ISSN: 2229-6956 (ONLINE) ICTACT JOURNAL ON SOFT COMPUTING, APRIL 2023, VOLUME: 13, ISSUE: 03
DOI: 10.21917/ijsc.2023.0412
SPECIAL ISSUE ON INNOVATIONS AND INTELLIGENCE FOR COMPUTING AND TECHNOLOGIES (IICT-2023)

PREDICTION OF AQI USING HYBRID APPROACH IN MACHINE LEARNING

Reema Gupta and Priti Singla
Department of Computer Science and Engineering, Baba Mastnath University, India

Abstract

Forecast of urban air pollutant concentrations must deal with a growth in the volume of environmental monitoring data as well as complex changes in the air pollutants. This initiates the need of efficient prediction techniques to increase prediction accuracy and stop pollution causing issues. In this paper, the combination of Random Forest Regression (RFR) and Support Vector Regression (SVR) based machine learning model is proposed to predict the AQI values. Daily meteorological and air pollution data between March 2020 and June 2022 from Jind city in Haryana were used for model training and test the model's results. First, the important factors affecting air quality are extracted and null values are replaced by their mean values to handle the irregularities in the dataset. Second, the Random Forest regression machine is used for model training and prediction of the value, the SVR model is used in correction of residual items, and finally the predicted values of AQI are obtained. The experimental results showed that the proposed prediction model of RFR-SVR had a better prediction result than the standard Random Forest Regression, support vector regression machine learning. RMSE, R2, MAPE are used as evaluation indicators to evaluate the performance of the proposed model.

Keywords:

Air Quality Index, RFR, Support Vector Regression, Hybrid Model, Coefficient Metric

1. INTRODUCTION

Air pollution is a recurring problem that is affecting human life and other resources continuously. The main sources of air pollution are combustion of fuels, heavy factories (industrial areas) and vehicle movement. Most of the people residing in densely populated areas are affected by excessive air pollution from these sources and are exposed to poisonous haze for extended time-span. The increases, particularly in fine particulate matter (PM2.5), have led to school closures, a state government health emergency declaration, protests by civil society and frightening media coverage. According to Teri's study the vehicle pollution causes 28% of PM2.5 emission and out of this, 9% is caused by truck and tractors while least by light vehicle emission [1] [2].These recurring bouts are now a yearly occurrence at the start of winter, made worse by the holiday season. It has also been observed that air pollution continues to have a strong correlation with fatalities from cancer, cardiovascular, and respiratory disorders [3] [4] studied the relation between Lung cancer and

PM2.5 pollutant and further ozone component was also added to analyze its effect and impact on the human health with variable parameters. The goal of air quality control is to maintain an environment that is safe for human health and the environment. For the air quality control there is need of monitoring and prediction of air quality index.

This paper proposed a model that can predict the air quality index using machine learning techniques and performance of proposed model is compared with the classic ones with various measurements to reflect proposed model's effectiveness. AQI calculation is based on various pollutants such as Oxides, Dioxides, Ozone, Particulate matters, etc. existing in the air and some effective meteorological parameters that has correlation with AQI and value of AQI varies with the change in the value of these parameters.

This paper is divided into five sections wherein first section presents the introduction; second section discussed about the related works carried out; third section presents the proposed model for the prediction of Air Quality Index. Further, results evaluation is presented in Section 4 and finally, the paper is concluded in last section.

2. RELATED WORKS

The population of metropolitan areas is growing quickly, and this has brought a number of environmental issues. Systems for monitoring air quality are a significant contribution to the endeavor of lessening the effects of pollution in air on the living beings. AQI is a measurement which indicates the presence of pollutants and other harmful gases in our air. It educates the humans about the air quality level in their immediate environment and, consequently, the health risks associated with it, especially for exposed populations like children, the elderly, and those with respiratory or cardiovascular problems [5].

