Air Quality Monitoring And Prediction Using IOT And Machine Learning Approaches

T.S. Kitchilan[*], M.D.B.C.K. Abeyratne[**], Prof. E.P.S.K Edirisweera[***]

[*] Department of Computer Science, Uva Wellassa University of Sri Lanka
[**] Department of Computer Science, Uva Wellassa University of Sri Lanka
*** Faculty of Applied Sciences, Uva Wellassa University of Sri Lanka

DOI: 10.29322/IJSRP.12.03.2022.p12307
http://dx.doi.org/10.29322/IJSRP.12.03.2022.p12307

Paper Received Date: 7th February 2022
Paper Acceptance Date: 26th February 2022
Paper Publication Date: 4th March 2022

Abstract- Air pollution today is on the rise in urban areas mainly due to vehicular emissions, manufacturing plants and industries accumulating harmful pollutants such as particulate matter, carbon monoxide and ozone posing a serious health risk to the people. The nature of air is also influenced by multi-dimensional elements including area, time, period during the year and many other factors. Implementing a cost effective, efficient air quality monitoring and forecasting system which gather information and provide evaluations on air pollution is hence required, with numerous experts utilizing the big data analytics approach for contemplating, assessing, and predicting air quality recently. In this study an air quality monitoring device using arduino based sensors was developed to capture air quality parameters followed by the calibration of sensors to improve accuracy. Using data from the air quality monitor over 3 months, forecast models for pollutant concentrations were built using random forest as the base model to forecast PM2.5 air quality index. The IoT device developed performed with a significant improvement in accuracy after calibration of sensors. Using random forest prediction model as base model outperformed the linear regression, decision tree and neural network models in prediction accuracy while the runtime performance improved significantly with split data technique and Bayesian approach used for parameter tuning. This study demonstrates the potential of cost effective IoT based, real-time air quality monitoring solution and the potential of forecasting using the air quality prediction model.

Index Terms- Air Quality, IoT, Prediction Model

I. Introduction

Air quality refers to the degree to which the air is free of pollutants in a particular location at a given time. Natural pollution sources include wildfires, volcanoes and dust storms while man-made sources include emissions from vehicular exhaust, power plants factories etc. The concerns over air quality is increasing daily due to the pollution of air caused mainly due to human action which leads to major health concerns. It is attributed as the main environmental health issue and as a major cause of premature human death causing heart disease, stroke, lung disease and cancer [1].

Current air quality monitoring methods include continuous monitoring methods, gravimetric particulate methods and passive monitoring methods. Air pollutants require continuous monitoring due to its high resolution to obtain hourly or daily average concentrations.

Air quality prediction using conventional mathematical and statistical approaches in the past have been inefficient until the advancements in the field of big data and machine learning. With the development in this field more research adopted this methodology to efficiently forecast air quality. Current data driven methods of forecasting readings from an air quality monitoring station include the use of neural network for global factors and the use of linear regression based temporal predictor for local factors [2].

In order to bring the level of pollution to a minimum, there is a necessity of cost-effective, accurate mechanisms to monitor air quality in real time and to forecast air quality parameters to conduct studies on ambient air quality.

II. Literature Review

A. IoT for monitoring air quality

IoT can be defined as “An open and comprehensive network of intelligent objects that have the capacity to auto-organize, share information, data and resources, reacting and acting in face of situations and changes in the environment” [3].

IoT is a popular concept in providing successful solutions to air quality monitoring. The “GasMobile” system, which is a participatory air quality monitoring system, has been implemented using a self-developed air quality sensing device in combination with android smartphones. The concept is that the citizen participates in the collection of air quality information [4].

Devarakonda proposed an air quality system consisting of sensing devices deployed on vehicles such as public transportation and personal vehicles which consists of an arduino microcontroller, PM sensor, CO sensor and a modem which will upload data to the server [5].

Yang & Li developed an air quality system with a microprocessor, CO sensor and VOC sensor which takes readings when the user executes a mobile application which is then transmitted to the mobile phone [6]. However, the system measures air quality parameters on demand only.

B. Air quality prediction

Fluctuation of air quality parameters is a complex phenomenon since it is a result of a combination of emissions, meteorological, demographic and terrain factors. Pollutants due to vehicular emissions such as particulate matter (PM10, PM2.5) and Carbon Monoxide (CO) show temporal variations throughout the day with time and the peak in vehicular activity.

