Atmospheric Environment 323 (2024) 120396

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Predicting $PM_{2.5}$ levels and exceedance days using machine learning methods

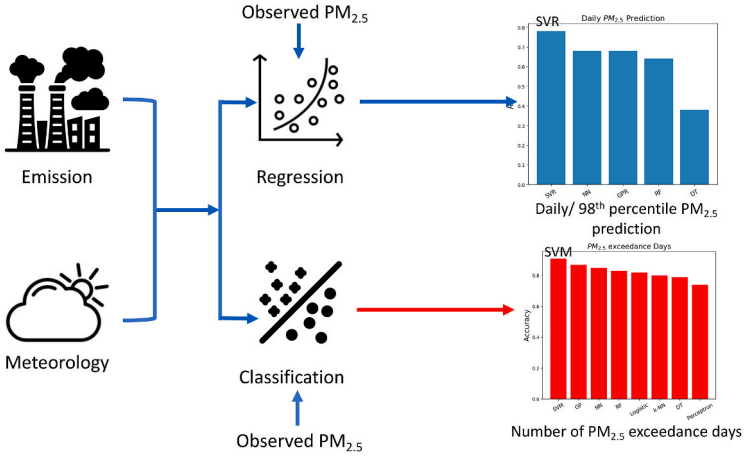
Ziqi Gao [a,d,*], Khanh Do [b,c], Zongrun Li [a], Xiangyu Jiang [d], Kamal J. Maji [a], Cesunica E. Ivey [b,e], Armistead G. Russell [a]

[a] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA
[b] Department of Chemical and Environmental Engineering, University of California, Riverside, CA, USA
[c] Center for Environmental Research and Technology, Riverside, CA, USA
[d] Georgia Environmental Protection Division, Atlanta, GA, 30354, USA
[e] Now at Department of Civil and Environmental Engineering, University of California, Berkeley, CA, USA

HIGHLIGHTS

- SVR is more accurate on daily $PM_{2.5}$, particularly $PM_{2.5}$ exceedances predictions.
- Surface RH was the most important meteorological factor for $PM_{2.5}$ prediction.
- The impact of emissions on $PM_{2.5}$ was significant before 2010 but reduced thereafter.
- ML models predict past better than future; all the ML models are limited at extremes.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
$PM_{2.5}$
South coast air basin
Support vector machine
Random forest
Neural network

ABSTRACT

Machine learning methods are increasingly being used in the field of air quality research to investigate the relationship between air pollutant levels, emissions, and meteorological changes over time. This research is used for both scientific investigation, and policy assessment and development. However, there is a lack of studies that have compared the performance of different machine learning methods. To address this gap, this paper employed various machine learning techniques, including decision tree, random forest (RF), support vector machine (SVM), support vector regression (SVR), k-nearest neighbor, neural network, and Gaussian process regression, to predict daily average $PM_{2.5}$ levels and the number of days with $PM_{2.5}$ exceedance in the South Coast Air Basin of California from 2000 to 2019. The models were trained using meteorological factors, estimated emissions, and large-scale climate indices as inputs. The SVR model demonstrated the highest predictive accuracy for $PM_{2.5}$

* Corresponding author. School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA.
E-mail address: zgao71@gatech.edu (Z. Gao).

https://doi.org/10.1016/j.atmosenv.2024.120396
Received 26 October 2023; Received in revised form 3 February 2024; Accepted 6 February 2024
Available online 10 February 2024
1352-2310/Published by Elsevier Ltd.

levels and the SVM model gave the most accurate results for predicting the number of days with $PM_{2.5}$ exceedances. Conversely, the decision tree model performed the least accurately. The results also showed that emissions have a greater impact on $PM_{2.5}$ levels over time compared to meteorological factors, though meteorology is responsible for daily variability. The most important meteorological factors were identified as surface relative humidity and relative humidity at 850 mbars, which are related to partitioning, cloud cover and wet deposition. We conducted sensitivity tests on the model's response to emissions and meteorological factors. The predicted $PM_{2.5}$ from RF and SVR showed large correlations with emissions at the early period (2000–2010). However, the changes were minimal in more recent years (2011–2019), implying that there are biases in machine learning models, in which the models consistently predict the minimum $PM_{2.5}$ levels at a baseline.

1. Introduction

Fine particulate matter ($PM_{2.5}$) is one of the criteria pollutants regulated by the National Ambient Air Quality Standard (NAAQS) and poses significant threat to human health and the climate (Dockery et al., 1993; Gurgueira et al., 2002; Pinault et al., 2016; Pope et al., 2002; Schwartz, 1994). Predicting $PM_{2.5}$ levels is a complex task, as it can be emitted directly from sources or formed in the atmosphere through chemical reactions between precursor pollutants, making it challenging to predict $PM_{2.5}$ concentrations and attribute levels to specific sources or processes.

Two methods commonly used to predict $PM_{2.5}$ concentrations are chemical transport models (CTMs) (e.g., the Community Multiscale Air Quality (CMAQ) and GEOS-Chem models) and empirical methods (e.g., traditional regression models and machine learning methods). These two methods have their own advantages and limitations. CTMs are designed to capture complex atmospheric processes and chemical reactions using first principal relationships, following compounds from their emission to ultimate fate. However, the dynamics of $PM_{2.5}$ formation are not fully understood, and there are uncertainties in the input meteorological and emissions data (Jiang and Yoo, 2018; Rybarczyk and Zalakeviciute, 2022; Vlachogianni et al., 2011; Xu et al., 2021). Furthermore, CTM performance in predicting daily $PM_{2.5}$ levels shows both bias and variance and is also impeded by the significant information and computational requirements, particularly for long-period application. In contrast, machine learning methods, such as decision tree (DT), random forest (RF), Gaussian process regression (GPR), support vector machine (SVM), support vector regression (SVR), neural network (NN), and k-nearest neighbor (KNN), are computationally faster and more easily constructed, but require a large amount of training data to train, have low interpretability and do not provide first-principle relationships between emissions and air quality (Bi et al., 2022; Gao et al., 2022). Machine learning models are challenging to represent with explicit mathematical equations compared to statistical models, largely due to their complex structures and the numerous parameters they include. For example, the RF algorithm is composed of multiple decision trees of varying depths, making it is hard to explain by reviewing each tree's structure. Also, the number of parameters in a random forest can range from thousands to millions, depending on the number of trees and the depth of each tree.

In this study, we focus on using machine learning models to predict $PM_{2.5}$ levels and compare the performance and computational requirements for DT, RF, GPR, SVM, SVR, NN, and KNN methods, and investigate how emissions and meteorology influence daily and annual $PM_{2.5}$ levels. Previous studies have shown that machine learning-based models can accurately predict $PM_{2.5}$ levels using emissions and meteorological data (Chen et al., 2020; Gao et al., 2023a, 2023b; Gupta et al., 2021; Kleine Deters et al., 2017; Kumar et al., 2020; Minh et al., 2021). We use binary classification models to predict the number of $PM_{2.5}$ exceedance days (defined as concentrations above 12 $\mu g/m^3$ for annual average $PM_{2.5}$ and above 35 $\mu g/m^3$ for 98th percentile daily average $PM_{2.5}$ based on NAAQS), which related to human health (e.g., respiratory and cardiovascular diseases, cancer and mortality rate). The predictive capabilities of these models can offer insights into the spatial-temporal patterns in $PM_{2.5}$ exceedance days and the variable importance of these models can show a potentially effective way to reduce $PM_{2.5}$ concentrations to policymakers.

