Synthetic Data Generation for Volatile Organic Compounds Recognition

Mahdia Ahmadi[1]
mahdia@ipb.pt

Felipe Merenda Izidorio[2]
felipeizidorio@ipb.pt

Ahmad Gamal Ibrahim[3]
ahmed@ipb.pt

Rui Pedro Lopes[4]
rlopes@ipb.pt

Mariam Jvarsheishvili[5]
mariam@ipb.pt

Caio Filipe Soares[6]
caiosoares@ipb.pt

Getúlio Igrejas[7]
igrejas@ipb.pt

Pedro João Rodrigues[8]
pjsr@ipb.pt

Research Centre in Digitalization and Intelligent Robotics (CeDRI)
Instituto Politécnico de Bragança, Portugal

Abstract

The fact that machine learning (ML) models to recognize volatile organic compounds (VOC) are typically developed with limited datasets and can be expensive to gather scaled sensor data is an obstacle in their development. The Bosch BME688 is a multi-gas sensor that can give detailed environmental data, but needs large experimental campaigns to construct representative data sets. To overcome this issue, we introduce a Python library on synthetic data generation to the BME688. The tool uses the Kernel Density Estimation (KDE) to generate an empirical gas resistance distribution according to various heater profiles and uses mathematical gas mixing to generate self-configurable multi-gas simulations. Experiments by validation on coffee and oil gases show that the resulting datasets retain the statistical characteristics of actual measurements, both at the stepwise level of gas resistance distributions and at the multivariate level with Principal Component Analysis (PCA). The library generates machine learning reproducible experimentation, machine learning algorithm prototyping on mixtures of percentages, and provision of systematic evaluation of VOC recognition systems. The contribution of the work is a modular and lightweight framework to address the problem of the lack of data, facilitate the reproducible research and speed up the creation of air quality monitoring solutions based on ML.

1 Introduction

The growing demand for reliable air quality monitoring and volatile organic compound (VOC) detection has increased interest in low-cost, portable gas sensors such as the Bosch BME688. These sensors provide multi-parameter environmental data, including temperature, humidity, pressure, and gas resistance, which can be exploited for machine learning (ML)-based classification of VOCs in applications such as indoor air quality, health diagnostics, and environmental monitoring [2, 3].

A significant barrier to the effective development of ML models in this domain is the scarcity and high cost of collecting large-scale, labeled sensor datasets. Experiments require controlled environments, repeatability across conditions, and long acquisition times, which limit the availability of representative data. Consequently, researchers and developers face difficulties in training, testing, and validating algorithms for VOC recognition.

To address this challenge, we developed a Synthetic Data Generator, a Python-based library that simulates realistic BME688 sensor outputs for multiple gases and heater profiles. The tool leverages Kernel Density Estimation (KDE) to model empirical distributions from collected measurements (e.g., coffee and oil vapors), and incorporates mathematical gas mixing to reproduce mixtures under configurable proportions. The generated datasets include both simulated sensor values and mixture-percentage columns, ensuring they are directly usable for ML workflows. This approach provides a reproducible, lightweight, and modular framework to overcome data scarcity, facilitate algorithm prototyping, and enable systematic testing of VOC recognition systems [1, 4].

2 Methodology

The proposed methodology consisted of five main stages: data acquisition, device adaptation, dataset construction, library development, and validation. The BME688 sensor, a multi-gas device capable of measuring temperature (*temp_c*), relative humidity (*humidity_pct*), atmospheric pressure (*pressure_hpa*), and gas resistance (*gas_resistance_ohm*), was employed as the core sensing platform. Its configurable heater profiles, defined by sequences of heating temperatures, influence the sensitivity to different gases, while the *measurement step* attribute specifies the exact point in the heating cycle at which the data is recorded, providing finer control of the sensing process.

Data acquisition was conducted in a controlled environment using sealed plastic containers, where the sensor was alternately exposed to coffee vapors and oil vapors. This approach ensured repeatability across experiments and minimized uncontrolled environmental interference.An example of the experimental setup is shown in Fig. 1.

