2025 European Conference on Computing in Construction

CIB W78 Conference on IT in Construction

Porto, Portugal
July 14-17, 2025

SELF-SUPERVISED LEARNING FOR OCCUPANT ACTIVITY RECOGNITION IN BUILDING ENVIRONMENTS USING BMS DATA

Haoxi Wei, Xiang Xie, Sneha Verma, Philip James, and Mohamad Kassem
School of Engineering, Newcastle University, Newcastle upon Tyne, UK

Abstract

Balancing energy efficiency and occupant comfort is key to maintaining the sustainability of buildings. Understanding occupant activities is essential for optimising energy use without compromising comfort. This paper proposes a self-supervised learning approach for recognising occupant activity patterns using indoor environmental data from the Building Management System (BMS). A modified Transformer Masked Autoencoder (Ti-MAE) is adopted to extract latent representations of data, followed by the K-means Clustering Algorithm for clustering typical occupant activity patterns. Experiments using real-life building data demonstrate its robust performance in occupant activity recognition, even without specific sensors. The approach optimises energy efficiency while preserving privacy.

Introduction

Buildings are essential to human life, with people spending nearly 80 % of their time indoors (Fernández, 2007). Indoor Environmental Quality (IEQ), including thermal comfort, air quality and acoustic conditions, is critical to health and productivity (Geng et al., 2017). Meanwhile, buildings contribute profoundly to global energy consumption and associated carbon emissions, primarily due to electricity use for lighting, heating, cooling, ventilation, and other building services (Shoemaker, 2023). However, reducing energy consumption often compromises occupant comfort, as conventional building control systems rely on predetermined schedules and setpoints (Naylor et al., 2018), failing to accommodate the diverse and dynamic needs of occupants. This trade-off between energy efficiency and occupant comfort has driven the emergence of the occupant-centric control strategy, designed to create a more adaptive and responsive balance based on individual behaviour or collective activities. The proliferation of sensors and BMS in modern buildings has enabled finer-grained data collection for human behaviour identification. In fact, occupant activity plays a dominant role in shaping indoor environmental conditions, making indoor sensor data a highly indicative source for detecting human activities (Al Horr et al., 2016).

Based on massive time series data collected from building sensors, machine learning and deep learning models have become promising approaches for identifying occupant behaviour. Although these models can achieve sufficient accuracy in occupant activity recognition, they typically rely on large-scale labelled datasets for training (Chen et al., 2021). Labelling occupant behaviour data is often impractical, as it can disrupt daily activities and raise significant privacy concerns (Kanthila et al., 2021). Meanwhile, BMS continuously generates unlabelled data, highlighting the need for methods to learn from such data without manual labelling.

Unsupervised and self-supervised learning offers a viable alternative, which learns latent representations of unlabelled data that can effectively distinguish occupant activity patterns. Despite their potential, these approaches remain underexplored due to training instability, resulting from collapsing representations, sensitivity to hyper-parameters, and difficulties in maintaining consistent gradient updates (Wang and Biljecki, 2022). However, in most buildings, broader activity patterns can already provide sufficient contextual information for effective decision-making. Moreover, the historical data collected from various sensors during building operations mitigates this limitation by enabling robust pattern analysis. This paper proposes a self-supervised learning approach to identify occupant activities using unlabelled and sometimes corrupted BMS data. The approach captures robust latent representations that indicate occupant activity and distinguish different activity patterns without relying on extensive manual labelling.

The remainder of the paper is organised as follows: Section 2 reviews research on human activity recognition using unsupervised or self-supervised approaches and sensor-based building behaviour/activity recognition; Section 3 presents the methodology of the proposed self-supervised approach; Section 4 analyses the results based on real-world cases; finally, Sections 5 and 6 summarise the implications and outlines future research directions.

Literature review

Unsupervised learning autonomously identifies patterns in data without the need for labelled examples, which makes it particularly useful when labelled data is limited or expensive to obtain. Clustering is widely used as an unsupervised technique for Human Activity Recognition (HAR). Ariza-Colpas et al. (2022) review the application of clustering in HAR, highlighting the vital role of feature extraction and cluster validation in keeping recognition accurate. Algorithms, such as K-means and DBSCAN, have proven effective for grouping similar activities based on sensor data. While their performance largely depends on feature quality and the selection of similarity metrics, dimensionality reduction techniques like Principal Component Analysis (PCA) are valuable for noise reduction and improving clustering accuracy.

