Indoor Air Quality Prediction Using Machine Learning Techniques

Dina Hamad Alghurair[1], Meshal Mansour Alnasheet[2]

[1]Higher Institute of Telecommunication and Navigation, HITN, PAAET, Shuwaikh Industrial, Kuwait (Specialist trainer A)
dh.alghurair[at]paaet.edu.kw

[2]Higher Institute of Telecommunication and Navigation, HITN, PAAET, Shuwaikh Industrial, Kuwait (Specialist trainer A)
mm.alnasheet[at]paaet.edu.kw

Abstract: ***Monitoring indoor environment parameters became one of a required service nowadays. Because of the higher consumption of fossil fuel and less depending on renewable energy, climate is changing into a negative and dangerous direction and caused climate overheating as well as increased air pollution rates to extreme levels. From here on, monitoring the indoor air quality and having a kind of evaluation system to predict the air quality in advance is required and can provide a great added value for people living in closed environments and under strong dusty, polluted and highly toxic weather conditions. In this research we provide a mixed based solution between indoor air environment measurements system and machine learning based solution to predict and evaluate the current indoor air quality. Several ML models like Support Vector Machine (SVM) and Artificial Neural Networks (ANN), Random Forest, etc are used to implement the machine learning models and a simple hardware implementation is designed using AT - Mega microcontroller attached with Raspberry - Pi 4 to finalize indoor parameters measurements and machine learning processes. Temperature, Humidity, Dust and Gas sensors were used to measure indoor environment parameters. Our system could reach a classification accuracy of 100% using RF, 96% using SVM and 99% using ANN.***

Keywords: Machine Learning, Feature Extraction, AT - Mega, Air Quality, Indoor Environment

1. Introduction

Because of the higher consumption of fossil fuel and less depending on the renewable energy resource, the climate is being changed into a very negative and dangerous direction and caused climate overheating as well as increased the air pollution rates to very extreme levels. From here on, metropolitan cities started to consider this issue as a major public health issue and started seeking for an appropriate and sustainable solution in order to control the pollution levels and start to fix the problem. Using the most developed communication technologies, the task of having fully automated and distributed systems became very easy and simple for implementation. So that, a kind of simple weather conditions, air quality, water quality and soil quality sensors attached to a simple microcontroller distributed in many different areas, can communicate with each other and share all of their readings with a centralized unit running over a cloud server, where all of these data can be analyzed, studied and then finally manipulated in order to give a very clear vision about the current air quality in a specific predefined area. This kind of solutions can help the responsible authorities to take an immediate and sustainable action in order to control the pollution levels and reduce it. In this research we are targeting to design a kind of microcontroller based solution collecting air quality parameters in indoor environment using (Temperature sensor, Humidity sensor, Gas sensor, and Dust sensor) attached to a single microcontroller device in closed area in order to send all of these readings to a specific cloud server using the Internet of Things communication protocols (IoT). The data will be collected at the server side, validated, studied, analyzed and visualized in order to help monitoring people to finalize reports about the air quality in specific closed area and then finalize the main cause of the pollution in order to offer some proper solutions for that issue. At server side, we have implemented a machine learning based model trained over a set of sensors readings collected from the hardware setup in order to predict the indoor air quality and classify it into 3 main categories (Bad, Medium and Good) air quality. A kind of wireless communication between the data measurements (Arduino and Sensors) and data processing unit (Raspberry) is being established and implemented using Xbee - S1 models in order to ensure low energy consumption for wireless network communication and to ensure also enough communication distance in indoor environment with possible noisy obstacles up to 400 meters. In the next sections of this research project, we will show the entire system design and implementation in details.

