buildings

MDPI

Article

LSTM Deep Learning Models for Virtual Sensing of Indoor Air Pollutants: A Feasible Alternative to Physical Sensors

Martin Gabriel * and Thomas Auer *

Building Technology and Climate Responsive Design, TUM School of Engineering and Design, Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany
* Correspondence: martin.gabriel@tum.de (M.G.); thomas.auer@tum.de (T.A.)

Abstract: Monitoring individual exposure to indoor air pollutants is crucial for human health and well-being. Due to the high spatiotemporal variations of indoor air pollutants, ubiquitous sensing is essential. However, the cost and maintenance associated with physical sensors make this currently infeasible. Consequently, this study investigates the feasibility of virtually sensing indoor air pollutants, such as particulate matter, volatile organic compounds (VOCs), and CO_2, using a long short-term memory (LSTM) deep learning model. Several years of accumulated measurement data were employed to train the model, which predicts indoor air pollutant concentrations based on Building Management System (BMS) data (e.g., temperature, humidity, illumination, noise, motion, and window state) as well as meteorological and outdoor pollution data. A cross-validation scheme and hyperparameter optimization were utilized to determine the best model parameters and evaluate its performance using common evaluation metrics (R^2, mean absolute error (MAE), root mean square error (RMSE)). The results demonstrate that the LSTM model can effectively replace physical indoor air pollutant sensors in the examined room, with evaluation metrics indicating a strong correlation in the testing set (MAE; CO_2: 15.4 ppm, $PM_{2.5}$: 0.3 µg/m^3, VOC: 20.1 IAQI; R^2; CO_2: 0.47, $PM_{2.5}$: 0.88, VOC:0.87). Additionally, the transferability of the model to other rooms was tested, with good results for CO_2 and mixed results for VOC and particulate matter (MAE; CO_2: 21.9 ppm, $PM_{2.5}$: 0.3 µg/m^3, VOC: 52.7 IAQI; R^2; CO_2: 0.45, $PM_{2.5}$: 0.09, VOC:0.13). Despite these mixed results, they hint at the potential for a more broadly applicable approach to virtual sensing of indoor air pollutants, given the incorporation of more diverse datasets, thereby offering the potential for real-time occupant exposure monitoring and enhanced building operations.

Keywords: machine learning; deep learning; virtual sensing; LSTM; IAQ; monitoring

Citation: Gabriel, M.; Auer, T. LSTM Deep Learning Models for Virtual Sensing of Indoor Air Pollutants: A Feasible Alternative to Physical Sensors. *Buildings* **2023**, *13*, 1684. https://doi.org/10.3390/buildings13071684

Academic Editor: Etienne Saloux

Received: 27 May 2023
Revised: 28 June 2023
Accepted: 29 June 2023
Published: 30 June 2023

1. Introduction

Indoor air pollutants are of different sizes and types, are harmful at different concentrations, and have different intake pathways and effects. The major groups of pollutants are inorganic gases, organic gases, particulates, microbial pollutants, and viral and bacterial infections [1]. Controlling indoor air pollutant exposure is especially relevant since we spend up to 87% of our time in buildings [2], rendering them the most important environments. In efforts to make buildings more energy efficient, they have become better sealed and indoor spaces more dependent on HVAC systems [2]. Studies have shown that bad indoor air quality leads to a multitude of different symptoms and health impacts. The gravity of these impacts depends on the pollutants, their concentration, the exposure time, and individual factors such as age, constitution, and health [3]. Most frequently, occupants experience tiredness, burning eyes, headaches, and concentration problems [3]. Prolonged exposure may also lead to respiratory syndromes and immune system reactions such as asthma, especially in vulnerable groups such as children or elderly persons [3]. According to a study from the WHO, air pollution is a significant health threat and a primary environmental factor in causing premature deaths in Europe [4]. Exposure to fine particulate matter

has been linked to over 400,000 premature deaths in European countries [5]. Therefore, many countries are taking steps to reduce indoor air pollutant concentration by enforcing exposure limits. An in-detail summary of exposure limits in different countries is given in Abdul et al. [6]. The European Union largely adopted the exposure limits suggested by the WHO in [7]. Effective strategies for reducing indoor air pollutant exposure include source control, which involves identifying and minimizing sources of pollution, mitigation measures such as removing pollutants and introducing clean air, and monitoring indoor air pollutant concentrations [1]. These measures are situated at different points of the building life cycle: source control is relevant in the planning and construction phase of buildings, and mitigation and monitoring are required during the use of the building. Building codes progressively ensure source control in new buildings by restricting harmful, pollutant-emitting materials. However, the majority of existing building stock was built without awareness of indoor air quality concerns. Therefore, improving indoor air quality in existing buildings is key. Monitoring and mitigation measures are especially relevant in the non-residential sector since occupants have little influence on the indoor environment, as opposed to residential buildings. Therefore, the following study will focus on non-residential typologies in the existing building stock; specifically, typology—with high occupant density and prolonged exposure will be part of the study.

2. Indoor Air Pollutants in the Non-Residential Building Stock

Several studies have examined the spatiotemporal distribution of indoor air pollutants within rooms in non-residential building stock. Szigeti et al. [8] examined the spatiotemporal distribution of particulate matter in European office buildings and found a significant variation between buildings. Within buildings, temporal variation is more pronounced than spatial distribution variations [8]. According to Szigeti et al. [8], occupants may be exposed to significantly different pollutant concentrations in different rooms within a building. Li et al. [9] examined the spatiotemporal distribution of particulate matter within one room (workshop) with localized sources and found high spatial and temporal variations within a single room. Sahu et al. [10] examined the distribution of indoor air pollutants—CO_2, particulate matter, VOC—in a multi-story library with shared air volume. They found high spatial and temporal variations within the library, with temporal variations mainly driven by the number of occupants and spatial variation more pronounced between the different stories. Studies Szigeti et al. [8], Li et al. [9], and Sahu et al. [10] show that indoor air pollutants show significant spatiotemporal variations. Therefore, a continuous and spatiotemporally high-resolution monitoring is required in order to evaluate occupant exposure and control ventilation units. The state of research in continuous and spatiotemporal high-resolution indoor air pollutant monitoring was analyzed, looking at fifteen studies, monitoring indoor air pollutants in non-residential buildings: [9,11–24]. Over all studies, the most measured pollutant is particulate matter, examined by 14 out of 15 publications. In 8 of 15 studies measured one or multiple volatile organic compounds, and 7 out of 15 studies measured CO_2. Carbon monoxide and nitrogen dioxide were assessed in 5 out of 15 studies, and ozone and sulfur dioxide in 2 studies. Several other pollutants as nitrogen oxide and fungi, are only considered in 1 publication. The studies evaluated the importance of different pollutants in non-residential buildings and found that carbon monoxide and radon only accumulate in particular spaces, such as kitchens with gas ovens and basements, and are insignificant in typical non-residential buildings [11,25]. Irga et al. [14] found the levels of nitrogen oxide, volatile organic compounds, fungi, and sulfur dioxide to be harmless in 11 office buildings. According to Challoner et al. [22], the most problematic pollutant in non-residential buildings is fine particulate matter, exceeding health thresholds in 10% of the measured time. Additionally, carbon dioxide regularly exceeds the threshold of 1000 ppm in naturally ventilated buildings [15]. Likewise, volatile organic compounds are effectively controlled by mechanical ventilation systems but can reach problematic concentrations if the ventilation system is switched off or buildings are naturally ventilated [15]. The reviewed articles exclusively use on-site measurement technology since laboratory

analysis is not able to gather continuous, high-resolution measurements. Most studies used MOS technology VOC sensors, NDIR technology CO_2 sensors, and OPC technology particulate matter sensors.

3. Virtual Sensing for Indoor Air Pollutants

Ubiquitous monitoring of indoor air pollutants requires durable, low-cost, low-maintenance, low-energy sensor equipment. NDIR and OPC technologies are optical measurement principles requiring fragile components and suffer from measurement drift and longevity issues due to the build-up of foreign particles in the measurement chambers [26]. Furthermore, optical measurement methods have higher energy consumption and are less suited for battery operation. Even though MOS-based VOC sensors are low cost and low energy, MOS-based VOC measurements are prone to drift and suffer from reproducibility issues [27]. These drawbacks necessitate careful maintenance, diligent monitoring for failures, and thorough post-processing of data, which is often overlooked in building operations. A mitigation of these shortcomings is presented in previous studies [28,29] that developed calibration models to improve accuracy and reduce the drift of IAP sensors. However, those systems have not found their way into practice yet. Therefore, alternatives to the measurement of particulate matter (PM), volatile organic compounds (VOC), and carbon dioxide (CO_2) have to be found.

