DOI 10.5162/17dss2024/P26

Robust Indoor Air Quality Monitoring: Out-of-Distribution Detection using Ensemble Neural Networks

Payman Goodarzi, Dennis Arendes, Christian Bur, Andreas Schütze

Lab for Measurement Technology, Saarland University, Saarbrücken, Germany

(p.goodarzi, d.arendes, c.bur, schuetze) @lmt.uni-saarland.de

Abstract

Data-driven indoor air quality (IAQ) monitoring systems have demonstrated strong performance; however, detecting out-of-range data is essential for reliable monitoring. This study proposes an out-of-distribution (OOD) detection method to identify out-of-range conditions and temporal drift in real-time applications. Our approach utilizes an ensemble of convolutional neural networks (CNNs) optimized via Bayesian hyperparameter tuning. The method achieved robust results, with an area under the receiver operating characteristic curve (AUC) of 93% for out-of-range gas detection and AUCs of 95% and 99% for identifying temporal drift at six and ten weeks post-calibration, respectively. Integrating this method into real-time IAQ monitoring systems enhances model reliability under real-world conditions.

Introduction

Indoor air quality (IAQ) measurement and the detection of volatile organic compounds (VOCs) are essential for healthy indoor air and accurate demand-controlled ventilation [1]. This can be achieved with low-cost sensor systems based on metal oxide semiconductor (MOS) gas sensors which make use of advanced operating modes, such as temperature-cycled operation, and machine learning (ML) to evaluate the complex sensor response [2].

Domain shift is a key challenge in real-world ML applications [3], particularly in IAQ monitoring with environmental variations. It occurs when the underlying data distribution changes due to factors such as sensor drift, exposure to gas concentrations beyond the calibration range, exposure to gases not included during calibration, or sensor poisoning [4]. This issue is intensified by limited observations or the influence of covariates on data distributions [5]. Despite sophisticated calibration processes for gas sensors, the number of gases and the range of gas concentrations used during calibration is often restricted [6]. As a result, models may encounter out-of-distribution (OOD) inputs, data outside the calibration range, that can lead to inaccurate predictions.

Dealing with domain shift is a challenging task. Several approaches can address this issue, such as transfer learning and domain adaptation [7]. However, these methods require data from the new working conditions, such as a different sensor, range, or gases, and are typically limited to those specific target conditions [8]. As a result, an additional model is often needed to monitor the supervised ML model's performance after deployment. There are several techniques available for validating the predictions of ML models, including uncertainty estimation [9], extrapolation detection [10], anomaly detection [11], and OOD detection [12].

OOD detection is an important technique for validating ML models in real-world deployment scenarios. A key advantage of OOD detection methods is their model-agnostic nature, often requiring no modifications to the underlying predictive model [12]. Numerous approaches have been proposed, particularly for neural networks (NNs), representing a state-of-the-art model class in a wide range of ML tasks [13].

Early OOD detection methods primarily leveraged the softmax confidence score in classification settings, where in-distribution (ID) samples typically yield higher predicted class probabilities compared to OOD samples [14]. Later, subsequent work introduced more sophisticated techniques [15]. Lakshminarayanan et al. proposed a robust and generalizable approach based on deep ensembles, which extends beyond classification tasks [16]. This method exploits the variance in predictions from an ensemble of models: ID samples tend to produce consistent predictions across models, while OOD samples demonstrate greater predictive variance.

Materials and Methods

Dataset

The IAQ dataset [2] used in this study simulates a complex mixture of typical indoor air conditions. It includes various VOCs, along with hydrogen, carbon monoxide, and relative humidity as background interference gases. The recorded signals were obtained from a low-cost system utilizing SGP30 sensors (Sensirion AG, Stäfa, Switzerland), which are equipped with four gas-sensitive layers [2]. The output signals represent the resistance of these four layers over time. Each observation of gas concentrations consists of 4 × 2400 data samples, collected at a sampling rate of 20 Hz. We treated each of the four layers as an independent sensor, thus framing this as a multi-sensor dataset. Besides normal gas concentration ranges, the dataset includes extended concentrations of acetone, ethanol, toluene, and hydrogen beyond typical levels.

In this study, we simulated OOD scenarios by applying out-of-range gas concentrations and drifted signals to the ML model. Out-of-range concentrations refer to extended VOC levels beyond typical ranges. We defined ID conditions as instances where all VOCs in the dataset fall within the normal concentration of their respective ranges. Observations outside this threshold were classified as OOD samples.

