily with limited samples [55].

The key concept used in designing the SVR model is structural risk minimization. The SVR model seeks the combination of different features to approximate the target value by support vectors, which is formulated in Eq. (3) [55], where ϵ means the tolerated error, that is, the soft margin; ζ_i, ζ_i^* are the introduced slack variables to transform the inequality optimization into equality optimization. The solution to Eq.(3) and its kernel selection methods can be found at [55].

$$\begin{aligned} \min_{\mathbf{w},\xi,\xi^*} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi^*) \\ \text{subject to} \quad & \begin{cases} \mathbf{y}_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b \le \epsilon + \xi_i \\ \mathbf{w}^T\phi(\mathbf{x}_i) + b - y_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1, \ldots, n \end{cases} \end{aligned} \tag{3}$$

There are several types of kernel functions, including linear, polynomial, radial basis function (RBF), and sigmoid. The linear kernel simply computes the dot product between two data points, while the polynomial kernel calculates the dot product raised to a power. The RBF kernel, on the other hand, measures the distance between two data points using a Gaussian distribution. The choice of kernel function can significantly impact the performance of the machine-learning model. For example, the linear kernel is often used for linearly separable data, while the RBF kernel is more effective for nonlinearly separable data. However, selecting the appropriate kernel function is not always straightforward and often requires trial and error. Overall, the SVR provides satisfactory performance with limited samples [55]. However, the SVR has some limitations, including difficulty in selecting the kernel to model nonlinearity and the complexity of $O(n^2)$.

MLP models

A multilayer perceptron (MLP) is a type of feedforward neural network consisting of an input layer, multiple hidden layers, and an output layer. The design objective of an MLP is to enhance prediction accuracy for nonlinear functions by stacking multiple layers of linear transformations and activation functions in the hidden layers. MLPs are effective models for approximating nonlinear functions and have been successfully applied in

various applications, such as image recognition and natural language processing. The ability to learn complex relationships between input and output data and the flexibility to adjust the number of hidden layers and nodes makes MLPs a popular choice for solving complex problems. In [56], an MLP with two hidden layers exhibited good capability in nonlinear predictions. Importantlly, MLP is seeking approximate the target value in nonlinear function by stacking the multiple linear transformations and activation functions layer by layer. The MLP's predictive Equation can be expressed as shown in Eq. (4).

$$y_i = f_m(g_m(f_{m-1}(g_{m-1})...f_1(g_1(\mathbf{x}_i)))),\ i = 1, 2, 3, \ldots, n, \tag{4}$$

where f and g represent activation and linear transformation functions, respectively. The target value y_i is estimated by the MLP for each input vector $\mathbf{x}_i$, with the weights and biases learned during the training process. Stochastic gradient descent and Adam are common optimization algorithms used in training MLPs [57].

Ensemble models

Ensemble methods train multiple models to solve the problem, where there are well-trained weak and strong learners. How to combine and select the strong and weak learners divides the ensemble methods into three parts: bagging, stacking, and boosting [58]. The boosting builds a strong learner from several base learners, where training samples are allocated different weights based on their prediction, truly predicted samples with small weight and falsely predicted samples with large weight, until the boosting algorithm converges. The bagging combines bootstrapping and aggregation, where several base learners are trained before aggregating multiple base learners. The stacking method contains a cardinal learner and meta learner, where meta learners receive the data preprocessed by the cardinal learner with the training dataset before giving the output [59]. While having the advantages the easy deployment and fast computation, the over-fitting and under-fitting are still haunting many researchers.

Evaluation metrics

The metrics used in our evaluation are as follows: n means the number of samples in the training dataset, ϵ is a small value set as 10^{-8} in case of the denominator as 0. The small value means good performance in terms of the RMSE, MAPE, MAE and J^2, while the large value represents a good performance for R^2. In addition, for RMSE and MAE, the value solely indicates the absolute difference, while MAPE measures the relative difference. R^2 gives an illustration of predicted variance compared to the average, but cannot guarantee good performance on the prediction task. J^2 has a measure of the performance of models but cannot give a concrete value. Thus, it is reasonable to consider several metrics together when evaluating the performance of algorithms.

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}, \tag{5}$$

$$\mathrm{MAPE}(y,\hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}\frac{\|y_i-\hat{y}_i\|_1}{\max(\epsilon,\|y_i\|_1)}, \tag{6}$$

$$\mathrm{MAE}(y,\hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}\|y_i-\hat{y}_i\|_1, \tag{7}$$

$$R^2 = 1-\frac{\sum_{i=0}^{n-1}(\hat{y}_i-\bar{y}_i)^2}{\sum_{i=0}^{n-1}(y_i-\bar{y}_i)^2}, \tag{8}$$

$$J^2 = \frac{RMSE^2_{test}}{RMSE^2_{train}}. \tag{9}$$

Algorithm framework

This study compares the prediction accuracy and efficiency of nineteen machine learning models for ozone concentration prediction. The whole process in our experiment is shown in Fig. 1. At first, the data was preprocessed by eliminating outliers and imputing missing values. Specifically, due to the sensors malfunction, it failed to obtain the ozone concentration from May 2020 to September 2022; we thus chose to remove this part of the data from the training even though we have the data for other variables. After carrying out an exploratory data analysis, we designed three experiment cases: ozone prediction using 1) only environmental factors, 2) environmental and pollutant factors with the selection method, and 3) the second case with time-lagged ozone data. For each case study, we construct machine-learning models using training data and evaluate the accuracy of the trained models for ozone prediction. Five evaluations scores are considered for the comparison. In each experiment, we implement the machine learning algorithms, such as linear model, GRP, SVR, MLP, ensemble methods and different metrics.

Results and discussion

This section presents the investigated data and some exploratory data analysis. Moreover, the performance of the studied prediction methods is compared and results are discussed.

Data description and analysis

The ambient air pollution datasets used in this study were gathered at KAUST (located in Thuwal, Saudi Arabia) by the KAUST weather team [60]. These data comprise ambient pollution data (PM_{10}, $PM_{2.5}$, CO, NO_2, O_3, SO_2) and weather data (i.e., absolute humidity, air temperature, ambient pressure, global radiation, precipitation, and wind speed). Table 1 lists the collected variables and their units of measurement. The monitoring periods start from May 20, 2020, to Dec 20, 2020, and from Jan 21, 2021, to Oct 21, 2021. The data are collected every fifteen minutes. Figure 2 exhibits the monitor equipment used to colled pollution data at KAUST.

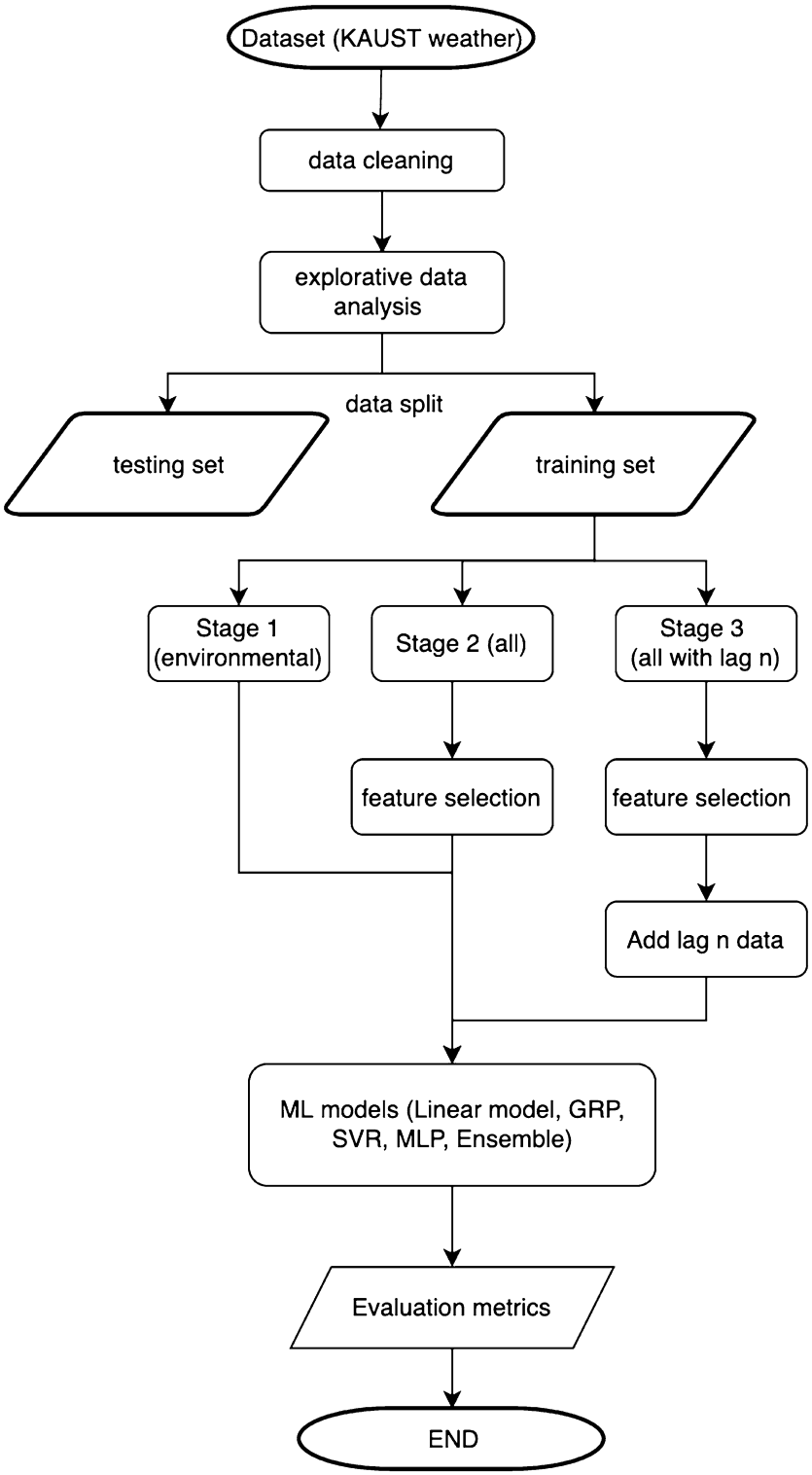

Fig. 1 Data prepossessing and modeling procedure

Fig. 2 One of the equipment for monitoring the local weather index at KAUST, Thuwal, Saudi Arabia [60]

Table 1 Measured Variables and units of measurement

Variable	Unit
PM_{10}	$\mu g/m^3$
$PM_{2.5}$	$\mu g/m^3$
CO	ppm
NO_2	ppb
O_3	ppb
SO_2	ppb
absolute humidity	g/m^3
air temperature	$°C$
ambient pressure	hPa
global radiation	W/m^2
precipitation	$l/m^2/h$
relative humidity	%
wind speed	m/s
Wind Direction Correction	Deg

Data analysis

The investigated datasets comprise some measurements with outliers. Outliers in pollution and weather data could be caused by malfunctioning sensors. Thus, the first essential step before building machine learning models is data cleaning. Essentially, outliers are removed and missing data points are imputed for enhancing data quality. Moreover, eliminating outliers helps increase the prediction accuracy of the considered models [61]. In this study, we replaced outliers with the median of the training dataset.

