Using Machine Learning to Predict Air Quality Index in New Delhi

Samayan Bhattacharya Department of Computer Science and Engineering Jadavpur University samayan.bhattacharya@gmail.com Sk Shahnawaz
Department of Computer Science and Engineering
Jadavpur University
skshahnawaz2909@gmail.com

Abstract—Air quality has a significant impact on human health. Degradation in air quality leads to a wide range of health issues, especially in children. The ability to predict air quality enables the government and other concerned organizations to take necessary steps to shield the most vulnerable, from being exposed to the air with hazardous quality. Traditional approaches to this task have very limited success because of a lack of access of such methods to sufficient longitudinal data. In this paper, we use a Support Vector Regression (SVR) model to forecast the levels of various pollutants and the air quality index, using archive pollution data made publicly available by Central Pollution Control Board and the US Embassy in New Delhi. Among the tested methods, a Radial Basis Function (RBF) kernel produced the best results with SVR. According to our experiments, using the whole range of available variables produced better results than using features selected by principal component analysis. The model predicts levels of various pollutants, like, sulfur dioxide, carbon monoxide, nitrogen dioxide, particulate matter 2.5, and ground-level ozone, as well as the Air Quality Index (AQI), at an accuracy of 93.4 percent.

Keywords—air quality index, support vector regression, radial basis function

I. INTRODUCTION

The sharp rise in air pollution in recent years, due to industrial and agricultural activities, as well as increased number of vehicles using internal combustion engines, has caught the attention of the scientific community [1, 2, 3]. Air pollution has significant impact on human health and may cause long-term health issues in children. The significant rise in air pollution in New Delhi is attributed to increased vehicular emissions, burning of fossil fuels at power plants, and other local industries and burning of fields by farmers in neighboring states [4].

Air quality is being monitored in New Delhi for about two decades. This has allowed a better understanding of the changes in air pollution in response to particular activities and government regulations, but the air pollution in New Delhi remains a problem [5].

Air pollution is responsible for 30 percent of lower-respiratory tract infections and is linked with 91 percent of premature deaths, from lung cancer, heart disease, acute respiratory infections, stroke and chronic obstructive pulmonary disease. It contributes to 20 percent of infant mortality worldwide and causes numerous short- and long-term illnesses in children. Exposure of the mother to high levels of air pollution can lead to adversely affect immune status, brain development, respiratory systems, and cardiometabolic health of the child. Air pollution has also been linked to low birth weight and stunted growth in children.

Air pollution is estimated to be responsible for one in ten deaths of children under five years of age. In elder people, air pollution causes high rates of asthma, with decreased cognitive performance.

Presently, the government implements regulations after the air quality reaches hazardous levels. If there is a way to foresee the air quality reaching hazardous levels, the government can implement such regulations early, potentially preventing further degradation of air quality and being able to shield those, most vulnerable, from getting exposed to such air quality. This study aims to build a model that can look at previously recorded air quality data and predicts levels of different pollutants as well as air quality index. For this we use a variation of Support Vector Machines (SVM), called Support Vector Regression (SVR).

The paper is organized as follows. We state the motivations of this work and frame our work in section 2, stating the potential impact of being able to successfully predict the air quality. We provide a critical revision of related work, done previously, in Section 3. We explain how Support Vector Machines (SVM), particularly Support Vector Regression works in Section 4. We describe the datasets used in this work and the data preprocessing steps used to produce a more efficient input for the SVR, in sections 5 and 6 respectively. In Section 7, we present details of the experiment performed, divided into subsections describing the experimental setup and the results obtained. In Section 8, we conclude the paper and discuss ideas for future work in this area.

II. BACKGROUND AND MOTIVATION

Air pollution is caused by the introduction of harmful or excessive quantities of certain substances into the atmosphere. Such substances include solid particles, liquid droplets and gases. Air pollutants are classifieds into primary and secondary pollutants. Primary pollutants are the pollutants that are directly released from their source directly into the atmosphere. Sources of primary air pollutants may be natural, like volcanic eruptions, sand storms, etc. or man-made, like burning of fossil fuels, leaking gases from appliances, etc. Primary pollutants include sulfur dioxide (SO₂), oxides of nitrogen (NO_x), particulate matter (PM) and carbon monoxide (CO). Secondary pollutants are formed in the atmosphere due to chemical or physical interactions between primary pollutants. Secondary pollutants include photochemical oxidants and secondary particulate matter.