An air quality forecasting model gives a deterministic description of the greater air problem and a tool used to explain the relationship between air quality determining parameters [7].

Air quality prediction using a hidden semi-Markov model for the prediction of PM2.5 proved to provide a reasonable accuracy [8] while the Bayesian approach to air quality prediction was better than the semi-Markov approach [9].

Overfitting is a challenge in machine learning models where the model gives highly accurate results during training but poor results during testing. Non-linear models are however superior to linear models because they capture non-linear relationships in the data set of pollutant concentrations [10].

III. Methodology

A. Air quality monitoring device

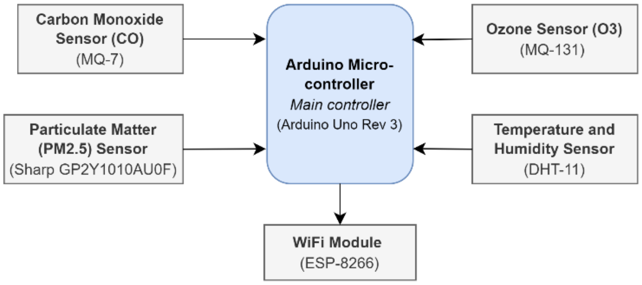

Carbon Monoxide Sensor (CO) (MQ-7)
Ozone Sensor (O3) (MQ-131)
Arduino Micro-controller
Main controller
(Arduino Uno Rev 3)
Particulate Matter (PM2.5) Sensor (Sharp GP2Y1010AU0F)
Temperature and Humidity Sensor (DHT-11)
WiFi Module (ESP-8266)

Figure. 1: Device design

Figure 1 shows the air quality monitoring device design, with an arduino uno rev3 microcontroller and 4 other sensors to capture data relevant to ozone (MQ-131 sensor), carbon monoxide (MQ-7 sensor), particulate matter 2.5 (Sharp GP2Y1010AU0F sensor), temperature and relative humidity (DHT11 sensor). A Wi-Fi module (ESP 8266) is used to transmit data directly to the server.

The sensors used may contain errors which cause differences between the sensor output and the actual output, which can be eliminated/reduced to improve sensor performance referred to as sensor calibration. Calibration process was carried out for each individual sensor to improve the accuracy of sensor output.

MQ gas sensors calibration

The 2 MQ sensors were initially exposed to a very high concentration of N2 (Nitrogen) gas in an enclosure to observe the reading they produce at ~0 ppm of CO and Ozone and noted.

Next the MQ7, CO sensor was exposed to known concentrations of CO gas within the enclosure and the readings for each concentration level were recorded.

A similar process was followed for the MQ131, sensor where the sensor was exposed to known concentrations of ozone produced by an ozonizer within an enclosure recording the output.

Values from the above procedure for each sensor were used to plot the concentration vs. sensor output graph, of which coordinates were used in the coding of the two sensors.

Sharp GP2Y1010AU0F – PM2.5 sensor

An approach different to MQ gas sensors had to be followed in order to calibrate the PM2.5 sensor, due to the difficulty in producing specific concentrations of particulate matter with a specific size range (below 2.5µm). Therefore, a software-based technique was followed for calibrating the PM2.5 sensor considering the following 2 types of errors.

1. Offset Correction 2. Scalar Coefficient

1. Offset correction

Offset is the sensor output voltage produced even during a clean, zero particulate matter environment referred to as Voc. The data sheet suggests a value of 0.6V as the typical Voc which may differ due to production errors.

The determination of Voc value during this study was done by using the value of 0.6V initially and updating it dynamically when the lowest output voltage is detected.

2. Scalar coefficient

The typical variation of output voltage vs. dust density of the sensor follows a linear relationship up to around 500µg/m^3.

