Airborne Insights: Predicting Occupancy and More from Air Measurements in Enclosed Environments

Oytun Kuday Duran
oytunkuday@sabanciuniv.edu

***Abstract*—This paper explores the use of machine learning algorithms and air quality sensor data from the Internet of Things (IoT) to predict occupancy and associated activities in large, enclosed environments such as auditoriums. The study leverages data from sensors placed throughout various classrooms at Sabancı University, aiming to develop models that predict the number of occupants during specific time frames and detect potential information leaks. Our work demonstrated promising results, suggesting that variations in air quality can reliably predict occupancy status, approximate number of occupants, and potentially infer the department of courses being conducted. Although the project faces challenges such as variations in classroom environments and external factors impacting air quality and occupancy, it provides valuable insights into the potential privacy implications and energy optimization opportunities of such IoT devices. Furthermore, the paper underscores the need for continued refinement and testing of these prediction models in different settings and under varying conditions.**

***Index Terms*—Air quality, occupancy, applied machine learning, sensors, privacy, Internet of Things.**

I. Introduction

THE rapid advancement and integration of smart devices into everyday life has precipitated a shift towards an increasingly interconnected world. The proliferation of the Internet of Things (IoT) has transformed the fabric of domestic spaces into intricate networks of smart homes, facilitating communication between devices to manage and improve various aspects of human life. One salient application of this technological evolution is observed in smart ventilation systems, which meticulously monitor air quality, tracking parameters such as temperature, humidity, and levels of airborne pollutants.

However, these conveniences offered by smart devices are not without potential risks. The interconnectedness that renders our lives more comfortable also unveils considerable privacy vulnerabilities. As such, as we progressively embrace the comfort afforded by IoT, it becomes critical to simultaneously consider and implement robust security measures, effectively safeguarding privacy within this extensively connected landscape.

This study aims to explore the privacy implications associated with IoT devices by employing air quality data obtained from devices installed in various classrooms at Sabancı University. Utilizing machine learning models, we aim to predict the number of occupants within a particular classroom during a specific timeframe, along with a possible identification of the nature of the ongoing event. The study employs time-series data comprising temperature, humidity, CO2 levels, and other pertinent air quality measures acquired from IoT devices that has an air sensor. Furthermore, various Python tools were employed to engineer and analyze data retrieved from the course schedule websites, providing early-stage labels for the measurement dataset, thereby expediting development and feature extraction.

Our primary motivation stems from understanding the correlation between air quality and occupancy and the potential for privacy breaches through seemingly inconsequential IoT devices. Preliminary analyses of the initial features and ensuing predictions underscore the presence of strong correlations, even within such uncontrolled environments, thus emphasizing the potential privacy threats intrinsic to IoT integration.

II. Related Work

In terms of literature and existing solutions, a gap is perceptible. A limited number of studies have utilized air quality data to predict occupancy and assess potential privacy threats within given environments but also mentioning possible benefits of employing such a tool. However, these studies often lack a realistic scenario, demonstrating a rigorous protection against external factors in the experiments. In [6], researchers use temperature, humidity, CO2 and light to predict binary indoor occupancy in a small room with 99% accuracy. Similarly, in a student dorm of 4, [8] predicts binary occupancy and headcount with an accuracy of 81.1% and 64.7%, under strict conditions such as ventilation of room for 10 minutes before every measurement. Dutta and Joy [2] utilizes CatBoost algorithm using interior data with multiple distinct exterior factors with an accuracy of 99.85% for occupancy, and 93.20% for headcount up to 6. Moreover, some of these approaches have proposed the experiments we have conducted as future work, emphasizing the relevance of using a real-life setting and higher amount of headcount. Our study aims to address this gap, proposing a novel model that takes advantage of IoT device air quality data to predict classroom occupancy and possible events, evaluating the extent of potential privacy leakage. This investigation sets a new trajectory in this field by dealing with real-life conditions, enhancing the practicality of the proposed solution by only using the air quality information.