The dataset was recorded in three calibration phases: the initial calibration phase lasted one week, followed by the first recalibration after four weeks of field testing, and the second recalibration three weeks later (Figure 1). In this study, the first and second recalibration phases are treated as potential drifted data.

Figure 1. The complete experiment over ten weeks, including calibration phases and field tests.

In this study, acetone was selected as the target gas for the regression task. An observation is classified as ID if:

$$x_i \notin E_i \quad \forall i$$

where x_i represents the concentration of $VOC(i)$, and E_i represents the extended concentration range of $VOC(i)$. Figure 2 illustrates the concentration ranges for acetone, ethanol, and formaldehyde. The vertical lines indicate the boundaries of the normal concentration ranges.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are powerful ML models that have produced outstanding results across various applications. In particular, 1D-CNNs are theoretically well-suited for signal processing tasks due to their inherent ability to filter temporal data [17]. In this study, we use a configurable CNN architecture, where the number of convolutional layers, kernel sizes, and number of filters are adjustable [18]. These architecture hyperparameters (HP) are optimized through a Bayesian hyperparameter (HP) tuning process. Figure 3 illustrates the CNN architecture, where each convolutional (conv) block includes a convolutional layer, batch normalization, and ReLU activation.

Ensemble-based Out-of-distribution Detection

The uncertainty in an ensemble of models is a strong indicator for detecting OOD samples [16]. Intuitively, the prediction variance between different models is lower when the input data is from the same distribution as the training data. The initial method proposing the use of ensemble neural networks (deep ensembles) relies on variations due to random initialization in each model. Since the initial weights of neural networks are randomly assigned, the weights of the trained networks differ, resulting in each model in the ensemble learning slightly different representations of the data. As a result, the

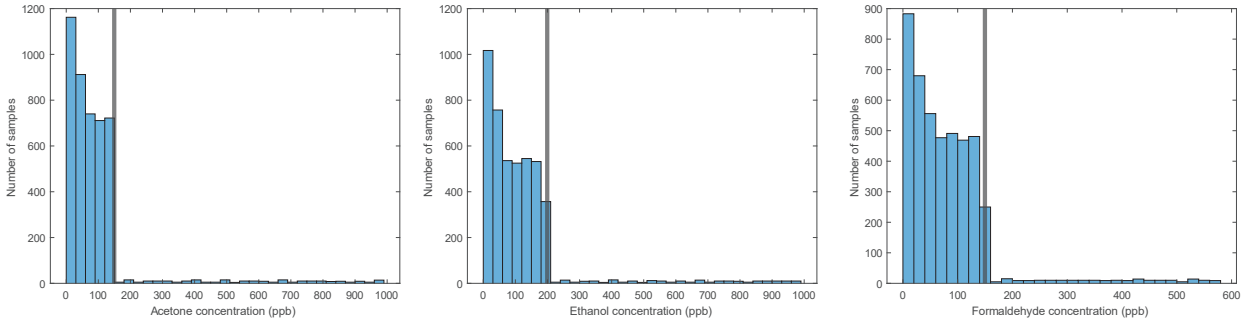

Figure 2. Distributions of acetone, ethanol, and formaldehyde in the dataset. Vertical lines indicate the upper limits of the normal concentration ranges, with higher concentrations considered as out-of-range conditions.

ensemble can capture diverse predictions for OOD samples, which often results in higher prediction variance compared to ID samples. It has been shown that increasing model diversity through data selection and varying architecture hyperparameters can enhance the estimation of prediction uncertainty [19], [20]. In this study, we treat each observation as four separate sensor signals. We conduct 50 trials of Bayesian HP tuning for each sensor individually and, finally, construct the ensemble by aggregating the predictions from the 10 best-performing models. The optimal number of models for the ensemble is determined based on the ID validation accuracy.

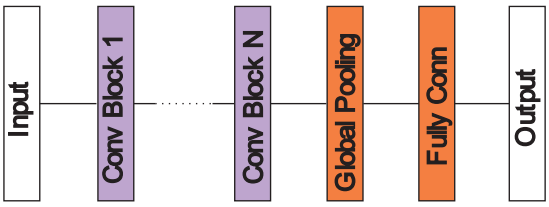

Figure 3. Parametric CNN architecture. The convolutional (conv) block consists of a convolutional layer, batch normalization layer, and ReLU activation function [18].