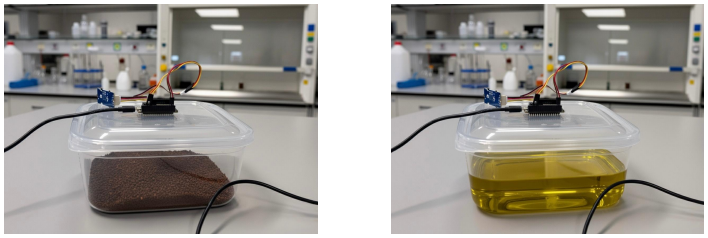

Figure 1: Experimental setup for data acquisition with the BME688 sensor inside a sealed container with coffee and olive oil.

To extend its applicability, the sensing module was adapted with a LoRa communication interface, enabling wireless transmission and supporting scenarios of remote monitoring. The collected measurements were then organized into structured datasets that combined environmental conditions with sensor-specific parameters. Each entry included metadata such as the heater profile and measurement step, offering a detailed representation of the sensing conditions.

Building on this dataset, a Python library was implemented to generate synthetic samples. The library allows users to define inputs such as gas type, heater profile, and number of samples, and it also incorporates mathematical gas mixing, where multiple gases can be combined through weighted formulas to simulate realistic mixtures. The generated datasets are machine learning–ready, including both synthetic sensor values and numerical percentage columns for each gas in the mixture.

Finally, validation of the synthetic data was performed using statistical distribution analysis and Principal Component Analysis (PCA). These methods were employed to evaluate similarity and consistency between synthetic and real measurements, ensuring that the generated datasets accurately preserved the statistical properties of empirical data.

3 Results and Discussion

The synthetic data generator was validated by comparing the statistical characteristics of synthetic data to empirical data of two representative volatile organic compounds: coffee and oil, using the identical heater profile (HP-354).

Figures 2 and 3 indicate the distribution of the gas resistance of 10 steps of the measurement of both original and synthetic data. The synthetic distributions (orange) in the case of coffee effectively model the

empirical histograms (blue) in terms of the overall shape, mean and variance of the sensor responses. Even though small deviations can be noticed in the tails of some of the steps, the KDE-based modeling made sure that the produced samples had the same multimodal properties as the original distributions. The same case can be said about the oil: the synthetic datasets recreate the peak structure and variance of the empirical data, although the density curves are a bit less smooth because of the statistical approximation. These findings show that the generator preserves the step-dependent sensitivity patterns of the BME688 sensor that are important in the VOC recognition tasks.

To further assess global similarity, the Principal Component Analysis (PCA) was used on both the original and the synthetic data (Fig. 4). For coffee vapors, the two datasets occupy overlapping areas in the reduced feature space, indicating that the generator successfully reproduces the dominant directions of variance of the original data. Moreover, the centroids of both datasets were computed, and their Euclidean distance was found to be zero, meaning that their global centers of mass coincide perfectly. For oil vapors, synthetic samples align with the general distribution of empirical data, although some dispersion occurs along the second major component. This suggests that while the dominant statistical structure is well captured, subtle fine-grained variability present in the original data is not entirely reproduced.

Generally, the outcomes indicate that the suggested library creates artificial datasets, which are close to the statistical behavior of actual VOC measurements at the single distribution level and at the multivariate space. This is a combination of realism and controllability, which makes the tool highly applicable in machine learning applications where reproducibility and the possibility to simulate other scenarios are also more useful than the ability to replicate empirical noise.

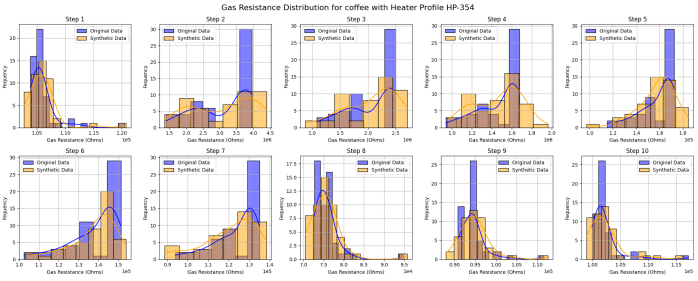

Figure 2: Coffee gas resistance distributions across 10 steps (HP-354): original (blue) vs. synthetic (orange).

