








Article

Real-Time In-Vehicle Air Quality Monitoring System Using Machine Learning Prediction Algorithm

Chew Cheik Goh ^{1,2}, Latifah Munirah Kamarudin ^{1,2,*} , Ammar Zakaria ^{2,3} , Hiromitsu Nishizaki ⁴ , Nuraminah Ramli ¹ , Xiaoyang Mao ⁴, Syed Muhammad Mamduh Syed Zakaria ^{1,2} , Ericson Kanagaraj ^{1,2} , Abdul Syafiq Abdull Sukor ^{2,3}  and Md. Fauzan Elham ⁵

- ¹ Faculty of Electronic Engineering Technology, Universiti Malaysia Perlis (UniMAP), Arau 02600, Malaysia; ccgoh@studentmail.unimap.edu.my (C.C.G.); nuraminah@unimap.edu.my (N.R.); smmamduh@unimap.edu.my (S.M.M.S.Z.); ericsonkanagaraj@gmail.com (E.K.)
- ² Advanced Sensor Technology, Centre of Excellence (CEASTech), Universiti Malaysia Perlis (UniMAP), Arau 02600, Malaysia; ammarzakaria@unimap.edu.my (A.Z.); abdulsyafiq@unimap.edu.my (A.S.A.S.)
- ³ Faculty of Electrical Engineering Technology, Universiti Malaysia Perlis (UniMAP), Arau 02600, Malaysia
- ⁴ Graduate Faculty of Interdisciplinary Research, University of Yamanashi, 4-3-11 Takeda, Kofu, Yamanashi 400-8511, Japan; hnishi@yamanashi.ac.jp (H.N.); mao@yamanashi.ac.jp (X.M.)
- ⁵ Selangor Industrial Corporation Sdn Bhd, Seksyen 14, Shah Alam 40000, Malaysia; fauzan@sic.com.my
- * Correspondence: latifahmunirah@unimap.edu.my



Citation: Goh, C.C.; Kamarudin, L.M.; Zakaria, A.; Nishizaki, H.; Ramli, N.; Mao, X.; Syed Zakaria, S.M.M.; Kanagaraj, E.; Abdull Sukor, A.S.; Elham, M.F. Real-Time In-Vehicle Air Quality Monitoring System Using Machine Learning Prediction Algorithm. *Sensors* **2021**, *21*, 4956. <https://doi.org/10.3390/s21154956>

Academic Editors: Jude Hemanth, Jose Garcia Rodriguez, Epameinondas Kapetanios, Peter M. Roth and Anastassia Angelopoulou

Received: 2 June 2021
Accepted: 16 July 2021
Published: 21 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: This paper presents the development of a real-time cloud-based in-vehicle air quality monitoring system that enables the prediction of the current and future cabin air quality. The designed system provides predictive analytics using machine learning algorithms that can measure the drivers' drowsiness and fatigue based on the air quality presented in the cabin car. It consists of five sensors that measure the level of CO₂, particulate matter, vehicle speed, temperature, and humidity. Data from these sensors were collected in real-time from the vehicle cabin and stored in the cloud database. A predictive model using multilayer perceptron, support vector regression, and linear regression was developed to analyze the data and predict the future condition of in-vehicle air quality. The performance of these models was evaluated using the Root Mean Square Error, Mean Squared Error, Mean Absolute Error, and coefficient of determination (R^2). The results showed that the support vector regression achieved excellent performance with the highest linearity between the predicted and actual data with an R^2 of 0.9981.

Keywords: internet of things (IoT); machine learning prediction; in-vehicle air quality; smart mobility; smart city

1. Introduction

One of the main aims of smart cities is to reduce the fatalities and injuries due to traffic accidents. According to transport statistics in Malaysia, the total vehicles that are involved in road accidents increased yearly from 2008 to 2017. In 2017, the total road accidents reported were 533,875 cases and the total casualties and damages caused by traffic accidents were 16,589 cases [1]. The Royal Malaysia Police has stated that the leading causes of a road crash are drivers in fatigued conditions and distracted drivers [2]. The American Automobile Association (AAA) estimates that one out of every six deadly traffic accidents as well as one out of eight crashes requiring hospitalization is due to drowsy drivers [3]. In fact, the air inside the vehicle cabin has a significant impact on the cognitive capability of the occupants without noticeable discomfort that would put them on alert [4].

Most indoor air quality studies are focused on the inside of a building. The main components of indoor air contamination are carbon monoxide (CO), formaldehyde, ozone (O₃), total volatile organic compounds (TVOC), and particulate matter (PM) which can highly affect human health [5]. A straightforward method of mitigating the hazardous gases is by closing all windows and doors to prevent the pollutants from the outdoors.

Furthermore, a similar indoor environment such as the vehicle cabin that has been equipped with a heater, ventilation, and air conditioning system (HVAC) can be categorized as an indoor space. HVAC systems use the recirculation mode (RC) that could mitigate the penetration of pollutants such as particulate matter and hazardous gases from the vehicle's exhaust system [6,7]. Nonetheless, the human occupant inhales the oxygen then replaces it with carbon dioxide (CO₂) which acts as contamination known as the human bio-effluent. The high concentration of CO₂ reduces human cognitive ability, causes drowsiness, dizziness, and fatigue [8,9]. These are dangerous consequences to the occupants as well as potentially to other road users. Moreover, the vehicle speed might affect the gas concentration inside the vehicle cabin [10]. The undesirable effect of this condition is not limited to gas concentration only. The particle count (PM_{2.5} and PM₁₀) circulating within the air can also affect the occupant's health status [11]. The diameter of dust smaller than 2.5 µm can impact heart and lung health if inhaled by the occupants [12].

Thus, there is a clear need to perform an early prediction of the in-vehicle air quality which can be used to alert the occupants before the air quality becomes worse and affects the driver's health condition while they are driving. Most of the previous studies only focused on classifying the hazardous gasses without having the ability to predict the future condition [13,14]. Furthermore, most studies are limited to a few hazardous gasses such as carbon dioxide (CO₂) and oxygen (O₂). In this respect, there are several methods from machine learning (ML) techniques such as artificial neural network (ANN) and regression algorithms that are applicable for air quality predictions. Furthermore, since the current reading of the air quality data depends on previous data, a time-sequence supervised learning air quality data can be used as the input structure [15].

This paper presents the design and development of low-cost sensor hardware for an in-vehicle air quality monitoring system with cloud-based storage and prediction on the current and future air quality. ML prediction methods are developed using several approaches such as multi-layer perceptron (MLP), support vector regression (SVR), and linear regression algorithms. These algorithms are compared to determine which is the best model by considering six inputs including the vehicle speed, CO₂, temperature, humidity, PM_{2.5}, and PM₁₀. The contribution of this paper is the prediction system that includes the development of sensor hardware and cloud-based predictive analysis for an in-vehicle air quality monitoring system. The system is essential for future smart cities and smart mobility applications which can help to reduce fatality and injuries due to road accidents.

This paper is organized as follows. Section 2 discusses the related studies on air quality and prediction applications using ML algorithms. Section 3 describes the system development of an in-vehicle air quality monitoring system. Section 4 introduces the methods of collecting data. Section 5 shows the procedures of the data processing and Section 6 presents the proposed ML methods applied to the in-vehicle air quality data and the process flow of the ML analysis. The predicted result is then discussed in Section 7. Finally, the last section of this paper is the conclusion of the research.