2. Literature Survey

In 2018, Jouvan Chandra Pratama Putra et. al. [1] described 'The prediction of indoor air quality in office room using artificial neural networks'. In this paper, authors have designed and implemented a machine learning based model using ANN feed forward structure in order to predict the indoor office air quality using mainly the value of CO2 collected by LGA meter. The authors created their own dataset for training and testing purposes and implemented the ANN architecture using the ANN toolbox provided by Matlab framework. The classification accuracy reached 93% during the evaluation phase. **In December, 2015,** Avril Challoner **et. al. [2] described 'Prediction of Indoor Air Exposure from Outdoor Air Quality Using an Artificial Neural Network Model for Inner City Commercial Buildings'.** In this research paper, NO2 values and density were measured and predicted in order to classify the air quality inside city buildings in Dublin, Ireland. Different machine learning models were implemented and evaluated in order to classify and predict the air quality inside buildings, so that different measurements end points were placed in different buildings f the city in order to collect a set of NO2 and PM2.5 measurements, during different day times and under different weather and overall buildings conditions. ANN model with different architectures FFNN and BPNN are being trained and evaluated in order to finalize a model which can be used for the real life scenario indoor air quality prediction. Both ANN models were able to predict the air quality and the expected NO2 and PM2.5 concentrations with approximately 92%. **In Feb, 2022, Himawan Nurcahyanto et. al. [3] described 'Multilevel RNN - Based PM10 Air Quality Prediction for Industrial Internet of Things Applications in Clean room Environment'.** They designed a stable RNN Deep Neural Network model to predict the annual and seasonal indoor air quality based on the PM2.5 and PM10 concentrations. They have proposed the IAQP method for air quality management systems that combines indoor air quality forecasting based on real - time data. To measure indoor air quality, they have predicted the outcome from the IIoT sensor and LoRa sensor. The result of prediction is that the multilevel RNN model outperformed the LSTM, as the model demonstrated excellent results and feasibility. Accuracy and performance of RNN classifier could reach 78%.

3. Methodology

3.1 Design Overview

In this paper, we propose a kind of wireless based communication Air Quality monitoring system using Arduino and Raspberry Pi communicating with each other over an Xbee communication channel, where the Arduino microcontroller is being attached with different Environment / Air Quality parameters measurements sensors and sends all of the sensors reading over Xbee channel to the Raspberry Pi, where the RPI device can also accept another set of readings from another Arduino endpoint. The RPI device will create a kind of training dataset including all sensors' readings (Temperature, Humidity, Gas, and Dust) labeled by a pre - defined category identifying the current readings air quality in the room. The label of each reading will be distributed over 3 main categories (Bad, Medium and Good) air qualities. This dataset will be used by different machine learning models (Support Vector Machine (SVM), Artificial Neural Network, and Random Forest) models to be trained over the dataset and approximate an accurate training model to be used further on by detecting and classifying the new incoming unlabeled sensors readings from the MCU and sensors. The data collected by the RPI device will be posted over WiFi channel to a dedicated Cloud server running on Google Cloud (Firebase) where the data will be stored on the Firebase server and will be made available for any third party application to be monitored and observed. A dedicated Android mobile application will be developed in order to access the Cloud Server Firebase and collect the sensors and parameters data points with the predicted air quality value and represent these Data for responsible people in order to take an action back in case of any extreme data reading by the data measurement end. The data measurement station (Arduino + Sensors) will be powered using a continuous DC power source using battery. Below we list all proposed components of our solution:

