data

MDPI

Data Descriptor

A Comprehensive Indoor Environment Dataset from Single-Family Houses in the US

Sheik Murad Hassan Anik [1,*], Xinghua Gao [2] and Na Meng [3]

[1] Department of Computer Science, Auburn University at Montgomery, Montgomery, AL 36117, USA
[2] Myers-Lawson School of Construction, Virginia Tech, Blacksburg, VA 24061, USA
[3] Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA; nm8247@vt.edu
* Correspondence: sanik1@aum.edu

Abstract: The paper describes a dataset comprising indoor environmental factors such as temperature, humidity, air quality, and noise levels. The data were collected from 10 sensing devices installed in various locations within three single-family houses in Virginia, USA. The objective of the data collection was to study the indoor environmental conditions of the houses over time. The data were collected at a frequency of one record per minute for a year, combining to a total over 2.5 million records. The paper provides actual floor plans with sensor placements to aid researchers and practitioners in creating reliable building performance models. The techniques used to collect and verify the data are also explained in the paper. The resulting dataset can be employed to enhance models for building energy consumption, occupant behavior, predictive maintenance, and other relevant purposes.

Dataset: https://doi.org/10.17605/OSF.IO/BAEW7.

Dataset License: CC0

Keywords: indoor environment dataset; remote sensing; IoT data collection; distributed data infrastructure

Academic Editor: Jamal Jokar Arsanjani

Received: 24 January 2025
Revised: 19 February 2025
Accepted: 24 February 2025
Published: 5 March 2025

Citation: Anik, S.M.H.; Gao, X.; Meng, N. A Comprehensive Indoor Environment Dataset from Single-Family Houses in the US. *Data* **2025**, *10*, 35. https://doi.org/10.3390/data10030035

1. Summary

Data generated in a building environment can provide a comprehensive understanding of various aspects of a building, enabling informed decision-making, improving occupant experiences, and enhancing overall building performance and sustainability. Data related to energy consumption; heating, ventilation, and air conditioning (HVAC) systems; lighting; and water usage can help identify inefficiencies and opportunities for optimization. This information allows for more informed decision-making regarding system upgrades, retrofits, or maintenance schedules. Data on how occupants interact with a building's spaces, systems, and technologies can reveal patterns, preferences, and habits. This information can be used to improve occupant comfort, well-being, and productivity by adjusting environmental factors, such as temperature, lighting, and air quality, according to their needs and preferences.

The accessibility of open datasets is of paramount importance in the investigation of indoor residential environments and their respective habitats. Such availability expedites research on housing performance and fosters the development of advanced energy analysis methodologies. The building environment not only exerts active influences but also yields passive effects on human productivity. It bears a significant impact on the health and

comfort levels of the inhabitants within the residential setting [1–6] as individuals spend 87% of their time in indoor environments [6]. Consequently, it is imperative to maintain optimal conditions to enhance the overall experience of both the built environment and its inhabitants.

Open indoor environmental data possess considerable potential to enhance architectural designs and simulations and inform future decisions pertaining to building construction, operations management, and innovative design approaches. It facilitates the establishment of performance evaluation benchmarks across diverse geographic locations, building ages, and typologies [7,8]. These datasets can further contribute to the benchmarking of machine learning models in the context of building and habitat data analysis, thereby promoting the development and evaluation of more accurate and robust predictive algorithms [9,10]. Consequently, indoor environmental datasets have emerged as a critical component with the potential to enhance the overall built environment by providing novel and innovative development guidelines. These improvements encompass reducing building operating expenses and elevating the living experience within indoor environments. In recent years, the trend of developing open-source building performance data (BPD) and the public dissemination of such datasets have garnered increasing interest and momentum [11–19]. Advancements in Internet of Things (IoT) technologies [20–22] have brought forth new techniques and methods for leveraging indoor environmental datasets. Air quality sensing and monitoring systems proposed in Zakaria et al. [23] and Marques et al. [24] can be useful for habitats in multiple manners like detection of harmful gases, measuring optimal oxygen levels, creating alerts for the extensive presence of carbon monoxide, etc.

The present study introduces a dataset comprising one year of indoor environmental data, totaling over 2.5 million records from three single-family houses in Virginia, USA. This dataset is part of a larger longitudinal study investigating the relationship between indoor environmental conditions and occupant behavior. The dataset serves as an initial release, providing a rich foundation for researchers working on building performance modeling, occupant comfort, and energy efficiency.

All participating households gave their informed consent to partake in the study, and the research was conducted in compliance with all Institutional Review Board (IRB) protocols. The indoor environmental data were collected using 10 sensing devices across the three households, deployed through the Building Data Lite (BDL) system [25]. These devices integrate multiple sensors to capture temperature, humidity, air quality, noise levels, and light intensity. Data collection began in the summer of 2021 and continued for approximately a year at a frequency of one record per minute, resulting in a large-scale dataset suitable for time-series analysis. The subsequent sections of this paper delineate the data collection procedures, the characteristics of the dataset, its potential applications, and the technical validation methods used to ensure data reliability.

2. Data Description

The dataset has been made publicly available on the Open Source Framework repository [26]. It includes a text guide on the data organization. The data are located in the data folder, which includes two sub-folders named “plus” and “reg” respectively containing data from Enviro Plus and Enviro sensor arrays. The dataset includes a folder named “meta_data” which contains separate metadata files of corresponding data files. Each metadata file is named after the corresponding data file with a trailing “meta” keyword. The dataset also includes a folder named “code” that contains two Python Notebook files. One is used for generating the metadata files from the data file and another is used for validation

of the collected data. The remainder of this section describes the data and corresponding metadata present in the dataset.

2.1. Background and Summary

The data presented here have been collected through the Building Data Lite (BDL) sensing system [25,27]. A total of 10 sensing nodes were deployed on specific locations of three households. Table 1 provides detailed sensor placement information. It also includes information on the date range of the placement of the sensing devices along with the number of records each device collected during that period. Out of the 10 sensing nodes, 5 used the Enviro+ [28] sensor array, and the remaining 5 used the regular Enviro board [28].