Another important and often unavoidable challenge impacting the data quality when working with real-world data is missing data [62]. Missing values frequently happen in air pollutant measurement. Different factors can cause missing data, such

as sensor malfunction, incorrect data recording, power outages, and faults in data acquisition [62, 63]. The presence of missing values can affect study understandings and findings and influence the operation of air quality-related public services. To alleviate this problem, various procedures were employed in the literature for missing value imputation [64]. We can distinguish two types of missing values in air pollution data: long-consecutive periods or short gaps. Generally, short periods of missing values can be caused by routine maintenance and temporary power outages; while long gaps of missing values could be due to sensor malfunctions or other critical failures in the data acquisition process [65].

Here, missing values exist in each dimension of the dataset, accounting for approximately 1%-10% of each variable. In the short term, there are no significant polynomial and linear trends in the dataset, so the mean values are used to fill in the missing values. In the long term, there is a huge missing portion of the ozone timeline due to equipment failure, so the latter half of the dataset is chosen for the experiment in this study.

Figure 3 depicts the distribution of the collected data, which indicates that these datasets are non-Gaussian distributed.

The temporal evolution of ozone concentrations is generally controlled by seasonal and diurnal factors, such as weather conditions, and industrial activities. Figure 4 presents the variation of ozone concentration per hour for each day. We observe the presence of a daily cycle in the concentration of ozone [66]. Specifically, we observe the formation of high concentrations of ozone in the heat of the afternoon and early evening and the destruction during the night.

Essentially, the gradual formation of ozone results from chemical reactions of nitrogen oxides (NOx) and volatile organic compounds (VOC) under the action of solar radiation [67]. On the other hand, the destruction of ozone at night is mainly due to

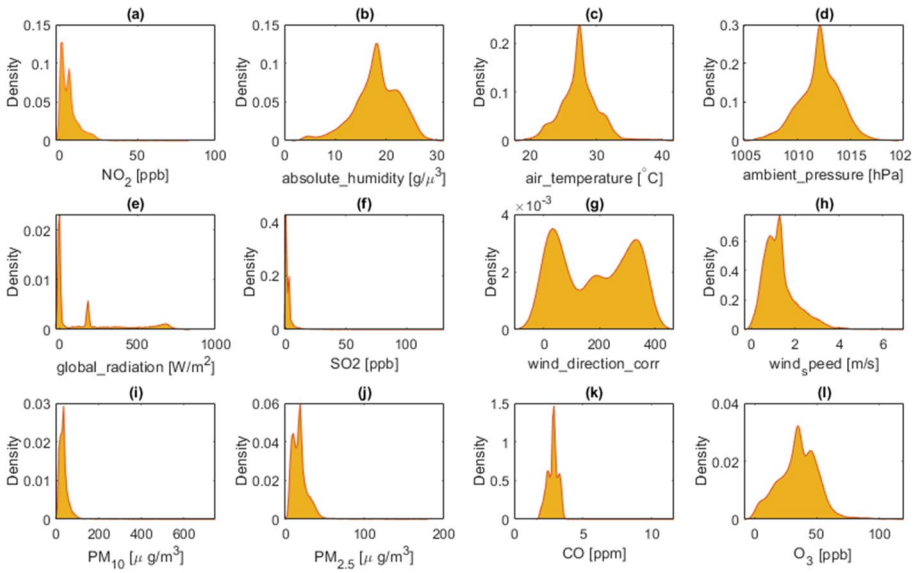

Fig. 3 Distribution of the KAUST pollution dataset

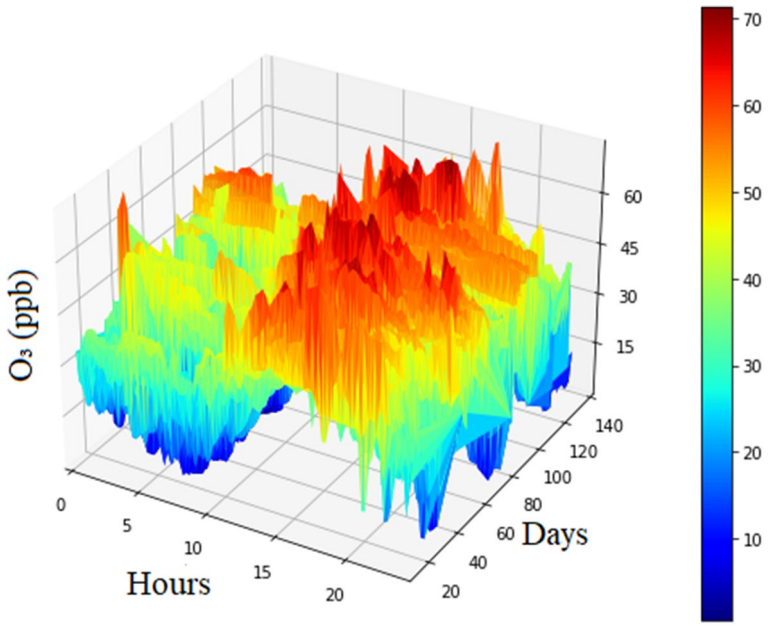

Fig. 4 Data analysis for ozone in 12 months

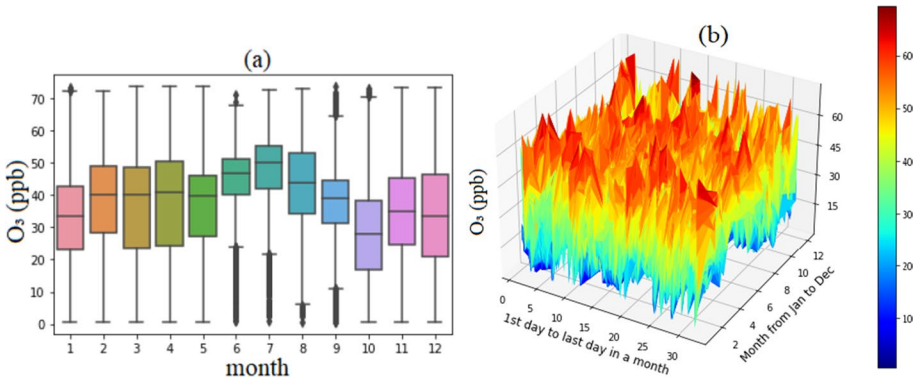

Fig. 5 **a** Boxplot for ozone in 12 months; **b** 3D line for ozone in 12 months

its reaction with nitric oxide, which is generally emitted by vehicles, to produce nitrogen dioxide.

Now, to visually analyze the monthly ozone trend Fig. 5a-b the monthly distributions of ozone concentrations by boxplot and line plot, respectively. From Fig. 5a, we observe that important peaks mainly appear in June, July, August, and September. Indeed, the weather in Saudi Arabia during summer is characterized by high temperatures and extreme weather conditions, which enable the formation of photochemical ozone pollution. Broadly speaking, the photochemical formation of ozone achieves peak levels in warmer weeks and hours (Fig. 5).

The correlation matrix between the studied pollution and weather variables is visualized via heatmap in Fig. 6, where the linear correlation is stronger as the color grows darker green or purple. We observe a high correlation between PM_{10} and $PM_{2.5}$ with a

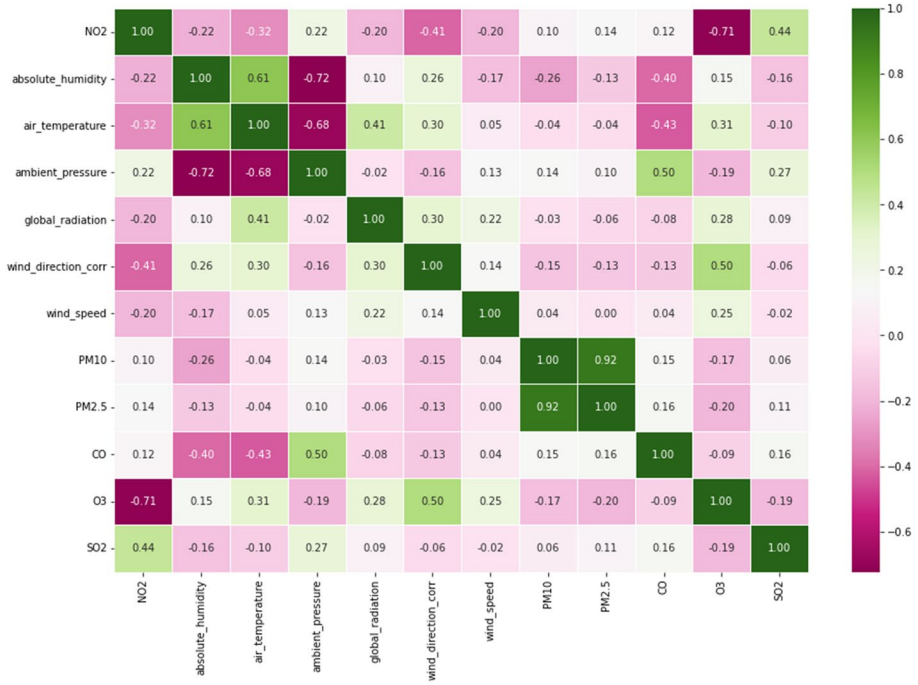

Fig. 6 Heatmap of correlation matrix between considered variables

coefficient of 0.92. This could reveal the presence of the same source for the PM_{10} and $PM_{2.5}$, like road traffic and combustion of diesel fuel. This information can be exploited during model construction by dropping one of PM_{10} and $PM_{2.5}$ in the experiment configuration to get a parsimonious model. Additionally, the relationship between ozone and NO_2 conforms to the formation of the ozone caused by nitrogen ozone reaction considering the correlation (−0.71) [66, 68]. This is expected as O_3 formation relies on NO_2. Besides, there are six pairs with a mild correlation around the absolute coefficient of 0.5 - 0.7. We also observe a weak negative correlation between $PM_{2.5}$ and PM_{10} and O_3. This is because $PM_{2.5}$ and PM_{10} are involved in slowing down the aerosol sink of hydro-peroxy (HO2) radicals, which is one of the precursors of ozone [69].

