Urban Air Pollution Monitoring System With Forecasting Models

Khaled Bashir Shaban, *Senior Member, IEEE*, Abdullah Kadri, *Member, IEEE*, and Eman Rezk

***Abstract*— A system for monitoring and forecasting urban air pollution is presented in this paper. The system uses low-cost air-quality monitoring motes that are equipped with an array of gaseous and meteorological sensors. These motes wirelessly communicate to an intelligent sensing platform that consists of several modules. The modules are responsible for receiving and storing the data, preprocessing and converting the data into useful information, forecasting the pollutants based on historical information, and finally presenting the acquired information through different channels, such as mobile application, Web portal, and short message service. The focus of this paper is on the monitoring system and its forecasting module. Three machine learning (ML) algorithms are investigated to build accurate forecasting models for one-step and multi-step ahead of concentrations of ground-level ozone (O_3), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2). These ML algorithms are support vector machines, M5P model trees, and artificial neural networks (ANN). Two types of modeling are pursued: 1) univariate and 2) multivariate. The performance evaluation measures used are prediction trend accuracy and root mean square error (RMSE). The results show that using different features in multivariate modeling with M5P algorithm yields the best forecasting performances. For example, using M5P, RMSE is at its lowest, reaching 31.4, when hydrogen sulfide (H_2S) is used to predict SO_2. Contrarily, the worst performance, i.e., RMSE of 62.4, for SO_2 is when using ANN in univariate modeling. The outcome of this paper can be significantly useful for alarming applications in areas with high air pollution levels.**

***Index Terms*— Air quality monitoring, forecasting, wireless sensors network, and machine learning algorithms.**

I. Introduction

IT is widely believed that urban air pollution has a direct impact on human health especially in developing and industrial countries, where air quality measures are not available or minimally implemented or enforced [1]. Recent studies have shown substantial evidences that exposure to atmospheric pollutants has strong links to adverse diseases including asthma and lung inflammation [2]–[4]. In terms of economic impact, the association between air pollution and human health inevitably results in increase of healthcare services costs in terms of hospital admissions and emergency room visits [5]. Considering the significance of air quality on human lives, the World Health Organization (WHO) has developed guidelines for reducing the health effects of air pollution on public health by setting the limits of the concentrations of various air pollutants, some of which are ground–level ozone (O_3), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2) [6]. Traditionally, the concentrations of air pollutants are measured using air quality monitoring (AQM) stations that are highly reliable, precise, accurate, and are able to measure a wide spectrum of pollutants using standardized analyzers. Different types of these gas analyzers are provided by various vendors such as Thermo Scientific,[1] envirocon instrumentation,[2] and CEREX.[3] However, these stations have three main drawbacks: 1) the significant infrastructure needed for installation due to their bulky size, 2) the complicated operational requirements, e.g. access to grid power, heating/cooling, and secure shelters, and 3) the prohibitive costs of acquiring, setting up, and performing regular maintenance and calibration. These drawbacks reduce the number of installations and result in sparsely distributed AQM networks with limited spatial resolution air pollution data [7].

Recently, the landscape of traditional AQM networks is being changed due to advances in sensing and monitoring technologies. The trend is moving towards the employment of the Next Generation of Air Monitoring (NGAM) that has the potential to complement the traditional AQM stations with small sized and inexpensive AQM motes that incorporate an array of gaseous sensors. The aim of this configuration is to move from meso–scale to micro–scale coverage that drastically improves the spatiotemporal resolution of the collected air pollution data. Consequently, NGAM can help reduce the costs of AQM networks, improve the public health by providing communities with better air pollution data [7].

Recent development of electronics has realized the vision of using wireless communication in devices used for monitoring wide range of real life parameters, such as temperature, pressure, and air pollution. These devices send their measurements wirelessly to a database hosted on a remote server for further processing and analysis [8]. The concept of using small size, inexpensive AQM motes that wirelessly communicate their air pollution measurements has been widely studied and implemented [9]–[12]. In [9], it is shown that a miniature,

Manuscript received November 3, 2015; revised December 21, 2015; accepted December 23, 2015. Date of publication January 4, 2016; date of current version February 24, 2016. The associate editor coordinating the review of this paper and approving it for publication was Dr. Themis Prodromakis.

K. Bashir Shaban is with Qatar University, Doha 2713, Qatar (e-mail: khaled.shaban@qu.edu.qa).

A. Kadri is with the Qatar Mobility Innovations Center, Doha 210531, Qatar (e-mail: abdullahk@qmic.com).

E. Rezk is with the College of Engineering, Qatar University, Doha 2713, Qatar (e-mail: er1201046@qu.edu.qa).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSEN.2016.2514378

[1] http://www.thermoscientific.com/

[2] http://www.envirocon.co.za/

[3] http://cerexms.com/

low-cost AQM devices based on electrochemical gas sensors can be used for urban AQM in the range of parts-per-billion (ppb) levels when suitably configured and operated. In [10], the authors proposed a distributed air pollution monitoring system that uses General Packet Radio Service (GPRS) modem to send the gathered air pollutants levels. A network based on smart sensors communicating using wireless local area network for air quality monitoring applications is presented in [11] and [25]. This network uses artificial neural networks (ANN) to study the effects of temperature and humidity on pollutants concentrations. In [12], a wearable wireless sensor system is developed for real-time monitoring of volatile organic compounds. This system uses Bluetooth interface for connectivity with the database. More similar work can be found in [13]–[17].

The work presented in this paper focuses on presenting a pilot NGAM, and on the development of accurate forecasting models for predicting future average concentrations of some urban air pollutants, namely: O_3, NO_2, and SO_2, all of which are mentioned as being harmful in the WHO's guidelines. Three machine learning (ML) algorithms are investigated, i.e. support vector machines (SVM), model trees (M5P), and ANN, with two types of time series data modeling: univariate and multivariate. These algorithms produce models that are considered as nonlinear estimators, with good predictive and generalization abilities, which are successfully applied in various fields. The performance of the forecasting models is measured using two metrics: prediction trend accuracy and root mean square error. The obtained results show that most ML algorithms forecasting performance is enhanced when multivariate modeling is used.