Table 1. Sensor placements and record summary.

SENSOR_TYPE	RPI_ID	ROW_COUNT	START_DATE	END_DATE	LOCATION
Enviro Plus	20	283,074	June 16, 2021	January 17, 2022	House B—Bedroom
Enviro Plus	21	309,591	June 16, 2021	March 11, 2022	House B—Kitchen
Enviro	22	279,801	June 16, 2021	January 12, 2022	House C—Room B
Enviro	23	280,030	June 16, 2021	January 12, 2022	House C—Room A
Enviro Plus	30	438,090	July 12, 2021	July 1, 2022	House C—Room A
Enviro	37	85,669	August 3, 2021	July 6, 2022	House A—Guest Room
Enviro Plus	39	242,101	August 3, 2021	July 6, 2022	House A—Kitchen
Enviro	41	43,465	August 3, 2021	September 7, 2021	House A—Guest Room
Enviro	45	352,087	August 3, 2021	July 6, 2022	House A—Living Room
Enviro Plus	50	222,585	August 3, 2021	July 6, 2022	House A—Master Bedroom

Each record in this dataset either contains 12 attributes for Enviro boards or 15 for Enviro+ boards. The attributes are represented as columns in the CSV files. Columns 1 to 3 represent unique identification information and timestamps. Each of the rest of the columns represents an environmental attribute captured by the sensors. Columns 4 and 7 contain proximity and light data. The LTR-599 sensor is used to measure surrounding proximity and light level. Without making any physical contact, the sensor is capable of identifying the existence of objects that are close by. Proximity is recorded in the nanometer (nm) unit and light is measured in the Lux unit. Columns 5, 6, and 8 represent humidity, air pressure, and temperature data, respectively. A BME280 sensor is used to measure surrounding humidity, air pressure, and temperature. These are measured in relative humidity (%RH), hectopascal (hPa), and degrees Celsius (°C), respectively.

The Enviro boards feature a microelectromechanical systems (MEMS) microphone designed to capture sound events. The recorded sound data are represented in columns 9 through 11, which, respectively, correspond to high, mid, and low sound levels. Column 12 represents the amplitude of sound. The recorded sound data can be utilized for different purposes depending on the application. All sound levels are measured in decibels (dB) unit, which is the standard unit for measuring sound intensity.

Columns 13 to 15 are only available in the Enviro+ CSV files. Columns 13, 14, and 15 represent oxidised, reduced, and ammonia (NH3) data collected from the MICS6814 sensor. The Enviro+ board includes an analog-to-digital converter (ADC) which is useful for interpreting gas sensor readings. The (ADC) takes the voltage readings generated by the sensor, and converts them into resistances that can vary from several hundred Ohms to several tens of thousands of Ohms based on the different gas levels that are present. Gas particle levels are recorded in forms of resistance (Ohms).

Table 2 presents attribute details such as the sensors used to capture different environmental data and the units in which the data were recorded. Additionally, it shows the median values of each attribute from three devices located in three different households. The table reveals that the median temperature of the three houses falls within the range of 26 °C to 29 °C. Moreover, similarities in other attributes are also visible, such as the median sound amplitude ranging between 20 dB to 30 dB. House A (device 39) shows a median proximity reading of 0, indicating that during that reading, there was no movement around the sensing device resulting in zero proximity values. The low Lux level in the light recordings of all three houses indicates that the devices were possibly kept in shadows inside the houses. Although the median light reading of House C (device 30) is 0 Lux, the reading varies widely from 0 Lux to 500 Lux over the recording period of approximately one year. Figure 1 presents a light heatmap of device 30, confirming this variation.

The sensing devices used in this study were supposed to work 24/7 throughout the entire duration. However, after analyzing the retrieved data, it came to light that, there were some time-frames when no data were collected. For example, Figure 2 shows the heatmap of the count of daily entries from sensing device 30. Each device is expected to record 1440 (24×60) entries every day throughout the collection period except for the first and the last day. This is because the devices were programmed to capture one record every minute. In Figure 2, it can be seen that most of the days are bright yellow, representing the desired 1440 entry count. However, there are a few days showing a lower entry count, somewhere around 1000 marked in green. There is also a period from December 2021 to January 2022, where there are zero entries recorded. The sensing devices are programmed to continue data collection in the event of network failure or other connectivity issues. The missing segments of Figure 2 indicate that the sensing device was manually turned off.

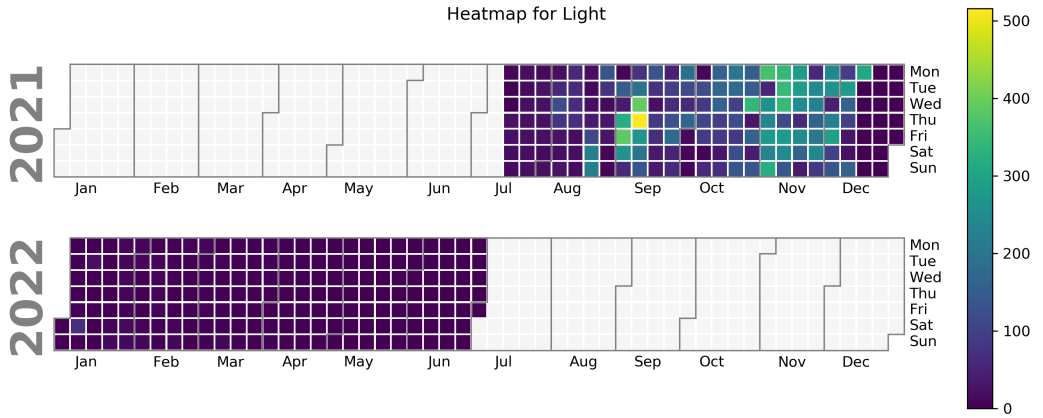

Figure 1. Light heatmap data from device #30.

