A Review of Machine Learning Models in the Air Quality Research

Huda W Ahmed[1], Dr Jameelah H Alamire[2]

[1] Institute of informatics for Postgraduate studies, Iraq.

[2] Computer Science Department, College of Science,

Mustansiriyah University, Iraq.

*Email: huda_wadah@yahoo.com ; dr.jameelahharbi@gmail.com

Abstract— **The mean objective of this research is to study and evaluate models of machine learning and the applications of it's that dealing with specific field which is the Air Quality and find the strengths of different machine learning and provide a background about it. Also, the research aims to emphasize the fundamental techniques of machine learning and their importance in improving the performance of predictions and the importance of input predictors in improving predictive accuracy. This study shows how merging of the machine learning techniques with prediction of the Air Quality is an effective and useful approach to solving some related environmental issues. In this study we produce two categories of the methods or models which are the single ML models and the hybrid models. The Single and hybrid Machine Learning was observed to offer better performance and higher accuracy in the Air Quality domain.**

Index Terms— **Air Quality (AQ), Machine Learning Applications, Machine Learning Models.**

I. INTRODUCTION

Number of Air Quality (AQ) models has been established in recent decades using an analytical deterministic technique because AQ plays a main role in urban environmental management and in the contemporary growth of smart cities [1][2][3]. The AQ models have been built by using the Machine Learning Techniques.

The study of Machine learning (ML) and its algorithms is scientific study and the statistical models which use the systems of computer to offer accurate predictions and take decision making without being specifically programmed for the carry out the task [4]. ML has obtained enormous fame for the rapid predictions and strong, especially with the huge data that require further real time for the analysis to obtain better understanding of their unseen values [5].

The important approaches used to predict or forecast or estimate the concentrations of pollutants based on Linear Regression, Neural Network, Vector Machine Support, or Ensemble Learning Algorithms. The number of European and American-based research articles is higher than that of Asia. Europe alone accounts for almost 40 per cent of the total jobs, followed by 33 per cent and 24 per cent respectively in America and Asia as shown in Fig. 1.

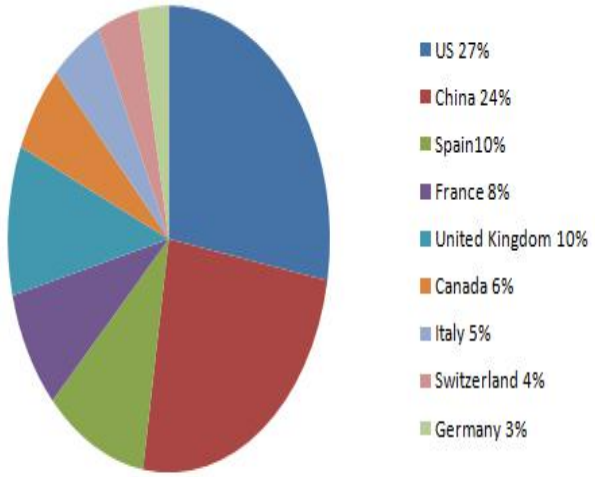

Fig. 1 Distribution of Air Quality Studies

The number of works that implemented the machine learning algorithms to air pollution modelling during the period 2013 to the period 2018 showing in the Fig.2. It is appears a continuously increasing trend in the number of studies published over the last three years (2016-2018) [6].

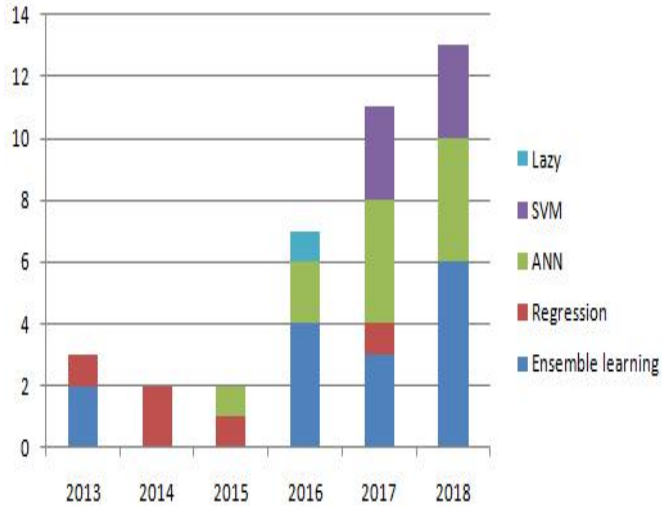

Fig. 2: Number of works which applied machine learning algorithms in Air pollution during 2013 to 2018

