Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

p-ISSN: 0972-6268
(Print copies up to 2016)
e-ISSN: 2395-3454

Vol. 23 | No. 1 | pp. 401-408 | 2024

Original Research Paper

https://doi.org/10.46488/NEPT.2024.v23i01.034

Open Access Journal

Machine Learning-based Calibration Approach for Low-cost Air Pollution Sensors MQ-7 and MQ-131

L. R. S. D. Rathnayake*, G. B. Sakura*†, N. A. Weerasekara* and P. D. Sandaruwan**

*Civil and Environmental Department, Faculty of Technology, University of Sri Jayewardenepura, Sri Lanka

**Department of Computer Science, University of Ruhuna, Matara, Sri Lanka

†Corresponding author: G.B. Sakura; sakurabogoda@sjp.ac.lk

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com

Received: 24-12-2022
Revised: 21-02-2023
Accepted: 22-02-2023

Key Words:
IoT
MQ-7
MQ-131
ThingSpeak
Machine learning
Neural network

ABSTRACT

Air quality is a vital concern globally, and Sri Lanka, according to WHO statistics, faces challenges in achieving optimal air quality levels. To address this, we introduced an innovative IoT-based Air Pollution Monitoring (APM) Box. This solution incorporates readily available Commercial Off-The-Shelf (COTS) sensors, specifically MQ-7 and MQ-131, for measuring concentrations of Carbon Monoxide (CO) and Ozone (O3) ,Arduino and "ThingSpeak" platform. Yet, those COTS sensors are not factory-calibrated. Therefore, we implemented machine learning algorithms, including linear regression and deep neural network models, to enhance the accuracy of CO and O3 concentration measurements from these non-calibrated sensors. Our findings indicate promising correlations when dealing with MQ-7 and MQ-131 measurements after removing outliers.

INTRODUCTION

Over time, the Earth's atmosphere has undergone changes, influenced by both natural events and human activities. Unfortunately, these alterations have led to an increase in air pollution, impacting humans and plant life negatively. The concerning shift is gradually making the Earth's atmosphere less conducive to the well-being of both humans and other living organisms (Choudhary & Garg, 2013). Addressing these challenges is crucial for a more optimistic environmental future.

As air pollution is a common problem that affects almost all the countries in the world, continuously measuring air pollutants keeps track of the well-being of the public, animals and plants, etc. Usually, economically well-established countries are concerned with measuring air quality to obtain sustainable goals with highly accurate real-time or conventional air quality monitoring systems. In Sri Lanka, air quality is mostly monitored by the Central Environmental Authority (CEA) and the National Building Research Organization (NBRO) using conventional chemical methods and the Mobile Ambient Air Quality Monitoring Lab (MAAQML).

It is possible to reduce air pollution by studying the changes in the composition of different types of gasses in the air and taking appropriate measurements using conventional air quality measuring equipment (Yi et al. 2015). However, due to their high cost, low and middle-income countries tend to use cost-effective sensors to measure air pollution and implement IoT devices (Yi et al. 2015). The air quality of a particular area can be monitored using sensors (gaseous and meteorological) and Arduino/Raspberry Pi (Malleswari & Mohana 2022). In research works carried out by Bathiya et al. (2016), Dhingra et al. (2019), Kennedy et al. (2018), Malhotra et al. (2020), Perumal et al. (2021), and Poonam et al. (2017), The authors presented recently developed systems to monitor air pollution using Arduino and Wireless Sensor Network (WSN) Technology. Here, we try to use the COTS sensors (Karagulian et al. 2019) in the market to measure the composition of the air accurately.

Si et al. (2020), have evaluated and calibrated low-cost particle sensors in ambient conditions using machine learning algorithms. In 2017, machine learning-based calibration methods have been used for COTS temperature sensors. Non-linear calibration models can be used for those non-linear relations shown between the reference instrument and the sensor (Yamamoto et al. 2017). Research works (Chen et al. 2018, Kumar & Sahu 2021, Okafor et al. 2020, Zimmerman et al. 2018) showcase the recently developed ML calibration approaches for COTS sensors for air quality monitoring.

While the most affordable sensors currently available in the market can measure air composition, they often suffer from quality and accuracy issues. As a result, accurate measurements remain challenging without appropriate calibration. Therefore, we tried to calibrate these sensors using machine learning methods, hoping to provide a cost-effective and easily approachable way to use low-cost sensors in air quality measurements.

