An IoT-Based Pollution Monitoring System Using Data Analytics Approach

Harshit Srivastava, Shashidhar Mishra, Santos Kumar Das, and Santanu Sarkar

Abstract Air pollution occurs when harmful gases such as CO and NH_3 concentration levels increase above the threshold level specified by the World Health Organization (WHO). Among this, one of the very important parameters is particulate matter. These are tiny particulate that they reach directly to the lungs and cause breathing problems. The standard level of range for pollution is already given by the central governing body of India, i.e., Central Pollution Control Board (CPCB) in terms of the air quality index (AQI). In this paper, a system has been developed for detecting the air pollution index with the help of Raspberry Pi based on IoT technology which sends an emergency notification (EN) if there are any chances that the air pollution may raise above the given threshold in the future is developed which measures physical parameters like temperature, humidity, dew point, wind speed and pollutants parameters like suspended particulate matter (SPM) and carbon monoxide (CO) are monitored, and the effect of these parameters in pollution level is being predicted for pollution monitoring. The main objective of this is to apply the machine learning algorithm for the prediction and analysis of gas sensors concentration levels, the effect of physical environmental parameters so that we can analyze the future concentration levels (AQI) level of the gaseous pollutant, and based on this, an emergency notification (EN) is sent to the public as well as the concerning authorities. A system is developed for monitoring and alerting in real time. We are discussing the different methods used in machine learning algorithm, i.e., support vector machine (SVM) and random decision forests (RDF) to predict the multivariate time series for forecasting and to use these predicted values to send an emergency notification (EN).

H. Srivastava (✉) · S. Mishra · S. K. Das · S. Sarkar
National Institute of Technology Rourkela, Rourkela, Odisha 769008, India
e-mail: harshitsrivastava2345@gmail.com

S. Mishra
e-mail: 218ec5339@nitrkl.ac.in

S. K. Das
e-mail: dassk@nitrkl.ac.in

S. Sarkar
e-mail: sarkars@nitrkl.ac.in

P. K. Mallick et al. (eds.), *Electronic Systems and Intelligent Computing*, Lecture Notes in Electrical Engineering 686, https://doi.org/10.1007/978-981-15-7031-5_18

Keywords Internet of things (IoT) · Raspberry Pi 3 · Support vector machine (SVM) · Random decision forests (RDF) · Central Pollution Control Board (CPCB) · Emergency notification (EN)

1 Introduction

Over the century, there has been a significant rise in the industries to fulfill the needs of the evergrowing population. As a result of large numbers of industries, it has caused a lot of major problems for the environment. As we know few of the impacts of air quality on the animals, human and environment, the governing body of the world, i.e., World Human Organization (WHO) has created guidelines to curb the health-related effects of air pollution on public by setting up the guidelines for the concentration of harmful air pollutants likes particulate matter, carbon monoxide, wind speed, temperature, humidity, etc. The first and the foremost effects that can be seen are the environmental pollution due to which there is a degradation of the atmosphere (breathable air) along with climate change with stratospheric ozone depletion and reduction in biodiversity, hydrological imbalance, freshwater, etc. Mobile sources and stationary releases chemical pollutants due to which suspended particulate matter (SPM), carbon monoxide (CO) and other toxic gases.

There are three important parts of integral design architecture in a system.

1.1 Perception Layer

This network layer primarily includes a stationary field sensor that works on a front-end acquisition device. As it can be realized by establishing a reliable as well as stable monitoring system, which includes site selection, sensor deployment over an area, and meteorological sensor deployment.

1.2 Network Layer

The importance of this is conveying the data related to the environment after assimilating all the sensor nodes which are deployed along an area for monitoring after this transmits the data received by the microcontroller to the data center in real time.

1.3 Application Layer

One of the primary focuses of this is to analyze and then process the pollutants. Then, calculate the air quality after which it should predict the trend which is to be followed. From the working prospective, the layer means pollution forecasts and air quality evaluation and then generating a notification to the concerning authorities.