AQI forecasting is not only one of the requirements for promoting urban public health, but also essential for environmental sustainability under the adverse effects of air pollution. When discussing air quality issues, linear regression analysis models are frequently utilized [6]. This model's benefits include easy calculation, and it is appropriate for interpreting regression coefficients and producing distinctive output outcomes, it is not appropriate for solving nonlinear issues. [7] proposed hybrid model for the better performance for the prediction which fuses the advantages of optimization algorithm, AI techniques.

Based on labeled training data, supervised learning maps the input values to the output values. The actual reaction is provided by the learning system. The discrepancy between the desired response and the actual response is the error signal [8]. Comparative analysis of the existing techniques under classification, regression and Ensemble techniques was done and it has been concluded that the DT, SVR and stacking ensemble methods are better among others in terms of their performances [9] In a study, Support Vector regression and Random Forest regression based model were used for the prediction of air quality of Beijing. Performance of Random Forest regression was better as compared to the Support Vector regression as with increasing number of samples the processing time of SVR increased cubically [10].

Along with the prediction of Air quality index prediction model can be used to predict the various pollutants such as PM2.5,

PM10, Sulphurdioxide etc. [11] used the SVR model with Radial Basis Function kernel to predict the concentrations of pollutants along with the AQI. Contaminants and particles are volatile, dynamic, and highly variable in both place and time, predicting the quality of the air is a difficult task. Simultaneously, due to the essential effects of air pollution on the human health and the environment, it is becoming more and more crucial to model, predict, and monitor air quality, particularly in urban areas. For accurate hourly prediction of pollutants and AQI levels in California RBF kernel of SVR was used [12]. The level of only one pollutant i.e., CO was assessed using Random Forest Regression (RFR), Decision Tree Regression (DTR) and Linear Regression (LR) algorithms but meteorological parameters was not considered in the process [13].

Multi-label classification approach was used by [14] for pollutant prediction and it has been concluded that this approach has more accuracy as compared to the independent ones. Appropriate combination of input parameters is most important for the accurate results [15] considers the one, two, three air pollutions input and used it for the prediction process and it has been concluded that one of the particulate matter with Ozone is the best combination of the input parameters for global and diffuse solar radiation prediction. In a study, Hybrid deep learning model based on CNN and LSTM was proposed and used to predict future PM2.5 concentrations and the results are compared with the results obtained from classical numerical model [16].

Six types of pollutants prediction were carried out with the help of combination of principal component analysis, support vector regression, autoregressive moving average model. The combination approach has better results as compared to the existing ones [17]. Prediction of air quality of Gurugram (Haryana) 2016 to 2019 was carried out using hybrid approach k means clustering followed by SVM and results were compared with the traditional SVM. It has been concluded that the accuracy of hybrid model i.e., 91.25%while the traditional SVM algorithm's accuracy is 65.93% [18]. Decisions about the resources of the atmosphere require careful consideration of concentration forecasts. This has detrimental implications on one's health, including increased cardiovascular and respiratory mortality due to disease. It is essential to improve air quality and provide efficient environmental monitoring to prevent air pollution in advance [19].

3. PROPOSED HYBRID METHODOLOGY

Meteorological parameters and pollutants are considered as input variables for the prediction. For prediction of Air Quality Index, RFR and SVR are used in the hybrid approach. After prediction, the performance is evaluated using performance metrics and compared based on the parameters. The process is shown in Fig.1.

Dataset of Jind city is considered for the prediction of AQI of approx. two years from 1st March 2020 to 30th June 2022. Dataset is collected from the open-source Central pollution Control board website which contains numerous pollutant values and meteorological parameters. The pollutants like PM2.5, PM10, NO_2, SO_2 and Ozone were gathered every day from Jind city. The dataset details used in this research are shown in Table.1.