MATERIALS AND METHODS

APM box was implemented with COTS sensors (MQ-7, MQ-131, DHT11) and controlling components as per the three-level architecture, such as the application layer, control layer (Arduino), and sensor layer (COTS sensors) (Fig. 2). Block diagram of the overall system is in Fig. 4. Components are mounted in a non-transparent plastic box with PVC pipe arms that includes a fan for enabling airflow from inside to outside (Fig. 1). It was designed to provide consistent airflow similar to the outside of the air quality measuring box. A9G Module was used to implement the GPRS connection using AT commands (Bogdanov & Mitrev 2021). The Real-Time Clock (RTC) module was used to record date and time for the best practice of recording data as the A9G module provides the current time obtained from the telecommunication network, and it is not very reliable due to the failures of the mobile network connection and module resets.

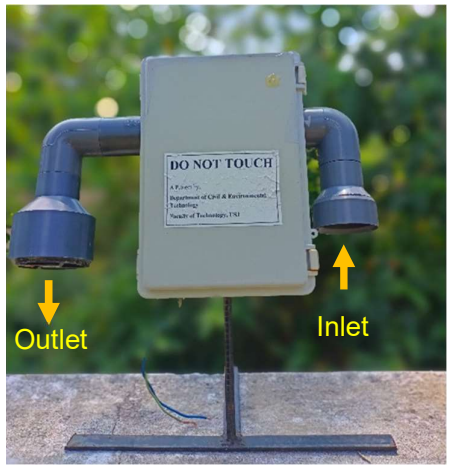


Fig. 1: APM box.

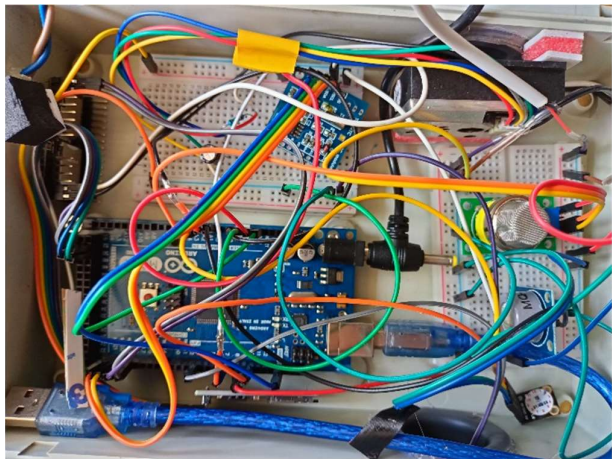


Fig. 2: Prototype of APM box.

Data Acquisition

Data collection was done using an APM box that contains sensors to measure CO, O_3, temperature, and humidity. COTS sensors are MQ-7, MQ-131, and DHT11, respectively. The air quality box was assembled with an Arduino mega microcontroller (Ismailov & Jo`rayev 2022). APM box was co-located with the National Building Research Organizational (NBRO) Automated Mobile Ambient Air Quality Monitoring (MAAQML) System at Colombo Municipal Council (CMC) Sri Lanka in Fig. 3 to record parallel readings. Thus, collected datasets were fed to the machine learning model to calibrate the COTS sensors in the APM box. The data was collected for 3 months approximately.

Data Storing and Transferring

The collected measurements by APM box were uploaded to the "ThingSpeak" (https://thingspeak.com/) databases using a GPRS connection and the IoT device. APM box is utilized to store a backup of the recorded data in case of any emergency, such as communication issues of mobile telecommunication networks.

An Arduino micro SD card module was used to store the backup data. The recorded dataset was sent to

Fig. 3: NBRO MAAQML at Colombo Municipal Council (CMC).

the "ThingSpeak" platform using its Representational State Transfer (REST) based web service using General Packet Radio Service (GPRS) of the (A9G Module) with 5 minutes frequency. This method allowed us to monitor the functionality of the A9G module and also get real-time measurements for the analysis.

Data Visualization and Analysis

"ThingSpeak" provides real-time data visualization using different kinds of graphs and allows users to customize the visualization by supporting plugins with user-defined calculations. It also provides MATLAB support for data analysis. However, we used Google Colaboratory (https://colab.research.google.com/) and conducted the data preprocessing analysis prior to applying ML algorithms.

Data cleaning was performed as the first step to eliminate possible errors once analysis started using Google Co-lab with pandas (https://pandas.pydata.org/) and "NumPy" (https://numpy.org/doc/stable/) library support.

Prediction Methods

Linear regression: Linear regression is a simple machine learning algorithm, and it is important to predict the association of ≥1 independent (predictor) variable with a continuous dependent (outcome) variable (Schober & Vetter 2021).

Simple linear regression: Simple linear regression is performed to determine the association between two quantitative variables. It can be represented as a straight line, as shown in Equation 1.

$$y = mx + c \quad \ldots(1)$$

In this equation, y stands for the outcome variable and x for the predictor. The slope and the interception are denoted by m and c, respectively. In sensor calibration, this method can be used to fit a linear equation to the NBRO measures of the gas and the gas measures of the air quality measuring device that is collected in parallel at the same time.