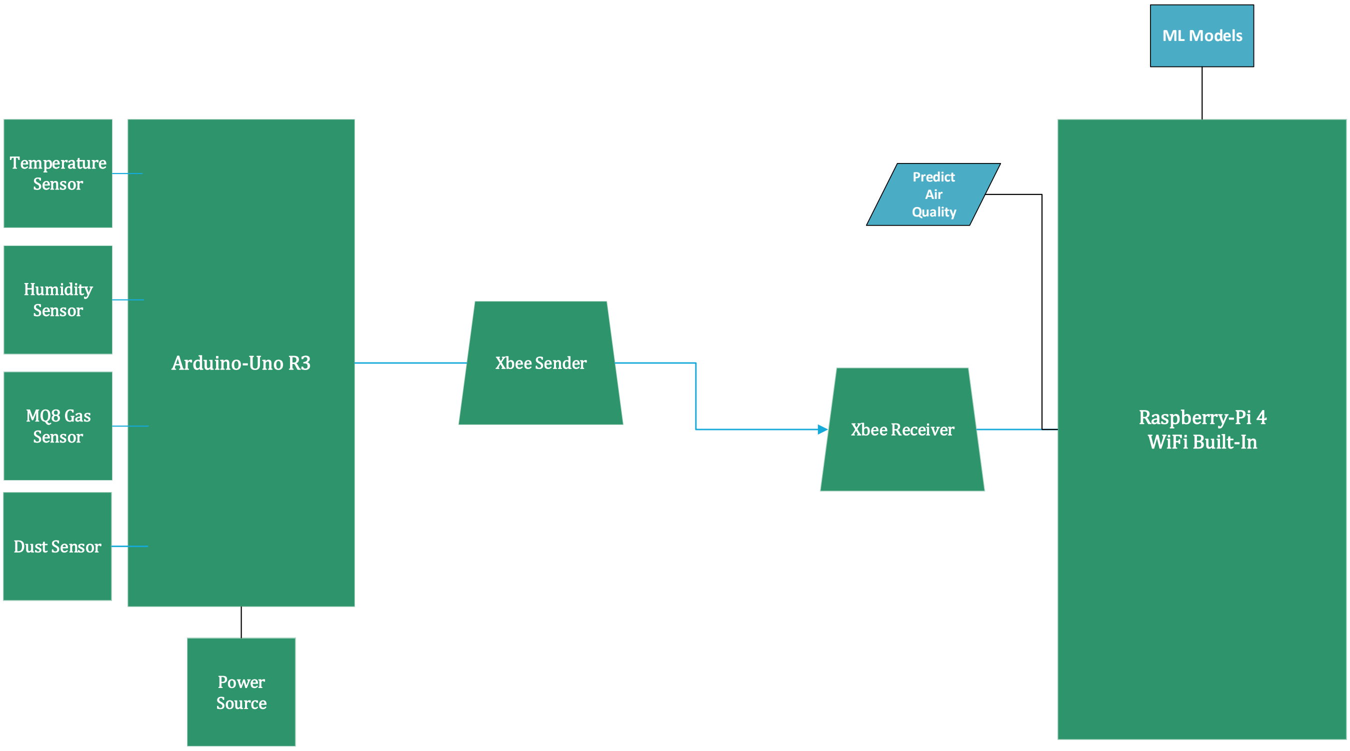


Figure 1: System High Level Diagram

3.2 Data Collection Unit

In this terminal, we have an Arduino Uno microcontroller attached with different sensors (Temperature Sensor, Humidity Sensor, Dust Sensor, MQ8 Gas Sensor) Where each sensor is considered as an input unit to the Arduino and send its reading and sensing values to Arduino on different Pins. Arduino is getting these values and using an Xbee - S1 module Arduino sends this data over Xbee channel to the second terminal (Raspberry Pi Unit). The following figure

(2) shows the circuit connections including all components used in our data collection model:

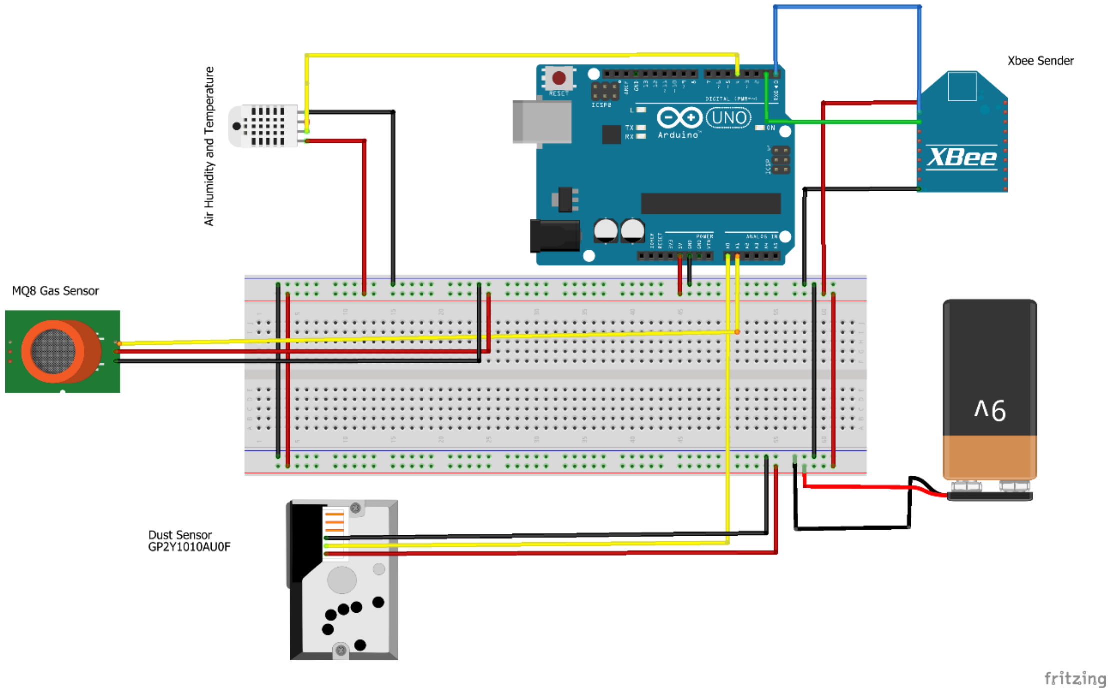


Figure 2: Circuit Diagram

below we list all hardware components in our solution with all details.