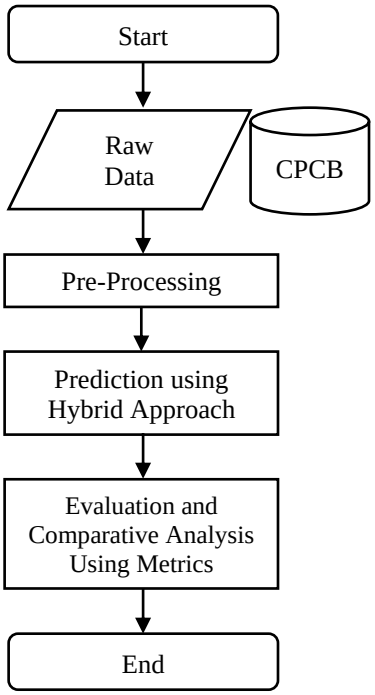

Fig.1. Process used for prediction using Hybrid Approach

Table.1. Description about the dataset

Parameters	Value
Dataset year	March 2020 to June 2022
Train: Test	80:20
Dataset source	Central Pollution Control Board
Instances	851 each
Location	Jind
Pollutants	PM2.5, PM10, NO2, SO2, Ozone
Meteorological Parameters	Temperature, Relative humidity, Wind speed and Wind direction
Target Variable	AQI

3.1 PREPROCESSING

This step deals with the irregularities within the dataset for example missing values or data having values 'None' or 'NA' are replaced with the mean of the corresponding column. Dataset is represented as Dn.

Dn = Dn.replace (to_replace='NA', value=np.nan)

Dn = Dn.replace (to_replace='None ', value=np.nan)

Dn['PM2.5'] = Dn['PM2.5'].as type('float64')

Dn['PM2.5'].replace({np.nan: Dn['PM2.5'].mean()},inplace =True)

This task is carried out on all components (Pollutants and other input parameters)

3.2 HYBRID APPROACH

A Hybrid model with combination of Support Vector Regression and Random Forest Regression techniques is proposed for the prediction of air quality index of Cities in Haryana, India.

Dataset is splitted using

X_train_1 = dataset.iloc[:, 2:7].values

y_train = dataset.iloc[:, 11].values

X_train_2 = dataset.iloc[:, 7:11].values

Model 1 is trained using X_train_1 and y_train

Model 2 is trained using X_train_2 and residual of y_train and y_pred_1 is the predicted value calculated using model1.

3.2.1 RFR (Random Forest Regression):

RFR [20] is a supervised learning technique which leverages the ensemble learning approach for regression. The ensemble learning method combines predictions from various machine learning algorithms to provide predictions that are more accurate than those from a single model.

3.2.2 SVR (Support Vector Regression):

SVR is an approach for supervised learning which is used to forecast discrete values. The SVMs and SVR both operate on the same theory. The hyperplane with the most points is the best-fitting line in SVR.

3.3 EVALUATION

The performance of proposed hybrid model are evaluated through Metrics such as Mean absolute error, Mean Square error, Root mean square error, and coefficient R^2 are calculated and shown in Eq.(1) to Eq.(4).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|z_i - z_i'\right| \tag{1}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(z_i - z_i'\right)^2 \tag{2}$$

$$RMSE = \sqrt{MSE} \tag{3}$$

$$R^2 = \frac{\sum_{i=1}^{n}\left(z_i - z_i'\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(z_i - z_i''\right)^2} \tag{4}$$

Here n is the sample size, z_i and z_i' represents actual and predicted value respectively whereas z_i'' denotes mean of the values.

4. RESULTS

Numpy, Matplotlib, Pandas and Sklearn are the major libraries used in the implementation using python for the air quality prediction. Fig.2-4 showing the predicted values of AQI of Jind city using RFR, SVR and proposed hybrid model. The performance analysis of the predicted AQI value over the time period for RFR, SVR and Hybrid one is shown in Fig. 2-4. The applied methods predict the AQI value for the given dataset of Jind city.

The summary of actual value of test dataset (80:20 train: test) is described in Table.2. Values are predicted using approach 1 i.e., Random Forest regression and the predicted values are described in column 2. Similarly, prediction of values is done using approach 2 i.e. Support Vector Regression and the values are represented in Column 3, Column 4 contains predicted values using proposed hybrid approach in Table.2.

