2026 European Conference on Computing in Construction
Corfu, Greece
July 12–15, 2026

ANOMALY DETECTION IN INDOOR ENVIRONMENTAL SENSOR DATA BASED ON A DIGITAL TWIN

Arezoo Fathollahizenooz, Cornelius Preidel, and Simon Vilgertshofer
HM Munich University of Applied Sciences, Munich, Germany

Abstract

Anomaly detection in building sensor networks often relies on static thresholds that fail to capture temporal dependencies and system changes. This work combines a baseline model of expected indoor environmental behavior with anomaly detection to form a data-driven Digital Twin. In this framework, a regression model learns expected behavior from historical data, while deviations are quantified as multivariate residuals and analyzed using unsupervised and semi-supervised methods. Preliminary analysis on a representative campus sensor shows detection of anomalous behavior and associated distributional shifts, with sensor relocation serving as an illustrative system change.

Introduction

Sensor networks are used in modern buildings to monitor indoor environmental quality (IEQ). However, due to the scale of the data, manual inspection is insufficient to detect anomalies. Consequently, research has shifted toward data-driven anomaly detection. Unsupervised and semi-supervised methods are used to detect IEQ deviations without the need for large labeled datasets (Chen et al., 2023; Bi et al., 2024). Although methods such as Long Short-Term Memory can capture complex temporal dynamics, they often lack an integrated analysis of normal system operation (Wei et al., 2022; Rajabi and McArthur, 2023). To address this gap, this work investigates whether DT-based anomaly detection can identify environmental irregularities in indoor sensor data. Residual-based anomaly detection is integrated with a Digital Twin (DT) representing expected indoor environmental behavior, while anomalies are identified as deviations between physical observations and digital expectations, enabling detection of patterns suggestive of HVAC malfunctions or sustained humidity conditions associated with increased mold risk.

Dataset

The proposed framework incorporates indoor environmental data from a campus-wide sensor infrastructure that offers ongoing time-series data from various rooms in university buildings. Temperature, humidity, CO_2 concentration, illumination, and motion are included in the dataset. The possibility of sensor relocation, which adds uncertainty about the spatial context, is an important feature of this dataset.

Digital Twin

The physical layer in DT represents the building space, while sensors measuring indoor environmental variables serve as the interface between the physical asset and its digital representation. The digital layer consists of the recorded time-series data that describe the state of the system. A Random Forest regression model (Breiman, 2001) serves as the analysis component of the Digital Twin, producing predictions $\hat{y}_t = f_{\mathrm{RF}}(\mathrm{x}_t)$, where x_t comprises time-based and lagged features capturing temporal dependencies and non-linear relationships. Hyperparameters (n_estimators = 300, max_depth = 10, min_samples_leaf = 5) are selected with GridSearchCV with 5-fold time-series cross-validation, and the predictive performance is reported using MAE and R^2 on the final 20% of baseline data. This coupled physical-digital representation enables detection of distributional shifts indicative of structural changes, such as sensor relocation, while providing a learned baseline that adapts as system conditions evolve.

Residual-Based Anomaly Detection

Anomaly detection operates on residuals $r_t = y_t - \hat{y}_t$, where y_t are the observed measurements and $\hat{y}_t$ is the DT prediction at time t. Multivariate residuals represent these deviations, enabling identification of abnormal behavior relative to learned expectations. Anomaly detection is implemented using two methods: Unsupervised approaches based on Isolation Forest (Liu et al., 2008) are applied to identify deviations from learned temporal patterns in multivariate sensor data without requiring labeled anomalies. In addition, a semi-supervised setting using One-Class Support Vector Machines (OCSVM) (Schölkopf et al., 2001) is explored. For OCSVM training, domain knowledge determines inlier conditions (20–25 °C temperature, 20–60% humidity), whereas observations outside these ranges are used to assess anomaly-detection behavior.

Results and Discussion

Period segmentation is applied in the initial analysis of a particular sensor (S12) from June to October 2023. Gaps longer than six hours, which are unlikely under normal operations, are used as indicators of potential system disruption and divide the data into baseline (Jun–Jul; 56,738 measurements), relocation/sensor outage (July–October; 5,510 measurements), and post-relocation (October; 3,884 measurements) as shown in figure 1. The trained multivariate DT baseline model exhibits substantial increases