Table 1: Hardware Components Details

Component	Input / Output	Digital / Analog	Pin on Arduino	Reading Value Range	Power Required
Arduino - Uno	Both	Both	-	-	5v
DHT11 (Temperature and Humidity)	Input	Digital	D4	T [- 50 to +50] H [0 - 100]%	5v
Dust Sensor	Input	Analog	A0	0 - 1023	5v
MQ8 Sensor	Input	Analog	A2	0 - 1023	5v
Xbee - S1	Output	Both	RX - D0 TX - D1	-	5v
RPI - 4	Both	Both	-	-	12v

a) Data Processing and IoT Communication Unit

In this terminal, we have a Raspberry Pi - 4 attached with an Xbee - S1 receiver model, which receives the collected sensors readings from Arduino terminal. Once the data is received at the Raspberry Pi end, the Raspberry sends the data over WiFi network to the third terminal which is the IoT Firebase Cloud Server. On Data processing station, we have implemented a SVM, RF and ANN models in order to train over the collected and designed training dataset. Both models are being trained on RPI device and saved locally to be used during the evaluation and testing phase. All sensors data with predicted air quality category will be posted from the RPI device into the Firebase cloud sever.

b) IoT Cloud Server

In this terminal, Firebase Storage and Cloud Database Engine running on Google Cloud service is used to receive the data from Raspberry device and store them directly in a dedicated Database. Firebase will provide a functionality to have an access to the Database from a Third - Party Monitoring system running on Android Mobile App which actually considered as fourth terminal.

c) Android Mobile Application: Using MIT - App - Inv2

A monitoring application is designed to communicate with Firebase Cloud Server and fetch the data/sensor readings from the server and display them on the Mobile App GUI. As well as the application will give the users the ability to send feedback commands back to the RPI device and from RPI device to Arduino end in order to actions for any extreme case.

4. Results and Discussion

4.1 Dataset Description

A customized dataset has been created using the output readings from the Arduino data collection unit. Our dataset includes 4 features (Temperature, Humidity, Dust Level and CO Level) attached with a class/category value (0 Bad, 1 Medium and 2 Good) air qualities. We have created 6200 instances in our dataset divided into 70% 4240 training instances and 30% 1860 evaluation instances.

4.2 Machine Learning Models

A pipeline based model has been created including (SVM, RF and ANN) in order to train over the training dataset and derive 3 trained models to be evaluated over the validation dataset. Since we focus on models that provide high precision ($\frac{TP}{TP+FP}$) and high recall ($\frac{TP}{TP+FN}$); hence, our choice of F1-score ($2\frac{Precision\ *Recall}{Precision\ +Recall}$) as a performance measurement criteria. The area under the curve (AUC) represents the recall or true positive rate (TPR) and the false positive rate ($FPR = \frac{FP}{FP+TN}$). A high AUC value represents a model with high TPR and low FPR. A value of 0.5 represents a random guess; the larger than 0.5 the AUC value is, the better performance of the model. Using 10-fold cross validation, we run our training pipeline including (Random Forest, SVM, and ANN). In Table 2 we list the F1-Scores with the related AUC values.

Table 2: Result of the Cross - validation learning

Model	F1 - Sore	AUC
SVM	0.95	0.96
Random Forest (RF)	1	1
ANN	0.99	0.99

Below we show all figures describing the accuracy curves with related confusion matrixes for each individual model.

Table 3: Learning Curves (Accuracy)