The rest of the paper is organized as follows: Section II presents an overview of the ML approach, and reviews related techniques. The AQM system design and architecture is presented in Section III. Section IV explains the experiments of univariate and multivariate ML-based modeling for forecasting of gas concentrations. The analysis and discussion of results are given in Section V. Section VI concludes the paper by highlighting findings and future extensions of the work.

II. Machine Learning Approach

ML involves computational methods that improve the performance of mechanizing the acquisition of knowledge from experience [18]. Machines learn from complex data to be able to solve problems, answer questions and be more intelligent. One of the tasks that highly involve learning is forecasting, in which the forecasting model is built through training from data that is generally nonlinear in the case of air quality [30]. Therefore, approached based on linear modeling may not be suitable for such data [31]. After training, the model is ready to predict unseen data and hence can answer the forecasting question, for example: "What will be the next hour value of NO_2 gas concentration in air?"

Specifically, we aim to accurately predict concentrations of O_3, NO_2, and SO_2 as they are considered to be the most harmful gases [6]. Before employing the nonlinear modeling methods, the nonlinear structure of the data is verified for all gases. Here, Brocke-Decherte-Scheinkman (BDS) method, proposed in [32], is used. The BDS statistic, $\omega_{\mathrm{m,n}}(\varepsilon)$, is computed and the nonlinearity in data is verified if the null hypothesis of linearity is rejected at the 5% significance level. This condition is applicable if $|\omega_{\mathrm{m,n}}(\varepsilon)| > 1.96$. If the time series data comprises more than 7500 observations, as in our case, the BDS statistic is derived in terms of the correlation integral, $\mathrm{c_{m,n}}(\varepsilon)$, using the formula:

$$\omega_{m,n}(\varepsilon) = \sqrt{n}\frac{c_{m,n}(\varepsilon) - c_{1,n}^{m}(\varepsilon)}{\sigma_{m,n}(\varepsilon)} \tag{1}$$

where n is the sample size (8832 observations for all gases, in our case), m is the embedding dimension and it takes a discrete value in the range [2]–[5] at the big sample size. σ is the standard deviation of time series and ε takes a recommended value in the range from 0.5σ to 2σ based on the assumption that samples have normal or near-normal distribution.

For our samples, $\omega_{\mathrm{m,n}}(\varepsilon)$ is computed, by equation (1), for SO_2, NO_2, and O_3 are found to take values in the ranges of [207.36-266.12], [140.72-176.42], and [191.43 -224.35], respectively. These values are extremely greater than 1.96, which reveal the sharp nonlinearity in data.

Our methodology consists of the following steps:

1. *Data Preprocessing:* here, data are cleaned, such as by removing outliers and anomalies. Data are also prepared to be in proper format for the ML algorithms to use.
2. *Feature Engineering:* this step is concerned with selecting the features to be included in the prediction process along with each target gas, such as temperature, humidity, and day of the week.
3. *Time Windowing:* this is a fundamental task with time series forecasting, in which a number of time-lagged features for each input attribute is generated in order clarify the time dependency between consecutive data points. Window size, step size and horizon are key parameters that control time windowing. Window size is the number of generated features (i.e., generating multi-dimensional vectors) from the single-dimensional data. Step size is the number of instances between windows. Horizon is the number of steps in the future to be forecasted.
4. *Building Forecasting Models:* in this step, models are learned to be used to assign future values to a target feature of unseen data instances based on historical data values of discriminant features.

The process encompassing the above steps to construct and apply ML-based models for predicting values of unseen target data is depicted in Fig. 1.

In training, data with known target values are collected; a subset of feature is selected, and then used to construct a forecasting model. There are many subsets of features selected and various ML algorithms used; therefore, there are various predictors that can be trained.

In testing, the produced models from the training phase are validated and evaluated. Several methods are used in model validation, such as different sliding windows, in which two windows are used for training and testing and each has its own size, step size, and horizon. This validation method guarantees that instances used for testing are not known before to the

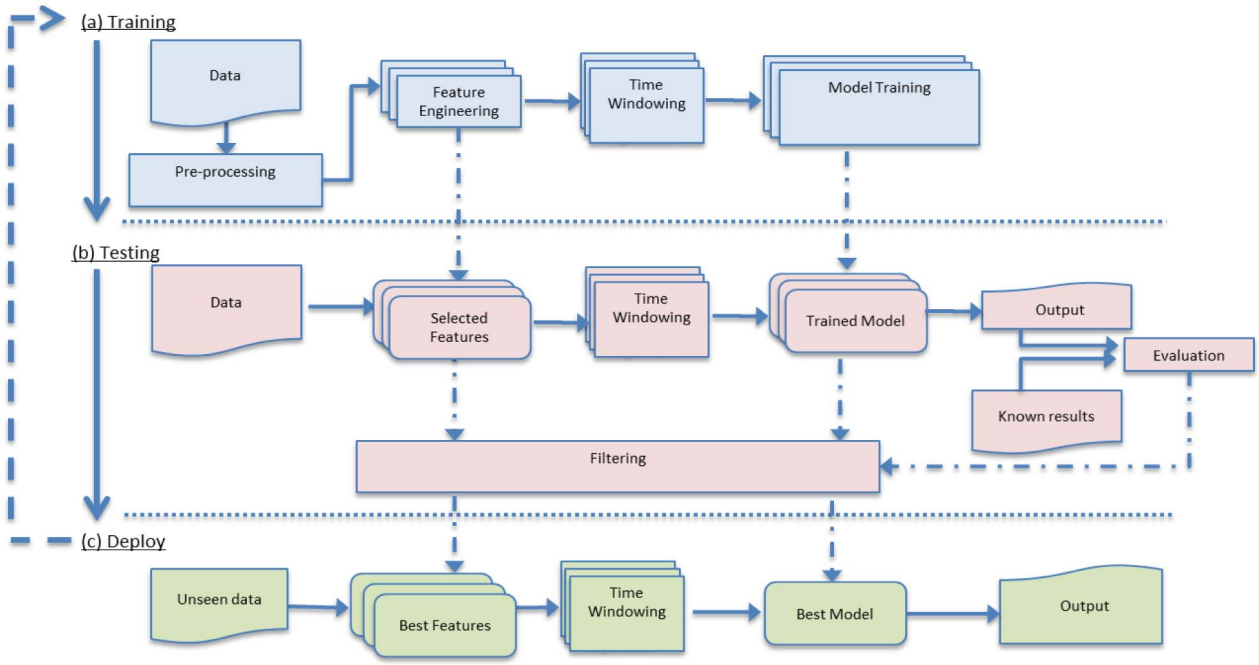


Fig. 1. Iterative process of constructing and applying ML-based prediction models.

model through training, hence reliable performance measures are calculated such as prediction trend accuracy (PTA) and root mean square error (RMSE). *PTA* is a time series measurement of how close is the predicted data trend from the trend of the actual data.