The most common pollutants are called criteria pollutants and correspond to the most prevalent health threats. These include SO₂, ground level ozone (O₃), NO₂, lead and PM. It

has been demonstrated that there is a correlation between exposure to such pollutants for a short while and health issue like inflamed respiratory track in healthy people, increased respiratory symptoms in people with asthma, difficulty meeting high oxygen requirements while exercising and critical respiratory situations, especially in children and the elderly [6].

National agencies like EPA, EU, and many others have set standards for acceptable levels of air quality. Air quality index (AQI) is used to indicate the levels of the criteria pollutants in the air. The overall AQI is the maximum of the AQI recorded for the individual criteria pollutants. AQI levels also indicate the health risks associated with exposure to the particular air quality. Such health symptoms may be experienced shortly after exposure to polluted air or in the long run. Such symptoms may also vary based on the age and health conditions of the particular person being exposed.

Thus, it is vital that we have a system to forecast increases in air pollution levels, so that government organizations may be able to counter further increase through on-demand pollution control mechanisms or an emergency response [7]. This would make AQI more controllable to suit the overall needs of the population. If the rise in pollution cannot be curbed, the authorities may issue warning to the population about the AQI forecast in order to shield the most vulnerable from getting exposed in case the AQI forecast exceeds the permitted limit.

III. PREVIOUS AND RELATED WORK

The best statistical method for predicting time series data is the autoregressive integrated moving average model (ARIMA) [8]. It has several advantages in terms of its statistical properties [9], potential for a wide range of applications and extendibility. With the rise in importance for predicting air quality levels, ARIMA was applied to this task as well. It was demonstrated to reach accuracies around 95% [10] for forecasting AQI monthly values. [11] compared the performance of ARIMA with a Holt exponential smoothing model and proved the superiority of the ARIMA model for forecasting AQI daily values. However, this method requires extensive manual intervention in terms of selecting the data fed into the system, as it has a low tolerance towards outliers. The features to be considered must also be selected.

The availability of large quantities of archive data made it convenient to use Machine Learning (ML) [12] models for the time series prediction of AQI. ML models are able to automatically look at large amounts of data and select important features, thus reducing the need for human intervention. ML models are able to achieve higher accuracies with large datasets, than classic statistical methods. Such models have long been used for AQI forecasting tasks.

ML models are nonlinear, nonparametric in nature and hence are better able to handle the complexity of nonlinear elements like pollutant levels in the air [13]. Hence they outperform statistical methods like ARIMA, Winter exponential smoothing, and multivariate regression, which work well only with linear systems [14, 15].ML models like Artificial Neural Networks (ANN), Genetic programming and Support Vector Machine (SVM) are able to find hidden patterns in vast quantities of data [7].

Several works have used Support Vector Machines (SVM) for predicting time series data. Some works have also used

SVM for forecasting air quality. [16] proposed the use a variant of SVM for regression tasks and called it Support Vector Regression (SVR). [17] proved the superiority of SVR over Artificial Neural Networks (ANN). [18] showed that a hybrid of ANN and SVM produced better results. They used ANN for partitioning the input space and the SVM to model each portion.

Some of the works that used SVM for time series air quality forecasting include, (1) the model for the prediction of air quality in downtown Hong Kong by [19], showing that SVMs perform better than other Machine Learning approaches, (2) SVM model by [20] for air quality (PM10) forecasting in Bangkok, (3) SVM for air quality forecasting in Macau by [21] (4) work of [22] for forecasting in Mexico City, that lead to the conclusion that SVMs are more scalable and flexible for nonlinear, dynamic data, (5) the hybrid model proposed by [23], combing the advantages of SVM and flower pollination algorithm, which was shown to outperform any particular model.

IV. SUPPORT VECTOR MACHINE

Support Vector Machines (SVM) were introduced by [25] as a classification technique. The objective is to use the hyperplane to separate the data, represented as support vectors, belonging to the different classes. When the original data is not linearly separable, it is projected to a higher dimension, using a kernel function. This makes the nonlinearly separable data linearly separable.

Support Vector Regression (SVR) was introduced by [24]. This allowed Support Vector Machines to be applied to regression using a new loss function. SVR has been used for time series forecasting by several works [16, 17, 26]. It has been demonstrated that SVR models offer faster training and better forecast ability while using smaller number of parameters.