II. REVIEWING MACHINE LEARNING IN AIR QUALITY RESEARCH

The ML methods are investigated to advancing further by using the hybrid and the ensemble techniques. In this section, defined and examined the most common ML methods in the research of air quality. We focus and made two classifications of the methods in this research, i.e. single ML methods, and hybrid models.

A. Single Machine Learning

In this study we identified the popular single machine learning like the Support Vector Machine (SVM), Decision Tree , Random forest , Logistic regression ,Neural Network (NN), Naïve Bayes, k-means, k-nearest neighbors, …, etc. for the classification ,clustering and regression applications as show in the below Table 1.

Y. Rybarczyk et al**. [7]** were solved the problem of forecasting the fine particulate matter (PM2.5) in the light of a combination of weather conditions in their research. In city of Quito, Ecuador, several years of meteorological data and machine learning method are used to create a model. In this research introduce algorithm of Decision Tree that need to characterize the concentrations to the two classifications (> 15µg / m3 vs. < 15µg / m3), from the small number of the Parameters like precipitation level and wind speed and the direction of the wind. Models resulting involving few rules can predict the concentration result with good accuracy. The results of classification obtained with the decision tree are compared with five other popular classifier (rules.OneR, rules.ZeroR,.Naive Bayes, lazy.IBk and finally the functions.SMO) to compare any major variations in the classification between all the classifier models .The prediction of PM2.5 concentrations with Decision Tree based on a threshold value of 15µg/m3 and the relatively high rate of correctly classified instances (above 65 per cent) compared with other classification models.

R. Waman et al. [8] presents design a system for classifying the risks of healthy for air pollutants depend on AQI standards and emphasizes on air quality based on data from different air pollutants (NO2, SO2, CO and O3). Their research applies the algorithm of Naive Bayes and the algorithm of Decision Tree to forecast the health problem. Air Quality Index (AQI) classifications are: good, moderate, (unhealthy for sensitive groups), unhealthy, very unhealthy. For classifying the health risks for air pollutants depend on AQI standards which in this analysis, the classifier are: for AQI values for the range 0 to the 50, the level of risk is "GOOD," in the range of 51 to the 100 is "MODERATE " the range 101 to the 150 is "unhealthy for sensitive groups," the range 151 to the 200 is "UNHEALTHY," the range 201 to the

300 is "VERY UNHEALTHY" and over 300 the level is very dangerous. Show the result that the decision tree algorithm Provides an accuracy of 91.9978% which is more the algorithm of Naïve Bayes viz. 86.663%.

T. Walelign et al.[9] were proposed research of Air Pollution monitoring and predict system which allows with the aid of IoT devices to monitor the air quality. This system uses air sensors for detect and transfer the data to the microcontroller that is finally sends data to Web server. The Long Short Term Memory algorithm (LSTM) was introduced for prediction. It has a fast convergence and sufficient precision and decreases the training cycles. The appropriate parameters monitored remotely by uses the internet, and information gathered from sensors are saved in cloud and the evaluate drift is expanded to forecast pollution using the algorithms (LSTM) of machine learning.

O. Kisi et al.[10] their research discusses the forecasting of SO2 concentration by using three various soft computing approaches, the least square support vector regression (LSSVR)algorithm , multivariate adaptive regression splines (MARS), and M5 model tree (M5-Tree).
All the above models are implemented to the monthly data collected in Delhi, India, Nizamuddin ,Janakpuri, and finally Shahzadabad. By use the (RMSE) root mean square error, (MAE) mean absolute error, and the correlation coefficient, all the models are compared and evaluated with each other.
Based on the result of the comparison, LSSVR offered superior in the concept of accuracy than all the other models, whereas MARS model considered being second model in predict of the SO2. The results offered that all models provided better accuracy in the forecasting in Janakpuri station than the other stations.