2 Literature Survey

Hu et al. [1] presented an idea of how to trade off among the monitoring data quality and the cost of the notification message using the adaptive approach by taking the variance of the data received from the sensing nodes and adjusting the sensing rates. Yi et al. [2] where they have tackled the main problem with a pollution monitoring system is that, the presence of the limited amount of data available and non-scalability of the system available this been removed in our work by using advance sensing techniques like Wireless Sensor Networks (WSN). Hu et al. [3] presented an idea of the Haze Est-a model, which links with fixed-station data and data from the sensor to measure air quality, and the finding shows that decision tree and estimation accuracy of support vector regression (SVR) are equivalent. Rybarczyk et al. [4] presented an idea for the statistical correlation of particulate matter and wind after which it proposes a simple machine learning model to predict the level of particulate matter. Zhang et al. [5] presented an idea where they have used the advantage of a support vector machine of the classification and algorithm to overcome practical problems with which an algorithm proves the effectiveness of the algorithm. Saha et al. [6] presented an idea for a region implementing an IoT-stationed method to monitor the noise intensity and air quality index, a simple machine learning model to predict the level of particulate matter using Raspberry Pi Wi-Fi-enabled module. Zhang et al. [7] presented an idea of predicting the future outcome helps in reducing the harmful, so in this paper, they have used prediction techniques to avoid the effects of air pollution on the environment. Zhao et al. [8] presented a monitoring system for air pollution using sensors that can detect particulate matter, temperature and humidity, and when it crosses a safe range of the above parameters, then an alert signal is emitted. Yu et al. [9] presented an idea of the random forest in which due to its recursive partitioning to get many trees and then aggregate the total output to reduce the error. A Lyons et al. [10, 11] presented an idea that an alert notification will help in reducing the effects of air pollution. As they have used an alert notification to the people for high levels of air pollution so as to reduce the damage caused on the health and can be used to avoid such instances. Ayele et al. [12] presented a system that focuses on monitoring of air pollutants having an IoT-enabled system using recurrent neural network specifically Long Short-Term Memory (LSTM) which is a machine learning algorithm that shows a quick convergence and reduction in the training cycles that too with good accuracy.

3 Methodology

3.1 Architecture Model

The architecture model that we are using for our system will describe how the whole system is going to work. As shown in Fig. 1, the model will be used to take the environmental parameters reading, such as the particulate matter, temperature and few gas parameters. Now, these data will be taken up by our microcontroller and then will be sent to the cloud server through a modem. The model also has a display device that can be used to see real-time data. It also helps to detect the exact location of the device. The data sent in the server can be seen in the GUI interface that is a web-page that can be accessed from anywhere across the world.

Sensors Nodes: The different kind of parameters which are analyzed is taken by the different kind of the sensors, i.e., particulate matter sensing PM2.5 and PM10 (HPMA115S0), temperature sensor and humidity sensor (SHT10) and CO (carbon monoxide). The reason why these parameters have been considered is because we want to predict if there is an alarming situation in the near future so that necessary steps can be taken by the concern authorities.

Control and Processing Unit: It contains a microcontroller which is used to control the sensors to take values in real time. The microcontroller that we are using is Raspberry Pi 3 Model B. It features Broadcom BCM2837 SOC, which has an ARM Quad Cortex processor A53 with 1.2 GHz clock speed, and in addition to this, we are using an ARPI600 chip is ADC which converts analog to digital values.

Cloud Database and Cloud Server: The server will be used to store, host and manipulate the database along with a push notification that will be released when the future values increase a threshold which is being defined by the Central Pollution Control Board (CPCB) and to send an emergency notification to the concerned authorities. The server will store the pollution data received from the sensor node

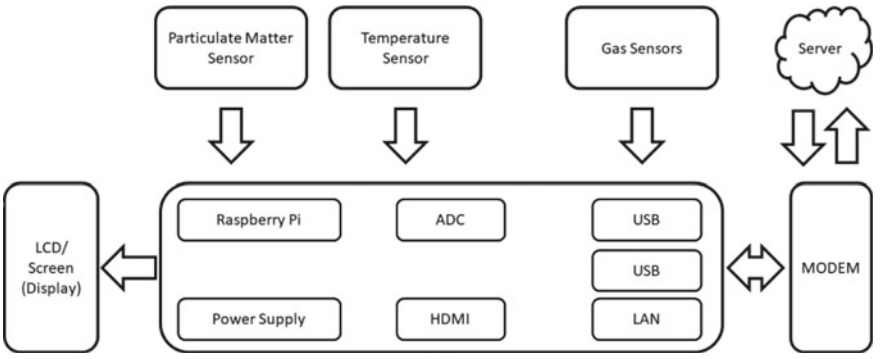

Fig. 1 System architecture model

through wireless links, and this data can be used by the front-end application which will then be used for the analytical purpose.