2. Methods and data

We applied various common machine learning techniques including DT, RF, GPR, SVR, and NN to predict daily $PM_{2.5}$ concentrations. In addition, we used eight classification methods to predict the number of $PM_{2.5}$ exceedance days, namely perceptron, logistic regression, and KNN, SVM, DT, RF, GPR, NN.

2.1. Methods

2.1.1. Decision tree and random forest

The DT model is a commonly used machine learning method for classification and regression, which is capable of capturing non-linear relationships between the dependent variable and independent indicators (Breiman et al., 2017; Hastie et al., 2009; Quinlan, 2014). A DT consists of three parts: a root node, leaf nodes, and branches. To build a DT, the model considers all the features at the root node and selects the split that yields the highest accuracy (least cost using the sum of the difference between the observations and predictions for regression or the Gini score for classification). The feature at the root node is the most important feature of the predicting dependent variable. The data is then split using the value of this feature, and the process is repeated recursively on each subset until further splitting does not improve the model or the predictions at each leaf node are identical. The DT model is easy to build and visualize, and can provide feature importance through the order of the nodes. Data preparation before building the DT is simple: indicators do not need to be on the same scale using standardization and feature selection is not required. The DT model can handle multi-dimensional data and numerical and categorical features. However, DT can suffer from overfitting and lack of stability when the number of indicators is large. This issue can be mitigated by adjusting hyperparameters such as increasing the minimum amount of data in each leaf node and reducing the maximum depth of the tree. Additionally, pruning can be used to remove branches with low variable importance. The main limitations of DT are instability and the potential for suboptimal model selection due to the method used to choose the root node feature.

A RF is a combination of multiple DTs (Tin Kam, 1995) that addresses some of the limitations of a single DT model. Unlike DT, the RF model randomly selects the subsets from the training dataset to train each tree and randomly selects features at each leaf node, reducing variance and increasing model stability. The final prediction of the random forest model is the average of all trees' prediction in regression, or the majority vote of all DTs is the final RF in classification. Hyperparameters such as the number of trees and number of features at each leaf node can be tuned to improve the model performance of an RF model. However, the tuning process may result in overfitting, so cross-validation is necessary to identify the 'optimal' RF model. The RF model generally provides more accurate and stable predictions than a single DT and can automatically consider feature interactions. The main

limitation of the RF model is increased computational requirements compared to a single DT. The "rpart" and "randomForest" packages were used in R program to build the DT and RF model (Liaw and Wiener, 2002; Loh, 2011).

2.1.2. Gaussian process regression

The GPR model is a nonparametric and kernel-based approach that can be applied to both classification and regression problems (Zhang et al., 2018). GPR utilized the Bayes rule to make predictions. One of the key factors that impacts the GPR model's performance is the choice of kernel function (covariate function). There are various kernel functions available for use in the GPR model, such as linear, radial basis functional (RBF), white noise, exponentiated quadratic, rational quadratic, and periodic kernels. An advantage of the GPR model is that it can be applied even when there is no specific relationship defined between the response variable and predictors. Moreover, the predictions tend to be smooth and flexible if the kernel function is appropriately selected. Additionally, the GPR model does not require a large amount of data for model training. However, the choice of the kernel is critical for the model's performance, and the model may perform poorly with an incorrect kernel. Furthermore, the running time of this model can be longer than other complex machine learning models when applied to datasets of similar sizes (Belyaev et al., 2014). We built the GPR model using the GaussianProcessRegressor from the sklearn.gaussian_process library in Python (Rasmussen and Williams, 2006).

2.1.3. Support vector machine and support vector regression

The SVM model is used for classification, while the SVR model is used for regression. Both models are capable of handling nonlinear relationships. SVM and SVR primarily work by categorizing data points into distinct groups. In the case of SVM, this is achieved by finding an optimal hyperplane, which can be thought of as a boundary line in two dimensions or a plane in three dimensions that best separates the different classes. For SVR, the model identifies a line or curve depending on the dimension of data that best predicts the target values. This hyperplane is a high-dimensional plane with the furthest distance to the closest data point in each class. A soft margin is utilized to avoid overfitting by finding the minimum value after adding a loss function to the distance between the hyperplane and the data point. To address nonlinear relationships, the kernels are introduced to map the features to the high-dimensional data and make it possible to separate the datasets into classes with a hyperplane. Kernels include linear and nonlinear types, such as polynomial, Gaussian, hyperbolic tangent, RBF, and sigmoid kernels. In this study, we used RBF. The RBF kernel was chosen due to its flexibility in handling non-linearities and its capability to approximate a wide variety of functions with fewer hyperparameters, which is widely used. Two hyperparameters are tuned to enhance model performance: cost and epsilon. Epsilon is a regularizer that defines the magnitude of the margins, while the cost determines the number of data points outside the margins. The SVM model is cost and memory efficient and performs well with nonlinear relationships. It is ideal for small datasets, compared to neural network and other complex machine learning methods. The e1071 package was used with R program to build the SVR and SVM models (Chang and Lin, 2011; Fan et al., 2005).

2.1.4. Perceptron and neural network

A perceptron is a type of supervised machine learning model used for classification (Rosenblatt, 1958). A simple neural network model that is capable of classifying input data into two categories, consisting of a single-layer neural network with a linear binary classifier. The algorithm behind the perceptron is straightforward: it computes the weighted sum of the input data and their weights and applies an activation function to it. The activation function introduces a nonlinear factor to the weighted sum, which is otherwise a linear equation. The choice of activation function is critical in the development of the perceptron model. There are many types of activation functions to choose from, including linear, exponential, sign, sigmoid, hyperbolic tangent, logistic, and rectified linear unit (ReLU).

Apart from the perceptron, other neural network methods include feedforward, multiple layer perceptron, convolutional, RBF/recurrent, sequence to sequence, and modular neural network. The primary differences among these methods are the data flow (i.e., the sequence in which data move through the neural network), the structure between the input and output data (e.g., the number of the hidden layer and recurrent layer, and the number of neurons in each layer), and the choice of the activation function. The multiple-layer perceptron neural network is the foundation of all neural network methods. It contains multiple hidden layers between the input and output layers, with each hidden layer having multiple perceptrons. One advantage of this model is that it offers greater flexibility in terms of its structure, enabling researchers to design different structures for solving different problems. However, this model is complex and requires finding the optimal activation function and the number of layers. Depending on the number of hidden layers, the running time of the model can be relatively long compared to other methods. We built the NN model using the sklearn.neural_network module in Python (He et al., 2015; Kingma and Ba, 2014).