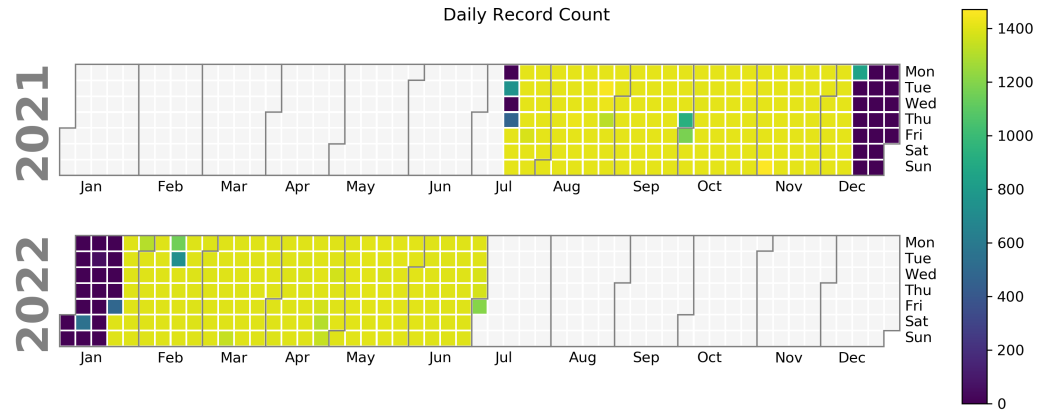

Figure 2. Daily entry count heatmap of device #30 data.

Table 2. Reading units and median values of 3 devices in 3 houses.

Attribute	Sensor	Unit	House A (39)	House B (21)	House C (30)
proximity	LTR-559	nm	0	5	11
humidity	BME280	%RH	20.44	33.14	23.63
pressure	BME280	hPa	1006.85	941.22	939.91
light	LTR-559	Lux	4.22	2.33	0
temperature	BME280	°C	29.01	26.55	26.59
sound_high	MEMS	dB	30.43	30.02	30.03
sound_mid	MEMS	dB	34.08	31.9	31.81
sound_low	MEMS	dB	99.24	54.01	53.14
sound_amp	MEMS	dB	27.52	20.28	20.12
oxidised	MICS6814	kΩ	46.32	114.64	126.79
reduced	MICS6814	kΩ	268.21	240.15	171.31
nh3	MICS6814	kΩ	78.2	89.63	99.16

2.2. Metadata

The dataset includes additional metadata files to provide supplementary information about the data collected by each sensing device. The metadata are in text files named with the identification number of the corresponding sensing device. Each metadata file contains the following six types of information:

- ID: the ID section of the metadata files represents the unique identification number assigned to the sensing device from which the data were collected. This number is recognized by the BDL central system and is specific to a single sensing device.
- Data preview: this section of the metadata file displays the first five rows of data collected by the sensing device, along with the names of each attribute. This provides a quick glimpse into the data recorded by the device.
- Columns: this section of the metadata file lists the names of the attributes recorded in the corresponding data file, separated by commas. For example, it may include attribute names such as 'id', 'date_time', 'rpi_id', 'proximity', 'humidity', 'pressure', 'light', 'oxidised', 'temperature', 'sound_amp'. This section provides a quick reference for the types of data collected by the sensing device.
- Data info: the data info section provides information on the data structure of each attribute, including the index range and total number of entries. It also includes information on the data type and memory consumption of each attribute. This information can be useful for understanding the size and format of the data, as well as for optimizing memory usage and data processing.
- Data description: the data description section provides a statistical summary of the collected data for each attribute in a numeric manner. It includes the minimum, maximum, mean, median, and standard deviation information for each attribute. Table 3 shows an example of the data description section for one of the data files.
- Date range: this section in the metadata file provides information on the time span of the data collected by the particular sensing device. Specifically, it shows the date and time of the first and last recorded data points. This information is important for understanding the temporal scope of the dataset and for identifying any potential gaps in the data collection. The date–time data are provided in the following format: "YYYY-MM-DD HH:MM:SS".

Table 3. Sample data description section of a data file.

	Proximity	Humidity	Pressure	Light	Oxidised	Reduced	nh3	Temperature
mean	10.43935	22.52375	857.1613	59.54044	205.5385	183.4451	103.5953	26.49848
std	5.972486	6.419771	101.0621	87.02963	326.5619	36.8538	24.45422	2.037828
min	0	4.543291	649.5737	0	0.721915	7.505155	3.230769	11.40154
25%	7	17.88105	742.3814	0	92.79227	154.2389	88.2623	25.1387
50%	11	23.63057	939.9159	0	126.7893	171.3063	99.16373	26.59042
75%	14	27.52975	947.0916	8.13295	222.733	212.9956	111.8474	27.89083
max	55	62.58616	961.5713	692.7939	3567.529	2877.333	1931.097	33.36806

The authors have developed a Python script for generating the metadata files from the CSV files downloaded from the BDL server. The script employs Pandas and CSV libraries for analyzing the data and generating the metadata file in text format. The script has been made available within the dataset [26].

2.3. Data Verification

Ensuring the reliability of readings is of paramount importance in cases where data are obtained through an individual case study. This is due to the lack of similar studies for comparison for the identification of any potential anomalies. The data presented in this description encompass multiple domains of building performance data (BPD) and require the use of multiple instruments. However, due to the discrete nature of the experiment, it was not possible to directly verify every category of data presented in this dataset against real-world data. To maintain the integrity of the data, the authors frequently monitored the data being collected through the user interface of the BDL system [27] during the recording period. In addition, the authors obtained actual weather data from Weather Underground [29] and conducted a side-by-side comparison with the data collected by the BDL system to ensure its validity.

Figures 3 and 4 depict a comparison between the indoor temperature readings obtained from the placed sensing devices and the actual outside temperature records acquired from the Richmond International Airport Station through Weather Underground [29]. Since capturing indoor temperature readings directly is not feasible, the authors used outside temperature as a reference scale. It is reasonable to expect a correlation between outdoor and indoor temperatures. The weather data used in this study were collected from the Weather Underground database [29].