2. Related Works

The literature studies have shown that humans spend up to 70%–90% of the time inside an environment with closed air circulation daily, including vehicle cabins [16,17]. The studies have shown that the air quality inside the vehicle cabin possibly contains polluted air [18–20]. They have also determined existing hazardous gases inside the cabin such as VOC, CO, CO₂, nitrogen dioxide (NO₂), sulphur dioxide (SO₂), and other pollutants. What is worse is that the concentration level of those gases might be higher than the standards established by the World Health Organization (WHO) and other governmental health organizations. The effect might cause occupants to experience immediate health issues, including impaired vision and coordination, nose and throat irritations, headaches, dizziness, drowsiness, and fatigue to the occupants [8]. The combinations of these effects on the occupants' health are not ideal for operating a vehicle.

The fresh air mode of the HVAC system triggered inside a vehicle can introduce air pollution from the outside environment such as: PM, NO₂, SO₂, and CO into the vehicle cabin. This can happen regularly especially in the urban and industrial areas. The RC mode significantly helps to reduce air pollution by circulating the air inside the vehicle cabin and increasing the passengers' comfort experience. However, the RC mode can build up the CO₂ concentration and accumulate rapidly due to the existence of the passengers.

Moreover, a statement has been declared by the Malaysian industry code of practice on the indoor environment that the CO₂ concentration limit should not exceed 1000 ppm at any time. One study found that if the concentration of CO₂ reached 2500 ppm in a room the size of 50.78 m³, the occupants' decision-making capability on primary activity, initiative, information usage, breadth of approach, and basic strategy fall into a range of marginal and dysfunctional [16]. An average sedan vehicle's interior space is around 2.72 m³ which is 18 times smaller than the experiment environment. A high CO₂ concentration will reduce the O₂ concentration and can cause permanent damage to organs, including the brain and heart [21].

In this paper, we focus entirely on the RC ventilation mode to operate inside the vehicle cabin. The primary pollutant source inside a cabin is the occupant. So, the CO₂ concentration level is the main parameter to be observed as well as the PM concentration level. The fact of the matter is that different countries and organizations have different standards of the air quality index, even for the same type of pollutant. Some critical pollutants such as SO₂ are not even taken into consideration in the air quality index for certain countries [22]. In addition, no established standard has showed the breakpoint concentration that is specifically for the in-vehicle air quality environment. Hence, Table 1 shows the combined in-vehicle air quality index breakpoints of the Environmental Protection Agency (EPA) standard, indoor air quality guideline in Malaysia [23], Occupational Safety and Health Administration (OSHA) [24], and the experiment that had been done by [16].

Table 1. Breakpoint concentration of the in-vehicle air quality.

CO ₂ (ppm)	PM _{2.5} (µg/m ³) ^d	PM ₁₀ (µg/m ³) ^d	IV-AQI	Five Bands of IV-AQI
C _{low} –C _{high}	C _{low} –C _{high}	C _{low} –C _{high}	I _{low} –I _{high}	
340–600 ^a	0.0–12.0	0–54	0–50	Good
601–1000 ^b	12.1–35.4	55–154	51–100	Moderate
1001–1500	35.5–55.4	155–254	101–150	Unhealthy for sensitive group
1501–2500 ^a	55.5–150.4	255–354	151–200	Unhealthy
2501–5000 ^c	150.5–250.4	355–424	201–500	Very unhealthy

a: Associations of the cognitive function scores with carbon dioxide, ventilation, and volatile organic compound exposures in office workers: A controlled exposure study of green and conventional office environments (USA) [16]. b: Industry code of practice on indoor air quality 2010 (Malaysia) [23]. c: Occupational safety and health administration (OSHA): carbon dioxide in workplace atmospheres (US) [24]. d: Environmental protection agency (EPA) [25].

The US EPA has introduced the individual pollutant index, also known as the air quality index (AQI) as in Equation (1). The AQI acts as an indicator of reporting the air quality of the targeted environment. Equation (1) calculates each observed parameter in a time series. The highest individual index among other air parameters for pollutants will stand as the air quality of the vehicle's cabin.

$$I_p = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}} (C_p - BP_{Lo}) + I_{Lo} \quad (1)$$

where,

I_p = index for pollutant p

C_p = the rounded concentration of pollutant p

BP_{Hi} = the breakpoint that is greater than or equal to C_p

BP_{Lo} = the breakpoint that is less than or equal to C_p

I_{Hi} = the AQI value corresponding to BP_{Hi}

I_{Lo} = the AQI value corresponding to BP_{Lo}

With respect to prediction systems, artificial intelligence algorithms are widely used in smart city applications for classification prediction and regression prediction such as human activity classification [26,27], transportation [28], and air quality prediction [29–32]. In [33] the authors applied the ML algorithms to predict the air quality by using the data from 750 observations with 0.95 accuracy and their prediction was successful. [34] focused on predicting air pollution in Canada using an MLP, and the prediction model performed on $PM_{2.5}$ had 4.5 of MAE.

Meanwhile, [35] have addressed the challenges in real-time air quality predictions, namely, the aspect of realistic real-time air quality monitoring devices, online systems, and predictive models in a review paper perspective. The real-time air quality system should provide an online user interface that allows the user to observe the air quality from anywhere. The support vector regression (SVR) is one of the most successful prediction models with a low root mean squared error (RMSE) (0.939) in forecasting the air quality in Japan [36,37] collected different types of gases and sent them to the cloud database. However, the authors addressed that the ML is used only for sensor calibration, not for in-vehicle air quality prediction. All these research studies are focused on indoor or outdoor air quality prediction, but not targeted for the air quality inside the vehicle cabin.

The research gap of this work is implementing the classification of the in-vehicle air quality together with the prediction of future conditions to monitor drivers' dizziness and fatigue while they are driving. Hence, this work will be focused on system development and regression prediction for an in-vehicle air quality system.

3. System Design and Development

Based on the related works, the essential element to integrate a gas sensor into the hardware system is the type of sensor selection and targeted gas [38–42]. CO_2 was found to be the most critical pollutant in the vehicle cabin with the RC ventilation mode. Moreover, the CO_2 concentration is affected by the speed of the vehicle [43,44]. Thus, an integrated in-vehicle air quality monitoring system was developed for this work. The developed system has an integrated GPS tracking device as well as a CO_2 gas sensor. Additional sensors such as particulate matter, temperature, and humidity are also embedded into the system. The overall system architecture of the in-vehicle air quality is illustrated in Figure 1. The system design is separated into four parts which are hardware development (device node), cloud database, software development (user interface), and an in-vehicle air quality prediction model. An IoT-oriented transportation system is applied in this system by connecting the device node to the internet in order to push real-time data into the cloud database [45].

The SIM808 (GPRS/GSM) communication module was chosen for this application. It not only provides better wireless regional coverage for up to 70 km but also has the feature to provide GPS coordinates. The raw GPS signal is cascaded with additional information which needs to be removed before storing only the latitudinal and longitudinal information. The microprocessor begins by initializing all the peripherals and sensors on the device node. Thirty seconds of initialization time is given to ensure all the sensors have been initialized properly. After initialization, a connection will be established between the device node and the cloud database by using the Message Queuing Telemetry Transport (MQTT) messaging protocol. Once the MQTT protocol connection has been established, the microprocessor begins gathering all the sensor data. Finally, the sensor data is aggregated into the buffer and encapsulated into the MQTT protocol format, and published into the cloud database using the brute force method. If the publishing is unsuccessful, the microcontroller checks the network of the MQTT connection and continues the sensor sampling process.