Virtual sensing of PM, VOC, and CO_2 is an alternative to ubiquitous sensor deployment. Virtual sensing "aims to approximate unmeasured physical quantities in a dynamic system using existing sensor information. This is especially beneficial when important locations of the system are difficult to instrument, or the cost of sensors is very high" [30]. "A virtual sensor uses low-cost measurements and mathematical models to estimate a difficult to measure or expensive quantity" [31]. The models are based on related physical measurements, control signals, operation information, and design information [32]. Virtual sensing finds widespread application in the domains of process control, automotive, avionics, and robotics [31]. However, with the exponential rise of available data points through developments in IoT and cost reduction of sensors, virtual sensing has been increasingly adopted in the building industry [32]. The application of virtual sensing in the building industry is manifold. Buildings gather many data points, and nearly every physical sensor can provide additional information for virtual sensing. Li et al. [31] gives an example of the potential of virtual sensing in buildings: "A 'smart' lighting fixture could provide power, lighting, and heat gain outputs based on the input control signal. A 'smart' window could provide estimates of heat gain and even solar radiation based on low-cost measurements and a model". Application scenarios in buildings include HVAC operation monitoring [33,34], indoor infiltration rate [35], zone temperature distribution [36], zone occupancy estimation [37] and indoor air pollutant monitoring [38–40]. Generally, virtual sensors can be differentiated into three application scenarios: replacement and backup, observation, and assistance [32]. Replacement and backup virtual sensors are deployed in parallel to their physical counterparts and, by computing the residuals between physical and virtual measurement, are able to detect sensor faults or calibration drifts [41], and can replace their physical counterparts if needed [32]. Observation virtual sensors estimate a data point without their physical counterpart using other measurements and mathematical models [32]. Assistance virtual sensors do not estimate a physical quantity but are integrated into other virtual sensors to improve accuracy. The output of assistance virtual sensors is often normalized [32]. Virtual sensors can further be differentiated regarding their modeling method and the underlying measurement characteristics [31]. Modeling methods are white-box, grey-box, and black-box models [31]. Measurement characteristics can be differentiated in transient (e.g., power usage, indoor temperature) or steady state (e.g., system failure state) data [31]. White and grey-box models require in-depth knowledge of the building. These approaches are infeasible for older buildings due to unavailable planning documents, undocumented changes, and performance deterioration. One alternative would be a black-box model, which requires extensive measurement data.

To summarize, virtual sensing for indoor air pollutant prediction in non-residential typologies requires transient-state observational virtual sensors created using a black-box modeling method.

Other studies have already examined the applicability of virtual sensing for indoor air pollutant prediction. A study by Gabriel and Auer [42] used multi-layer perceptron (MLP) artificial neural networks and support vector machines (SVM) to create an observational virtual particulate matter sensor based on available Building Management System (BMS) data (temperature, pressure, humidity, sound, illumination, window opening state, and printer power consumption). Six months of measurement data were used to train and test the two machine-learning models. Gabriel and Auer [42] found that MLPs performed best, and the results indicated that physical particulate matter sensors could be replaced by virtual sensors based on BMS data. Kusiak et al. [41] created virtual replacement sensors for temperature, humidity, and CO_2 with four modeling approaches for the calibration and monitoring of physical sensors using HVAC, climate, and other indoor air pollutant data. MLPs were found to perform best in modeling the physical sensors. Kusiak et al. [41] conclude that the virtual sensors are able to detect failures of their corresponding physical sensors and replace them if necessary. Skoen et al. [38] used MLPs to create an observational virtual sensor using temperature and humidity as input for modeling CO_2. Skoen et al. [38] concludes that estimating CO_2 only based on temperature and humidity is difficult and requires additional measurements to support the black-box model. Leidinger et al. [43] created a virtual replacement sensor for selective VOC sampling of formaldehyde, benzene, and naphthalene using an array of low-cost MOS sensors as input. Leidinger et al. [43] used linear discriminant analysis to estimate the target variables. Under laboratory conditions, the study achieved a classification ratio of over 99%. However, in field tests, the classification ratio significantly dropped (83%) due to VOC emissions of the hardware [43]. A summary of Literature on Virtual Sensor Creation is given in Table 1. Research in other domains showed that long short-term memory (LSTM) recurrent neural networks are suited for time-series data in virtual sensor creation due to their ability to incorporate measurements from a lookback window into their model. LSTMs are recurrent neural networks specialized in time series data by incorporating memory cells in their network architecture, which enable them to identify and remember patterns in time series data [30]. In the building industry, LSTMs have already been applied to building load management for forecasting energy consumption [44,45] and predicting occupancy [46]. LSTMs have not yet been applied to modeling virtual indoor air pollutant sensors.

Table 1. Summary of Literature on Virtual Sensor Creation.

Study	Virtual Sensor Type	Methods Used	Main Findings
Gabriel and Auer [42]	Particulate Matter	MLP, SVM	MLPs performed better than SVM; results show the potential of virtual sensors to replace physical ones
Kusiak et al. [41]	Temperature, Humidity, CO_2	MLP, SVM, Pacereg, RBF	MLP outperformed other models; Virtual sensors can detect and replace failing physical sensors
Skoen et al. [38]	CO_2	MLP	Estimating CO_2 based only on temperature and humidity is challenging
Leidinger et al. [43]	VOC Sampling	Linear Discriminant Analysis	99% lab accuracy, 83% field accuracy due to hardware VOC emissions
Karijadi et al. [44] and Jang et al. [45]	Energy Consumption	LSTM	LSTMs have been successfully applied in energy consumption forecasting
Qolomany et al. [46]	Occupancy	LSTM	LSTM can be used for predicting occupancy

4. Study Definition

The literature indicates that monitoring occupant pollutant exposure and mitigating indoor air pollutant concentrations is important to ensure health and well-being. Since occupants in non-residential typologies have no or low possibility of intervention regarding indoor air quality, particular care has to be taken in providing adequate indoor conditions in these typologies. Due to the high spatiotemporal variation of indoor air pollutants in non-residential buildings, high-resolution monitoring is required. However, ubiquitous sensing of indoor air pollutants is infeasible due to the high cost and time required for installation, operation, and maintenance. Therefore, virtual sensing is suggested as an alternative to the ubiquitous deployment of physical sensors. Previous studies have already been conducted, applying virtual sensing to indoor air pollutant estimation. However, these studies mostly assessed only one air pollutant, even though exposure monitoring requires the estimation of multiple pollutants. Furthermore, all currently available studies build virtual sensors based on data from one zone in a single typology and do not consider the transferability of the models to other zones and/or typologies. Additionally, all studies reviewed here used less than a year of measurement data to build the models, thus introducing significant bias into the models. Despite the demonstrated effectiveness of LSTM in handling time-series problems across various domains, it is observed that none of the known virtual sensing approaches to indoor air pollutants have adopted LSTM as their modeling approach [47,48].

Therefore, our study examines the feasibility of observational virtual sensors for PM, VOC, and CO_2 based on an LSTM modeling approach. The study uses multiple years of accumulated measurement data from multiple zones and typologies to build the virtual-sensor model and check its transferability to other zones and typologies. The capability of the virtual sensor was evaluated independently for the room where the model was trained and on unknown rooms.

Figure 1 gives a visual overview of the study definition.

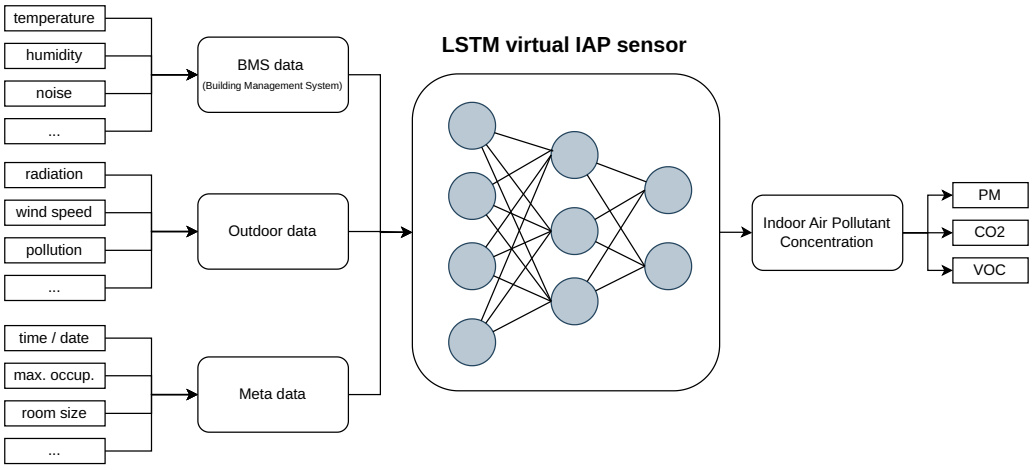

Figure 1. Flowchart of the study definition with a black box LSTM model and input/output components (Own representation).

5. Methods

In this study, we employed an LSTM model trained on collected measurement data to predict indoor air pollutant concentrations using BMS, outdoor meteorological, and outdoor pollution data as model inputs. In the following sections, we detail the methods used. Section 5.1 describes the steps performed in order to build the dataset, including the measurement equipment, the measurement setup, and the measurement location. Section 5.2 presents the steps to preprocess the data for machine learning model training. Section 5.3 encompasses the training of the models, while Section 5.4 focuses on model evaluation. Finally, the transferability tests are detailed in Section 5.5.

A graphical representation of the method is illustrated in Figure 2.

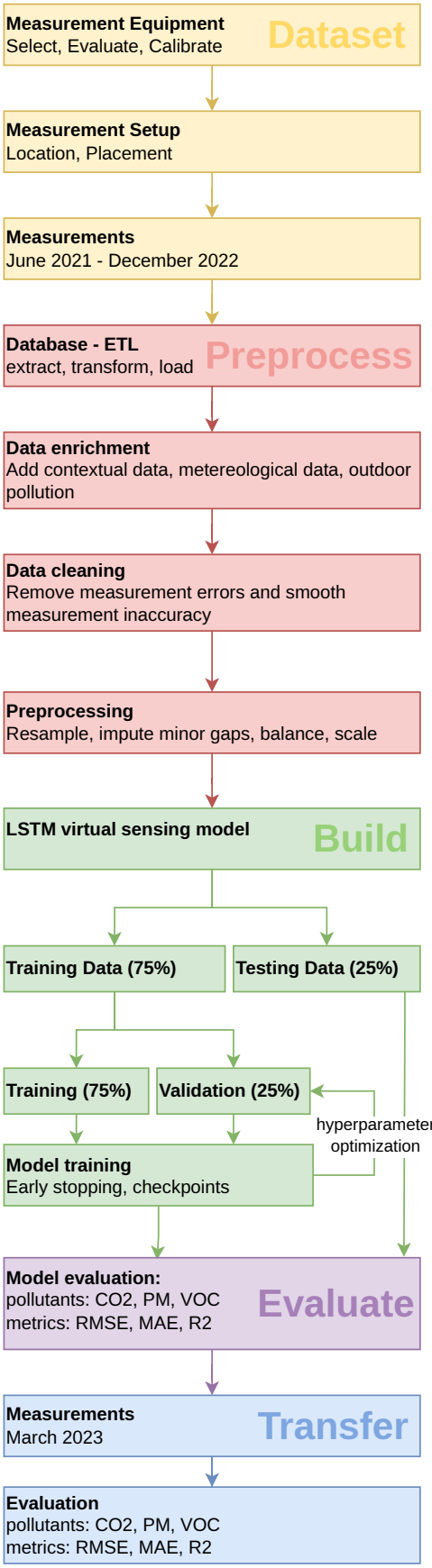

Figure 2. Flowchart of the implemented data processing and machine learning pipeline (Own representation).