Figure 3 depicts the temperature recordings from June 2022 obtained from four sensing devices located in different areas of the house. The green line represents device 39 in the kitchen, the violet line shows device 37 in the guest room, the pink line represents device 45 in the living room, and the red line represents device 50 in the master bedroom. The bold black line indicates the actual outdoor temperature. The figures illustrate that the temperature readings of different devices demonstrate parallel patterns. Additionally, the data show a striking resemblance between the actual outdoor temperature readings and the indoor temperature readings from different rooms. Moreover, temperature peaks observed in the outside temperature around 10th, 12th, and 17th of June are also visible in the indoor sensor readings.

Figure 4 presents a comparison between the actual outdoor temperature and indoor temperature readings from two distinct houses (House A and House B) situated in Richmond, VA. The graph displays data from three sensing devices, specifically device 20 located in the bedroom of House B (blue line), device 21 located in the kitchen of House B (orange line), and device 45 placed in the living room of House A (pink line). The outside temperature is indicated by a bold black line. The data in this figure pertain to a 4-month period spanning from September 2021 to the end of December 2021. A noteworthy

observation from this figure is that, despite the decline in the outdoor temperature during this period, the indoor temperature did not decrease significantly, presumably because of the use of room heaters. The indoor temperature lines also demonstrate a parallel nature, indicating a consistent pattern. Comparison of the collected data with the actual outdoor temperature displayed in Figures 3 and 4 shows a plausible change pattern in temperature over different time periods, thus substantiating the reliability of the collected data.

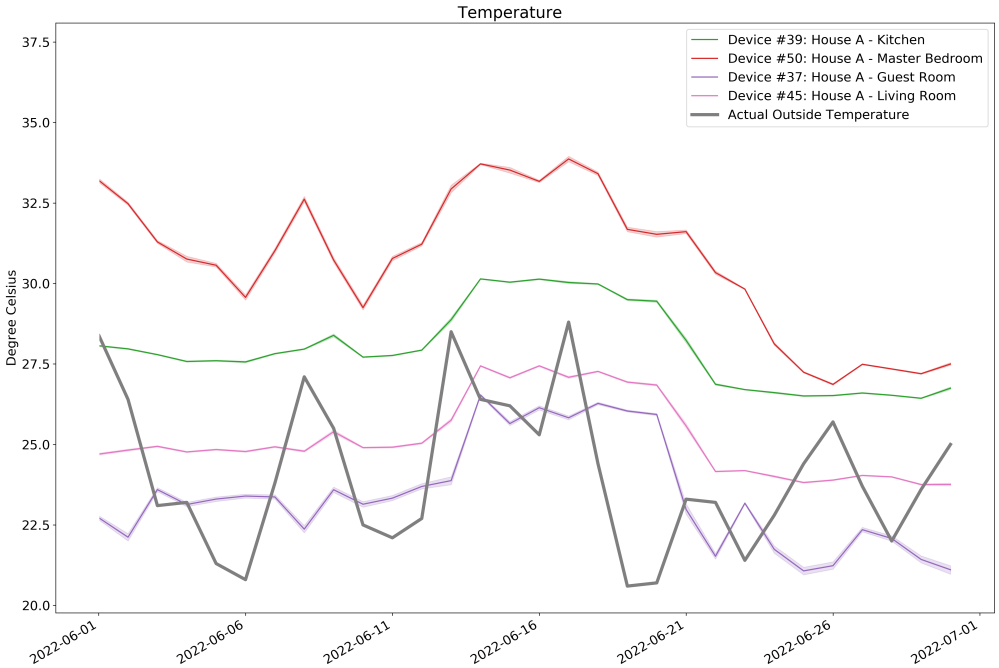

Figure 3. Comparison of outside temperature data of June 2022 in Richmond, VA with indoor temperature records collected by the deployed sensors.

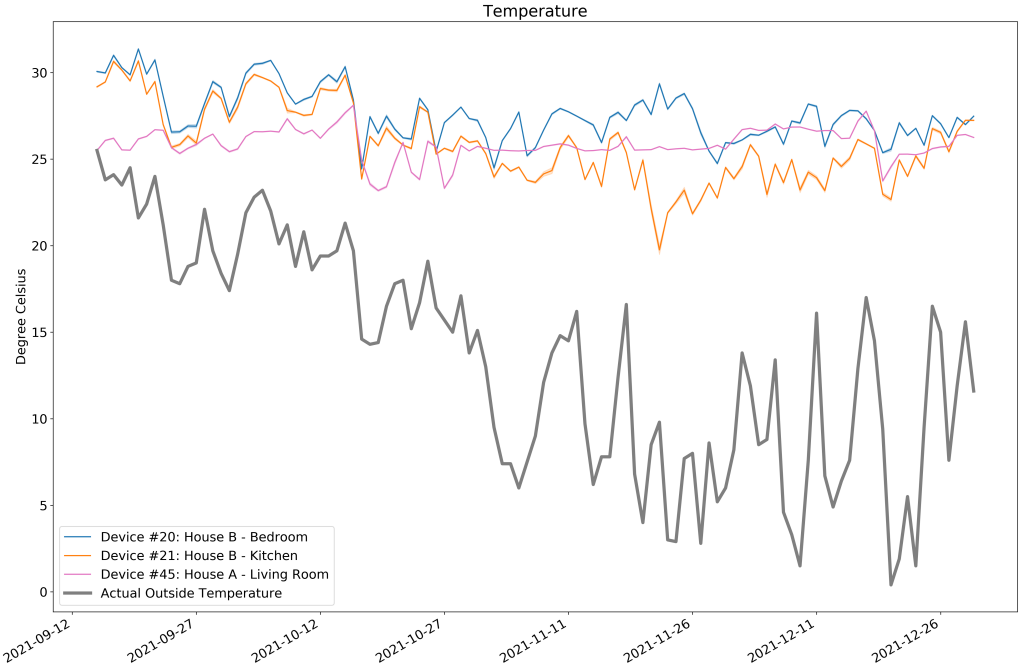

Figure 4. Comparison of outside temperature data from September 2021–December 2021 in Richmond, VA with indoor temperature records collected by the deployed sensors.

