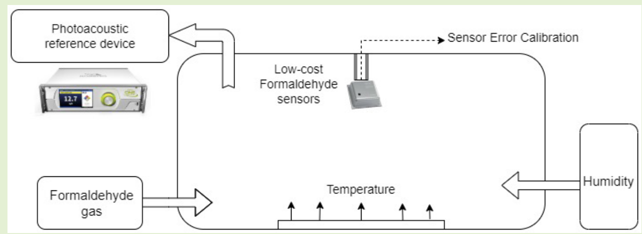
Low-Cost Formaldehyde Sensor Evaluation and Calibration in a Controlled Environment

Arnab Chattopadhyay, Andres Huertas, Andrew Rebeiro-Hargrave, Pak Lun Fung, Samu Varjonen, Tuomas Hieta, Sasu Tarkoma, *Senior Member, IEEE*, and Tuukka Petäjä

***Abstract*—Formaldehyde is a carcinogenic indoor air pollutant emitted from common wood-based materials. Low-cost sensing of formaldehyde is difficult due to inaccuracies in measuring low concentrations and susceptibility of sensors to changing indoor environmental conditions. Currently gas sensors are calibrated by manufacturers using simplistic models which fail to capture their complex behaviour. We evaluated different low-cost gas sensors to ascertain a suitable component to create a mobile sensing node and built a calibration algorithm to correct it. We compared the performance of 2 electrochemical sensors and 3 metal oxide sensors in a controlled chamber against a photo-acoustic reference device. In the chamber the formaldehyde concentrations, temperature and humidity were varied to assess the sensors in diverse environments. Pre-calibration, the electrochemical sensors (mean absolute error (MAE) = 70.8 ppb) outperformed the best performing metal oxide sensor (MAE = 335 ppb). A two-stage calibration model was built, using linear regression followed by random forest, where the residual of the first stage acted as a input for the second. Post-calibration, the metal oxide sensors (MAE = 154 ppb) improved compared to their electrochemical counterparts (MAE = 78.8 ppb). Nevertheless, the uncalibrated electrochemical sensor showed overall superior performance hence was selected for the mobile sensing node.**



***Index Terms*—Air quality monitoring, formaldehyde detection, metal oxide semiconductor, electrochemical sensors, machine learning, gas sensors.**

I. Introduction

THE World Health Organisation (WHO) has designated indoor air pollution to be the leading cause for 4.3 million premature deaths per year [1]. Today, majority of people are spending more than 90 percent of their time in indoor environments [1] and health problems and diseases associated with poor indoor air quality (IAQ) can cause a variety of adverse health effects to them [2]. Additionally, over the past two years there have been numerous lockdowns and 'work from home' recommendations issued due to the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic. This has resulted in an increase in time spent by people indoors [3]. Thus there is a need to better assess major indoor pollutants in home environments and develop methods of sensing it to avoid health degradation due to indoor air quality issues [4]. Low-cost gas sensors (<100€) embedded in a mobile sensing node could be a great cost effective method of sensing personal exposure to these pollutants. However, low-cost gas sensors often give inaccurate results and are equipped with a built-in calibration algorithm which fail to capture the complexity of their behaviour [5], [6]. If calibrated properly they can be a great resource in measuring personal exposure to pollutants in indoor environments and cities [7], [8]. The data from a network of these mobile sensing nodes can equip the governments and private sector with vital information needed to introduce targeted pragmatic measures to solve this complex problem.

Some volatile organic compound (VOC) are considered to be indoor air pollutants and are highly hazardous to human health, even when present in small concentrations [9] [10]. Formaldehyde is a volatile organic compound which is an omnipresent indoor air pollutant. It is emitted mostly from common wood-based home materials e.g. furniture, coatings, insulation and flooring materials, where it is used in the

Manuscript received April 8, 2022; accepted April 30, 2022. Date of publication May 4, 2022; date of current version June 14, 2022. This work was supported in part by the European Union through the Urban Innovative Action Healthy Outdoor Premises for Everyone under Project UIA03-240, in part by the Academy of Finland under Grant 307537, in part by the MegaSense Growth Engine through the Air Quality Monitoring funded by Business Finland under Grant 7517/31/2018, in part by the Trans-National ERA-PLANET Project SMURBS under EU Horizon 2020 Framework Program under Grant 689443, and in part by the Academy of Finland Flagship under Grant 337549. The associate editor coordinating the review of this article and approving it for publication was Prof. Kea-Tiong (Samuel) Tang. *(Corresponding author: Arnab Chattopadhyay.)*

Arnab Chattopadhyay, Andrew Rebeiro-Hargrave, Samu Varjonen, Sasu Tarkoma, and Tuukka Petäjä are with the University of Helsinki, 00014 Helsinki, Finland (e-mail: arnab.chattopadhyay@helsinki.fi).

Andres Huertas is with Accenture, Finland.

Pak Lun Fung is with the Faculty of Science, Institute for Atmospheric and Earth System Research (INAR), University of Helsinki, 00010 Helsinki, Finland.

Tuomas Hieta is with Gaseera Oy, 20520 Turku, Finland.

Digital Object Identifier 10.1109/JSEN.2022.3172864

adhesives. It has been found that short term exposures towards formaldehyde causes eye and nose irritation [7], [11] and on medium and sustained long term exposure it can also cause asthma [12], [13] and nasopharyngeal cancer [14], [15]. It has been designated as a category 1B carcinogen by the European Union [16]. Lack of affordable and accurate options for mobile formaldehyde sensing make it important to carefully study them in order to make them available to the general public [17]. Formaldehyde is difficult to measure using low-cost devices because they start harming humans even when present in small concentrations [18], detecting which have been known to be challenging for low-cost gas sensors [19], [20]. Furthermore formaldehyde emissions are often triggered by changing environmental conditions which has been shown to be a leading cause of errors in low-cost gas sensors [7], [21]. Therefore understanding of this complex phenomenon of sensor interaction with formaldehyde in varying environmental conditions is an important topic which needs to be researched in order to build an affordable and accurate mobile sensing unit.

The main goal of this study is to evaluate the efficacy of the promising low-cost formaldehyde sensors to be embedded in a mobile sensing node. The evaluation of individual low-cost sensor is done in a controlled environment chamber by comparing them against an accurate reference device in varying environmental conditions and varying concentrations of formaldehyde. The evaluation of the sensors is done in two phases: firstly, pre-calibration wherein we evaluate the sensor's inherent capabilities and effectiveness of its built-in calibration algorithm and secondly, post-calibration after building a two-stage machine learning calibration model to correct its data-stream. By doing this our objective is to gauge two important parameters: the level of errors innately present in the sensors and the ease with which they can be modelled and calibrated. The sensors will be subjected to differing concentrations of only formaldehyde therefore their selectivity or performance in presence of other cross-sensitive gases is beyond the scope of this paper.

A secondary goal of the study is to compile and open source the comparative sensor data to serve as a high quality training data-set for building of machine learning models [22]. Gas sensor data-sets are often created in ambient conditions which make the models fail on deployment in extreme conditions. The use of the controlled environment chamber enables us tackle this challenge by mapping the sensor performance against the entire state space of the environmental variables creating a more complete data-set for model training. There have been some previous studies where gas sensor evaluation was done via environmental chamber [23]–[26]. However these studies were mostly focused on common outdoor pollutants and particulate matter and did not use the chamber as a source for generating diverse training data for aiding model building.

A. Formaldehyde Emission Regulations and Environmental Factors

Understanding concentrations of formaldehyde considered to be dangerous is key to understanding the range in which optimal sensor performance is needed. In regulations around the world, two limiting parameters for formaldehyde exposure are specified: time weighted average for 8 hours (TWA) and short term exposure limits for 15 minutes (STEL). Values have been converted to parts per billion from parts per million for consistency throughout the paper. Table I lists the different formaldehyde exposure limits set in different countries [27]. Fig. 1 maps these legislated exposure limits against formaldehyde concentrations measured in ambient conditions at homes [28], mobile homes [29] and during some common activities like smoking [30] and oven cleaning [31]. Amongst these, the highest ambient formaldehyde concentration of 800 ppb is measured in mobile homes due to their enclosed structure and high use of pressed wood products [29].