Before we start model construction, we analyze the time-dependent in the considered time-series data by computing the empirical autocorrelation function (ACF). Importantly, the ACF is employed to measure the self-similarity of the time-series data over distinct delay times [70]. Figure 7a-b depicts the ACF of the meteorological and pollution time-series data, which is divided into the strong periodicity Fig. 7a and the mild periodicity Fig. 7b. We observe a clear seasonality of 24 h from the ACF plot of ozone and NO_2 and solar irradiance (Fig. 7a). As discussed before, the ozone seasonality is due to the diurnal formation cycle of ozone caused by the diurnal temperature cycle. Then, the difference in periodicity is supposed to be taken into consideration in the modeling.

Ozone prediction results

In this study, three experiments are conducted to design parsimonious and efficient machine-learning models for ozone concentration prediction.

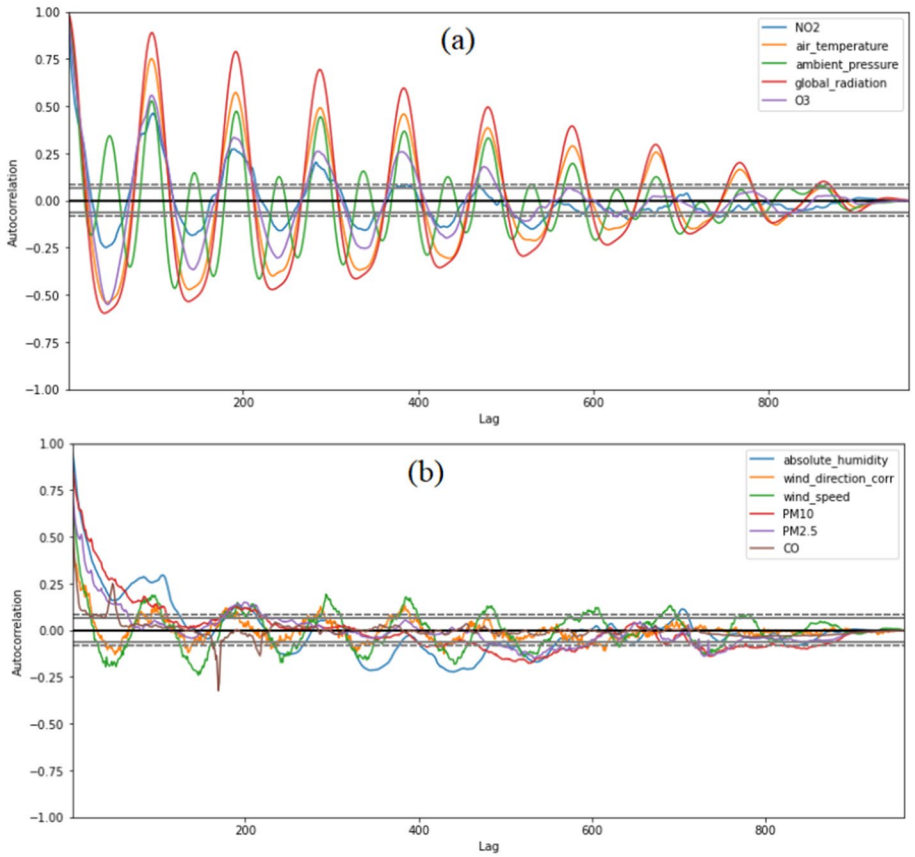

Fig. 7 Autocorrelation of the used time-series data: **a** strong periodicity in the dataset, **b** mild periodicity in the dataset

- In the first experiment, we consider only weather variables (i.e., air temperature, ambient pressure, global irradiance, absolute humidity, wind direction, and wind speed) to predict ozone pollution via several machine learning models.
- In the first experiment, to further improve the prediction performance, in addition to the meteorological variables, we incorporate also pollution data (CO, $PM_{2.5}$, PM_{10}, NO_2). Here, we considered variables selection to choose only important variables for ozone prediction.
- In the final experiment, we include lagged data to further improve the prediction accuracy of the considered machine learning models.

Experimental configurations

The models in the experiment are categorized into five parts, linear models, GPR, SVR, MLP, and ensemble methods, where each part has its own variants and parameter tuning setting (scikit-learn and LightGBM in python), shown in the Table 2. To this end, the best performance for each model can be observed in our experiment of the real dataset.

Table 2 The description of the machine learning methods adopted in this paper, (**X**, **Y**) means the dataset in pair; β the parameters of the models, λ the hyperparameter

Model	Formulation	Kernel
Linear Regression	$\mathbf{Y} = \mathbf{X}\beta$	None
Linear Regression with *l*2 regularizer	$\mathbf{Y} = \mathbf{X}\beta + \lambda\|\beta\|^2$	None
Lasso	$\mathbf{Y} = \mathbf{X}\beta + \lambda\|\beta\|^1$	None
Partial Linear Least Square Regression	Regression of decomposed **Y** and **X**	None
GRP (Exponential kernel)	$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \Sigma)$	Eq. 10
GRP (DotProd kernel)	$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \Sigma)$	Eq. 13
GRP (Matérn kernel)	$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \Sigma)$	Eq. 12
SVR (linear kernel)	Eq. 3	Eq. 14
SVR (Polynomial kernel)	Eq. 3	Eq. 15
SVR (Radial basis kernel)	Eq. 3	Eq. 11
SVR (Sigmoid kenerl)	Eq. 3	Eq. 16
MLP_1	Layer shape [10, 5, 1]	None
MLP_2	Layer shape [10, 5, 2, 1]	None
RF	Depth 7, Criterion: squared error	None
Bagging	10 decision trees as base learner	None
GBoost	Criterion: Friedman MSE	None
AdaBoost	50 decision trees as base learner	None
HistGBoost	Criterion: squared error	None
LightGBM	31 leaves and Criterion: squared error	None

- Linear models: no penalty, l_1 norm (Lasso) penalty, and l_2 norm penalty are considered in the linear regression where the p in the Eq. (1) is set 0, 1 and 2 respectively and t is set as 1, as well as partial linear least square regression which is used for multicollinearity issues [47].
- GPR: the kernels used are radial basis function kernel (Eq. (11)), Matern kernel (Eq. (12)) and Dotproduct kernel (Eq. (11)), where $l = 1$, $d(,)$ means Euclidean distance in the Eq. (11); $\nu = 1.5, l = 1$, and $K_\nu()$ is a modified Bessel function, $\Gamma()$ is the gamma function in the Eq. (12); and $\sigma_0 = 1$ in the Eq. (11).

$$k(x_a, x_b) = \sigma^2 \exp\left(-\frac{\|x_a - x_b\|^2}{2\ell^2}\right) \tag{10}$$

$$k(x_i, x_j) = \exp\left(-\frac{d(x_i, x_j)^2}{2l^2}\right), \tag{11}$$

$$k(x_i, x_j) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l}d(x_i, x_j)\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{l}d(x_i, x_j)\right) \tag{12}$$

$$k(x_i, x_j) = \sigma_0^2 + x_i \cdot x_j, \tag{13}$$

- SVR: the kernels used are linear kernel (Eq. (14)), polynomial kernel (Eq. (15)), radial basis function kernel (Eq. (11)), and sigmoid kernel (Eq. (16)), where the $\gamma = 1, d = 3$ and $r = 0$ in the Eq. (15) and Eq. (16).

$$k(x_i, x_j) = \langle x_i, x_j \rangle \tag{14}$$

$$k(x_i, x_j) = (\gamma \langle x_i, x_j \rangle + r)^d \tag{15}$$

$$k(x_i, x_j) = \tanh(\gamma \langle x_i, x_j \rangle + r) \tag{16}$$

- MLP: the layer's shape for MLP in our setting is [10, 5, 1] and [10, 5, 2, 1], respectively. In this way, we can explore the influence of depth on the model's performance.
- Ensemble: RF with the maximum depth of 7, bagging with 10 decision trees as base learners, GBoost with criterion as Friedman MSE, AdaBoost with 50 decision trees as base learners, HistGBoost with loss of squared error, LightGBM with 31 leaves and objective loss as a squared error.

In our dataset, the data from the last week is used as a test set (672 samples) and the rest as a train set (39328 samples). For this study, we tested the models using one week of data with a 15-minute time resolution, which was deemed sufficient. We utilized a large training dataset to capture the variability and dynamics in the data. Due to there are two main features, environmental variables and pollutant variables, we conduct the experiments in three stages, where the first experiment only contains the environmental variables, such as wind direction, absolute humidity, ambient pressure, air temperature, and wind speed; the second stage adds the rest of pollutant variables, such as NO_2, PM2.5, CO, and SO_2 into the experiment and uses feature selection methods to select some features with high importance due to the large dimensions; lastly, the third stage considered lag-n information of ozone based on the selected features in the second stage.

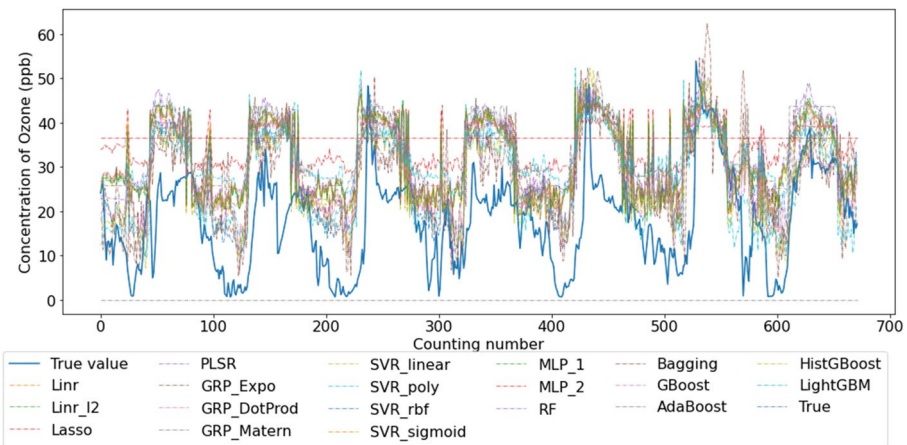

Fig. 8 Measured and predicted ozone concentrations using the investigated machine learning models

Case study 1: Ozone concentrations prediction using meteorological factors

In the first experiment, we solely consider the meteorological factors to predict the concentrations of ozone, specifically, absolute humidity, air temperature, ambient pressure, global radiation, wind direction, and wind speed. Importantly, this experiment aims to evaluate the capacity of machine learning models to predict ozone concentrations based on meteorological factors. Comparison between the observed values and predicted ozone levels from the considered twenty machine learning models are shown in Fig. 8. Here, we omitted the results from the GPR_{Matern} and MLP_2 because they delivered unsatisfactory results in the testing phase. On the other hand, Fig. 8 indicates that most of the considered models can well capture the trend of the ozone levels during the testing seven days.