First, actual trend (AT) and predicted trend (PT) are calculated as:

$$AT = Label\,[i] - Label\,[i - horizon] \quad (2)$$

$$PT = Predicted\,[i] - Label\,[i - horizon] \quad (3)$$

where *Label* is the target feature, i is the instance number, and *horizon* is the number of steps forecasted in the future.

Trends are then multiplied by each other. If the result is greater than or equal to zero, then the actual and predicted trends have the same sign, hence have the same trend, so a counter is incremented. This process is repeated for all data, and finally, the counter is divided by the total number of instances.

RMSE is a common performance metric in model evaluation, and it is calculated as [29]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2)} \quad (4)$$

where n is the number of instances, y is the actual value of the target feature and $\hat{y}$ is the predicted value of it.

Normalized RMSE (NRMSE) is used to compare the performance of different models predicting different target variables and it is calculated as [30]:

$$NRMSE = \frac{RMSE}{(y_{max} - y_{min})} \quad (5)$$

where y_{max} and y_{min} are the maximum and minimum values of collected data respectively.

In the deployment phase, the best model and features will be used to process unseen data and produce prediction results. The model performance is kept on check to validate its prediction results. Practically, and especially in changing environments, the process of training, testing, and deployment are periodically repeated to maintain high accuracy of results. Moreover, this iterative process can be performed to improve performance of the models as historical data become increasingly available.

There exists a plethora of algorithms to build ML-based forecasting models that may behave differently to the given data. Among the most successfully used models are SVM, ANN, and M5P.

SVM is a supervised learning method that can be used for solving classification and regression problems. Our utilization of SVM is in regression. SVM regression aims to find an approximation to a non-linear function that maps the input data into high dimensional space. In this space, a hyper-plane is constructed in a way that it separates the data points with maximal margin with linear regression [19].

ANN is a network of nodes connected via different layers; input layer, hidden layer(s), and output layer. In a feed forward neural network, input data are fed to the nodes in the input layer, and then it is propagated through the network passing by hidden layer nodes and then to the output layer [20]. The input to every node in the hidden layer is the sum of all input values transferred from all connections multiplied by its weight. The value of each hidden node is calculated as the activation function of total weighted input of that node [21]. The activation function used is the sigmoid function. ANN uses a back-propagation algorithm to train the network. In this phase, the output of the network is compared to the actual correct

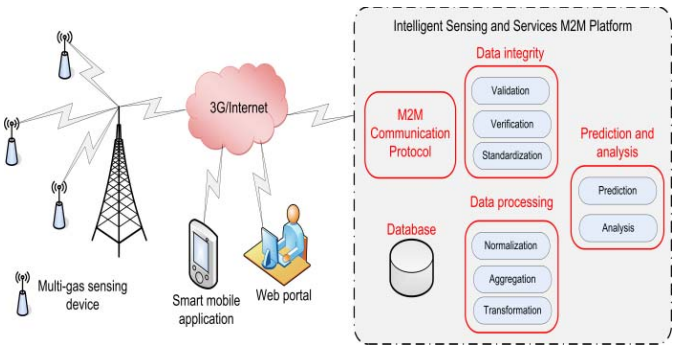


Fig. 2. The architecture of the AQM system.

output and an error is calculated. This error is propagated from the output back to the hidden and input layer and adjusting the weights of each connection in the network layers. This procedure is repeated to minimize the error.

M5P is a model tree algorithm that combines trees and regression models. It has the structure of tree with linear regression models at the leaves. These regression models ae linear multivariate regression equations that can be solved to find the predicted values [22]. M5P integrates the key benefits of tress and regression equations. Trees are usually much larger and more complex than the regression equations, but they are more accurate. However, trees are cumbersome and difficult to interpret because of its large size. That is why model trees combine the accuracy of trees and simplicity of regression models [23].

III. System Design and Architecture

The architecture of the AQM system used for data collection, communication, storage, processing and analysis is shown in Fig. 2. The AQM system, a pilot initiative of Qatar Mobility Innovations Center (QMIC), consists of multi-gas sensing (MGS) devices and an intelligent sensing and services machine-to-machine (M2M) platform.

A. Multi-Gas Sensing Devices

MGS devices consist of several gas and meteorological sensors, in addition to the data logging, communication, and controlling boards. Each device can house up to four gas sensors where the selection of the combination of the gas sensors depends on the installation location and the purpose of the monitoring application. All of these gas sensors are based on either electrochemical or metal-oxide semiconductor sensing technology; similar sensors of these technologies are reviewed in [24].

These sensors are exposed to a sample of the ambient air every 15 minutes where each sensor generates an electrical analog signal proportional to the amount of the gas exists in that sample. The electrical signals are then fed to the analog inputs of the controlling board that converts these analog outputs, measured in volte, to digital values that are mapped to the concentrations of the monitored gases. Accuracy of the sensors is $\pm 10\%$ and resolution is within 20 ppb.

