PREDICTION OF POLLUTANT CONCENTRATIONS BY METEOROLOGICAL DATA USING MACHINE LEARNING ALGORITHMS

K. Alpan [1], B. Sekeroglu [1]*

[1] NEU, Information Systems Engineering, 99138 Nicosia, TRNC, Mersin 10 Turkey - (kezban.alpan, boran.sekeroglu)@neu.edu.tr

KEY WORDS: Air Pollution, Prediction, Machine Learning, Pollutant Concentrations, Meteorological Data, Smart City

ABSTRACT:

Air pollution, which is one of the biggest problems created by the developing world, reaches severe levels, especially in urban areas. Weather stations established at certain points in countries regularly obtain data and inform people about air quality. In Smart City applications, it is aimed to perform this process with higher speed and accuracy by collecting data with thousands of sensors based on the Internet of Things. At this stage, artificial intelligence and machine learning plays a vital role in analyzing the data to be obtained. In this study, six pollutant concentrations; particulate matters ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), Ozone (O_3), and carbon monoxide (CO), were predicted using three basic machine learning algorithms, namely, random forest, decision tree and support vector regression, by considering only meteorological data. Experiments on two different datasets showed that the random forest has a high prediction capacity (R^2:0.74-0.86), and high-accuracy predictions can be performed on pollutant concentrations using only meteorological data. This and further studies based on meteorological data would help to reduce the number of devices in Smart City applications and will make it more cost-effective.

1. INTRODUCTION

Increasing air pollution has reached a level that may endanger human life. Even if developed countries applied different precautions against this threat, the increase in the number of vehicles in cities, forest fires, and especially industrialization still cause the release of harmful substances. Therefore, air quality indexes are followed up daily, and notifications or warnings are announced to people living in cities.

World Health Organization (WHO) announced that (World Health Organization, 2018) air pollution is a leading cause of chronic or noncommunicable diseases (NCDs), causing over one-third of deaths from stroke, lung cancer, and chronic respiratory disease, and one-quarter of deaths from ischaemic heart disease (Martinez et al., 2018).

With the new applications and development in the Internet of Things (IoT) in Smart City applications, instead of the data received from the weather stations deployed in a fixed location, thousands of more cost-effective and highly accurate sensors are designed and started to be implemented in the Smart Cities (Catlett et al., 2017).

Concentrations which are used as indicators of air pollution are nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), carbon monoxide (CO), and other small invisible particles that are named as particulate matter 2.5 ($PM_{2.5}$) and particulate matter 10 (PM_{10}). Data mining, artificial intelligence, and machine learning are of great importance at this point, as it is not possible to process the data obtained from thousands of sensors to be used to measure the levels of all these concentrations in Smart Cities.

Estimating air pollution in advance, analyzing data, and achieving desired information with less data will increase the efficiency of Smart Cities, and a cost-effective system can be created, minimizing the unnecessary sensors or data. For this reason, many researchers have conducted studies on air pollution and performed different experiments on different components.

Eslami et al. (Eslami et al., 2019) implemented a deep convolutional neural network (CNN) for real-time hourly ozone concentration prediction over the city of Seoul, South Korea. Several predictors (attributes) such as O_3, and NO_x concentrations, temperature, relative humidity, precipitation, etc. were considered as inputs for the model. They implemented a deep CNN with five convolutional layers (32 filters for each) and a fully connected layer with 256 neurons. They used several metrics such as mean absolute error (MAE), correlation coefficient (r), etc. to evaluate the results obtained, and the comparison was performed using other artificial neural networks (ANN) models as long-short term memory (LSTM), stacked autoencoder (SAE) and multi-layer perceptron (MLP). It was concluded that deep CNN outperformed other ANN models with r = 0.74–0.81; however, it was also noticed that the underprediction of the peak ozone was the limitation of the study.

Leong et al. (Leong et al., 2019) used Support Vector Machine (SVM) to predict the air quality index (AQI) of Malaysia instead of predicting the concentrations. They used the data from 2009 to 2014, collected at the air quality monitoring stations of two states of Malaysia, and they concluded that SVM with radial-basis function achieved an R^2 score of 0.9843.

Qadeer et al. (Qadeer et al., 2020) implemented the LSTM network to predict the $PM_{2.5}$ concentration over two cities in South Korea. They provided the past 24h data of 16 predictors as measured concentration values and meteorological measurements, to predict the next 1h of $PM_{2.5}$ concentration. They used four metrics, such as MAE, mean squared error

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIV-4-W3-2020-21-2020 | © Authors 2020. CC BY 4.0 License.