Closed indoor environments are places where build up of harmful volatile gases can occur easily hence understanding these environments are essential for accurately detecting formaldehyde. Formaldehyde emission in indoor environments is dependant on environmental parameters like relative humidity and temperature. Relative humidity (RH) is one of the main environmental factors affecting the emission behaviours of formaldehyde from building materials. In one of the early studies measuring this correlation, a doubling of emission rate of formaldehyde from a particleboard was observed when RH increased from 30 % to 70 % at constant temperature [32]. Similarly recent studies also displayed similar results with medium density fiberboard, another common wood-based products found at home [33]. In addition to these results, it was also found that in the temperature range of 14 to 35° C formaldehyde emissions doubled for every 7° C increase

TABLE I
PERMISSIBLE FORMALDEHYDE EXPOSURE LIMITS OF DIFFERENT COUNTRIES EXPRESSED IN TIME WEIGHTED AVERAGE FOR 8 HOURS (TWA) AND SHORT TERM EXPOSURE LIMITS OF 15 MINUTES (STEL) [27]

Countries	TWA (8 hrs)		STEL (15 min)	
	ppb	$\mu g/m^3$	ppb	$\mu g/m^3$
Finland	300	370	1000	1200
Germany	300	370	600	740
France	500	625	1000	1250
Sweden	300	370	600	740
United Kingdom	2000	2500	2000	2500
Japan	100	120		
Australia	1000	1200	2000	2500

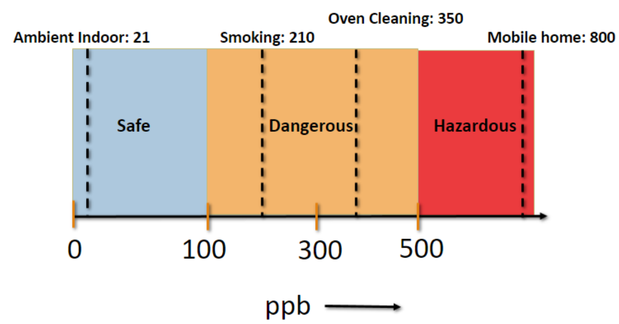


Fig. 1. Ranges of concentration of Formaldehyde as per time weighted averages (TWA) mentioned in table I along with formaldehyde concentrations measured after emission from common activities.

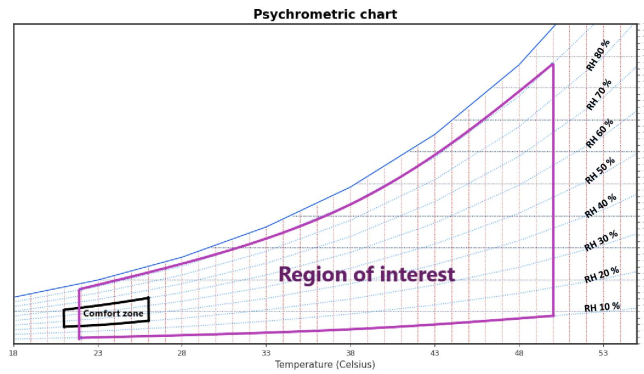


Fig. 2. Psychometric chart used for designing Heating, ventilation, and air conditioning (HVAC) systems showing region of interest of indoor environment conditions.

[32], [34]. Therefore studying the sensor performance in varying conditions like at higher temperature and humidity is of great importance as that is when formaldehyde emissions peak. Psychometric chart is often used for designing indoor heating, ventilation, and air conditioning (HVAC) systems. In Fig. 2 the 'comfort zone' denotes the range of optimal indoor environmental parameters defined by different European union countries as their national recommendations [35]. For this study, after taking into account open ventilation systems and varying indoor environments around the world we have used a larger range of environmental conditions as the boundary conditions which is denoted as 'region of interest' in Fig. 2. The initial boundary for our region of interest are seen to be slightly higher in Fig. 2 due to the higher recommended ambient temperature in Finland [35]

B. Measuring Technologies

Currently the most accurate way of measuring formaldehyde is batch sampling followed by off-site analysis using either chromatography or spectroscopy [36]. This procedure is expensive, time consuming and cannot be done in real time [36]. One of the newer development in technology is the use of infrared spectroscopy for accurate real-time detection of formaldehyde [37]. This will be the principle of operation of the reference device used in our study [38]. However, current infrared spectroscopy devices are also highly expensive and bulky to act as a compact mobile gas sensing node.

In our study, low-cost sensors measuring gas concentration using electrochemical sensing and a metal oxide film are studied. Electrochemical (EC) and metal oxide (MOS) sensors are the two most common sensing principles used in low-cost gas sensors. This is primarily because both have good sensitivity, fast response time, portable size and with slightly different configuration can be optimized for a different gas. However each of these sensing techniques have their advantages and disadvantages:

1) Electrochemical (EC) Sensor: EC sensors sense gas concentrations based on current measurements between electrodes in an electrolytic cell. The main benefits of EC sensors are their resistance to environmental changes and low power draw due to lack of need of an electric heater. However their operating range is narrower than in MOS sensors. A combination of low humidity and high temperature is particularly problematic to EC sensors, as it can dry out the sensor's electrolyte and break the sensor [39]. The EC sensors also tend to degrade faster than MOS sensors over time, especially when in contact with high concentrations of pollutants [40].

2) Metal Oxide (MOS) Sensor: These sensors sense gas by measuring the change in resistance of metal oxide sensing element which reacts to the target gas on heating up to 300–500° C. MOS sensors have several advantages, such as low-cost and compact size. The biggest disadvantage of MOS sensors is their accuracy and tendency to drift [41]. Additionally, higher humidity results in higher sensor error [42]. Heating elements in MOS sensors make them highly power inefficient which needs to be considered especially while designing a mobile device.

These sensors have improved over the past few years with respect to their underlying material science [43], [44] and signal processing [45], [46]. However calibration using machine learning when done effectively has also shown encouraging results [47], [48].

II. Methods and Materials

This section describes the sensors and the apparatus used for data generation along with the calibration algorithm used for modelling the low-cost sensor errors. The experimental setup and its components used for generation of the comparative data of the low-cost sensors and reference device are described in section II-A, II-B and II-C. The sampling strategy used to generate the data from entire range of independent variables is detailed in section II-D, followed by the specifics of the machine learning calibration algorithm used to calibrate the data-stream in section II-E.

A. Low-Cost Sensors

There were three different models of low-cost gas sensors used in the experiments: two MOS sensors i.e. Bosch BME680, Sensiron SGP30 and one EC sensor i.e. Winzen ZE08-CH2O. In the data acquisition unit there were two units of the Bosch BME680, two units of Winzen ZE08-CH2O and one unit of Sensiron SGP30 (due to fixed I2C address limitation). The BME680 and SGP30 are micro-electromechanical system (MEMS) based MOS sensors for detecting TVOC (Total Volatile Organic Compounds) whereas the ZE08-CH2O is an EC sensor designed for measuring formaldehyde. More established and tested metal-oxide TVOC sensors were chosen for the experiments after poor results shown in preliminary testing by a number of low-cost metal-oxide (MOS) sensors marketed as formaldehyde sensors. Lack of high number of units of individual sensors is a limitation of this study which must be taken into account while interpreting the results.

B. Reference Device

To act as a reference and to get accurate formaldehyde concentration values throughout the experiments the Gasera One Formaldehyde device is used [49]. This unit senses the gas concentration by a combination of photo-acoustic

TABLE II
LOW-COST GAS SENSORS USED IN THE STUDY

Sensor	Sensor Type	Target gas	Price (€)	Form factor (mm^3)
 BME680	MOS	TVOC	10-20	3.0 × 3.0 × 1.0
 SGP30	MOS	TVOC	10-20	2.45 x 2.45 x 0.9
 ZE08-CH2O	EC	Formaldehyde	20-30	25.5 x 23 x 4.4

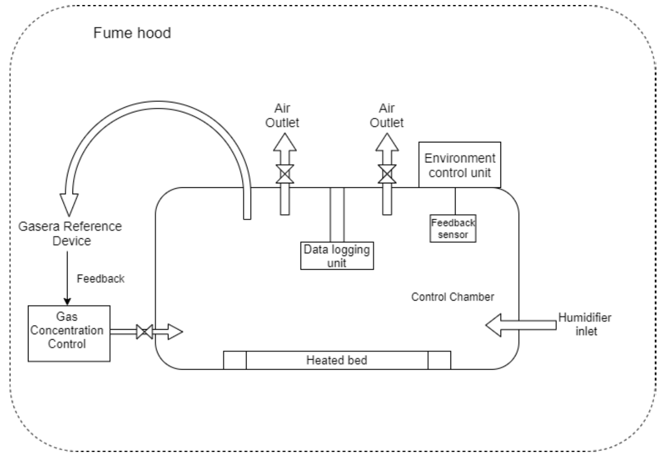


Fig. 3. Schematic of the entire test setup.