The verification of certain attributes, such as oxidised, reduced, or sound levels, is technically challenging due to the absence of a benchmark dataset for comparison. However, it is plausible for some attributes to have similar readings among the devices of the same house and also among devices of different houses located nearby. Humidity is such an attribute. Figure 5 shows the humidity records collected by three sensing devices placed inside House A and House B. Device 20 is located in the bedroom of House B (blue line), device 21 is located in the kitchen of House B (orange line), and device 45 is located in the living room of House A (pink line). Humidity records are recorded in the relative humidity (% RH) unit. The indoor humidity is in general lower than the outside humidity. Here, a similarity in the humidity readings can be seen among different devices throughout the 4 months. Even devices placed in different houses demonstrate resemblance which indicates the validity of the collected data.

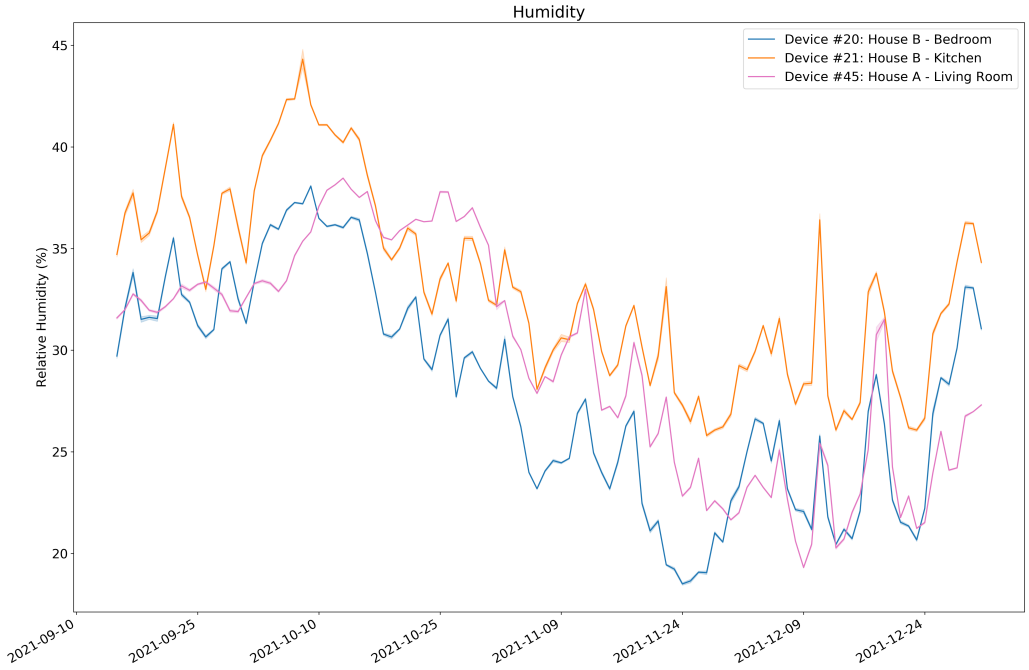

Figure 5. Comparison of humidity records collected by the deployed sensors from September 2021 to December 2021 in Richmond, VA.

While direct validation of oxidized gas, reduced gas, and sound levels remains challenging due to the lack of established reference datasets, future research efforts could explore alternative validation methods, such as sensor calibration experiments, controlled environment testing, or comparisons with external datasets from similar studies. Additionally, statistical anomaly detection methods could help assess data consistency over time. We acknowledge this as an important area for future work, particularly as the dataset expands in subsequent phases of the study.

2.4. Limitation and Threats to Validity

While this dataset provides a comprehensive record of indoor environmental conditions, certain limitations must be acknowledged. One key challenge is the presence of missing data, which may occur due to temporary power loss, sensor disconnections, or hardware malfunctions. Although the data collection system supports both offline and online modes, instances where devices lose power or are inadvertently removed from their placement result in data gaps. To mitigate this, multiple sensors were deployed in each household, allowing for redundancy in data collection. Furthermore, the dataset is limited

to three single-family houses within the same geographic region, which may affect the generalizability of findings to other climates or building types. However, this study is part of a larger data collection effort involving more households and sensing devices, and future expansions will address this limitation.

3. Methods

This section elaborates on the different components of the data collection methods used in this study. Beginning with the data collection architecture, sensors used in the process, sensing device deployment, and participant declaration. The study utilized the Building Data Lite (BDL) [25] system to collect data. A unique set of sensors were used to capture the different information from the surroundings. The sensing devices were deployed to multiple individual houses with the informed consent of the occupants.

3.1. Building Data Lite

The data presented in this paper were obtained using the Building Data Lite (BDL) system [25]. BDL is a distributed, portable, scalable, and cost-effective indoor environment sensing system. The BDL system is an open-source platform [30] that facilitates the development of customizable and portable sensing devices capable of connecting to a central server. These sensing devices continuously transmit collected data to the central server via the internet. However, in case of a connection interruption, the data are stored on the local storage available in each sensing device. The BDL system features a web interface through the central server to access and download the collected data.

The data collected in the BDL system are first stored in the local database of the sensing devices and then transmitted to the central database located in the central server. Both of these databases are relational databases and share similar characteristics. The Raspberry Pi (RPi) devices by Raspberry Pi Foundation from Cambridge, United Kingdom, use the inbuilt MariaDB [31] to store offline data. The central server utilizes MySQL [32] to organize all collected data and share them through the visualization interface.

3.2. Sensor Array

The BDL sensing nodes used for data collection consisted of either an Enviro or an Enviro Plus sensor array [28] along with a Raspberry Pi Zero. These sensor arrays were developed by the company Pimoroni located in Sheffield, Yorkshire, United Kingdom. Both of these sensor boards function similarly and include the same sensors except for the analog gas sensor which is exclusive to the Plus edition of Enviro. The following are the descriptions of the different sensors used in this study.