(MSE), etc. to evaluate the obtained results. The comparison was performed using five other models, and they concluded that the LSTM outperformed other models and capable of predicting the $PM_{2.5}$ concentration value more accurately.

Shishegaran et al. (Shishegaran et al., 2020) developed an ensemble model to predict the air quality index in Tehran, Iran. Daily air pollution data such as O_3, NO_2, $PM_{2.5}$, and PM_{10} for five years (2012-2016) and meteorological data such as radiation, visibility, pressure, wind speed and sunshine hours were considered in their study. They concluded that the proposed model was capable of predicting AQI and outperformed other considered models with an R^2 score of 0.983.

Su et al. (Su et al., 2020) proposed a method for predicting ozone O_3 concentration based on kernel extreme learning machine (KELM) and support vector machine regression (SVR). The pre-processing was applied using wavelet transformation (WT) and partial least squares (PLS). The meteorological data and hourly O_3 concentrations of the years 2014 – 2016 summer were considered in a city of China. Several metrics, such as mean absolute error, root mean squared error (RMSE), and coefficient of determination (R^2) was used to evaluate the proposed model. They concluded that the proposed method outperformed the backpropagation neural network and linear regression models in the comparison.

In addition to these studies, Aljanabi et al. (Aljanabi et al., 2020) performed another study using machine learning models for the prediction of ground-level ozone in Amman, Jordan. They concluded that the multi-layer perceptron outperformed other considered models using pre-processing filters. Zhu et al. (Zhu et al., 2020) proposed a hybrid model for air quality index prediction, and Ho et al. (Ho et al., 2020) proposed several models on data obtained by Airbox micro-sensors to predict ground-level $PM_{2.5}$ levels in Taiwan.

Most of the studies mentioned above proposed ML models to predict particular pollutant concentrations and air quality index using meteorological data and measured concentrations values together. A few of them were considered only meteorological data to predict the values of six primary pollutant concentrations. In this paper, we present the implementation of three ML algorithms to predict the values of six concentrations, namely, particulate matters ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), Ozone (O_3), and carbon monoxide (CO) by considering only the meteorological data to minimize the use of sensors or measurement devices for Smart City applications.

This paper is organized as follows: Section 2 introduces the considered dataset, machine learning models, and design of experiments in detail, and Section 3 presents the obtained results in this study. Discussions and conclusions of the study are presented in Section 4 and Section 5, respectively.

2. MATERIALS AND METHODS

This section presents the details of the considered dataset, used machine learning algorithms, and the design of performed experiments.

2.1 Dataset

The publicly available Beijing Multi-Site Air-Quality Data Dataset (Zhang et al., 2017) was used in this study. The data included the air quality and the meteorological data of twelve stations in Beijing, China, between March 1st, 2013 to February 28th, 2017. It consisted of 18 attributes and a total of 420,768 instances. The list of all attributes of the dataset is shown in Table 1. In this study, two air quality weather stations (Guanyuan and Wanshouxigong) of the Beijing Multi-Site Air-Quality Data Dataset were considered to perform preliminary experiments and to observe the ability for predicting the levels of concentrations separately that primarily cause air pollution, using meteorological data and time information.

Attributes	
Number	Station
Year	Pressure
Month	Dew Point Temperature
Day	CO
Hour	NO_2
Wind Direction	SO_2
Precipitation	PM_{10}
Wind Speed	$PM_{2.5}$
Temperature	O_3

Table 1. All attributes of the dataset

2.2 Machine Learning Algorithms

Three machine learning algorithms, namely a decision tree (DT), support vector regression (SVR), and random forest (RF), were considered in this study.

2.2.1 Decision Tree: The decision tree is a unique structure that is constructed using instances and attributes of the dataset. The constructed tree form starts with a root node, and decisions are performed in decision nodes. It can be used for both classification (Vanfretti, Arava, 2020) (Yilmaz, Sekeroglu, 2020) and prediction problems (Oytun et al., 2020). The main advantage of DT is the minimized computation time once it is constructed. However, it is possible to construct several DTs from a single dataset, and it is a challenging task to determine which tree is superior. For that reason, different algorithms, such as Gini, entropy, ID3, etc. were proposed to construct an optimal tree.