Android Emergency Notification: The data that is stored in the cloud database is stored from the sensor node through wireless links, and this data is used by the front-end application to analyze in which prediction algorithm is applied to trigger an Android emergency notification (AEN) with the help of an application program interface (API), which will be used to take proper precautions and to avoid the harmful effects of air pollutions.

3.2 Flowchart Model

The primary goal of the model is for monitoring the air quality index over an area and predicting the value to eliminate the problem of air pollution. As shown in Fig. 2, the data is taken from a different kind of the sensor and then given to the Raspberry Pi board and further, the AQI value is being calculated, and finally, the data is uploaded to the cloud where the prediction models will be used to forecast the future values, and if the values cross a certain threshold given by an IoT-based pollution Alert will be shown using Data Analytics Approach the CPCB (Central Pollution Control Board) in there website so that the concerned authorities can take the necessary steps.

3.3 Prediction Model

The prediction model can be explained by Fig. 3 whose working for each block is given below.

Data Preprocessing: First of all, remove all the irrelevant features and missing values from the database. Remove those values that are considered outside of the range to make sure that the credibility of the pollution dataset is not reduced.

Input Feature Engineering: After taking the sensor values from the ground station data splitting of the dataset into the hour, weekday/weekend and season by dividing it into different rows and columns so that as per the need, prediction model can be applied to get the desired output.

Target Preparation: In target preparation, we take an average of the data point received over an hour as the data we are getting is coming after 6 s, so there will not be a significant parameter change over such a short time, and then, the normalization is done on this data to avoid over-fitting.

Model Training: Here, we use Random Forest and SVM for classification and predicting the values with the help of Regression models based on the above methods

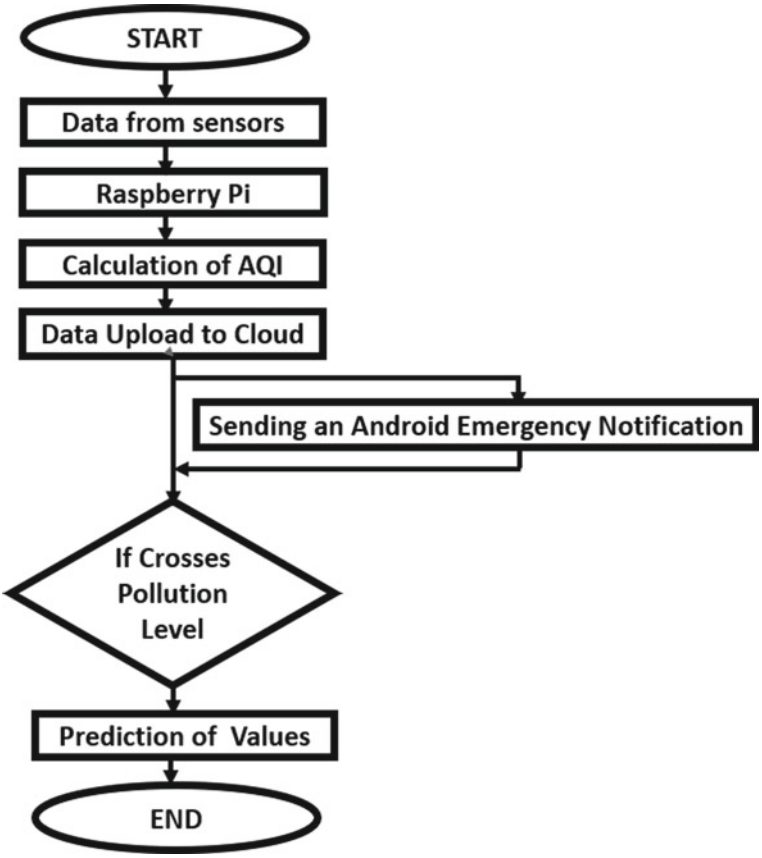

Fig. 2 Flowchart model diagram

now based on the output decision that will be taken to change the Hyper-parameters to improve the accuracy of the model.