5.1. Measurements

Measurement equipment was developed for indoor air pollutant (IAP) measurements. The measurement infrastructure for the BMS, meteorological, and outdoor pollution data was already in place. The BMS is implemented based on the LoRaWAN standard. LoRaWAN is a wireless IoT standard that achieves long-range communication with low power consumption, thus enabling battery-powered nodes. Due to the minimal installation effort and battery-powered nodes, it is applicable as a retrofit solution. The BMS data recorded measurements at a 1 min interval.

The IAP nodes are required to measure CO_2 concentrations in parts per million (ppm), particulate matter concentrations in micrograms per cubic meter ($\mu g/m^3$), and total volatile organic compound concentrations as Indoor Air Quality Index (IAQI). Furthermore, the IAP nodes must achieve continuous, automated measurements with a high sampling rate (10 s) over a prolonged period of time and should account for measurement drift by frequent recalibration. While the initial data sampling rate is 10 s, these measurements will be resampled to a one-minute interval later. This higher sampling rate allows for smoother and more reliable data, as it enables using more data points for each resampled data point. Due to the high volume of data collected, data must be stored centrally rather than locally on the measurement nodes. Therefore, a communication infrastructure supporting high data rates and low latencies was required. Since tuning periods will be needed in subsequent deployments, sensor costs must be low to achieve the goal of a ubiquitous deployment.

No currently available commercial system fulfilled these requirements. Therefore, custom indoor air pollutant nodes (see Figure 3) were developed in order to meet the requirements. Sensors were selected based on their evaluation in the literature. For particulate matter measurements, we selected the Sensirion SPS30 sensor (Sensirion AG, Switzerland, Stäfa) based on its evaluation in previous studies [49]. Ref. [49] ascertained a very strong correlation with the reference instrument for fine particulate matter [49]. The Sensirion SPS30 utilizes the optical particles counter measurement principle, which has been shown to have good accuracy in measuring particulate matter of varying diameters [27].

For VOC measurements, we chose the Sensirion SGP30 (Sensirion AG, Switzerland, Stäfa) sensor. The sensor employs a metal oxide sensing (MOS) element, which is able to detect a wide range of volatile organic compounds through changes in the material's resistance due to chemical reactions with the pollutants. However, due to their broad sensitivity, it is not possible to identify the pollutant concentrations of individual VOCs, which means the output value of these sensors is qualitative. However, ref. [49] evaluated a range of VOC MOS sensors under different pollution events and, in the case of the Sensirion SGP30, performed well compared to reference instruments, thus making it viable for a qualitative evaluation of VOC pollution.

For CO_2 measurements, we selected the Sensirion SCD30 (Sensirion AG, Switzerland, Stäfa) due to its proven accuracy [27]. This sensor uses the optical NDIR measurement principle, which is the common standard in accurately measuring CO_2 concentration [50]. The sensors were connected to a microcontroller, which performs continuous measurements in the defined interval, automatic recalibration, and upload the data to a central database via WiFi connectivity.

Figure 3. Custom-built IAP sensor node (Own representation).

Since multiple indoor air pollutant nodes would be deployed, it was important to reduce sensor bias. Therefore, a cross-calibration scheme was introduced in this study. Cross-calibration is a method used to reduce sensor bias and improve accuracy by comparing the readings of individual sensors to a chosen reference sensor. In our case, one sensor was selected as the reference, and all other sensors were calibrated to perform like the reference sensor. This approach ensures consistency among the sensor readings.

The cross-calibration procedure was conducted over a 24 h period, during which a wide range of environmental conditions were introduced to test sensor response over the entire measurement range. Based on the gathered data, calibration curves for each individual sensor are generated using regression analysis. By applying these calibration curves to the input data of the latter measurements, we could reduce the influence of sensor biases.

Table 2 gives an overview of the used measurement equipment.

Table 2. Overview of the measurement equipment and sensors used.

Property	Sensirion SPS30	Sensirion SGP30	Sensirion SCD30
Parameter	Particulate Matter	Volatile Organic Compounds (VOC)	Carbon Dioxide (CO_2)
Measurement Principle	Optical Particle Counter	Metal Oxide Sensing (MOS)	Optical NDIR
Evaluation Source	[27,49]	[49]	[27,50]
Measurement Interval	10 s	10 s	10 s
Use in Literature	[51,52]	[53,54]	[55,56]

Measurements were taken in a high-rise office building in the center of Munich with 23 stories and 130,000 m^2 floor area that accommodates about 2500 employees. The building is supplied with heating and cooling through thermally activated ceilings (concrete core activation) supplied by groundwater heat pumps. A central mechanical ventilation system supplies the building with fresh air introduced into the room through induction units and extracted through exhaust outlets in the center of the zones. The ventilation system is not designed to supply heating or cooling energy. The ventilation operates at a constant schedule of 1.6 air changes per hour between 5.15 am and 8 pm. In addition to the mechanical ventilation systems, rooms in the lower stories also have operable windows. All rooms have radiation-controlled shading systems that can be overridden by the occupants. The building is in close proximity to much-frequented roads and railway tracks.

The examined office (Office 1) is located on the third floor of the building. It has two external façades, which are orientated toward the northwest and southeast. The room provides workplaces for about thirty-five employees and features operable windows. Measurements were taken in Office 1 from June 2021 to December 2022 with three independent indoor air pollutant nodes. The placement of the IAP nodes is in accordance with the guidelines for monitoring indoor air pollutants of the United States Environmental Protection Agency (EPA):

- Installation of the nodes in the breathing zone (1.10 m height)
- More than 0.5 m away from walls, corners, and windows
- More than 1 m away from local pollutant sources and occupants
- Not in front or below air supply units
- Not exposed to direct sunlight

The floorpolan of the office as well as the sensor node setup is shown in Figure 4.

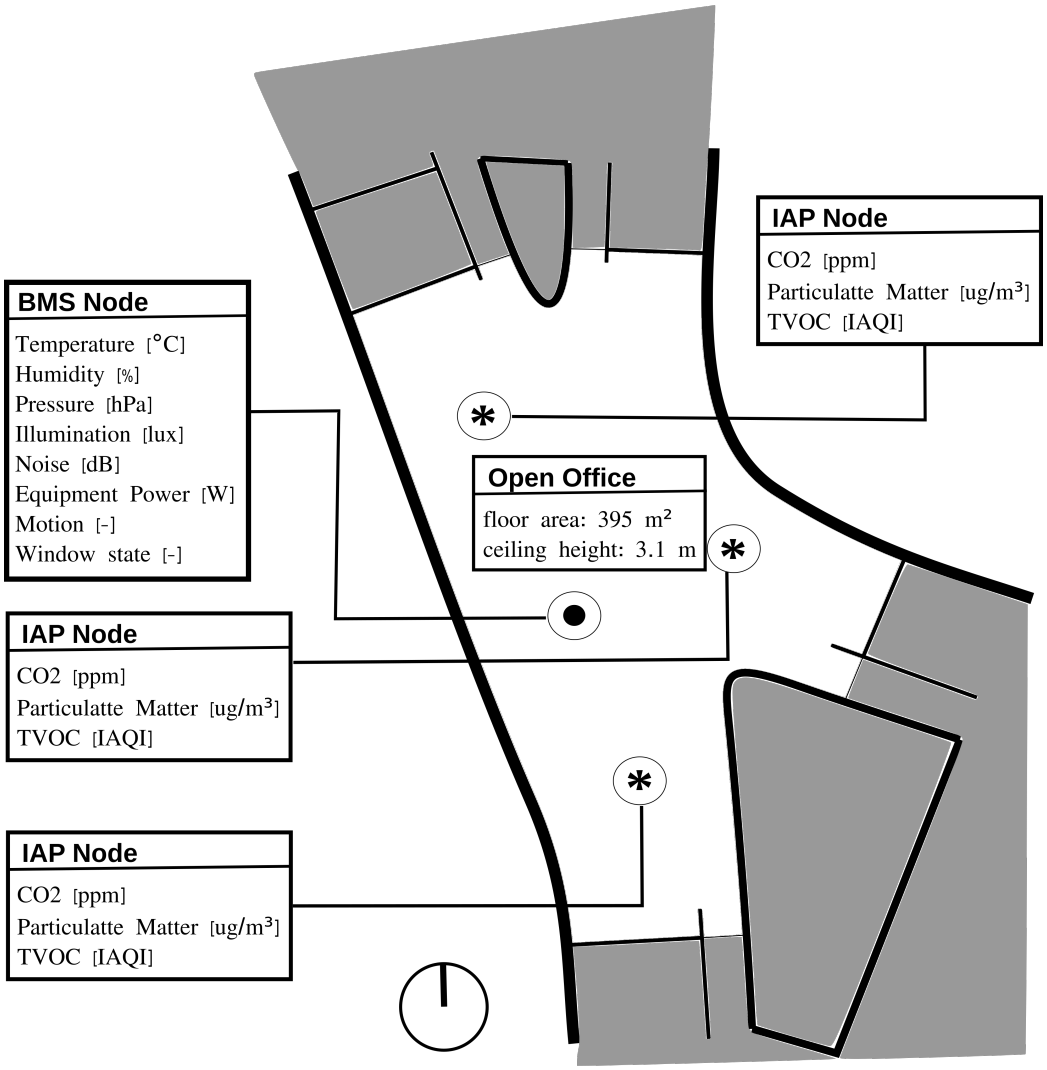

Figure 4. Floorplan illustrating sensor placement and room layout of Office 1 (Own representation).