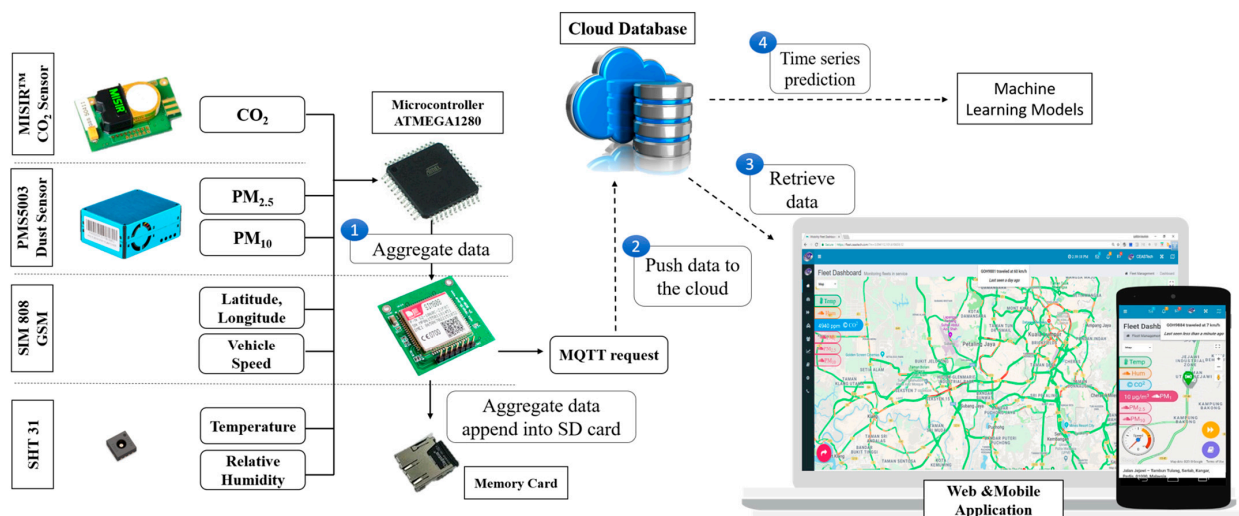


Figure 1. System architecture design for the in-vehicle air quality monitoring system.

Furthermore, the data will be processed and sorted in the cloud database. Each of the device nodes is assigned with a unique identifier (ID) to avoid mis-location of the data entry. Then, a database handler is designed to discard the distorted data entry and invalid ID. For example, some published data may contain unidentified ASCII characters. Meanwhile, a web page and a mobile application are developed, which is capable of viewing the real-time data of the in-vehicle air quality status. The visualization is for the convenience of the user to understand and learn the patterns of the in-vehicle air quality. The data is illustrated in the form of Google Maps. There are several features available in the interface such as real-time view, data export, playback of the daily route, and view of historical data according to date.

The primary power source for the sensor device node is obtained from the in-car charger. The voltage range of a car battery is from 11.9 V to 14.8 V, where most of the time it is in the fluctuation mode. Therefore, several stages of voltage step-down are necessary due to the different operating voltages of the sensors. Figure 2 shows the final design of the sensor device baseboard with the specific voltage requirement and interface of the device node labelled. As for the sensor validation, the gas sensor data have been calibrated and verified with the established portable gas sensor device with the model Aeroqual, Series-500.

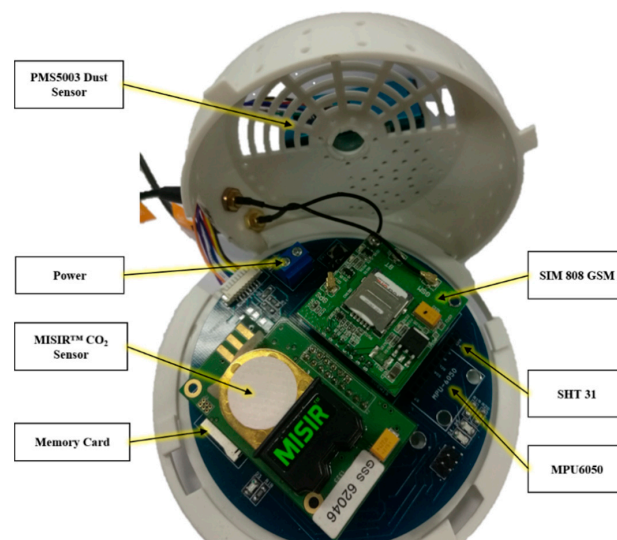


Figure 2. The final design for the device node.

4. Real-Time Data Collections

In the preliminary study, the fresh air (FA) and RC ventilation mode are selected in order to observe the gas compositions inside the vehicle cabin. The eight gas parameters selected to be sampled inside the cabin are O₃, CO, VOC, NO₂, SO₂, CO₂, PM_{2.5}, and PM₁₀. The time taken for one set of data collection is forty minutes and will be repeated sixty times for each experiment. There are two types of cars used in this experiment that represents the indoor vehicle environment: the Nissan Grand Livina and the Toyota Vios. The experiment is conducted under an average vehicle speed of about 70 km/h.

Figure 3 shows the data collection of the common hazardous gases inside the vehicle cabin with the FA mode. For the FA mode, the quality of air inside the vehicle cabin is highly dependent on the outdoor air quality. Results show that most of the parameters observed exceeded the recommended limit established by the DOSH standard. Only two parameters—CO₂ and SO₂—are below the recommended limit. In the FA mode, the CO₂ gas detected in the vehicle cabin was in the range of 600–900 ppm and the data collection showed a similar range with the outdoor air. It can be assumed that the low SO₂ obtained is due to the experiment routes performed outside the petroleum refineries, chemical manufacturing industries, mineral ore processing plants, and power station areas. Meanwhile, the RC ventilation mode presented lesser hazardous gases existing inside the vehicle cabin. The only three parameters, which are CO₂, PM_{2.5}, and PM₁₀ exceeded the recommended limit with reference to the DOSH standard. The air quality inside the vehicle cabin was not affected by the outdoor environment.

After finishing the preliminary experiment, we identified the essential observation parameters for further investigation. Next, the experiment is conducted on a real-time traffic basis. The experiment is performed for two months entirely under RC ventilation mode conditions. The travelling time can be separated into three slots, which are morning (06:00–08:00), afternoon (11:20–13:30), and evening (16:00–18:00). Travelling distance in June 2019 is 2306 km for 14 days, and July 2019 is 2494 km for 19 days. The daily average travelling distance is approximately 164.7 km, as shown in Figure 4. The minimum occupant is one and the maximum occupants are five. The experiment vehicle is a sedan car type with 2.75 m³ of space. Figure 5 shows the daily travelled path in this experiment. On the other hand, Table 2 shows the size of the data samples that have been collected throughout the experiments. Time series data in the air quality system has a parameter dedicated to counting the number of packets received by the cloud database. Each time the device is powered up, the count will be set as one and increases according to the subsequent data packet. Thus, the section data can be sorted using the count parameter. The average acquisition time for each data is 4 s. Once the data entered into the cloud database is completed, the vehicle speed is computed using the latitude, longitude, and time data. Then, the data labelling for the ML algorithms is generated. The data labelling is used to help the algorithms to train better and produce reliable results.