Now, to further evaluate the performance for the other 17 modes, which well capture the trend, we examined the distribution of the prediction errors from each model. The prediction error is the deviation separating the observed from the predicted ozone values. Figure 9 displays the boxplot for prediction error of the investigated models, which indicates the poor training of GPR_{Matern} but failed to show the over-fitting of MLP_2. Broadly speaking, the more compact the boxplot of the prediction error is, the more precise the prediction is. Meanwhile, according to Figs. 9 and 8, there exists a bias in all of the models, because the mean of the error cannot be around zero. Visually, we can see that the distributions of the prediction errors of GPR_{Expo}, SVR_{poly} and SVR_{rbf}, and HistGBoost are more close to zero compared to the other models (Fig. 9).

The prediction performance of the trained model is listed in Table 3 in terms of different metrics, namely RMSE, MAE, MAPE, R^2, J2 and execution time. Based on the performance measures in Table 3, it can be seen that the GPR_{expo} model provides the best score on RMSE, MAE and MAPE, AdaBoost has the best performance on R^2 which is verified from Fig. 9, and SVR_{poly} is the best regarding the J^2. The results indicate that the best model on the absolute/relative metrics such as RMSE and MAPE is not necessarily the most efficient training method or the most stable method with the least variance.

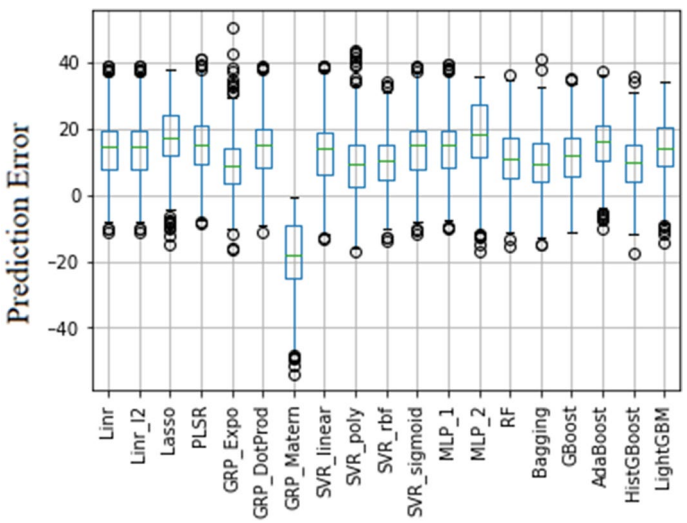

Fig. 9 Boxplot of prediction errors of each model based on test data

Table 3 Stage 1, the comparison of machine learning methods on ozone prediction, where the value before slash represents the loss value in the training dataset and the latter in the testing dataset (training/testing), the best performance is bold and the line colored grey means over-fitting

Model	rMSE ↓	MAE ↓	MAPE ↓	R-square ↑	J2 ↓	Time (min)
Linr	12.677/16.196	10.09/14.137	0.945/2.713	0.332/0.305	1.632	0.01
Linr_l2	12.677/16.199	10.09/14.141	0.945/2.714	0.332/0.305	1.633	0
Lasso	13.642/19.561	10.91/17.674	1.114/3.589	0.227/0.267	2.056	0
PLSR	12.777/17.336	10.243/15.319	0.934/2.901	0.322/0.331	1.841	0.01
GRP_Expo	**9.117/12.612**	**6.847/10.237**	**0.543/1.941**	0.655/0.387	1.914	9.64
GRP_DotProd	12.706/16.754	10.121/14.705	0.963/2.847	0.329/0.306	1.739	121.44
GRP_Matern	0.0/20.868	0.0/17.887	0.0/1.0	1.0/0.0	inf	25.47
SVR_linear	12.795/15.982	9.99/13.821	0.997/2.69	0.323/0.218	1.56	0.01
SVR_poly	11.858/13.271	9.358/10.827	0.873/2.002	0.416/0.177	**1.253**	5.45
SVR_rbf	10.66/13.062	8.202/10.981	0.685/2.007	0.528/0.428	1.501	8.57
SVR_sigmoid	12.764/16.48	10.124/14.4	0.973/2.778	0.323/0.278	1.667	9.77
MLP_1	12.681/16.561	10.111/14.523	0.939/2.755	0.332/0.316	1.706	0.6
MLP_2	15.514/21.591	12.682/19.384	1.283/4.199	0.0/0.0	1.937	1.35
RF	10.116/13.874	7.801/11.75	0.657/2.188	0.575/0.42	1.881	0.06
Bagging	3.722/13.254	2.497/10.781	0.174/2.121	0.942/0.363	12.681	0.18
GBoost	10.272/14.359	7.948/12.269	0.66/2.306	0.562/0.398	1.954	0.28
AdaBoost	11.704/17.821	9.465/16.113	0.852/3.047	**0.431/0.449**	2.318	0.1
HistGBoost	9.036/13.009	6.899/10.831	0.538/2.147	0.661/0.431	2.073	13.07
LightGBM	11.532/16.499	9.204/14.699	0.892/3.051	0.447/0.422	2.047	0.01

In addition, the ↑ means the model is better when the value is larger, and the ↓ means the model is better when the value is smaller

Moreover, the time consumed by GPR_{Expo} (9.64 mins) is nearly × 40 times than the second/third best models such as RF and Bagging (0.06 and 0.18 mins) with only merely 5% loss of predicting performance with regard to RMSE and MAE. Therefore, GRP_{Expo} is revealed as the best prediction model in this case study. Overall, using machine learning for predicting ozone concentrations based only on meteorological variables does not provide promising prediction accuracy. Therefore, we should take the different factors into consideration when using machine learning methods for ozone prediction.

Case study 2: Ozone concentrations prediction using meteorological factors and other pollutants with features selection

The main objective of this experiment is to construct parsimonious machine-learning models to predict ozone concentrations. To this end, we select important variables from all meteorological and pollutant variables. There are various techniques for identifying feature importance in the literature, including Random Forest feature selection, Principal Component Analysis (PCA) [71], and mutual information [72]. In the case of this study, the random forest feature selection technique was explicitly adopted because it is well-suited for nonlinear and high-dimensional data [73]. Additionally, Random Forest feature selection can provide information about the importance of individual features, which is useful for understanding the underlying relationships between the features and the target variable. Specifically, we adopt the random forest method [73] to select a subgroup of features to reduce the heavy computation. Indeed, non-informative and redundant input variables will be ignored in building a predictive model to reduce the number

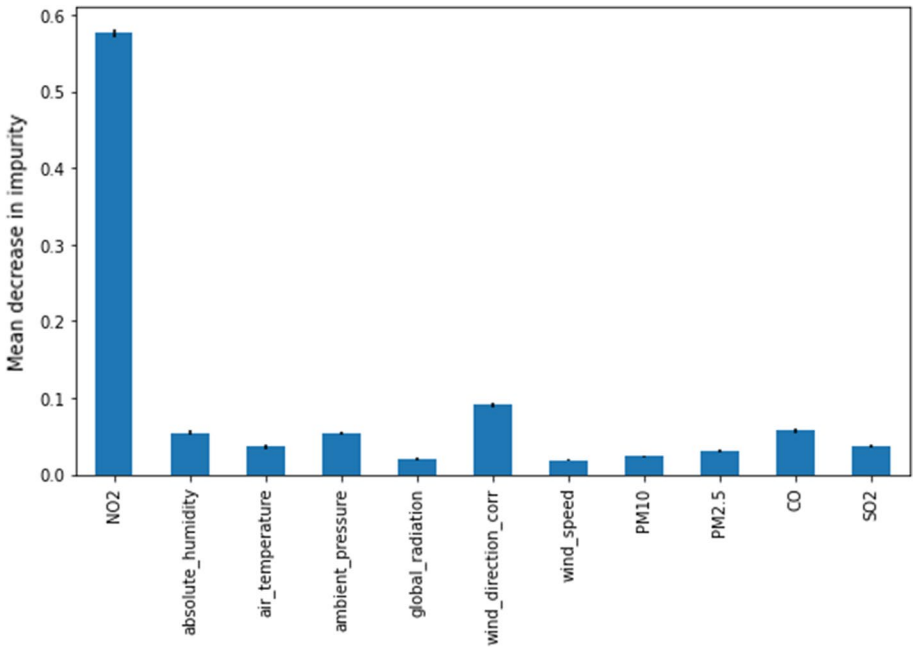

Fig. 10 Feature selection based on the random forest

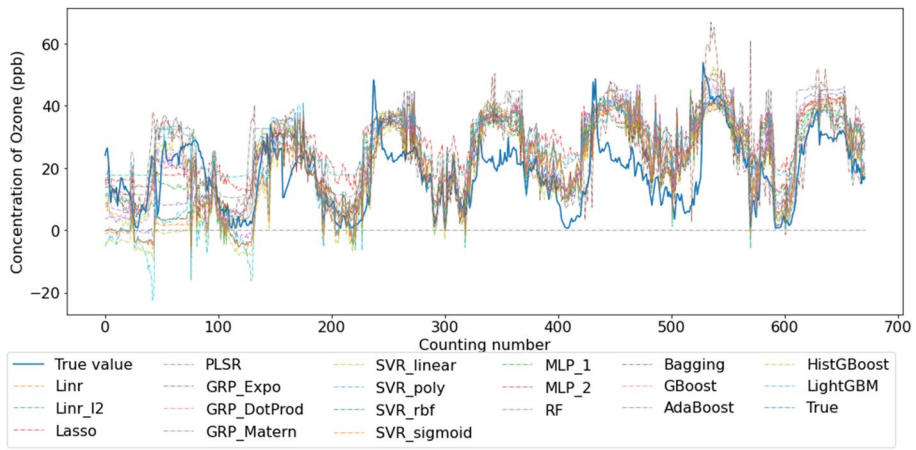

Fig. 11 Comparison between observed and predicted O_3 concentrations from the investigated models when using the selected features from meteorological and pollutant variables

of input variables. In this case, inputs are meteorological data (i.e., air temperature, ambient pressure, global irradiance, absolute humidity, wind direction, and wind speed) and (CO, PM2.5, PM10, SO_2, NO_2). The larger amplitude of input for feature importance means the greater the influence of that variable is on the ozone prediction.