2.1.5. k-nearest neighbors

The k-nearest neighbors (k-NN) method is a non-parametric, supervised machine learning algorithm used for classification (Cover and Hart, 1967). Unlike other methods, it is a "lazy" learning algorithm that does not train a model between the response variable and independent features. Instead, it stores the data during the training process. The k-NN algorithm classifies data by measuring the similarity between the input data and the data from the training set in the relevant classes. This is determined by the "distance" between them, which can be computed using various methods such as Euclidean, Manhattan, Minkowski, and Hamming distances (Dudoit et al., 2002; Jaskowiak et al., 2012). The key parameter to consider when building the k-NN algorithm is the value of k, which refers to the number of neighbors in a class with the closest distance to the new data that can assign to that class. Generally, a large k value is recommended for the classification problem, especially when dealing with high variance data. However, different k values can result in underfitting and overfitting, so a cross-validation test is necessary to determine the optimal k value. Compared to other complex machine learning algorithms, such as neural network and random forest, the k-NN method is relatively simple and only requires consideration of two hyperparameters during development: the method used to calculate the distance and the choice of k value. However, it can be prone to overfitting or underfitting depending on the k value and may require a large memory to store the dataset. The class package was used with R program to build the k-NN model (Ripley, 2007; Venables and Ripley, 2013).

2.1.6. Logistic regression

Logistic regression is a supervised machine learning model and can be used to solve classification problems based on probability. Three types of logistic regression are binomial, multinomial, and ordinal. The sigmoid function is applied to the logistic regression model, and the output is the probability, which is in the range of 0 and 1:

$$y = \frac{1}{1 + e^{-x}} \quad \text{(Equation 1)}$$

This method is easier to build than other complex machine learning methods and has a low computational cost. It is better for a linear relationship between the response variable and independent indicators (Cramer, 2002). The coefficient of each independent variable can be used to evaluate the variable importance. This model has two main limitations: one is that this model does not work well with nonlinear relationships, and the other is that the model performance is generally worse than other complex machine learning methods. We built the

Logistic regression model using glm function in R program.

2.2. Data

Observed $PM_{2.5}$ levels in the South Coast Air Basin (SoCAB) were obtained from the California Air Resources Board (CARB) archives (CARB, 2020b). Historical $PM_{2.5}$ mass concentrations from 2000 to 2019 were used to train all the empirical models. The Rubidoux site was the primary focus of this study due to its longer $PM_{2.5}$ record and relatively higher $PM_{2.5}$ levels.

Surface maximum/average/minimum temperature, average wind speed, wind direction, and average relative humidity (RH) data were obtained from CARB and National Centers for Environmental Information (NCEI) (CARB, 2020a; NCEI, 2020). The maximum and average solar radiation (SR) were obtained from a composite of SR data acquired from CARB, U.S. Environmental Pollution Agency (EPA) Air Quality System, and the National Solar Radiation Database (NSRD). The upper-level height, temperature, RH, wind speed, and wind direction at 500 mbars (mb) and 850 mb were obtained from National Oceanic and Atmospheric Administration (NOAA). These factors are associated with synoptic scale weather, and are the indicators of the expected local temperature and precipitation. Large-scale climate indices were obtained from the NOAA Climate Prediction Center (CPC, 2020). Local meteorological conditions were related to large-scale climate patterns, including temperature, rainfall, and wind speed, which have an effect on the air pollutants formation.

Estimated emissions for 2000 to 2019 were calculated for nitrogen oxides (NO_x), sulfur dioxides (SO_2), primary $PM_{2.5}$, ammonia (NH_3), and volatile organic compounds (VOCs) emissions (CARB, 2022). In the performance evaluation, we focus on daily $PM_{2.5}$ rather than annual predictions, as that is a more demanding application, although we provide annual statistics.

2.3. Sensitivity test

We adjusted the key emissions and meteorological indicators (excluding maximum temperature) by 20%, 50%, 80%, 120%, and 150% to assess their response to $PM_{2.5}$ levels using RF and SVR, based on the rank of variable importance. These indicators include daily average relative humidity, wind speed, relative humidity at 850 mb, and emissions of NH_3, NOx, primary $PM_{2.5}$, SO_2, and VOC. Considering the sensitivity of maximum temperature, we only increased it by 1, 2, and 3 °C and decreased it by 1 and 2 °C to test its impact.

3. Results

3.1. Model performance

We evaluated the model performance based on two sets of evaluation metrics and several common metrics, including the coefficient of determination (R^2), mean bias (MB), and root mean square error (RMSE) for regression (SI Eqns. 1-3) and accuracy, precision, F1 score, and probability of detection (POD) for classification (SI Eqns. 4-8). The equations for these metrics are shown in supplementary material. We trained all the machine learning models using the whole dataset in the period from 2000 to 2019. To assess overfitting, we applied 10-fold cross-validation, where the input data was shuffled and 90% was randomly selected as the training dataset. The remaining 10% was used for testing.

3.2. Comparison among regression models

To build the regression models for daily $PM_{2.5}$ mass concentrations at the Rubidoux site, we used two sets of independent indicators. The first set (VAR1) was created by excluding strongly correlated independent variables (Fig. S1) and using stepwise regression and F statistics to select significant indicators (Pope and Webster, 1972). The second set (VAR2) included all available indicators (Table S1).

3.2.1. Decision tree

We employed the selected indicators from VAR1 and VAR2 to build the DT model. After pruning, the model performance of these two models was almost identical for the daily average $PM_{2.5}$ levels, with an R^2 value of 0.40, and an RMSE value of 19.00 μg/m^3 (Table S3). Both models exhibited stability during the 10-fold cross-validation test, with similar performance on the training and testing dataset. The R^2 value of the testing dataset was around 0.03 lower than that of the training data, and the RMSE value of the testing dataset was about 4.5% higher than that of the training data (Table S4). This result indicated that the DT automatically selected the important features. In addition, this model performed well in predicting the annual average and 98th percentile daily average $PM_{2.5}$ levels with R^2 values of 0.87 and 0.83, and RMSE values of 3.05 μg/m^3 and 20.10 μg/m^3, respectively (Table S5).

3.2.2. Random forest

Initially, we used VAR1 to construct the RF model. Following this, we tuned the RF model through a grid search that involved two user-defined hyperparameters aimed at improving model performance: the number of trees was set to 500, and the number of predictors at each leaf node was 4. As a result, the RF model exhibited an R^2 of 0.60, and the RMSE value of 8.10 μg/m^3. This model effectively explained the majority of annual average and 98th percentile daily average $PM_{2.5}$ levels (R^2 = 1.00, RMSE = 1.94 μg/m^3 and R^2 = 0.96, RMSE = 15.00 μg/m^3) (Table S5). Furthermore,10-fold cross-validation showed the model performance of this RF model with the training and testing datasets were similar, indicating the absence of overfitting.

Next, we included all the available features we had to the RF model without considering the correlation between independent variables (VAR2). After tuning this RF model, the model performance for the daily average $PM_{2.5}$ levels predictions improved. Specifically, the optimal number of random indicators at each split was 8, and the number of trees was 414. The R^2 value equaled 0.70, and the RMSE value was 7.44 μg/m^3 (Table S3). Moreover, the R^2 and RMSE values of the training and testing dataset in the 10-fold validation test were almost the same using the RF model (Table S4). While the R^2 value for the annual average $PM_{2.5}$ levels predictions using this RF were slightly worse than those using the RF model built with the indicators after feature selection, but RMSE value was slightly better (R^2 = 0.99, and RMSE = 1.90 μg/m^3). The model performance for the 98th percentile daily average $PM_{2.5}$ is similar using these two models. The R^2 value of the RF model with VAR1 is slightly better than that using RF model with VAR2, but the RMSE value is worse (R^2 = 0.95, RMSE = 14.10 μg/m^3) (Table S5). Therefore, the RF model includes more indicators suggesting a better model performance for daily average $PM_{2.5}$ predictions, but the RF model built with the independent variables after feature selection works well with annual average and 98th percentile daily average $PM_{2.5}$ predictions. However, $PM_{2.5}$ predictions using both RF models were biased low.