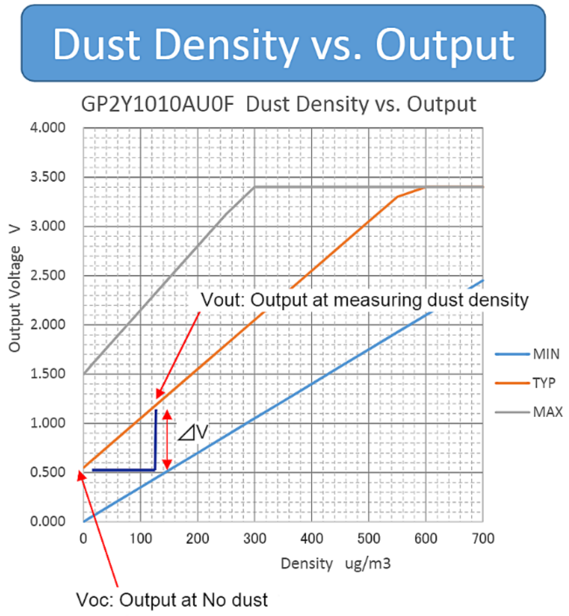


Figure 2: Dust density vs. pm2.5 sensor output

Figure 2 shows the graph of minimum, typical and maximum output voltages against the perceived PM2.5 concentration.

The PM2.5 concentration is calculated by,

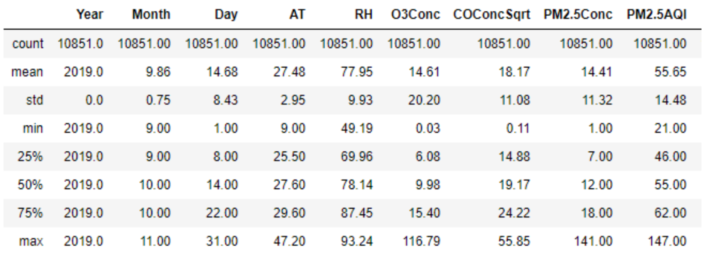
PM2.5 concentration = (Vout – Voc)/K*100 where,

- Vout : Output at no dust
- Voc : Output at measuring dust
- K : Sensitivity

The output voltage for 2 separate concentrations ~100µg/m^3 and ~400µg/m^3 were obtained as ~0.7V and ~2.8V with the help of MAAQML respectively thereby providing sensitivity value (K) of 0.7.

B. Prediction Model

Ambient air quality is a result of a set of complex processes occurring together throughout the day. The dataset used in this study were obtained from the air quality monitoring device developed and placed in Colombo for monitoring. The raw data obtained were normalized into hourly mean values as it was assumed sufficient to capture temporal changes in air quality and split into two as 75% training data and 25% for testing and model validation. The prediction model takes 8 input parameters in order to predict the air quality index value for a particular hour.



	Year	Month	Day	AT	RH	O3Conc	COConcSqrt	PM2.5Conc	PM2.5AQI
count	10851.0	10851.00	10851.00	10851.00	10851.00	10851.00	10851.00	10851.00	10851.00
mean	2019.0	9.86	14.68	27.48	77.95	14.61	18.17	14.41	55.65
std	0.0	0.75	8.43	2.95	9.93	20.20	11.08	11.32	14.48
min	2019.0	9.00	1.00	9.00	49.19	0.03	0.11	1.00	21.00
25%	2019.0	9.00	8.00	25.50	69.96	6.08	14.88	7.00	46.00
50%	2019.0	10.00	14.00	27.60	78.14	9.98	19.17	12.00	55.00
75%	2019.0	10.00	22.00	29.60	87.45	15.40	24.22	18.00	62.00
max	2019.0	11.00	31.00	47.20	93.24	116.79	55.85	141.00	147.00

Figure 3: Descriptive statistics

Figure 3 shows descriptive statistics of the input and output parameters of the prediction model with Year, Month, Day, Atmospheric Temperature (AT), Relative Humidity (RH), Ozone (O3Conc), Square root of Carbon Monoxide (COConcSqrt) and PM2.5 concentrations (PM2.5Conc).

Machine learning model selection

The process of selecting a machine learning model, based upon factors such as performance, complexity, maintainability and available resources is referred to as "Model Selection". The main techniques of model selection are probabilistic and resampling methods. Using the probabilistic approach several basic machine learning models were adopted such as decision trees and random forest where random forest provided the best accuracy during the initial basic implementation. Therefore, "random forest" model was selected as the most suitable base model for implementation.

Prediction model implementation

Random forest is an ensemble learning technique used for both classification and regression tasks, that incorporates the construction of multiple decision trees either predicting the class in classification or the value in regression of individual trees. It is known to have a high prediction accuracy among all machine learning techniques.