Multiple linear regression: This regression refers to a regression model that contains multiple independent variables. In sensor calibration, temperature, humidity, and wind flow can be used as multiple predictor variables, as shown in Equation 2.

$$y = m_0\, x_0 + m_1 x_1 + \ldots + m_n x_n + c \quad \ldots(2)$$

Where x_n represents the number of multiple independent variables, and y represents the dependent variable.

In this research work, First, we used the simple linear regression model to identify the correlation between only the NBRO sensor reading and the COTS sensor readings. Next, we used temperature and the COTS sensor readings as independent variables, while the NBRO sensor readings are the dependent variables of a multiple linear regression model.

Feed-forward neural network: A feed-forward neural network is a method of ML that does not have any cycle in the connections between the nodes. As input is only processed in one direction, the feed-forward model is the simplest NN model (Fig. 5). The data may go via a number of hidden nodes, but it always moves forward and never backward.

In this research, first, we designed a feed-forward neural network with an input layer containing one node to represent non-calibrated sensor value. The input layer contained 01 Node and 02 hidden layers, with 64 per layer added next to the input layer. Next, another hidden layer was added with 64 before the output layers that consisted of 01 Node. Each

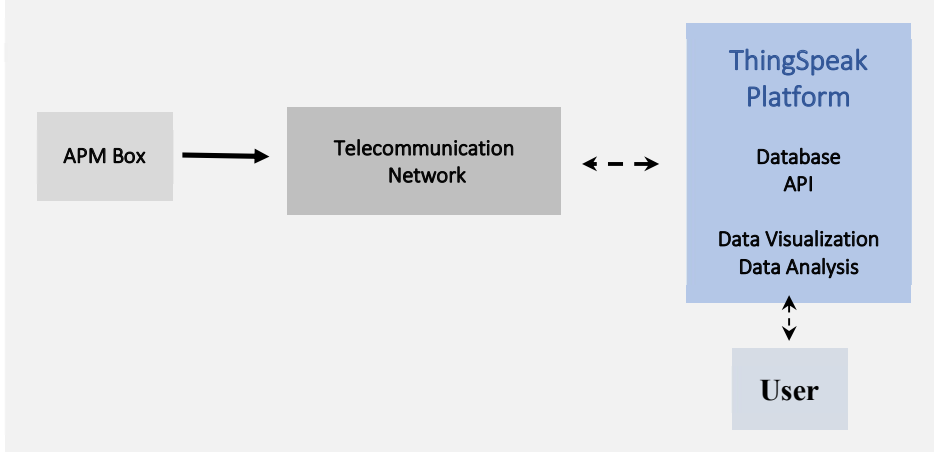


Fig. 4: Block diagram of the system.