3.4 Regression Models

Support Vector Regression (SVR): It is one of the much blooming supervised methods for the regression, support vector tries to give a mapping function which is nonlinear for plotting the given training data points, i.e., *D*: ($x1$, $y1$), ($x2$, $y2$), …, (xi, yi) to a higher dimension. Over this higher dimension, a separating decision boundary hyperplane is defined which distinguishes all the dataset over the maximum margin function.

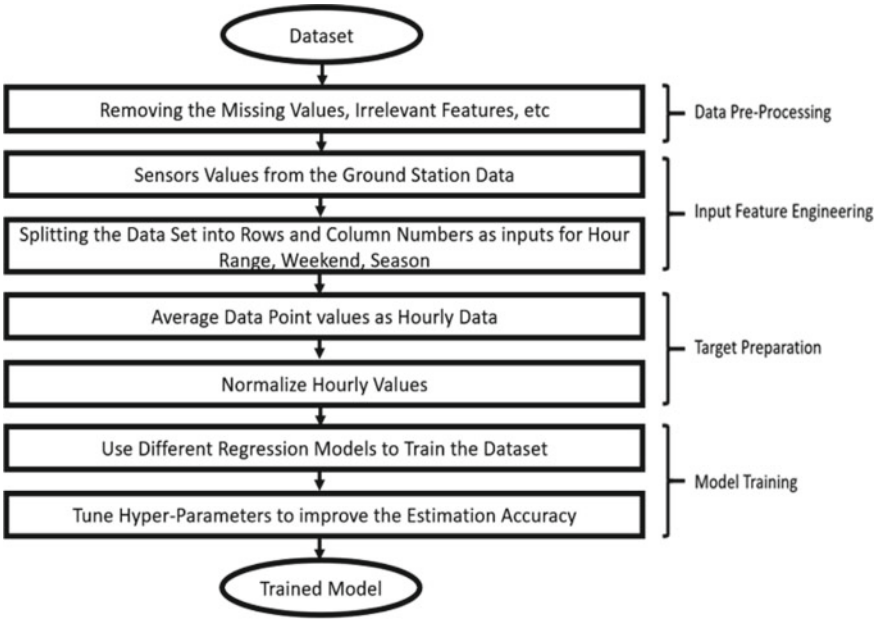

Fig. 3 Flowchart of prediction model

Random Forest Regression (RFR): It is an ensemble learning algorithm that is based on decision tree learning algorithm and bootstrap aggregating, which can be used for regression, classification and other tasks. The vital and important concept is to improve the prediction accuracy where subsamples play an important role. It is done by fitting a large number of decision trees on random subsets to the features available as a result of which it avoids over-fitting. As in this case, bootstrap aggregating is used (or named bagging) to continuously train decision or regression trees, with random feature subsets and sample subsets over which the algorithm is applied. Finally, after this, it predicts all the samples which have been taken into consideration by using the averaging technique and implying it to all the prediction from trees that have been trained.

4 Results

4.1 Firebase Server Database

As shown in Fig. 4, we have obtained some of the real-time parameters like temperature, humidity, dew point and particulate matter (PM2.5 and PM10) along with CO gas sensor values stored in the Firebase database after this prediction model is used.