The results of the selection are shown in the Fig. 10. It can be seen that six features are used, NO_2, absolute humidity, air temperature, absolute pressure, wind direction and SO_2, which are the most important features for ozone pollution prediction. We can notice that the selected variables contain some meteorological variables (i.e., absolute humidity, air temperature, absolute pressure, and wind direction) and

pollutants variables (i.e., NO_2 and SO_2). Figure 10 indicates that the NO_2 plays a dominant role in ozone prediction because it is the primary component in ozone formation. We observe also that CO has some interaction with ozone and can be used as input for predicting ozone concentrations due to the chemical equilibrium of the reaction $CO + O_3 \leftrightarrows CO_2 + O_2$ [13]. For meteorology variables, we can see that wind direction has a role in impacting ozone concentrations [74].

Here, we consider only the most important variables as inputs into the machine-learning methods to reduce the computation and get parsimonious models. Figure 11 shows the observed and predicted ozone concentrations from the investigated machine-learning models. Similar conclusions hold as in the previous case study, most models can well capture the trend of the ozone in the testing data except GPR_{Matern} which fails to be well trained. Importantly, based on Fig. 11 and Fig. 12, the bias between predicted and observed ozone values are alleviated compared to that of the previous case study. Specifically, we observe that the mean value prediction error observed is much more close to 0 in stage 2 than in the previous case (Fig. 12). Importantly, we conclude that including both meteorological and pollutant features enhances the prediction performance of the machine learning models (Fig. 12).

Table 4 compares the prediction results obtained by the nineteen machine learning models in terms of different metrics. The best score in terms of RMSE, MAE, and J^2 is attributed to the SVR_{rbf}, HistBoost has the best performance on MAPE, and LightGBM is the best regarding the R^2 (Table 4). We observe that the GPR_{Matern} achieved the highest R^2, but it comes with very high values of J^2, indicating overfitting. The results indicate that the best model on the absolute/relative metrics, such as RMSE and MAE still has the best training efficiency in the J^2. Moreover, the time consumed by SVR_{rbf} (9.24 mins) is nearly $\times$ 80 times than the second best models such as RF (0.06 min) with only merely 8% loss of predicting performance with regard to RMSE and MAE. Therefore, if the accuracy does not rank first in the priority of predicting the concentration of ozone in the target task, it would be better to use the bagging or forest methods to

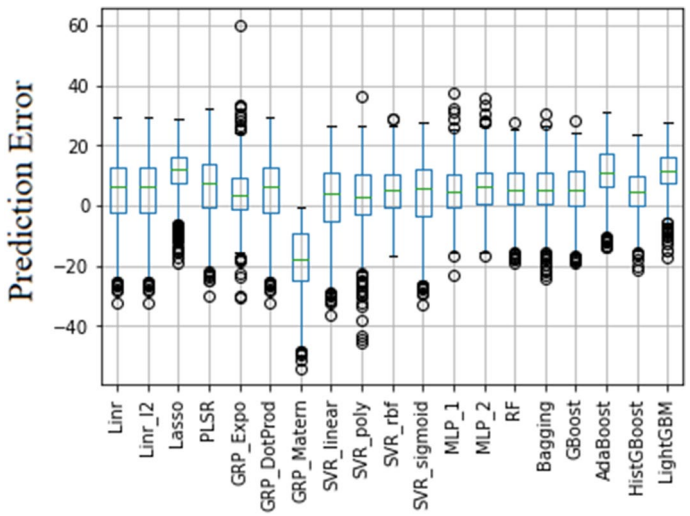

Fig. 12 Boxplot of prediction errors of each model based on test data

Table 4 Stage 2, the comparison of machine learning methods on ozone prediction, where the value before slash represents the loss value in the training dataset and the latter in the testing dataset (training/testing), the best performance is bold and the line colored grey means over-fitting

Model	rMSE ↓	MAE ↓	MAPE ↓	R-square ↑	J2 ↓	Time (min)
Linr	9.847/11.958	7.634/10.116	0.484/1.052	0.597/-0.111	1.475	0.01
Linr_l2	9.847/11.956	7.635/10.114	0.484/1.051	0.597/-0.11	1.474	0
Lasso	11.035/13.558	8.867/12.345	0.77/2.111	0.494/0.379	1.51	0.01
PLSR	9.956/12.148	7.761/10.339	0.496/1.067	0.588/0.013	1.489	0.01
GRP_Expo	7.558/9.968	5.696/7.383	0.31/1.051	0.763/0.302	1.739	9.04
GRP_DotProd	9.848/11.94	7.635/10.09	0.482/1.049	0.597/-0.111	1.47	132.69
GRP_Matern	0.0/20.868	0.0/17.887	0.0/1.0	1.0/0.0	inf	25.77
SVR_linear	9.901/12.106	7.582/9.98	0.464/1.075	0.593/-0.247	1.495	0.01
SVR_poly	9.937/11.656	7.663/9.096	0.56/0.949	0.59/-0.141	1.376	6.82
SVR_rbf	**8.54/9.19**	**6.437/7.302**	0.38/1.047	0.697/0.499	**1.158**	9.24
SVR_sigmoid	9.947/12.026	7.69/10.106	0.484/1.052	0.589/-0.163	1.462	10.24
MLP_1	8.821/9.505	6.703/7.642	0.41/1.056	0.677/0.416	1.161	1.33
MLP_2	8.809/9.985	6.704/8.022	0.397/1.23	0.678/0.457	1.285	1.42
RF	8.059/9.921	6.153/7.874	0.357/0.953	0.73/0.436	1.515	0.06
Bagging	2.919/10.164	1.94/7.99	0.097/0.918	0.965/0.377	12.124	0.1
GBoost	8.191/9.856	6.218/7.926	0.366/0.863	0.721/0.394	1.448	0.26
AdaBoost	9.413/13.832	7.46/12.113	0.561/1.758	0.638/0.445	2.159	0.11
HistGBoost	7.214/9.509	5.458/7.53	**0.305/0.84**	0.784/0.419	1.737	8.42
LightGBM	9.712/12.974	7.721/11.671	0.69/2.167	**0.608/0.586**	1.785	0.01

In addition, the R^2 of the orange lines is less than 0, representing that the performance of the corresponding model is worse than that of directly using the mean value. In addition, the ↑ means the model is better when the value is larger, and the ↓ means the model is better when the value is smaller

let the model as much more efficient with a mere loss of accuracy. Noticeably, in this experiment, the most important factor to influence the performance of *SVR* is the kernel, where the performance of SVR is even worse than the average-value model if we set the kernel as linear, polynomial or sigmoid, which are all mark as orange color in Table 4. In addition, the best model in the previous experiment, GPR_{Expo}, is outperformed by the SVR_{rbf} in every metric.

Overall, this experiment showed that by involving important meteorological and pollutants variables the machine learning models are expected to incorporate more information, which results in improved prediction performance. However, there is still a large gap for improvement as the prediction accuracy is not within a satisfactory range in terms of the considered evaluation metrics. This could be due to the construction of machine learning models by ignoring information from past data. The next experiments are dedicated to a comparative assessment taking into account lagged ozone data in building machine learning models.

Case study 3: Ozone prediction with lagged data

In the two previous experiments, the machine learning models have been developed without considering information from past data, making them difficult to capture dynamics in data. Due to there still being a mild bias in the previous experiment, we attempt to make the algorithm as unbiased as possible by considering the lag n ozone

Table 5 Time-lagged ozone data

Variable	Description
O_3.Lag1	15-minutes lagged ozone value
O_3.Lag2	30-minutes lagged ozone value
O_3.Lag3	45-minutes lagged ozone value

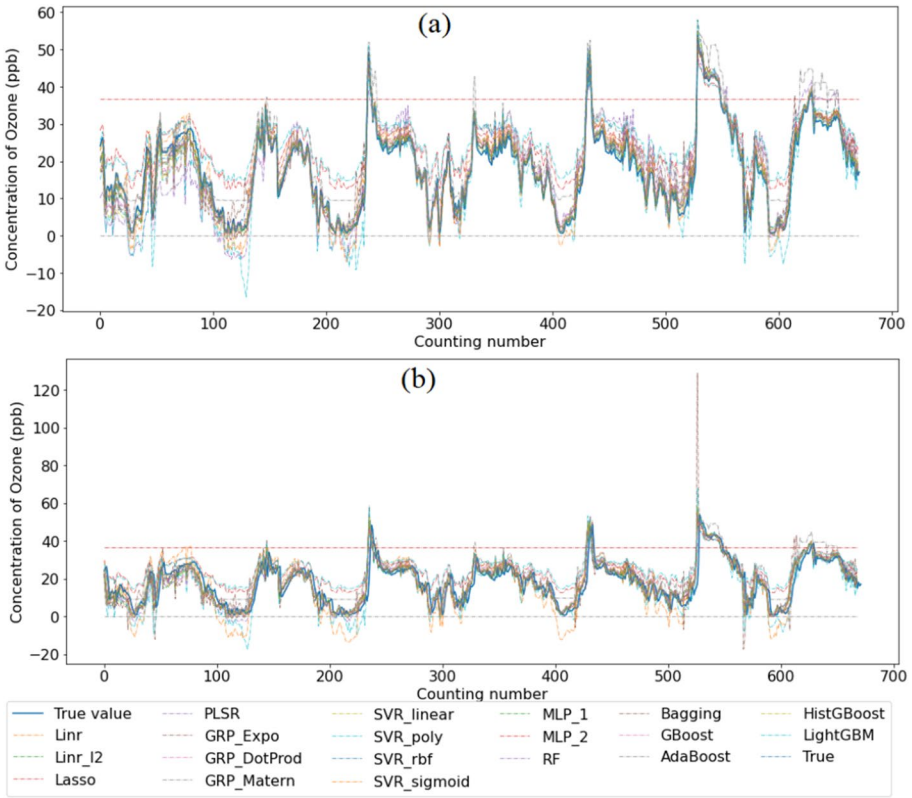

Fig. 13 Observed and predicted ozone concentrations based on testing data including **a** lag 1 ozone data, and **b** lag 3 ozone data

data. In addition, ozone time-series data is characterized by a dynamic nature, as indicated in ACF plots (Fig. 7). Hence, in this experiment, we consider lagged ozone data when predicting the concentration of ozone. In short, this experiment aims to investigate the impact of considering lagged data on ozone prediction accuracy. To this end, we consider adding the lag 1 and lag 3 ozone data when predicting the concentration of ozone. In the first scenario with lag 1 O_3, the input data contains seven variables (NO_2, absolute humidity, air temperature, absolute pressure, wind direction, SO_2, O_3.Lag1). In the second scenario with lag 3 O_3, there are nine input variables (NO_2, absolute humidity, air temperature, absolute pressure, wind direction, SO_2, O_3.Lag1, O_3.Lag2, and O_3.Lag3). The considered time-lagged ozone data are defined in Table 5.