With the advancement of machine learning, unsupervised generative models like Autoencoder (AE), Generative Adversarial Network (GAN), and Convolutional Neural Network (CNN) have been applied for accurate HAR. Almaslukh et al. (2017) demonstrate that using AEs to learn a compact representation of sensor data allows raw data compression and reconstruction to classify typical activities. Not necessarily using environmental data, Abedin et al. (2021) applied GANs to human behaviour recognition using wearable device data, achieving performance comparable to supervised learning models. Similarly, Ordóñez and Roggen (2016) use CNN to capture temporal dependencies in wearable device data, enabling HAR in home environments.

Self-supervised learning, a unique form of unsupervised learning, generates supervisory 'signals' from data by designing predefined auxiliary tasks. These models demonstrate strong potential in minimising the reliance on labelled data while achieving state-of-the-art performance. Contrastive learning approaches, such as SimCLR and MoCo, learn discriminative representations by comparing similar and dissimilar pairs of behavioural data. Haresamudram et al. (2022) demonstrate the effectiveness of contrastive learning in HAR using wearable sensor data, showing improved robustness and generalisability with limited annotations. However, a recent study by Ek et al. (2024) found that comparative learning, as well as predictive learning (e.g., Data2Vec), struggle with frozen feature extraction, whereas generative learning, such as Masked Autoencoder (MAE), performs better across different datasets, and integrates well with architectures like CNNs. Further, studies validate that transformer-based models (e.g., MAE) outperform other architectures in processing time series data, as they process entire time-series data in parallel while capturing long-range dependencies (Dirgová Luptáková et al., 2022). Their ability to model complex temporal patterns makes transformers particularly effective for handling heterogeneous sensor data with missing or corrupted values.

Focusing on building occupant behaviour and activity recognition, the use of environmental and wearable sensor data is gaining popularity. Nesa and Banerjee (2017) propose a flexible framework for combining various data (e.g., temperature, humidity, CO_2) to enable occupancy detection based on Dempster-Shafer evidence theory. Anguita-Molina et al. (2025) developed a deep learning-based multi-occupancy activity recognition model by integrating ultra-wideband (UWB) positioning and sensor data in complex indoor environments. With the growing adoption of unsupervised and self-supervised learning in occupant activity recognition, the vast amount of unlabelled data in buildings can be better utilised. However, some studies place excessive dependence on wearable devices, which are often costly and potentially intrusive to individual privacy. Internet-of-Things (IoT) enabled indoor environmental sensors and BMS embedded in buildings offer a less invasive and more cost-effective solution. Therefore, applying self-supervised learning for occupant behaviour and activity recognition using only building environmental sensors or BMS data is an emerging challenge that requires further study.

Methodology

This section presents the self-supervised approach for recognising occupant activity patterns based on BMS data. As illustrated in Figure 1, the model tokenises the pre-processed BMS data into a ternary consisting of the timestamp, the feature/sensor ID and the feature/sensor