III. Challenges

Our work that is using data inferred from the daily life, due to the real-life setting without any restrictions (artificial con-

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

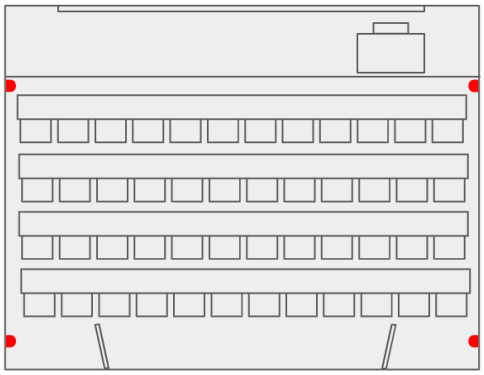

Fig. 1. Sketch of a classroom (capacity of 125), red nodes represent sensors

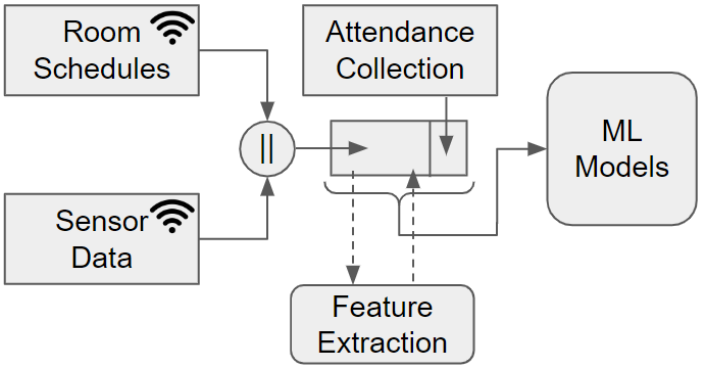

Fig. 2. An abstract diagram of the experiment

trols), faced many challenges. Lower attendance in classrooms due to hybrid education, technical problems such as sensors being damaged or disconnected affected the data collection. Moreover, the ventilation systems in the classrooms were not operational outside of regular class hours, and sometimes just spontaneously turned off, which resulted in pollutant build-up and significantly worse air quality. This prevented establishing a baseline for prediction and led to less reliable air quality data. Additionally, there were fluctuations and noise in measurements which stemmed from the air flow due to the doors and windows that could be opened and closed by anyone, adding even more complexity to our data. On the administrative side, we faced delays in accessing past attendance records due to privacy concerns and legal issues.

IV. Methodology

A. Measurements

The measurements for this project were collected using Bosch BME680 air quality sensors, which were strategically placed in classrooms and auditoriums to ensure air quality during the Covid period. These sensors were designed to measure a variety of air quality metrics and performed periodic measurements at one minute intervals. The measurements captured data based on 11 different metrics, which are detailed below:

- **Raw Temperature:** Corresponds to the unprocessed, direct temperature measurement obtained from the sensor, devoid of any compensatory adjustments.
- **Temperature:** Refined temperature measurement, adjusted to account for the internal cross-influences induced by the sensor's operational characteristics.
- **Pressure:** Provides quantifiable information pertaining to the atmospheric pressure in the sensor's immediate environment.
- **Raw Humidity:** Represents the unprocessed humidity measurement, directly acquired from the sensor without any compensatory adjustments.
- **Humidity:** Refined relative humidity measurement, adjusted to account for internal cross influences induced by the operational characteristics of the sensor.
- **Raw Gas Sensor Resistance:** Quantifies the resistance offered by the gases present in the environment, measured directly by the sensor.
- **Air Quality:** Comprehensive metric providing a general indication of the air quality in the environment, as measured by the sensor.
- **Static Air Quality:** Specialized air quality metric, processed specifically for stationary devices, such as those utilized in this project for data collection.
- **Air Quality Accuracy:** Quantifies the degree of accuracy of the air quality measurement provided by the sensor.
- **Breath VOC Equivalent:** Quantifies the concentration of a mixture of gases typically produced by respiration present in the environment.
- **CO2 Equivalent:** Estimated metric of CO2 content in the environment, expressed in parts per million (ppm), derived from the other measurements.