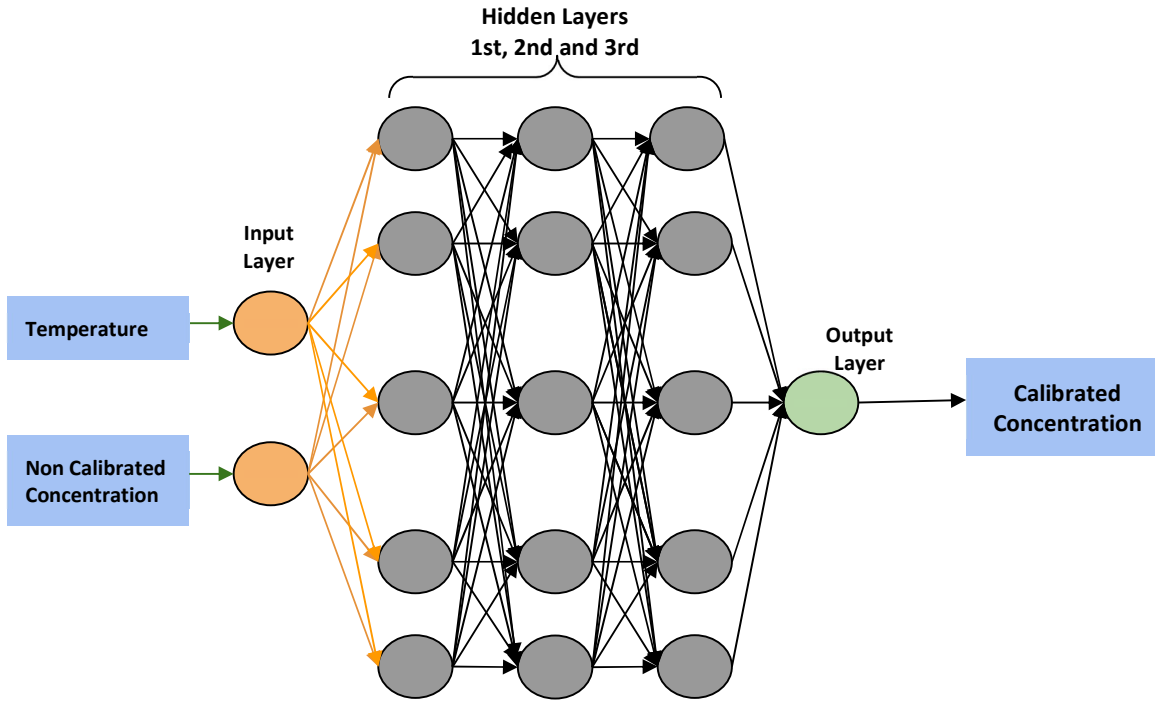


Fig. 5: Illustration of the neural network (NN) layers.

hidden layer has the Rectified Linear Unit ("ReLu") activation as the activation function. Layer weight initializers define the approach for setting the initial random weights of Keras layers. Then, we set the layer initializer as normal to initialize the weights of the hidden layers. Finally, In the output layer, we used the linear function as the activation function, as it needs to get the calibrated concentration as the output value of the NN. There were 8577 total trainable parameters in this model.

As the second approach, we changed the input layer by adding two nodes for the non-calibrated sensor reading and temperature.

RESULTS AND DISCUSSION

The data collection phase was continued for approximately 3 months starting from June 2022, with dynamically parallel to NBRO MAAQML at CMC for CO and O_3 gasses. As these are low-cost, non-factory calibrated COTS sensors, their readings include a higher number of outliers than the amount that can be expected from calibrated higher-accuracy devices. Therefore, we implemented the following procedures before utilizing ML algorithms for readings of the low-cost sensors, specifically the MQ-7 and MQ-131.

First, the measurements were taken at a similar meteorological environment during the same time with all the sensors. Next, we compared the sensor values and selected the sensors that gave the most similar results to build up the APM boxes. This procedure was conducted to identify the malfunctioned sensors before locating them in the field.

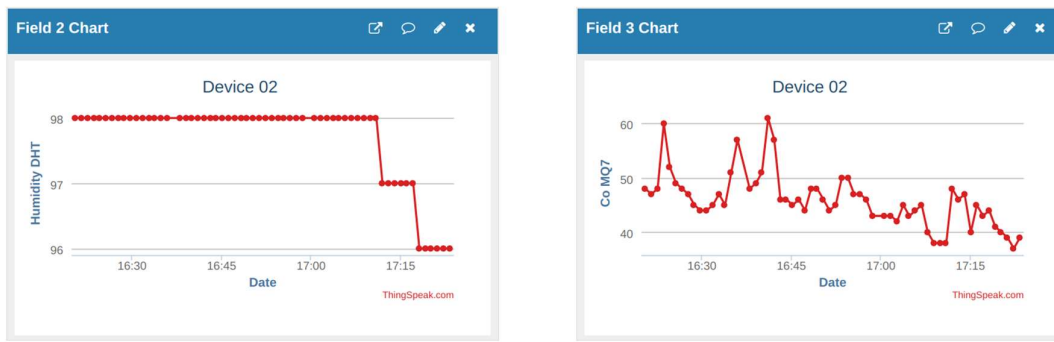


Fig. 6: Real-time data visualization of the "ThingSpeak" platform.

Fig. 6 shows the "ThingSpeak" visualizing the data reading in real time with/without "ThingSpeak" plugins. "ThingSpeak" plugins can be used to set custom functions to visualization that help to compare the analog readings and the values determined by those readings in real-time. After collecting the data, data preprocessing was done, and the datasets were prepared for applying ML algorithms after eliminating the outliers using the IQR method and standard deviation method. The results were evaluated using the Mean Squared Error (MSE) and Coefficient of determination (R-squared value).

MQ Sensors

Usually, low-cost sensors such as MQ-7 and MQ-131 are not accurate enough to use in monitoring pollutant concentrations as they are not factory-calibrated. Here, we have implemented a calibration equation for both MQ-7 and MQ-131 sensors using the direct-taking sensor raw values that had been collected for presenting a calibration method to take accurate CO and O_3 concentrations using ML methods. Fig. 7 shows the generalized circuit for MQ sensors which was used with the Arduino Mega microcontroller.

The following formula, Equation 3 can be used to derive the values of sensor resistance for different gasses using MQ Sensors in modules that use an MOS (Metal Oxide Semiconductor) sensor.

