WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

SCALABLE APPROACH FOR DETECTING AIR QUALITY INFERENCE USING ENSEMBLE REGRESSORS

Mrs. Preethy Jemima P[1], Singa Selvamani S[2]

[1] Dept of CSE SRM Institute of Science and Technology preethyj@srmist.edu.in Samarendra T Dept of CSE SRM Institute of Science and Technology st4516@srmist.edu.in

[2] Dept of CSE SRM Institute of Science and Technology ss2360@srmist.edu.in Vimal Aditya Raj Dept of CSE SRM Institute of Science and Technology vv5019@srmist.edu.in

ABSTRACT—

Air pollution prediction is not exempt from the fast-expanding effect of machine learning technologies, which are present in nearly every field. And delves into research concerning air quality prediction utilizing machine learning models, concentrating in particular on sensor data in the context of smart cities. By leveraging prominent databases and implementing targeted filtration processes, the most pertinent papers were identified and scrutinized. Subsequently, through a comprehensive review, the primary features were extracted, forming the foundation for linking and comparing them to one another. The most important stage of data preprocessing is feature selection. But none of the earlier research on concentration prediction employed machine-learning techniques; instead, it was restricted to statistical method feature selection.. The aim is to investigate the influence of various input variables on training diverse air quality indices. Normalization of data was achieved by using correlation analysis in conjunction with the Min-Max normalization method to determine which variables are highly connected.. Subsequently, essential features were extracted from these highly correlated variables using an optimization algorithm. To validate the results, several measures were calculated, the correlation coefficient and the mean absolute error.

As a result of swift industrialization and urban expansion, The concentration of air pollutants rises significantly as an outcome of unavoidable emissions of production-related gases into the ambient air. Excessive air pollution concentrations are hazardous to respiratory health since they agitate the respiratory system and could lead to damage to it. This might even have a direct impact on a person's ability to breathe on themselves and increase their risk of developing lung cancer. Accurate air pollution concentration modeling helps to provide prompt notifications for levels that differ. Authorities can take proper remedial measures based on expected pollution levels, and individuals can modify their activities accordingly. This may significantly mitigate the adverse impacts that pollutants in the air has on people.

Keywords—Air Quality Index, correlation coefficients

INTRODUCTION

Air pollution presents immense challenges to public health and environmental stability in cities globally. As noted by the World Health Organization (WHO It exerts an enormous strain on society and increases the risk of numerous diseases among populations. As such, analysis of air quality is now an important tool for people and society combined. On the one hand, precise monitoring of air pollution empowers those who make decisions to establish efficient laws regarding the environment and specific measures meant to reduce emission levels of pollutants. But it can also enable individuals to make educated decisions to avoid exposure to harmful pollutants, such as changing travel routes or cutting back on outdoor activities. As a consequence, atmospheric analytics has drawn a lot of attention in recent decades, which has given rise to plenty of research topics and uses, like forecasting, pollution pattern mining, and air quality inference.

The aforementioned changes have facilitated a deeper comprehension of air pollution, fostering the creation of more precise air quality monitoring and forecasting systems. There has been a growing focus on the deterioration of air quality, with particulate matter (PM) is being warned for its considerable harmful effects on human health. Due to its small dimensions, fine PM can impede gas exchange in the lungs by permeating deeply into the alveoli and even reaching the bronchioles. According to several studies, breathing in particulate matter over an extended period raised the risk of lung cancer, respiratory disorders, and cardiovascular disease. An increasing number of cities have set up sites for air quality monitoring due to concerns about public health. Nevertheless, the majority of services just display the air quality as it is right now; they do not predict it.

Anticipating the state of the surrounding environment is crucial for directing people's behaviors toward minimizing their exposure to PM2.5, including deciding indoor or outdoor activities.

RELATED WORKS

"The authors present sensor validation methods and data analysis for a dense air quality sensor network. They show solutions to challenges in a large-scale sensor network deployment. They use data from a dense air quality sensor network deployment, located in Nanjing downtown, China, that comprises 126 LCSs and 13 reference stations. Since the majority of sensors deployed in the network are based on LCSs, they are prone to have low-quality data. Therefore, they propose three methods of sensor validation. First, The authors perform a reliability investigation to evaluate all LCSs in the network to observe if they provide reliable measurements as a whole in comparison to the measurements of all reference stations. Thus, they compare the measurements between all LCSs and the reference stations through statistical properties and correlation coefficients between pollutant variables measured at both sensing units. Second, they perform accuracy tests on a few of the LCSs that are nearest to the reference stations. The accuracy tests are generalized to the remaining LCSs in the sensor network as the LCSs are based on the same sensing technology, as they are identical units." [1]