3.2.3. Support vector regression

We developed two of SVR models, one with the VAR1 and the other using all the available predictors (VAR2). The SVR model performances were very similar using VAR1 before and after tuning (the default hyperparameter settings: cost = 1 and epsilon = 0, and cost = 4 and epsilon = 0.3 after tuning) (R^2 = 0.64 and RMSE = 7.59 μg/m^3 (before tuning); R^2 = 0.68 and RMSE = 7.23 μg/m^3 (after tuning)), with little indication of overfitting through the 10-fold validation. The SVR model explains most of the annual average and 98th percentile daily average $PM_{2.5}$ concentrations with R^2 values of 1.00 and 0.94, and RMSE values of 1.13 μg/m^3 and 10.50 μg/m^3, respectively (Table S5).

The SVR model with more independent variables showed better model performance than the one with the indicators after feature selection, especially after tuning this SVR model to find the best

hyperparameter selection (cost = 2 and epsilon = 0.3; R^2 = 0.81 and RMSE = 5.55 μg/m^3) (Table S3). The predicted annual average and 98th percentile daily average $PM_{2.5}$ concentrations using this SVR model also had a better agreement with the observations (R^2 = 1.00, RMSE = 0.80 μg/m^3 and R^2 = 0.97, RMSE = 6.17 μg/m^3) (Table S5). Thus, the SVR model with more indicators exhibited better model performance when using the optimized value for hyperparameters.

3.2.4. Gaussian process regression

The GPR model uses different kernel functions to estimate covariance between any pair of data points. We multiplied constant kernel and radius basis function kernels to develop Gaussian process regression model (Eqn. (2)) and tuned the hyperparameters (intensity value (r) and variance (σ)) in this study:

$$K(x_1, x_2) = r^2 \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right) \quad \text{(Equation 2)}$$

We conducted a10-fold cross-validation test to assess different combinations of hyperparameters, and we selected r = 19 and σ = 2000 for both VAR1 and VAR2. The model trained by VAR2 had an R^2 of 0.73 and RMSE of 6.53 μg/m^3 for daily $PM_{2.5}$ predictions (Table S3). The 10-fold cross validation for testing data had R^2 = 0.68 and RMSE = 7.06 μg/m^3 (Table 1). The model trained by VAR1 has a lower R^2 value (0.59) and a higher RMSE value (8.10 μg/m^3) compared to the one built with more indicators, in which the 10-fold cross validation for testing data showed less overfitting (R^2 = 0.58 and RMSE = 8.22 μg/m^3) (Table 1, S3 and S4).

Model performance of these two models was virtually the same for annual $PM_{2.5}$ predictions and very similar for the 98th percentile daily $PM_{2.5}$ predictions (Table S5). The R^2 was the same when predicting the peak $PM_{2.5}$ levels, but the RMSE value of the model with more features was much lower than that of the model using the selected variables. We mainly focused on the GPR model built with the variables after feature selection because the one with all the variables may be overfitting.

3.2.5. Neural network

We varied the number of hidden layers from 2 to 5 and the number of nodes in each hidden layer from 40 to 200 (with a step of 10) to tune different neural network (NN) structures using a combination of hyperparameters. For the gradient descent process, we used mean squared error as the loss function, and an Adam (Kingma and Ba, 2014) optimizer was utilized for NN training. After evaluating different structures through cross-validation testing, we developed the final NN models using VAR1 features and VAR2 features with 3 hidden layers and 10 neurons for each hidden layer. To avoid the gradient vanishing when training the neural network, we included normalization layers after each hidden layer. The model trained by all the available indicators (VAR2) had an R^2 of 0.76 and RMSE of 6.11 μg/m^3. The 10-fold cross-validation for testing data had an R^2 of 0.68 and RMSE of 7.07 μg/m^3, indicating little overfitting. The VAR1-based model performed slightly worse than VAR2 (R^2 = 0.57 and RMSE = 8.24 μg/m^3), with little indication of overfitting, which is expected since the feature selection is typically conducted to mitigate overfitting. Additionally, the neural network structure for VAR2 is more complex and has more parameters than the model for VAR1 which can lead to overfitting.

The predicted annual average and peak $PM_{2.5}$ concentrations using the NN models fit observations well (Table S5). The two models (VAR1, VAR2) performed similarly, although the model with more indicators had a higher R^2 and lower RMSE.

Table 1
Summary of statistical results of the daily average $PM_{2.5}$ predictions at the Rubidoux site using different regression models with the testing dataset (10% of the complete dataset).

Method	R^2	RMSE (μg/m^3)
Decision Tree	0.38	10.00
RF	0.64	7.56
SVR	0.78	5.68
GPR	0.68	7.07
NN	0.68	7.07

3.2.6. Variable importance

We used the univariate R^2 value between the daily average $PM_{2.5}$ levels at Rubidoux and each indicator to assess the impact of each independent indicator on predicted $PM_{2.5}$ levels, (Fig. 1). PM precursor emissions, such as NO_x, SO_2, VOCs and NH_3, had the most significant importance on $PM_{2.5}$ levels at the Rubidoux site and had positive contributions, followed by primary $PM_{2.5}$ emissions. Among the meteorological factors, the average surface RH and average RH at 850 mb were the most important indicators, followed by maximum temperature and average wind speed. A high RH from 850 mb to 500 mb can trigger cloud formation and precipitation, which can washout PM (Haby, 2022). The positive correlation between surface RH and $PM_{2.5}$ levels suggests that the increased water content can enhance heterogeneous formation of $PM_{2.5}$, including increased ammonium nitrate formation and faster oxidation of SO_2 and NO_x (Jiang et al., 2019; Sun et al., 2019). Besides the above-mentioned variables, the impact of the remaining variables on the $PM_{2.5}$ formation is small. The wind direction at 500 mb and 850 mb had little influence on $PM_{2.5}$ levels, with an importance value close to 0.