TABLE III
REFERENCE DEVICE: GASERA ONE FORMALDEHYDE SPECIFICATIONS

	 Gasera One Formaldehyde
Detection range	100 ppm (100,000 times the detection limit)
Detection limit	1 ppb (0.001 ppm)
Response time	minimum 10 seconds (user configurable)
Principle of operation	Infrared spectroscopy and photo-acoustic effect
Weight	13 kg
Operating conditions:	
Sample gas temperature range	0 °C to +50 °C
Humidity	below 90% RH, non-condensing
Pressure range	ambient level

infrared spectroscopy and an ultra-sensitive cantilever sensor. The sample gas is essentially stored in a photo-acoustic measuring chamber wherein it is excited using a mid-IR light of frequency corresponding to the resonant frequency of the formaldehyde molecule. The emitted energy is converted to an equivalent electrical signal using the ultra-sensitive cantilever sensor. Hence it is a highly accurate and selective device for measuring formaldehyde (see Table III) [49].

C. Experimental Setup

Environmental chambers are used to study the complex interactions between sensors and their surroundings in a controlled manner. These chambers can be an asset as they can give us an overview of sensor performance in many diverse environments representing their potential environment of deployment. They are also a great source of generating sensor training data synthetically in a situation where enough high quality training data might not be present. Our current experimental setup was designed to serve both the above mentioned goals. The entire experimental setup can be essentially divided into three parts: the control chamber, environmental control unit and data-logging unit.

1) Control Chamber: To ensure safety against poisonous formaldehyde fumes the entire experimental setup was built inside a fume-hood. The fume hood had a maximum exhaust airflow of 225 m/s which also represents the rate at which air intake takes place from the lab environment. The lab environment did not have any formaldehyde sources based on ambient condition measurements by the reference device and had a ventilation factor of 3-5 which represents the number of time the air in the room is changed per hour. A 52L polypropylene box served as the control chamber inside which the varying combinations of environmental parameters and gas concentrations were created. The use of polypropylene as a material was due to ease of access, sufficient resistance to formaldehyde and formic acid and its experimental non-interference due to its inability to emit formaldehyde. The control chamber has two 10 mm openings at the top to provide ventilation needed to reach equilibrium. These openings are equipped with manually actuated pneumatic shut-off valves. The control chambers also contains two 10 mm openings at the bottom of the side faces for entry of the formaldehyde gas and humid air as seen in Fig. 3.

2) Environmental Control Unit: The control of the humidity and temperature was carried out by the environmental control unit. It consisted of an Arduino Uno, that controlled a printed circuit board (PCB) heated bed and a 3L portable humidifier. This module is given a feedback through an Adafruit BME280 temperature humidity sensing module, which is placed inside the control chamber. Based on the feedback, the control of the heated bed and humidifier was carried out using an on-off control via an Arduino shield consisting of multiple mechanical relay units. The reference device is present outside the control chamber as seen in Fig. 3 and 4, it draws in a sample every minute for measurement. A Raspberry pi is used to log down the readings of the reference unit. This Raspberry pi also controls the gas concentration in the control

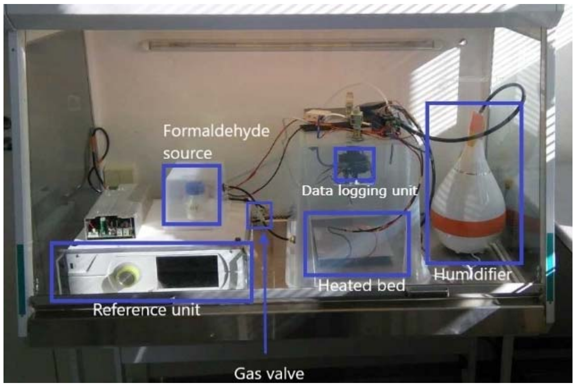


Fig. 4. Image of the actual setup.

TABLE IV
SUMMARY OF EXPERIMENTS CONDUCTED IN THE CONTROLLED CHAMBER

Experiment set	Duration	Concentration	Temperature (°C)	RH(%)
1: Low formaldehyde concentration	145 hours	10 - 275 ppb	22 (constant)	12 (constant)
2: High formaldehyde concentration	48 hours	100 - 800 ppb	22 (constant)	12 (constant)
3: Step temperature rise	113 hours	100 - 800 ppb	Step increase	12 (constant)
4: Random temperature	22 hours	100 - 600 ppb	Random	12 (constant)
5: Step humidity rise	94 hours	100 - 600 ppb	22 (constant)	Step increase

chamber by using the reference unit data as a feedback and a DC fan as an actuator. The DC fan is connected to a separate, smaller 1.5L polypropylene box as seen in Fig. 3 labelled as 'Gas concentration control', where a beaker of formalin is kept whose constant evaporation acts a source of formaldehyde fumes. The intake of formaldehyde gas into the control chamber is done by opening the gas valve and actuating the fan at a fixed speed using a simple on-off controller in order to achieve the desired concentrations mentioned in the sampling strategy (in section II-D). The discharge of the formaldehyde is done gradually via the outlets shown in Fig. II-C to get maximum comparative data points for the entire concentration range of interest.

3) Data Logging Unit: A data logging unit is suspended from the centre of the top face of the control chamber. This data logging unit consists of an Arduino Mega which has all the low-cost sensors attached to it. It is also provided with a real time clock module and an SD card module for logging of the data with corresponding timestamps. The data is taken out at the end of every experiment and stored in a work station for analysis.

D. Sampling Strategy for Experimental Data

A machine learning model's performance is highly dependent on quality of training data fed into it. In order to perform well in all possible indoor environments post deployment, some data depicting sensor performance in these environments is needed. As stated in section I, the secondary goal of the research was to generate a high-quality training data-set of sensor performance over the entire range of temperature, relative humidity and formaldehyde concentration. The sampling strategy was designed to help achieve this goal. Five sets of experiments were carried out, wherein one or two of the independent variables were varied to understand their effect on sensor performance. The nature of variation and range of the independent variables is as mentioned below:

1) Formaldehyde Concentration: After understanding the common governmental regulations against VOCs and formaldehyde a range of 0-800 ppb was chosen. This range comprises of prescribed limits of gas exposure and commonly experienced formaldehyde concentrations as described in Fig. 1. During the experiments the concentration in the chamber was oscillated from minimum to maximum value of this range to get a uniform distribution of data-points in the entire range. The intake of formaldehyde is done in a much more rapid rate using the fan at a constant speed to avoid stagnation at an intermediate steady state. The discharge is done at a slower pace to get sufficient comparative data points for the entire concentration range. This process is repeated for the time period of the experiment once the concentration comes back to the lower limit specified for the particular experiment in Table IV.

2) Temperature: VOCs are dominant and harmful primarily in indoor environments. Therefore the temperature range of $22-50°$ C was chosen. This represents the range of temperature experienced in indoor environments in a number of places in the world (as seen in Fig. 2). To explore the behaviour of the sensors in this temperature range, the temperature was changed in two ways: step-wise and random. For step-wise change: the temperature was increased or decreased in step manner wherein after each increment of 1° C a period of constant temperature of 4500 seconds was maintained. For random: a random target temperature in the specified temperature range was chosen which was reached and maintained for 1800 seconds following which a new target temperature was chosen. Due to the actuation being done by an on-off controller this period of constant temperature shows oscillations about the target value. For every constant temperature step, the concentration was varied from 0-800 ppb in order to get an average effect of the changing temperature on the entire concentration range.

3) Relative Humidity: The relative humidity level was varied from extremely low values of 8% up to the highest amount of 85% in accordance with the conditions shown in Fig. 2. Variation of the humidity was done primarily in the form of steps wherein every increment is of 5% RH followed by a period of 4500 seconds of steady state. Similar to the temperature measurements here too an on-off controller was used hence oscillations about target humidity value can be seen. Additionally, for every constant relative humidity step the concentration was varied from 0-800 ppb in order to get an average effect of the changing relative humidity on the entire concentration range.

4) Cross Correlation: There were some interdependence observed in the independent variables especially between temperature and formaldehyde emissions from the formalin container due to environment dependent adsorption properties of formaldehyde on surfaces because of its high dipole moment [50]–[52]. Due to these interdependencies there was some correspondence between the variation of these 'independent variables' which have to be taken into account while evaluating the data.