During training, the ID data is divided into three subsets: 70% of the observations are used as training data, 10% as validation, and 20% as test data. The performance of the models in HP tuning and training are evaluated based on the ID validation accuracy.
Prediction uncertainty can be estimated using various methods, such as calculating the variance of predictions across multiple models. In this study, we apply a distance-based method to detect anomalies in the predictions. Specifically, we use the predictions as feature inputs for a k-nearest neighbors (kNN) approach to identify OOD samples [21]. The results from all models are fused at the decision level, with a maximum of 4 × 10 features for each observation.

Metrics

The area under the receiver operating characteristic curve (AUC) is a commonly used metric to assess the accuracy of OOD detection methods [22]. As a threshold-independent metric, an AUC of 100% indicates a perfect classifier. To further evaluate the OOD detection model, we also report the false positive rate (FPR) at a true positive rate (TPR) of 95%, a metric referred to as FPR95.

Results and Discussion

Figure 4 illustrates the prediction results of the ensemble model, which combines 40 distinct networks and achieves a root mean square error (RMSE) of less than 12 ppb on the test data. The ensemble model achieves accuracy comparable to state-of-the-art methods [23], despite a limited number of observations, as many samples in the designed scenario are classified as OOD data.
By applying kNN (k = 5) to the 40 predictions, we can classify ID and OOD data effectively. The 5NN score distinguishes test ID data from OOD data with AUC of 93%, indicating strong predictive capability. For this OOD scenario, FPR95 is 20%.
The method effectively detects temporal drift, achieving an AUC of 95% when distinguishing test data from data collected after six weeks, and 99% for data collected after ten weeks. The FPR95 for these cases is 18% and 3%, respectively. Table 1 provides a summary of results for the designed OOD detection scenarios.

Table. 1: OOD detection results

Scenario	AUC (%)	FPR95 (%)
Out-of-range inputs	93	20
Drift after 6 weeks	95	18
Drift after 10 weeks	99	3

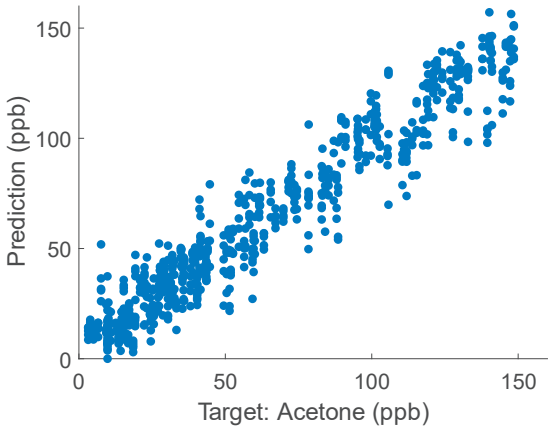

Figure 4. Ensemble model predictions for the ID test data.

Conclusion

In this work, we presented an effective approach for detecting OOD data in IAQ monitoring applications. Our method accurately identifies out-of-range conditions and temporal drift in sensor data through an ensemble model constructed with Bayesian hyperparameter tuning. The ensemble model demonstrated reliable prediction of ID data, achieving high

AUC values in OOD detection with a 93% AUC for out-of-range data and AUCs of 95% and 99% for data collected after six and ten weeks, respectively. Integrating this model into real-time IAQ monitoring systems promises enhanced reliability under dynamic, real-world conditions.

Future work could extend this framework to include additional target gases and assess model performance across different gas types. Additionally, studies on optimizing the ensemble size and selection method could further refine the model's effectiveness.

References

[1] World Health Organization, "Combined or multiple exposure to health stressors in indoor built environments: an evidence-based review prepared for the WHO training workshop 'Multiple environmental exposures and risks': 16–18 October 2013, Bonn, Germany," World Health Organization. Regional Office for Europe, 2014.

[2] T. Baur, J. Amann, C. Schultealbert, and A. Schütze, "Field Study of Metal Oxide Semiconductor Gas Sensors in Temperature Cycled Operation for Selective VOC Monitoring in Indoor Air," *Atmosphere (Basel)*, vol. 12, no. 5, p. 647, May 2021, doi: 10.3390/atmos12050647.

[3] P. Goodarzi, A. Schütze, and T. Schneider, "Comparison of different ML methods concerning prediction quality, domain adaptation and robustness," *Technisches Messen*, vol. 89, no. 4, pp. 224–239, 2022, doi: doi:10.1515/teme-2021-0129.

[4] D. Arendes, Y. Robin, J. Amann, A. Petto, A. Schütze, and C. Bur, "Transfer Learning Between Two Different Datasets of MOS Gas Sensors," in *2024 IEEE International Symposium on Olfaction and Electronic Nose (ISOEN)*, IEEE, May 2024, pp. 1–3. doi: 10.1109/ISOEN61239.2024.10556179.