2.2.2 Support Vector Regression: Support vector regression is the improved version of support vector machines to accept and produce real-valued inputs and outputs (Ever et al., 2019). Input features are mapped into higher-dimension to provide a linear relationship that is not possible in lower-dimensions. This mapping process is performed by the kernels such as radial-basis function, quadratic, etc. used in SVR, and support vectors, which are the closest points to the data, are used to determine the regression line. It has been widely implemented in regression studies (Oytun et al., 2020) (Nourali, Osanloo, 2018).

2.2.3 Random Forest: The random forest was first proposed by Ho (Ho, 1995), and its' extended version was created by Breiman (Breiman, 2001). It constructs several individual decision trees using bagging and feature randomness, and its' aim is to create an uncorrelated forest. This provides more accurate predictions than any of the individual trees. It can be used for both classification and regression tasks (Wu et al., 2018) (Zhang et al., 2019).

2.3 Design of Experiments

Totally 36 experiments were performed for the data of two air quality weather stations (Guanyuan and Wanshouxigong) of the considered dataset using three ML algorithms described above. The aim was to predict the levels of concentrations separately that primarily cause air pollution, using meteorological data and time information.

Totally ten attributes, namely year, month, day, hour, temperature, pressure, dew point temperature, rain (precipitation), wind direction, and wind speed, were considered as the inputs of machine learning algorithms for each station, separately. Then, six concentration levels, as particulate matters ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), Ozone (O_3), and carbon monoxide (CO), were predicted.

Missing values of each station were removed from the dataset, and all instances were normalized between 0 and 1 using min-max normalization in order to minimize the complexity of the data and computational cost. The formula of min-max normalization is shown in Equation 1. Table 2 shows the details of the training data used in this study.

$$X_i = \frac{t_i - \min(t)}{\max(t) - \min(t)} \tag{1}$$

where X_i = normalized value
t_i = data point
min(t) = minimum value of the attribute
max(t) = maximum value of the attribute

Number of Training Instances	
Guanyuan WS	32,263
Wanshouxigong WS	32,768
Training Attributes	
Year	Pressure
Month	dew point temperature
Day	Rain (precipitation)
Hour	Wind Direction
Temperature	Wind Speed

Table 2. Properties of training data

Training was performed for each concentration separately using the 80% of instances for each weather station. The rest of the data (20%) were used in the testing phase. R^2 (coefficient of determination) and the mean-squared error (MSE) metrics were used to evaluate the prediction ability of the considered algorithms. R^2 score is a statistical measure of the predicted and observed points to determine how well the regression line fits data. A higher value (maximum=1) indicates the better-fitted data. Equation 2 shows the formula of the coefficient of determination (R^2).

$$R^2 = 1 - \frac{\sum O_i - \hat{O}_i}{\sum O_i - \bar{O}_i} \tag{2}$$

where O_i = observed data
$\hat{O}_i$ = predicted value
$\bar{O}_i$ = mean value of all observed data

Mean-squared error calculates the average of the squared errors, where the errors are obtained from the difference between the observed and the predicted values. The lowest MSE demonstrates the algorithm with the best prediction ability. Equation 3 shows the formula of MSE.

All parameters of the ML algorithms were decided by trial and error during the experiments, while there are no exact criteria to determine the parameters for the ML models.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(O_i - \hat{O}_i)^2 \tag{3}$$

where N = number of instances
O_i = observed data
$\hat{O}_i$ = predicted value

In the decision tree, the construction was performed using the MSE, and two samples were used to split the internal nodes.

The radial-basis function kernel was used in SVR, and the γ and ε values were initially set as 0.005 and 0.001, respectively. Then, the grid search was performed for optimal parameters, and the algorithm was tested using the optimal hyperparameters. In the random forest algorithm, 500 trees were used to build the forest. Mean Absolute Error criterion was used to construct the algorithms, and similar to DT, two samples were used to split the internal nodes.

3. EXPERIMENTAL RESULTS

As mentioned above, 36 experiments were performed for the data of two weather stations using three ML algorithms to predict six concentration levels.

3.1 Results of Guanyuan Weather Station Experiments

In the prediction of $PM_{2.5}$, which was performed on the first considered weather station (Guanyuan), the SVR produced the worst results both in terms of MSE and R^2 score. Even SVR obtained 0.0106 MSE value; it could not produce an acceptable R^2 score 0.2768. DT had better results than SVR in MSE and R^2 score 0.0045 and 0.6894, respectively); however, the optimal results were achieved by random forest for both MSE and R^2 scores (0.0020 and 0.8588, respectively).