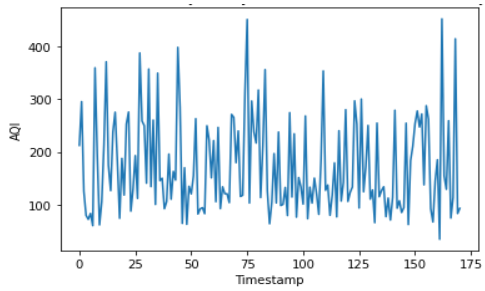

Fig.2. Predicted AQI Time series using RFR

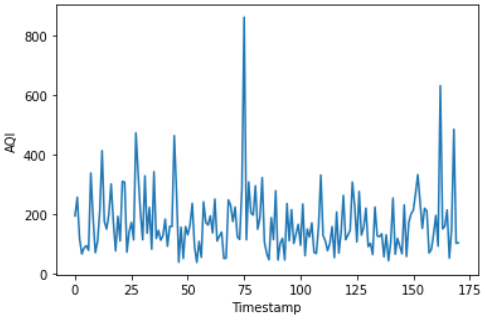

Fig.3. Predicted AQI Time series using SVR

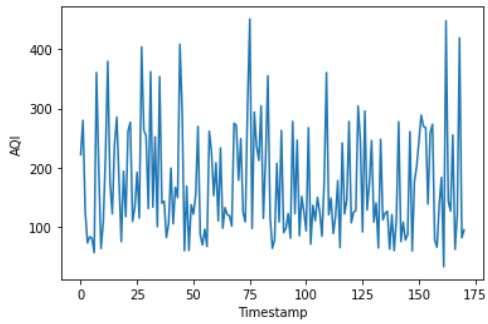

Fig.4. Predicted AQI Time Series Using Hybrid Model

Table.2. Actual and Predicted values [171 rows] using RFR, SVR and Proposed Approach

Actual Value	Predicted Value		
	RFR	SVR	Proposed Approach
219.0	212.507812	193.559585542	221.780885
301.0	295.600000	256.151537	279.899239
94.0	125.600000	118.228334	126.803107
58.0	80.061719	65.012347	72.789054
54.0	72.407813	85.052516	82.967045
56.0	74.807812	50.926027	61.942690
114.0	114.200000	151.601683	121.782276
366.0	414.500000	485.734213	419.299594
32.0	83.457813	100.778401	81.422748
77.0	93.407813	102.677661	94.680046

In Fig.5-Fig.7, the actual value is shown in blue color circle for the AQI of Jind city, the predicted AQI values are shown as red color cross using RFR, SVR and proposed approach respectively. It is observed that the proposed model fits for both training as well as test data.