The objective of SVR model is to learn a nonlinear mapping $\phi_i(X)$ of the data to a high dimensional vector space such that the projections can be used to train a linear regression model. The trained linear regression model is then used to forecast in the high dimensional space after mapping the input to the high dimensional space using the kernel function.

The SVR model uses a combination of the training error and a regularization term in the loss function. Apart from this, other interesting properties arise from the use of kernel function, enabling it to be used for both linear and nonlinear forecasting and the convex nature of the fitness function and its constraints.

Let, the training set with m data points be represented as

$$T = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), \dots (x_m, y_m)\}$$
 (1)

where, $x \in X \subset \mathbb{R}^n$ are the inputs in the training set and $y \in Y \subset \mathbb{R}$ are the corresponding expected outputs in the training set.

A nonlinear kernel function is represented as

$$f(x) = \omega^{\mathrm{T}} \Phi(x_{i}) + b \tag{2}$$

Equation (2) can be written in the form of a constrained convex optimization problem as:

minimize
$$\frac{1}{2}\omega^T\omega$$

(3)

subject to

$$\int y_i - \omega^T \Phi(x_i) - b \le \varepsilon$$
$$\omega^T \Phi(x_i) + b - y_i \le \varepsilon$$

The aim of this objective function is to minimize ω , while satisfying the other constraints, under the assumption that convex optimization problem is feasible, that is, f(x) exists. In case this assumption is not true, errors can be traded off for the flatness of the estimate. Thus, (3) can be reformulated as (4).

minimize
$$\frac{1}{2}\omega^T\omega + C\sum_{i=1}^m (\epsilon_i^+ + \epsilon_i^-)$$

(4)

subject to

$$\begin{cases} y_i - \omega^T \Phi(\mathbf{x}_i) - \mathbf{b} \le \varepsilon + \epsilon_i^+ \\ \omega^T \Phi(\mathbf{x}_i) + \mathbf{b} - y_i \le \varepsilon + \epsilon_i^- \\ \epsilon_i^+ \epsilon_i^- \ge 0 \end{cases}$$

where C<0 represents weights of the loss function and is an initialized constant. $\omega^T \omega$ is the regularized term and $C \sum_{i=1}^m (\epsilon_i^+ + \epsilon_i^-)$ is the empirical term, measuring the ε -insensitive loss function.

While solving (4) Lagragian multipliers $(\alpha_i^+, \eta_i^+, \alpha_i^-, \eta_i^-)$ may be used to eliminate some of the primal variables. The final equation translating the dual optimization problem is (5).

minimize
$$\frac{\frac{1}{2}\sum_{i,j=1}^{m}K(x_{i},x_{j})(\alpha_{i}^{+}-\alpha_{i}^{-})(\alpha_{j}^{+}-\alpha_{j}^{-})+}{\varepsilon\sum_{i=1}^{m}(\alpha_{i}^{+}+\alpha_{i}^{-})-\sum_{i=1}^{m}(\alpha_{i}^{+}-\alpha_{i}^{-})}$$
subject to
$$\sum_{i=1}^{m}(\alpha_{i}^{+}-\alpha_{i}^{-})=0$$

$$\alpha_{i}^{+},\alpha_{i}^{-}\in[0,C]$$

where $K(x_i,x_j)$ represents the kernel function, allowing the application of SVR to nonlinear functions. The performance of the SVR model depends on the kernel function, the regularization parameter (C) and the insensitive parameter (ϵ). There are many options for the kernel function [27]. In this work we study radial basis function (RBF) and polynomial kernel.

V. DATA DESCRIPTION

The datasets used in this study were obtained from the archive data provided by the US Embassy in Delhi and the Central Pollution Control Board. These datasets are described below.

The data from the US Mission in India consists of hourly concentrations of particulate matter of sizes less than or equal to 2.5 microns (PM2.5) and particulate matter of diameter less than or equal to 10 microns (PM10). They derive the Air Quality Index based on these values. The values are recorded using device located on the campus of the embassy and are thus, highly local and different from those recorded by the Central Pollution Control Board.

The data from Central Pollution Control Board are daily recordings of the concentrations of Sulfur dioxide (SO₂), Nitrogen dioxide (NO₂), PM2.5, PM10, and suspended particulate matter (SPM). The data is of relatively poor quality with significant amount of missing values. The recordings are from 4 recording stations in different parts of Delhi but are reasonably different from one another. Since the data is recorded daily instead of hourly, the continuity of the time series representation of the data is adversely affected and must be handled in the pre-processing step.