In the prediction of PM_{10} of Guanyuan, it was observed that the decrement in the prediction levels occurred for all algorithms. Similar to PM_{10} experiments, SVR produced the lowest prediction results (MSE=0.0067 and R^2 score = 0.2362). It was followed by DT with higher scores, but the highest prediction results were obtained by RF for both evaluation metrics (MSE = 0.0017 and R^2 score = 0.7971).

In CO prediction, SVR increased its performance in terms of R^2 score when compared to the first two prediction experiments. However, even it produce relatively better results, it was not able to achieve the highest prediction levels. The MSE and R^2 scores of SVR were obtained as 0.0088 and 0.3512, respectively. DT achieved 0.0046 and 0.6623 MSE and R^2 scores, respectively. However, similar to other experiments, optimum results were achieved by RF (MSE = 0.0023 and R^2 score = 0.8306).

In NO^2 predictions, while the prediction levels of DT and RF decreased, a little increment in the performance of SVR was observed. However, similar results as in previous experiments were obtained, and worse and optimum results were achieved by SVR and RF, respectively. It should be noticed that even the highest prediction level was achieved by RF in NO^2 predictions, it was the lowest score that was achieved for Guanyuan data (MSE=0.0044 and R^2 score= 0.7442).

In ozone (O_3) prediction experiments, even SVR achieved its highest performance for Guanyuan data (MSE=0.0069 and R^2 score=0.6342); it could not produce optimum results. DT produced close but higher results than SVR for both MSE and R^2 score (0.0059 and 0.6862, respectively). Optimum results in O_3 predictions were achieved by RF (MSE=0.0027 and R^2 score=0.8538)

In the last experiment for Guanyuan city, which is the prediction of SO_2, similar results to all other experiments were obtained. SVR produced the lowest prediction results, followed by DT, and RF achieved the highest results. However, it should be noticed that the minimum MSE value for Guanyuan experiments was obtained in SO_2 prediction by RF (0.0012), even its R^2 score (0.7918) was not the highest for all the experiments.

Table 3 shows all obtained results for Guanyuan city weather station in detail. Figure 1 demonstrates the regression lines of RF for $PM_{2.5}$ and O_3.

	R^2 Score		
	DT	SVR	RF
$PM_{2.5}$	0.6894	0.2768	0.8588
PM_{10}	0.5784	0.2362	0.7971
CO	0.6623	0.3512	0.8306
NO_2	0.4616	0.4255	0.7442
O_3	0.6862	0.6342	0.8538
SO_2	0.6159	0.3195	0.7918
	MSE		
	DT	SVR	RF
$PM_{2.5}$	0.0045	0.0106	0.0020
PM_{10}	0.0037	0.0067	0.0017
CO	0.0046	0.0088	0.0023
NO_2	0.0093	0.0099	0.0044
O_3	0.0059	0.0069	0.0027
SO_2	0.0023	0.0041	0.0012

Table 3. Results Obtained for Guanyuan City Weather Station Data

3.2 Results of Wanshouxigong Weather Station Experiments

In the Wanshouxigong Weather Station Experiments, all ML algorithms produced closer prediction results for particulate matters, $PM_{2.5}$ and PM_{10}, in terms of MSE values. The highest and the worst MSE results were obtained by SVR (0.0077 for both $PM_{2.5}$ and PM_{10}), and followed by DT with 0.0033 and 0.0053 for $PM_{2.5}$ and PM_{10}, respectively. RF achieved the minimum MSE value for both fine particulate matters (0.0018 and 0.0024 for $PM_{2.5}$ and PM_{10}, respectively). Similar results to MSE values were obtained by the considered ML algorithms in terms of R^2 scores for both $PM_{2.5}$ and PM_{10} concentrations. SVR produced the lowest prediction levels (0.3124 and 0.2691 for $PM_{2.5}$ and PM_{10}, respectively). DT achieved higher scores than SVR (0.7008 and 0.5006 for $PM_{2.5}$ and PM_{10}, respectively); however, the optimum R^2 scores were achieved by RF (0.8362 and 0.7687 for $PM_{2.5}$ and PM_{10}, respectively).

In CO prediction, DT, SVR, and RF produced R^2 scores in the same order as in other experiments with 0.6389, 0.3887, and 0.8185, respectively. The algorithms also minimized errors as in the same success level of R^2 scores obtained, and the minimum error was achieved by RF (0.0030).