"This paper develops and tests a linear adaptive model for the RSSI signal at the AQIMoS (Mobile Air Quality IPB Monitoring System) sensor node based on GSM/GPRS cellular communication. The proposed model can adaptively the response time limit based on the RSSI and adjust from Atang Sanjaya Airport, Bogor City. vehicles, especially diesel trucks carrying sand, as shown in monitoring equipment during the test were considered. The complete experimental results regarding the application of AQIMoS require further testing. Pollutant sensor data sent from AQIMoS to the server were downloaded using a web application and averaged into hourly concentration used to determine the fluctuation in the concentration pattern change in the measurement time range". [2]

"In this paper, the authors propose a deep learning approach for anomaly detection by considering the spatial correlation, temporal correlation, and multivariate features of air quality data. The essential idea of this approach is to combine the temporal correlation and spatial correlation of air quality data, use the node information correlation degree for feature fusion, represent the spatiotemporal correlation of air quality data in the spatiotemporal graph structure data, and use them for anomaly detection." [3]

"In this paper, the authors employ the explainable deep learning method, Shapely Additive explanations, to reveal the influence of meteorological conditions on air quality prediction. The essential idea is to use the SHAP interpretation method to interpret the established LSTM and GRU air quality prediction models and analyze the influence of meteorological conditions on air quality prediction. The results show that (1) in both the LSTM and GRU models, the prediction accuracy is not improved by considering only meteorological conditions. However, when considering other air pollutants, the prediction accuracy is improved, and when combining meteorological conditions with others the prediction accuracy is even higher. air pollutants, (2) Whether only considering meteorological conditions or combining meteorological conditions and other air pollutants for PM2". [4]

"This study examines the problem of missing data recovery, using air pollution data recovery as a case study. The problem has been formulated and two widely used data recovery approaches, namely, the Interpolation approach and the Matrix Completion approach, have been introduced. Given the heterogeneous distribution of monitoring stations for air pollution and meteorology, a new strategy to reconstruct the data matrix to recover the missing air pollution data has been proposed. Next, the low-rank property of a newly constructed Y. Yu et al.: Novel Interpolation-SVT Approach for Recovering Missing Low Rank Air Quality Data matrix has been introduced. The formulated AQDR problem can be transformed into an LRMC problem". [5]

"In this paper, the authors use the Light GBM model to process the high-dimensional data to predict the PM2.5 concentration in 24 hours based on the historical datasets and predictive datasets. The authors proposed a predictive data feature exploration-based air quality prediction approach. The approach enables to deeply mine and explore the high-dimensional time-related features and statistical features based on the exploratory analysis of big data. The authors utilize the sliding window mechanism of increasing the amount of training data to improve the training effect of the model and employed the air quality historical dataset of Beijing to evaluate the prediction model. The experimental results show that the approach outperforms the other baseline models". [6]

"It examines the process of building a spatial and temporal network model of air quality using the intricate Innovative Spatial-Temporal Network Modeling and Analysis Method of Air Quality, G. Zhao et al. "[7]

"In this paper, the authors proposed the n-step AAQP, which is an attention-based seq2seq model, for air quality prediction. The n-step AAQP had better performance than the seq2seq models. To accelerate the training process of seq2seq with attention, an FC encoder replaced the RNN encoder of seq2seq. In addition, position embedding was introduced to help the FC encoder extract the sequential information. Moreover, the performance of the AAQP was close to the seq2seq with attention at the Olympic Center station and is even better at Dongsi station. To overcome the shortcomings of the accumulated errors as the time step grows, the n-step recurrent prediction was applied. Through n-step recurrent prediction, the performance of the AAQP was significantly improved. In addition, the training of seq2seq was further accelerated. The two promotions make seq2seq have more accurate predictions and higher training speed. Particularly, the AAQP can give a trustworthy alert 2 hours in advance before sudden air pollution strikes. Additionally, the weather forecast data are essential to improve the accuracy of air quality prediction". [8]