The observed versus predicted ozone values are shown in Fig. 13 and their prediction errors are shown in Fig. 14. Visually we can see that as expected considering information from lagged ozone data enables improving the prediction of machine learning models

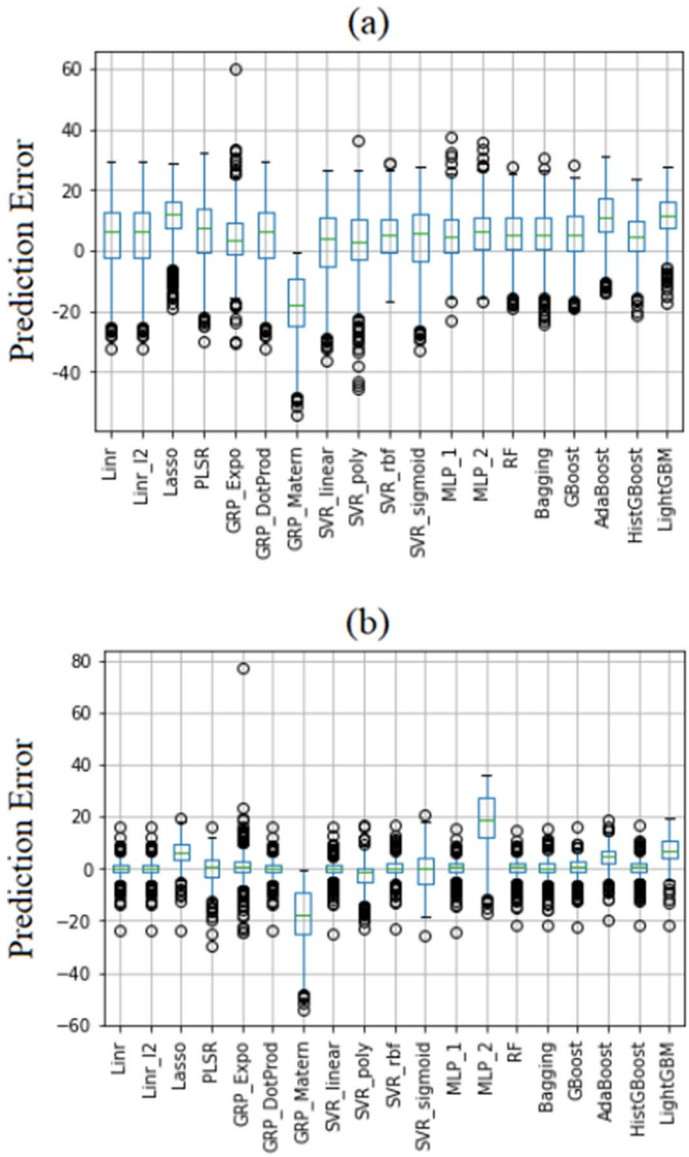

Fig. 14 Boxplot for prediction error of each model **a** by adding one time-lagged ozone value, and **b** three time-lagged ozone values in input data

(Fig. 13). Visually, from Fig. 14 we can see that the prediction errors have been significantly reduced by considering lagged ozone data with lag 3. Specifically, the mean error of some models, such as RF and SVR_{rbf} in this experiment with lag 3 are fluctuating nearly at zero and there is obvious mitigation on the error fluctuation. This clearly shows the improvement in the prediction quality of ozone concentrations when incorporating information from past data (Fig. 13a-b). It can be seen that by including lagged data with lag 3 the prediction from the considered models becomes very close to those observed data.

Table 6 and Table 7 provide the prediction results obtained by the investigated models using testing data with lag 1 and lag 3 ozone data, respectively. Results in Table 6 indicate that the SVR_{linear} model achieves the best score on every metric, and importantly, the

Table 6 Statistic indicators of the nineteen machine learning methods on ozone prediction for testing data with one-time-lagged ozone values

Model	RMSE ↓	MAE ↓	MAPE ↓	R-square ↑	J2 ↓	Time (min)
Linr	3.315/3.217	2.171/2.289	0.12/0.248	0.954/0.911	0.942	0.01
Linr_l2	3.315/3.22	2.173/2.293	0.121/0.249	0.954/0.911	0.944	0
Lasso	5.867/7.614	4.624/6.66	0.458/1.448	0.857/0.836	1.684	0.01
PLSR	4.987/6.558	3.849/5.405	0.213/0.63	0.897/0.629	1.729	0.01
GRP_Expo	2.685/4.166	1.797/3.086	0.097/0.414	0.97/0.86	2.407	7.79
GRP_DotProd	3.315/3.215	2.17/2.286	0.12/0.249	0.954/0.911	0.941	15.59
GRP_Matern	0.0/20.863	0.0/17.878	0.0/1.0	1.0/0.0	inf	21.61
SVR_linear	**3.388/3.032**	**2.087/1.979**	**0.114/0.213**	**0.952/0.921**	**0.801**	0.01
SVR_poly	3.903/5.351	2.678/3.778	0.222/0.624	0.937/0.801	1.88	5.37
SVR_rbf	3.097/3.189	1.998/2.277	0.12/0.298	0.96/0.914	1.06	6.04
SVR_sigmoid	4.501/4.079	3.352/3.241	0.215/0.521	0.916/0.86	0.821	6.92
MLP_1	3.315/3.199	2.174/2.234	0.126/0.232	0.954/0.912	0.931	0.28
MLP_2	15.514/21.707	12.668/19.503	1.287/4.222	0.0/0.0	1.958	1.09
RF	3.07/3.416	2.002/2.451	0.114/0.272	0.961/0.9	1.238	0.05
Bagging	1.381/3.731	0.863/2.756	0.046/0.281	0.992/0.881	7.299	0.16
GBoost	3.068/3.663	2.001/2.719	0.115/0.278	0.961/0.885	1.425	0.29
AdaBoost	4.908/6.396	3.942/5.485	0.326/0.985	0.912/0.844	1.698	0.12
HistGBoost	2.806/3.604	1.855/2.667	0.102/0.284	0.967/0.889	1.65	0.89
LightGBM	6.568/8.834	5.241/7.857	0.527/1.682	0.821/0.804	1.809	0.01

The value before slash represents the loss value in the training dataset and the latter in the testing dataset (training/testing), the best performance is bold and the line colored grey means over-fitting. In addition, the ↑ means the model is better when the value is larger, and the ↓ means the model is better when the value is smaller

Table 7 Statistic indicators of the nineteen machine learning methods on ozone prediction for testing data with three time-lagged ozone values

Model	RMSE ↓	MAE ↓	MAPE ↓	R-square ↑	J2 ↓	Time (min)
Linr	3.274/3.117	2.132/2.233	0.118/0.236	0.955/0.916	0.906	0.01
Linr_l2	3.275/3.126	2.135/2.242	0.118/0.237	0.955/0.916	0.911	0.01
Lasso	5.867/7.624	4.624/6.672	0.458/1.452	0.857/0.836	1.689	0.02
PLSR	4.334/4.817	3.134/3.663	0.18/0.419	0.922/0.8	1.235	0.01
GRP_Expo	2.21/5.826	1.506/3.593	0.077/0.447	0.98/0.713	6.95	9.7
GRP_DotProd	3.274/3.121	2.133/2.235	0.118/0.235	0.955/0.916	0.909	15.86
GRP_Matern	0.0/20.846	0.0/17.854	0.0/1.0	1.0/0.0	inf	21.98
SVR_linear	3.351/2.979	2.053/1.957	0.112/0.209	0.953/0.924	0.79	0.01
SVR_poly	3.859/5.315	2.636/3.747	0.22/0.637	0.938/0.804	1.897	6.99
SVR_rbf	3.045/3.13	1.952/2.207	0.117/0.288	0.962/0.916	1.057	6.32
SVR_sigmoid	6.106/6.862	4.568/5.585	0.35/1.179	0.845/0.6	1.263	11.4
MLP_1	3.24/3.185	2.105/2.275	0.125/0.252	0.956/0.913	0.966	0.37
MLP_2	15.514/21.658	12.678/19.454	1.284/4.221	0.0/0.0	1.949	1.39
RF	3.024/3.292	1.977/2.323	0.113/0.266	0.962/0.907	1.185	0.05
Bagging	1.346/3.453	0.833/2.433	0.045/0.264	0.992/0.897	6.581	0.17
GBoost	3.033/3.42	1.987/2.58	0.115/0.274	0.962/0.902	1.271	0.3
AdaBoost	4.629/5.911	3.663/5.04	0.311/0.92	0.918/0.855	1.631	0.17
HistGBoost	2.722/3.425	1.813/2.53	0.101/0.273	0.969/0.9	1.583	2.35
LightGBM	6.371/8.558	5.065/7.574	0.514/1.652	0.831/0.808	1.804	0.01

In addition, the ↑ means the model is better when the value is larger, and the ↓ means the model is better when the value is smaller

consuming time is only 0.01 min. In this case study, the SVR_{linear} is the best model with regard to accuracy and efficiency. In terms of adjustment, the SVR_{linear} model achieved successful results with the highest $R^2 of$ 0.924 and with the lowest RMSE of 2.979. Except for the GRP_{Matern}, all other models can catch the ozone trends with reasonable prediction errors. Thus, this result is confirming that the third-order lagged ozone data (O_3.Lag1, O_3.Lag2, and O_3.Lag3) are sufficient to enhance the prediction quality of the studied machine learning models. In addition, the most important factor to influence the performance of *SVR* is still the kernel, where the performance of *SVR* is worse if we use other non-linear kernels such as polynomial or sigmoid and their efficiency is reduced as well.

Results in Table 7 indicate that conclusions hold the same when considering the input with three time-lagged ozone values for ozone prediction. Specifically, the SVR_{linear} dominates the other models for ozone prediction with regard to accuracy and efficiency. Besides, the performance model becomes more stable by introducing three time-lagged ozone values. More precisely, the index for every metric is improved to a mild degree, and the prediction error range in Fig. 13 is much more narrow and concentrated. Therefore, we can conclude that the best model is SVR_{linear} in accuracy and efficiency.

Ensuring that the model with the best prediction results is statistically more significant than the other models is crucial. One common way to do this is to use statistical tests to compare the performance of the models, such as the Chow test, the Granger causality test, and the Breusch-Pagan test [75, 76]. In this study, we adopted the Diebold–Mariano test [75] to compare the prediction accuracy of the investigated models. It is commonly used in economics and finance to evaluate the performance of different forecasting models. The advantage of using the Diebold-Mariano Test is that it is a robust statistical test that can be used to compare the forecasting accuracy of different models without assuming a specific distribution for the forecast errors. Furthermore, the test is relatively simple to implement and interpret, making it a popular choice for comparing forecasting models.