In the NO_2 prediction of Wanshouxigong Weather Station Experiments, DT obtained its minimum prediction levels in this study both in terms of MSE and R^2 score (0.0122 and 0.4535, respectively). SVR obtained closer results to DT for both MSE and R^2 scores (0.0123 and 0.4490, respectively), but it could not outperform any of the considered algorithms. Even also RF achieved its lowest prediction level in this study, it achieved the optimum results in NO_2 predictions (MSE = 0.0058 and R^2 score = 0.7411).

In ozone (O_3) prediction, all algorithms reached their highest prediction levels in terms of R^2 scores. However, the success rate of the algorithms did not differ, and the lowest and the highest prediction rates were achieved by SVR and RF, respectively. The RF produced 0.8654 of the R^2 score and 0.0034 MSE value.

In the last experiment of this study, SO_2 prediction was performed for Wanshouxigong Weather Station Dataset. The lowest prediction level was obtained by SVR and followed by DT as it was in all experiments of this study. The highest results for both MSE and R^2 scores were achieved by RF (0.0013 and 0.8301, respectively).

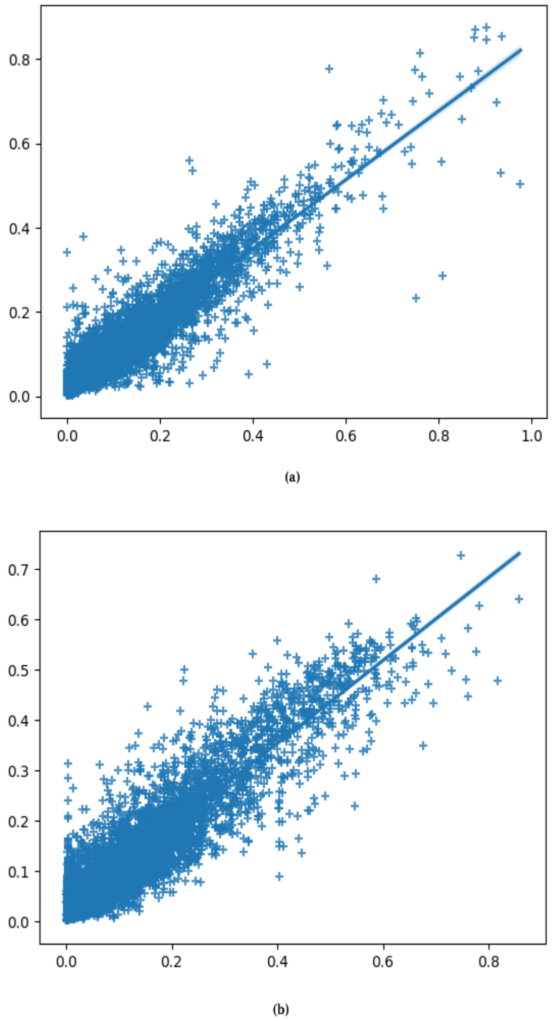
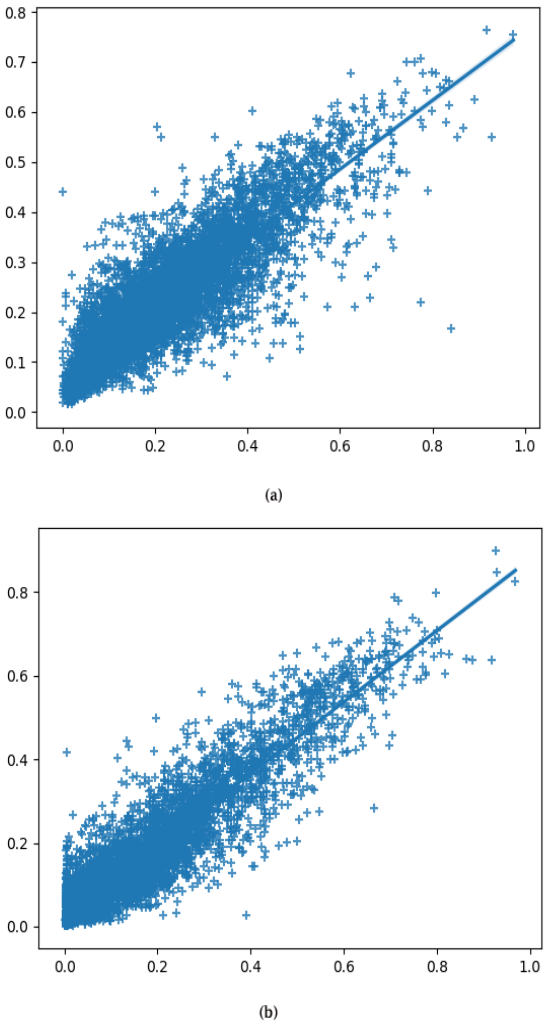


Figure 1. Regression lines of RF for Guanyuan City Weather Station Data (a) for $PM_{2.5}$ and (b) for O_3

Table 4 shows all obtained results for Wanshouxigong weather station in detail, and Figure 2 demonstrates the regression lines of RF for NO_2 and O_3.