"In this paper, the authors propose two approaches for AQI estimation and prediction, both based on meteorological and historical pollutant data; one learns a model based on the previous AQI and meteorological data to predict AQIs, and the other learns models based on the previous pollution data and meteorological data to predict pollution concentrations and then compute AQIs. Both approaches can get good band accuracy (over 75%), as shown in the evaluations conducted across various datasets. The best approach is the latter approach combined with a neural network, which achieves the lowest RMSE and MAPE across most of the evaluated datasets. This approach gets very good band accuracies (more than 81%) on all the datasets. However, by further analyzing the individual pollutant value prediction step, the authors found that a neutral network-based method is not the optimum for predicting PM10 data. Therefore, the authors recommend using linear regression to predict AQI if the dominant pollution is PM10 in the area of interest. In summary, the results show the feasibility of our proposed approaches for predicting AQIs based on meteorological data and the historical pollutant data/AQIs." [9]

"The main contribution of this work is the definition of a development process based on big data and intelligent systems concepts for a traffic regulation system according to air quality data. The authors have, through this paper presented the implementation of an air quality system for recommendation and traffic regulation over distributed data gathered from different air quality sensors, user devices, and other external databases, that are managed using Hadoop to ensure fast data Agent-Based Traffic Regulation System for the Roadside Air Quality Control". [10]

THE PROBLEM STATEMENT OF THE PROJECT

In order to enhance forecasting performance, many complicated machine learning models have been proposed lately. However despite the fact they might improve prediction accuracy, humans continue to encounter major difficulties in interpreting the model the results due to the ever-increasing complexity of the model. Consequently, improving models' interpretative ability has become a vital topic. Interpretability, on the one hand, increases decision-makers' faith in the models by giving them supporting data regarding the models' outcome. On the other hand, interpretability empowers researchers and data scientists to have a comprehensive comprehension of the models' strengths and weaknesses, casting light on model optimization and design.

PROPOSED WORKS

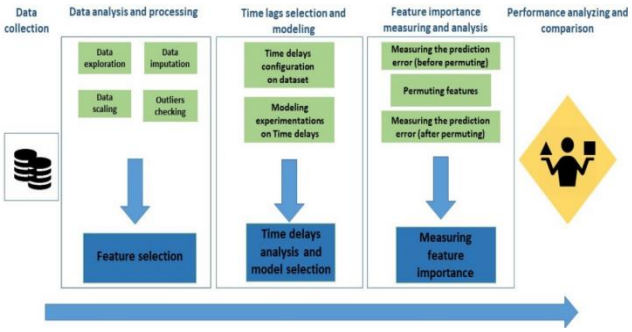

Figure 1. Proposed architecture diagram.

Outlier and Missing Data Processing:

Outliers and values that were missing in the original data need to be evaluated. The data has been reviewed and came from the official platform, subsequently no anomalous values—like readings for air pressure or atmospheric concentration that were in fact negative—were observed. The data are regarded to be accurate within acceptable boundaries, therefore the analysis of outliers does not appear mandatory.

A Min-Max normalization technique was used for standardizing the data. This makes it easier in minimizing scale inconsistencies or units from the collected data. Taking advantage of the Min-Max normalization approach, data values are scaled within an established range (zero to one). The Min-Max normalization method splits data points by their range following the fact that initially omitting the minimal value.

Feature selection is the process of minimizing the number of input variables when building a predictive model Stepwise methods of regression which includes forward selection begins with a null model. Each time implementing this method, the model starts out empty and gradually includes variables until no variable remaining of the model can meaningfully impact the model's outcome. During forward selection, the variable with the lowest p-value below the cut-off or the highest test statistic above the cut-off is chosen and added to the model.

The most basic selection of variables technique, backward elimination, begins with a complete model which incorporates every variable. The variable with the lowest test statistic or the greatest p-value over the cutoff is the one the fact that gets eliminated as well. This process continues until all remaining variables are statistically significant at the designated cut-off.

For tasks like regression and classification, ensemble learning techniques called random forests (RFs) or random decision forests are used. The Random Forest (RF) algorithm generates a large number of decision trees across various training cases. The class that represents the mean prediction (in tasks including regression) or the mode of classes (in tasks involving classification) produced from the individual trees is then output.

The comparative results in mean square error (MSE), mean absolute error (MAE), and coefficient of determination (R2) all outperform the comparison models. The following lists our paper's primary contributions:

- Initially, a Min-Max normalization strategy was employed in this investigation, which successfully maintains the low standard deviation relationship between the data values. Using varying sliding window sizes, the Min-Max normalization technique minimizes the impact of outliers in time series forecasting by forecasting the concentration for the following hour.

- Multivariate prediction considers both the overall prediction accuracy of multiple pollutants and the prediction accuracy of each pollutant, which is a significant improvement over based models.