E. Data Calibration Algorithm

This section describes the data pre-processing and machine learning algorithm we use in the second phase of our evaluation to measure the ease of modelling of errors in the different sensors. The calibration algorithm will be created over the data output from the built-in sensor calibration algorithm with a goal of improving its performance. In this study, a single calibration model is presented which is a potential limitation as it prevents us to see the effect of varying algorithms of differing methodologies on sensor calibration.

Before pre-processing of the data-set, the various experiments are compiled into three data-sets based on the primary controlled target variable i.e. concentration, temperature and humidity. Followed by that, pre-processing of the entire data-set is carried out, which comprises of synchronizing the time series resolutions, standardizing the data set such that every feature has a mean of zero and a standard deviation of 1 and transforming the data set into a normal distribution using its quantile information. The pre-processed data was fed into the calibration model.

The model used for calibration is a two-stage model. The first stage of the model made use of a linear regression layer. Based on the residuals from the first stage, we then implemented random forest regression as the second stage. In the end, we presented the accuracy of the calibration model as a whole. The two-stage model has been previously used for air quality sensing of $PM_{2.5}$ [53]. The need for a two-stage regression model is explained by the limitations of random forest regression [54]. Theoretically, the linear regression layer not only captures linear relationship between the reference sensor and the individual low-cost sensors, but also help address the generalization problem that random forest algorithms fail to accomplish when dealing with regression tasks [53]. The random forest layer is expected to capture the nonlinear association between the input and output variables [54]. The data set was then divided such that the first 70% of the data was used as a training set and the remaining 30% as a testing set. The training procedures were carried out with a walk-forward time series cross validation [55] which has been demonstrated to work well in time series data set.

III. Experimental Data

During the data acquisition phase using the experimental setup about 422 hours of comparative data of the collocated reference and low-cost sensing devices were obtained. In these experiments, three independent variables i.e. concentration, temperature and humidity were varied based on the boundary conditions mentioned in section II-D. The data-set is open source and available to be used [22].

A. Experiments

The reference device was sampled per minute for its data. The low-cost sensors data was initially sampled every 3 seconds and later averaged to every minute to get rid of its inherent variability. A total of 5 experiments were conducted as listed in table IV. The ambient room temperature and relative humidity in the experimental setup was was 22° C and 12% respectively. In the first two experiments, concentration is the only variable which is varied over time. In the first experiment the comparative sensor data is obtained for lower concentration range whereas for the second one it is obtained for a higher concentration range. Experiment sets 3 and 4 are where temperature is varied along with concentration. The temperature variation is carried out in a step fashion in experiment set 3 whereas in a random manner in experiment set 4. The Experiment 5 is an experiment wherein the humidity is increased in a step wise manner to see the effect on sensor performance. Snippets of the time series data generated from the experiments are presented in Fig. 5 for reference.

Conversion of Bosch sensor data from resistance to ppb was done by an open source code [56] because of the closed source nature of the BSEC library, its inability to work with Arduino Mega and lack of ppm value output.

IV. Results

In this section we will present the results obtained from the two phases of evaluation: pre-calibration and post-calibration. During the pre-calibration analysis in section IV-A we analyse the data output of the sensors from its built in calibration algorithm. The error or difference between the value of the reference device and the low-cost sensor is the main metric analyzed. In the results from post-calibration phase presented in section IV-B we examine the change in errors in the low-cost sensor after implementing our machine learning calibration algorithm over the pre-existing data stream. The results of the pre-calibration phase give us an idea of the inherent errors in the different sensors and the post-calibration results give us an idea of the ease with which these errors can be modelled and corrected. Hence presenting a complete evaluation of the low-cost sensing units.

A. Pre-Calibration Results

The main characteristic observed during the analysis of the data was error, which is basically the difference between the reading of the reference device and the low-cost sensor at a particular time step when placed in identical conditions. This parameter gives us a good estimate about the comparative performance of the low-cost sensor with respect to the reference device. Its magnitude tells us about the value by which the low-cost sensor differs from the reference. The sign of the error tells us if it underestimates or overestimates the concentration of formaldehyde.

Throughout the analysis the error of the reference device with respect to the changing environmental conditions is assumed to be zero. The insights and preliminary analysis based on this characteristic are listed below:

1) On comparing the data of the multiple units of ZE08-CH2O and BME680 sensor: the ZE08-CH2O sensors

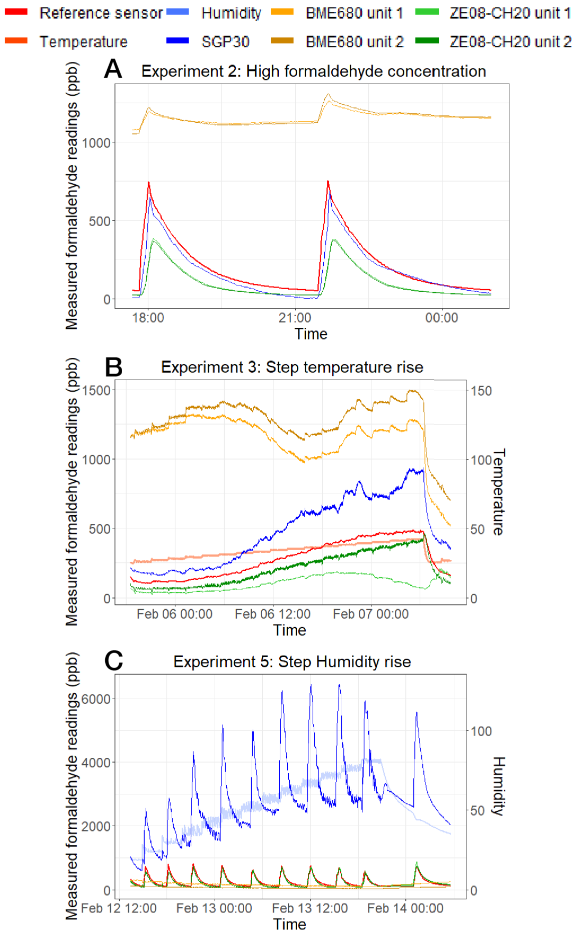


Fig. 5. Time series snippets of sensor readings for some experiments mentioned in table IV following the sampling strategy of section II-D, A: High concentration experiments B: Step temperature experiments C: Step humidity experiments.

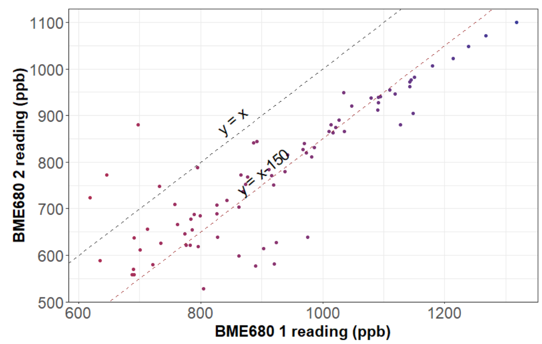


Fig. 7. BME680 sensor unit comparison.

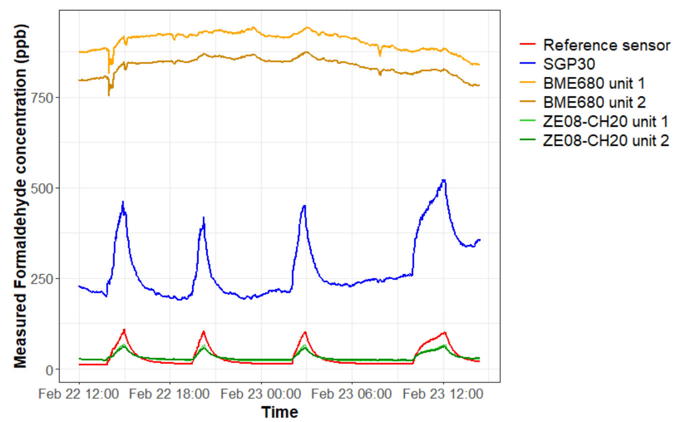


Fig. 8. Response of low-cost sensors on gradually raising the concentration of formaldehyde from 10-125 ppb (snippet of experiment 1: low concentration experiment).

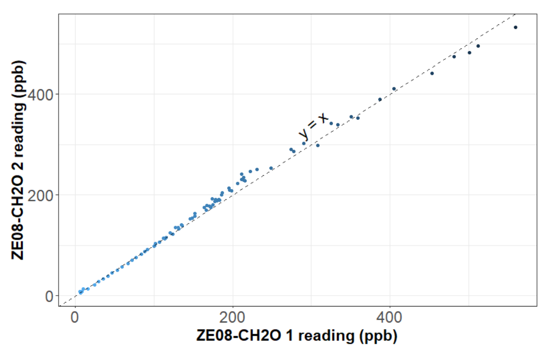


Fig. 6. ZE08-CH2O sensor unit comparison.