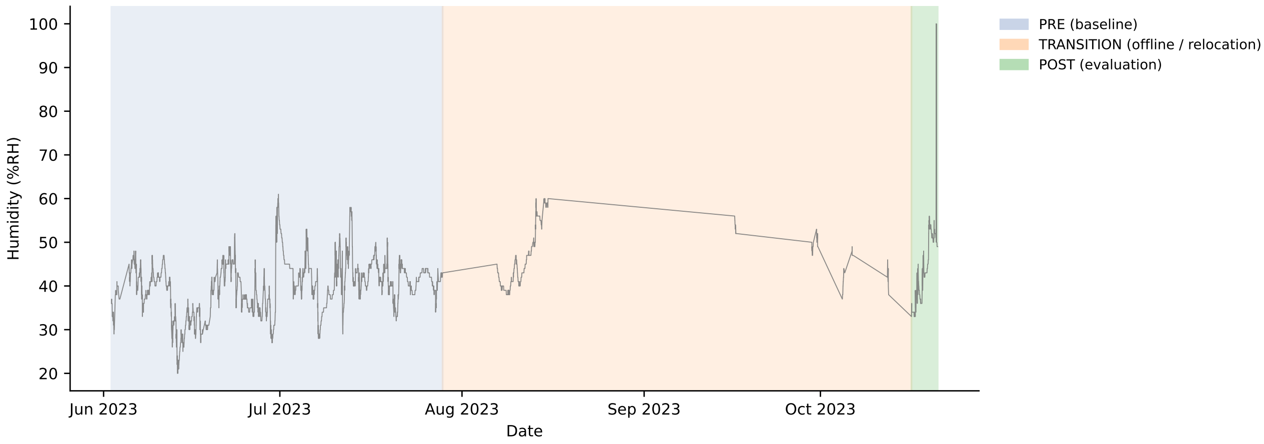

Figure 1: Period segmentation based on sensor gaps

in prediction error in all variables (temperature, humidity, and CO_2) when applied to post-relocation data, confirming a significant shift in distribution. Figure 2 shows this quantitative degradation by a clear shift in the residual distributions across the variables. During the first unsupervised phase of this framework, an Isolation Forest model identified every post-relocation observation as anomalous, indicating a distributional shift. Next, semi-supervised learning methods will be used to enhance the classification of these deviations. Although a representative, operationally stable sensor (S12) is selected from the network of 82 sensors to illustrate the framework, extension to the full sensor network is part of ongoing work.

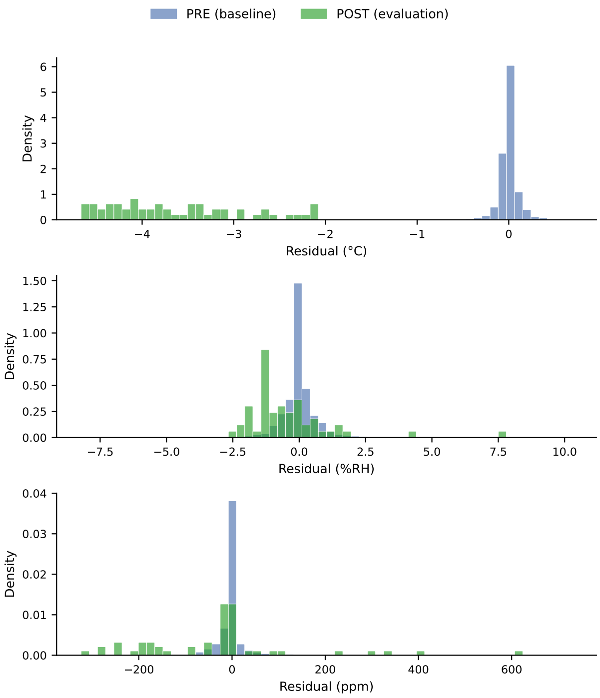

Figure 2: Residual distribution shift across variables

Summary

This research presents a DT-based framework for anomaly detection in indoor environmental sensor data. A Random Forest baseline learns expected behavior from historical data, and residual analysis detects distributional shifts with Isolation Forest. Semi-supervised OCSVM for environmental anomaly detection and adaptive DT updating are the next steps in ongoing work.

References

Bi, J., Wang, H., Yan, E., Wang, C., Yan, K., Jiang, L., and Yang, B. (2024). AI in HVAC fault detection and diagnosis: A systematic review. Energy Review, 3.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1):5–32.

Chen, Z., O'Neill, Z., Wen, J., Pradhan, O., Yang, T., lu, X., Lin, G., Miyata, S., Lee, S., Shen, C., Chiosa, R., Piscitelli, M., Capozzoli, A., Hengel, F., Kührer, A., Pritoni, M., Liu, W., Clauß, J., Chen, Y., and Herr, T. (2023). A review of data-driven fault detection and diagnostics for building HVAC systems. Applied Energy, 339:121030.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation Forest. In 2008 Eighth IEEE International Conference on Data Mining.

Rajabi, F. and McArthur, J. J. (2023). Unsupervised Learning for Fault Detection of HVAC Systems: An optics -based approach for terminal air handling units.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural Computation.

Wei, Y., Jang-Jaccard, J., Xu, W., Sabrina, F., Camtepe, S., and Boulic, M. (2022). LSTM-Autoencoder based Anomaly Detection for Indoor Air Quality Time Series Data.