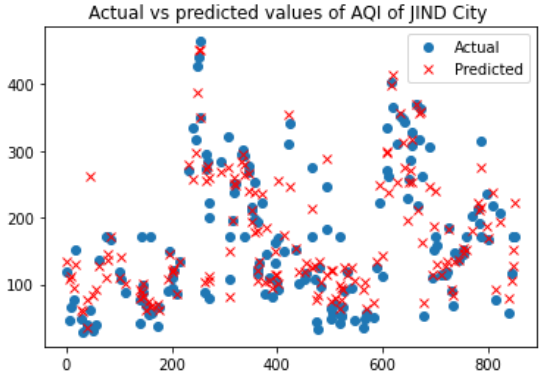

Fig.5. RFR Approach

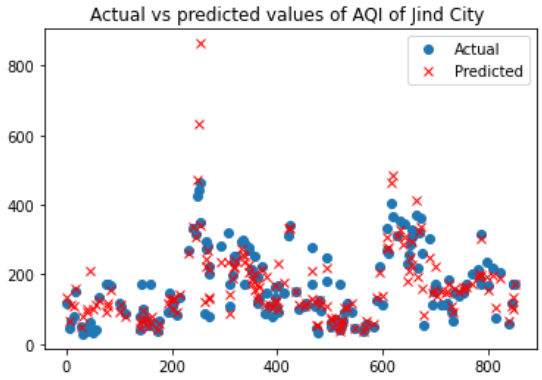

Fig.6. Using SVR approach

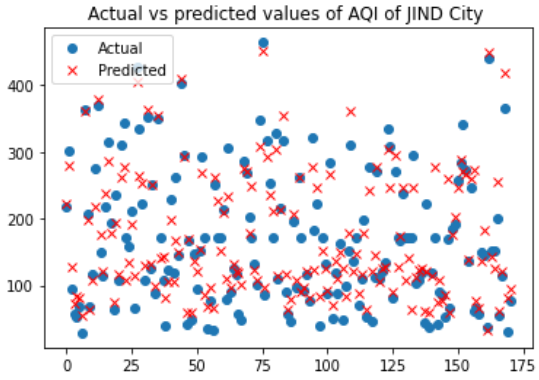

Fig.7. Proposed Hybrid Approach

Evaluation of different approaches were done using Mean absolute error, Root mean squared error and Coefficient R^2values and described in Table.3.

Table.3. Performance analysis using Evaluation Metrics

Evaluation Metrics	Approach		
	RFR	SVR	Proposed Model
MAE	27.87	32.63	**26.30**
MSE	1552.26	2759.80	**1421.86**
RMSE	39.40	52.53	**37.71**
Coeff. R^2	0.85	0.73	**0.86**

5. CONCLUSION

Contaminants are volatile, dynamic, and highly variable in both place and time, predicting the quality of the air is a difficult task. Due to the crucial effects of air pollution on populations and the environment that have been seen, it is becoming more and more relevant to model and predict. AQI are widely used to assess the air based on the gases CO, NO2, SO2, PM2.5, PM10, Ozone and geographic parameters i.e., wind speed, wind direction, humidity, and temperature. At first, data has been collected from the Central Pollution Control Board website and then the collected Dataset has undergone pre-processing so as to deal with the missing values and redundant data which are then followed by features selection. In this work various gaseous input variables and climatic conditions were considered for AQI prediction. Then this paper proposed the combined air quality model based on RFR and SVR to predict the AQI. The Proposed model we gave is very effective in prediction, the performance of proposed hybrid model is compared with the classic ones, and it has been concluded that the proposed model has good prediction accuracy as compared to the existing ones. Evaluation was carried out using metrics and Value of Coefficient R^2 of hybrid proposed model is 0.859293403768103 which is better as compared to the others i.e., 0.8463898171811104 in case of RFR and 0.7268919074512392 in case of SVR.

REFERENCES

[1] R. Gupta and P. Singla, "An Analysis on Degrading Air Quality Index of Metropolitan Cities", *International Journal of Novel Research and Development*, Vol. 7, No. 6, pp. 647-656, 2022.

[2] R.K. Grace and S. Manju, "A Comprehensive Review of Wireless Sensor Networks Based Air-Pollution Monitoring Systems", *Wireless Personal Communications*, Vol. 87, pp. 2499-2515, 2019.

[3] R. Januszek, B. Staszczak, Z. Siudak, J. Bartus, K. Plens, S. Bartus and D. Dudek, "The Relationship between increased Air Pollution expressed as PM10 Consentration and the Frequency of Percutaneous Coronary Interventions in Patients with Accute Coronary Syndromes - A Seasonal Differences", *Environmental Science and Pollution Research*, Vol. 27, pp. 21320-21330, 2020.

[4] L. Gharibvand, D. Shavlik, M. Ghamsary, W.L. Beeson, S. Soret, R. Knutsen and S.F. Knutsen, "The Association between Ambient Fine Particulate Air Pollution and Lung Cancer Incidence: Results from the AHSMOG-2 Study", *Environmental Health Perspect*, Vol. 125, No. 3, pp. 378-384, 2017.

[5] R. Janarthanan, P. Partheeban, K. Somasundaram and P.N. Elamparithi, "A Deep Learning Approach for Prediction of Air Quality Index in a Metropolitan City", *Sustainable Cities and Society*, Vol. 67, pp. 1-11, 2021.