Initially data cleaning was done in order to remove features that have less than 50% values followed by removal of rows that contain null values in order improve training and prediction accuracy. One hot encoding was applied to categorical variable "Day_Of_Week" in order to use it in the machine learning model.

Mon	1	0	0	0	0	0	0
Tue	0	1	0	0	0	0	0
Wed	0	0	1	0	0	0	0
Thu	0	0	0	1	0	0	0
Fri	0	0	0	0	1	0	0
Sat	0	0	0	0	0	1	0
Sun	0	0	0	0	0	0	1

Table: 1

Table 1 shows the sample data after one hot encoding for day_of_week parameter.

Next the data set was divided into 2 sets as original dataset and expanded dataset.

- Original data set: Contains only one half of the total data set
- Expanded data set: Contains the entire data set.

Each dataset was then divided into 75% training and 25% testing data, building initial models on the "original" data set, and then on "expanded" data set to improve the initial model. The 75% training data was used to train the prediction model and was tested by making predictions on the 25% testing data. The training process generated several trees based on the dataset.

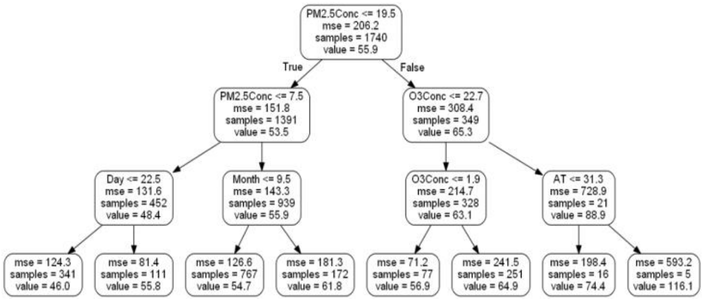


Figure 4: Simplified generated tree

Figure 4 shows a simplified version of one of the several hundreds of trees generated by the prediction model where the decision-making process is based on each parameter value at each split node.

The prediction model with all parameters is named as "Expanded_all_features" and was analyzed to obtain the features with 95% importance in order perform feature reduction and build a tuned model named "Exapanded_reduced_features" thereby increasing the efficiency of the prediction model.

The variables which are used as the input for the prediction model, used to generate the predicted output "Air Quality Index" is affected by the input parameters in varying degrees.

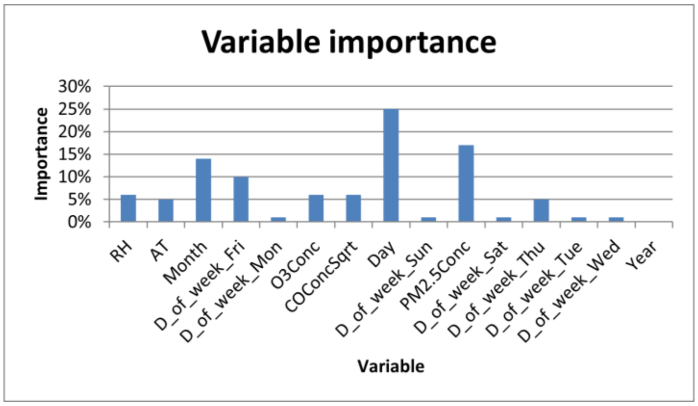


Figure 5: Variable Importance

Figure 5 shows the variable importance of the features that are used to build the prediction model. “Day” variable has the highest importance (25%) followed by “PM2.5Conc” (17%) and “Month” (14%) variables which have the greatest impact on the predicted PM2.5AQI value.

IV. RESULTS

A. Air quality device testing and Results

The air quality monitoring device must be tested for accuracy of its readings after the calibration process to verify functionality and reliability of the system before deploying it to the actual monitoring site to gather air quality information. Testing for accuracy was carried out with “The Mobile Ambient Air Quality Monitoring Lab” (MAAQML) of the National Building Research Organization of Sri Lanka.

The device was placed at the same location with the MAAQML, and random readings were taken throughout the day at the same time against MAAQML. The two readings from the device and the MAAQML were compared against to obtain an average percentage accuracy value.