3.3. Comparison among classification models

We followed the EPA guidelines for $PM_{2.5}$ exceedances to train our supervised learning models. We considered daily average exceedances to be when the level of $PM_{2.5}$ is 35 μg/m^3 or greater (Fig. 2). To increase data availability for testing, we relaxed the threshold to 12 μg/m^3 (the annual standard) for exceedances. Using the daily $PM_{2.5}$ standard for machine learning classification may lead to limited model robustness due to the small dataset size, potentially resulting in overfitting or poor generalization to new data. While relaxing the threshold from the daily to the annual standard is necessary to increase data availability, this method might introduce biases or inaccuracies, especially in distinguishing between daily and annual exceedance levels in $PM_{2.5}$

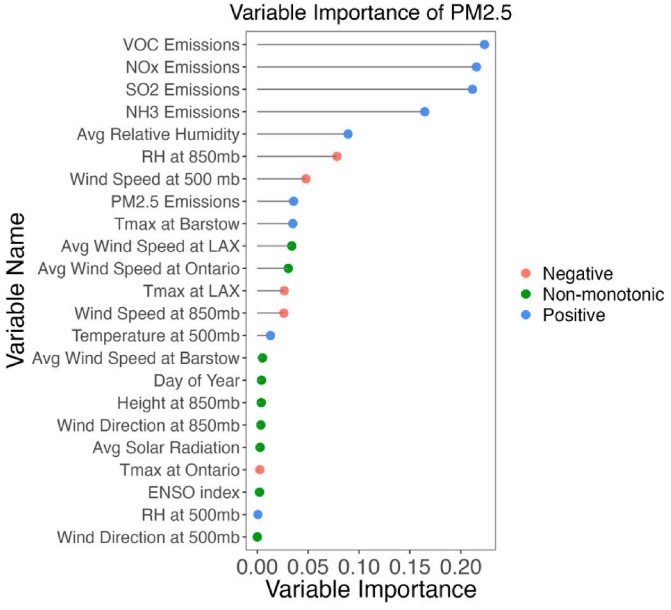


Fig. 1. Univariate R^2 value for 23 features with daily average $PM_{2.5}$ levels at the Rubidoux site. The blue color shows the positive contribution, the red color indicates the negative contribution, and the green color is the non-monotonic relationship.

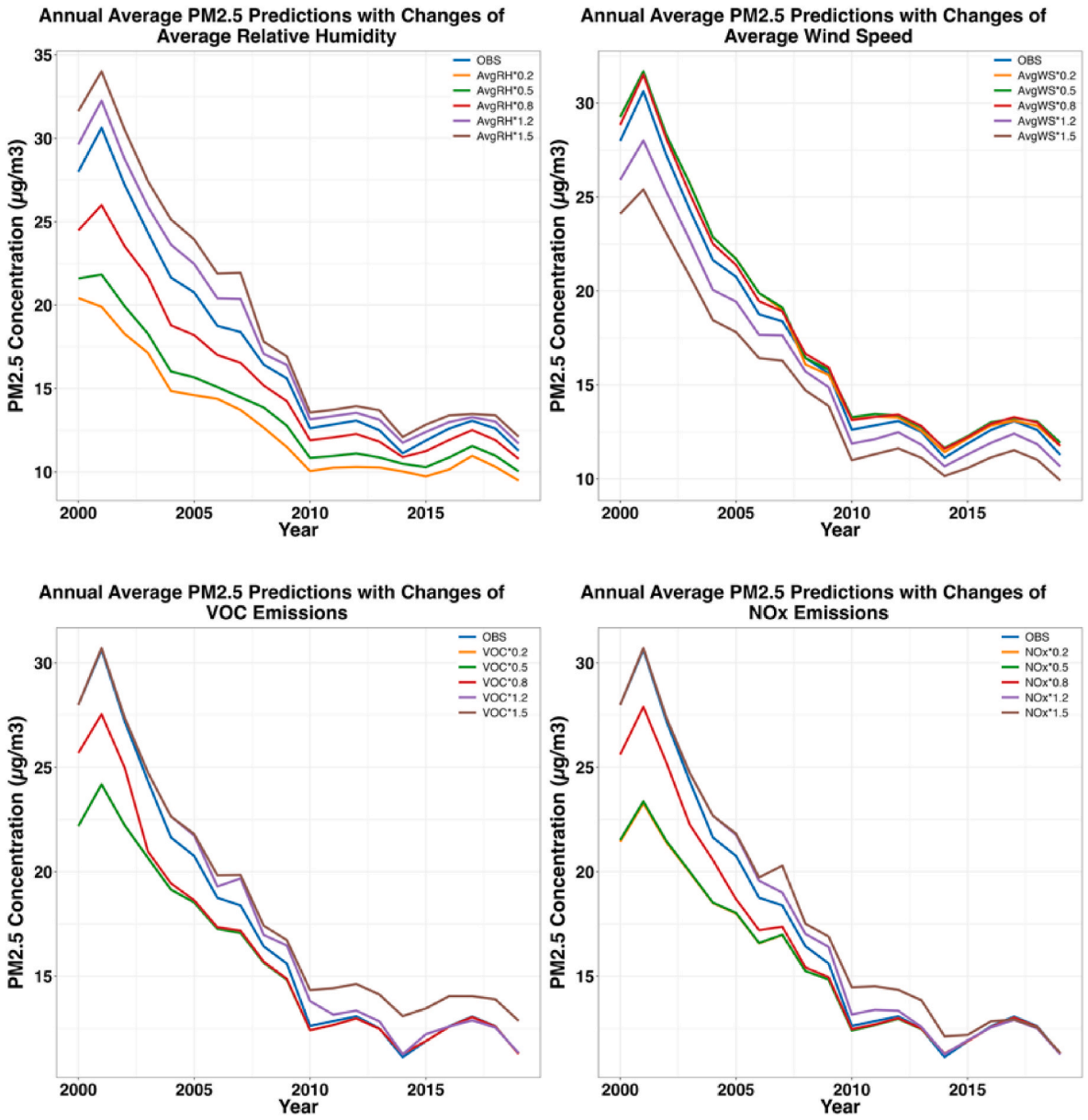


Fig. 2. The observed and counterfactual $PM_{2.5}$ concentrations with the small changes to surface relative humidity, wind speed, VOC and NOx emissions using random forest model. The counterfactual concentrations are computed by (observations-simulations) + predictions (to reduce the uncertainty).

measurements. We labeled $PM_{2.5}$ levels as 0 for non-exceedances ($[PM_{2.5}] < \text{threshold}$) and 1 for exceedances ($[PM_{2.5}] \geq \text{threshold}$). In this study, the total number of predictions was 5465 with 3299 exceedances above 12 µg/m^3 and 489 exceedances above 35 µg/m^3. We combined the confusion matrices (which visualize the actual and predicted values) with the evaluation metrics described in Section 3.1 to assess the models' ability to predict $PM_{2.5}$ exceedances to evaluate the machine learning models' classification performance.

3.3.1. *Decision tree and random forest*

We used the RF model developed above, optimizing the hyperparameters following the method presented in section 3.2.2. To check for overfitting, we subjected the model with the optimized hyperparameters to a 10-fold cross-validation test. The number of predictors chosen at each leaf node was 4 for with a threshold of 12 µg/m^3 and 8 with a threshold of 35 µg/m^3. Also, we applied the decision tree model developed in section 3.2.1.

3.3.2. *Gaussian process classification*

The Gaussian process classification used the same process as the GPR while projecting the regression results from the real number domain (-inf, inf) to the probability domain [0, 1]. We chose the same kernel and hyperparameters used in GPR in section 3.2.4. The uncertainty of this method is evaluated by 10-fold cross-validation.

3.3.3. *Support vector machine*

The primary factor impacting the SVM model's performance is the selection of kernels. There are four common kernels to be chosen from: linear, polynomial, RBF and sigmoid. We developed SVM models with all four kernels (linear, polynomial, sigmoid, and radial) and evaluated their accuracy and precision (Table S2). The RBF kernel had the highest accuracy and precision values, so we opted for the RBF kernel to build the SVM model in this study. Also, we tuned the SVM model to achieve the optimal cost value. We selected a cost value of 4 for the SVM model when the threshold was 12 µg/m^3 and 1 when the threshold was 35 µg/m^3. Moreover, we applied min-max normalization to standardize all the

feature values.