The values are then stored in the data logging board and transmitted wirelessly using GPRS protocol to the central

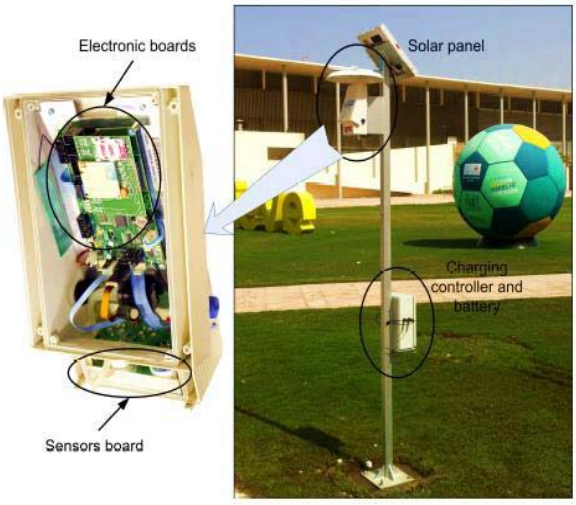


Fig. 3. Multi-gas sensing device.

platform through the GPRS modem. In addition to the gas sensors, each device is equipped with temperature, relative humidity, and wind speed and direction sensors. Finally, these devices are powered by solar energy system composed of a solar panel, a battery, and a charging controller. The battery's voltage level is sent along with the data every 15 minutes. The intelligent platform has a mechanism that sends alert should this level goes below a threshold due to any environmental conditions, such as dust accumulation on the solar panel. A picture of one of the current installations is shown in Fig. 3.

B. Intelligent Sensing and Services M2M Platform

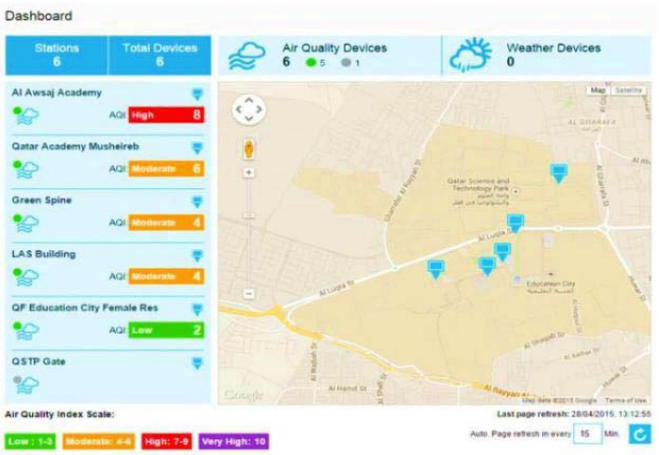
This platform operates in a backend server located at QMIC premises. The platform consists of five modules: 1) the M2M communication protocol, 2) data integrity, 3) data processing, 4) prediction and analysis, and 5) database management.

The main function of M2M communication module is to connect with all multi-gas sensing devices for data transfer using TCP/IP protocol. The data integrity module is responsible for handling missing, erroneous, and noisy data and passes the clean data in the database management module for storage. Data processing module applies statistical operations, such as hourly, daily, and monthly averaging, on the stored data in order to present it in a friendly manner to the user by the prediction and analysis module. Fig. 4 (a) and (b) show screenshots of the user interface of the module, where (a) shows the dashboard indicating air quality index (AQI) of each station as located in the map and calculated as in [25] and (b) shows a plot of the wind speed over time. In addition, this module runs the trained forecasting models to predict, the hourly, 8-hour, 12-hour, and 24-hour averages of each gas as per the adopted regulation.

IV. Univariate and Multivariate ML-Based Forecasting

To carry out experiments, historical data are utilized from one MGS device station during three months from June to August of 2013. This station measures O_3, NO_2, SO_2, H_2S,

(a)



(b)

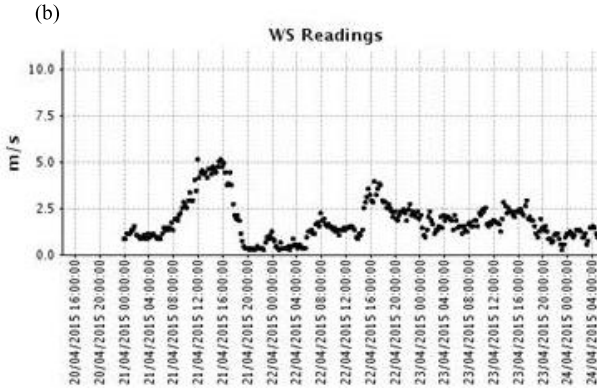


Fig. 4. User interface screenshots. (a) Dashboard. (b) Detailed wind speed measurements.

temperature, humidity, and wind speed. The measurements are taken and sent to the backend server every 15 minutes. The data size is 8833 readings that are processed through the ML steps explained in Section II. In our work, we focus on the prediction of O_3, NO_2, and SO_2 because they are used to calculate the AQI of the station [28], and are identified as the most harmful gases by WHO [6]. Specifically, the steps taken are as follows:

1. *Data Preprocessing*: in this step, the erroneous and missing data are estimated and replaced with new data points using interpolation process. The historical data used in these experiments has less than 0.5% erroneous data and less than 0.45% missing data. In addition to these processes of the data integrity and processing modules, the collected data are aggregated and averaged per hour to reduce the data size (from 8833 instances to 2208), and hence lower the computational cost of training the prediction models, while preserving trends of the data.
2. *Feature Engineering*: the included features in the experiments are of two kinds depending on the modeling type:
   a) *Univariate Modeling:* in this setting, only the concentration value of one gas is used, i.e. the target gas.
   b) *Multivariate Modeling:* here, different features are incorporated, to help predicting the target gas future values, including:

H	O_3
0	25.3
1	25.8
2	25.0
3	20.5
4	19.8
5	15.0
6	14.8
7	15.3
8	**17.5**

L	O_3-7	O_3-6	O_3-5	O_3-4	O_3-3	O_3-2	O_3-1	O_3-0
17.5	25.3	25.8	25.0	20.5	19.8	15.0	14.8	15.3
19.5	25.8	25.0	20.5	19.8	15.0	14.8	15.3	17.5
19.3	25.0	20.5	19.8	15.0	14.8	15.3	17.5	19.5
17.8	20.5	19.8	15.0	14.8	15.3	17.5	19.5	19.3
17.5	19.8	15.0	14.8	15.3	17.5	19.5	19.3	17.8
15.5	15.0	14.8	15.3	17.5	19.5	19.3	17.8	17.5
11.3	14.8	15.3	17.5	19.5	19.3	17.8	17.5	15.5
9.3	15.3	17.5	19.5	19.3	17.8	17.5	15.5	11.3
14.5	17.5	19.5	19.3	17.8	17.5	15.5	11.3	9.3

Fig. 5. Windowing example.