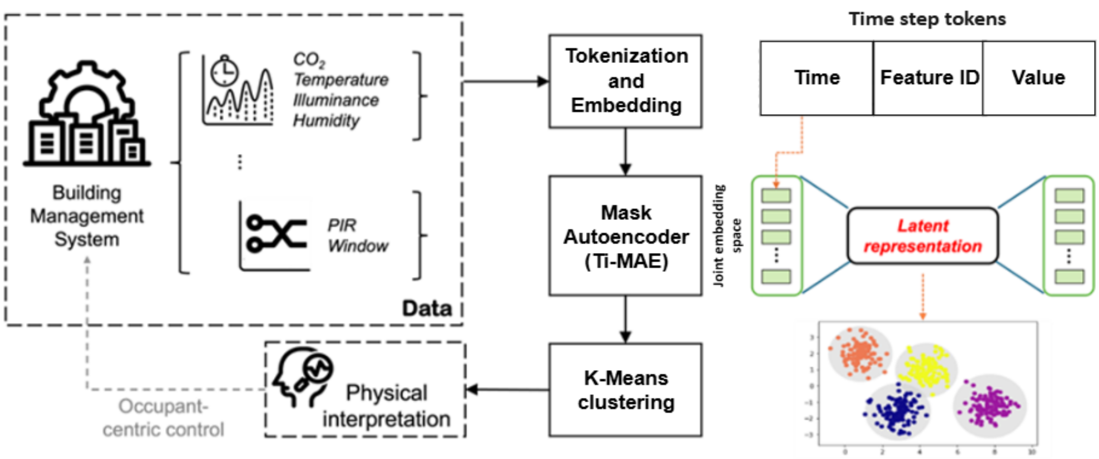

Figure 1: Self-supervised learning approach for occupant activity pattern recognition

value. Each component is projected into the latent space through a dedicated feed-forward or embedding layer, where learnable mask tokens replace missing values. To learn latent representations, these token embeddings are forwarded into a Transformer-based mask autoencoder (Ti-MAE). Besides, a multitasking framework is adopted to simultaneously optimise reconstruction, clustering and time series forecasting. This architecture allows the model to capture both local variations and global dependencies among parameters, ensuring reliable performance even when critical sensor data (e.g., CO_2 and PIR sensors) are partially missing or corrupted. The model outputs clusters of the latent space, identifying different patterns of occupant activity in a self-supervised manner. These clusters are analysed and interpreted using prior knowledge of occupant schedules and behavioural modes, providing meaningful insights into occupant activities and supporting occupant-centric control strategies.

Data pre-processing: BMS data are typically collected over extended periods with a fine-granular temporal resolution. These datasets often include measurements from multifunctional integrated environmental sensors (e.g., temperature, humidity, illuminance and CO_2 concentration), Passive Infrared (PIR) occupancy sensors, and window open/close sensors. BMS data are recorded as time series with typical sampling intervals of 15 minutes, 30 minutes, or an hour. BMS data are pre-processed, more specifically synchronised and linearly interpolated to handle missing values. Raw sensor data is transformed into a triplet-form token (time, sensor ID, value) for embedding into the model. Tokenisation eliminates tedious feature engineering, allowing the model to automatically learn feature representations directly from the raw data, giving the model flexibility to capture the complex relationships and temporal dynamics between sensors.

Masked autoencoder (Ti-MAE): Ti-MAE is a self-supervised framework for time series representation learning (Li et al., 2023). Ti-MAE reconstructs the time series by randomly masking portions of the raw time series and using AE to learn latent representations. The architecture in Figure 2 includes the Token Embeddings, the Multi-Head Self-Attention Block (Transformer), and the Multitasking Block. A token, the fundamental unit of information, is derived from the raw data through a Feed-Forward Network (FFN). Each token encodes a timestamp, sensor identifier, and corresponding value, which are integrated using an embedding layer and the Gaussian Error Linear Unit (GELU) activation function. These tokens serve as input to the Transformer encoder (Kämäräinen, 2025), enabling them to capture feature-specific information and broader contextual relationships. Through the transformer's self-attention mechanism, tokens interact with one another to learn global dependencies embedded within the data.

In traditional self-supervised training, masked inputs are typically replaced with a fixed placeholder vector, which restricts the model to relying solely on surrounding unmasked tokens from the same feature or sensor for reconstruction. To address this limitation, a learnable mask token is employed (Hu et al., 2023). Unlike static placeholders, the learnable mask token has configurable parameters that are updated during training. When used to replace a masked feature, it allows the model to infer missing values by leveraging contextual information across different sensors through iterative learning. The reconstruction loss is backpropagated throughout the model, including the mask token, enabling its representation to be progressively refined. Over time, the mask token converges to an optimal embedding that enhances the model's capacity to reconstruct missing information under diverse conditions. The resulting latent representations are processed by a Transformer encoder, which uses a self-attention mechanism to compute pairwise relationships among all tokens, capturing both local variations and global dependencies within the time series.