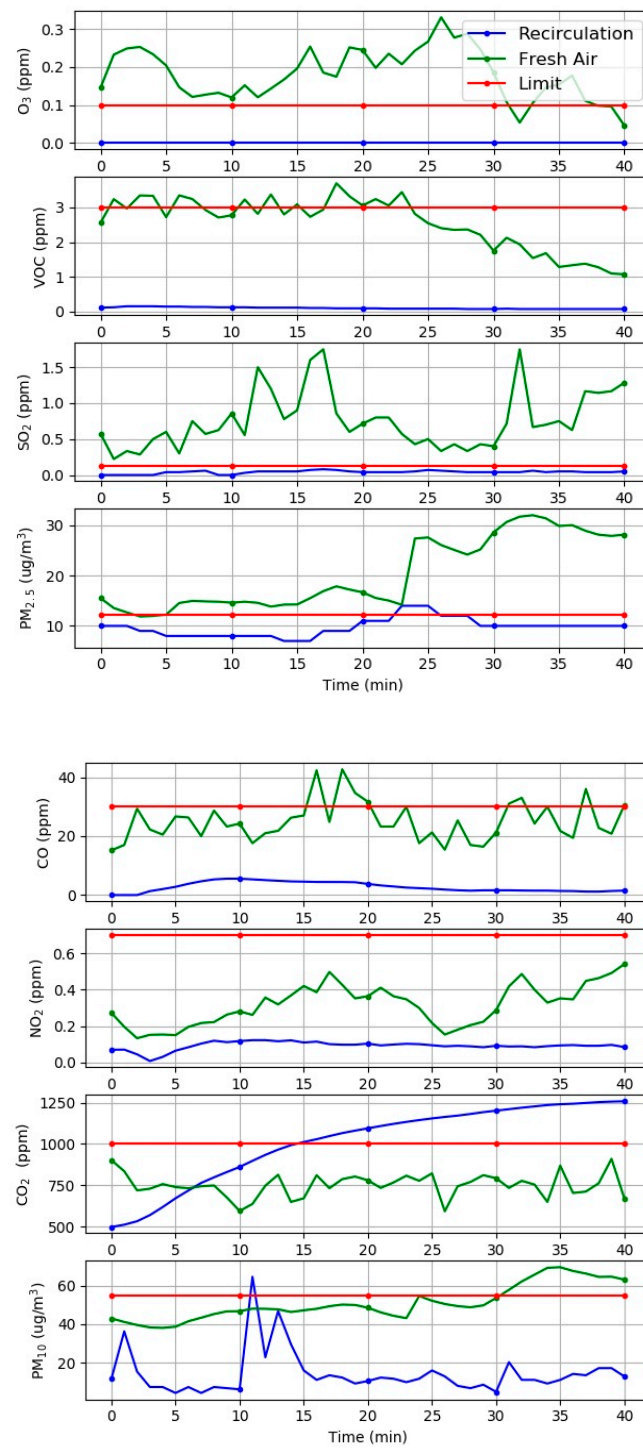


Figure 3. The common hazardous gases inside the vehicle cabin with the FA and RC ventilation modes.



Figure 4. The daily travelled path in the experiment.

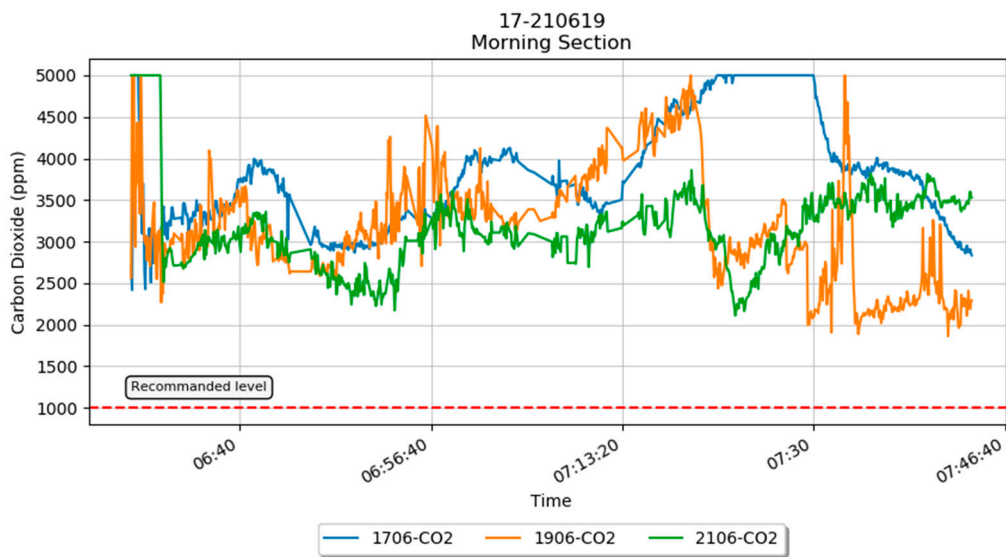


Figure 5. The raw carbon dioxide sensor data plotted for the morning slot.

Table 2. Data features and samples size for the in-vehicle air quality.

Collection Site		Parameter
Monthly	Number of records	48,816
	Size	3,593,866 bytes (3.59 MB)
One section	Number of records	1184
	Size	92,160 bytes (0.09 MB)
Value types		Twelve air quality variables (Time, latitude, longitude, speed, CO ₂ , temperature, humidity, PM ₁ , PM _{2.5} , PM ₁₀ , count, label)

Figure 5 shows the raw CO₂ sensor data on three different days for the morning slot. The fluctuation in the graph is due to the variation of the vehicle's speed as well as the number of occupants in the car. The graph shows that the CO₂ concentration level is higher than the recommended level by the Department of Occupational Safety and Health (DOSH). Other parameters of the morning slot such as vehicle speed, temperature, humidity, CO₂, PM_{2.5}, and PM₁₀ are illustrated in Figure 6.

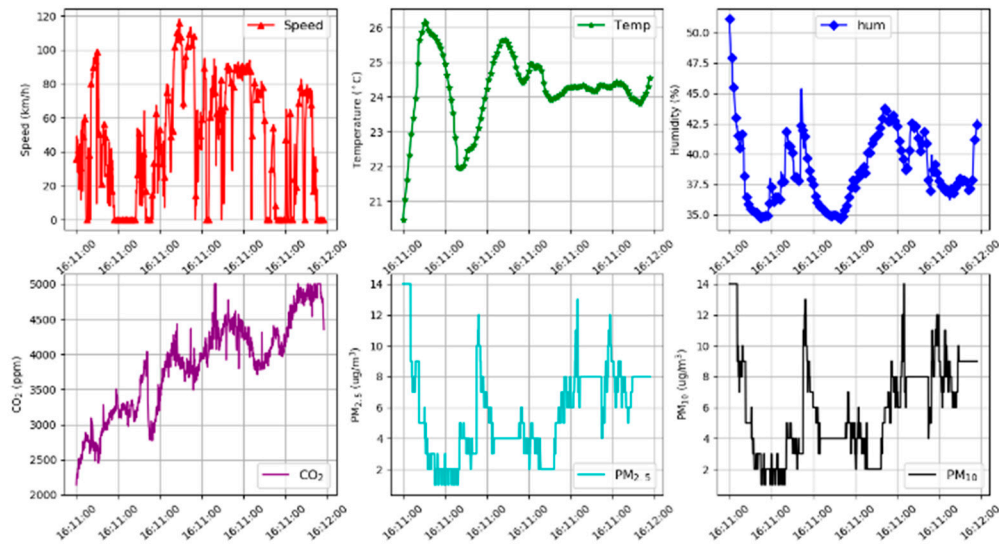


Figure 6. The raw section sensor data plotted for each parameter.

5. Data Processing for Time Series Data

This section will briefly introduce the flow of preparing the real-time data, labelling the data, and performing the data normalization procedure. This study uses time-series data, which is an ordered sequence data. The time interval between the data point is continuous and each time unit observation has at most, one data point.

The real-time sensor data might have data errors such as sensor error and outlier data prone to a false trend. The data preprocessing method is introduced to reduce the training complexity and to increase the accuracy while feeding the data into the prediction algorithms. The next step of the data preparation is labelling the sensor data. After completing the labelling process, the data will be run through a series of ML experiments to figure out the most compatible ML parameters for the air quality system. There are a total of six input data used for the ML, which are CO₂, PM_{2.5}, PM₁₀, vehicle speed, temperature, and humidity.