3.3.4. Perceptron and neural network

The perceptron is a type of one-layer neural network that comprises a linear layer and an activation function. The predictions using the perceptron can be used for binary classification by applying a threshold (for instance, values above 0 can be classified as the first class, and those below 0 as the other class). To prevent overfitting, a regularization term in the loss function can be used as a penalty to reduce the weights of unimportant features. In this study, we tuned different penalty methods including the L1 and L2 norm of weights. Surprisingly, the models without any penalty had the best performance on both training and testing datasets. This can be explained by the perceptron's simplicity, which may not capture the non-linear relationships between input and output even without penalties.

The neural network (NN) approach used the softmax function (a normalized exponential function) to map results from the real number domain (-inf, inf) to the probability domain (0, 1). Unlike regression applications, we used mean square error as the loss function to train the NN model for classification. The classifier utilized cross entropy as the loss function, which measured the differences between predicted and observed probability and trained the model to decrease the differences. We built the NN model for classification according to the specifications outlined in Table S7.

3.3.5. k-nearest neighbors

The model performance of k-NN is primarily influenced by the selection of the number of nearest neighbors (k), which generally prefers an odd number and should not exceed the square root of the number of data points. As we have approximately 5500 observations, the largest k value is around 73. We tested all the odd numbers between 1 and 73 and assessed their stability using 10-fold cross-validation. After comparing the accuracy and precision values, we chose k = 11 to build the k-NN model for the threshold = 12 μg/m^3, and k = 21 for the model with the threshold = 35 μg/m^3. Additionally, we applied min-max normalization to scale the values of each indicator.

3.3.6. Diagnostic results

We applied the confusion matrix to summarize the total number of correct and incorrect predictions to assess the performance of the classification models. The subdiagonal of the confusion matrix represents true positive and true negative, which shows the correct predictions for $PM_{2.5}$ exceedances and non-exceedances, while the main diagonal (false negative and false positive) shows the incorrect predictions (Fig. S3 shows the annual average $PM_{2.5}$ exceedances (levels larger than 12 μg/m^3) and Fig. S2 shows the daily average $PM_{2.5}$ exceedances (levels larger than 35 μg/m^3), the SI Table 8 includes a comparison for predictions of annual average exceedances). Similar to predicting annual average concentration trends, the classification models also predicted daily and annual average $PM_{2.5}$ exceedances effectively (Table S6).

In this study, we made a total of 5465 daily predictions, with 3299 exceedances (larger than 12 μg/m^3) and 489 exceedances (higher than 35 μg/m^3). The SVM model has the highest accuracy for predicting the annual average $PM_{2.5}$ exceedances with the most correct predictions for both exceedances (true positives) and non-exceedance predictions (true negatives), and the least incorrect predictions (false positives and false negatives), followed by the neural network, Gaussian process classification, and random forest. Compared to the classification results using a threshold of 12 μg/m^3, the efficacy of these eight classification methods reduces when the threshold is increase to 35 μg/m^3 (Fig. S2). This decline in performance is characterized by most of the values in the confusion matrices in the lower left quadrant (true negatives), and all the methods having fewer predicted $PM_{2.5}$ exceedances than were observed. This is partly due to not capturing very transient emission events such as wildfires. The perceptron had the worst performance for predicting the exceedances in that the number of correct predictions was the least and the most incorrect predictions, although it has the most correct daily average exceedances predictions (which is 354 days).

We used multiple approaches (section 3.1 and Table 2, S8) in addition to the confusion matrix to evaluate the accuracy and precision of our models. The precision value is used to determine whether the labels of the predictions are correct, while the POD value assessed the model's ability to detect the exceedances and ranges from 0 to 1. A high POD indicated that exceedances are correctly, but this value did not consider false negatives in its calculation. Hence, a POD value of 1 may indicate poor performance if the labels all data as exceedances when that may not be the case. The POD value of the perceptron classifier was the lowest, consistent with the result from the confusion matrix (Figs. S2 and S3), which showed that this model's performance was the worst among all the machine learning models.

The F1 score is computed from the harmonic mean of the precision and POD values, serving as a key criterion for comparing model performance. In this study, we combined the POD, precision, accuracy, and F1 score values to do the assessment in order to avoid bias. The SVM model has an overall best performance among all the models (the highest accuracy, the second highest precision, the third highest F1 score, and the 5th POD values), followed by Gaussian process classification and logistic regression (which also has the highest accuracy, even has a better F1 score and POD values, but a lower precision). The decision tree, perceptron and k-NN method has a slightly worse performance for predicting daily $PM_{2.5}$ exceedances (Table 2). We also built these classification models at different sites in Southern California. Although the performance of the SVM, RF, and Logistic models varies by site, the accuracy and precision of most SVM models at various sites are generally the highest. This is followed by the Logistic model and then the Random Forest model, aligning with the performance of different classification models observed at Rubidoux.

3.4. Discussions

3.4.1. Performance of machine learning models for predicting the future $PM_{2.5}$

We applied the built machine learning models to predict $PM_{2.5}$ concentrations with the observed meteorological data and projected emissions in 2020 to further assess the predictive accuracy of each machine learning model. R^2 values ranged between 0.08 (Decision Tree) and 0.37 (GPR). In 2020, the Rubidoux site recorded some extreme $PM_{2.5}$ values, likely due to wildfires. After excluding data from those particular days, the gaussian process regression model was the most accurate in predicting future scenarios among all the models, followed by random forest and support vector regression (Table 3 and Fig. S7). However, the predictions using SVR are closer to the observations compared to RF and GPR based on the slope and RMSE value (Table 3 and Fig. S7). This suggests machine learning models can predict historical values well, but hard to predict the forecasted results accurately. RF and SVR captured some of the complex interactions of factors influencing $PM_{2.5}$ concentrations and are better applied to predict $PM_{2.5}$ levels. However, none of these models accurately predicted the extreme values, suggesting a potential integration with a chemical transport

Table 2
Summary of evaluation matrices of the daily average $PM_{2.5}$ exceedances at the Rubidoux site using different classification models (The threshold is 35 μg/m^3).

Model	Accuracy	Precision	F1 Score	POD	FTP
Decision tree	0.89	0.42	0.44	0.45	0.55
K-NN	0.92	0.63	0.52	0.44	0.56
Gaussian process	0.94	0.76	0.65	0.57	0.43
Logistic	0.94	0.76	0.63	0.53	0.47
Neural network	0.92	0.60	0.59	0.58	0.42
Perceptron	0.91	0.52	0.52	0.51	0.49
Random forest	0.93	0.85	0.53	0.38	0.62
SVM	0.94	0.83	0.60	0.46	0.54

Table 3
Summary of statistical results of the daily average $PM_{2.5}$ predictions at the Rubidoux site using different regression models in 2020.