C. Feng et al. [11] proposed estimation method based on the random forest algorithm for the fine-grained PM2.5 without any measurement devices for the PM2.5. They assess their work depend on the five data sources: which are meteorological and traffic data, records from the monitoring sites, POIs and finally the photos that are collected. The results of the work compared with other methods and appeared that they have high level of accuracy when implemented the random forest method on estimation of PM2.5 (the precision is : 87.5% and the recall is : 87.2%), which superior to the other methods (Logistic, Naïve Bayes, Random Tree, and BP ANN)

K. Hu et al.[12] suggested new Support Vector Regression (SVR) based on the model to predict the Air Pollution at fine spatial granularity for any hour and day in Sydney city. They use the historical data depend on the concentration (CO) readings that comes from the wireless sensor network and government monitoring sites. The SVR-based estimation model contains of five major steps.
They run tests and results compared to data based on estimates of both models ANN and SVR. The results showing that predict of air pollution from the estimation system that uses novel (SVR) have extreme resolution and are high accurate than predict from the Artificial Neural Network (ANN) model. The MAE and RMSE are the best in the SVR. In a fine spatial granularity (10,000 grids) the suggested model can estimate the pollution data more accurately than ANN.

K. Bashir et al.[13] presented the emphasis of their document is on the monitoring system and its forecasting module. In this study introduced three algorithms of machine learning (ML) examined to build predict models ahead of the ground level of the ozone (O3), the sulfur dioxide (SO2) and finally the nitrogen dioxide (NO2) levels. Such Machine Learning algorithms are M5P model trees, support vector machine and lastly the artificial neural networks (ANNs). In this study there are two kinds of modeling which are followed:
1- univariate just one gas concentration value has been used which is the target gas .
2-multivariate various features are implemented to assist in predicting future target gas values.
The measurements used for performance evaluation are firstly the accuracy and secondly the root means square error (RMSE) for the prediction. Their study shows that the use of various features with M5P algorithm in multivariate modeling returns the best performance for forecasting. M5P outperforms on the other algorithms for all gases. But at the other hand, ANN showed the worst results when working on a small dataset with more attributes leading to a complex network that overfiting the data due to its poor generalization ability, whereas SVM has the better results than of the ANN.

B. Hybrid Machine Learning

There are more sophisticated machine learning methods that are combinations together and which may include various methods as shown below in Table 2.

W. Tamas et al.[14] presents base approach which integrates Artificial Neural Networks (ANNs) and the clustering to detect the peaks of pollutants. The optimized air quality prediction models use the machine learning methods

that were implement 24 hours ahead to hourly concentrations of ozone (O3), nitrogen dioxide (NO2), and particulate matter (PM10). Multilayers Perceptron (MLP) was used by default, and then subsequently hybridized with the hierarchical clustering and a mixed of self organizing map and clustering of k-means. A first model use full training data set (FMLP) and test to provide good global accuracy (AI is 87% for the O3, 74% for the PM10 and 80% for the NO2). There were also two hybrid models built, which combined MLP and clustering methods. Two approaches for the clustering studied are the hierarchical clustering (HMLP) and the clustering of k-means coupled with Self Organization Map (KMLP). The precision of the Hybrid models evaluated with Index of Agreement (IA) which had weaker global precision in terms of (IA) but displayed better ability to properly predict high concentration events. This is the most important ability to forecast operational air quality. The clustering methods hierarchical and SOM / k-means seemed successful, depending on the situation. Its use typically increased the rate of identification of high pollution events relative to the traditional PM10 and O3 MLP. Nevertheless, in global results classical MLP still performed better than hybrid versions. As the clustering process reduces the size of the training set,
when more data is available, an improvement of hybrid models can be waited for. Hybrid models can be used to deal with high concentration events and classic MLP for the prediction of air quality regardless of high pollution.