   i) Temporal features such as, the hour of the day and the day of the week as they are indicative of traffic volume which directly increase pollutants in air.
   ii) Meteorological features such as, temperature, humidity, and wind speed. These measures play roles in the formation of certain pollutants such as O_3 [1], [26].
   iii) Measured gases collected from the same sensor because some gases are byproducts of other gases' interactions such as SO_2 results from interaction of H_2S and O_2 [27].
3. *Time Windowing:* windowing of size 24, 8 steps, and horizons of 1, 8, 12, and 24 are applied. This generates 24 time-lagged features and moves every 8 hours to start generating new instances. It allows training the forecasting models to predict average gas counteractions for the next 1, 8, 12 and 24 hours ahead. An example of a univariate time windowing on O3 is shown by
4. Fig. 5, in which the window size (WS) is 8, and step size and horizon is 1, H is for hour, L is for label. Eight time-lagged features from O3-0 to O3-7 are created to be one instance in the windowed data on the right side of the figure. A horizon of 1 denotes that one-step in the future is to be predicted, i.e. setting the very next one value of O3 as a label for the corresponding instance. Finally, the step size is set to 1 in order to move one step for the next window. Each data point dt_i in the original data is mapped to dt_{jk}, where i is the row index and has values from 0 to n (original data size), j is the row index of the windowed data and has values from 0 to $\frac{n}{WS}$, and k is the column index that has values from 0 to WS-1.
5. *Model Training:* the collected, preprocessed data and engineered features are used to train prediction models based on ML algorithms including ANN, SVM, and M5P. The algorithms are optimized through parameter tuning in order to produce the most reliable and accurate models.
6. *Model Testing:* the produced models are validated using sliding window validation method with training window of size 30, testing window of size 30, step of size 1, and horizon of size 1. These parameters are optimized to achieve best prediction accuracy while requiring reasonable processing time.

TABLE I
SUMMARY STATISTICS

Feature (unit)	Min	Max	Avg	SD
H_2S (ppb)	0.00	135.54	19.51	13.51
NO_2 (ppb)	0.00	199.94	86.78	29.37
O_3 (ppb)	0.00	55.60	23.34	6.70
SO_2 (ppb)	0.00	700.31	117.49	112.53
Temperature (Cel.)	0.00	51.17	37.28	5.19
Humidity (%)	0.00	97.94	44.96	23.03
Wind Speed (m/sec)	0.49	11.75	3.16	2.11

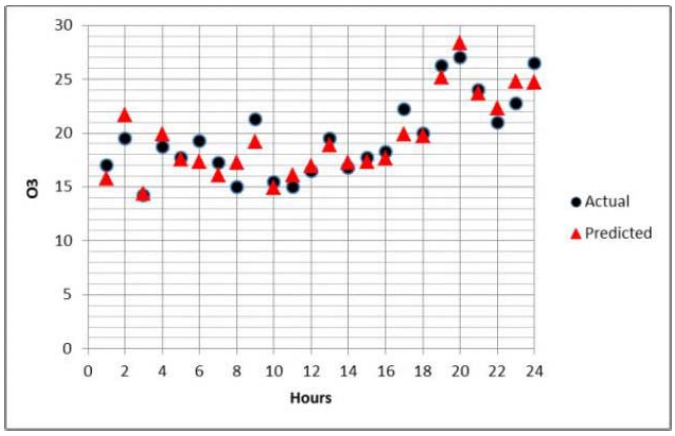


Fig. 6. O_3 actual and predicted values by univariate modeling.

7. *Performance Measures:* the performances of the produced forecasting models are measured using PTA and RMSE. Moreover, results are visualized for easy comparisons and testing of the output from the different models. The obtained results and the discussion thereof are given in the next section.

V. RESULTS AND DISCUSSIONS

This section shows the results obtained using the three ML algorithms for univariate and multivariate modeling after preprocessing, feature engineering, and windowing. O_3, NO_2, SO_2, and H_2S are all in units of ppb, temperature in Celsius, humidity in percentages, and wind speed in meters per second. TABLE I shows summary statistics of the data including the values of the minimum, maximum, average, and standard deviation for each.

The concentrations of O_3, NO_2, and SO_2 are predicted for the next 1, 8, 12, and 24 hour(s) either using values of the target gas (univariate), or using other features such as temporal, meteorological, and measured gases (multivariate). The ML algorithms used for building the forecasting models are fed with time windowed data, in which time is implicitly represented, based on the future steps to be predicted. These models are tuned to enhance results as summarized in TABLE II.

A. Ground-Level Ozone (O_3)

Fig. 6 shows the measured and predicted values of O_3 during 24 hours using M5P to predict the next hour.

TABLE II
PARAMETERS TUNING FOR ML ALGORITHMS

Alg.	Parameter	Value	Rationale
ANN	No. of hidden layers	1	The more complex the network, the worse the results, and the longer the training time
	No. of nodes in the hidden layer	Number of attributes	It is changing dynamically with number of features
SVMs	Kernel type	Dot	Produces the best RMSE
M5P	Min. No. of instances in leaf node	4	It is dependent on data size, so it is tuned to improve accuracy
	Pruning	Allowed	Reduces tree complexity, improves prediction accuracy, avoids over fitting and poor generalization
	Smoothing	Allowed	Reduces predictions extremeness

TABLE III
BEST FEATURES WITH EACH GAS BASED ON NRMSE IN MULTIVARIATE MODELING FOR DIFFERENT HORIZONS

Horizon	SO_2	NO_2	O_3
1	H_2S	Wind speed	NO_2
8	NO_2	SO_2	Humidity
12	Humidity	SO_2	Humidity
24	Day	Humidity	Temperature

TABLE IV
BEST FEATURES WITH EACH GAS BASED ON PTA BY MULTIVARIATE MODELING FOR DIFFERENT HORIZONS

Horizon	SO_2	NO_2	O_3
1	H_2S	Hour	Humidity
8	Humidity	Temperature	Humidity
12	Wind speed	Temperature	Hour
24	Day	SO_2	Hour

In univariate modeling, M5P algorithm achieved the highest accuracy and NRMSE in horizons 1, 8, and 24, while SVM achieved the best results in horizon 12. On the other hand, ANN achieved the worst NRMSE for all horizons compared to M5P and SVM. For example, in horizon 24, M5P achieves RMSE of 5.8; SVM achieves 6.4, while ANN achieves 16.4. Using M5P reduces RMSE by 35.4% compared to ANN in this horizon.