5.2. Preprocessing

The following section outlines the steps that were taken to bring the raw datasets into a form that can be used as input for a machine-learning model. These include filtering or selecting relevant data, handling missing or corrupted values, normalizing the data, and splitting the data into training, validation, and test sets.

The initial data preparation involved extracting measurement data from IAP and the BMS node from the database and transforming the data from a long format to a wide format. This data was then loaded into a Pandas data frame for further processing. Pandas is a widely used library in Python for data analysis. It provides a structure for storing and manipulating the data in preparation for machine learning tasks.

The available measurement data was enriched by adding contextual and outdoor environmental data. For contextual data, date and time tags were added, with hours and days encoded as continuous sinus. Workdays, weekends, holidays, and seasons were added as boolean tags. Furthermore, information on the HVAC operation schedule, room size, and number of occupants were integrated into the dataset.

Additionally, outdoor environmental data from a local meteorological station was added to the dataset. The outdoor environmental data encompasses air temperature, ground temperature, dew point temperature, global and diffuse radiation, humidity, illumination, air pressure, precipitation, sunlight hours, wind- direction and -speed, as well

as outdoor particulate matter concentration. The meteorological station is at a distance of 5 km.

We noted that measurements after power cycling the nodes, e.g., after a power outage, showed elevated values for temperature and humidity for a short timespan after. In order to avoid model bias, measurements up to 15 min after a power cycle were excluded from the data set. Furthermore, random measurement fluctuations due to sensor inaccuracies were removed programmatically from the data set by smoothing measurements.

The final pre-processing steps involved optimizing the dataset for machine learning. We resampled the data to a one-minute frequency, a balance between attaining high accuracy, capturing brief temporal fluctuations, and ensuring smooth, even data. Missing data up to 15 min was input due to transient sensor response post power cycling, which we found to normalize after this interval. Overall, the input data amounted to less than 16 h for the whole measurement period. Finally, the datasets were balanced and normalized using a min-max scaler for each feature.

5.3. LSTM Setup and Training Protocol

We opted for a deep learning approach utilizing a recurrent neural network architecture, specifically an LSTM with two hidden layers. Data were fed into the LSTM as a three-dimensional input tensor, with the first dimension representing the length of the input variables (temperature, humidity, etc.), the second dimension being the lookback period (number of past timesteps), and the third dimension representing the batch size, which indicates the number of input sequences processed concurrently during training and inference. The learning rate, batch size, lookback period, and the number of neurons in the two hidden layers were determined through hyperparameter optimization.

The model's hyperparameters were optimized using Bayesian optimization, a method that uses a Gaussian process objective function and utilizes probabilistic reasoning to optimize the model's hyperparameters with the goal of minimizing the model error. An early stopping function was implemented to prevent model overfitting by monitoring the validation loss and terminating model training if the validation loss did not improve for five consecutive runs. The overall training of the LSTM took 28 min on a GPU. An overview of the model input is provided in the Appendix A in Table A1, and the model output is shown in Table A2.

Data collected from Office 1 were used to train the machine learning model. To ensure that the model could generalize and predict indoor air pollutant concentrations, a cross-validation scheme was employed. This process involved reserving 25% of the data for testing purposes and using the remaining data to train the models. This training data were further divided into a training set (75%) and a validation set (25%), with the latter being employed to trigger the early stopping algorithm to prevent overfitting, determine the best epoch, and perform hyperparameter optimization.

5.4. Evaluation

Model predictions were evaluated using the set-aside testing dataset, employing the R^2, mean absolute error (MAE) and root mean squared error (RMSE) metrics for quantification. Metrics were calculated individually for each model, pollutant, and room. We selected the MAE, RMSE, and coefficient of determination (R^2) metrics to assess model performance, as they are widely used in model performance evaluation [57]. MAE and RMSE are not dimensionless and are expressed in the units of the evaluated target. The MAE metric output represents the mean absolute difference between predicted and true values for all tested timesteps. Due to its quadratic component in the RMSE calculation, larger errors are weighted more heavily than smaller ones [57]. Consequently, MAE provides a good indication of the overall error in target units, while RMSE indicates the number of high deviations. R^2 is a dimensionless metric that measures the proportion of the total variance in the dependent variable that is predictable from the independent variables. Smaller

values for both RMSE and MAE signify a better fit, while a higher value for R^2 indicates a more accurate fit.

$$RMSE(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}} \quad (1)$$

$$MAE(y, \hat{y}) = \frac{\sum_{i=0}^{N-1}|y_i - \hat{y}_i|}{N} \quad (2)$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{N-1}(y_i - \bar{y})^2}. \quad (3)$$

5.5. Transferability Testing

To evaluate the model's ability to predict indoor air pollutant concentrations in other rooms and environments, the trained and assessed model was transferred to an unseen office room (Office 2) in the same building with a different layout, occupancy patterns, density, and orientation.

Office 2 is located on the third floor of the same building as the previous room. It has one external façade, which is oriented towards the east. The room accommodates ten employees and features operable windows. Measurements were taken in Office 2 in March 2023 using one IAP node. The same BMS data points used in the training of the LSTM model in Office 1 are available in Office 2. The outdoor meteorological and pollution data were retrieved from the same source, as both rooms are located in the same building.

The locations of the nodes and the room layout are depicted in Figure 5.

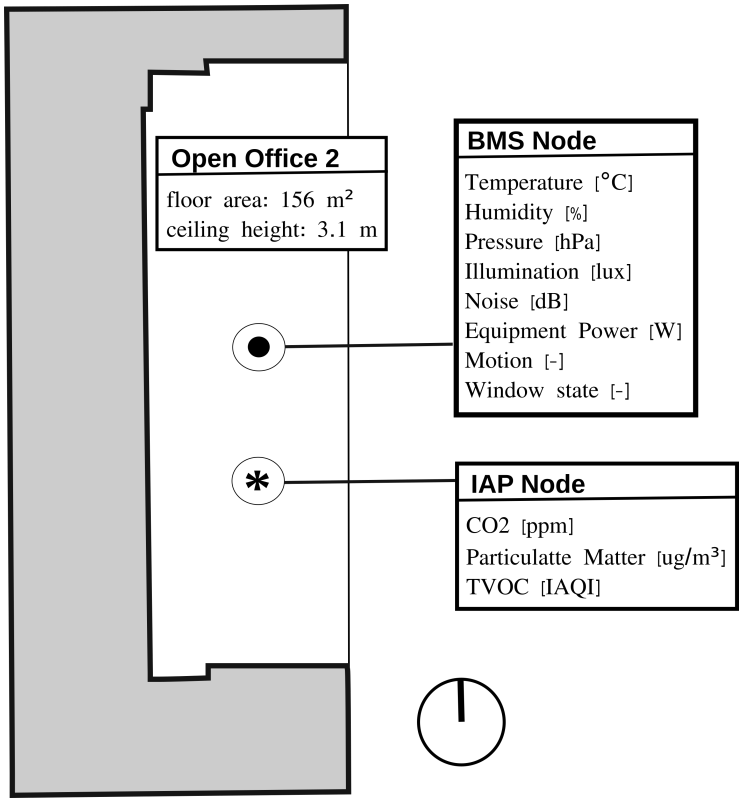

Figure 5. Floorplan illustrating sensor placement and room layout of Office 2 (Own representation).

The measurements of the IAP nodes were solely used for evaluation in Office 2. The trained model was used as is.

The model inputs, as specified in Table A1, were provided by the BMS-node as well as outdoor and metadata. The model then predicts the indoor air pollutant concentrations

for each timestep (1 min). In the evaluation, the predicted values were then compared to the actual measurements of the IAP nodes for March 2023. As previously, three prediction metrics were calculated: MAE, RMSE, and R^2.

6. Results

In this section, we present and discuss the results of our machine-learning model. Section 6.1 presents the results of the model training and its evaluation metrics. Section 6.2 reports the results of transferring the model to Office 2.

6.1. Model Evaluation

Figure 6 displays the predictions of the trained LSTM model for the testing set in Office 1 (yellow) and the measured truth (blue) for each indoor air pollutant. The evaluation metrics are calculated individually for each pollutant and shown in the top-left corner of each plot. The testing was conducted for three months, from March 2022 to May 2022. A visual assessment of the time series plots reveals a high correlation between the truth and prediction. The most significant deviations between truth and prediction are identified for CO_2 predictions. In the case of CO_2, the model tends to slightly overestimate the CO_2 concentration during low concentration periods, whereas high concentration events show a closer fit. However, the model occasionally predicts pollutant peaks incorrectly during low concentration periods and vice versa. In the case of CO_2 predictions, they appear to be more accurate during the second half of the testing period. For particulate matter, the visual assessment shows an excellent fit between prediction and truth. All peaks are identified correctly. The time series plot demonstrates a slight underestimation of peaks and high pollution events by the prediction compared to the truth. In the case of VOC, the visual assessment of the time series plots reveals an excellent fit between prediction and truth. The model can detect all concentration peaks, even though VOC concentration is highly dynamic. However, a slight underestimation of pollutant peaks can be observed in the time series, especially during the first half of the testing period.