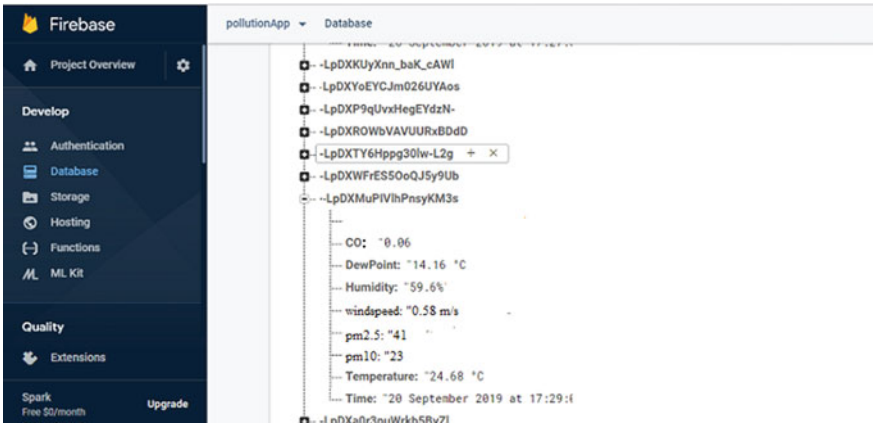

Fig. 4 Real-time data packets in database

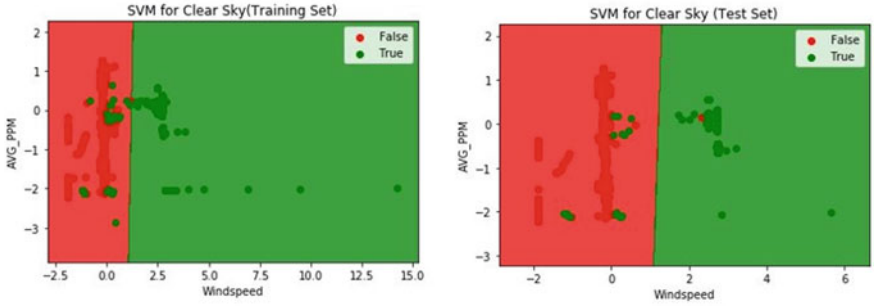

Fig. 5 Classification model of SVM

4.2 Support Vector Machine (SVM)

The primary function of this support is to identify a hyperplane in an *N*-dimensional hyperspace, which classifies the data points to distinguish the data points in between two classes, as it is visible in Fig. 5. Our main goal is to find a hyperplane that has the maximum margin as it is clear from the plot as the distance between data points of both the classes should increase significantly over the testing, so the accuracy is high. As shown in Fig. 6, the accuracy in our case is around 90%.

4.3 Random Forest Regression (RFR)

This technique uses a large number of decision trees in which these trees break a class of prediction, and then, the class over which maximum votes are considered

```
[[1551   12]
 [ 149   61]]
              precision    recall  f1-score   support

       False       0.91      0.99      0.95      1563
        True       0.84      0.29      0.43       210

    accuracy                           0.91      1773
   macro avg       0.87      0.64      0.69      1773
weighted avg       0.90      0.91      0.89      1773

accuracy =  90.91934574168077
```

Fig. 6 The accuracy of classification model of SVM

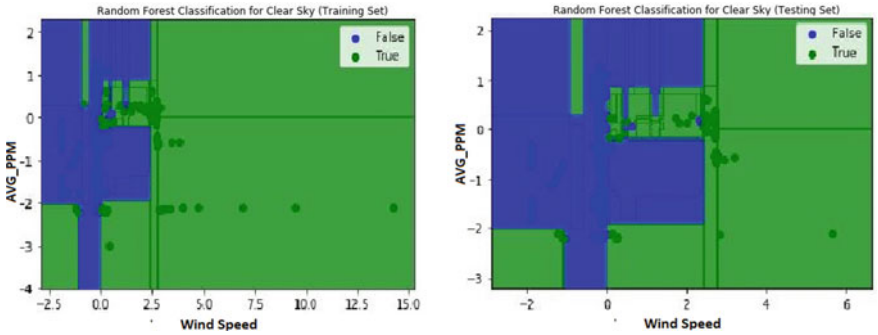

Fig. 7 Classification model of random forest regression

is considered as it is visible in our model for prediction in Fig. 7. The basic principle behind why the random forest model gives high accuracy is that when several uncorrelated trees will outperform any individual models. This can be explained as a wonderful effect of an individual tree that protects each other from their errors. There may be some trees that are wrong, and many others may be right, so overall, these trees can move in the right direction. As shown in Fig. 8, it is visible the accuracy that we are getting is around 99% which is indeed very good. Figure 9 shows the testing values and the predicted values of classification model of random forest.