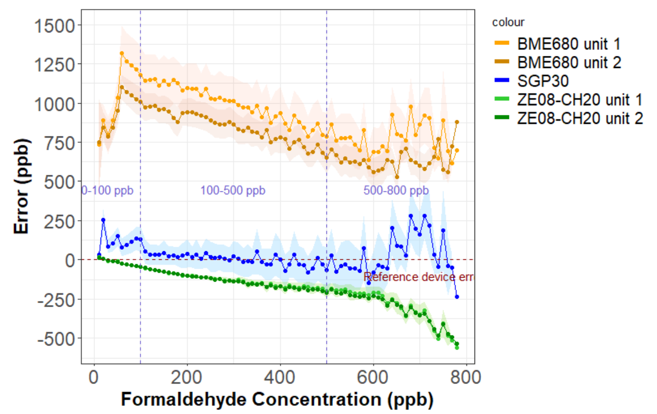


Fig. 9. Sensor error for different concentration ranges, values averaged for every 10 ppb (solid line), shaded area represents standard deviation (SD).

showed very good correspondence between the readings (Fig. 6) by showing a high Pearson correlation coefficient (r) of 0.99 under constant temperature and humidity. Whereas there was a significant but relatively constant offset seen in the measurements of the two BME680 sensors, as seen in the Fig. 7. The r for the BME680 for the corresponding dataset was also slightly lower at 0.85.

2) Fig. 8 shows a part of the time series of the experiment 1 as mentioned in table IV. Here the concentration of formaldehyde is maintained at a low level and very slowly increased from 10-125 ppb. The sensors SGP30 and ZE08-CH2O respond well to changes at these low concentrations of formaldehyde, showing their effectiveness and low limit of detection. This can be seen in Fig. 8 where the ZE08-CH2O demonstrates a r of

0.99 and the SGP30 shows 0.88. They also respond very promptly to the rise and fall of the gas concentration showing their high sensitivity and response, however as per our observations, the BME680 sensor is unable to detect changes of concentration at such low concentrations. The similarity in time series profiles of the ZE08-CH2O, SGP30 and the reference device also show their quick recovery rate as they are able to follow the change in gas concentration as adeptly as the reference device.

3) On comparing the behaviour of the sensors over the entire concentration range of 0-800 ppb we can see how their performance varies with different concentration ranges: at low concentration (0-100 ppb), the ZE08-CH2O outperforms the other two sensors by showing a very low mean error (ME) of −17.49 ppb when compared to the SGP30 (ME = 118.29 ppb) and BME680 (ME = 1055.84 ppb). Whereas, in the dangerous range (100-500 ppb), the SGP30 shows a considerably superior performance to the ZE08-CH2O (The SGP30 shows a lower mean error of 6.97 ppb when compared to −130.04 ppb of ZE08-CH2O but a higher variability). At high concentrations (500-800 ppb), all the sensors show comparatively high error. The BME680's decreasing error also becomes comparable to the other error shown by the other two sensors.

4) All the sensors show a certain degree of sensitivity to changing temperature but the metal oxide sensors (MOS) especially the SGP30 show maximum deviation at higher temperatures (>45° C) where the mean error (ME) is on an average as high as 433 ppb. However in the range of temperature usually observed indoors i.e. from 20-40° C, the performance and magnitude of error of both the SGP30 (ME = 140.55 ppb) and ZE08-CH2O (ME = −137.16 ppb) sensors are very similar, even though we see them showing the opposite signs (as seen in Fig. 10). The raw BME680 sensor readings show the highest error and variability amongst all three for any given temperature range.

5) The performance of the EC sensors (ZE08-CH2O) stands out when we look at the comparative performance of the sensors with respect to humidity. Over the entire range of humidity they show a consistently low mean error of −29.43 ppb (as seen in Fig. 11). The BME680 sensors also show low mean error and variance at higher humidity (above 45% RH) of 48.33 ppb. However the SGP30 shows extremely high sensitivity to humidity and shows an increasing trend of errors as the mean error below 40% RH is 375.39 ppb but above 40 % RH it reaches as high as 2526.65 ppb.

B. Post-Calibration Results

For evaluation of the post calibration performance mean absolute error (MAE) and standard deviation (SD) is the metric used. MAE helps us evaluate the individual performance of sensors in experiments and also their overall performance by preventing the occurrence of inaccurately optimistic results due to cancelling out of errors with opposite signs. SD helps us see if the model is able to reduce the variation present in the sensor errors.

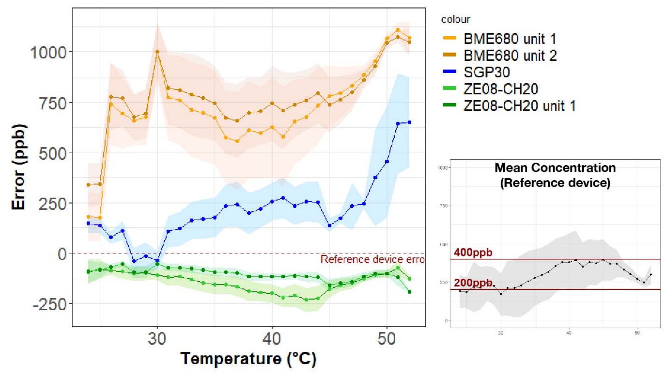


Fig. 10. Sensor error (difference between sensor reading and reference device) for different temperatures, values averaged for every 1° C (solid line), shaded area represents SD.

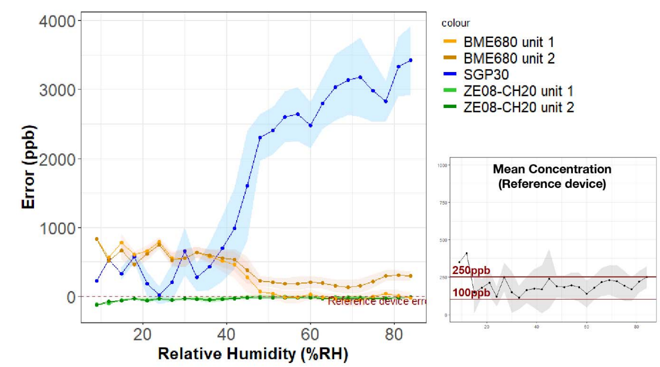


Fig. 11. Sensor error for different values of relative humidity, values averaged for every 5% RH (solid line), shaded area represents SD.

TABLE V
SENSOR PERFORMANCE PRE-CALIBRATION FOR TEST DATA-SET

Experiment	BME680		SGP30		ZE08-CH2O	
	MAE	SD	MAE	SD	MAE	SD
Concentration	446.93	159.32	243.24	104.32	42.82	56.61
Temperature	310.55	208.63	196.46	274.33	128.59	64.40
Humidity	249.20	186.20	1305.15	1518.63	41.17	41.32

Table V presents the sensor's average performance pre-calibration for different experiments and table VI shows the performance on the test data set after calibrating using the machine learning model described in section II-E. Additionally, we use target diagrams (Fig. 12) to assess the model's effect on the bias and the variance of the measurements. The bias is simply defined as the difference between the mean of the output values and the mean of the input reference value as seen in equation 2, while the variance was demonstrated by root-mean-square deviation (RMSD), which can be calculated as follows:

$$RMSD = \frac{1}{\sigma_{ref}} \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(y_i - x_i)^2]} \quad (1)$$

TABLE VI
SENSOR PERFORMANCE POST-CALIBRATION FOR TEST DATA-SET

Experiment	BME680		SGP30		ZE08-CH2O	
	MAE	SD	MAE	SD	MAE	SD
Concentration	138.23	114.44	206.99	104.10	21.76	26.31
Temperature	202.08	136.77	113.67	91.51	130.87	68.05
Humidity	173.81	124.90	141.50	127.13	82.30	88.57