5.1. Data Preprocessing

The raw data of the sensor is collected without a filtering process. The filtering process is implemented in the cloud rather than on the embedded device to reduce the complexity in the embedded system. The common data errors of the real-time monitoring application as expected are outliers and data missing [46]. Three types of common data pre-processing are: filling the not-a-number (NAN) data into zero, dropping the NAN data, or data interpolation before feeding the data into the ML algorithms. In this research, data interpolation is conducted using the nearest-neighbour method. This method is suitable for datasets that have missing values or outlier conditions [47]. Equation (2) shows the nearest-neighbour mathematical equations. When the outlier occurs at the position x_i , the value of the closest known neighbour is used to replace the outlier value. There are four states of different formulas that are used in this method. If the position of x_i is greater than 5, the average of the five previous data will be used to replace the outlier position. When the outlier position is less than five, an average value will be used by taking as much

historical data that it has. The reason for taking previous data and not using the future data is because the data collection is in real-time in time-series form.

$$x_i = \begin{cases} \frac{x_{i-1}+x_{i-2}}{2}, & \text{if } i = 2; \\ \frac{x_{i-1}+x_{i-2}+x_{i-3}}{3}, & \text{if } i = 3; \\ \frac{x_{i-1}+x_{i-2}+x_{i-3}+x_{i-4}}{4}, & \text{if } i = 4; \\ \frac{x_{i-1}+x_{i-2}+x_{i-3}+x_{i-4}+x_{i-5}}{5}, & \text{otherwise} \end{cases} \quad (2)$$

where as x_i is the outlier value.

5.2. Data Labelling

Before feeding the sensor data into the prediction algorithm, a set of data should be labelled as the output in supervised machine learning. In fact, the air quality index (AQI) is an index approach to categorize the quality of the air in a specific environment. The AQI is usually separated into a few ranges and each range is assigned a color code as well as a description. It provides a public health advisor for each range [48]. The breakpoint concentration of the in-vehicle air quality for different types of pollutants has been discussed previously in Table 1. There are various versions of standards and guidelines which depend on the international agencies [22]. Current air quality standards do not provide a breakpoint for CO₂. So, the CO₂ breakpoint listed in Table 1 is obtained from different organizations and research groups [16,23–25].

Figure 7 shows the flowchart of labeling the in-vehicle air quality index. The parameters of CO₂, PM_{2.5}, and PM₁₀ are selected to compute the index for pollutants. The highest index represents the AQI at that time. A large pressure will be created against the vehicle's body when the vehicle travels at a high speed and leakages will occur between the joints [44]. The higher the vehicle speed, the more outdoor air will be penetrating the vehicle cabin [10]. Thus, other parameters such as temperature, humidity, and vehicle speed may also affect the AQI inside the vehicle cabin.

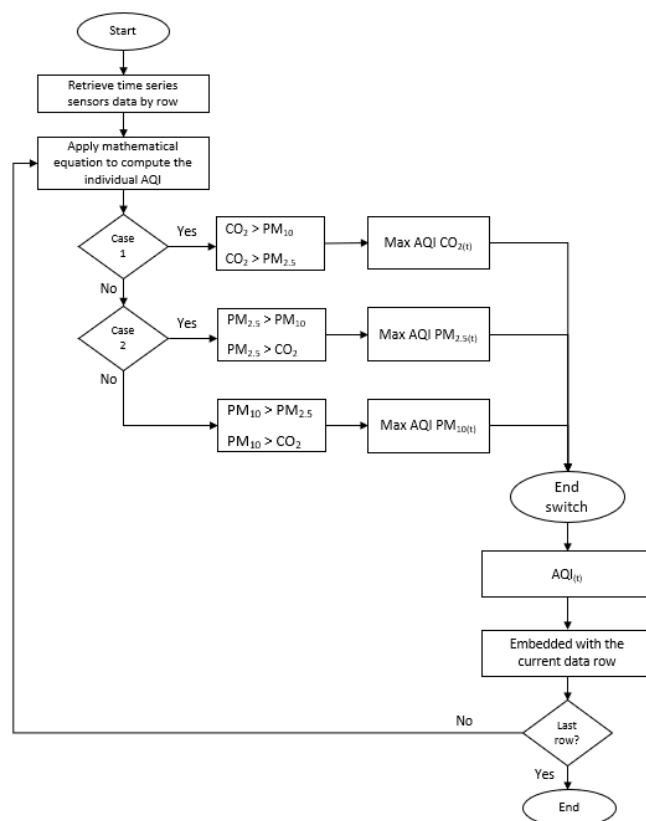


Figure 7. The time series data labelling of the AQI.

5.3. Normalization

The data normalization method is often implemented in the dataset to ease the data processing time. The function of the normalization is to change the numeric values in the dataset into a 0 to 1 range without changing the original range values of the dataset. In AI prediction, not every dataset requires normalization. However, the sensor dataset contains several features in a different range. For instance, the CO₂ sensor consisting of three digits might reach to four digits and the PM sensor consisting of one digit might reach three digits. Hence, the min-max scalar is selected due to the sensor dataset in a time-series form with a short interval. If the originality of the trend is not preserved, the learning process in the ML prediction models will be affected.

6. Prediction Analysis using Machine Learning Algorithms

This section describes the algorithms that are used to predict the future condition of the air quality.

6.1. Linear Regression

A linear regression model is capable of time series prediction [49]. This is because the model makes a prediction by simply computing a weighted sum of the input features, plus a constant called the bias term, as shown in Equation (3).

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n \quad (3)$$

where,

\hat{y} —the predicted value

n —the number of features

x_n —the n th feature value

θ_j —the j th model parameter

6.2. Support Vector Machine

The support vector machine (SVM) can perform linear or nonlinear classification. Besides that, SVM also supports linear and nonlinear regression applications, known as SVR [50]. In the SVR, there are three parameters that need to be appropriately selected to achieve higher prediction accuracy and better performance. These are the insensitive loss coefficient (ϵ), error penalty factor (C), and kernel function coefficient (γ). The complexity of the model is dependent on these parameters. These three parameters are highly inter-related and affect the SVR model. The grid search method provides the best combination for the three mentioned parameters. By implementing the GridSearchCV function in the sklearn library, the grid search range for both ϵ and C is set $(-3, 3, 21)$ with the logspace function. Thus, the best hyperparameters found for the model of ϵ , C, and γ are at 0.001, 501, and a radial basis function kernel (RBD kernel), respectively.

6.3. Multilayer Perceptron

The MLP is the most commonly used model in the feed-forward neural network. The basic MLP has three layers which are the input layer, hidden layer, and output layer as shown in Figure 8. A grid search method is implemented to search the fine-tuned hyperparameters in the MLP. The range of hyperparameters for hidden nodes, learning rate, optimizer, and activation function are 2^3-2^{10} , 0.001–0.05, adam or stochastic gradient descent, and relu or tanh, respectively. The fine-tuned hyperparameters in the MLP structure applied in this research have a single hidden layer of 128 hidden nodes, a 0.001 learning rate, tanh activation function, and stochastic gradient descent optimizer.