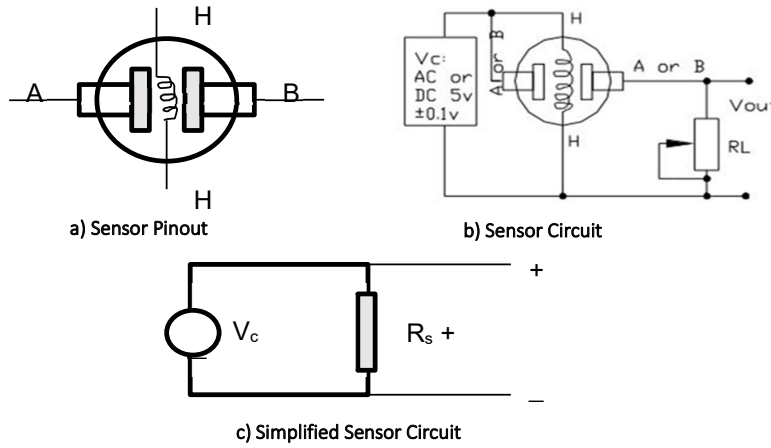
Using V = I x R (Ohm's Law)

$$V_{RL} = [V_C/(R_S + R_L)] \times R_L$$

$$R_S = [(V_C/V_{RL}) - 1] \times R_L \quad \ldots(3)$$

Where, V_C = Circuit Voltage

V_{RL} = Changing Analog Voltage depending on gas Concentration



*Rs – resistance of sensor that changes depending on concentration of gas

*R_L – Resistance of sensor at a known concentration without the presence of other gases or fresh air

Fig. 7: Generalized MQ sensor circuit.

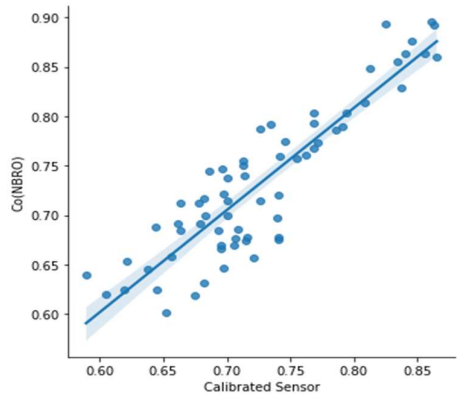


Fig. 8: Simple linear regression model performance-predicted calibrates sensor values Vs NBRO reference data for CO.

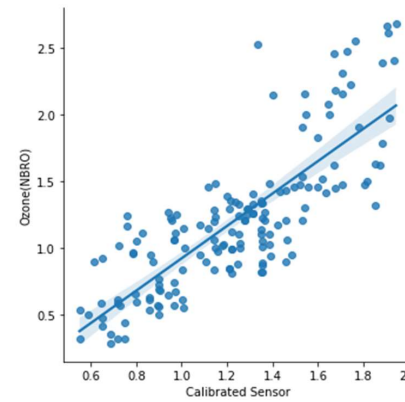


Fig. 9: Multiple linear regression model performance - predicted calibrates sensor values Vs NBRO reference data for CO.

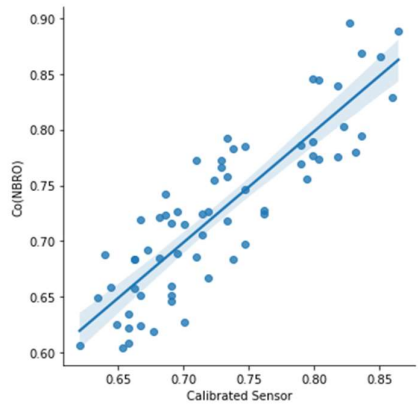


Fig. 10: Simple linear regression model performance - predicted calibrates sensor values Vs NBRO reference data for ozone.

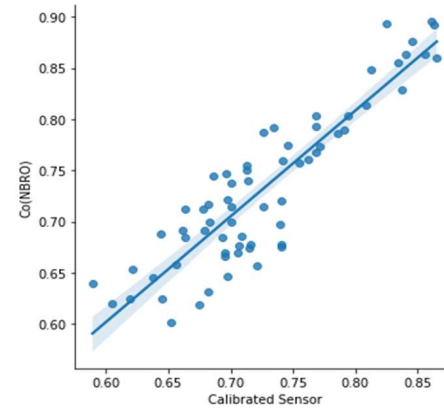


Fig. 11: Multiple linear regression model performance - predicted calibrates sensor values Vs NBRO reference data for ozone.