VI. DATA PREPROCESSING

Data quality and effective data representation are of paramount importance in ensuring good performance of a forecasting model and its generalizability [28] of the SPM data. The standard data processing steps include (1) preparing more accurate and complete datasets by imputing missing data and removing or modifying outlier data points, (2) ensuring data is uniformly distributed by normalization and standardization of data, (3) creating a smaller and compact dataset by extraction and selection of features. We perform these steps on our data as follows.

Imputation of missing data: We found that more than 50% data was missing, so we removed the SPM field from the Central Pollution Control Board data. For the other fields in the US embassy data and the Central Pollution Control Board data, we substituted missing data by second order polynomial estimation using nearest available data points. It gave better results than using series mean or linear interpolation.

Removing or modifying outliers: An irregular pattern was observed in pollutant data between August and October 2020 in both US embassy data and the Central Pollution Control Board data. Thus, these data were removed. For data modification, we used the power transformation method [29]. This provides a nonlinear transformation that is more robust to noise and hence produces better data.

Feature extraction: The date component in the Central Pollution Control Board data and the date-time component in the US embassy data were used to produce new features. The date component was used to obtain a field called seasons. Four seasons were used (Summer, Fall, Winter, Spring). The cyclic nature of the time component was exploited to obtain two fields $\{\sin(2\pi hour/24), \cos(2\pi hour/24)\}$. The date component was also used to obtained fields for day, month and year.

Feature selection: From the features obtained in the previous step using feature engineering, a few variables were selected to reduce dimensionality of the dataset and remove collinearity. Correlation-based feature selection was used [30], to check for collinearity among features. It was observed that the concentration of some of the pollutants had an almost linear correlation. For example, NO2 and CO concentrations were almost linearly related and so were CO and PM2.5 concentrations. Based on the remarkable correlation [31], it was decided to keep all pollutants in the dataset. In spite of SVR models being robust against collinearity and multicollinearity [32], we dropped some variables showing strong correlation n with some other variable. For example, the season variable had strong correlation with the month variable, hence we dropped the month variable, also, the hour variable was dropped due to strong correlation with the hour sin and hour cos cyclic variables. We also used Principal Component Analysis (PCA) [33] for reducing the dimensionality of the dataset. It enabled us to reduce the

number of variables for each pollutant by about 76%. We compare the results with and without PCA in the next section.

VII. EXPERIMENTAL STUDY

Experimental settings: There are three user defined hyperparameters in an SVR model, the maximum allowed deviation ϵ , the regularization constant C, and the kernel type function. For determining C and ϵ , time-series split, combined with random grid search was employed [34, 35]. The range of C was extended to be 1 to 100, to allow wider exploration, as opposed to the 10 to 100 range [26]. The range of ϵ was taken to be between 0.001 and 0.1, with a step of 0.001. The most popular kernel functions being RBF and polynomial, we compare results for both of them. The optimum number of iterations for the random search was taken to be 60 [36].

Experimental results: Here we discuss the performance of the SVR models in forecasting the levels of 4 pollutants, nitrogen dioxide (NO₂), sulfur dioxide (SO₂), particulate matter 2.5 (PM_{2.5}) and particulate matter 10 (PM₁₀) with and without Principal Component Analysis (PCA).

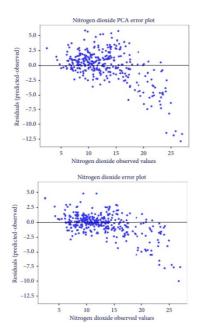
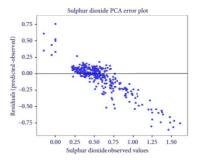


Figure 1: (a) PCA SVR-RBF forecasting errors plotted against observed NO2 values and (b) SVR-RBF forecasting errors plotted against observed NO2 values

TABLE I. ERROR METRICS OF THE FORECASTING MODEL FOR NO_2 LEVEL DETECTION

Error	PCA S	VR-RBF	SVR-RBF		
Metrics	Training Set	Validation Set	Training Set	Validation Set	
MAE	0.235	0.460	0.228	0.413	
\mathbb{R}^2	0.788	0.024	0.831	0.272	
RMSE	0.363	0.753	0.351	0.702	
nRMSE	0.037	0.054	0.034	0.048	



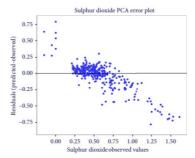
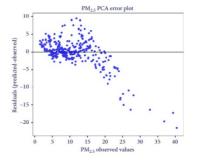


Figure 2: (a) PCA SVR-RBF forecasting errors plotted against observed SO2 values and (b) SVR-RBF forecasting errors plotted against observed SO2 values.