[5] Z. Zhang, Z. Zhao, X. Zhang, C. Sun, and X. Chen, "Industrial anomaly detection with domain shift: A real-world dataset and masked multi-scale reconstruction," *Comput Ind*, vol. 151, p. 103990, 2023.

[6] T. Baur, M. Bastuck, C. Schultealbert, T. Sauerwald, and A. Schütze, "Random gas mixtures for efficient gas sensor calibration," *Journal of Sensors and Sensor Systems*, vol. 9, no. 2, pp. 411–424, Nov. 2020, doi: 10.5194/jsss-9-411-2020.

[7] W. M. Kouw and M. Loog, "An introduction to domain adaptation and transfer learning," *arXiv preprint*, Dec. 2018, doi: 10.48550/arXiv.1812.11806.

[8] I. Gulrajani and D. Lopez-Paz, "In Search of Lost Domain Generalization," *arXiv preprint arXiv:2007.01434*, Jul. 2020, doi: 10.48550/arXiv.2007.01434.

[9] A. Loquercio, M. Segu, and D. Scaramuzza, "A General Framework for Uncertainty Estimation in Deep Learning," *IEEE Robot Autom Lett*, vol. 5, no. 2, pp. 3153–3160, Apr. 2020, doi: 10.1109/LRA.2020.2974682.

[10] D. Madras, J. Atwood, and A. D'Amour, "Detecting extrapolation with local ensembles," in *International Conference on Learning Representations*, 2019. doi: 10.48550/arXiv.1910.09573.

[11] A. A. Cook, G. Mısırlı, and Z. Fan, "Anomaly Detection for IoT Time-Series Data: A Survey," *IEEE Internet Things J*, vol. 7, no. 7, pp. 6481–6494, 2020, doi: 10.1109/JIOT.2019.2958185.

[12] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized Out-of-Distribution Detection: A Survey," *Int J Comput Vis*, Jun. 2024, doi: 10.1007/s11263-024-02117-4.

[13] J. Gawlikowski *et al.*, "A survey of uncertainty in deep neural networks," *Artif Intell Rev*, vol. 56, no. S1, pp. 1513–1589, Oct. 2023, doi: 10.1007/s10462-023-10562-9.

[14] D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," in *International Conference on Learning Representations*, 2022. doi: 10.48550/arXiv.1610.02136.

[15] J. Yang *et al.*, "OpenOOD: Benchmarking Generalized Out-of-Distribution Detection," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Curran Associates, Inc., 2022, pp. 32598–32611. doi: 10.48550/arXiv.2210.07242.

[16] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. doi: 10.48550/arXiv.1612.01474.

[17] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech Syst Signal Process*, vol. 151, p. 107398, 2021, doi: 10.48550/arXiv.1905.03554.

[18] P. Goodarzi, A. Schütze, and T. Schneider, "Comparing AutoML and Deep Learning Methods for Condition Monitoring using Realistic Validation Scenarios," *arXiv preprint*, 2023, doi: 10.48550/arXiv.2308.14632.

[19] S. Zaidi, A. Zela, T. Elsken, C. C. Holmes, F. Hutter, and Y. Teh, "Neural Ensemble Search for Uncertainty Estimation and Dataset Shift," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 7898–7911. doi: 10.48550/arXiv.2006.08573.

[20] F. Wenzel, J. Snoek, D. Tran, and R. Jenatton, "Hyperparameter ensembles for robustness and uncertainty quantification," *Adv Neural Inf Process Syst*, vol. 33, pp. 6514–6527, 2020, doi: 10.48550/arXiv.2006.13570.

[21] Y. Sun, Y. Ming, X. Zhu, and Y. Li, “Out-of-distribution detection with deep nearest neighbors,” in *International Conference on Machine Learning*, 2022, pp. 20827–20840. doi: 10.48550/arXiv.2204.06507.

[22] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognit*, vol. 30, no. 7, pp. 1145–1159, 1997, doi: 10.1016/S0031-3203(96)00142-2.

[23] Y. Robin, J. Amann, P. Goodarzi, T. Schneider, A. Schütze, and C. Bur, “Deep Learning Based Calibration Time Reduction for MOS Gas Sensors with Transfer Learning,” *Atmosphere (Basel)*, vol. 13, no. 10, p. 1614, Oct. 2022, doi: 10.3390/atmos13101614.