Parameter	Sensor	No. of readings	Mean % accuracy	Mean % error
O3	MQ-131	770	90.54	9.46
CO	MQ-7	760	91.89	8.11
PM2.5	Sharp GP2Y1010AU0F	810	89.77	10.23
Temperature	DHT-11	800	91.55	8.45
Relative Humidity	DHT-11	800	89.59	10.41

Table: 2

Table 2 shows the results of sensor accuracy validation for the air quality monitoring device for each sensor. It is seen that the overall device accuracy is ~91% which is a significant improvement over current non-calibrated devices which rates approx. 80-85% accuracy.

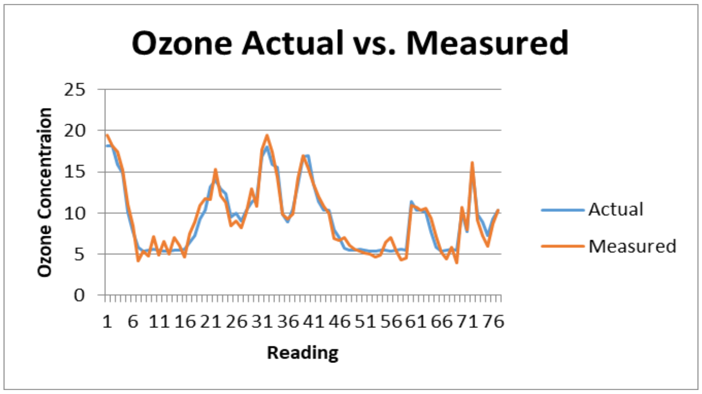


Figure 6: Ozone actual Vs. measured

Figure 6 shows the graph of ozone gas concentration measured by the air quality monitoring device and the actual recorded ozone gas concentration by the MAAQML, based on 770 random readings taken throughout the day which exhibits a mean percentage accuracy of 90.54%.

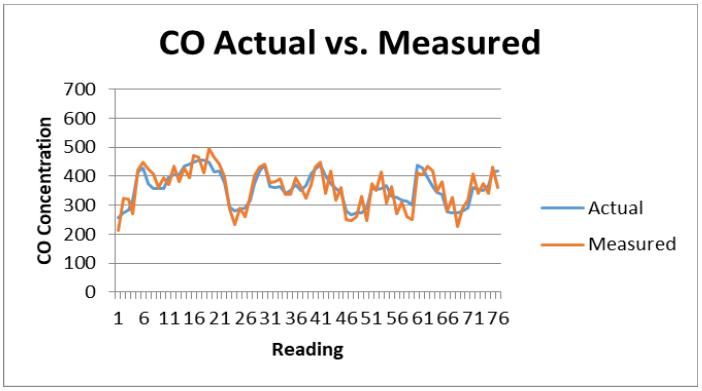


Figure 7: CO actual Vs. measured

Figure 7 shows the graph of carbon monoxide gas concentration measured by the air quality device and the actual recorded carbon monoxide concentration by the MAAQML, based on 760 random readings taken throughout the day which exhibits a mean percentage accuracy of 91.89%.

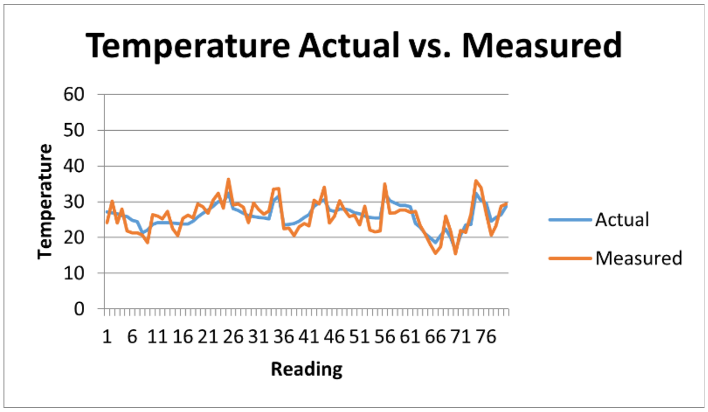


Figure 8: Temperature actual Vs. measured

Figure 8 shows the graph of temperature measured by the air quality monitoring device and the actual recorded temperature by the MAAQML, based on 810 random readings taken throughout the day which exhibits a mean percentage accuracy of 89.77%.