TABLE II. ERROR METRICS OF THE FORECASTING MODEL FOR SO₂
LEVEL DETECTION

Error	PCA SV	R-RBF	SVR-RBF		
Metrics	Training Set	Validatio n Set	Training Set	Validation Set	
MAE	0.105	0.228	0.094	0.161	
R ²	0.974	0.884	0.980	0.938	
RMSE	0.151	0.315	0.133	0.237	
nRMSE	0.028	0.050	0.024	0.037	



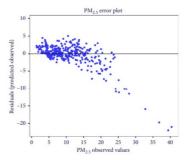


Figure 3: (a) PCA SVR-RBF forecasting errors plotted against observed PM2.5 values and (b) SVR-RBF forecasting errors plotted against observed PM2.5 values.

TABLE III. Error Metrics of the forecasting model for $PM_{2.5}$ Level detection

Error	PCA S	VR-RBF	SVR-RBF		
Metrics	Training Set	Validation Set	Training Set	Validation Set	
MAE	0.201	0.381	0.145	0.330	
\mathbb{R}^2	0.881	0.562	0.936	0.646	
RMSE	0.274	0.577	0.204	0.511	
nRMSE	0.511	0.073	0.030	0.067	

0.2 -	: .*		Ö.		
-0.2 -					n <u>.</u>
-0.2 -			Congress of		Ċ.
-0.4 -					·
-0.6 -	0.2	0.4	0.6	0.8	1.0

TABLE IV. Error Metrics of the forecasting model for PM_{10} level detection

Error	PCA S	VR-RBF	SVR-RBF		
Metrics	Training Set	Validation Set	Training Set	Validation Set	
MAE	0.084	0.191	0.040	0.089	
\mathbb{R}^2	0.988	0.922	0.997	0.981	
RMSE	0.111	0.261	0.061	0.134	
nRMSE	0.024	0.051	0.012	0.024	

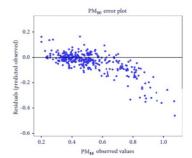


Figure 4: (a) PCA SVR-RBF forecasting errors plotted against observed PM10 values and (b) SVR-RBF forecasting errors plotted against observed PM10 values.

Air quality index value range	Levels of health concern	Description			
0 to 50	Good	Air quality is considered satisfactory.			
51 to 100	Moderate	Air quality is acceptable; however, for some pollutants, there is a more health concern for a small number of people, namely, those that experiments respiratory problems.			
101 to 150	Unhealthy for sensitive groups	Although for most of the people, the health concern is moderate, for groups with lung diseases, the elderly, and children, there is a great risk of exposure to some pollutants and particulates.			
151 to 200	Unhealthy	Health side effects for all the affected area population. Sensitive groups may experience more serious effects.			
201 to 300	Very unhealthy	Health alerts would be triggered as all the affected area population would experience serious health effects.			
301 to 500	Hazardous	Health alerts with emergency warnings would be triggered. The entire area population would be severely affected.			

Figure 5: Six AQI categories defined by EPA

TABLE V. CONFUSION MATRIX FOR THE AQI CLASSIFICATIONS OBTAINED WITH BOTH MODELS FOR THE TRAINING SET

Training	PCA SVR-RBF			SVR-RBF		
Dataset	Good	Moderate	Unhealthy	Good	Moderate	Unhealthy
Good	2509	159	0	2547	106	0
Moderate	306	992	0	252	1045	0
Unhealthy	6	24	0	9	19	0

TABLE VI. CONFUSION MATRIX FOR THE AQI CLASSIFICATIONS OBTAINED WITH BOTH MODELS FOR THE VALIDATION SET

Training	PCA SVR-RBF			SVR-RBF		
Dataset	Good	Moderate	Unhealthy	Good	Moderate	Unhealthy
Good	1166	54	0	1179	45	0
Moderate	61	425	0	49	439	0
Unhealthy	0	2	0	9	3	0

VIII. CONCLUSION

The task of forecasting pollutant levels is inherently hard because of the volatile and dynamic nature of the data and its variability in space and time. However, the task of forecasting pollutant levels has been increasing in importance due to the effects of pollution on the population and the environment. In this work we used SVR for forecasting levels of pollutants like NO₂, SO₂, PM_{2.5} and PM₁₀, and Air Quality Index (AQI), using publicly available data for New Delhi.