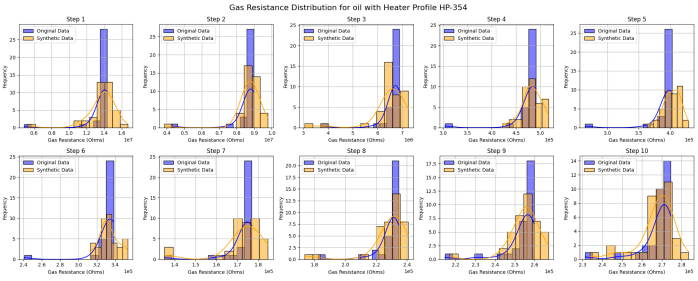

Figure 3: Oil gas resistance distributions across 10 steps (HP-354): original (blue) vs. synthetic (orange).

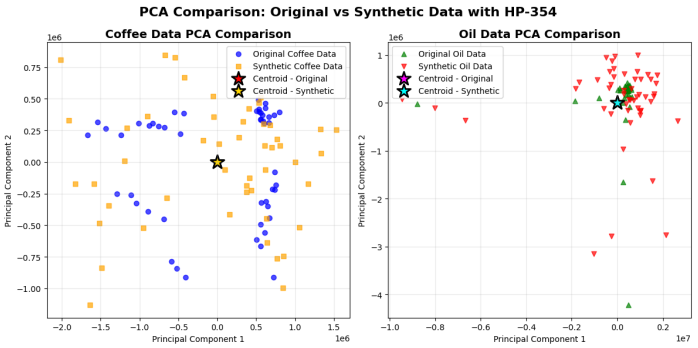

Figure 4: PCA comparison of original vs. synthetic data for coffee (left) and oil (right) under HP-354.

4 Conclusion and Future Work

This work presented a Python-based library for generating synthetic data from BME688 gas sensors, addressing the scarcity of labeled datasets in sensor-based research. By modeling empirical distributions with Kernel Density Estimation (KDE) and supporting mathematical gas mixing under configurable heater profiles, the tool provides realistic, flexible, and machine learning-ready datasets that approximate the statistical properties of real measurements.

The generator enables diverse applications across research and development, facilitating rapid algorithm prototyping without extensive data collection campaigns and enabling simulation of edge cases and rare events. In machine learning contexts, the library produces structured outputs with explicit percentage columns for each gas, directly supporting supervised learning tasks, balanced dataset creation, and controlled model evaluation. For system validation, it enables reproducible testing of calibration strategies, stress testing under hypothetical environmental conditions, and performance benchmarking. The flexible mixing ratios support exploration of complex multi-gas interactions that are difficult to replicate experimentally.

Beyond research and engineering, the tool provides educational value by lowering barriers for students and practitioners to experiment with data analysis, pattern recognition, and machine learning techniques in gas sensing contexts. This supports reproducible experimentation, reduces dependency on costly acquisition campaigns, and accelerates VOC recognition system development.

Future work will focus on three key directions: expanding the dataset with additional VOCs, incorporating dynamic environmental conditions to capture temporal variations, and validating the synthetic data quality through comprehensive evaluation across diverse machine learning models. These improvements will enhance realism and broaden applicability, supporting more sophisticated machine learning pipelines and enabling large-scale deployment in air quality monitoring and machine olfaction systems.

Acknowledgments

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: 2024.07316.IACDC/2024 and by national funds: UID/05757 – Research Centre in Digitalization and Intelligent Robotics (CeDRI); and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

References

[1] André Goncalves, Prithwish Ray, Darren Soper, and et al. Generation and evaluation of synthetic data for machine learning applications: a survey. *Journal of Big Data*, 7(1):1–48, 2020. doi: 10.1186/s40537-020-00353-5.

[2] Markus Rupp, Andreas Fuchs, and et al. Low-cost air quality measurement system for urban environments based on bme680 sensor. *Sensors*, 20(17):4811, 2020. doi: 10.3390/s20174811.

[3] Bosch Sensortec. Bme688 environmental sensor: Gas sensor with ai. `https://www.bosch-sensortec.com/products/environmental-sensors/gas-sensors/bme688/`, 2021.

[4] Lei Xu, Yufeng Liu, Yisen Zhou, and et al. A comprehensive survey on tabular data synthesis. *arXiv preprint arXiv:2310.12491*, 2023.