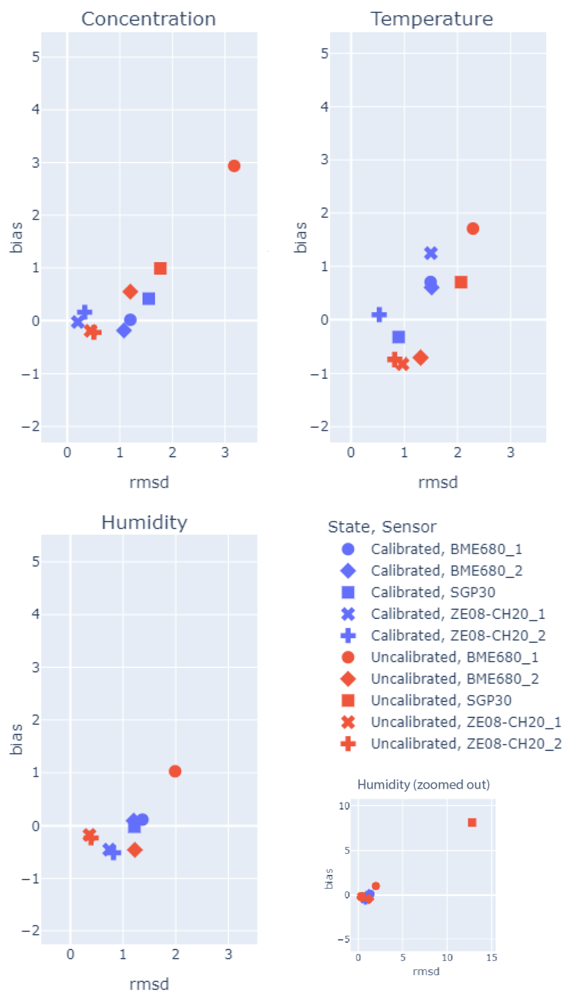


Fig. 12. Target diagrams for analyzing change in bias and RMSD post-calibration.

$$Bias = \frac{1}{\sigma_{ref}} \left(\frac{\sum y}{N} - \frac{\sum x}{N} \right) \tag{2}$$

where x, y represent low-cost sensor and reference values respectively and x_i, y_i represent their i^{th} measurement. N is the total number of observations and σ_{ref} is the standard deviation of the reference measurements. Both the metrics are normalized by σ_{ref} (the standard deviation of the reference values). The closer the points are to the origin, the better the calibration models perform.

From the target diagrams (Fig. 12) and table V and VI we can see that after calibration the mean absolute error and the bias is highly reduced all across the board except in the case of the humidity experiment of sensor ZE08-CH2O (doubling of MAE from 41.17 ppb to 82.30 ppb). This could possibly be due to the already low error exhibited by the ZE08-CH2O sensor for humidity which at their current state are difficult to model. The MOS sensors show a very high reduction in both RMSD and bias all throughout the three experiments (in Fig. 12) which does show that even though their error is high they are easy to model. SGP30 benefits the most from the calibration as seen in table V and VI, it shows a very high reduction in its errors especially with respect to humidity (89% reduction in MAE from 1305.15 ppb to 141.50 ppb) bringing it at par with the other sensors. BME680 also shows a high level of reduction in its errors in all experiments post calibration (49% reduction in average MAE from 335.56 ppb to 171.37 ppb). The reduction in SD of the sensors errors is mainly in regions with high sensor variation like the humidity experiment of SGP30. However, RMSD (Fig. 12) is highly reduced all across the board apart from some exceptional cases like the BME680 sensor 2 and humidity experiment of the ZE08-CH2O.

On comparing the overall performance throughout the different experiments the uncalibrated ZE08-CH2O (MAE = 70.86 ppb) shows the best performance when compared to the calibrated SGP30 (MAE = 154.54 ppb) and BME680 (MAE = 171.37 ppb) sensors and even the calibrated ZE08-CH2O (MAE = 78.81 ppb).

V. DISCUSSION

Main objective defined for this paper was to evaluate the efficacy of low-cost formaldehyde sensors for creation of a mobile sensing node. There were two types of sensors under consideration: an EC sensor (ZE08-CH2O) and two MOS sensors (BME680 and SGP30). The study began with a detailed study of regulations and indoor environment conditions to define the boundary conditions by understanding at what indoor environmental conditions and concentrations formaldehyde is dangerous. The evaluation of these sensors was done by looking at their performance against a reference device in varying conditions created in a controlled environment. An additional machine learning calibration algorithm was also created to test the ease of modelling of the sensor's errors.

The results of low-cost sensor performance from both the phases of evaluation were promising. All the sensors exhibited a good level of accuracy in certain parts of the entire domain of the variables of interest. The raw data analysis showed that the EC sensors being evaluated (ZE08-CH2O) performed better when low concentration of formaldehyde are present in accordance to some previous research [57], [58]. Additionally it also exhibited a high level of resilience to changing environmental conditions, especially to varying humidity which is critical component triggering formaldehyde emission. The SGP30 metal oxide sensor does show promising results in moderate concentration of formaldehyde and is fairly resilient to changes in temperature. However as per our experiments it also displays a high level of susceptibility to relative humidity over 40%. The BME680 metal oxide sensor demonstrates comparable results to other sensors at higher concentration of formaldehyde and humidity but shows poor results overall which has been seen in a previous paper as well [59]. This posits that certain sensors might be more suitable for certain use cases depending on the possible environmental conditions and formaldehyde concentrations the sensor would be exposed to. In our particular use case, the mobile sensing node was being designed to be deployed in an indoor environment to detect possible exposure of a person to dangerous concentrations of formaldehyde. For this particular scenario both SGP30, ZE08-CH2O are suitable based on their raw data results.

Post calibration of the data stream we can see that the performance of both the metal oxide sensors (SGP30 and BME680) showed a very high improvement. This characteristic has been observed in some previous studies [60], [61]. However, the improvement in the performance of ZE08-CH2O sensor was minimal. These results also show that the EC sensor being evaluated does perform better by itself in the absence of any calibration algorithm as seen particularly in varying humidity experiments. This is probably because its errors are difficult to be modelled with our current modelling algorithm. Overall the uncalibrated ZE08-CH2O performs the best. Therefore even though the EC sensor is difficult to model its good performance and innate resistance to environmental conditions due to most likely due to its underlying EC sensing principle makes it the leading choice. Therefore after considering the overall performance of all the sensors the ZE08-CH2O was chosen for building the mobile sensing node. However, one interesting result worth mentioning is that the SGP30 also does show considerably good resistance to temperature and humidity post calibration. It also performs well in the concentration range of formaldehyde deemed dangerous (100-500 ppb) as per its raw data results. The fact that it has a much smaller form factor compared to the ZE08-CH2O makes it a choice worth considering for a mobile sensing node where there are multiple gas sensors and a space constraint.

VI. Conclusion

The study evaluated various low-cost sensors with differing sensing principles for detection of formaldehyde in indoor environment. The sensors being evaluated measured the gas concentration via two techniques i.e. electrochemical sensing and metal oxide film. The results indicated a domain specific behavior where the electrochemical sensors and metal oxide sensors can be seen performing well in certain regions of the domain of the independent variables (formaldehyde concentration, humidity and temperature). The results showed that the electrochemical sensors were possibly more robust and resistant to fluctuations to the environmental parameters. However the errors of the metal oxide sensors were easier to model and calibrate. The study also demonstrated a new way of evaluating sensors in a controlled environment where their performance is evaluated by subjecting them to sampled diverse environment representing their environment of deployment. This methodology can also serve as a very effective tool for generation of synthetic training data for machine learning sensor models. Furthermore, the diverse data-set generated from this setup was made open-source in order to enable future machine learning models to work well in all kinds of locations of deployment with varying environmental conditions. There are however some limitations to this methodology as the sensor evaluation takes place only in the presence of the target gas which excludes possible errors caused by cross-interfering gases which are generally present in real world environments. However, dealing with selectivity challenges is a problem which might need to be dealt in a more comprehensive way and can be a topic for future studies.

There were some more limitations to the studies mainly in two fronts: hardware and modelling. On the hardware front a higher number of units of each sensor might have helped yield a more conclusive inference of the performance of the sensors and their respective sensing principles. The lack of enough sensing units was primarily due to to the global electronics supply crunch occurring during the time of conduction of the experiments. On the modelling front use of a number of different models would help us understand if the ease of modelling of the errors of the various sensors differs with varying modelling techniques and algorithms. Additionally in this study a diverse data set was used for training of the machine learning model but there were no control data-sets created to measure the actual advantage achieved by the diverse data-set. Tackling some of these limitations might become subjects for future work.

Acknowledgment

The authors would like to thank Gasera Oy and Tuomas Hieta for being a part of this project and would also like to thank Prof. Kalevi Ekman and the Faculty of Aalto Design factory for their support.