Table 3 summarizes the evaluation metrics (R^2, RMSE, MAE) for each pollutant. Overall, the model exhibits a low error for all pollutants, as demonstrated by the MAE and RMSE performance metrics. In the case of CO_2, the mean absolute error amounts to 15.4 ppm for the testing period, while a slightly increased RMSE value of 20.2 ppm indicates that no outliers significantly impact the model's predictions. The CO_2 measurements ranged from 380 to 560 ppm during the measurement period. For particulate matter, the errors amount to 0.3 and 0.5 $\mu g/m^3$ for MAE and RMSE, respectively, indicating consistently low error rates without outliers. The measurements ranged from 0 to 13 $\mu g/m^3$ during the measurement period. In the case of volatile organic compounds, MAE and RMSE errors amounted to 20.1 IAQI and 31.4 IAQI, respectively, demonstrating low error rates without major deviations. The measurements for VOC ranged from 0 to 450 IAQI. The R^2 performance metric is a statistical measure representing the goodness of fit of the LSTM model and indicates the percentage of variance in the truth data that can be explained by the LSTM model. In the case of CO_2, an R^2 value of 0.47 indicates that the model explains a substantial part of the variance in CO_2 concentration and has a reasonably good fit, providing meaningful predictions. For PM, a high R^2 of 0.88 was identified, indicating that the model explains a significant percentage of the variability in particulate matter measurements. Furthermore, it shows a very good fit and indicates that the model is highly predictive of PM. In the case of VOC, a high R^2 of 0.87 was identified, which shows that a significant percentage of VOC volatility is explained by the LSTM virtual sensing model. Furthermore, the R^2 indicates a very good fit for VOC and that the model is highly predictive.

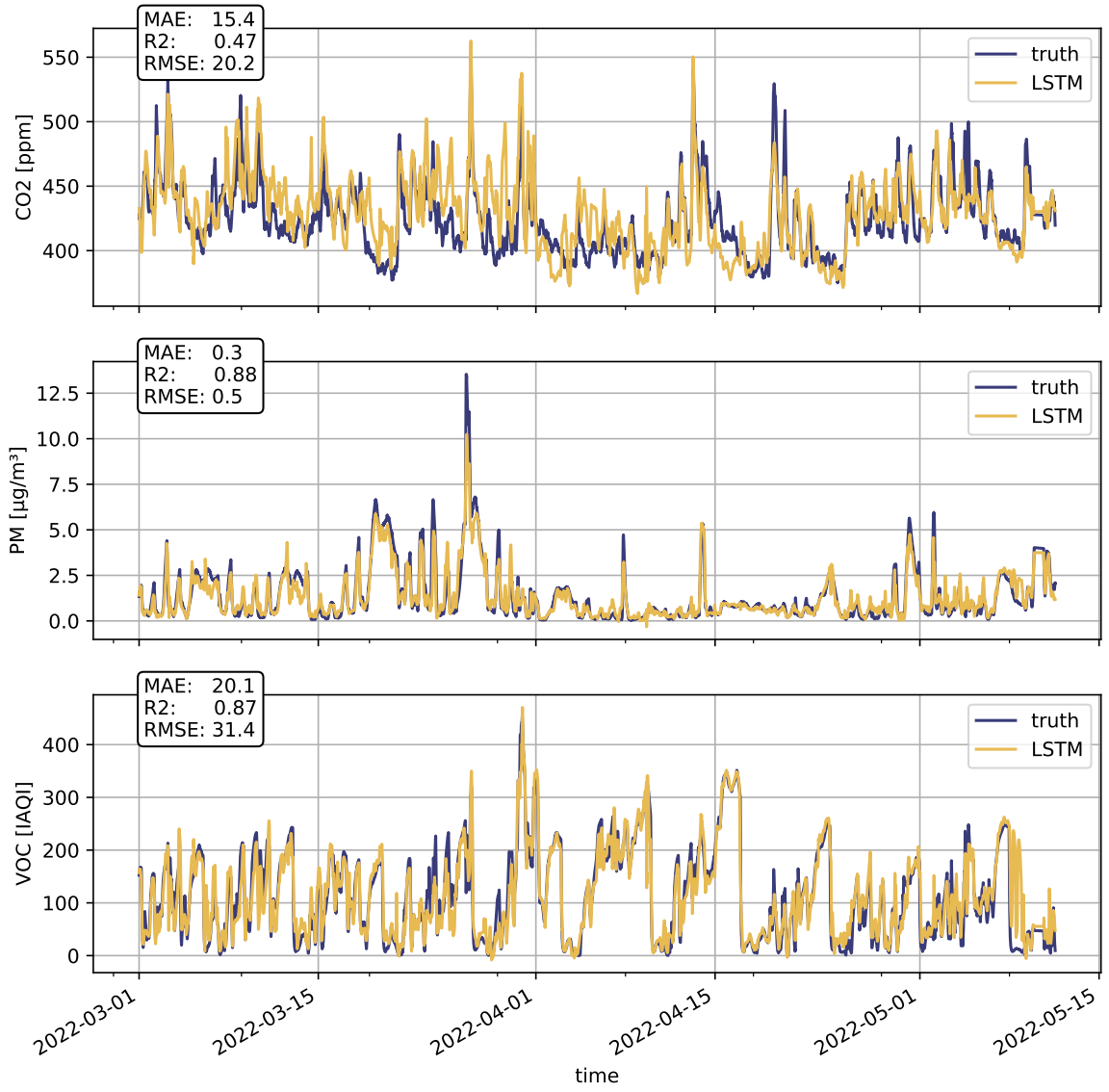

Figure 6. Comparison of virtual indoor air pollutant sensors (yellow) and physical indoor air pollutant sensors (blue) with overlayed evaluation metrics for Indoor Air Pollutants: VOC (**bottom**), PM (**middle**), and CO_2 (**top**) for Office 1 (Own representation).

Table 3. Evaluation metrics LSTM virtual sensing model in Office 1.

Pollutant	MAE	R^2	RMSE
CO_2	15.4	0.47	20.2
$PM_{2.5}$	0.3	0.88	0.5
VOC	20.1	0.87	31.4

The model successfully identified all pollutant peaks during the testing period, with the only error being a slight underestimation of peak concentrations. For CO_2, a less ideal but still satisfactory prediction result was achieved. This led to minor errors and a less accurate representation of the variability in actual concentrations, resulting in some erroneous predictions, such as misidentified pollutant peaks during the testing period. Nevertheless, the predictions yielded a mean absolute error within the range of measurement inaccuracies for most sensors.

The performance metrics of MAE = 15.4 ppm, RMSE = 20.2 ppm, and R^2 = 0.47 for CO_2 showed very low errors with insignificant outliers. The R^2 value indicated a reasonably good

fit and predictive capability. For PM, the metrics MAE = 0.3 µg/m^3, RMSE = 0.5 µg/m^3, and R^2 = 0.88 signified a strong prediction capability with minimal errors and an excellent fit. Similar results were observed for VOC, with MAE = 20.1 IAQI, RMSE = 31.4 IAQI, and R^2 = 0.87. Overall, the LSTM model demonstrated strong performance in predicting indoor air pollutant concentrations, with some room for improvement in CO_2 predictions. Based on the findings from this study, the LSTM model shows promise to potentially replace physical sensors, contributing to more cost-effective and efficient air quality monitoring solutions.

6.2. Transferability Evaluation

Figure 7 displays the predictions of the trained LSTM model (yellow) for Office 2, as well as the measured actual values (blue) for each indoor air pollutant. The test took place in March 2023.