The measurements were sent to the database periodically using Wi-Fi linking the devices to the database with MAC addresses. Lastly, while a measurement entry is being formed, it received a timestamp. The measurements collected from multiple sensors in the same classroom were combined using a weighted average, with the weights determined by the accuracy metric of each sensor.

B. Data Collection and Pre-processing

The Data Collection and Pre-processing phases involved the collection of data from air quality sensors and classroom occupancy counts, the transformation of these raw data into a structured format, and the integration and normalization of these data to create a comprehensive dataset. The data collection period spanned the Fall 2022 and Spring 2023 semesters, specifically from the 3rd of October, 2022, to the 16th of May, 2023. Measurements during the period yielded approximately 10 million entries, each entry representing a measurement value of a specific measurement type from a sensor at a specific timestamp.

To ensure the consistency and workability of the collected data, Python libraries that are specialized in data processing was used. Raw data were cleaned and transformed to achieve a uniform format. The raw sensor data was merged with the measurement type information data and the sensor address

information, resulting in a uniform and representable dataset. The data was then filtered based on their timestamps to include only a specific time interval, this interval being between 8.30 AM and and 7.30 PM, which is the interval that the University has scheduled courses at. There are also important reason why the interval between 7.30 PM to 8.30 AM was not considered:

1) The air conditioning in the buildings were automatically turned off after 7.30 PM, resulting in spikes in air quality that are undesired.
2) As there was no lecture after 7.30 PM, inclusion of such data would increase the bias significantly as there would be no one in classrooms during the discluded interval.

The entries were then formatted again such that there would be one entry for each 10-minute period since 1-minute measurements were not needed as predictions would take part at 50 minute interval.

For the purpose of binary classification of occupancy and prediction of course departments, data was gathered pertaining to events occurring in classrooms and auditoriums from two primary sources responsible for course scheduling, namely the Mysu website and Bannerweb. The Mysu website offers a room scheduling tool that displays information on time slots and assigned classrooms for lectures, recitations, and labs. To extract the relevant data, web scraping techniques utilizing Python libraries such as BeautifulSoup were employed.

For predicting the approximate count of attendants, the dataset was concatenated with manually collected attendance to include information about the number of students attending lectures in specific classrooms and time slots. Furthermore, the dataset was enriched with attendance sheets from some courses offered during previous semesters.

In summary, the data collection and processing stages of this study involved web scraping techniques using Python libraries for extracting data from the University's course schedule website. The extracted data was transformed, cleaned, and organized, resulting in a structured dataset ready for analysis, feature engineering, and feeding the learning models. The inclusion of attendance sheets and physical and manual counting of attendants made by students added valuable information to the dataset for multi-class classification tasks. By undertaking these rigorous data collection and processing procedures, we created a solid foundation for accurate prediction modeling.

C. Feature Engineering and Modelling

In this study, a few different targets for the prediction of models were present. The predictions were made can be categorized to 4 stages:

1) Time-Series Binary Occupancy Prediction
2) Time-Series Multi-Class Occupancy Prediction
3) Interval Multi-Class Occupancy Prediction
4) Faculty & Course Prediction

We applied numerous modifications to the dataset, features and the models across each stage. The primary reason behind this categorization was our strategic approach to modeling. The experiment started with more abstract predictions based on limited information and gradually advanced towards more detailed predictions with the potential to extract more valuable insights.

TABLE I
CORRELATIONS BETWEEN MEASUREMENTS AND BINARY OCCUPANCY

Measurement	Correlation
Humidity	-0.15
Temperature	0.54
Pressure	0.04
bVOC	0.34
CO_2	0.37
gasR	0.02
rawH	-0.11
rawT	0.52

The first target was to see whether it is possible to predict if there is a lecture in the classrooms. The labels for prediction were arranged as '1' if there is a lecture and '0' if there is no scheduled lecture in the specific classroom. First, the correlations of measurements extracted. The initial correlations of the parameters with the binary occupancy can be seen in Table I. Redundant features, such as raw measurements that were highly correlated with other corresponding measurements, were removed from the dataset to minimize the risk of overfitting.