References

[1] O. S. Brittain, H. Wood, and P. Kumar, "Prioritising indoor air quality in building design can mitigate future airborne viral outbreaks," *Cities Health*, pp. 1–4, Jul. 2020.

[2] A. J. Koivisto *et al.*, "Source specific exposure and risk assessment for indoor aerosols," *Sci. Total Environ.*, vol. 668, pp. 13–24, Jun. 2019.

[3] H. Omidvarborna, P. Kumar, J. Hayward, M. Gupta, and E. G. S. Nascimento, "Low-cost air quality sensing towards smart Homes," *Atmosphere*, vol. 12, no. 4, p. 453, Apr. 2021.

[4] V. Goletto *et al.*, "Formaldehyde and total VOC (TVOC) commercial low-cost monitoring devices: From an evaluation in controlled conditions to a use case application in a real building," *Chemosensors*, vol. 8, no. 1, p. 8, Jan. 2020.

[5] D. Hasenfratz, O. Saukh, and L. Thiele, "On-the-fly calibration of low-cost gas sensors," in *Proc. Eur. Conf. Wireless Sensor Netw.* Berlin, Germany: Springer, 2012, pp. 228–244.

[6] P. Nowack, L. Konstantinovskiy, H. Gardiner, and J. Cant, "Machine learning calibration of low-cost NO_2, and PM_{10} sensors: Non-linear algorithms and their impact on site transferability," *Atmos. Meas. Techn.*, vol. 14, no. 8, pp. 5637–5655, 2021.

[7] T. Godish, *Indoor Environmental Quality*. Boca Raton, FL, USA: CRC Press, 2016.

[8] C. Pijolat, C. Pupier, M. Sauvan, G. Tournier, and R. Lalauze, "Gas detection for automotive pollution control," *Sens. Actuators B, Chem.*, vol. 59, nos. 2–3, pp. 195–202, 1999.

[9] *Indoor Air Quality: Organic Pollutants*, World Health Org., Geneva, Switzerland, 1989.

[10] P. Kumar *et al.*, "Real-time sensors for indoor air monitoring and challenges ahead in deploying them to urban buildings," *Sci. Total Environ.*, vols. 560–561, pp. 150–159, Aug. 2016.

[11] T. J. Kulle, "Acute odor and irritation response in healthy nonsmokers with formaldehyde exposure," *Inhalation Toxicol.*, vol. 5, no. 3, pp. 323–332, Jan. 1993.

[12] A. Casset *et al.*, "Inhaled formaldehyde exposure: Effect on bronchial response to mite allergen in sensitized asthma patients," *Allergy*, vol. 61, no. 11, pp. 1344–1350, Nov. 2006.

[13] G. McGwin, J. Lienert, and J. I. Kennedy, "Formaldehyde exposure and asthma in children: A systematic review," *Environ. Health Perspect.*, vol. 118, no. 3, pp. 313–317, Mar. 2010.

[14] International Agency for Research on Cancer. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Volume 88: Formaldehyde, 2-Butoxyethanol and 1-Tert-Butoxypropan-2-OL.* Accessed: Feb. 30, 2012. [Online]. Available: http://www.Monographs/vol100F/mono100F-29

[15] IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, "Formaldehyde, 2-butoxyethanol and 1-tert-butoxypropan-2-ol," *IARC Monographs Eval. Carcinogenic Risks Hum.*, vol. 88, p. 1, Jun. 2006.

[16] V. J. Cogliano *et al.*, "Meeting report: Summary of *IARC Monographs* on formaldehyde, 2-butoxyethanol, and 1-*tert*-butoxy-2-propanol," *Environ. Health Perspect.*, vol. 113, no. 9, pp. 1205–1208, Sep. 2005.

[17] A. Weber. (2021). *Final Report: Inexpensive Formaldehyde Sensors for Indoor Air Quality (IAQ) Applications*. [Online]. Available: https://cfpub.epa.gov/ncer_abstracts/index.cfm/fuseaction/display.abstractDetail abstract_id10921/report/F

[18] R. Golden, "Identifying an indoor air exposure limit for formaldehyde considering both irritation and cancer hazards," *Crit. Rev. Toxicol.*, vol. 41, no. 8, pp. 672–721, Sep. 2011.

[19] W. T. Honeycutt, M. T. Ley, and N. F. Materer, "Precision and limits of detection for selected commercially available, low-cost carbon dioxide and methane gas sensors," *Sensors*, vol. 19, no. 14, p. 3157, Jul. 2019.

[20] M. Kamionka, P. Breuil, and C. Pijolat, "Calibration of a multivariate gas sensing device for atmospheric pollution measurement," *Sens. Actuators B, Chem.*, vol. 118, nos. 1–2, pp. 323–327, Oct. 2006.

[21] A. Samad, D. R. O. Nuñez, G. C. S. Castillo, B. Laquai, and U. Vogt, "Effect of relative humidity and air temperature on the results obtained from low-cost gas sensors for ambient air quality measurements," *Sensors*, vol. 20, no. 18, p. 5175, Sep. 2020.

[22] Megasense. (2021). *Open Source Formaldehyde Database*. [Online]. Available: https://megasense-server.cs.helsinki.fi/datasets.html

[23] V. Papapostolou, H. Zhang, B. J. Feenstra, and A. Polidori, "Development of an environmental chamber for evaluating the performance of low-cost air quality sensors under controlled conditions," *Atmos. Environ.*, vol. 171, pp. 82–90, Dec. 2017.

[24] H. Omidvarborna, P. Kumar, and A. Tiwari, "Envilution' chamber for performance evaluation of low-cost sensors," *Atmos. Environ.*, vol. 223, p. 117264, Feb. 2020.

[25] Z. Wang, W. W. Delp, and B. C. Singer, "Performance of low-cost indoor air quality monitors for $PM_{2.5}$ and PM_{10} from residential sources," *Building Environ.*, vol. 171, Mar. 2020, Art. no. 106654.

[26] J. Kuula *et al.*, "Laboratory evaluation of particle-size selectivity of optical low-cost particulate matter sensors," *Atmos. Meas. Techn.*, vol. 13, no. 5, pp. 2413–2423, May 2020.

[27] H. Bolt, G. Johanson, G. Nielsen, D. Papameletiou, and C. Klein, "Scoel/rec/125 formaldehyde," Tech. Rep.

[28] D. A. Kaden, C. Mandin, G. D. Nielsen, and P. Wolkoff, "Formaldehyde," in *WHO Guidelines for Indoor Air Quality: Selected Pollutants*. Geneva, Switzerland: World Health Organization, 2010.

[29] W. Europe, "Air quality guidelines 2nd edition regional office for Europe," Tech. Rep., 2000.

[30] T. Godish, "Formaldehyde exposures from tobacco smoke: A review," *Amer. J. Public Health*, vol. 79, no. 8, pp. 1044–1045, Aug. 1989.

[31] R. Fortmann, P. Kariher, and R. Clayton, "Indoor air quality: Residential cooking exposures," Tech. Rep., 2001.

[32] I. Andersen, G. R. Lundqvist, and L. Mølhave, "Indoor air pollution due to chipboard used as a construction material," *Atmos. Environ.*, vol. 9, no. 12, pp. 1121–1127, Jan. 1975.

[33] S. Huang, J. Xiong, C. Cai, W. Xu, and Y. Zhang, "Influence of humidity on the initial emittable concentration of formaldehyde and hexaldehyde in building materials: Experimental observation and correlation," *Sci. Rep.*, vol. 6, no. 1, pp. 1–9, Sep. 2016.

[34] C.-C. Lin, K.-P. Yu, P. Zhao, and G. W.-M. Lee, "Evaluation of impact factors on VOC emissions and concentrations from wooden flooring based on chamber tests," *Building Environ.*, vol. 44, no. 3, pp. 525–533, Mar. 2009.

[35] N. Brelih, "Thermal and acoustic comfort requirements in European standards and national regulations," *Fed. Eur. Heating, Ventilation Air Conditioning Associations J.*, p. 4, Mar. 2013.

[36] T. Salthammer, "The formaldehyde dilemma," *Int. J. Hygiene Environ. Health*, vol. 218, no. 4, pp. 433–436, Jun. 2015.

[37] M. Winkowski and T. Stacewicz, "Optical detection of formaldehyde in air in the 3.6 μm range," *Biomed. Opt. Exp.*, vol. 11, no. 12, pp. 7019–7031, 2020.