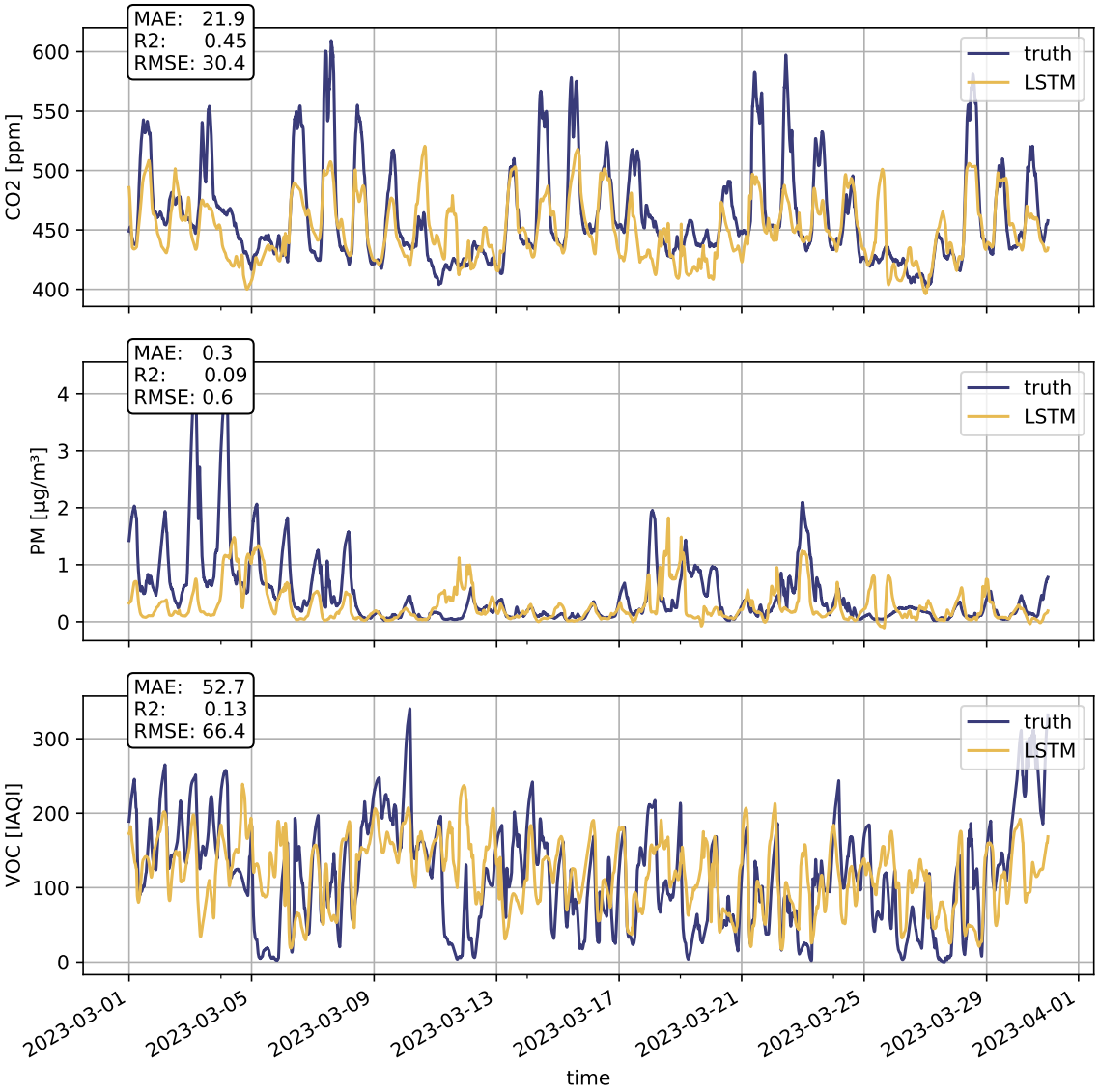

Figure 7. Comparison of virtual indoor air pollutant sensors (yellow) and physical indoor air pollutant sensors (blue) with overlayed evaluation metrics for Indoor Air Pollutants: VOC (**bottom**), PM (**middle**), and CO_2 (**top**) for Office 2 (Own representation).

Visual assessment of the time series plots revealed a correlation between actual values and predictions for all pollutants, albeit with varying degrees of fit. The highest correlation

between actual and predicted values was observed for CO_2 predictions. The prediction model successfully identified pollutant peaks, albeit with underestimation. During low pollutant events, such as weekends or nights, the model results were less smooth and tended to overestimate variability in pollutant concentrations. Occasionally, the model predicted pollutant peaks under unpolluted conditions.

For particulate matter, the visual assessment showed a general fit between the magnitudes of predicted and actual concentrations. However, the prediction failed to detect some peaks and underestimated all others. In some cases, the prediction exhibited a phase shift, resulting in delayed identification of rising concentrations.

For VOC, a visual assessment of the time series plots revealed that the model could identify some concentration peaks. However, the model frequently and erroneously detected pollutant peaks when none were present.

Table 3 summarizes the evaluation metrics (R^2, RMSE, MAE) for each pollutant. Overall, the model exhibited very low errors for all pollutants, as evidenced by the MAE and RMSE performance metrics. For CO_2, the mean absolute error was 21.9 ppm during the testing period, while a slightly increased RMSE value of 30.4 ppm indicated that no outliers affected the model's predictions. CO_2 measurements ranged from 420 ppm to 610 ppm during the measurement period.

For particulate matter, errors amounted to 0.3 and 0.6 $\mu g/m^3$ for MAE and RMSE, respectively, indicating consistently low error rates without outliers. Measurements ranged from 0 to 4 $\mu g/m^3$ during the measurement period. For volatile organic compounds, MAE and RMSE errors were 52.7 IAQI and 66.4 IAQI, respectively, demonstrating very low error rates without significant deviations. Measurements for VOC ranged from 0 to 330 IAQI.

For CO_2, an R^2 value of 0.45 indicated that the model accounted for a substantial portion of the variability in CO_2 concentrations, exhibited a reasonably good fit, and provided meaningful predictions. For PM, a low R^2 of 0.09 suggested that the model explained a smaller percentage of the variability in particulate matter measurements, demonstrated a less accurate fit, and was less predictive. For VOC, a low R^2 of 0.13 indicated that the model explained a smaller percentage of VOC variability and was less accurate and less predictive. The evaluation metrics are summarized in Table 4.

Table 4. Evaluation metrics LSTM virtual sensing model transfer in Office 2.

Pollutant	MAE	R^2	RMSE
CO_2	21.9	0.45	30.4
$PM_{2.5}$	0.3	0.09	0.6
VOC	52.7	0.13	66.4

The LSTM-based virtual indoor air pollutant sensor was tested for Office 2 using the testing dataset for March 2023. The evaluation results indicated varying degrees of correlation between the actual and predicted pollutant concentrations. For CO_2, the model successfully identified pollutant peaks, albeit underestimated, and exhibited an MAE of 21.9 ppm, RMSE of 30.4 ppm, and R^2 of 0.45, indicating a reasonably good fit and predictive capabilities. For particulate matter and volatile organic compounds (VOC), the model showed less accurate predictions in terms of R^2 values; however, the MAE and RMSE errors remained low. For PM, despite the model's failure to detect some pollutant peaks and a low R^2 of 0.09, the MAE and RMSE were consistently low at 0.3 $\mu g/m^3$ and 0.6 $\mu g/m^3$, respectively, indicating a relatively low error rate without significant outliers. Similarly, for VOC, the model erroneously detected pollutant peaks in some cases and showed a low R^2 of 0.13. Yet, the MAE and RMSE remained low at 52.7 IAQI and 66.4 IAQI, respectively, demonstrating low error rates without major deviations. In conclusion, the LSTM model exhibits varying performance in predicting indoor air pollutant concentrations for Office 2, with better results for CO_2 predictions and low error rates in terms of MAE and RMSE for PM and VOC predictions. However, there is room for improvement in capturing the variability of PM and VOC concentrations, as indicated by the low R^2 values.

7. Discussion

The findings of this study indicate that machine learning models, particularly LSTM networks, are effective in predicting indoor air pollutants, especially particulate matter, and VOC, as demonstrated by the low error rates achieved in the testing set of Office 1. The testing results from Office 1 indicate certain limitations of the virtual sensing model in capturing the full range of variability in CO_2 concentrations. This limitation may be attributed to the model's reduced precision in predicting occupancy and occupant count. Skoen et al. [38] previously noted similar findings when applying Multi-Layer Perceptron (MLP) models for virtual sensing of CO_2, notably, even though the R^2 values from Skoen et al. (0.39) closely match the 0.47 achieved in this study, and a significantly lower Root Mean Square Error (RMSE) of 31.4 was obtained in this study, compared to Skoen et al.'s 122.85 [38]. This suggests that, despite the model's inability to capture full variability with Long Short-Term Memory (LSTM), the error margins remained relatively low, particularly when compared to other models. When the pre-trained models were applied to other rooms of identical typologies, they still exhibited predictive capacity. However, these models demonstrated a decreased ability to explain the variability of pollutant concentrations as well as increased errors. This suggests a limitation in model transferability to different rooms, with a significant decline in the model's predictive capability noticed, particularly in terms of capturing the ground truth variability. It is postulated that this decrease in performance is attributable to the limitations of the training dataset, which was exclusively trained in Office 1. Given that occupancy and numerous other dynamic factors influence indoor air pollutants, indoor environments can significantly differ from each other. They may also display vastly different pollutant dynamics, as previously demonstrated by Szigeti et al. [8]. It is anticipated that the model's performance will be reduced when applied to rooms in other buildings or those belonging to different typologies, as these environments may present conditions not encountered during model training. Consequently, it is crucial to enhance the transferability and performance of the virtual sensing LSTM model by generating larger and more diverse datasets.

While current results do not yet allow for a complete replacement of physical sensors with LSTM models, the promising predictions of IAP concentrations in the training room, along with the successful prediction of CO_2 levels in a separate office, demonstrate potential. The general application of this model is not yet feasible, but, given more diverse data, the outlook for the full replacement of physical sensors with such models becomes more attainable.

The use of machine learning techniques to create virtual sensors for monitoring indoor air pollutants has the potential to provide real-time data and improve building operations. Further research and development may lead to the use of virtual sensors for wider application in building environments, potentially allowing for the optimization of mechanical ventilation systems and operable window usage. It is important to continue exploring the potential of virtual indoor air pollutant sensors as a tool for improving indoor air quality and the overall comfort and health of building occupants.

Further research is needed in expanding the training data for LSTM models for virtual sensing of indoor air pollutants and testing their generalizability across various typologies and buildings in different climate zones. Additionally, future studies could investigate the integration of these models into Heating, Ventilation, and Air Conditioning (HVAC) systems and evaluate their performance when only a fraction of the given input data is available. This would help advance the practical implementation of virtual sensing in real-world scenarios and contribute to the field of indoor air quality monitoring.