3.2.1. Temperature, Pressure, and Humidity Measurement

The BME280 sensor [33,34] on the Enviro+ board is a high-precision sensor that can measure temperature, pressure, and humidity. The placement of the sensor was deliberately made on the left side of the board with the intention of preventing any potential heat produced by the Raspberry Pi's CPU from reaching the sensor, which could interfere with the precision of the readings. The BME280 is commonly used for indoor environmental monitoring and can provide valuable information on the conditions of a home or other indoor space. There is a little slot right next to the sensor which can be useful to lessen the amount of heat that is generated from the Enviro+ board towards the direction of the sensor, further improving the accuracy of the readings.

3.2.2. Light, and Proximity Measurement

The LTR-559 sensor [34] on the Enviro and Enviro Plus sensor array can detect the amount of light in Lux. Lux represents the measure of the intensity of visible light. This

sensor is useful for monitoring lighting conditions in indoor environments. Additionally, the LTR-559 sensor includes a proximity sensor, which can detect the presence of objects within a certain distance from the sensor. This feature can be used to create proximity-sensitive inputs, which can be useful in applications such as touchless control interfaces.

3.2.3. Sound Measurement

The Enviro+ device features a miniature Micro-Electro-Mechanical System (MEMS) microphone [28], which facilitates the recording of audio and detection of noise levels [35]. This functionality is particularly useful for monitoring and assessing the degree of noise pollution. The device is capable of detecting high, mid, and low sound levels, as well as measuring the amplitude of the sound.

3.2.4. Gas Measurement

The MICS6814 [36] is an analog gas sensor that is exclusively available in the Plus variant of the Enviro boards. This sensor can detect three distinct groups of gases categorized as reducing, oxidizing, and NH3 according to the datasheet. Notably, the MICS6814 can detect the presence of major gas or vapors like carbon monoxide (reducing), nitrogen dioxide (oxidizing), and ammonia (NH3), in addition to other gases including hydrogen, ethanol, and hydrocarbons. The recommended approach to interpret MICS6814 gas data is to record readings until they reach a steady state, establish a baseline, and then examine changes in relation to that baseline. This method provides a general indication of the air quality trend. Further information on how to interpret these readings and the chart is available in the guide provided by Pimoroni [28].

3.3. Sensing Device Deployment

The dataset presented in this study is part of a larger ongoing study primarily located in Virginia, USA. In this larger study, a total of 48 sensing devices were deployed in different locations of 12 individual households. Four sensing devices were placed in each house. The devices were placed with the acknowledged consent of the occupants living in the houses following the regulations of the Institutional Review Board (IRB), Virginia Tech.

The deployment of the sensing devices took place during the summer of 2021 and the devices have been collecting data since. However, not all the sensing devices are connected to the internet and such devices are collecting data in offline mode. The data from the offline devices can be retrieved manually with physical access to the devices once they are brought back. The current study presents data from 10 devices of three individual houses. Some of these devices were retrieved from the households and some were online throughout the time span of a year. The rest of the deployed sensing devices are still in the occupant houses and are still collecting data.

The dataset presented in this study was derived from a singular, non-random case study. The data collection process took place in three separate housing units, namely House A and House B, both located in Richmond, Virginia, and House C situated in Christiansburg, Virginia. Information regarding the placement locations of the sensing devices can be found in Table 1. The sensors were installed in specific areas of each site, such as living rooms, kitchens, and bedrooms.

Figure 6 shows an example of sensor deployment in House A floor layout. The green rectangles in the figure symbolize the sensing devices, while the light green circle that surrounds each sensing device represents the hypothetical coverage area of the device. These devices are directly connected to the central server through a Wi-Fi internet connection. Additionally, each device includes a local data backup in case of connectivity issues. These sensing devices are also portable and can be easily relocated to any location with access to power and network connection.

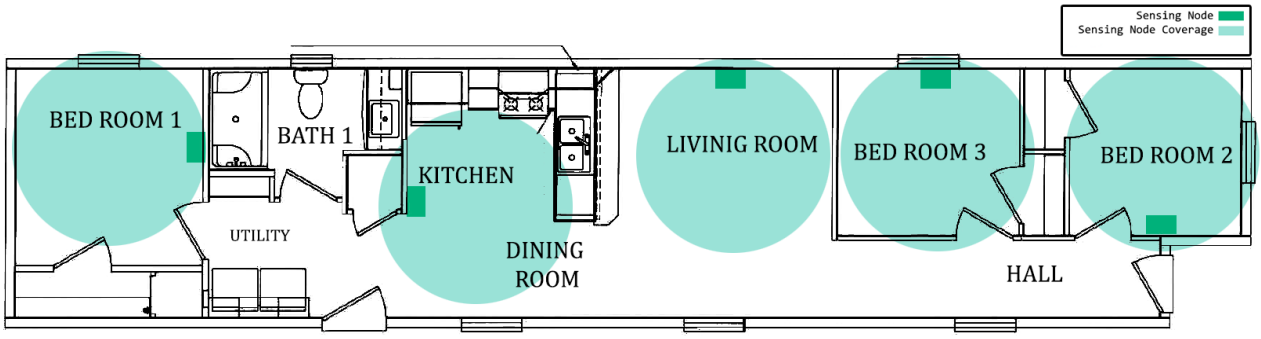

Figure 6. Sensor deployment plan of House A.

4. User Notes

This dataset and the accompanying methodology can serve as a valuable resource for researchers and practitioners working in various domains, including building energy efficiency, occupant behavior modeling, and predictive maintenance. Researchers may leverage the dataset to validate building performance models, facilitating studies on the interaction between indoor environmental conditions and occupant behavior [37]. The dataset is well-suited for educational purposes in interdisciplinary fields such as environmental science, data analytics, and smart building technology. It can serve as a foundation for future studies on the interaction between indoor environmental conditions and residential health outcomes. Researchers can investigate how temperature stability, humidity control, and air quality influence occupant well-being, including respiratory health and sleep quality [38]. Incorporating data on insulation and airtightness in future expansions of this dataset can further enhance its utility in evaluating energy efficiency and thermal comfort [39]. This information can also be utilized for predictive smart home automation modeling, and HVAC control methods are realized depending on environmental trends. Through the facilitation of cross-disciplinary research in areas such as environmental health, smart buildings, and human-oriented energy management, this dataset provides a wealth of resource material to attain sustainable and intelligent living environments.