Over 30 models were trained at the beginning including Support Vector Machines (SVMs), Decision Trees, Naïve Bayes classifiers, neural networks and other well-known regression, boosting, bagging models. 20% of the available dataset was separated for testing, and a 10-fold cross-validation was applied during training to find suitable models. Hyperparameter tuning was applied to tune optimal model parameters to enhance performance in each iteration. These parameters guide the learning process, shaping aspects such as the optimization rate or complexity of the model. It is crucial as correct tuning aids model accuracy, while poor choices risk underfitting or overfitting, impairing the model's applicability to new data. In average, 9600 iteration was done for training of each model including cross -fitting. In each stage of our evaluation, models that performed poorly were systematically eliminated.

Initially, the dataset was utilized for eight separate classrooms for binary occupancy prediction. However, due to the availability of attendance information and reliability of the collected data, for subsequent multi-class occupancy evaluations the classroom count was reduced to two, one semi-auditorium and one auditorium with total capacities of 125 and 200, respectively.

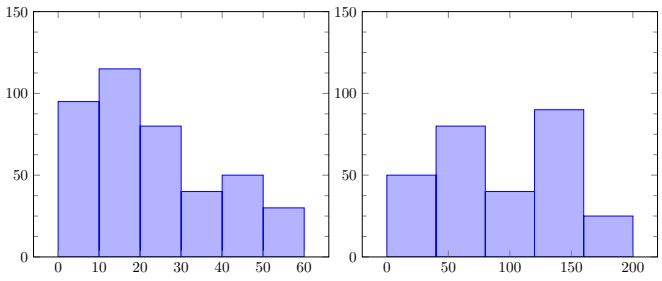

Fig. 3. Attendance distributions of datasets in two classrooms, respectively.

During the transition to multi-class occupancy detection,

training a regression model for occupancy detection in such an uncontrolled environment was not feasible. This was due to the variability and unpredictability of occupancy patterns and random events. Therefore, we chose to divide the number of occupants into$n$n bins, thus creating n distinct categories for a more effective and manageable analysis. The value of n was determined by the following, where Res_p corresponds to the resolution (granularity) of the prediction:

$$n = \lceil \frac{Max_{attendances}}{Res_p} \rceil \tag{1}$$

After the predictions utilizing time-series information, it was desired to understand the relation of air quality trends over time with number of occupants and the predictability using only the changes in air measurements. Feature vectors were generated for each lecture block, each of which consists of 50 minutes. Temporal index, denoted as $T_{HH.MM}$, where $HH.MM$ refers to hours and minutes respectively, is defined:

$$T_{08.40} = \{08.40, 08.50, 09.00, 09.10, 09.20, 09.30\} \tag{2}$$

The temporal index $T_{08.40}$ incorporates time points ranging from 08.40 to 09.30, inclusive, in increments of 10 minutes. This was applied for considering each $HH.40$ as the temporal index. For each inclusive time-point, the measurements were added to their respective array in corresponding temporal index.

To extract features from the arrays, statistical properties of each measurement, such as mean, were calculated and stored as features. Next, the differences between consecutive measurements in each array were computed, providing an additional set of characteristics reflecting the temporal variance within the data. From these differences, their mean, median, standard deviation, minimum, maximum, and range were extracted. Subsequently, the first and second derivatives of each array are computed, encapsulating the information of the rate of change as features. This is followed by the calculation of mean values for both the first and second derivatives, further enriching the feature set. The first and second derivative mean and standard deviation can be expressed as:

$$\bar{\Delta} = \frac{1}{n-1} \sum_{i=1}^{n-1} \Delta x_i \tag{3}$$

$$\sigma_{\Delta} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} (\Delta x_i - \bar{\Delta})^2} \tag{4}$$

$$\bar{\Delta^2} = \frac{1}{n-2} \sum_{i=1}^{n-2} \Delta^2 x_i \tag{5}$$

$$\sigma_{\Delta^2} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n-2} (\Delta^2 x_i - \bar{\Delta^2})^2} \tag{6}$$