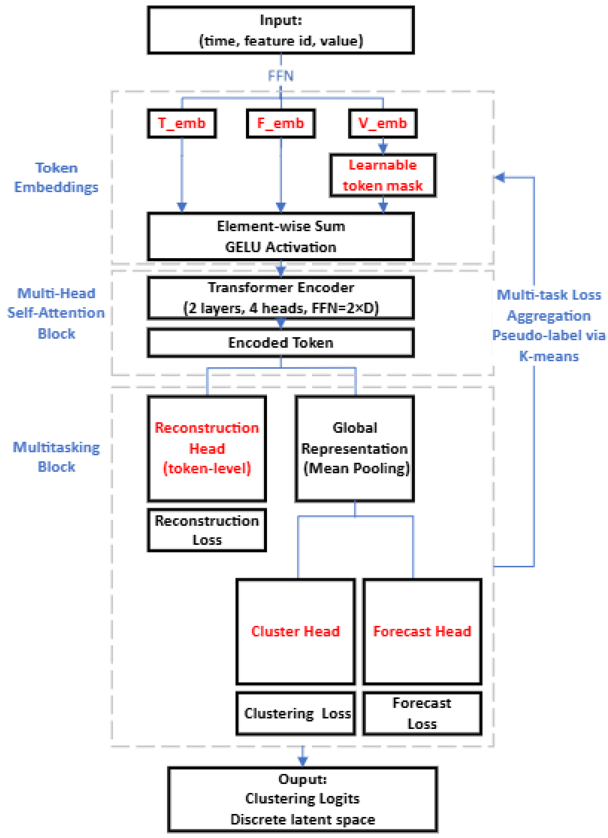

Figure 2: Self-supervised learning architecture (Ti-MAE)

The model incorporates a Multitasking Block to enhance the model's robustness to missing sensor data and improve overall representational quality and generalisation performance. This block comprises three heads. First, the reconstruction head individually reconstructs the token-level representations output by the Transformer encoder, compelling the model to recover the original features from masked inputs. This process enables the capture of fine-grained local details and contextual dependencies. Second, the clustering head aggregates global hidden vectors by pooling the mean of all token representations, which are then mapped to clustering logits. These logits are used to iteratively update pseudo-labels through K-means clustering, with the Silhouette score used to determine the optimal number of clusters, thereby revealing the intrinsic distributional structure of the data. Last but not least, the prediction head utilises the global hidden vectors to forecast future sensor values, aiding in the modelling of temporal dynamics and enhancing the model's capacity to impute missing data through prediction. These heads operate collaboratively, allowing the model to concurrently learn local and global, temporal and contextual dynamics, leading to improved recognition of occupant activity patterns. The reconstruction loss is dominant, while clustering and prediction losses are weighted appropriately to ensure balanced optimisation.

Case study

This section presents a case study conducted in a building owned by Newcastle University, with data collected from a meeting room on the 4th floor. Faculty and students regularly use the room for meetings, discussions, and group work. A multifunctional sensor (RC1), integrating temperature, illuminance, humidity, CO_2 concentration measurements, and PIR-based occupancy, was installed to capture environmental and occupancy-related data. Besides, window open/close sensors were fitted on the arm of the south-east facing window. The floor plan of the meeting room is shown in Figure 3. To ensure that the dataset captured a representative range of occupant activities, data were retrieved from the BMS over fourteen weeks during the term, from 16 September to 23 December 2024. This period was chosen due to the space's high occupancy and activity level, providing a comprehensive dataset for model training and validation.

Data pre-processing

The retrieved BMS data are pre-processed by aligning different sensor datasets to a uniform time frame, interpolating scattered missing values, and distinguishing between binary and numerical data types. The process involves defining consistent time ranges with specified start and end dates and uniform intervals. In this study, data from 16 September 2024 at 00:00:00 UTC to 1 December 2024 at 23:59:59 UTC were used as the training set, while data from 2 December 2024 at 00:00:00 UTC to 23 December 2024 at 23:59:59 UTC were used as the testing set. The acquisition interval for all sensors was uniformly aligned to 15 minutes, and any missing or null values were filled using linear interpolation.