The presented data have been collected through multiple similar sensing devices thus the data files include similar categories of data. The authors used the Pandas (Open-source, Version 2.2.3) [40] and Matplotlib (Open-source, Version 3.9) [41] libraries of Python (Open-source, Version 3.10) for data analysis and visualization. An example is available in the form of Jupyter Notebook (Open-source, Version 7.2.2) [42] on Open Science Framework (OSF) [26]. Given the rapid growth of smart building technologies, this dataset holds the potential to drive advancements in understanding and optimizing indoor environmental quality. It may also inform the development of innovative solutions for energy management, occupant comfort, and health. By making this dataset publicly available, the authors aim to contribute to the broader scientific community, enabling researchers to explore novel perspectives and conserve significant time and effort.

Author Contributions: N.M. and X.G. supervised the project, provided insights on analyzing the collected data, and guided the manuscript writing. X.G. installed the sensing devices in occupant housing. S.M.H.A. integrated the BDL system with the sensing devices, developed the scripts to validate the data, generate metadata information, and led the writing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science Foundation (NSF-1845 446 and NSF-1929 701 grants).

Institutional Review Board Statement: This research received approval from the Institutional Review Board of Virginia Tech (VT IRB 20-784 and 21-507), ensuring that the research coheres with ethical standards. Participants of this study were not subject to any known risks and informed consent was provided from all subjects.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original data described in the study are openly available at https://doi.org/10.17605/OSF.IO/BAEW7 [26] (accessed on 23 January 2025).

Acknowledgments: The authors would like to express their gratitude towards Zack Miller, Marion Cake, and Madeline Petrie at project:HOMES for supporting the case study. The authors would like to thank National Science Foundation, Virginia Housing Development Authority, and Virginia Center for Housing Research for the research funding support.

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

Abbreviations

The following abbreviations are used in this manuscript:

HVAC	Heating, Ventilation, and Air Conditioning
BDL	Building Data Lite
BPD	Building Performance Data
IoT	Internet of Things
IRB	Institutional Review Board
CSV	Comma Separated Values
MEMS	Microelectromechanical systems
ADC	Analog-to-Digital Converter
CPU	Central Processing Unit
RPi	Raspberry Pi
OSF	Open Science Framework

References

1. Hedge, A.; Miller, L.; Dorsey, J. Occupant comfort and health in green and conventional university buildings. *Work* **2014**, *49*, 363–372. [CrossRef] [PubMed]
2. Mirzaei, N.; Kamelnia, H.; Islami, S.G.; Kamyabi, S.; Assadi, S.N. The impact of indoor environmental quality of green buildings on occupants' health and Satisfaction: A systematic review. *J. Community Health Res.* **2020**, *9*, 54–65. [CrossRef]
3. Andargie, M.S.; Touchie, M.; O'Brien, W. A review of factors affecting occupant comfort in multi-unit residential buildings. *Build. Environ.* **2019**, *160*, 106182. [CrossRef]
4. Zhang, Y.; Tzortzopoulos, P.; Kagioglou, M. Healing built-environment effects on health outcomes: Environment–occupant–health framework. *Build. Res. Inf.* **2019**, *47*, 747–766. [CrossRef]
5. Ghodrati, N.; Samari, M.; Shafiei, M.W.M. Green buildings impacts on occupants' health and productivity. *J. Appl. Sci. Res.* **2012**, *8*, 4235–4241.
6. Mujan, I.; Anđelković, A.S.; Munćan, V.; Kljajić, M.; Ružić, D. Influence of indoor environmental quality on human health and productivity-A review. *J. Clean. Prod.* **2019**, *217*, 646–657. [CrossRef]
7. Li, Z.; Han, Y.; Xu, P. Methods for benchmarking building energy consumption against its past or intended performance: An overview. *Appl. Energy* **2014**, *124*, 325–334. [CrossRef]
8. Roth, J.; Lim, B.; Jain, R.K.; Grueneich, D. Examining the feasibility of using open data to benchmark building energy usage in cities: A data science and policy perspective. *Energy Policy* **2020**, *139*, 111327. [CrossRef]
9. Quevedo, T.; Geraldi, M.; Melo, A. Applying machine learning to develop energy benchmarking for university buildings in Brazil. *J. Build. Eng.* **2023**, *63*, 105468. [CrossRef]
10. Robinson, C.; Dilkina, B.; Hubbs, J.; Zhang, W.; Guhathakurta, S.; Brown, M.A.; Pendyala, R.M. Machine learning approaches for estimating commercial building energy consumption. *Appl. Energy* **2017**, *208*, 889–904. [CrossRef]
11. Pipattanasomporn, M.; Chitalia, G.; Songsiri, J.; Aswakul, C.; Pora, W.; Suwankawin, S.; Audomvongseree, K.; Hoonchareon, N. CU-BEMS, smart building electricity consumption and indoor environmental sensor datasets. *Sci. Data* **2020**, *7*, 241. [CrossRef] [PubMed]