J.C. Zapata-Hernandez et al.[15] proposed a methodology for forecasting (CAQE) which is the Critical Air Quality Events in the Aburr'a Valley depend on algorithm of the Support Vector Machines (SVM) developed with (PSO) Particle Swarm Optimization and a characterization scheme to evaluate current and past patterns of pollutants and weather behaviour, evaluating statistical behaviour at various time intervals. The authors offer three-stage approach composite of pre-possessing, characterizations and CAQE prediction. Current method provides better result with an error of 30% for ozone CAQE prediction. Because of low sensitivity among pollutants, further machine learning technique is required to guarantee robust behaviour in unbalanced data.

P. J. García Nieto et al. [16] in this work introduced a hybrid method of particle swarm optimization depend on support vector regression is introduced based on the experimental data set (specifically, nitrogen oxides, carbon monoxide, sulfur dioxide, ozone, and dust) to predict the air quality and this data collected in the Oviedo metropolitan area from 2013 through 2015. In addition, with multilayer perceptron network (MLP) and the model tree M5 the dataset are fitted for the comparison purpose. At last forecasted results indicate that suggested hybrid model was high robust in terms of test performance than the other models (MLP and M5 model tree).

W. Li et al.[17] solved the increasing environmental problems, a new hybrid model CI-FPA-SVM is used for the predict air concentration of PM2.5 and PM10 in two major cities in Yunnan Province, and China). The proposed model has two sections to it. Firstly because the lack of assessment of
possible correlation between various variables, is implemented the cointegration theory in order for obtain the input-output relationship and t gain the nonlinear dynamic system with SVM, so parameters (c, g) are developed by the (FPA) Flower Pollination Algorithm which is novel technique
of swarm intelligence (SI). The superiority of the suggested hybrid model is considered to be verified by uses six models for benchmark, including FPA-SVM, CI-SVM, CI-GA-SVM, CI-PSO-SVM, CI-FPA-NN and multiple linear regression models. The empirical results of their study show that the suggested model CI-FPA-SVM is better due to high predictive accuracy to all considered benchmark models.

III. CONCLUSION

The techniques of machine learning are popular and successfully used in all sciences fields. We especially handling in this study with the specific field which is the Air quality and in the recent decades improved different machine learning methods to obtain better performance. This study presents a review and analysis of both single model and the hybrid model. The review indicate that single ML methods are popular while hybrid models are combined various methods and continue for develop to the better precision and accuracy and higher performance. The different combinations of the hybrid methods can be offered most efficiently in dealing with the environmental issues related to air quality.

References	Contribution	Method	ML Application	Research domain
7	forecasting particulate matter (PM2.5) in the light of a combination of weather conditions	- Decision Tree - Rules.ZeroR - Rules.OneR - Naïve Bayes - Lazy.IBk - SMO	classification	- Machine learning - Forecast fine particulate matter (PM2.5)
8	Design system to classifying the risks of healthy for air pollutants based on AQI standards	- Decision Tree - Naive Bayes	classification	- Machine learning - classifying health risks for air pollutants
9	Air pollution monitoring and prediction system to forecast pollution rate.	- Long Short Term Memory (LSTM)	Regression	- Machine learning - Monitoring air pollution
10	forecasting of SO2concentration by using three various soft computing approaches	- least square support vector regression(LSSVR) - multivariate adaptive regression splines (MARS) - M5 model tree (M5-Tree)	Regression	- Machine learning - forecast of SO2 Concentration
11	depending on the Mobile Crowd Sensing estimate the Air Quality	- Random forest - Logistic - Naïve Bayes - Random Tree - BP ANN	Regression	- Machine learning - fine-grained PM2.5 estimation
12	Predict fine-spatial granularity of air pollution (CO) for any hour and day in the Sydney City	- Support Vector Regression (SVM)	Regression	- Machine learning - predict air pollution (CO)
13	Build accurate forecasting models to the (O3),(NO2), (SO2) with two kinds of modeling which are the univariate modeling and the multivariate modeling	- support vector machine (SVM) - M5P model trees - Artificial Neural networks (ANN).	Regression	forecasting models to the (O3),(NO2), (SO2)

Table 1. The top studies that identified popular Machine Learning in the Air Quality Field