	R^2 Score		
	DT	SVR	RF
$PM_{2.5}$	0.7008	0.3124	0.8362
PM_{10}	0.5006	0.2691	0.7687
CO	0.6389	0.3887	0.8185
NO_2	0.4535	0.4490	0.7411
O_3	0.7242	0.6524	0.8654
SO_2	0.6792	0.3303	0.8301
	MSE		
	DT	SVR	RF
$PM_{2.5}$	0.0033	0.0077	0.0018
PM_{10}	0.0053	0.0077	0.0024
CO	0.0060	0.0102	0.0030
NO_2	0.0122	0.0123	0.0058
O_3	0.0069	0.0087	0.0034
SO_2	0.0025	0.0053	0.0013

Table 4. Results Obtained for Wanshouxigong Weather Station Data



(a)

(b)

Figure 2. Regression lines of RF for Wanshouxigong Weather Station Data (a) for NO_2 and (b) for O_3

4. DISCUSSIONS

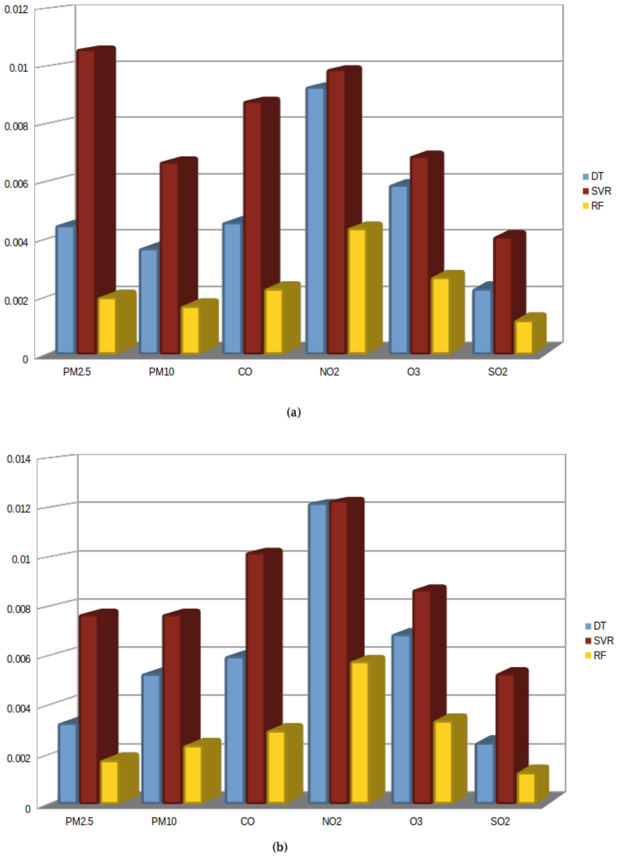
Three machine learning models were considered to predict six concentrations that indicate air quality in the cities, using ten meteorological and time attributes. The obtained results were evaluated using the MSE, which is sensitive to the outliers of the data, and the R^2 score that measures how well the algorithms' prediction line fits the observed data.

The results obtained in this study should be evaluated in two stages. The first is the comparison of machine learning algorithms considered in this study, and the second is to evaluate the optimal performances achieved on six concentrations.

When the MSE results were considered, it was observed that all considered machine learning algorithms were capable of minimizing the error between predicted and observed data. However, the obtained MSE results showed that the regression lines of SVR for all concentrations were not well-fitted while DT and RF were minimized MSE results more accurately for all experiments performed in this study. Even though the DT produced more steady and more accurate results than SVR, RF outperformed other considered algorithms in all experiments by minimizing the MSE.

The visualization of the obtained MSE results for the comparison of considered algorithms can be seen in Figure 3.