4.4 Android Emergency Notification (AEN)

An AEN is send to the public based on the prediction algorithm that it is not safe for them move out for tomorrow based on the past data that has been received from the sensing node, this notification is sent based on the data that is received by the

```
[[1594    1]
 [   3  175]]
              precision    recall  f1-score   support

       False       1.00      1.00      1.00      1595
        True       0.99      0.98      0.99       178

    accuracy                           1.00      1773
   macro avg       1.00      0.99      0.99      1773
weighted avg       1.00      1.00      1.00      1773

accuracy =  99.77439368302312
```

Fig. 8 Accuracy of classification model of random forest regression

Fig. 9 Testing values on the left side and predicted values on the right side of classification model of random forest

No.	y_test	No.	y_pred
0	38.4598	0	38.446
1	37.3202	1	37.2852
2	25.8971	2	25.9047
3	19.6317	3	19.6904
4	44.6964	4	44.6724
5	21.8763	5	21.8615
6	45.9081	6	45.9252
7	43.9372	7	43.959
8	41.7965	8	41.8002
9	30.7355	9	30.5352
10	43.2034	10	43.212
11	46.5866	11	46.575

Firebase, this data is being taken up with the help of the API key over which the data is being accessed, and then based on the threshold levels, a notification is sent as shown in Fig. 10.

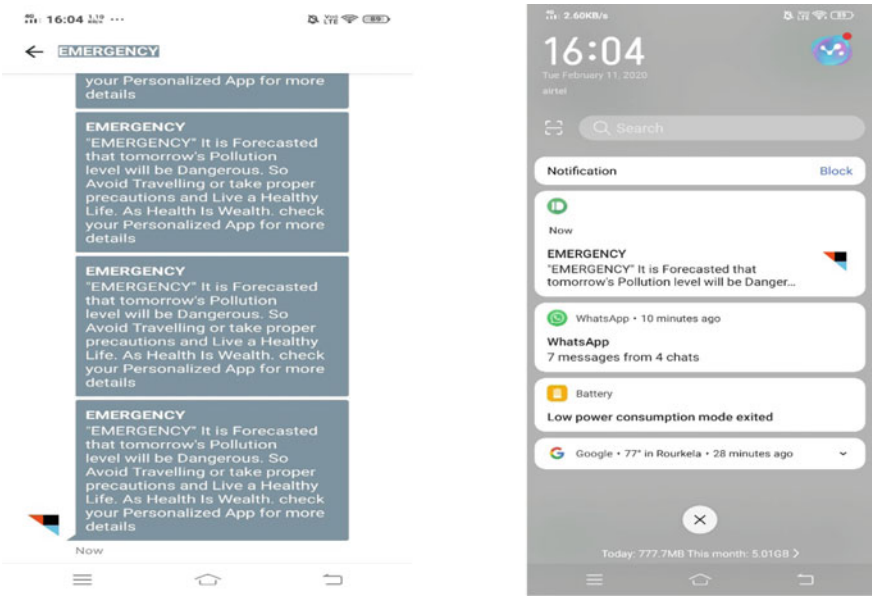

Fig. 10 Real-time images of emergency notification on left side and emergency pop-up on the right side

5 Conclusion

We have proposed a model which uses Raspberry Pi microcontroller along with the different kind of the sensors and then sending the data to the Firebase server over which prediction algorithm is used to predict the future values, and then based on the values, an emergency notification is sent. In an Internet of things (IoT)-based environment to measure air quality, this gives us a macro description of our proposed hardware and software module along with an explanation of the algorithm for calculation of values of different gas sensors, temperature sensors, humidity sensors, PPM data and predicts the value with the help of machine learning. Prediction model shows that RFR outperforms SVM which can be seen from Fig. 9 where the accuracy for this type of dataset is very high as it has to be the result because random forest model uses a lot of individual decision trees due to which it reduces the chances of predicting the wrong values overall due to considering an average result from a large number of decision trees of air pollution.