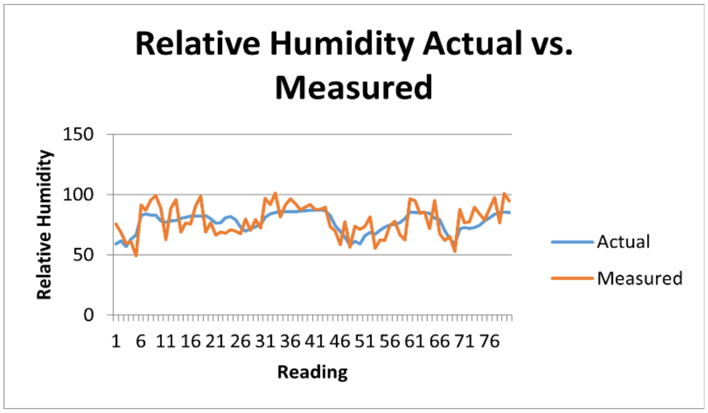


Figure 9: Relative humidity actual Vs. measured

Figure 9 shows the graph of relative humidity measured by the air quality device and the actual recorded relative humidity by the MAAQML, based on 800 random readings taken throughout the day which exhibits a mean percentage accuracy of 91.55%.

http://dx.doi.org/10.29322/IJSRP.12.03.2022.p12307

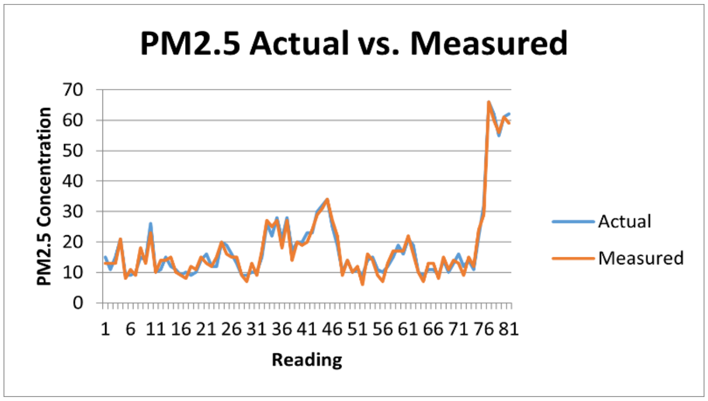


Figure 10: PM2.5 actual Vs. measured

Figure 10 shows the graph of PM2.5 concentration measured by the air quality monitoring device and the actual recorded PM2.5 values by the MAAQML, based on 800 random readings taken throughout the day which exhibits a mean percentage accuracy of 89.59%.

A. Prediction model testing and Results

The data set used as input for each of the prediction models were split as training and testing datasets at random locations to a percentage of 75% training data and 25% testing data so that the trained model can be run against the test data to compute how well the model performs and to avoid errors such as over-fitting.
The prediction models which were trained using the training dataset were used to make predictions on the test data set where the absolute error for each reading was computed thereby obtaining the mean absolute error and mean absolute percentage error to calculate the accuracy of the trained model.
Model runtime, which is the time taken by the processor to execute the machine code and a measure of performance of the prediction model was calculated by running 10 iterations of training and testing cycles and obtaining the average for each prediction model.

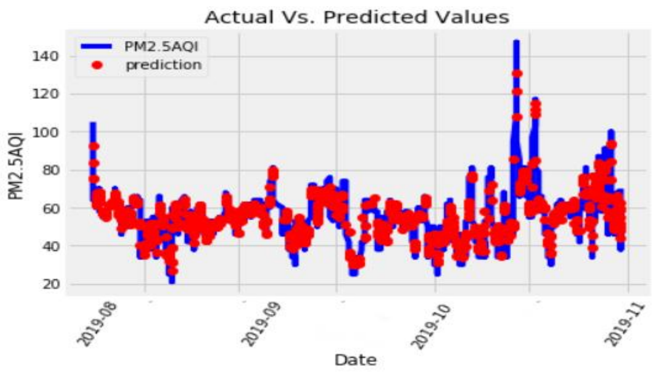


Figure 11: actual Vs. predicted values

Figure 11 shows the graph plotting actual PM2.5AQI values against the predicted PM2.5AQI values by the "Expanded_all_ features" prediction model. The PM2.5AQI value was predicted with an accuracy of 92.53% percent by the prediction model.