[38] J. Karhu, J. Kuula, A. Virkkula, H. Timonen, M. Vainio, and T. Hieta, "Cantilever-enhanced photoacoustic measurement of light-absorbing aerosols," *Aerosol Sci. Technol.*, vol. 56, no. 1, pp. 92–100, 2021.

[39] M. J. Tierney and H. O. L. Kim, "Electrochemical gas sensor with extremely fast response times," *Anal. Chem.*, vol. 65, no. 23, pp. 3435–3440, Dec. 1993.

[40] N. Castell *et al.*, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?" *Environ. Int.*, vol. 99, pp. 293–302, Feb. 2017.

[41] C. Wang, L. Yin, L. Zhang, D. Xiang, and R. Gao, "Metal oxide gas sensors: Sensitivity and influencing factors," *Sensors*, vol. 10, no. 3, pp. 2088–2106, 2010.

[42] J. H. Sohn, M. Atzeni, L. Zeller, and G. Pioggia, "Characterisation of humidity dependence of a metal oxide semiconductor sensor array using partial least squares," *Sens. Actuators B, Chem.*, vol. 131, no. 1, pp. 230–235, 2008.

[43] A. K. Farquhar, G. S. Henshaw, and D. E. Williams, "Understanding and correcting unwanted influences on the signal from electrochemical gas sensors," *ACS Sensors*, vol. 6, no. 3, pp. 1295–1304, Mar. 2021.

[44] G. Li, Z. Cheng, Q. Xiang, L. Yan, X. Wang, and J. Xu, "Bimetal PdAu decorated SnO_2 nanosheets based gas sensor with temperature-dependent dual selectivity for detecting formaldehyde and acetone," *Sens. Actuators B, Chem.*, vol. 283, pp. 590–601, Mar. 2019.

[45] S. Munir, M. Mayfield, D. Coca, S. A. Jubb, and O. Osammor, "Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities—A case study in Sheffield," *Environ. Monitor. Assessment*, vol. 191, no. 2, pp. 1–22, Feb. 2019.

[46] C. Malings *et al.*, "Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring," *Atmos. Meas. Techn.*, vol. 12, no. 2, pp. 903–920, Feb. 2019.

[47] P. Wei *et al.*, "Development and evaluation of a robust temperature sensitive algorithm for long term NO_2 gas sensor network data correction," *Atmos. Environ.*, vol. 230, Jun. 2020, Art. no. 117509.

[48] M. Si, Y. Xiong, S. Du, and K. Du, "Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine-learning methods," *Atmos. Meas. Techn.*, vol. 13, no. 4, pp. 1693–1707, Apr. 2020.

[49] G. Oy. (2021). *Gasera One Formaldehyde*. [Online]. Available: https://www.gasera.fi/product/gasera-one-formaldehyde-photoacoustic-gas-monitor/

[50] H. Dahnke, G. von Basum, K. Kleinermanns, P. Hering, and M. Mürtz, "Rapid formaldehyde monitoring in ambient air by means of mid-infrared cavity leak-out spectroscopy," *Appl. Phys. B, Lasers Opt.*, vol. 75, nos. 2–3, pp. 311–316, Sep. 2002.

[51] M. Horstjann *et al.*, "Formaldehyde sensor using interband cascade laser based quartz-enhanced photoacoustic spectroscopy," *Appl. Phys. B, Lasers Opt.*, vol. 79, no. 7, pp. 799–803, 2004.

[52] Y. W. Choi *et al.*, "Formaldehyde vapor characteristics in varied decontamination environments," *Appl. Biosafety*, vol. 26, no. 1, pp. 33–41, Mar. 2021.

[53] Y. Lin, W. Dong, and Y. Chen, "Calibrating low-cost sensors by a two-phase learning approach for urban air quality measurement," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–18, Mar. 2018.

[54] H. Zhang, D. Nettleton, and Z. Zhu, "Regression-enhanced random forests," Tech. Rep., 2017.

[55] M. Y. Hu, G. Zhang, C. X. Jiang, and B. E. Patuwo, "A cross-validation analysis of neural network out-of-sample performance in exchange rate forecasting," *Decis. Sci.*, vol. 30, no. 1, pp. 197–216, Jan. 1999.

[56] Jurgen. (2019). *Bme680 Arduino Library With TVOC*. [Online]. Available: https://github.com/juergs/BME680-Arduino-Library-with_TVOC-

[57] L. Ma *et al.*, "A low cost compact measurement system constructed using a smart electrochemical sensor for the real-time discrimination of fruit ripening," *Sensors*, vol. 16, no. 4, p. 501, Apr. 2016.

[58] P. Wei *et al.*, "Impact analysis of temperature and humidity conditions on electrochemical sensor response in ambient air quality monitoring," *Sensors*, vol. 18, no. 2, p. 59, 2018.

[59] G. Yurko *et al.*, "Real-time sensor response characteristics of 3 commercial metal oxide sensors for detection of BTEX and chlorinated aliphatic hydrocarbon organic vapors," *Chemosensors*, vol. 7, no. 3, p. 40, Aug. 2019.

[60] K. R. Mallires, D. Wang, V. V. Tipparaju, and N. Tao, “Developing a low-cost wearable personal exposure monitor for studying respiratory diseases using metal–oxide sensors,” *IEEE Sensors J.*, vol. 19, no. 18, pp. 8252–8261, Sep. 2019.

[61] C. Di Natale, S. Marco, F. Davide, and A. D’Amico, “Sensor-array calibration time reduction by dynamic modelling,” *Sens. Actuators B, Chem.*, vol. 25, nos. 1–3, pp. 578–583, Apr. 1995.

Arnab Chattopadhyay received the master’s degree in mechanical engineering with a specialization in mechatronics from Aalto University, Finland. He is currently pursuing the Ph.D. degree with the Department of Computer Science, University of Helsinki. His research interests include building ML systems for sensors and sensor networks and building of continuous machine learning (CML) applications for cyber–physical systems.

Andres Huertas received the degree in physics from the National University of Colombia in 2016 and the master’s degree in data science from the University of Helsinki in 2020. He is currently working as a Data Scientist at Accenture, Finland. His research interests include ML, deep learning, and AI in general.

Andrew Rebeiro-Hargrave received the Ph.D. degree from King’s College London in 2000. He worked on research and design at Nokia, Finland. He is a Nokia author with the Multimedia Group Communications from 2000 to 2012. He lectured in ICT at Kyushu University, Japan, and has published in telemedicine from 2012 to 2015. He is a Senior Researcher of Computer Science with the University of Helsinki. He has published on air quality monitoring and participatory sensing.

Pak Lun Fung received the Ph.D. degree in proxy derivation of air quality parameters in urban regions from the Institute for Atmospheric and Earth System Research (INAR), University of Helsinki, in 2022. He is now working as a Postdoctoral Researcher in INAR focusing on CO2 emission modeling under various autonomous driving scenarios. His research interests include low-cost sensor measurements and statistical modeling using machine learning techniques.

Samu Varjonen received the Ph.D. degree in computer science from the University of Helsinki, Finland, in 2012. He worked as a Researcher for five years with the Networking Research Group, Helsinki Institute for Information Technology (HIIT) focusing on the Host Identity Protocol (HIP) and trustworthy communication in distributed systems. At the present, he is working as a University Researcher with the University of Helsinki. The focus of his current work is on implementing a large, distributed testbed with mobile, stationary, and reference air quality sensors in Megasense Research Program.

Tuomas Hieta received the D.Sc. (Tech.) degree in the field of applied laser spectroscopy and electrical engineering from the Helsinki University of Technology, Finland, in 2011. He has worked as a Researcher and a Senior Research Scientist at MIKES Metrology and VTT in the field of industrial and environmental laser sensing applications before joining Gasera Ltd., in 2016. He is the author/coauthor of over 30 publications in peer-reviewed journals.

Sasu Tarkoma (Senior Member, IEEE) is a Professor of Computer Science with the University of Helsinki and the Dean of the Faculty of Science. He has authored four textbooks and has published over 240 scientific articles. He is a Fellow of IET and EAI. He has ten granted U.S. patents. His research interests are internet technology, distributed systems, data analytics, and mobile and ubiquitous computing. His research has received several best paper awards and mentions, for example, at IEEE PerCom, IEEE ICDCS, ACM CCR, and ACM OSR.

Tuukka Petäjä received the Ph.D. degree in physics from the University of Helsinki in 2006. He is a Full Professor with the Institute for Atmospheric and Earth System Research, University of Helsinki. His research interests include atmospheric aerosol particles and their role in climate change and air quality. He is a Clarivate Highly Cited Scientist.