Gas concentrations and analog voltage generated by sensors are proportionally variate and can be applied to a linear graph to showcase its ideal relation (Karamchandani 2016). Further, it specifies that the analog voltage increases when the gas concentration increases.

Linear Regression on Ozone (O_3) Data

A simple Linear regression method was applied to the Ozone dataset using the "Scikit-Learn" regression functions. The Fig. 8 shows the correlation between the NBRO Ozone data and the MQ-131 sensor data. MSE and R-squared values were 0.105 and 0.59, respectively. Lower MSE shows a lower average squared difference between the MQ-131 readings vs NBRO Ozone sensor data. Mid-level of R-squared shows a considerable correlation between the accurate readings and the low-cost sensor analog reading.

Using the "Scikit-Learn" regression library, a multiple regression model was trained using the particular temperature values along with the non-calibrated sensor readings as independent variables. Fig. 9 shows the result of the test dataset, which provided MSE, MAE, and R-squared values as 0.107, 0.260, and 0.625. Respectively. This shows that a better correlation can be identified when we use independent variables in multiple linear regressions with lower errors.

Linear Regression on Carbon Monoxide (CO) Data

Fig. 10 shows the line obtained from the simple linear regression model using the "Scikit-Learn" library. After fine-tuning the model, the MSE and R-squared values were 0.0012 and 0.80, respectively. This indicates that the average difference between observed and predicted values is much lower and that there is a higher correlation between the MQ-7 readings and the NBRO sensor readings.

Similar to the multiple linear regression model used for Ozone with temperature values, we trained multiple linear regression models for CO readings. As shown in Fig. 11, the model performed well compared to the simple regression model with higher co-relation and fewer errors, as shown in Table 1.

Deep Neural network for both MQ-7 and MQ-131

We trained a deep neural network with an input layer of size 2 for non-calibrated sensor reading and temperature, as deep neural networks can be used for regression problems with multiple input variables. We observed that the deep neural network model performed similarly to the regression models with lower MSE and higher R-squared value, as shown in Table 2 and Fig. 12. We trained to use a lesser number of nodes to avoid the overfitting of the model. After fine-tuning, the trained model was performed on

Table 1: Regression model performance analysis.

Sensor	Independent variables	Coefficients	Intercept	MSE	MAE	R-Squared
MQ-131 (Ozone)	1. Non-Calibrated Sensor value 2. Temperature	1) 0.0356 2) 0.0647	-3.2620	0.1067	0.2596	0.6254
	1. Non-Calibrated Sensor value	1) 0.0393	-1.6626	0.1053	0.2520	0.5895
MQ-7 (Carbon Monoxide)	1. Non-Calibrated Sensor value 2. Temperature	1) 0.0046 2) -0.0019	0.0877	0.0012	0.8010	0.8010
	1. Non-Calibrated Sensor value	1) 0.0047	0.0203	0.0013	0.0314	0.7799

Table 2: Similarities showcased in deep neural network model.

Sensor	The input layer of the model	MSE	MAE	R-Squared
MQ-131 (Ozone)	1. Non-Calibrated Sensor value 2. Temperature	0.1349	0.5265	0.2833
	1. Non-Calibrated Sensor value	0.1493	0.2880	0.4181
MQ-7 (Carbon Monoxide)	1. Non-Calibrated Sensor value 2. Temperature	0.0013	0.0303	0.7858
	1. Non-Calibrated Sensor value	0.0015	0.0321	0.7572