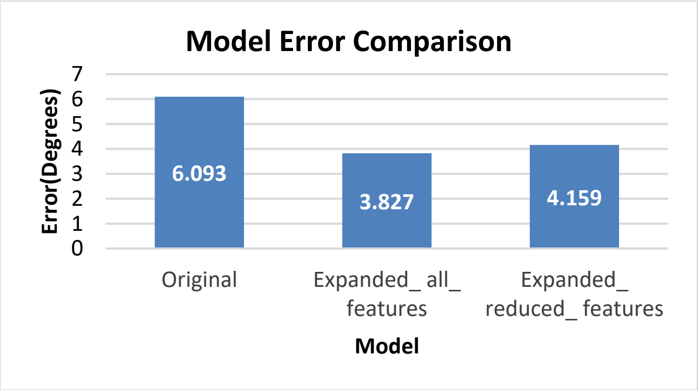


Figure 12: Model error comparison

Figure 12 shows the model error comparison where the model "original" has the highest error rate (6.093) and the model "Expanded_all_features" have the least error rate (3.827).

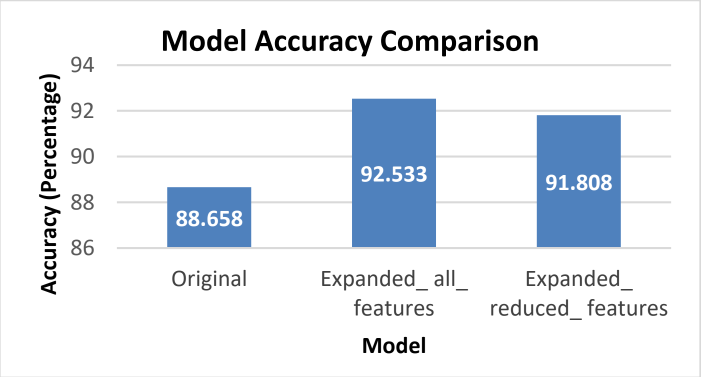


Figure 13: Model accuracy comparison

Figure 13 shows the prediction model accuracy comparison where the "Expanded_all_features" model have the best accuracy (92.533%) and the "original" model having the lowest accuracy (88.658%).

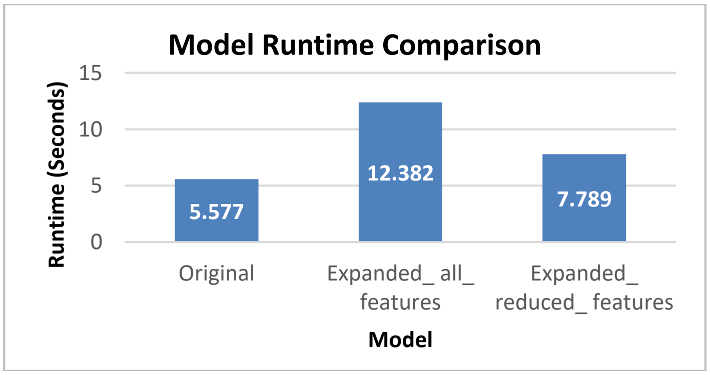


Figure 14: Model runtime comparison

Figure 14 shows the comparison of runtime performance of the 3 prediction models where the "original" model has the best runtime (5.557s) while the "expanded_all_features" model have the lowest runtime (12.382s). However, the model "expanded_reduced_features" (7.789s) have a better runtime and performance than "expanded_all_features" model while having a better accuracy of 91.808% than the "original" prediction model.

V. CONCLUSION

This study focuses on the development and implementation of a portable and low-cost device to measure and capture ambient air quality determining parameters, PM2.5, Ozone, Carbon Monoxide concentration, temperature and relative humidity using arduino uno as the main controller.

Testing and validation carried out for the device provided with an overall accuracy of ~ 91%. The overall average accuracy of similar IoT systems range from 80%-85%. The developed air quality system shows an improvement in accuracy by approximately 6%-11% which shows that the calibration procedures followed during the study has had a positive impact. It has proven that electro-chemical, metal-oxide sensors can be used to capture air quality parameters and the possibility of participatory air quality monitoring systems. In conclusion, it is evident that calibration techniques can be used to improve accuracy and overcome the accuracy issue of low-cost air quality sensors and similar monitoring devices.