The basic idea behind the Diebold-Mariano test is to compare the mean squared error of the forecasts produced by two models. In other words, the p-values generated from the test indicate whether there is a significant difference in the mean squared error between the forecasts produced by the two models. The null hypothesis is that there is no difference in the forecast accuracy between the two methods, while the alternative hypothesis is that one method is better than the other. To conduct the test, the differences between the two sets of forecast errors are calculated and then tested for statistical significance using a t-test. If the p-value is less than the significance level, it indicates that one method is significantly better than the other. Figure 15 shows the heatmap of the matrix of p-values generated from the Diebold-Mariano Test for each pair of models. The p-values range from 0 to 1, with values closer to 0 indicating stronger evidence to reject the null hypothesis and conclude that there is a significant difference in forecasting performance between the two models. A value less than 0.05 indicates that there is evidence to reject the null hypothesis and conclude that there is a significant difference in forecasting performance between the two models. Results in Fig. 15 show most of the p-values were found to be 0, indicating that there is strong evidence to reject the null hypothesis and conclude that there is a significant difference in forecasting performance

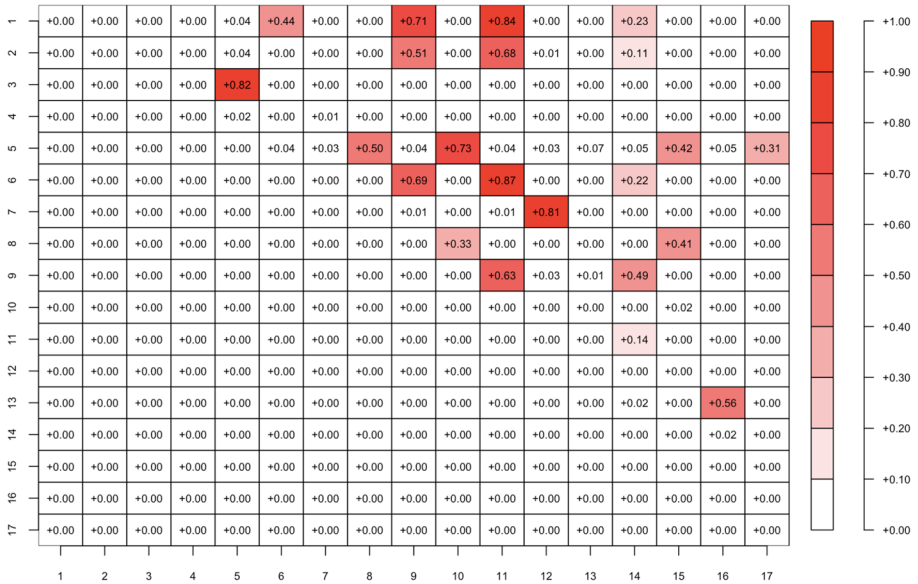

Fig. 15 DM test for the models (the number and corresponding model 1 : *Linr*, 2 : $Linr_{l2}$, 3 : *Lasso*, 4 : *PLSR*, 5 : GRP_{Expo}, 6 : $GRP_{DotProd}$, 7 : SVR_{linear}, 8 : SVR_{poly}, 9 : SVR_{rbf}, 10 : $SVR_{sigmoid}$, 11 : MLP_1, RF, 12 : *Bagging*, 13 : *GBoost*, 14 : *AdaBoost*, 15 : *HistGBoost*, 16 : *LightGBM*)

between most pairs of models. However, there are some exceptions, such as the pairs (Linr, GRP_*Expo*), (Linr_*l*2, GRP_*Expo*), and (PLSR, GRP_DotProd), where the p-values are relatively higher (0.04 or above), indicating that there may not be a significant difference in forecasting performance between these pairs of models. Notably, the results depicted in Fig. 15 highlight that the SVR model with a linear kernel is the best performer, with significant differences observed compared to the other models except for RF. Overall, this implies that the SVR with a linear kernel performed the best out of all the models tested, according to both the Diebold-Mariano test and evaluation metrics such as R^2 and MAPE. This model had the highest R^2 value of 0.924, indicating a strong correlation between the predicted and actual ozone concentrations. Additionally, it had the lowest MAPE of 0.209, indicating that the predicted values were, on average, very close to the actual values.

Conclusion

The detrimental impact of high ozone (O_3) pollution concentrations on human health and ecosystems underscores the importance of precise and efficient ozone concentration prediction for weather monitoring and environmental policymaking. In this study, we conducted a comparative analysis of various machine learning models to predict ozone concentrations. Real data collected at KAUST, including meteorological and pollution variables, were used to evaluate prediction accuracy. Our results showed that the investigated machine learning models failed to capture ozone trends when considering only meteorological variables as inputs. However, we demonstrated that incorporating meteorology and air pollutants as input significantly improves the prediction performance of machine learning models. We further identified that including time-lagged ozone data substantially enhances the prediction quality of the machine-learning models.

Specifically, by considering input with three time-lagged ozone values, the machine learning models provide ozone prediction more accurately than previous experiments, achieving an RMSE improvement of 300% and 200%, respectively. Importantly, with the lag information, the best model only needs 0.01 s, which is over 900 times faster than the other two best models in the first two experiments.

Overall, this study highlights the importance of incorporating time-lagged ozone data and suggests that SVR outperforms the other machine learning models for ozone prediction. This finding could significantly benefit air quality monitoring and management, improving public health outcomes. Our work also opens up new research problems, such as exploring the scalability of different models and the deployment of algorithms in real-life scenarios. In future studies, we plan to investigate the potential of graph models, such as Graph Convolutional Networks (GCN) [77], for ozone prediction when data from multiple stations are available. Additionally, we aim to exploit and deploy advanced spatiotemporal models for ozone prediction and monitoring.

Acknowledgements
The authors would like to express their gratitude towards Health, Safety and Environment (HSE) Department at KAUST, for providing the data used in this study.

Author contributions
These authors contributed equally to this work. P.Q.: Conceptualization, formal analysis, investigation, methodology, software, writing–original draft, and writing–review and editing. F.H.: Conceptualization, formal analysis, investigation, methodology, supervision, writing–original draft, and writing–review and editing. Y.S.: Investigation, conceptualization, formal analysis, methodology, writing–review and editing, funding acquisition, and supervision. All authors have read and agreed to the published version of the manuscript.

Funding
This publication is based upon work supported by King Abdullah University of Science and Technology (KAUST) Research Funding (KRF) from the Climate and Livability Initiative (CLI) under Award No. ORA-2022-5339.

Data Availability
The datasets generated and/or analyzed during the current study are not publicly available because the authors do not have permissions from the Health, Safety and Environment (HSE) Department at KAUST to share them publicly.

Declarations

Ethics approval consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no conflict of interest.

Received: 8 November 2022 Accepted: 2 May 2023
Published online: 15 May 2023

References

1. Wu A, Harrou F, Dairi A, Sun Y. Machine learning and deep learning-driven methods for predicting ambient particulate matters levels: a case study. Concurr Comput: Pract Exp. 2022;34(19): e7035.
2. Zhang JJ, Wei Y, Fang Z. Ozone pollution: a major health hazard worldwide. Fronti Immunol. 2019;10:2518.
3. Wilkinson S, Mills G, Illidge R, Davies WJ. How is ozone pollution reducing our food supply? J Exp Bot. 2012;63(2):527–36.
4. Poumadere M, Mays C, Le Mer S, Blong R. The 2003 heat wave in france: dangerous climate change here and now. Risk Anal: Int J. 2005;25(6):1483–94.
5. Board CAS. Ozone Effects - Overview of the harmful health effects of ground level ozone; 2016. https://ww2.arb.ca.gov/resources/fact-sheets/ozone-effects. Accessed 10 May 2023.
6. Yafouz A, Ahmed AN, Zaini N, El-Shafie A. Ozone concentration forecasting based on artificial intelligence techniques: a systematic review. Water Air Soil Pollut. 2021;232(2):1–29.