8. Conclusions

This study demonstrates the potential of machine learning models, specifically LSTM networks, to accurately predict indoor air pollutant concentrations in a range of environments. By using a large dataset with several years of accumulated data, we were able to build a virtual indoor air pollutant sensor that exhibited strong performance in

predicting indoor air pollutant concentrations for the room in which it was trained. The evaluation results indicated a very high correlation between the actual and predicted pollutant concentrations for particulate matter and VOC, with performance metrics such as MAE = 0.3 $\mu g/m^3$, RMSE = 0.5 $\mu g/m^3$, and R^2 = 0.88 for PM; and MAE = 20.1 IAQI, RMSE = 31.4 IAQI, and R^2 = 0.87 for VOC. These results show that the model was able to identify most pollutant peaks during the testing period with only a slight underestimation of peak concentrations. For CO_2, the model achieved less ideal but reasonable prediction results. The performance metrics of MAE = 15.4 ppm, RMSE = 20.2 ppm, and R^2 = 0.47 for CO_2 indicated very low errors with insignificant outliers, and the R^2 value suggested a reasonably good fit and predictive capabilities. However, the model was not able to explain the variability of the actual concentrations and showed some erroneous predictions, such as misidentified pollutant peaks during the testing period. Nevertheless, the predictions resulted in a mean absolute error within the range of the measurement inaccuracy of most sensors. When transferring the model to another room, the LSTM model demonstrated varying performance, with better results for CO_2 predictions and low error rates in terms of MAE and RMSE for PM and VOC predictions. Specifically, the CO_2 predictions exhibited a mean absolute error of 21.9 ppm, RMSE of 30.4 ppm, and R^2 of 0.45, indicating a reasonably good fit and predictive capabilities. However, there is room for improvement in capturing the variability of PM and VOC concentrations, as indicated by the low R^2 values of 0.09 for PM and 0.13 for VOC. Despite these challenges, the LSTM model shows its potential in generalizing its ability to predict indoor air pollutant concentrations in different rooms. To enhance the model's performance when transferring to other rooms, further research and optimization could focus on refining the LSTM architecture, incorporating additional features such as building materials, type of air distribution, and the distance of the nodes from vents and windows, or exploring other machine learning techniques to improve the model's ability to capture the variability of different pollutants. In summary, the LSTM-based virtual indoor air pollutant sensor presents a promising approach to monitoring air quality in indoor environments. With further refinement and optimization, this model could potentially replace physical sensors, contributing to more cost-effective and efficient air quality monitoring solutions. Ultimately, the development and deployment of accurate virtual sensing models can play a crucial role in addressing indoor air pollution, leading to improved public health and well-being.

Author Contributions: Conceptualization, M.G.; methodology, M.G.; software, M.G..; validation, M.G.; formal analysis, M.G.; investigation, M.G.; resources, M.G.; data curation, M.G.; writing—original draft preparation, M.G.; writing—review and editing, M.G.; visualization, M.G.; supervision, T.A.; project administration, M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data from the model evaluation is available upon request. The data used for creating the LSTM model, including the training, testing, and validation datasets, contain sensitive information and cannot be shared to ensure data privacy.

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

Abbreviations

The following abbreviations are used in this manuscript:

BMS	Building management system
IAP	Indoor air pollutants
IAQ	Indoor air quality
IoT	Internet of things
MAE	Mean absolute error
RMSE	Root mean squared error
LSTM	long short-term memory network
VOC	Volatile organic compounds
PM	Particulate matter
IAQI	Indoor air quality index
ppm	Parts per million
GPU	Graphics processing unit
HVAC	Heating, Ventilation, and Air Conditioning
MOS	Metal oxide sensing
OPC	Optial particle counter
NDIR	Non-dispersive infrared
SVM	Support vector machine
MLP	Multi-layer perceptron
ETL	Extract transfer load

Appendix A

Table A1. Input Features for LSTM Model.

Feature	Description	Dimension (Normalized)	Group
month_sin	Continuous sinusoidal encoding of month	0–1	Meta
hr_sin	Continuous sinusoidal encoding of hour	0–1	Meta
day_sin	Continuous sinusoidal encoding of day	0–1	Meta
workday	Boolean tag for workdays	0, 1	Meta
weekend	Boolean tag for weekends	0, 1	Meta
holiday	Boolean tag for holidays	0, 1	Meta
season	Boolean tags for each season	0, 1	Meta
hvac	Boolean tag for HVAC operation	0, 1	Indoor
room_size	Size of the room	0–1	Meta
occupants	Occupant density	0–1	Meta
temp	Outdoor air temperature	0–1	Outdoor
ground_temp	Outdoor ground temperature	0–1	Outdoor
dew_point_temp	Outdoor dew point temperature	0–1	Outdoor
global_rad	Outdoor global radiation	0–1	Outdoor
diffuse_rad	Outdoor diffuse radiation	0–1	Outdoor
humidity	Outdoor humidity	0–1	Outdoor
illumination	Outdoor illumination	0–1	Outdoor
air_pressure	Outdoor air pressure	0–1	Outdoor

Table A1. *Cont.*

Feature	Description	Dimension (Normalized)	Group
precipitation	Outdoor precipitation	0–1	Outdoor
wind_dir	Outdoor wind direction	0–1	Outdoor
wind_speed	Outdoor wind speed	0–1	Outdoor
particulate_matter	Outdoor particulate matter concentration	0–1	Outdoor
indoor_temp	Indoor air temperature	0–1	Indoor
indoor_humidity	Indoor humidity	0–1	Indoor
indoor_air_pressure	Indoor air pressure	0–1	Indoor
indoor_illum	Indoor illumination	0–1	Indoor
noise_level	Indoor noise level	0–1	Indoor
window_state	State of window (open/closed)	0, 1	Indoor
power_consumption	Power consumption of equipment	0–1	Indoor

Table A2. LSTM model output.

Output	Description	Dimension (Normalized)
pm	Particulate matter concentration	0–1
co2	CO_2 concentration	0–1
voc	Volatile organic compound concentration	0–1

References

1. Tham, K.W. Indoor air quality and its effects on humans—A review of challenges and developments in the last 30 years. *Energy Build.* **2016**, *130*, 637–650. [CrossRef]
2. Hasager, F.; Bjerregaard, J.D.; Bonomaully, J.; Knap, H.; Afshari, A.; Johnson, M.S. Indoor Air Quality: Status and Standards. *Air Pollut. Sources Stat. Health Eff.* **2021**, 135–162.
3. Berglund, B.; Brunekreef, B.; Knöppe, H.; Lindvall, T.; Maroni, M.; Mølhave, L.; Skov, P. Effects of indoor air pollution on human health. *Indoor Air* **1992**, *2*, 2–25. [CrossRef]
4. Henschel, S.; Chan, G.; World Health Organization. *Health Risks of Air Pollution in Europe-HRAPIE Project: New Emerging Risks to Health from Air Pollution-Results from the Survey of Experts*; WHO: Geneva, Switzerland, 2013.
5. Soares, A.G.O.G.J. *Air Quality in Europe—2020 Report*; Technical Report; European Environment Agency: Copenhagen, Denmark, 2020.
6. 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]
7. World Health Organization. *WHO Guidelines for Indoor Air Quality: Selected Pollutants*; World Health Organization. Regional Office for Europe: Geneva, Switzerland, 2010.
8. Szigeti, T.; Dunster, C.; Cattaneo, A.; Spinazzè, A.; Mandin, C.; Le Ponner, E.; de Oliveira Fernandes, E.; Ventura, G.; Saraga, D.E.; Sakellaris, I.A.; et al. Spatial and temporal variation of particulate matter characteristics within office buildings—The OFFICAIR study. *Sci. Total Environ.* **2017**, *587*, 59–67. [CrossRef]
9. Li, J.; Li, H.; Ma, Y.; Wang, Y.; Abokifa, A.A.; Lu, C.; Biswas, P. Spatiotemporal distribution of indoor particulate matter concentration with a low-cost sensor network. *Build. Environ.* **2018**, *127*, 138–147. [CrossRef]
10. Sahu, V.; Gurjar, B.R. Spatio-temporal variations of indoor air quality in a university library. *Int. J. Environ. Health Res.* **2021**, *31*, 475–490. [CrossRef]
11. Zhang, H.; Srinivasan, R.; Ganesan, V. Low Cost, Multi-Pollutant Sensing System Using Raspberry Pi for Indoor Air Quality Monitoring. *Sustainability* **2021**, *13*, 370. [CrossRef]
12. Kim, J.; Kong, M.; Hong, T.; Jeong, K.; Lee, M. The effects of filters for an intelligent air pollutant control system considering natural ventilation and the occupants. *Sci. Total Environ.* **2019**, *657*, 410–419. [CrossRef]
13. Saraga, D.; Maggos, T.; Sadoun, E.; Fthenou, E.; Hassan, H.; Tsiouri, V.; Karavoltsos, S.; Sakellari, A.; Vasilakos, C.; Kakosimos, K. Chemical characterization of indoor and outdoor particulate matter (PM2. 5, PM10) in Doha, Qatar. *Aerosol Air Qual. Res.* **2017**, *17*, 1156–1168.