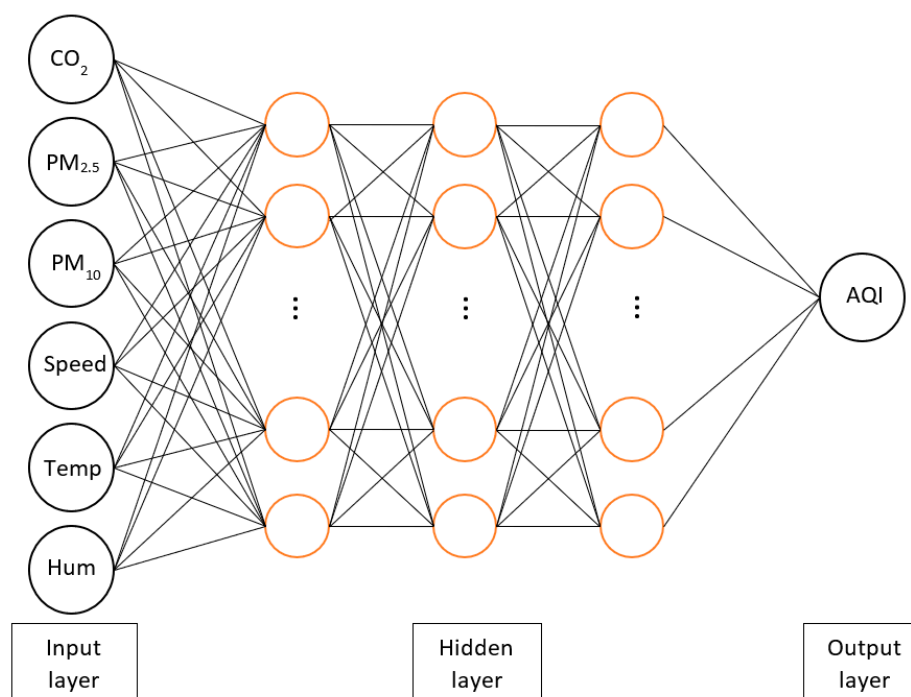


Figure 8. The MLP structure with various hidden node numbers.

6.4. Evaluation Methods

To evaluate the performance of the ML applied to the in-vehicle air quality monitoring system, *RMSE* [51], the mean squared error (*MSE*), mean absolute error (*MAE*), and coefficient of determination (R^2) are selected. Equations (4)–(7) present the formula for each of the evaluation metrics, respectively.

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

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (7)$$

where,

\hat{y}_i —predicted value of y

\bar{y}_i —mean value of y

7. Results and Discussions

All data were separated into two sets, which are 80% as the training data and 20% as the test data. The model scripts were executed using the Nvidia GeForce RTX 2080 Ti graphics card as the hardware accelerator. 75 data points, which represented data of 5 min, were predicted using the different ML prediction models mentioned in previous sections. There are two datasets to test, train, and evaluate which have different time slots and monthly data as mentioned in Table 2.

The evaluation results of the ML prediction models are shown in Table 3. The evaluation result is to verify the prediction capability. The accuracy of the MLP model had a significant improvement when the historical data was increased from 0.7151 to 0.9107

of R^2 . The SVR model with the RBF kernel had the highest R^2 and lowest MSE , $RMSE$, and MAE compared to other models. The SVR-RBF-based prediction model showed the highest prediction accuracy and had better generalization performance. The R^2 obtained was as high as 0.9890 (section) and 0.9981 (month).

Table 3. Prediction model results.

	Section					Month				
	R^2	MSE	$RMSE$	MAE	Computation Time (s)	R^2	MSE	$RMSE$	MAE	Computation Time (min)
SVM	0.9890	6.4513	2.5410	0.97194	1.6	0.9981	3.6168	1.9018	0.4101	44.5
LR	0.8137	109.9008	10.4833	5.1379	0.2	0.9946	10.1875	3.1917	2.1348	37.2
MLP	0.7151	212.4807	14.5767	11.5757	26	0.9107	100.0034	9.0589	5.0422	83.3

In addition, another important aspect is the computation time of the model for future implementation of the prediction algorithm in edge computing. In the section dataset, the LR had an outstanding computation. It only took 0.2 s for the prediction. However, the R^2 of the LR was lower than the SVR. The prediction model of the SVR had an acceptable computation time (1.6 s) with a high R^2 . For the real-time prediction model, high accuracy and low computation time are an important aspect that must be considered.

Figures 9 and 10 show the distributions of prediction results for easy interpretation. From the graph, the prediction model of the SVR-RBF shows similar shapes and tendencies to the actual data. The LR prediction model also shows a good fitting line. However, the MLP prediction model does not fit into the actual data. Hence, the SVR with the RBF prediction model is suitable for system prediction.

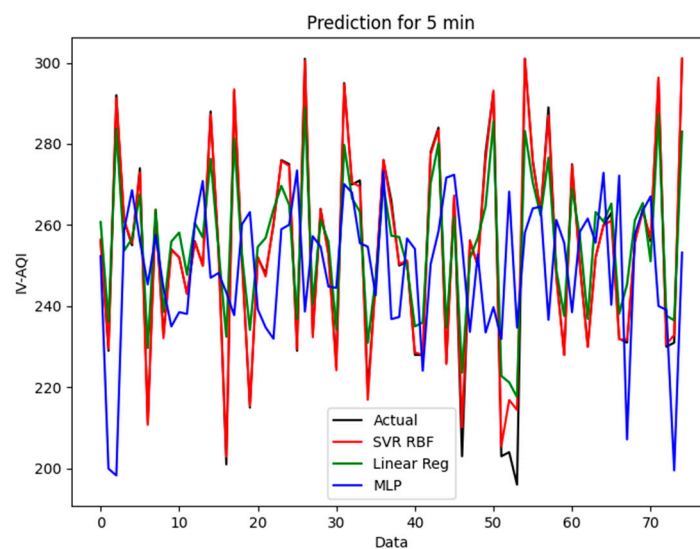


Figure 9. The prediction for the section event in SVR, LR, and MLP.

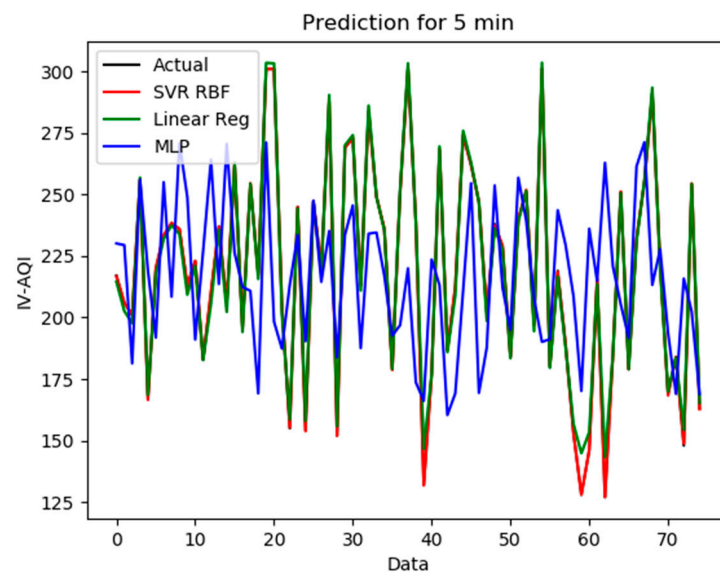


Figure 10. The prediction for the month event in SVR, LR, and MLP.

8. Conclusions

This research focused on the ML prediction model for an in-vehicle air quality application. A hardware testbed was developed to obtain sensor data in the in-vehicle indoor environment. Then, three predictive models of machine learning algorithms such as LR, SVR, and MLP were applied to the in-vehicle air quality prediction system to predict the air quality inside the vehicle cabin. This allowed the monitoring of the real-time air quality inside the car cabin. The system can be used as a potential measure to reduce traffic accidents due to driver drowsiness and fatigue. The results showed that the SVR had the highest performance rates in terms of R^2 and had less error rate. This indicates that the SVR model has an outstanding prediction performance as well as low computation time compared with the LR and MLP models.