6 Scope for Future

Future work is to predict the AQI value for analyzing the air quality using machine learning with different techniques like LSTM, KNN and MLP and comparing the

accuracy of prediction. The real-time data is to be uploaded on a cloud server using a LAMP server model, and routing algorithm is to be implemented for finding the safest path based on environmental parameters from the similar nodes deployed across different location.

References

1. Hu SC, Wang YC, Huang CY, Tseng YC (November, 2011) Measuring air quality in city areas by vehicular wireless sensor networks. Elsevier J Syst Softw 84(11):2005–2012
2. Yi WY, Lo KM, Mak T, Leung KS, Leung Y, Meng ML (December, 2015) A survey of wireless sensor network based air pollution monitoring systems 15(12):31392–31427
3. Hu K, Rahman A, Bhrugubanda H, Sivaraman V (1 June, 2017) Haze Est: machine learning based metropolitan air pollution estimation from fixed and mobile sensors. IEEE Sensors J 17(11):3517–3525
4. Rybarczyk Y, Zalakeviciute R (2016) Machine learning approach to forecasting urban pollution. IEEE Ecuad Tech Chap Meet (ETCM), pp 1–6
5. Zhang Y, Liu C, Wang L, Yang A (November, 2012) Support vector machine classification algorithm and its application. In: Springer information computing and applications (ICICA 2012) communications in computer and information science, vol 308, pp 179–186
6. Saha AK, Sircar S, Chatterjee P, Dutta S, Mitra A, Chatterjee A, Chattopadhyay SP, Saha HN (2018) A raspberry Pi controlled cloud based air and sound pollution monitoring system with temperature and humidity sensing. In: 2018 IEEE 8th annual computing and communication workshop and conference (CCWC). Las Vegas, NV, pp 607–611
7. Zhang J, Ding W (January, 2017) Prediction of air pollutants concentration based on an extreme learning machine: the case of Hong Kong. Int J Environ Res Public Health 14(2):114
8. Zhao X, Zuo S, Ghannam R, Abbasi QH, Heidari H (2018) Design and implementation of portable sensory system for air pollution monitoring. In: IEEE Asia Pacific conference on postgraduate research in microelectronics and electronics (prime Asia), Chengdu, pp 47–50
9. Yu R, Yang Y, Yang L, Han G, Move OA (January, 2016) RAQ-a random forest approach for predicting air quality in urban sensing systems. Sensors (Basel). Int J Environ Res Public Health 16(1):PMCID: PMC4732119
10. Lyons RA, Rodgers SE, Thomas S, et al (May, 2016) Effects of an air pollution personal alert system on health service usage in a high-risk general population: a quasi-experimental study using linked data. J Epidemiol Community Health 70(12):1184–1190
11. Chen H, Li Q, Kaufman JS, Wang J, Copes R, Su Y, Benmarhnia T (January, 2018) Effect of air quality alerts on human health: a regression discontinuity analysis in Toronto, Canada. Elsevier Lancet Planet Health 2(1):19–26
12. Ayele TW, Mehta R (2018) Air pollution monitoring and prediction using IoT. In: 2018 second international conference on inventive communication and computational technologies (ICICCT), Coimbatore, pp 1741–1745
13. Yang Y, Zheng Z, Bian K, Song L, Han Z (2018) Real-time profiling of fine-grained air quality index distribution using UAV sensing. IEEE Internet Things J 5(1):186–198
14. Shaban KB (2016) Urban air pollution monitoring system with forecasting models. IEEE Sensors J 16(8)
15. Martınez-Espana R, Bueno-Crespo A, Timon I, Soto J, Munoz A, Cecilia JM (2018) Air-pollution prediction in smart cities through machine learning methods: a case of study in Murcia Spain. J Univers Comput Sci 24(3):261–276
16. Xiao Jun C, Xian Peng L, Peng X (2015) IoT-based air pollution monitoring and forecasting system. In: International Conference on Computer and Computational Sciences (ICCCS), Noida, pp 257–260