Finally, running window statistics were computed for each array. A convolution operation is applied to calculate rolling means and standard deviations over different window sizes, extracting the mean, standard deviation, minimum, maximum, and range of these rolling statistics where the mean and standard deviation of rolling mean of window size w can be expressed as:

$$\bar{x}_{i,w} = \frac{1}{w} \sum_{j=0}^{w-1} x_{i+j} \tag{7}$$

$$\sigma_{i,w} = \sqrt{\frac{1}{w} \sum_{j=0}^{w-1} (x_{i+j} - \bar{x}_{i,w})^2} \tag{8}$$

After feature arrays were used to generate over 400 features, initial arrays were removed and correlation-based feature selection (CFS) was applied for each applicable model. To avoid overfitting, a maximum of 15 features that passed a threshold were selected, where at most three derived features from the same measurement were used.

For the department prediction and an attempt to exam prediction, the process was the same except for target labels.

V. Results

For binary occupancy classification predicting whether a lecture is in session, over 10 million data points were used from 8 classrooms, of which 55% were labeled as '1'. Among the 35 models that were used to train, the best performing were Decision Tree Classifier and Random Forest Classifier. Considering that there is a possibility of inconsistency between measurements and labels such that some lectures being cancelled or took place somewhere else, it is a very promising result.

TABLE II
Model performances on binary occupancy

Model	Accuracy	MSE
Decision Tree Classifier	0.953	0.0469
Random Forest Classifier	0.963	0.0365
XGBoost	-	0.028
GradientBoosting	0.924	0.215
LightGBM	-	0.038
AdaBoost	0.843	0.038
CatBoost	0.948	0.038
BaggingClassifier	0.963	0.038
KNN	0.754	0.2454
Naive Bayes	0.712	0.287

In the subsequent stage, while the initial data-point count from the 8 classrooms exceeded 2400, we faced imbalanced datasets. Consequently, we decided to narrow our focus to two specific classrooms, from which balanced attendance data spanning over 1000 time points was derived. For multi-class classification, the Res_p was set to 5 for the first classroom and 15 for the second, in accordance with equation (1). Models were trained individually for each classroom. Models not supporting multi-class classification or those that underperformed in the previous stage were excluded from the evaluation. As presented in table III, where 'L' stands for the larger classroom, the results demonstrate the feasibility of predicting the number of individuals in a room with a resolution as fine as 5 with significant accuracy. For the top-performing model in the smaller classroom, a deeper dive into the misclassified data points indicated that the category difference was no greater than 2 for 88% of the misclassifications. Moreover, 64% of these misclassifications were time-indexed at HH.*40*,

suggesting they occurred at the beginning of a lecture, before the air conditions were influenced by the attendees.

TABLE III
MODEL PERFORMANCES ON TIME-SERIES MULTI-CLASS CLASSIFICATION

Model	Accuracy	Accuracy (L)
Decision Tree Classifier	0.85	0.78
Random Forest Classifier	0.89	0.77
Bagging Classifier	0.92	0.83
GradientBoosting	0.90	0.85
CatBoost	0.84	0.78
SVM	0.42	0.52
Logistic Regression	0.36	0.38
NN	0.40	0.43
kNN	0.73	0.75
Naive Bayes	0.42	0.42

For the next stages, we generated arrays to derive features with respect to changes in air measurements. In our last experiments, we only conducted the experiment in classroom in figure 1 which had 492 data points. Due to the limited training set, test split was reduced to 15%. After constructing the arrays with respect to temporal indices and applying correlation based feature selection as we explained in the methodology part, results were evaluated in various Res_p using accuracy as a metric in table IV. We posit that one potential reason for the observed improvement in accuracy with increasing Res_p values is the consequent enhancement in the balance of the dataset distribution. A more evenly distributed dataset can often lead to better model performance, as it reduces the risk of overfitting to a particular class.