- The forecast models discussed above do not consider the correlation qualities between pollution and meteorological data. We provide a thorough investigation of prediction issues related to pollution and meteorological data in order to improve prediction accuracy by utilizing the intricate correlation patterns within the model's input data.

MODULE DESCRIPTION

Data Preprocessing:

For projects involving deep learning and machine learning, data pretreatment is crucial. Its goal is to convert unprocessed data into a format that can be used for cleaning, noise reduction, and model training in order to improve model performance. Transforming a dataset's data values with the goal of improving the information that is collected and processed is known as data preparation. Normalizing the data reduces the complexity of the method for this corresponding processing because the dataset's maximum and minimal values generally have a very wide contrast. The data normalization gives enough advantages for neural network-related algorithm classification. We loaded these datasets using the Pythons Pandas library.
Categorical coding needs to be performed for transforming wind direction from a non-numerical data type into a numerical one. For pollution and meteorological data, missing values are filled in using the average of the data collected before and after the missing time. Next, the Min-Max function was used to convert meteorological and pollution data to a range in order to remove the impact of numerical variations on the precision of forecasts.

Correlation Analysis:

The matrix displays the correlation between each possible pair of data in a table.. It serves as a valuable resource for compiling large datasets and visualizing data patterns. A correlation matrix streamlines the asset selection process by presenting correlations among various assets. Identifying correlations between PM concentrations and influencing Factors are critical for developing a robust prediction model. It ensures that the effective characteristics for AQP are used in the suggested regression model. There are various factors that affect PM2.5. elements, but each and every one of them is crucial for efficient AQP. Conversely, the time complexity of the suggested model is impacted by the inactive and irrelevant components. As a result, figuring out each factor's correlation coefficient (CC) is critical in order to help choose the best features needed for precise air pollution forecasting. Let's take into account characteristic time series data, represented as x = (x1, x2, x3,..., xn), and other data, represented as y = (y1, y2, y3,..., yn). Equation describes how to compute the CC between the factors.

$$\frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Where n is the number of samples, $0 < r < 1$ denotes a positive correlation, and $-1 < r < -1$ denotes a negative correlation. If the absolute value of r is closer to 1, the correlation is higher and the range between x and y is constrained.

Random Forest Regressor:

Growing classification and regression trees (CARTs) have been integrated into the RF method. Random vectors are used to construct each CART. The primary parameters of the RF-based classifier model were the quantity of decision trees and the number of features (NF) in the random subset

at every node in the increasing list. Choosing the appropriate number of decision trees was the first step in the model training process. A higher number is preferable in terms of trees, but it requires more time to calculate. increased variance reduction and increased bias are the results of a lower NF. The empirical formula NF = √M, where M is the total number of features, can be used to define NF. Regression and classification problems can both be handled by Random Forest (RF), depending on whether the underlying trees are regression or classification trees. The final output of the regression model is

H(x) = 1 =1 *h*i (x), where T is the number of regression trees and hi(x) is the output of the i-th regression tree (hi) on sample x. This is assuming that the model has T regression trees (learners) for regression prediction. As a result, the average of each tree's anticipated values is the RF forecast.

Support Vector Regression (SVR):

Support Vector Machine (SVM) is a machine learning approach that creates hyperplanes to divide classes. A popular method for evaluating data with categorical output variables is Random Forest (RF). However, for continuous numeric output variables, regression analysis, such as Support Vector Regression (SVR), is employed. Various SVR kernels, such as linear, polynomial, radial basis function (RBF), sigmoid, and precomputed kernels, were evaluated in this study. Out of all of them, the linear kernel function performed better, making it the model parameter for SVR in this investigation. SVR has a number of benefits, such as high prediction accuracy, simplicity of implementation, and resilience to outliers. SVR has been used to improve prediction accuracy by overcoming non-linear constraints and uncertainty. SVR has been effectively used to forecast Bangkok, Thailand's PM10 concentration levels using meteorological variables and data on air quality.

Performance Measures:

The effectiveness of the suggested model was assessed using a variety of loss functions, including MAE, SMAPE, RMSE, and MSE. The actual state of the forecasting error was accurately captured by the MAE performance metric. Furthermore, SMAPE and other performance measures efficiently assess the extent of data modification and gauge the accuracy of the suggested model's predictions. Conversely, the is calculated as the mean square difference, or average, between the estimated and actual values. The performance measurements' mathematical formulations, which include SMAPE, R2, MAPE, RMSE, MAE, and MSE, are given in The following are the formulas used to calculate the five evaluation metrics mentioned above.