For multivariate modeling, the results are listed in TABLE III and TABLE IV. We notice from TABLE III, that O_3 NRMSE is mostly affected by meteorological features such as humidity and temperature. On the other hand, O_3 PTA is affected by meteorological and temporal features such as humidity and hour as shown in TABLE IV.

As shown in Fig. 7 and TABLE V, multivariate modeling achieved lower NRMSE in most of the horizons. Combining O_3 with other features yields lower NRMSE.

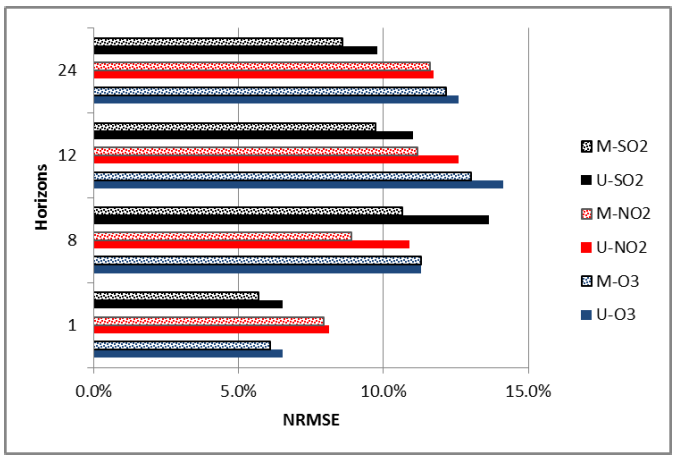


Fig. 7. Best values of NRMSE by univariate (U) and multivariate (M) modelling for all gases.

TABLE V
NRMSE OF UNIVARIATE AND MULTIVARIATE MODELING IN PERCENTAGE (U FOR UNIVARIATE AND M FOR MULTIVARIATE)

H	O_3		NO_2		SO_2	
	U	M	U	M	U	M
1	6.5	6.1	8.1	8.0	6.5	5.7
8	11.3	11.3	10.9	8.9	13.6	10.7
12	14.1	13.0	12.6	11.2	11.0	9.7
24	12.6	12.2	11.7	11.6	9.8	8.6

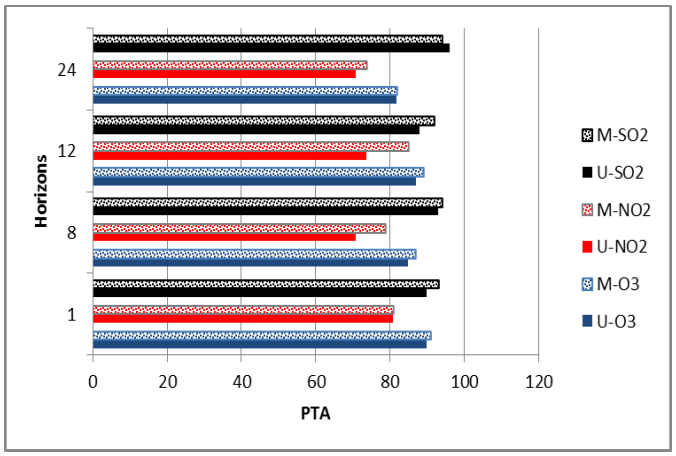


Fig. 8. Best values of PTA by univariate (U) and multivariate (M) modeling for all gases.

Fig. 8 and TABLE VI show PTA also increases with the multivariate modeling. All best results of multivariate modeling are produced using M5P algorithm. For example, RMSE in horizon 24 using temperature is 5.6, 6.6, and 11.6 using M5P, SVM, and ANN respectively. RMSE in horizon 24 without using temperature (univariate) is 5.8, 6.4, and 16.4 using M5P, SVM, and ANN respectively.

B. Nitrogen Dioxide (NO_2)

Fig. 9 shows the measured and predicted values of NO_2 during 24 hours using M5P for horizon 1.

TABLE VI
PTA OF UNIVARIATE AND MULTIVRIATE MODELING IN PERCENTAGE (U FOR UNIVARIATE AND M FOR MULTIVARIATE)

H	O_3		NO_2		SO_2	
	U	M	U	M	U	M
1	89.9	90.9	80.8	80.8	89.9	92.9
8	84.8	86.9	70.7	78.8	92.9	93.9
12	86.9	88.9	73.7	84.8	87.9	91.9
24	81.8	81.8	70.7	73.7	96	93.9

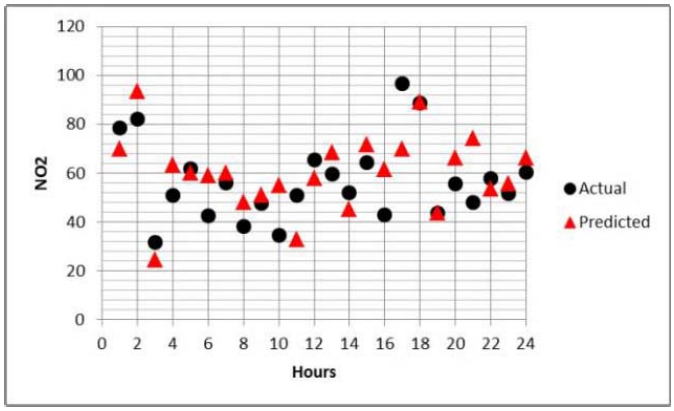


Fig. 9. NO_2 actual and predicted values of by univariate modeling.

In univariate modeling, M5P attained the best NRMSE for horizons 1 and 8, while SVM outperformed that for horizons 12 and 24. However, the best PTA of NO_2 is achieved using M5P for horizon 1 and SVM for horizon 8 and 12, both algorithms performed similarly for horizon 24. For example, in horizon 1, RMSE is 14.4, 15.5, and 25.7 for algorithms M5P, SVM, and ANN, respectively.