7. Cabaneros SM, Calautit JK, Hughes BR. A review of artificial neural network models for ambient air pollution prediction. Environ Model Softw. 2019;119:285–304.
8. Cheng Y, He L-Y, Huang X-F. Development of a high-performance machine learning model to predict ground ozone pollution in typical cities of china. J Environ Manag. 2021;299:113670.
9. Harrou F, Kadri F, Khadraoui S, Sun Y. Ozone measurements monitoring using data-based approach. Process Safety Environ Prot. 2016;100:220–31.
10. Bouyeddou B, Harrou F, Dairi A, Sun Y. Monitoring ground-level ozone pollution based on a semi-supervised approach. In: 2022 7th International Conference on Frontiers of Signal Processing (ICFSP); 2022. p. 194–198. https://doi.org/10.1109/ICFSP55781.2022.9924670
11. Li L, An J, Shi Y, Zhou M, Yan R, Huang C, Wang H, Lou S, Wang Q, Lu Q, et al. Source apportionment of surface ozone in the yangtze river delta, china in the summer of 2013. Atmos Environ. 2016;144:194–207.
12. Zhou G, Xu J, Xie Y, Chang L, Gao W, Gu Y, Zhou J. Numerical air quality forecasting over eastern china: An operational application of wrf-chem. Atmos Environ. 2017;153:94–108.
13. Feng R, Zheng H-J, Zhang A-R, Huang C, Gao H, Ma Y-C. Unveiling tropospheric ozone by the traditional atmospheric model and machine learning, and their comparison: A case study in hangzhou, china. Environ Pollut. 2019;252:366–78.
14. Seinfeld JH, Pandis SN. Atmos Chem Phys: Air Poll Clim Change. John Wiley & Sons; 2016.
15. Cook PA, Wheater P. Using Statistics to Understand the Environment. Abingdon: Routledge; 2005.
16. Smith RL. Statistics of extremes, with applications in environment, insurance, and finance. Extreme values in finance, telecommunications, and the environment. 2003;20–97.
17. Harrou F, Fillatre L, Bobbia M, Nikiforov I. Statistical detection of abnormal ozone measurements based on constrained generalized likelihood ratio test. In: 52nd IEEE Conference on Decision and Control. IEEE; 2013. p. 4997–5002
18. Duenas C, Fernandez M, Canete S, Carretero J, Liger E. Stochastic model to forecast ground-level ozone concentration at urban and rural areas. Chemosphere. 2005;61(10):1379–89.
19. Sousa S, Martins FG, Alvim-Ferraz MC, Pereira MC. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environ Model Softw. 2007;22(1):97–103.
20. Ezimand K, Kakroodi A. Prediction and spatio-temporal analysis of ozone concentration in a metropolitan area. Ecol indic. 2019;103:589–98.
21. Lv J, Xu X. Prediction of daily maximum ozone levels using lasso sparse modeling method; 2020. arXiv preprint arXiv:2010.08909. Accessed 10 May 2023.
22. Napi NNLM, Mohamed MSN, Abdullah S, Mansor AA, Ahmed AN, Ismail M. Multiple linear regression (mlr) and principal component regression (pcr) for ozone (o3) concentrations prediction. In: IOP Conference Series: Earth and Environmental Science, vol. 616. IOP Publishing; 2020, p. 012004
23. Harrou F, Nounou M, Nounou H. Statistical detection of abnormal ozone levels using principal component analysis. Int J Eng Technol. 2012;12(6):54–9.
24. Đorđević F, Kostić SM. Axial strength prediction of square cfst columns based on the ann model. In: First Serbian International Conference on Applied Artificial Intelligence, Kragujevac, Serbia; 2022
25. Đorđević F, Kostić SM. Estimation of ultimate strength of slender ccfst columns using artificial neural networks. In: 16th Congress of Association of Structural Engineers of Serbia; 2022
26. Mester G, Filipovi'c N. Editorial-computational modeling and machine learning in biomedical and engineering application. IPSI Bgd Trans Internet Res. 2022;18(1). http://ipsitransactions.org/journals/papers/tir/2022jan/fullPaper.pdf
27. Dairi A, Harrou F, Khadraoui S, Sun Y. Integrated multiple directed attention-based deep learning for improved air pollution forecasting. IEEE Trans Instrum Meas. 2021;70:1–15.
28. Harrou F, Dairi A, Sun Y, Kadri F. Detecting abnormal ozone measurements with a deep learning-based strategy. IEEE Sens J. 2018;18(17):7222–32.
29. Ettouney RS, Mjalli FS, Zaki JG, El-Rifai MA, Ettouney HM. Forecasting of ozone pollution using artificial neural networks. Management of Environmental Quality: An International Journal. 2009.
30. Su X, An J, Zhang Y, Zhu P, Zhu B. Prediction of ozone hourly concentrations by support vector machine and kernel extreme learning machine using wavelet transformation and partial least squares methods. Atmos Poll Res. 2020;11(6):51–60.
31. Chen G, Chen J, Dong G-H, Yang B-Y, Liu Y, Lu T, Yu P, Guo Y, Li S. Improving satellite-based estimation of surface ozone across china during 2008–2019 using iterative random forest model and high-resolution grid meteorological data. Sustain Cities Soc. 2021;69:102807.
32. Bhuiyan MAM, Sahi RK, Islam MR, Mahmud S. Machine learning techniques applied to predict tropospheric ozone in a semi-arid climate region. Mathematics. 2021;9(22):2901.
33. Jumin E, Zaini N, Ahmed AN, Abdullah S, Ismail M, Sherif M, Sefelnasr A, El-Shafie A. Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction. Eng Appl Comput Fluid Mech. 2020;14(1):713–25.
34. Yılmaz A. Ozone level prediction with machine learning algorithms. J Aeronaut Space Technol. 2021;14(2):177–83.
35. Bhuiyan MAM, Mahmud S, Sarmin N, Elahee S. A study on statistical data mining algorithms for the prediction of ground-level ozone concentration in the el paso-juarez area. Aerosol Sci Eng. 2020;4(4):293–305.
36. Jiang N, Riley ML. Exploring the utility of the random forest method for forecasting ozone pollution in sydney. J Environ Protect Sustainable develop. 2015;1:245–54.
37. Allu SK, Srinivasan S, Maddala RK, Reddy A, Anupoju GR. Seasonal ground level ozone prediction using multiple linear regression (mlr) model. Model Earth Syst Environ. 2020;6(4):1981–9.

38. Chelani AB. Prediction of daily maximum ground ozone concentration using support vector machine. Environ Monit Assess. 2010;162(1):169–76.
39. Arsić M, Mihajlović I, Nikolić D, Živković Ž, Panić M. Prediction of ozone concentration in ambient air using multilinear regression and the artificial neural networks methods. Ozone: Sci Eng. 2020;42(1):79–88.
40. Hoshyaripour G, Brasseur G, Andrade MDF, Gavidia-Calderón M, Bouarar I, Ynoue RY. Prediction of ground-level ozone concentration in são paulo, brazil: Deterministic versus statistic models. Atmos Environ. 2016;145:365–75.
41. Braik M, Sheta A, Al-Hiary H. Hybrid neural network models for forecasting ozone and particulate matter concentrations in the republic of china. Air Qual, Atmos Health. 2020;13(7):839–51.
42. Ren X, Mi Z, Georgopoulos PG. 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. Environ Int. 2020;142:105827.
43. Oufdou H, Bellanger L, Bergam A, Khomsi K. Forecasting daily of surface ozone concentration in the grand casablanca region using parametric and nonparametric statistical models. Atmosphere. 2021;12(6):666.
44. Juarez EK, Petersen MR. A comparison of machine learning methods to forecast tropospheric ozone levels in delhi. Atmosphere. 2021;13(1):46.
45. Marvin D, Nespoli L, Strepparava D, Medici V. A data-driven approach to forecasting ground-level ozone concentration. Int J Forecast. 2021;38(3):970–87.
46. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Stat Soc: Ser B (Methodological). 1996;58(1):267–88.
47. Hubert M, Branden KV. Robust methods for partial least squares regression. J Chemome: J Chemom Soc. 2003;17(10):537–49.
48. Harrou F, Sun Y, Hering AS, Madakyaru M, Dairi A. Linear latent variable regression (lvr)-based process monitoring. Amsterdam: Elsevier BV; 2021.
49. Cai H, Jia X, Feng J, Li W, Hsu Y-M, Lee J. Gaussian process regression for numerical wind speed prediction enhancement. Renew Energy. 2020;146:2112–23.
50. Harrou F, Saidi A, Sun Y, Khadraoui S. Monitoring of photovoltaic systems using improved kernel-based learning schemes. IEEE J Photovoltaics. 2021;11(3):806–18.
51. Bousquet O, von Luxburg U, Rätsch G. Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2–14, 2003, Tübingen, Germany, August 4–16, 2003, Revised Lectures vol. 3176. Springer; 2011. p. 63–72
52. Genton MG. Classes of kernels for machine learning: a statistics perspective. J Machine Learn Res. 2001;2(Dec):299–312.
53. Yu P-S, Chen S-T, Chang I-F. Support vector regression for real-time flood stage forecasting. J Hydrol. 2006;328(3–4):704–16.
54. Hong W-C, Dong Y, Chen L-Y, Wei S-Y. Svr with hybrid chaotic genetic algorithms for tourism demand forecasting. Appl Soft Comput. 2011;11(2):1881–90.
55. Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat comput. 2004;14(3):199–222.
56. Dong Z, Zhang Z, Dong Y, Huang X. Multi-layer perception based model predictive control for the thermal power of nuclear superheated-steam supply systems. Energy. 2018;151:116–25.
57. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
58. Zhou Z. Ensemble Methods: Foundations and Algorithms. Boca Raton: CRC press; 2012. p. 15–55.
59. Rincy TN, Gupta R. Ensemble learning techniques and its efficiency in machine learning: A survey. In: 2nd International Conference on Data, Engineering and Applications (IDEA), IEEE; 2020. p. 1–6
60. weather K. Weather Monitoring Equipment; 2022. https://hse.kaust.edu.sa/services/weather-monitoring-equipment. Accessed 10 May 2023.
61. Zou M, Djokic SZ. A review of approaches for the detection and treatment of outliers in processing wind turbine and wind farm measurements. Energies. 2020;13(16):4228.
62. Peña M, Ortega P, Orellana M. A novel imputation method for missing values in air pollutant time series data. In: 2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI). IEEE; 2019, p. 1–6
63. Wardana I, Gardner JW, Fahmy SA. Estimation of missing air pollutant data using a spatiotemporal convolutional autoencoder. Neural Comput Appl. 2022;34(18):16129–54.
64. Lin W-C, Tsai C-F. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev. 2020;53(2):1487–509.
65. Moshenberg S, Lerner U, Fishbain B. Spectral methods for imputation of missing air quality data. Environ Syst Res. 2015;4(1):1–13.
66. Fenn ME, Poth MA, Bytnerowicz A, Sickman JO, Takemoto BK. Effects of ozone, nitrogen deposition, and other stressors on montane ecosystems in the sierra nevada. Dev Environ Sci. 2003;2:111–55.
67. Brulfert G, Galvez O, Yang F, Sloan J. A regional modelling study of the high ozone episode of June 2001 in southern ontario. Atmos Environ. 2007;41(18):3777–88.
68. Bodor Z, Bodor K, Keresztesi Á, Szép R. Major air pollutants seasonal variation analysis and long-range transport of pm10 in an urban environment with specific climate condition in transylvania (romania). Environ Sci Poll Res. 2020;27(30):38181–99.
69. Qu Y, Wang T, Cai Y, Wang S, Chen P, Li S, Li M, Yuan C, Wang J, Xu S. Influence of atmospheric particulate matter on ozone in nanjing, china: observational study and mechanistic analysis. Adv Atmos Sci. 2018;35(11):1381–95.
70. Box GEP, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. John Wiley & Sons, 2015.
71. Song F, Guo Z, Mei D. Feature selection using principal component analysis. In: 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization, vol. 1. IEEE; 2010; p. 27–30

72. Gao L, Wu W. Relevance assignation feature selection method based on mutual information for machine learning. Knowl Based Syst. 2020;209:106439.
73. Hasan MAM, Nasser M, Ahmad S, Molla KI. Feature selection for intrusion detection using random forest. J Inf Secur. 2016;7(3):129–40.
74. Santurtún A, González-Hidalgo JC, Sanchez-Lorenzo A, Zarrabeitia MT. Surface ozone concentration trends and its relationship with weather types in spain (2001–2010). Atmos Environ. 2015;101:10–22.
75. Diebold FX, Mariano RS. Comparing predictive accuracy. J Bus Econ stat. 2002;20(1):134–44.
76. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Machine Learn Res. 2006;7:1–30.
77. Wang C, Zhu Y, Zang T, Liu H, Yu J. Modeling inter-station relationships with attentive temporal graph convolutional network for air quality prediction. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining; 2021. p. 616–634

Publisher's Note