Author Contributions: C.C.G. conceptualization, design of experiments, methodology, investigation and data collection, formal analysis, writing—original draft preparation and writing—review & editing. L.M.K. conceptualization, design of experiments, methodology, formal analysis, writing—original draft preparation, writing—review & editing, supervision and obtained funding. N.R. data collection. S.M.M.S.Z. formal analysis and writing—review & editing. E.K. formal analysis and writing—review & editing. A.Z. writing—review & editing, supervision, obtained funding. A.S.A.S. formal analysis, writing—review & editing. H.N. conceptualization, methodology, supervision and obtained funding. X.M. conceptualization, methodology, supervision and obtained funding. M.F.E. formal analysis and writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or appeared in this study are available upon requested by contact with the corresponding author. Moreover, the models and code used during the study cannot be shared at this time as the data also forms part of an ongoing study.

Acknowledgments: The authors wish to thank the MTUN Research Matching Grant under the Development and Deployment of Smart Community and Innovation Centre for Smart City 4.0 for providing the funding used in this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ministry of Transport Malaysia. Transport Statistics Malaysia. In *Statistik Pengangkutan Malaysia*; 2017. Available online: <http://www.mot.gov.my/my/StatistikTahunanPengangkutan/StatistikPengangkutanMalaysia2017.pdf> (accessed on 9 July 2019).
2. Kumar, M. Fatigue, Mobile Phone Use among Top Causes of Road Accidents. The Star Online. 2018. Available online: <https://www.thestar.com.my/news/nation/2018/06/08/fatigue-mobile-phone-use-among-top-causes-of-road-accidents> (accessed on 1 October 2019).
3. Tefft, B.C. Prevalence of motor vehicle crashes involving drowsy drivers, United States, 1999–2008. *Accid. Anal. Prev.* **2012**, *45*, 180–186. [[CrossRef](#)] [[PubMed](#)]
4. Constantin, D.; Mazilescu, C.-A.; Nagi, M.; Draghici, A.; Mihartescu, A.-A. Perception of Cabin Air Quality among Drivers and Passengers. *Sustainability* **2016**, *8*, 852. [[CrossRef](#)]
5. DOSH. *Industry Code of Practice on Indoor Air Quality 2010*; Ministry of Human Resources Department of Occupational Safety and Health: Putrajaya, Malaysia, 2010; pp. 1–50.
6. Grady, M.L.; Jung, H.; Kim, Y.C.; Park, J.K.; Lee, B.C. Vehicle Cabin Air Quality with Fractional Air Recirculation. *SAE Tech. Pap. Ser.* **2013**, *1*, 7. [[CrossRef](#)]
7. Thirumal, P.; Amirthagadeswaran, K.S.; Jayabal, S. Optimization of IAQ characteristics of an air-conditioned car using GRA and RSM. *J. Mech. Sci. Technol.* **2014**, *28*, 1899–1907. [[CrossRef](#)]
8. Hudda, N.; Fruin, S. Carbon dioxide accumulation inside vehicles: The effect of ventilation and driving conditions. *Sci. Total. Environ.* **2018**, *610–611*, 1448–1456. [[CrossRef](#)] [[PubMed](#)]
9. Satish, U.; Mendell, M.J.; Shekhar, K.; Hotchi, T.; Sullivan, D.; Streufert, S.; Fisk, W.J. Is CO₂ an indoor pollutant? Direct effects of low-to-moderate CO₂ concentrations on human decision-making performance. *Environ. Health Perspect.* **2012**, *120*, 1671–1677. [[CrossRef](#)]
10. Goh, C.; Kamarudin, L.; Shukri, S.; Abdullah, N.; Zakaria, A. Monitoring of carbon dioxide (CO₂) accumulation in vehicle cabin. In Proceedings of the 2016 3rd International Conference on Electronic Design (ICED), Phuket, Thailand, 11–12 August 2016; pp. 427–432. [[CrossRef](#)]
11. Mohd Firdaus, O.; Juliana, J. Exposure to Indoor Air Pollutants (PM₁₀, CO₂ And CO) and Respiratory Health Effects among Long Distance Express Bus Drivers. *Health Environ. J.* **2014**, *5*, 66–85.
12. Zheng, Y.; Liu, F.; Hsieh, H.-P. U-Air: When Urban Air Quality Inference Meets Big Data. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; pp. 1436–1444. [[CrossRef](#)]
13. Chiu, C.-F.; Chen, M.-H.; Chang, F.-H. Carbon Dioxide Concentrations and Temperatures within Tour Buses under Real-Time Traffic Conditions. *PLoS ONE* **2015**, *10*, e0125117. [[CrossRef](#)]
14. Lohani, D.; Acharya, D. Real time in-vehicle air quality monitoring using mobile sensing. In Proceedings of the 2016 IEEE Annual India Conference (INDICON), Bangalore, India, 16–18 December 2016; pp. 1–6. [[CrossRef](#)]
15. Glavas, S.D.; Sazakli, E. Ozone long-range transport in the Balkans. *Atmos. Environ.* **2011**, *45*, 1615–1626. [[CrossRef](#)]
16. Allen, J.G.; Macnaughton, P.; Satish, U.; Santanam, S.; Vallarino, J.; Spengler, J.D. Associations of Cognitive Function Scores with Carbon Dioxide, Ventilation, and Volatile Organic Compound Exposures in Office Workers: A Controlled Exposure Study of Green and Conventional Office Environments. *Environ. Health Perspect.* **2016**, *124*, 805–812. [[CrossRef](#)]
17. Jaimini, U.; Banerjee, T.; Romine, W.; Thirunarayan, K.; Sheth, A.; Kalra, M. Investigation of an Indoor Air Quality Sensor for Asthma Management in Children. *IEEE Sens. Lett.* **2017**, *1*, 1–4. [[CrossRef](#)]
18. Alameddine, I.; Esber, L.A.; Zeid, E.B.; Hatzopoulou, M.; El-Fadel, M. Operational and environmental determinants of in-vehicle CO and PM_{2.5} exposure. *Sci. Total. Environ.* **2016**, *551–552*, 42–50. [[CrossRef](#)] [[PubMed](#)]
19. Moreno, T.; Pacitto, A.; Fernández, A.; Amato, F.; Marco, E.; Grimalt, J.O.; Buonanno, G.; Querol, X. Vehicle interior air quality conditions when travelling by taxi. *Environ. Res.* **2019**, *172*, 529–542. [[CrossRef](#)] [[PubMed](#)]
20. Xu, B.; Chen, X.; Xiong, J. Air quality inside motor vehicles' cabins: A review. *Indoor Built Environ.* **2018**, *27*, 452–465. [[CrossRef](#)]
21. Shivani, P.; Julia, H.M.; Ekerm, Y.; Anya, A.; Sapan, H.M. *Physiology, Carbon Dioxide Retention*; StatPearls Publishing: Treasure Island, FL, USA, 2021.
22. Abdul-Wahab, S.A.; En, S.C.F.; Elkamel, A.; Ahmadi, L.; Yetilmezsoy, K. A review of standards and guidelines set by international bodies for the parameters of indoor air quality. *Atmos. Pollut. Res.* **2015**, *6*, 751–767. [[CrossRef](#)]
23. Department of Environment Ministry. Air Pollutant Index (API). Ministry of Energy, Science, Technology, Environment & Climate Change, 2019. Available online: <https://www.doe.gov.my/portalv1/en/info-umum/english-air-pollutant-index-api/100> (accessed on 17 July 2019).
24. OSHA. Carbon Dioxide in Workplace Atmospheres. United States Department of Labor, 1990. Available online: <https://www.osha.gov/dts/sltc/methods/inorganic/id172/id172.html> (accessed on 4 July 2019).
25. EPA. NAAQS Table. US Environmental Protection Agency, 2016. Available online: <https://www.epa.gov/criteria-air-pollutants/naaqs-table>. (accessed on 21 December 2019).
26. Sukor, A.S.A.; Zakaria, A.; Rahim, N.A. Activity recognition using accelerometer sensor and machine learning classifiers. In Proceedings of the 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA), Batu Feringghi, Malaysia, 9–10 March 2018; pp. 233–238. [[CrossRef](#)]