When R^2 scores were analyzed, similar results were observed for all algorithms. SVR could not produce accurate results to perform reliable predictions. DT produced more accurate results, and the optimum results were achieved by RF in all experiments for both weather stations.



(a)

(b)

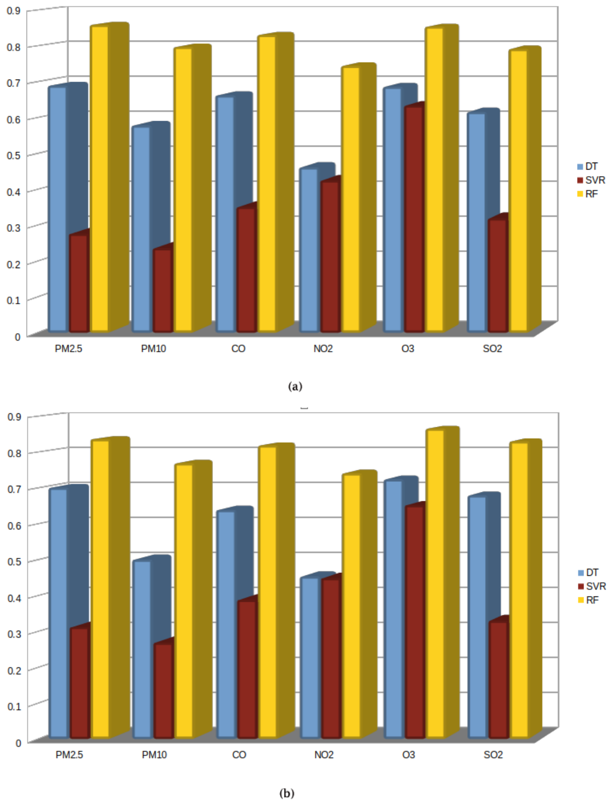
Figure 3. Visualization of MSE results obtained in all experiments, (a) for Guanyuan City Weather Station Data, and (b) for Wanshouxigong Weather Station Data

The general analysis of algorithms showed that fluctuated results occurred in SVR experiments, while the difference between the minimum and maximum R^2 scores of SVR obtained in this study was 0.4162. This difference decreased to 0.2707 and 0.1243 in DT and RF experiments, respectively. These results demonstrated the effectiveness and the steadiness of the RF for all kind of predictions considered in this study.

The visualization of the obtained R^2 scores for the comparison of considered algorithms can be seen in Figure 4.

The optimal results achieved by RF in all experiments showed that the NO_2 and PM_{10} were the least predictable concentrations for both datasets using the meteorological data. Increment of the data, the addition of new attributes, or implementing different ML algorithms not considered in this study may increase the prediction rates for these concentrations.

Besides, the prediction of SO_2 in the Guanyuan weather station dataset decreased to the prediction level of PM_{10}; however, a higher prediction level was achieved in the Wanshouxigong Weather Station Experiments dataset. This showed that the SO_2 level prediction is more sensitive to the data considered and may vary depending on the differences of meteorological data.
While $PM_{2.5}$, CO, and O_3 concentrations were predicted more consistently, especially the prediction levels of $PM_{2.5}$ and O_3 concentrations were at higher levels than other concentrations. This shows that the levels of the concentrations and, therefore, air quality index can be predicted using only meteorological data, and this would minimize the number of excessive sensors in smart cities. However, it should be noticed that this study was limited to the data collected by two weather stations, and further experiments are required to achieve more steady and generalized results.



(a)

(b)

Figure 4. Visualization of R^2 scores obtained in all experiments, (a) for Guanyuan City Weather Station Data, and (b) for Wanshouxigong Weather Station Data

5. CONCLUSION

Prediction of the pollutant concentrates values has crucial importance for human health as well as Smart City applications. More accurate predictions with minimized predictors led to minimized devices and the implementation of cost-friendly applications. In addition, simplified data would also be provided for both human researchers and interconnected devices.

Three basic machine learning algorithms, random forest, decision tree, and support vector regression, were implemented in this study to obtain preliminary results for the prediction of six primary pollutant concentrations by considering only meteorological data. Obtained results showed that random forest is capable of predicting concentrations with high R^2 scores (0.7411-0.8654), and there is a strong correlation between the pollutant concentrations and meteorological data.

Even the obtained preliminary results of this study are encouraging; further studies are required to improve the prediction levels to obtain more accurate results.

Future work will include the implementation of more machine learning algorithms such as neural network-based algorithms, and to consider more datasets from different weather stations. In addition, the most significant meteorological factors will be determined on pollutant concentrations using machine learning algorithms.