Method	Year 2020		Year 2020 (exclude extreme $PM_{2.5}$ values)	
	R^2	RMSE (μg/m^3)	R^2	RMSE (μg/m^3)
Decision Tree	0.008	8.47	0.27	5.07
RF	0.36	7.23	0.62	3.81
SVR	0.36	6.91	0.58	3.66
GPR	0.37	6.75	0.60	3.78
NN	0.32	7.06	0.52	4.19

model might be beneficial.

3.4.2. Performance of each machine learning models for sensitivity tests

We applied small changes to the key indicators to evaluate their response to $PM_{2.5}$ levels with RF and SVR, based on the rank of variable importance. Surface RH and maximum temperature were found to positively influence $PM_{2.5}$ levels. A correlation was observed between higher RH at 850 mb, stronger wind speed, and a decline in $PM_{2.5}$ concentrations (Fig. 2, S3 and S4). Most emission variables, particularly NH_3 and VOC emissions, were associated with a positively relationship with $PM_{2.5}$ levels. A higher NH_3 and VOC emissions appear to increase $PM_{2.5}$ concentrations (Fig. 2 and S3). NOx emissions positively affected $PM_{2.5}$ levels in the early years, but from 2015 onwards, a 1.5 multiplier for NOx emissions showed a negative impact on $PM_{2.5}$ (Fig. 2). This could be attributed to NO titration, which reduces ozone formation, impacting the oxidative capacity and subsequently, $PM_{2.5}$ formation.

There is a clearer divergence between various multipliers in 2000s for VOC, NOx, and SO_2 emissions compared to later years, indicating the sensitivity of $PM_{2.5}$ concentrations to these emissions changes (Fig. 2 and S4). The RF model indicated a consistent influence of NH_3 and $PM_{2.5}$ emissions on $PM_{2.5}$ levels, suggesting that even slight changes in these two emissions could predictably affect $PM_{2.5}$ concentrations over time (Fig. S4). However, the sensitivity of NH_3 and $PM_{2.5}$ using SVR showed varying responses similar to other emission variables. There is a variance in the response of each indicator to $PM_{2.5}$ concentrations when using RF and SVR (Fig. S6).

Meteorological factors' sensitivity differences are more uniform and less than those of emission variables (Fig. S6). This may be due to the broader temporal resolution of the emissions and their significance. The sensitivity differences for emission variables showed variations in the early years but stabilized after 2010. This suggested that while both models can capture most of $PM_{2.5}$ concentrations, their sensitivities to specific emission/meteorological variables differ. Such differences suggest the importance of model selection based on the specific research objectives, as the choice of model can influence the response of various factors on $PM_{2.5}$ concentrations. We may need to use the Decoupled Direct Method (DDM) in the chemical transport model, which can efficiently calculate the direct sensitivities of pollutant concentrations to various input parameters during each model simulation step by computing the first-order derivatives, to improve the precision of the sensitivity analyses.

3.5. Limitations

There are several limitations in the development of the models in this study. First, incorporating additional features such as number of detected fire events near the monitor could potentially improve the model's performance. Second, optimizing all the hyperparameters of each machine learning method can be a time-consuming and subjective process, so we only considered the common hyperparameters. The optimal machine learning method for predicting air pollutants may vary depending on the specific pollutant being analyzed. Also, the size of the dataset and the interaction terms can also affect the comparison between machine learning methods, as previous research has shown the relationship between the model performance and the size of the training dataset is not always linear (Bailly et al., 2022). Finally, combining machine learning models with chemical transport models could improve the accuracy of predictions and the sensitivity test.

4. Conclusions

This study developed and compared the machine learning-based models for predicting the $PM_{2.5}$ levels and the number of $PM_{2.5}$ exceedance days at the Rubidoux site in the South Coast Air Basin of California, using precursor emissions, meteorological factors, and climate indices. The statistical results showed that the support vector regression model, with all the available features and the Gaussian process regression model, with the selected features after tuning the hyperparameters worked best for the $PM_{2.5}$ predictions (including daily average, annual average, and the 98th percentile daily average $PM_{2.5}$ levels), although the performance of all methods, except the decision tree model, were relatively similar. The decision tree model had the worst performance in capturing $PM_{2.5}$ levels, although it is less complex and computationally faster than other methods. The support vector regression model, on the other hand, requires fewer computational resources than other complex machine learning methods (e.g., the random forest, and neural network models), with relatively short running time. However, computational time could become more critical with larger datasets. In summary, the support vector regression model had high predictive accuracy and good computational efficiency. The variable importance analysis showed that precursor emissions had a greater impact on $PM_{2.5}$ levels over time than meteorology, though meteorology caused large day-to-day variations. The support vector machine model has the highest accuracy and precision for predicting the number of $PM_{2.5}$ exceedances days, followed by the Gaussian process classification model, neural network, and random forest models.

This work supports the idea that advanced machine learning methods can effectively capture daily and annual $PM_{2.5}$ levels and attribute such concentrations to emissions and meteorological factors (including climate impacts) over time, and can be used to provide daily predictions for health analyses and policy assessment and formulation, including capturing the non-linear responses to further emissions reductions and to assess how $PM_{2.5}$ will respond under different meteorological regimes, including changing climate.

Disclaimer

The views expressed in this paper are those the authors and do not necessarily reflect the views or policies of the Georgia EPD.

CRediT authorship contribution statement

Ziqi Gao: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Khanh Do:** Formal analysis, Methodology, Software, Validation. **Zongrun Li:** Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing. **Xiangyu Jiang:** Writing – original draft. **Kamal J. Maji:** Writing – original draft. **Cesunica E. Ivey:** Conceptualization, Funding acquisition, Supervision, Writing – original draft. **Armistead G. Russell:** Conceptualization, Formal analysis, Software, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This study was supported, in part, by the South Coast Air Quality Management District (#20058), Coordinating Research Council, Health Effects Institute and Phillips 66 company.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2024.120396.

References

Bailly, A., Blanc, C., Francis, É., Guillotin, T., Jamal, F., Wakim, B., Roy, P., 2022. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. Comput. Methods Progr. Biomed. 213, 106504.

Belyaev, M., Burnaev, E.V., Kapushev, Y., 2014. Exact Inference for Gaussian Process Regression in Case of Big Data with the Cartesian Product Structure arXiv: Methodology.

Bi, J., Knowland, K.E., Keller, C.A., Liu, Y., 2022. Combining machine learning and numerical simulation for high-resolution PM2.5 concentration forecast. Environ. Sci. Technol. 56 (3), 1544–1556.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 2017. Classification and Regression Trees. Routledge.

CARB, 2020a. Air Quality and Meteorological Information System (AQMIS). Retrieved from. https://www.arb.ca.gov/aqmis2/metselect.php. (Accessed 27 May 2020).

CARB, 2020b. Air Quality and Meteorological Information System (AQMIS). Retrieved from. https://www.arb.ca.gov/aqmis2/aqdselect.php. (Accessed 27 May 2020).

CARB, 2022. CEPAM: 2019 SIP - standard emission tool emission projections by summary category base year: 2017. Retrieved from. https://ww2.arb.ca.gov/applications/cepam2019v103-standard-emission-tool. (Accessed 21 July 2022).

Chang, C.-C., Lin, C.-J., 2011. LIBSVM. ACM Transactions on Intelligent Systems and Technology 2 (3), 1–27.