27. Sukor, A.S.A.; Zakaria, A.; Rahim, N.A.; Kamarudin, L.M.; Setchi, R.; Nishizaki, H. A hybrid approach of knowledge-driven and data-driven reasoning for activity recognition in smart homes. *J. Intell. Fuzzy Syst.* **2019**, *36*, 4177–4188. [CrossRef]
28. Amita, J.; Jain, S.; Garg, P. Prediction of Bus Travel Time Using ANN: A Case Study in Delhi. *Transp. Res. Procedia* **2016**, *17*, 263–272. [CrossRef]
29. Bin, M.; Site, L.; Shijin, Y. Air Quality Forecast Based on Principal Component Analysis-Genetic Algorithm and Back Propagation Model. *Int. J. Environ. Ecol. Eng.* **2016**, *10*, 899–906. [CrossRef]
30. Ong, B.T.; Sugiura, K.; Zettsu, K. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5. *Neural Comput. Appl.* **2016**, *27*, 1553–1566. [CrossRef]
31. Jin, X.-B.; Jeremiah, R.R.; Su, T.-L.; Bai, Y.-T.; Kong, J.-L. The New Trend of State Estimation: From Model-Driven to Hybrid-Driven Methods. *Sensors* **2021**, *21*, 2085. [CrossRef]
32. Jin, X.-B.; Yu, X.-H.; Su, T.-L.; Yang, D.-N.; Bai, Y.-T.; Kong, J.-L.; Wang, L. Distributed Deep Fusion Predictor for a Multi-Sensor System Based on Causality Entropy. *Entropy* **2021**, *23*, 219. [CrossRef]
33. Amado, T.M.; Cruz, J.D. Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization. In Proceedings of the TENCON 2018—2018 IEEE Region 10 Conference, Jeju, Korea, 28–31 October 2018; pp. 0668–0672. [CrossRef]
34. Peng, H.; Lima, A.R.; Teakles, A.; Jin, J.; Cannon, A.; Hsieh, W.W. Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods. *Air Qual. Atmos. Health* **2017**, *10*, 195–211. [CrossRef]
35. Hable-Khandekar, V.; Srinath, P. Machine Learning Techniques for Air Quality Forecasting and Study on Real-Time Air Quality Monitoring. In Proceedings of the 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 17–18 August 2017; pp. 1–6. [CrossRef]
36. Zhu, H.; Hu, J. Air Quality Forecasting Using SVR with Quasi-Linear Kernel. In Proceedings of the 2019 International Conference on Computer, Information and Telecommunication Systems (CITS), Beijing, China, 28–31 August 2019; pp. 1–5. [CrossRef]
37. Yang, J.; Chen, Y.; Liu, Y.; Makke, O.; Yeung, J.; Gusikhin, O.; Macneille, P. The effectiveness of cloud-based smart in-vehicle air quality management. In Proceedings of the 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Xi'an, China, 3–5 October 2016; pp. 325–329. [CrossRef]
38. Abdullah, A.; Adom, A.; Shakaff, A.; Ahmad, M.; Zakaria, A.; Saad, F.; Isa, C.; Masnan, M.J.; Kamarudin, L. Hand-Held Electronic Nose Sensor Selection System for Basal Stamp Rot (BSR) Disease Detection. In Proceedings of the 2012 Third International Conference on Intelligent Systems Modelling and Simulation, Kota Kinabalu, Malaysia, 8–10 February 2012; pp. 737–742. [CrossRef]
39. Abdullah, A.H.; Shakaff, A.Y.; Adom, A.H.; Zakaria, A.; Saad, F.S.; Kamarudin, L.M. Chicken Farm Mal-odour Monitoring Using Portable Electronic Nose System. *Chem. Eng. Trans.* **2012**, *30*, 55–60.
40. Kamarudin, K.; Shakaff, A.Y.M.; Bennetts, V.H.; Mamduh, S.M.; Zakaria, A.; Visvanathan, R.; Yeon, A.S.A.; Kamarudin, L.M. Integrating SLAM and gas distribution mapping (SLAM-GDM) for real-time gas source localization. *Adv. Robot.* **2018**, *32*, 903–917. [CrossRef]
41. Thriumani, R.; Zakaria, A.; Hashim, Y.Z.H.-Y.; Jeffree, A.I.; Helmy, K.M.; Kamarudin, L.M.; Omar, M.I.; Shakaff, A.Y.M.; Adom, A.H.; Persaud, K.C. A study on volatile organic compounds emitted by in-vitro lung cancer cultured cells using gas sensor array and SPME-GCMS. *BMC Cancer* **2018**, *18*, 362. [CrossRef] [PubMed]
42. Yusuf, N.; Zakaria, A.; Omar, M.I.; Shakaff, A.Y.M.; Masnan, M.J.; Kamarudin, L.M.; Rahim, N.A.; Zakaria, N.Z.I.; Abdullah, A.A.; Othman, A.; et al. In-vitro diagnosis of single and poly microbial species targeted for diabetic foot infection using e-nose technology. *BMC Bioinform.* **2015**, *16*, 1–12. [CrossRef] [PubMed]
43. Grady, M.L. *On-Road Air Quality and the Effect of Partial Recirculation on In-Cabin Air Quality for Vehicles*; University of California Riverside: Riverside, CA, USA, 2013.
44. Matton, T.J.P. *Simulation and Analysis of Air Recirculation Control Strategies to Control Carbon Dioxide Build-Up Inside a Vehicle Cabin*; University of Windsor: Windsor, ON, USA, 2015.
45. Hussein, W.N.; Kamarudin, L.M.; Hussain, H.N.; Hamzah, M.R.; Jadaa, K.J. Technology Elements that Influence the Implementation Success for Big Data Analytics and IoT- Oriented Transportation System. *Int. J. Adv. Trends Comput. Sci. Eng.* **2019**, *8*, 2347–2352. [CrossRef]
46. Ni, K.; Ramanathan, N.; Chehade, M.N.H.; Balzano, L.; Nair, S.; Zahedi, S.; Kohler, E.; Pottier, G.; Hansen, M.; Srivastava, M. Sensor network data fault types. *ACM Trans. Sens. Networks* **2009**, *5*, 1–29. [CrossRef]
47. Lepot, M.; Aubin, J.B.; Clemens, F.H.L.R. Interpolation in time series: An introductory overview of existing methods, their performance criteria and uncertainty assessment. *Water (Switz.)* **2017**, *9*, 796. [CrossRef]
48. Mintz, D. Technical Assistance Document for the Reporting of Daily Air Quality—The Air Quality Index (AQI). United States Environmental Protection Agency, 2013; pp. 1–26. Available online: <https://permanent.fdlp.gov/gpo50063/aqi-technical-assistance-document-dec2013.pdf> (accessed on 2 June 2021).
49. Géron, A. *Hands-On Machine Learning with Scikit-Learn & Tensor Flow*, 1st ed.; Tache, N., Ed.; O'Reilly, Media, Inc.: Sebastopol, CA, USA, 2017.
50. Vladimir, N.V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 2013. [CrossRef]
51. Willmott, C.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]