REFERENCES

Aljanabi, M., Shkoukani, M., Hijjawi, M., 2020: Ground-level Ozone Prediction Using Machine Learning Techniques: A Case Study in Amman, Jordan. *International Journal of Automation and Computing*, 1-11. https://doi.org/10.1007/s11633-020-1233-4

Breiman, L., 2001: Random Forests. *Machine Learning*, 45, 5-32.

Catlett, C., Beckman, P., Sankaran, R., Galvin, K., 2017. Array of things: a scientific research instrument in the public way: platform design and early lessons learned. *SCOPE '17: Proceedings of the 2nd International Workshop on Science of Smart City Operations and Platforms*, 26–33. https://doi.org/10.1145/3063386.3063771

Eslami, E., Choi, Y., Lops, Y., Sayeed, A., 2019: A real-time hourly ozone prediction system using deep convolutional neural network. *Neural Computing and Applications*, 1-15. 10.1007/s00521-019-04282-x

Ever, Y. K., Dimililer, K., Sekeroglu, B., 2019. Comparison of machine learning techniques for prediction problems. *Advances in Intelligent Systems and Computing*, 927, 713–723.

Ho, C., Chen, L., Hwang, J., 2020: Estimating ground-level PM2.5 levels in Taiwan using data from air quality monitoring stations and high coverage of microsensors. *Environmental Pollution*, 264, 114810.

Leong, W., Kelani, R., Ahmad, Z., 2019: Prediction of air pollution index (API) using support vector machine (SVM). *Journal of Environmental Chemical Engineering*, 8, 103208.

Martinez, R., Bueno-Crespo, A., Timon, I., Soto, J., Munoz, A., Cecilia, J., 2018: Air-pollution prediction in smart cities through machine learning methods: A case of study in Murcia, Spain. *Journal of Universal Computer Science*, 24, 261-276.

Nourali, H., Osanloo, M., 2018: Mining capital cost estimation using Support Vector Regression (SVR). *Resources Policy*. 62, 527-540. https://doi.org/10.1016/j.resourpol.2018.10.008

Oytun, M., Tinazci, C., Sekeroglu, B., Acikada, C., Yavuz, H. U., 2020: Performance Prediction and Evaluation in Female Handball Players Using Machine Learning Models. *IEEE Access,* 8, 116321-116335.

Qadeer, K., Rehman, W., Sheri, A., Park, I., Kim, H., Jeon, M., 2020: A Long Short-Term Memory (LSTM) Network for Hourly Estimation of PM2.5 Concentration in Two Cities of South Korea. *Applied Sciences*, 10, 3984.

Shishegaran, A., Saeedi, M., Kumar, A., Ghiasinejad, H., 2020: Prediction of air quality in Tehran by developing the nonlinear ensemble model. *Journal of Cleaner Production*, 259, 120825.

Su, X., Junlin, A., Zhang, Y., Zhu, P., Zhu, B., 2020: Prediction of ozone hourly concentrations by support vector machine and kernel extreme learning machine using wavelet transformation and partial least squares methods. *Atmospheric Pollution Research*, 11(6), 51-60.

Tin K. H., 1995. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition,* 1, 278–282.

Vanfretti, L., Arava, N., 2020: Decision tree-based classification of multiple operating conditions for power system voltage stability assessment. *International Journal of Electrical Power and Energy Systems*, 123, 106251.

World Health Organization, 2018. World Health Organization (WHO) air pollution programme. http://www.who.int/airpollution/en/.

Wu, Q., Zhong, R., Zhao, W., Song, K., Du, L., 2018: Landcover classification using GF-2 images and airborne lidar data based on Random Forest. *International Journal of Remote Sensing*, 40, 1-17.

Yilmaz, N., Sekeroglu, B., 2020. Student performance classification using artificial intelligence techniques. *Advances in Intelligent Systems and Computing*, 1095.

Zhang, J., Ma, G., Huang, Y., Aslani, F., Nener, B., 2019: Modelling uniaxial compressive strength of lightweight selfcompacting concrete using random forest regression. *Construction and Building Materials*, 210, 713-719.

Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., Chen, S., 2017. Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 473, 20170457.

Zhu, S., Sun, J., Liu, Y., Lu, M., Liu, X., 2020: The air quality index trend forecasting based on improved error correction model and data pre-processing for 17 port cities in China. *Chemosphere*, 252, 126474.