TABLE IV
MODEL PERFORMANCES ON INTERVAL MULTI-CLASS CLASSIFICATION

Model	Accuracy		
	$Res_p = 15$	$Res_p = 12$	$Res_p = 10$
Decision Tree Classifier	0.615	0.56	0.54
Random Forest Classifier	0.63	0.47	0.53
Bagging Classifier	0.69	0.44	0.40
GradientBoosting	0.85	0.67	0.67
CatBoost	0.77	0.46	0.46
SVM	0.53	0.27	0.30
AdaBoost	0.66	0.43	0.49
NN	0.37	0.33	0.24
kNN	0.53	0.43	0.30
Naive Bayes	0.53	0.33	0.20

VI. CONCLUSION

Our experiments indicate that an attacker can exploit air measurements, or fluctuations in air data, to deduce information about the occupancy and approximate number of individuals in an enclosed space. Moreover, subsequent experiments suggest that one can also may leak the nature of an ongoing event with high accuracy, even when many unpredictable elements are present in the environment. These measurements are used and stored using a network in many IoT sensors in smart houses which also have less uncertainty in environment compared to a classroom. Nevertheless, it is also important to note that predicting occupancy through air measurements can have significant benefits such as building management systems that can optimize HVAC operations based on real-time occupancy, ensuring comfort while conserving energy. It can also bolster safety and emergency responses; knowing the occupancy can aid first responders in emergencies like fires or ensure optimal air quality during pandemics. Furthermore, facility managers can gain insights into space utilization, leading to efficient room bookings and improved architectural designs.

As a future work, it would be valuable to replicate these experiments with a larger dataset and aim for a more balanced distribution, especially for departmental predictions by selecting attendances that are as close in count as possible.

REFERENCES

[1] Bosch Sensortec, "BME680 Datasheet," [Online]. Available: https://www.boschsensortec.com/media/boschsensortec/downloads/datasheets/bst-bme680-ds001.pdf. [Accessed April 16, 2023].

[2] J. Dutta and S. Roy, "OccupancySense: Context-Based Indoor Occupancy Detection & Prediction Using CatBoost Model," Appl. Soft Comput., vol. 119, 2022, Art. no. 108536, doi: 10.1016/j.asoc.2022.108536.

[3] J. Dutta, F. Gazi, S. Roy, and C. Chowdhury, "AirSense: Opportunistic crowd-sensing based air quality monitoring system for smart city," in Proc. IEEE SENSORS, Orlando, FL, USA, 2016, pp. 1-3, doi: 10.1109/ICSENS.2016.7808730.

[4] J. Kröger, "Unexpected Inferences from Sensor Data: A Hidden Privacy Threat in the Internet of Things," in IFIP Advances in Inf. Commun. Technol., Jul. 2019, pp. 147–59, doi: 10.1007/978-3-030-15651-0_13.

[5] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 359–366.

[6] N. Nesa and I. Banerjee, "IOT-Based Sensor Data Fusion for Occupancy Sensing Using Dempster–Shafer Evidence Theory for Smart Buildings," IEEE Internet Things J., vol. 4, no. 5, pp. 1563–70, 2017, doi: 10.1109/jiot.2017.2723424.

[7] W. Zhang, Y. Wu, and J. K. Calautit, "A Review on Occupancy Prediction through Machine Learning for Enhancing Energy Efficiency, Air Quality and Thermal Comfort in the Built Environment," Renew. Sustain. Energy Rev., vol. 167, 2022, Art. no. 112704, doi: 10.1016/j.rser.2022.112704.

[8] L. Zimmermann, R. Weigel, and G. Fischer, "Fusion of Nonintrusive Environmental Sensors for Occupancy Detection in Smart Homes," IEEE Internet Things J., vol. 5, no. 4, pp. 2343–52, 2018, doi: 10.1109/jiot.2017.2752134.

[9] Y. Yang, Y. Yuan, T. Pan, X. Zang, and G. Liu, "A framework for occupancy prediction based on image information fusion and machine learning," Build. Environ., vol. 207, 2022, Art. no. 108524, doi: 10.1016/j.buildenv.2021.108524.