1. MEAN ABSOLUTE ERROR (MAE):

$$\text{MAE} = \sum_{i=1}^{n} \frac{P_i - r_i}{n}$$

where Pi and ri stand for the set of actual values and the set of anticipated values, respectively, and n indicates the number of sets.

2. ROOT MEAN SQUARE ERROR (RMSE):

$$\text{MAE} = \sqrt{\sum_{i=1}^{n} \left(\frac{P_i - r_i}{n}\right)^2}$$

where Pi and ri denote the sets of predicted and actual values, respectively, and n denotes the number of sets.

3. NORMALIZED MEAN AVERAGE ERROR (NMAE)

$$NMAE = \frac{MAE}{r_{max} - r_{min}}$$

where rmin denotes the target column's minimum value and rmax denotes its maximum value.

RESULT

	CO_GT	PT08_S1_CO	NMHC_GT	C6H6_GT	PT08_S2_NMHC	Nox_GT	PT08_S3_Nox	NO2_GT	PT08_S4_NO2	PT08_S5_O3	T	RH	AH
count	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000
mean	-34.207524	1048.990061	-159.090093	1.865683	894.595276	168.616971	794.990168	58.148873	1391.479641	975.072032	9.778305	39.485380	-6.837604
std	77.657170	329.832710	139.789093	41.380206	342.333252	257.433866	321.993552	126.940455	467.210125	456.938184	43.203623	51.216145	38.976670
min	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000
25%	0.600000	921.000000	-200.000000	4.000000	711.000000	50.000000	637.000000	53.000000	1185.000000	700.000000	10.900000	34.100000	0.692300
50%	1.500000	1053.000000	-200.000000	7.900000	895.000000	141.000000	794.000000	96.000000	1446.000000	942.000000	17.200000	48.600000	0.976800
75%	2.600000	1221.000000	-200.000000	13.600000	1105.000000	284.000000	960.000000	133.000000	1662.000000	1255.000000	24.100000	61.900000	1.296200
max	11.900000	2040.000000	1189.000000	63.700000	2214.000000	1479.000000	2683.000000	340.000000	2775.000000	2523.000000	44.600000	88.700000	2.231000

Figure 2: Results of training a random Forest Regressor model.

Figure 2 Represents the results of training a Random Forest Regressor model to predict Relative Humidity (RH) based on other air quality measurements in the dataset.

- y_test: This column contains the actual Relative Humidity (RH) values from the testing set of the data. These are the real values the model was trying to predict.
- y_pred: This column contains the Relative Humidity (RH) values predicted by the Random Forest Regressor model for the testing set.
- error: This column represents the difference between the actual values (y_test) and the predicted values (y_pred) for each data point in the testing set. It essentially shows how far off the model's predictions were from the real values.

In essence, this table allows you to compare the model's predictions (y_pred) with the actual values (y_test) and see the corresponding errors (differences) for each data point in the testing set. This helps you evaluate how well the model performed in predicting Relative Humidity.

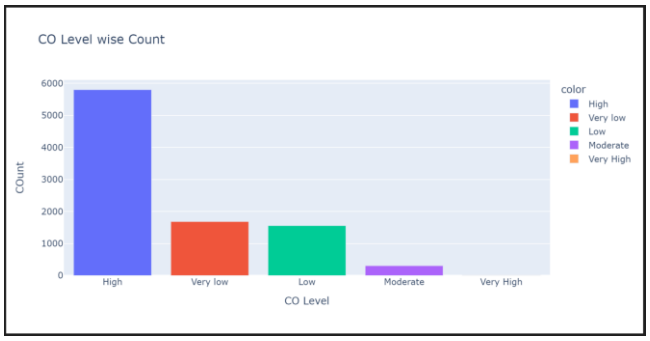

Figure 3: Scatter plot visualizing the performance of a machine learning model on a prediction task

Figure 3 The graph you sent appears to be a scatter plot visualizing the performance of a machine-learning model, likely the Random Forest Regressor you trained.

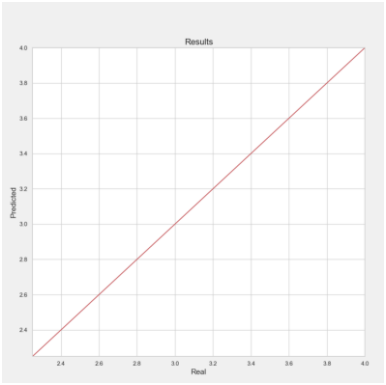

Figure 4: Final result

Figure 4 Explain the graph representation of predicted and real output. The Y-axis show predicted and x-axis shows real.