REFERENCES

- [1] U. A. Hvidtfeldt, M. Ketzel, M. Sørensen, O. Hertel, J. Khan, J. Brandt, and O. Raaschou-Nielsen, "Evaluation of the danish airgis air pollution modeling system against measured concentrations of pm2.5, pm10, and black carbon," Environmental Epidemiology, vol. 2, no. 2, p. e014, 2018.
- [2] Y. Gonzalez, C. Carranza, M. Iniguez, M. Torres, R. Quintana, A. R. Osornio-Vargas, C. Gardner, S. Sarkar, and S. Schwander, "Inhaled air pollution particulate matter in alveolar macrophages alters local pro-inflammatory cytokine and peripheral ifn production in response to mycobacterium tuberculosis," in B17. MYCOBACTERIAL HOST DEFENSES. American Thoracic Society, 2017, pp. A2901–A2901.
- [3] L. Pimpin, L. Retat, D. Fecht, L. de Preux, F. Sassi, J. Gulliver, A. Belloni, B. Ferguson, E. Corbould, A. Jaccard et al., "Estimating the costs of air pollution to the national health service and social care: An assessment and forecast up to 2035," PLoS medicine, vol. 15, no. 7, p. e1002602, 2018.
- [4] V. Kanawade, A. Srivastava, K. Ram, E. Asmi, V. Vakkari, V. Soni, V. Varaprasad, and C. Sarangi, "What caused severe air pollution episode of november 2016 in new delhi?" Atmospheric Environment, vol.222, p. 117125, 2020.
- [5] B. R. Gurjar, L. T. Molina, and C. S. P. Ojha, Air pollution: health and environmental impacts. CRC press, 2010.
- [6] D. C. Payne-Sturges, M. A. Marty, F. Perera, M. D. Miller, M. Swanson, K. Ellickson, D. A. Cory-Slechta, B. Ritz, J. Balmes, L. Anderko etal., "Healthy air, healthy brains: advancing air pollution policy to protect children's health," American journal of public health, vol. 109, no. 4, pp. 550–554, 2019.
- [7] M. Castelli, F. M. Clemente, A. Popovi'c, S. Silva, and L. Vanneschi, "A machine learning approach to predict air quality in california," Complexity, vol. 2020, 2020.
- [8] G. E. Box and D. A. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models,"

As the next step, we would like to investigate and compare the performance of other Machine Learning methods like Artificial Neural Network (ANN) and genetic algorithms, for this task. We would also like to explore the use other methods of hyperparameter optimization and other methods of variable selection for larger datasets.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of the Central Pollution Control Board (CBCP) and the Government of Delhi in collecting the data for this project

- Journal of the American statistical Association, vol. 65, no. 332, pp. 1509-1526, 1970.
- [9] C.-L. Hor, S. J. Watson, and S. Majithia, "Daily load forecasting and maximum demand estimation using arima and garch," in 2006 International Conference on Probabilistic Methods Applied to Power Systems. IEEE, 2006, pp. 1–6.
- [10] L. Y. Siew, L. Y. Chin, and P. M. J. Wee, "Arima and integrated arfima models for forecasting air pollution index in shah alam, selangor," Malaysian Journal of Analytical Sciences, vol. 12, no. 1, pp. 257–263, 2008
- [11] J. Zhu, R. Zhang, B. Fu, and R. Jin, "Comparison of arima model and exponential smoothing model on 2014 air quality index in yanqing county, beijing, china," Appl. Comput. Math, vol. 4, pp. 456–461, 2015.
- [12] I. Goodfellow, Y. Bengio, and A. Courville, "Machine learning basics," Deep learning, vol. 1, no. 7, pp. 98–164, 2016.
- [13] U. Brunelli, V. Piazza, L. Pignato, F. Sorbello, and S. Vitabile, "Three hours ahead prevision of so2 pollutant concentration using an elman neural based forecaster," Building and Environment, vol. 43, no. 3, pp. 304–314, 2008.
- [14] R. Sharda and R. Patil, "Neural networks as forecasting experts: an empirical test," in Proceedings of the International Joint Conference on Neural Networks, vol. 2. IEEE, 1990, pp. 491–494.
- [15] I. Alon, M. Qi, and R. J. Sadowski, "Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods," Journal of retailing and consumer services, vol. 8, no. 3, pp. 147–156, 2001.
- [16] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, V. Vapnik et al., "Support vector regression machines," Advances in neural information processing systems, vol. 9, pp. 155–161, 1997.
- [17] K.-R. M "uller, A. J. Smola, G. R atsch, B. Sch "olkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in International Conference on Artificial Neural Networks. Springer, 1997, pp. 999–1004.