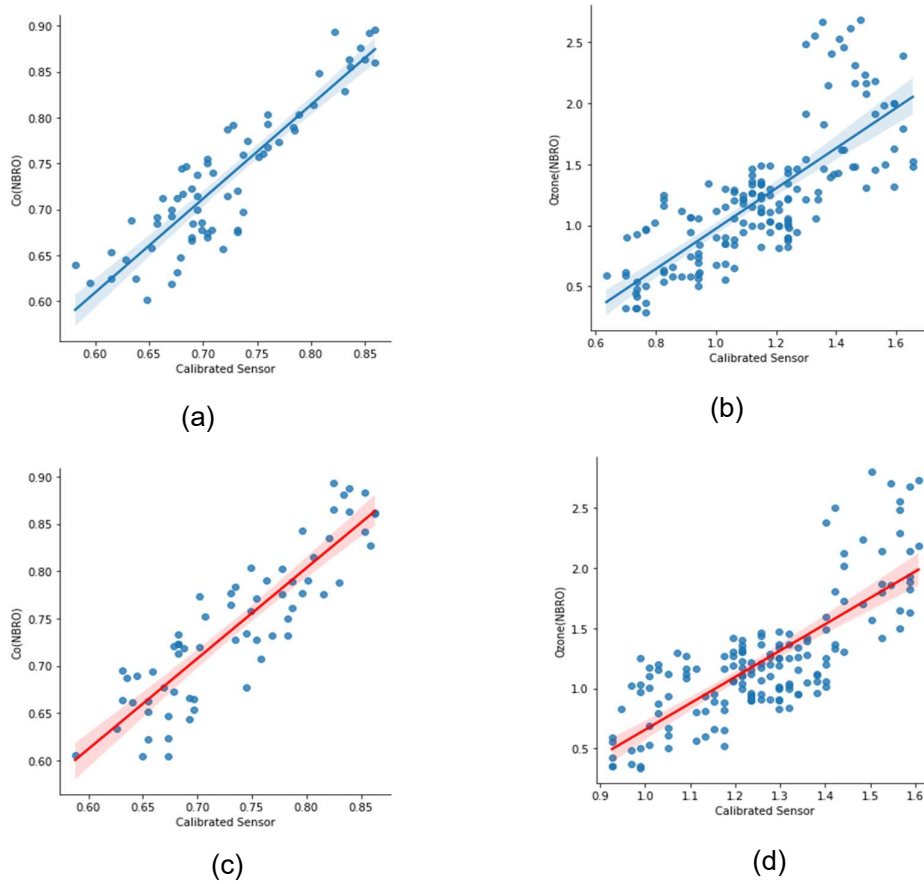


Fig. 12: Neural network (NN) performance analysis.
(a) Calibrated CO sensor values with temperature as inputs vs NBRO reference data.
(b) Calibrated Ozone sensor values with temperature as inputs vs NBRO reference data.
(c) Calibrated CO sensor values Vs NBRO reference data.
(d) Calibrated Ozone sensor values Vs NBRO reference data.

the test datasets of both CO and O_3 datasets separately. The MSE, MAE, and R-squared were calculated by using "Scikit-Learn" library metrics for ML. When there are more input variables, it shows an increase in both ML models.

MQ-7 (CO) sensor readings show a higher correlation between the observed and predicted sensor readings with a lower MSE, indicating a very lower average squared difference of the data. Also, higher R-squared values show the correlation between the sensors.

It is better to have a lesser number of nodes per hidden layer and total trainable parameters with respect to the size of the data set typically (Abiodun et al. 2018). The generalization of this NN is shown as it performed comparatively better for both MQ-7 and MQ-131 senor types with a lesser number of nodes (64) per hidden layer and with lesser total trainable parameters (8577).

CONCLUSION AND FUTURE WORKS

Here, our main goal was to find an ML method for calibrating the low-cost sensors MQ-7 and MQ-131 to monitor air pollution accurately using sensor analog readings. We observed that both simple linear regression and deep neural network models performed better for the calibration. It is more cost-effective than the usual conventional criteria followed using calibration gas. In this research work, we observed that the NN model (64 nodes per layer and 8577 trainable parameters) is more accurate than the linear regression. Also, both ML models performed the calibration better when tested with relevant temperature and sensor analog values. Therefore, this research can be extended using other co-related values such as wind speed, wind direction, and humidity as the input layer variables in the NN model. Also, getting simultaneous readings using a set of similar low-cost non-calibrated sensors in similar environmental conditions instead of getting readings with just one single sensor module is more convenient for monitoring the effects of sensor functionality changes over a longer time period. It would help to increase the reliance and reliability of the ML algorithm for the calibration of that kind of sensor.

ACKNOWLEDGEMENT

This study is supported by the University Research Grant, Faculty of Technology, University of Sri Jayewardenepura, Sri Lanka. (Grant No: ASP/01/RE/FOT/2017/76). And, National Building Research Organization (NBRO), Sri Lanka supported with appropriate certified reference air pollution data.

REFERENCES

Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A. and Arshad, H. 2018. State-of-the-art in artificial neural network applications: A survey. Heliyon, 4(11): e00938. doi:10.1016/j.heliyon.2018.e00938

Bathiya, B., Srivastava, S. and Mishra, B. 2016. Air Pollution Monitoring Using Wireless Sensor Network. 2016 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), 19-21 December 2016, Pune, Maharashtra, India, IEEE, Piscataway, NJ, pp. 1821-1829.

Bogdanov, L. and Mitrev, H. 2021. Flash Programming Microcontrollers Over the GSM Network. 2021 International Conference on Software, Telecommunications, and Computer Networks (SoftCOM), 23-25 September 2021, Hvar, Croatia, IEEE, NJ, pp. 541-551.