CONCLUSION

This work adds to a program that makes it easier to transform data, do linear regression, fit mathematical models, and process prediction functions. This makes it possible to display results showing pollutant correlations graphically. By demonstrating accurate predictions that closely match the data, it assesses the precision of linear regression functions. Predicting air quality accurately is essential for public health and disaster preparedness. Regression models predict air quality markers through machine learning. According to experimental results, the RFR model outperforms the SVR model while both yield respectable results. Large datasets are not a good fit for SVR because of its temporal complexity, which increases cubically with sample size. Accordingly upward, this study provides two prediction models that are customized for various situations, improving the accuracy of air quality predictions and providing information for the investigation of urban air quality.

REFERENCES

1. Martha Arbayani Zaidan,Yuning Xie,Naser Hossein Motlagh,Bo Wang,Wei Nie,Petteri Nurmi,Sasu Tarkoma,Tuukka PetÃ jÃ ,Aijun Ding,Markku Kulmala Dense Air Quality Sensor Networks: Validation, Analysis, and Benefits IEEE Sensors Journal, 2022

2. Rady Purbakawaca ,Arief Sabdo Yuwono,I. Dewa Made Subrata, Supandi, Husin Alatas Ambient Air Monitoring System With Adaptive Performance Stability pieces, 2022

3. Xiaoling Lin, Hong zhang Wang, Jing Guo, Gang Mei A Deep Learning Approach Using Graph Neural Networks for Anomaly Detection in Air Quality Data Considering Spatiotemporal Correlations IEEE Access, 2022

4. Yuting Yang, Gang Mei, Stefano Izzo Revealing Influence of Meteorological Conditions on Air Quality Prediction Using Explainable Deep Learning IEEE Access, 2022

5. Yangwen Yu, James J. Q. Yu, Victor O. K. Li, Jacqueline C. K. LamANovel Interpolation-SVT Approach for Recovering Missing Low-Rank Air Quality Data pieces, 2020

6. Ying Zhang, Yanhao Wang, Minghe Gao, Qunfei Ma, Jing Zhao, RongrongZhang, Qingqing Wang, Linyan Huang A Predictive Data Feature Exploration-Based Air Quality Prediction Approach IEEE Access, 2019

7. Guyu Zhao,Guoyan Huang,Hongdou He,Qian Wang Innovative Spatial-Temporal Network Modeling and Analysis Method of Air Quality IEEE Access, 2019

8. Bo Liu,Shuo Yan,Jianqiang Li,Guangzhi Qu,Yong Li,Jianlei Lang,Rentao Gu A Sequence-to-Sequence Air Quality Predictor Based on the n-Step Recurrent Prediction IEEE Access, 2019

9. Yuchao Zhou, Suparna De,Gideon Ewa, Charith Perera, Klaus Moessner Data-Driven Air Quality Characterization for Urban Environments: A Case Study IEEE Access, 2018

10. Abdelaziz El Fazziki, Djamal Benslimane, Abderrahmane Sadiq, Jamal Ouarzazi, Mohamed Sadgal An Agent-Based Traffic Regulation System for the Road side Air Quality Control IEEE Access, 2017 45

11. C. J. L. Murray et al., Global burden of 87 risk factors in 204 countries and territories 1990 2019: A systematic analysis for the global burden of disease study 2019, Lancet, vol. 396, no. 10258, pp. 1223-1249, 2020

12. Clean Air Strategy, DEFRA, London, U.K, 2019.

13. T.-M. Chen, W. G. Kuschner, J. Gokhale and S. Shofer, Outdoor air pollution: Nitrogen dioxide sulfur dioxide and carbon monoxide health effects, Amer. J. Med. Sci., vol. 333, no. 4, pp. 249-256, Apr. 2007.

14. C. Holman, R. Harrison and X. Querol, Review of the efficacy of low emission zones to improve urban air quality in European cities, Atmos. Environ., vol. 111, pp. 161-169, Jun. 2015.

15. J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, Data-driven intelligent transportation systems: A survey, IEEE Trans. Intel. Transp. Syst., vol. 12, no. 4, pp. 1624-1639, Dec. 2011.

16. A. I. Torre-Bastida, J. Del Ser, I. La na, M. Ilardia, M. N. Bilbao and S. Campos- Cordob, Big data for transportation and mobility: Recent advances trends and challenges, IET Intel. Transp. Syst., vol. 12, no. 8, pp. 742-755, Oct. 2018.