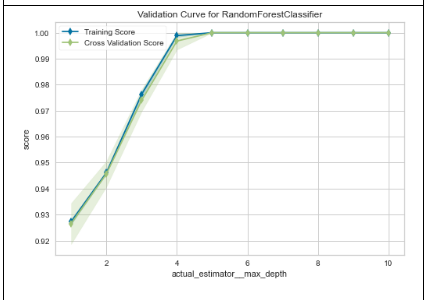
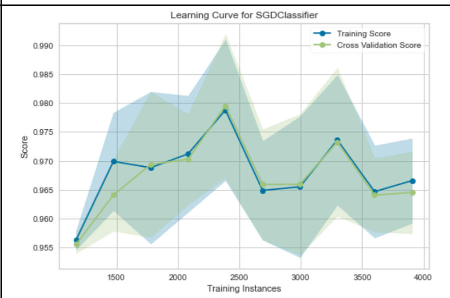
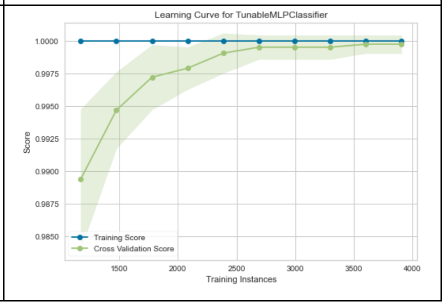
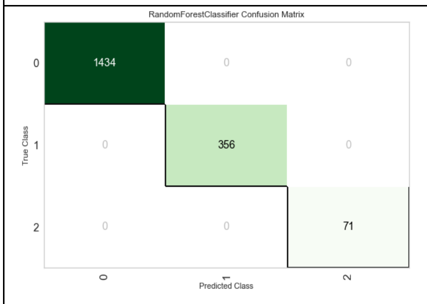
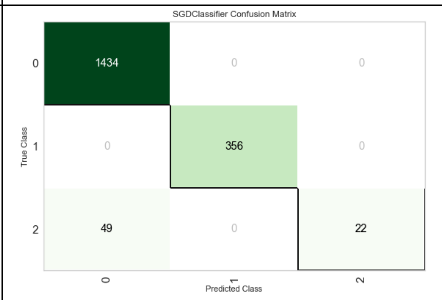
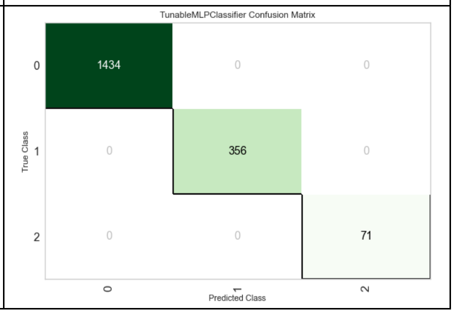
<table>
<tr><td colspan="3">Accuracy</td></tr>
<tr><td>RF</td><td>SVM</td><td>ANN</td></tr>
<tr><td>
</td><td>
</td><td>
</td></tr>
</table>

Table 4: Confusion Matrix

<table>
<tr><td colspan="3">Confusion Matrix</td></tr>
<tr><td>RF</td><td>SVM</td><td>ANN</td></tr>
<tr><td>
</td><td>
</td><td>
</td></tr>
</table>

Based on all previous evaluation criteria and parameters, we can conclude that all models are performing in a very good range with high accuracy and scoring. RF is showing more accuracy and robustness in terms of the AUC and this is because of the low features dimension and size.

5. Conclusion

In this paper, we have presented a comparison a mixed based design between hardware / data collection implementation and machine learning / air quality prediction model. Different ML models are trained and evaluated over our custom created dataset. RF model could provide the highest performance in terms of accuracy during the training and evaluation phases. Our design could be extended by adding more environmental measurements sensors in order to sense more air quality parameters. On other hand, more ML models could be trained over huge dataset in order to approximate the best possible prediction/classification model.

6. Future Scope

We would like to extend the current work by adding more environmental measurements sensors in order to sense more air quality parameters and upgrade the current Machine learning models by training them over huge dataset in order to approximate the best possible prediction/classification model and find a way to mass propagate the current work in all government sectors at the very least before going to the private sector.

References

[1] Hong Su and Gangtie Zheng, “A Partial Least Squares Regression - Based Fusion Model for Predicting the Trend in Drowsiness” IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL.38, NO.5, SEPTEMBER 2008.

[2] Fabian Friedrichs and Bin Yang, "Camera - based Drowsiness Reference for Driver State Classification under Real Driving Conditions" 2010 IEEE Intelligent Vehicles Symposium University of California, San Diego, CA, USA June 21 - 24, 2010.
[3] Zhang, Wei; Cheng, Bo; Lin, Yingzi," Driver drowsiness recognition based on computer vision technology." Published in: Tsinghua Science and Technology (Volume: 17, Issue: 3) Page (s): 354 - 362 Date of Publication: June 2012
[4] Eyosiyas Tadesse, Weihua Sheng, Meiqin Liu," Driver Drowsiness Detection through HMM based Dynamic Modeling. " 2014 IEEE International Conference on Robotics & Automation (ICRA) Hong Kong Convention and Exhibition Center May 31 - June 7, 2014. Hong Kong, China.

Author Profile

Dina Hamad Alghurair, works in Higher Institute for Communication and Training, PAAET, Kuwait, having Masters of Science in Computer Engineering and Information Systems, Gulf University, Bahrain, currently work as Specialist trainer (A) in Telecommunication and Navigation Institute in computer department since 1997.

Meshal Mansour Alnasheet works as Specialist Trainers (A) in Higher Institute for Communication and Training, PAAET, Kuwait, in Telecommunication and Navigation Institute in computer department since 1999.