14. Irga, P.; Torpy, F. Indoor air pollutants in occupational buildings in a sub-tropical climate: Comparison among ventilation types. *Build. Environ.* **2016**, *98*, 190–199. [CrossRef]
15. Montgomery, J.F.; Storey, S.; Bartlett, K. Comparison of the indoor air quality in an office operating with natural or mechanical ventilation using short-term intensive pollutant monitoring. *Indoor Built Environ.* **2015**, *24*, 777–787. [CrossRef]
16. Ha, Q.P.; Metia, S.; Phung, M.D. Sensing data fusion for enhanced indoor air quality monitoring. *IEEE Sens. J.* **2020**, *20*, 4430–4441. [CrossRef]
17. Kang, J.; Hwang, K.I. A comprehensive real-time indoor air-quality level indicator. *Sustainability* **2016**, *8*, 881. [CrossRef]
18. Mendoza, D.; Benney, T.M.; Boll, S. Long-term analysis of the relationships between indoor and outdoor fine particulate pollution: A case study using research grade sensors. *Sci. Total. Environ.* **2021**, *776*, 145778. [CrossRef]
19. Tiele, A.; Esfahani, S.; Covington, J. Design and development of a low-cost, portable monitoring device for indoor environment quality. *J. Sens.* **2018**, *2018*. [CrossRef]
20. Spinazzè, A.; Campagnolo, D.; Cattaneo, A.; Urso, P.; Sakellaris, I.A.; Saraga, D.E.; Mandin, C.; Canha, N.; Mabilia, R.; Perreca, E.; et al. Indoor gaseous air pollutants determinants in office buildings—The OFFICAIR project. *Indoor Air* **2020**, *30*, 76–87. [CrossRef]
21. Saini, J.; Dutta, M.; Marques, G. Indoor air quality prediction using optimizers: A comparative study. *J. Intell. Fuzzy Syst.* **2020**, *39*, 7053–7069. [CrossRef]
22. Challoner, A.; Gill, L. Indoor/outdoor air pollution relationships in ten commercial buildings: PM2. 5 and NO_2. *Build. Environ.* **2014**, *80*, 159–173. [CrossRef]
23. Challoner, A.; Pilla, F.; Gill, L. Prediction of indoor air exposure from outdoor air quality using an artificial neural network model for inner city commercial buildings. *Int. J. Environ. Res. Public Health* **2015**, *12*, 15233–15253. [CrossRef]
24. Ahn, J.; Shin, D.; Kim, K.; Yang, J. Indoor air quality analysis using deep learning with sensor data. *Sensors* **2017**, *17*, 2476. [CrossRef] [PubMed]
25. Ma, N.; Aviv, D.; Guo, H.; Braham, W.W. Measuring the right factors: A review of variables and models for thermal comfort and indoor air quality. *Renew. Sustain. Energy Rev.* **2021**, *135*, 110436. [CrossRef]
26. Kolarik, J.; Lyng, N.L.; Bossi, R.; Witterseh, T.; Smith, K.M.; Wargocki, P. 3.6 Response of commercially available Metal Oxide Semiconductor Sensors under air polluting activities typical for residences. *Indoor Air Qual. Des. Control. -Low-Energy Resid. Build. (Ebc Annex. 68)* **2020**, 47.
27. Frederickson, L.B.; Petersen-Sonn, E.A.; Shen, Y.; Hertel, O.; Hong, Y.; Schmidt, J.; Johnson, M.S. Low-Cost Sensors for Indoor and Outdoor Pollution. *Air Pollut. Sources Stat. Health Eff.* **2021**, 423–453.
28. Alhasa, K.M.; Mohd Nadzir, M.S.; Olalekan, P.; Latif, M.T.; Yusup, Y.; Iqbal Faruque, M.R.; Ahamad, F.; Abd. Hamid, H.H.; Aiyub, K.; Md Ali, S.H.; et al. Calibration model of a low-cost air quality sensor using an adaptive neuro-fuzzy inference system. *Sensors* **2018**, *18*, 4380. [CrossRef]
29. Topalović, D.B.; Davidović, M.D.; Jovanović, M.; Bartonova, A.; Ristovski, Z.; Jovašević-Stojanović, M. In search of an optimal in-field calibration method of low-cost gas sensors for ambient air pollutants: Comparison of linear, multilinear and artificial neural network approaches. *Atmos. Environ.* **2019**, *213*, 640–658. [CrossRef]
30. Heindel, L.; Hantschke, P.; Kästner, M. A Virtual Sensing approach for approximating nonlinear dynamical systems using LSTM networks. *PAMM* **2021**, *21*, e202100119. [CrossRef]
31. Li, H.; Yu, D.; Braun, J.E. A review of virtual sensing technology and application in building systems. *Hvac&R Res.* **2011**, *17*, 619–645.
32. Yoon, S. Virtual sensing in intelligent buildings and digitalization. *Autom. Constr.* **2022**, *143*, 104578. [CrossRef]
33. Wu, Q.; Cai, W.; Wang, X.; Chakraborty, A. Dehumidifier desiccant concentration soft-sensor for a distributed operating Liquid Desiccant Dehumidification System. *Energy Build.* **2016**, *129*, 215–226. [CrossRef]
34. Hong, Y.; Yoon, S.; Kim, Y.S.; Jang, H. System-level virtual sensing method in building energy systems using autoencoder: Under the limited sensors and operational datasets. *Appl. Energy* **2021**, *301*, 117458. [CrossRef]
35. Li, H.; Hong, T.; Sofos, M. An inverse approach to solving zone air infiltration rate and people count using indoor environmental sensor data. *Energy Build.* **2019**, *198*, 228–242. [CrossRef]
36. Alhashme, M.; Ashgriz, N. A virtual thermostat for local temperature control. *Energy Build.* **2016**, *126*, 323–339. [CrossRef]
37. Zhao, Y.; Zeiler, W.; Boxem, G.; Labeodan, T. Virtual occupancy sensors for real-time occupancy information in buildings. *Build. Environ.* **2015**, *93*, 9–20. [CrossRef]
38. Skön, J.; Johansson, M.; Raatikainen, M.; Leiviskä, K.; Kolehmainen, M. Modelling indoor air carbon dioxide (CO_2) concentration using neural network. *Methods* **2012**, *14*, 16.
39. Elbayoumi, M.; Ramli, N.A.; Yusof, N.F.F.M. Development and comparison of regression models and feedforward backpropagation neural network models to predict seasonal indoor PM2. 5–10 and PM2. 5 concentrations in naturally ventilated schools. *Atmos. Pollut. Res.* **2015**, *6*, 1013–1023. [CrossRef]
40. Khazaei, B.; Shiehbeigi, A.; Haji Molla Ali Kani, A. Modeling indoor air carbon dioxide concentration using artificial neural network. *Int. J. Environ. Sci. Technol.* **2019**, *16*, 729–736. [CrossRef]
41. Kusiak, A.; Li, M.; Zheng, H. Virtual models of indoor-air-quality sensors. *Appl. Energy* **2010**, *87*, 2087–2094. [CrossRef]

42. Gabriel, M.; Auer, T. Indoor air pollution estimation using machine learning (ANN and SVR) in smart buildings. In *BauSim Conference 2022, Proceedings of the 9th Conference of IBPSA-Germany and Austria, Weimar, Germany, 20–22 September 2022*; IBPSA-Germany and Austria: Dresden, Germany, 2022; Volume 9. [CrossRef]
43. Leidinger, M.; Sauerwald, T.; Reimringer, W.; Ventura, G.; Schütze, A. Selective detection of hazardous VOCs for indoor air quality applications using a virtual gas sensor array. *J. Sens. Sens. Syst.* **2014**, *3*, 253–263. [CrossRef]
44. Karijadi, I.; Chou, S.Y. A hybrid RF-LSTM based on CEEMDAN for improving the accuracy of building energy consumption prediction. *Energy Build.* **2022**, *259*, 111908. [CrossRef]
45. Jang, J.; Han, J.; Leigh, S.B. Prediction of heating energy consumption with operation pattern variables for non-residential buildings using LSTM networks. *Energy Build.* **2022**, *255*, 111647. [CrossRef]
46. Qolomany, B.; Al-Fuqaha, A.; Benhaddou, D.; Gupta, A. Role of deep LSTM neural networks and Wi-Fi networks in support of occupancy prediction in smart buildings. In Proceedings of the 2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Bangkok, Thailand, 18–20 December 2017; pp. 50–57.
47. Zhang, L.; Liu, P.; Zhao, L.; Wang, G.; Zhang, W.; Liu, J. Air quality predictions with a semi-supervised bidirectional LSTM neural network. *Atmos. Pollut. Res.* **2021**, *12*, 328–339. [CrossRef]
48. Bai, Y.; Zeng, B.; Li, C.; Zhang, J. An ensemble long short-term memory neural network for hourly PM2. 5 concentration forecasting. *Chemosphere* **2019**, *222*, 286–294. [CrossRef] [PubMed]
49. Demanega, I.; Mujan, I.; Singer, B.C.; Andelkovic, A.S.; Babich, F.; Licina, D. Performance assessment of low-cost environmental monitors and single sensors under variable indoor air quality and thermal conditions. *Build. Environ.* **2021**, *187*, 107415. [CrossRef]
50. Marinov, M.B.; Djermanova, N.; Ganev, B.; Nikolov, G.; Janchevska, E. Performance evaluation of low-cost carbon dioxide sensors. In Proceedings of the 2018 IEEE XXVII International Scientific Conference Electronics-ET, Sozopol, Bulgaria, 3–15 September 2018; pp. 1–4.
51. Hassani, A.; Castell, N.; Watne, Å.K.; Schneider, P. Citizen-operated mobile low-cost sensors for urban PM2. 5 monitoring: field calibration, uncertainty estimation, and application. *Sustain. Cities Soc.* **2023**, *95*, 104607. [CrossRef]
52. Kuula, J.; Friman, M.; Helin, A.; Niemi, J.V.; Aurela, M.; Timonen, H.; Saarikoski, S. Utilization of scattering and absorption-based particulate matter sensors in the environment impacted by residential wood combustion. *J. Aerosol Sci.* **2020**, *150*, 105671. [CrossRef]
53. Alonso, M.J.; Madsen, H.; Liu, P.; Jørgensen, R.B.; Jørgensen, T.B.; Christiansen, E.J.; Myrvang, O.A.; Bastien, D.; Mathisen, H.M. Evaluation of low-cost formaldehyde sensors calibration. *Build. Environ.* **2022**, *222*, 109380. [CrossRef]
54. Arsiwala, A.; Elghaish, F.; Zoher, M. Digital twin with Machine learning for predictive monitoring of CO_2 equivalent from existing buildings. *Energy Build.* **2023**, *284*, 112851. [CrossRef]
55. Trilles, S.; Juan, P.; Chaudhuri, S.; Fortea, A.B.V. Data on CO_2, temperature and air humidity records in Spanish classrooms during the reopening of schools in the COVID-19 pandemic. *Data Brief* **2021**, *39*, 107489. [CrossRef]
56. Toschke, Y.; Lusmoeller, J.; Otte, L.; Schmidt, J.; Meyer, S.; Tessmer, A.; Brockmann, C.; Ahuis, M.; Hüer, E.; Kirberger, C.; et al. Distributed LoRa based CO_2 monitoring network–A standalone open source system for contagion prevention by controlled ventilation. *HardwareX* **2022**, *11*, e00261. [CrossRef]
57. Willmott, C.J.; 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]