Based on Fig. 7 and TABLE III, the NRMSE by multivariate modeling of NO_2 is enhanced with meteorological features as, humidity and wind speed. It is also affected by SO_2. On the other hand, NO_2 PTA, as shown in Fig. 8 and TABLE IV, is enhanced through using temperature, hour of the day, and SO_2. For example, RMSE in horizon 1 using wind speed is 14.1, 18.1, and 20.3 for algorithms M5P, SVM, and ANN, respectively. RMSE in horizon 1 without using wind speed is 14.4, 15.5, and 25.7 using M5P, SVM, and ANN, respectively.

C. Sulfur Dioxide(SO_2)

Fig. 10 shows the measured and predicted values of SO_2 during 24 hours using M5P for horizon 1.

In univariate modeling, M5P outperformed other algorithms for all horizons in terms of NRMSE. However, SVM outperformed M5P only for horizons 8 and 12 in terms of PTA. For example, RMSE in horizon 8 is 71.2, 74.1, and 172.3 for algorithms M5P, SVM, and ANN, respectively.

In multivariate modeling, using the day of the week in forecasting SO_2 allowed reaching the best values of NRMSE and PTA for horizon 24 as shown in Fig. 7 and Fig. 9 along with TABLE III and TABLE IV. Adding humidity feature

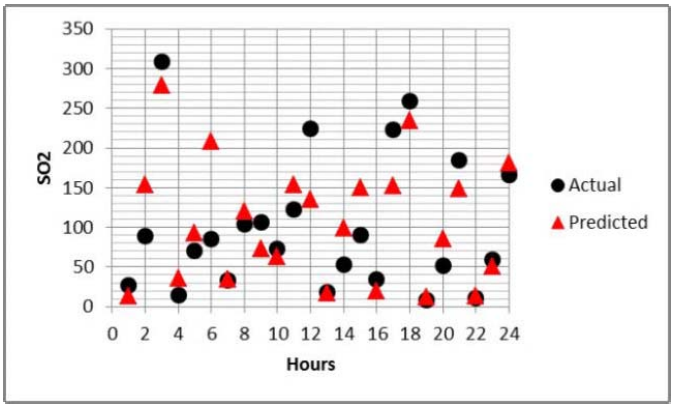


Fig. 10. SO_2 actual and predicted values of by univariate modeling.

results in the best NRMSE and PTA for horizon 12 and 8 respectively. Also, adding wind speed achieved the best PTA for horizon 12. Including H_2S and NO_2 as features enhanced the prediction of SO_2 and resulted in the best NRMSE for horizons 1 and 8. SO_2 also achieved the best PTA for horizon 1 when H_2S is added as a feature. For example, RMSE in horizon 8 using NO_2 is 58.8, 99.5, and 107.6 for algorithms, M5P, SVM, and ANN, respectively. RMSE in horizon 8 without using NO_2 is 71.2, 74.1, and 172.3 for M5P, SVM, and ANN, respectively.

Based on all experiments done on the 3 gases, ANN achieved the worst NRMSE and PTA for all horizons. ANN failed to properly produce good models for the data especially after windowing which increases the number data dimensions and reduces the training set size. This caused the ANN to poorly generalize the predictions and hence lead to worst results. On the other hand, SVM outperformed ANN because SVM is less resistant to training data dimensionality and size, so it can efficiently handle data with high dimensionality and small size [34]. These factors are the main causes of overfitting and hence poor generalization in ANN. In addition, M5P outperformed SVM and ANN due to its tree structure and high generalization ability [23].

VI. Conclusion

Air quality is an important problem that directly affects human health. Air quality data are collected wirelessly from monitoring motes that are equipped with an array of gaseous and meteorological sensors. These data are analyzed and used in forecasting concentration values of pollutants using intelligent machine to machine platform. The platform uses ML-based algorithms to build the forecasting models by learning from the collected data. These models predict 1, 8, 12, and 24 hours ahead of concentration values.

Based on extensive experiments, M5P outperforms other algorithms for all gases in all horizons in terms of NRMSE and PTA because of the tree structure efficiency and powerful generalization ability. On the other hand, ANN achieved the worst results because of its poor generalization ability when working on small dataset with many attributes that leads to a complex network that overfit the data, while having SVM better than ANN in our case due to its adaptability with high dimensional data.

Using multivariate modeling approach enhances the prediction accuracy and reduces error because of the dependency between target gases and other features included such as temperature, day of the week, and H_2S. For example, O_3 achieves the best NRMSE values when using NO_2, humidity, and temperature as multivariate modeling. NO_2 best PTA is achieved when using the hour, temperature, and SO_2. For SO_2, the best NRMSE results are achieved when using H_2S, NO_2, humidity, and day.

This work can be extended by considering data changes over time for real-time forecasting. This can be achieved by building online models that adapt automatically to changes in environment. Also, more data can be included to increase data seasonality.