Table 2. Studies of Hybrid Machine Learning methods in the Air Quality Field

Reference	Contribution	Method	ML Application	Research domain
14	Developed air quality prediction models by use the Artificial Neural Networks (ANNs) and clustering	-MLP -MLP and hierarchical clustering - MLP and K -means coupled with Self Organization Map (kMLP)	Clustering	- Air quality forecasting - Detect peaks of pollutants
15	Critical Air Quality Events (CAQE) are predicted	-Support Vector Machines(SVM) developed with the Particle Swarm Optimization (PSO)	Classification	- Prediction of Critical Air Quality event
16	predict the Air Quality	-particle swarm optimization based on evolutionary support vector regression (SVR)	Regression	- predict the Air Quality
17	introduce new hybrid of the CI-FPA-SVM to forecast the concentration of PM2.5 and the PM10	-(FPA) flower pollination algorithm with (SVM) support vector machine	Regression	-forecast concentration of the PM2.5 and PM10

REFERENCE

[1] Riffat, S., Powell, R., Aydin, D." Future cities and environmental sustainability. Future Cities Environ." 2,1 (2016).

[2] Webel, S." Forecasting Software that's a Breath of Fresh Air. Pictures of the Future Siemens Magazine" (2016). http://www.siemens.com/innovation/en/home/pictures-of-thefuture/infrastructure-and-finance/smart-cities-air-pollution-forecasting-models.html

[3] K . Karatzas, N . Katsifarakis, C . Orlowski , and A . Sarzyński, "Urban Air Quality Forecasting: A Regression and a Classification Approach", Springer, pp. 539–548, 2017.

[4] C.M. Bishop, "Pattern Recognition and Machine Learning"; Springer: New York, NY, USA, 2006.

[5] Machine Learning Algorithms. Available online: https://www.packtpub.com/big-data-and-businessintelligence
/machine-learning-algorithms-second-edition (accessed on 9 September 2019).

[6] A. Masih, "Machine learning algorithms in air quality modeling", Global Journal of Environmental Science and Management (GJESM), 5(4): 515-534, autumn 2019.

[7] Y. Rybarczyk and R. Zalakeviciute, "Machine Learning Approach to Forecasting Urban Pollution", IEEE, 2016.

[8] R. Waman Gore and D. S. Deshpande,"An Approach for Classification of Health Risks Based on Air Quality Levels", IEEE, 2017

[9] T. W. Ayele , R. Mehta ," Air pollution monitoring and prediction using IoT", IEEE, 2018.

[10] O. Kisi , K. S. Parmar , K. Soni and V. Demir ,"Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, andM5 model tree models", springer,2017

[11] C. Feng, W. Wang, Y. Tian, X. Que and X.Gong, "Estimate Air Quality Based on Mobile Crowd Sensing and Big Data",IEEE,2017.

[12] K. Hu, V. Sivaraman, H. Bhrugubanda, S. Kang, A. Rahman, "SVR Based Dense Air Pollution Estimation Model Using Static and Wireless Sensor Network", IEEE, 2016.

[13] K. B. Shaban, A. Kadri and E. Rezk ,"Urban Air Pollution Monitoring System With Forecasting Models",IEEE,2016.

[14] W. Tamas, G. Notton, C. Paoli1, M . Nivet, C. Voyant , "Hybridization of Air Quality Forecasting Models Using Machine Learning and Clustering: An Original Approach to Detect Pollutant Peaks, *Aerosol and Air Quality Research"*, 16: 405–416, 2016.

[15] J.C. Zapata-Hernandez, Y.K Rojas-Idarraga1, D.A Orrego and J. Murillo-Escobar, "Prediction of Critical Air Quality Events Using Support Vector Machines and Particle Swarm Optimization", Springer ,2017.

[16] P. J. García Nieto , E. García-Gonzalo1 , A. Bernardo Sánchez and A. Rodríguez Miranda ,"Air Quality Modeling Using the PSO-SVM-Based Approach ,MLP Neural Network, and M5 Model Tree in the Metropolitan Area of Oviedo (Northern Spain)", Springer , 2017.

[17] W. Li, D. Kong and Jinran W , "A New Hybrid Model FPA-SVM Considering Cointegration for Particular Matter Concentration Forecasting: A Case Study of Kunming and Yuxi, China", Hindawi Computational Intelligence and Neuroscience , 2017.