Chen, Z., Chen, D., Zhao, C., Kwan, M.-p., Cai, J., Zhuang, Y., Zhao, B., Wang, X., Chen, B., Yang, J., Li, R., He, B., Gao, B., Wang, K., Xu, B., 2020. Influence of meteorological conditions on PM2.5 concentrations across China: a review of methodology and mechanism. Environ. Int. 139, 105558.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theor. 13 (1), 21–27.

CPC, 2020. El niño-southern oscillation (ENSO). Retrieved from. https://www.cpc.ncep.noaa.gov/products/precip/CWlink/MJO/enso.shtml#references. (Accessed 27 May 2020).

Cramer, J., 2002. The Origins of Logistic Regression. Retrieved from. https://EconPapers.repec.org/RePEc:tin:wpaper:20020119.

Dockery, D.W., Pope, C.A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G., Speizer, F.E., 1993. An association between air pollution and mortality in six U.S. Cities. N. Engl. J. Med. 329 (24), 1753–1759.

Dudoit, S., Fridlyand, J., Speed, T.P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc. 97 (457), 77–87.

Fan, R.-E., Chen, P.-H., Lin, C.-J., 2005. Working set selection using second order information for training support vector machines. J. Mach. Learn. Res. 6, 1889–1918.

Gao, Z., Ivey, C.E., Blanchard, C.L., Do, K., Lee, S.-M., Russell, A.G., 2023a. Emissions and meteorological impacts on PM2.5 species concentrations in Southern California using generalized additive modeling. Sci. Total Environ. 891, 164464.

Gao, Z., Ivey, C.E., Blanchard, C.L., Do, K., Lee, S.-M., Russell, A.G., 2023b. Emissions, meteorological and climate impacts on PM2.5 levels in Southern California using a generalized additive model: historic trends and future estimates. Chemosphere, 138385.

Gao, Z., Wang, Y., Vasilakos, P., Ivey, C.E., Do, K., Russell, A.G., 2022. Predicting peak daily maximum 8 h ozone and linkages to emissions and meteorology in Southern California using machine learning methods (SoCAB-8HR V1.0). Geosci. Model Dev. (GMD) 15 (24), 9015–9029.

Gupta, P., Zhan, S., Mishra, V., Aekakkararungroj, A., Markert, A., Paibong, S., Chishtie, F., 2021. Machine learning algorithm for estimating surface PM2.5 in Thailand. Aerosol Air Qual. Res. 21 (11), 210105.

Gurgueira, S.A., Lawrence, J., Coull, B., Murthy, G.K., González-Flecha, B., 2002. Rapid increases in the steady-state concentration of reactive oxygen species in the lungs and heart after particulate air pollution inhalation. Environ. Health Perspect. 110 (8), 749–755.

Haby, J., 2022. THE RELATIVE HUMIDITY PROG. Retrieved from. http://www.theweatherprediction.com/habyhints/105/.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, vol. 2. Springer.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Paper Presented at the Proceedings of the IEEE International Conference on Computer Vision.

Jaskowiak, P.A., Campello, R.J.G.B., Costa, I.G., 2012. Evaluating Correlation Coefficients for Clustering Gene Expression Profiles of Cancer. Springer Berlin Heidelberg, pp. 120–131.

Jiang, F., Liu, F., Lin, Q., Fu, Y., Yang, Y., Peng, L., Lian, X., Zhang, G., Bi, X., Wang, X., Sheng, G., 2019. Characteristics and formation mechanisms of sulfate and nitrate in size-segregated atmospheric particles from urban guangzhou, China. Aerosol Air Qual. Res. 19 (6), 1284–1293.

Jiang, X., Yoo, E.-h., 2018. The importance of spatial resolutions of Community Multiscale Air Quality (CMAQ) models on health impact assessment. Sci. Total Environ. 627, 1528–1543.

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization arXiv preprint arXiv:1412.6980.

Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., Rybarczyk, Y., 2017. Modeling PM2.5 urban pollution using machine learning and selected meteorological parameters. Journal of Electrical and Computer Engineering 1–14, 2017.

Kumar, S., Mishra, S., Singh, S.K., 2020. A machine learning-based model to estimate PM2.5 concentration levels in Delhi's atmosphere. Heliyon 6 (11), e05618.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R. News 2 (3), 18–22.

Loh, W.Y., 2011. Classification and regression trees. Wiley interdisciplinary reviews: Data Min. Knowl. Discov. 1 (1), 14–23.

Minh, V.T.T., Tin, T.T., Hien, T.T., 2021. PM2.5 forecast System by using machine learning and WRF model, A case study: Ho chi Minh city, vietnam. Aerosol Air Qual. Res. 21 (12), 210108.

NCEI, 2020. Climate Data Online. Retrieved from. https://www.ncdc.noaa.gov/cdo-web/datasets. (Accessed 27 May 2020).

Pinault, L., Tjepkema, M., Crouse, D.L., Weichenthal, S., Van Donkelaar, A., Martin, R.V., Brauer, M., Chen, H., Burnett, R.T., 2016. Risk estimates of mortality attributed to low concentrations of ambient fine particulate matter in the Canadian community health survey cohort. Environ. Health 15 (1).

Pope III, C., Burnett, R., Thun, M., Calle, E., Krewski, D., Ito, K., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA 287 (9), 1132–1141. Find this article online.

Pope, P.T., Webster, J.T., 1972. The use of an F-statistic in stepwise regression procedures. Technometrics 14, 327–340.

Quinlan, J.R., 2014. C4. 5: Programs for Machine Learning. Elsevier.

Rasmussen, C.E., Williams, C.K., 2006. Gaussian Processes for Machine Learning, vol. 1. Springer.

Ripley, B.D., 2007. Pattern Recognition and Neural Networks. Cambridge university press.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. 65 (6), 386–408.

Rybarczyk, Y.P., Zalakeviciute, R., 2022. Editorial: statistical learning for predicting air quality. Frontiers in Big Data 5.

Schwartz, J., 1994. Air pollution and daily mortality: a review and meta analysis. Environ. Res. 64 (1), 36–52.

Sun, L., Xue, L., Wang, Y., Li, L., Lin, J., Ni, R., Yan, Y., Chen, L., Li, J., Zhang, Q., Wang, W., 2019. Impacts of meteorology and emissions on summertime surface ozone increases over central eastern China between 2003 and 2015. Atmos. Chem. Phys. 19 (3), 1455–1469.

Tin Kam, H., 1995. Proceedings of 3rd International Conference on Document Analysis and Recognition, pp. 14–16. Aug. 1995). Random decision forests. Paper presented at the.

Venables, W.N., Ripley, B.D., 2013. Modern Applied Statistics with S-PLUS. Springer Science & Business Media.

Vlachogianni, A., Kassomenos, P., Karppinen, A., Karakitsios, S., Kukkonen, J., 2011. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki. Sci. Total Environ. 409 (8), 1559–1571.

Xu, M., Jin, J., Wang, G., Segers, A., Deng, T., Lin, H.X., 2021. Machine learning based bias correction for numerical chemical transport models. Atmos. Environ. 248, 118022.

Zhang, N., Xiong, J., Zhong, J., Leatham, K., 2018. Gaussian Process Regression Method for Classification for High-Dimensional Data with Limited Samples, 2018.