References

[1] World Health Organization, "Monitoring ambient air quality for health impact assessment," WHO Regional Office Eur., Copenhagen, Denmark, Tech. Rep. 85, 1999.
[2] U. Gehring *et al.*, "Traffic-related air pollution and the development of asthma and allergies during the first 8 years of life," *Amer. J. Respiratory Critical Care Med.*, vol. 181, no. 6, pp. 596–603, 2010.
[3] L. E. Plummer, S. Smiley-Jewell, and K. E. Pinkerton, "Impact of air pollution on lung inflammation and the role of toll-like receptors," *Int. J. Interferon, Cytokine Mediator Res.*, vol. 4, pp. 43–57, May 2012.
[4] International Agency for Research on Cancer (IARC), "Outdoor air pollution a leading environmental cause of cancer deaths," World Health Org., Geneva, Switzerland, Tech. Rep. 221, 2013.
[5] J.-S. Hwang and C.-C. Chan, "Effects of air pollution on daily clinic visits for lower respiratory tract illness," *Amer. J. Epidemiol.*, vol. 155, no. 1, pp. 1–10, 2002.
[6] World Health Organization, "WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide," World Health Org., Geneva, Switzerland, Tech. Rep. WHO/SDE/PHE/OEH/06.02, 2005.
[7] US-EPA, "DRAFT roadmap for next generation air monitoring," U.S. Environ. Protect. Agency, Washington, DC, USA, Tech. Rep., 2013.
[8] M. F. Othmana and K. Shazali, "Wireless sensor network applications: A study in environment monitoring system," *Proc. Eng.*, vol. 41, pp. 1204–1210, Aug. 2012.
[9] M. I. Mead *et al.*, "The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks," *Atmos. Environ.*, vol. 70, pp. 186–203, May 2013.
[10] A. R. Al-Ali, I. Zualkernan, and F. Aloul, "A mobile GPRS-sensors array for air pollution monitoring," *IEEE Sensors J.*, vol. 10, no. 10, pp. 1666–1671, Oct. 2010.
[11] O. A. Postolache, J. M. D. Pereira, and P. M. B. S. Girao, "Smart sensors network for air quality monitoring applications," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 9, pp. 3253–3262, Sep. 2009.
[12] F. Tsow *et al.*, "A wearable and wireless sensor system for real-time monitoring of toxic environmental volatile organic compounds," *IEEE Sensors J.*, vol. 9, no. 12, pp. 1734–1740, Dec. 2009.
[13] S.-C. Hu, Y.-C. Wang, C.-Y. Huang, and Y.-C. Tseng, "Measuring air quality in city areas by vehicular wireless sensor networks," *J. Syst. Softw.*, vol. 84, no. 11, pp. 2005–2012, 2011.
[14] T.-C. Yu *et al.*, "Wireless sensor networks for indoor air quality monitoring," *Med. Eng. Phys.*, vol. 35, no. 2, pp. 231–235, Feb. 2013.
[15] C. B. D. Kuncoro, Armansyah, N. H. Saad, A. Jaffarb, C. Y. Low, and S. Kasolang, "Wireless e-nose sensor node: State of the art," *Proc. Eng.*, vol. 41, pp. 1405–1411, Aug. 2012.
[16] K. Aberer *et al.*, "OpenSense: Open community driven sensing of environment," in *Proc. ACM SIGSPATIAL Int. Workshop GeoStreaming (IWGS)*, 2010, pp. 39–42.
[17] E. Bales, N. Nikzad, N. Quick, C. Ziftci, K. Patrick, and W. Griswold, "Citisense: Mobile air quality sensing for individuals and communities Design and deployment of the Citisense mobile air-quality system," in *Proc. 6th Int. Conf. Pervasive Comput. Technol. Healthcare (PervasiveHealth)*, 2012, pp. 155–158.
[18] P. Langley and H. A. Simon, "Applications of machine learning and rule induction," *Commun. ACM*, vol. 38, no. 11, pp. 54–64, 1995.
[19] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.

[20] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon Press, 1995.
[21] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed. Cambridge, MA, USA: MIT Press, 2010, pp. 234–247.
[22] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. H. Witten, "Using model trees for classification," *Mach. Learn.*, vol. 32, no. 1, pp. 63–76, 1998.
[23] I. H. Witten, E. Frank, M. A. Hall, and G. Holmes, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Mateo, CA, USA: Morgan Kaufmann, 2011, pp. 252–259.
[24] A. Kumar, H. Kim, and G. P. Hancke, "Environmental monitoring systems: A review," *IEEE Sensors J.*, vol. 13, no. 4, pp. 1329–1339, Apr. 2013.
[25] J.-Y. Kim, C.-H. Chu, and S.-M. Shin, "ISSAQ: An integrated sensing systems for real-time indoor air quality monitoring," *IEEE Sensors J.*, vol. 14, no. 12, pp. 4230–4244, Dec. 2014.
[26] S. Rodopoulou, E. Samoli, A. Analitis, R. W. Atkinson, F. K. de'Donato, and K. Katsouyanni, "Searching for the best modeling specification for assessing the effects of temperature and humidity on health: A time series analysis in three European cities," *Int. J. Biometeorol.*, vol. 59, no. 11, pp. 1585–1596, 2015.
[27] B. Mattson and S. Mattson, *Microscale Gas Chemistry*, 4th ed. Bethel, CT, USA: Educational Innovations, 2006.
[28] D. Mintz, "Technical assistance document for the reporting of daily air quality—The air quality index (AQI)," US-EPA, Triangle Park, NC, USA, Tech. Rep. EPA-454/B-13-001, Dec. 2013.
[29] J. F. Kenney and E. S. Keeping, "Root mean square," in *Mathematics of Statistics*, 3rd ed. Princeton, NJ, USA: Van Nostrand, 1962, ch. 4, pp. 59–60.
[30] F. U. Dowla and L. L. Rogers, *Solving Problems in Environmental Engineering and Geosciences With Artificial Neural Networks*. Cambridge, MA, USA: MIT Press, 2003, ch. 3, sec 3.5.2, p. 54.
[31] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," *Atmos. Environ.*, vol. 80, pp. 426–437, Dec. 2013.
[32] K. P. Singh, S. Gupta, A. Kumar, and S. P. Shukla, "Linear and nonlinear modeling approaches for urban air quality prediction," *Sci. Total Environ.*, vol. 426, pp. 244–255, Jun. 2012.
[33] W. A. Broock, W. D. Dechert, J. A. Scheinkman, and B. LeBaron, "A test for independence based on the correlation dimension," *Econometric Rev.*, vol. 15, no. 3, pp. 197–235, 1996.
[34] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

Khaled Bashir Shaban received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada, in 2006. He is currently an Assistant Professor with the Computer Science and Engineering Department, College of Engineering, Qatar University, Doha, Qatar. His research experience in academic and industrial institutions covers a variety of domains in intelligent systems application and design.

Abdullah Kadri (M'09) received the M.E.Sc. and Ph.D. degrees in electrical engineering from the University of Western Ontario, London, ON, Canada, in 2005 and 2009, respectively. In 2009, he joined the Qatar Mobility Innovations Center, as a Research Scientist, and in 2013, he became a Senior Research and Development Expert and Technology Lead, focusing on research and development activities related to intelligent sensing and monitoring using mobility sensing. His research interests include wireless communications, wireless sensor networks, indoor localization, and smart sensing.

Eman Rezk received the M.Sc. degree from the Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar, in 2015. She is currently a Research Assistant with the Department of Computer Science and Engineering, College of Engineering, Qatar University. Her research experience in academic and industrial institutions involves sensors data analysis, hidden data analysis, machine learning, and semantic Web.