Chen, C.C., Kuo, C.T., Chen, S.Y., Lin, H., Chue, J.J., Hsieh, Y.J. and Huang, C.M. 2018. Calibration of Low-Cost Particle Sensors By Using the Machine-Learning Method. IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), 26-30 Oct. 2018, Shangri La, Chengdu, China, IEEE, NJ, pp. 111-114.

Choudhary, D.M. and Garg, V. 2013. Causes, Consequences, and Control of Air Pollution. Springer, Cham.

Dhingra, S., Madda, R.B., Gandomi, A.H., Patan, R. and Daneshmand, M. 2019. Internet of Things mobile–air pollution monitoring system (IoT-Mobair). IEEE Inter. Things J., 6(3): 5577-5584. doi:10.1109/JIOT.2019.2903821

Ismailov, A. and Jo`rayev, Z. 2022. Study of Arduino microcontroller board. Sci. Edu. J., 3(3): 61.

Karagulian, F., Barbiere, M., Kotsev, A., Spinelle, L., Gerboles, M., Lagler, F. and Borowiak, A. 2019. Review of the performance of low-cost sensors for air quality monitoring. Atmosphere, 10(9): 506.

Karamchandani, S. 2016. Pervasive Monitoring of Carbon Monoxide And Methane Using Air Quality Prediction. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 16-18 March 2016, New Delhi, India, IEEE, NY, pp. 3803-3807.

Kennedy, O., Etinosa, N., Odusami, M., Samuel, J. and Oluga, O. 2018. A smart air pollution monitoring system. Int. J. Civ. Eng. Technol., 9(9): 799-809.

Kumar, V. and Sahu, M. 2021. Evaluation of nine machine learning regression algorithms for calibration of low-cost PM2.5 sensor. J. Aerosol Sci., 157: 105809. doi:https://doi.org/10.1016/j.jaerosci.2021.105809

Malhotra, M., Aulakh, I.K., Kaur, N. and Aulakh, N.S. 2020. Air Pollution Monitoring Through Arduino Uno. ICT Systems and Sustainability, Singapore.

Malleswari, S.M.S.D. and Mohana, T.K. 2022. Air pollution monitoring system using IoT devices: A review. Mater. Today Proceed., 51: 1147-1150. doi:https://doi.org/10.1016/j.matpr.2021.07.114

NBROS. n.d. Strengthen the Air Quality Monitoring Capacity of NBRO with Automated Mobile Ambient Air Quality Monitoring System. National Building Research Organization, Ministry of Defence, Sri Lanka.

Okafor, N.U., Alghorani, Y. and Delaney, D.T. 2020. Improving data quality of low-cost IoT sensors in environmental monitoring networks using data fusion and machine Learn. Appr. ICT Expr., 6(3): 220-228. doi:https://doi.org/10.1016/j.icte.2020.06.004

Perumal, B., Deny, J., Alekhya, K., Maneesha, V. and Vaishnavi, M. 2021. Air Pollution Monitoring System by using Arduino IDE. Paper Presented at the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 4-6 August 2021, Coimbatore, Tamil Nadu, Hindustan Institute of Technology, Coimbatore, pp. 797-802.

Poonam, P., Ritik, G., Sanjana, T. and Ashutosh, S. 2017. IoT-based air pollution monitoring system using Arduino. IRJET, 4(10): 515-526.

Schober, P. and Vetter, T.R. 2021. Linear regression in medical research. Anesth Analg., 132(1): 108-109. doi:10.1213/ane.0000000000005206

Si, M., Xiong, Y., Du, S. and Du, K. 2020. Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine-learning methods. Atmos. Meas. Tech., 13(4): 1693-1707. doi:10.5194/amt-13-1693-2020

Yamamoto, K., Togami, T., Yamaguchi, N. and Ninomiya, S. 2017. Machine learning-based calibration of low-cost air temperature sensors using environmental data. Sensors, 17(6): 1290.

Yi, W.Y., Lo, K.M., Mak, T., Leung, K.S., Leung, Y. and Meng, M.L. 2015. A survey of wireless sensor network-based air pollution monitoring systems. Sensors, 15(12): 31392-31427.

Zimmerman, N., Presto, A.A., Kumar, S.P.N., Gu, J., Hauryliuk, A., Robinson, E.S. and Subramanian, R. 2018. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos. Meas. Tech., 11(1): 291-313. doi:10.5194/amt-11-291-2018.

ORCID DETAILS OF THE AUTHORS

L. R. S. D. Rathnayake: https://orcid.org/0000-0002-7405-7785
G. B. Sakura: https://orcid.org/0000-0003-0533-9380
N. A. Weerasekara: https://orcid.org/0000-0001-6809-8778
P. D. Sandaruwan: https://orcid.org/0009-0007-4819-8684