12. Tasgaonkar, P.; Zade, D.; Ehsan, S.; Gorti, G.; Mamnun, N.; Siderius, C.; Singh, T. Indoor heat measurement data from low-income households in rural and urban South Asia. *Sci. Data* **2022**, *9*, 285. [CrossRef]
13. Yoon, Y.; Jung, S.; Im, P.; Gehl, A. Datasets of a Multizone Office Building under Different HVAC System Operation Scenarios. *Sci. Data* **2022**, *9*, 775. [CrossRef] [PubMed]
14. Gao, N.; Marschall, M.; Burry, J.; Watkins, S.; Salim, F.D. Understanding occupants' behaviour, engagement, emotion, and comfort indoors with heterogeneous sensors and wearables. *Sci. Data* **2022**, *9*, 261. [CrossRef]
15. Thorve, S.; Baek, Y.Y.; Swarup, S.; Mortveit, H.; Marathe, A.; Vullikanti, A.; Marathe, M. High resolution synthetic residential energy use profiles for the United States. *Sci. Data* **2023**, *10*, 76. [CrossRef] [PubMed]
16. Schwee, J.H.; Johansen, A.; Jørgensen, B.N.; Kjærgaard, M.B.; Mattera, C.G.; Sangogboye, F.C.; Veje, C. Room-level occupant counts and environmental quality from heterogeneous sensing modalities in a smart building. *Sci. Data* **2019**, *6*, 287. [CrossRef]
17. Dong, B.; Liu, Y.; Mu, W.; Jiang, Z.; Pandey, P.; Hong, T.; Olesen, B.; Lawrence, T.; O'Neil, Z.; Andrews, C.; et al. A global building occupant behavior database. *Sci. Data* **2022**, *9*, 369. [CrossRef]
18. Agee, P.; Nikdel, L.; Roberts, S. A measured energy use, solar production, and building air leakage dataset for a zero energy commercial building. *Sci. Data* **2021**, *8*, 299. [CrossRef]
19. Paige, F.; Agee, P.; Jazizadeh, F. flEECe, an energy use and occupant behavior dataset for net-zero energy affordable senior residential buildings. *Sci. Data* **2019**, *6*, 291. [CrossRef]
20. Bashir, M.R.; Gill, A.Q. Towards an IoT big data analytics framework: Smart buildings systems. In Proceedings of the 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Sydney, NSW, Australia, 12–14 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1325–1332.
21. Baghalzadeh Shishehgarkhaneh, M.; Keivani, A.; Moehler, R.C.; Jelodari, N.; Roshdi Laleh, S. Internet of Things (IoT), Building Information Modeling (BIM), and Digital Twin (DT) in Construction Industry: A Review, Bibliometric, and Network Analysis. *Buildings* **2022**, *12*, 1503. [CrossRef]
22. Tang, S.; Shelden, D.R.; Eastman, C.M.; Pishdad-Bozorgi, P.; Gao, X. A review of building information modeling (BIM) and the internet of things (IoT) devices integration: Present status and future trends. *Autom. Constr.* **2019**, *101*, 127–139. [CrossRef]
23. Zakaria, N.A.; Abidin, Z.Z.; Harum, N.; Hau, L.C.; Ali, N.S.; Jafar, F.A. Wireless internet of things-based air quality device for smart pollution monitoring. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 65–69. [CrossRef]
24. Marques, G.; Pitarma, R. A cost-effective air quality supervision solution for enhanced living environments through the internet of things. *Electronics* **2019**, *8*, 170. [CrossRef]
25. Anik, S.M.H.; Gao, X.; Meng, N.; Agee, P.R.; McCoy, A.P. A cost-effective, scalable, and portable IoT data infrastructure for indoor environment sensing. *J. Build. Eng.* **2022**, *49*, 104027. [CrossRef]
26. Gao, X.; Anik, M.H. A Comprehensive Indoor Environment Dataset from Single-family Houses in the US. *OSF* **2023**. [CrossRef]
27. Anik, S.M.H. Building Data Lite. Available online: https://www.building-data-lite.com (accessed on 1 March 2023).
28. Macdonald, S. Getting Started with Enviro+. 2019. Available online: https://learn.pimoroni.com/article/getting-started-with-enviro-plus (accessed on 1 March 2023).
29. Underground, W. Henrico, VA Weather History. 2021. Available online: https://www.wunderground.com/history/monthly/us/va/henrico/KRIC/date/2021-9 (accessed on 1 March 2023).
30. Anik, S.M.H. BDL Project Repository. Available online: https://github.com/anik801/data_collection (accessed on 1 October 2022).
31. Kenler, E.; Razzoli, F. *MariaDB Essentials*; Packt Publishing Ltd.: Birmingham, UK, 2015.
32. Bartholomew, D. Mariadb vs. mysql. *Dostopano* **2012**, *7*, 2014.
33. Bosch Sensortec. *BME280—Datasheet: BME280 Combined Humidity and Pressure Sensor*; Version: 1.24; Bosch Sensortec: Reutlingen, Germany, 2015.
34. Riffelli, S. A Wireless Indoor Environmental Quality Logger Processing the Indoor Global Comfort Index. *Sensors* **2022**, *22*, 2558. [CrossRef]
35. Loeppert, P.V.; Lee, S.B. SiSonicTM-The first commercialized MEMS microphone. In Proceedings of the Solid-State Sensors, Actuators, and Microsystems Workshop, Hilton Head Island, SC, USA, 4–8 June 2006; pp. 27–30.
36. De Medeiros, H.P.L.; Girão, G. An iot-based air quality monitoring platform. In Proceedings of the 2020 IEEE International Smart Cities Conference (ISC2), Piscataway, NJ, USA, 28 September–1 October 2020; pp. 1–6.
37. Yan, D.; O'Brien, W.; Hong, T.; Feng, X.; Gunay, H.B.; Tahmasebi, F.; Mahdavi, A. Occupant behavior modeling for building performance simulation: Current state and future challenges. *Energy Build.* **2015**, *107*, 264–278. [CrossRef]
38. Wolkoff, P. Indoor air humidity, air quality, and health–An overview. *Int. J. Hyg. Environ. Health* **2018**, *221*, 376–390. [CrossRef]
39. Wąs, K.; Radoń, J.; Sadłowska-Sałęga, A. Thermal comfort—Case study in a lightweight passive house. *Energies* **2022**, *15*, 4687. [CrossRef]

40. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 56–61. [CrossRef]
41. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
42. Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.; Grout, J.; Corlay, S.; et al. Jupyter Notebooks—A publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*; Loizides, F., Schmidt, B., Eds.; IOS Press: Amsterdam, The Netherlands, 2016; pp. 87–90.