The study also focuses on the development of a prediction model to predict PM2.5 Air Quality Index value for a particular hour based on several input parameters. Most air quality prediction models use linear regression techniques to predict air quality parameters which may face difficulties to capture nonlinear relationship in data. Similar air quality prediction models using gradient boosting regression has shown an accuracy of 88.6% and multilayered perceptron model an accuracy of 91.6%. Prediction model implemented in this study using random forest as the base model could predict the AQI value for a particular hour with an accuracy of 92.53% which is an improvement of 0.93% over multilayered perceptron. For an accuracy tradeoff of 0.72%, the model performance can be improved by 37.1%. This improvement is seen due to the noise reduction caused by feature reduction and parameter tuning.

This study has demonstrated the potential of developing low cost, portable and accurate air quality monitoring devices for monitoring air quality efficiently in real-time.

During this study it has been evident that the non-linear ensemble technique implemented performs better, showing that it supports the theory that non-linear regression techniques are better in predicting air quality parameters. It also shows that Bayesian approach can be used to successfully implement an accurate air quality prediction model.

It has also demonstrated that the ensemble machine learning technique used performs with a higher accuracy and efficiency than traditional gradient boosting and regression machine learning techniques in the prediction of Air Quality Index.

APPENDIX

- MAAQML – Mobile ambient air quality monitoring lab
- PM 2.5 – Fine Particulate matter (Below 2.5 micron)
- PM 10 – Particulate matter (Below 10 micron)
- AQI – Air Quality Index

REFERENCES

[1] "Air pollution: how it affects our health," *European Environment Agency*, 03-Dec-2019. [Online]. Available: https://www.eea.europa.eu/themes/air/health-impacts-of-air-pollution

[2] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and A. G. Xie, "Air Quality Prediction: Big Data and Machine Learning Approaches," *Ijesd.org*, 2018. [Online]. Available: http://www.ijesd.org/show-103-1491-1.html.

[3] M. Sarika, A. Korade, V. Kotak, and M. A. Durafe, "A review paper on internet of Things (IoT) and its applications," *Irjet.net*. [Online]. Available: https://www.irjet.net/archives/V6/i6/IRJET-V6I6376.pdf.

[4] D. Hasenfratz and O. Saukh, "Participatory Air Pollution Monitoring Using Smartphones," *Researchgate.net*, 2012. [Online]. Available: https://www.researchgate.net/publication/267963506_Participatory_Air_Pollution_Monitoring_Using_Smartphones.

[5] S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode, and B. Nath, "Real-time air quality monitoring through mobile sensing in metropolitan areas January 2013," *Researchgate.net*, 2013. [Online]. Available: https://www.researchgate.net/publication/308054705_Real-time_air_quality_monitoring_through_mobile_sensing_in_metropolitan_areas.

[6] Y. Yang and L. Li, "A smart sensor system for air quality monitoring and massive data collection," in *2015 International Conference on Information and Communication Technology Convergence (ICTC)*, 2015, pp. 147–152.

[7] D. Luong Nguyen, "A brief review of air quality models and their applications," *Psu.edu*. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.974.8959&rep=rep1&type=pdf.

[8] M. Dong, D. Yang, Y. Kuang, D. He, S. Erdal, and D. Kenski, "PM2.5 concentration prediction using hidden semi-Markov model-based times series data mining," *Expert Syst. Appl.*, vol. 36, no. 5, pp. 9046–9055, 2009.

[9] Y. Dou, N. D. Le, and J. V. Zidek, "Temporal forecasting with a Bayesian spatial predictor: Application to ozone," *Adv. Meteorol.*, vol. 2012, pp. 1–13, 2012.

[10] X. Ren, Z. Mi, and P. G. Georgopoulos, "Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States," *Environ. Int.*, vol. 142, no. 105827, p. 105827, 2020.

AUTHORS

First Author – T.S. Kitchilan, Uva Wellassa University of Sri Lanka, tskitchilan123@gmail.com
Second Author – M.D.B.C.K. Abeyratne, Uva Wellassa University of Sri Lanka, buddhi.chamitha@gmail.com.
Third Author – Prof. E.P.S.K. Ediriweera, Uva Wellassa University of Sri Lanka, sisira@uwu.ac.lk.

Correspondence Author – T.S. Kitchilan, tskitchilan123@gmail.com, +9471-3349110

http://dx.doi.org/10.29322/IJSRP.12.03.2022.p12307