- [18] L. Cao, "Support vector machines experts for time series forecasting," Neurocomputing, vol. 51, pp. 321–339, 2003.
- [19] W.-Z. Lu and W.-J. Wang, "Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends," chemosphere, vol. 59, no. 5, pp. 693–701, 2005.
- [20] S. Arampongsanuwat and P. Meesad, "Prediction of pm10 using support vector regression," in International Conference on Information and Electronics Engineering, IACSIT Press. Singapore, vol. 6, 2011.
- [21] C.-M. Vong, W.-F. Ip, P.-k. Wong, and J.-y. Yang, "Short-term prediction of air pollution in macau using support vector machines," Journal of Control Science and Engineering, vol. 2012, 2012.
- [22] A. Sotomayor-Olmedo, M. A. Aceves-Fern andez, E. Gorrostieta-Hurtado, C. Pedraza-Ortega, J. M. Ramos-Arregu in, and J. E. Vargas-Soto, "Forecast urban air pollution in mexico city by using support vector machines: A kernel performance approach," 2013.
- [23] W. Li, D. Kong, and J. Wu, "A new hybrid model fpa-svm considering cointegration for particular matter concentration forecasting: a case study of kunming and yuxi, china," Computational intelligence and neuroscience, vol. 2017, 2017.
- [24] A. J. Smola et al., "Regression estimation with support vector learning machines," Ph.D. dissertation, Master's thesis, Technische Universit at M unchen, 1996.
- [25] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.
- [26] L. Cao and F. E. Tay, "Financial forecasting using support vector machines," Neural Computing & Applications, vol. 10, no. 2, pp. 184– 192, 2001.
- [27] R. G. Brereton and G. R. Lloyd, "Support vector machines for classification and regression," Analyst, vol. 135, no. 2, pp. 230–267, 2010.
- [28] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," International journal of computer science, vol. 1, no. 2, pp. 111–117, 2006.

- [29] I.-K. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," Biometrika, vol. 87, no. 4, pp. 954-959, 2000.
- [30] H.-x. Zhao and F. Magoul'es, "Feature selection for support vector regression in the application of building energy prediction," in 2011 IEEE 9th International Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE, 2011, pp. 219–223.
- [31] L. Cagliero, T. Cerquitelli, S. Chiusano, P. Garza, G. Ricupero, and X. Xiao, "Modeling correlations among air pollution-related data through generalized association rules," in 2016 IEEE International Conference on Smart Computing (SMARTCOMP). IEEE, 2016, pp. 1–6.
- [32] P. S. Gromski, E. Correa, A. A. Vaughan, D. C. Wedge, M. L. Turner, and R. Goodacre, "A comparison of different chemometrics approaches for the robust classification of electronic nose data," Analytical and bioanalytical chemistry, vol. 406, no. 29, pp. 7581–7590, 2014.
- [33] A. Azid, H. Juahir, M. E. Toriman, M. K. A. Kamarudin, A. S. M. Saudi, C. N. C. Hasnam, N. A. A. Aziz, F. Azaman, M. T. Latif, S. F. M. Zainuddin et al., "Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in malaysia," Water, Air, & Soil Pollution, vol. 225, no. 8, pp. 1–14, 2014.
- [34] Q. Huang, J. Mao, and Y. Liu, "An improved grid search algorithm of svr parameters optimization," in 2012 IEEE 14th International Conference on Communication Technology. IEEE, 2012, pp. 1022– 1026.
- [35] P. Hajek, V. Olej et al., "Predicting common air quality index-the case of czech microregions," Aerosol and Air Quality Research, vol. 15, no. 2, pp. 544–555, 2015.
- [36] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization." Journal of machine learning research, vol. 13, no. 2, 2012.