[6] L. Spinelle, M. Gerboles, M.G. Villani, M. Aleixandre and F. Bonavitacola, "Field Calibration of a Cluster of Low Cost Commercially available Sensors for Air Quality Monitoring. Part B: NO, CO and CO2", *Sensors and Actuators*, Vol. 238, pp. 706-715, 2017.

[7] Q. Wu and H. Lin, "A Novel Optimal- Hybrid Model for Daily Air Quality Index Prediction considering Air Pollutant Factors", *Science of the Total Environment*, Vol. 683, pp. 808-821, 2019.

[8] M. Mittal, L.M. Goyal, J.K. Sethi and D.J. Hemanth, "Monitoring the Impact of Economic Crisis on Crime in India using Machine Learning", *Computational Economics*, Vol. 53, pp. 1467-1485, 2019.

[9] J.K. Sethi and M. Mittal, "Ambient Air Quality Estimation Using Supervised Learning Techniques", *EAI Endorsed Transactions*, Vol. 6, No. 22, pp. 1-10, 2019.

[10] H. Liu, Q. Li, D. Yu and Y. Gu, "Air Quality Index and Air Pollutant Concentration Prediction based on Machine Learning Algorithms", *MDPI*, Vol. 4069, pp. 1-9, 2019.

[11] S. Bhattacharya and S. Shahnawaz, "Using Machine Learning to Predict Air Quality Index in New Delhi", *Proceedings of International Conference on Machine Learning*, pp. 1-7, 2021.

[12] M. Castelli, F.M. Clemente, A. Popovik, S. Silva and L. Vanneschi, "A Machine Learning Approach to Predict Air Quality in California", *Complexity*, Vol. 2020, pp. 1-23, 2020.

[13] V.A. Bhat, A.S. Manek and P. Mishra, "Machine Learning based Prediction System for Detecting Air Pollution", *International Journal of Engineering Research and Technology*, Vol. 8, No. 9, pp. 155-159, 2019.

[14] G. Corani and M. Scanagatta, "Air Pollution Prediction via Multi Label Classification", *Environmental Modelling and Software*, Vol. 80, No. 7, pp. 259-264, 2016.

[15] J. Fan, L. Wu, F. Zhang, H. Cai, X. Wang, X. Lu and Y. Xiang, "Evaluation the Effect of Air Pollution on Global and Diffuse Solar Rediation Prediction using Spot Vector Machine Modeling based on Sunshine Duration and Air temperature", *Renewable and Sustainable Energy Reviews*, Vol. 94, pp. 732-747, 2018.

[16] D. Qin, J. Yu, G. Zou, R. Yong, Q. Zhao and B. Zhang, "A Novel Combined Prediction Scheme Based on CNN and LSTM for Urban PM2.5 Concentration", *IEEE Access*, Vol. 7, pp. 20050-20059, 2019.

[17] B. Liu, Y. Jin and C. Li, "Analysis and Prediction of Air Quality in Nanjing from Autumn 2018 to Summer 2019 using PCR-SVR-ARMA combined Model", *Scientific Reports*, Vol. 348, pp. 1-14, 2021.

[18] J.K. Sethi and M. Mittal, "Prediction of Air Quality Index using Hybrid Machine Learning Algorithm", *Proceedings of International Conference on Machine Learning*, pp. 1-13, 2019.

[19] C. Wen, S. Liu, X. Yao, L. Peng, X. Li, Y. Hu and T. Chi, "A Novel Spatiotemporal Convolutional Long Short-Term Neural Network for Air Pollution Prediction", *Science of the Total Environment*, Vol. 654, pp. 1091-1099, 2019.

[20] J.K. Sethi and M. Mittal, "Monitoring the Impact of Air Quality on the COVID-19 Fatalities in Delhi, India: using Machine Learning Techniques", *Disaster Medicine and Public Health Preparedness*, Vol. 16, No. 2, pp. 604-611, 2022.