Latent representation by Ti-MAE

This study applies a uniform random masking strategy, with 70% of input tokens masked to enhance the model's robustness to missing data. The encoding module comprises two Transformer encoder layers, each equipped with four multi-head self-attention mechanisms and a feed-forward network with twice the number of hidden neurons. This architecture enables the model to capture complex temporal and contextual dependencies among the input features. The latent representations are mapped back to the original through a linear reconstruction layer. Model training is guided by a composite loss function, with reconstruction, clustering, and prediction losses jointly optimised, weighted at 0.59, 0.29, and 0.12, respectively, ensuring an appropriate balance between recovering input features, uncovering latent structures, and forecasting sensor values.

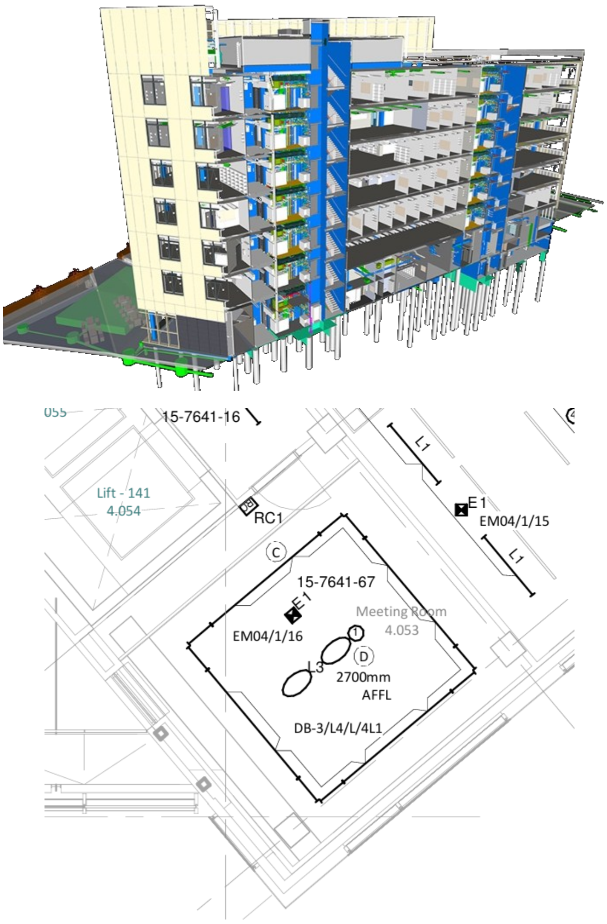

Figure 3: Studied building and floor plan of the floor-4 room

Figure 4 presents the results of the Silhouette method, indicating an optimal number of clusters as 9. The model is trained using the AdamW optimiser with a learning rate of 1e-4 through an iterative training process. The clustering pseudo-label is updated once per Iteration. Training begins with 10 epochs of pre-training focused solely on reconstruction loss, followed by up to 25 rounds of multitask training, each comprising 60 epochs. Early stopping is triggered if reconstruction and clustering losses, along with pseudo-label updates, do not improve within five consecutive epochs.

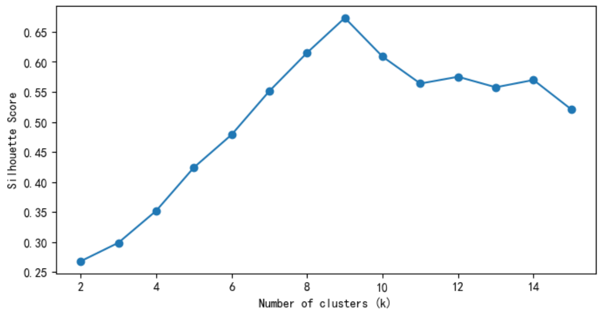

Figure 4: Silhouette Analysis

Given that some sensors may be unavailable in particular spaces, this study evaluated the expressiveness of the latent representations learnt from Ti-MAE by selectively 'masking' the input data from one of the sensors in the testing set. The reconstruction loss from the model after removing particular sensor data is shown in Table 1. Notably, the exclusion of the PIR sensors led to the most substantial decline in model accuracy, underscoring the critical importance of occupancy detection in recognising occupant activity. However, the model maintains robust performance despite the absence of some sensor inputs, demonstrating the flexibility and practical applicability of the Ti-MAE framework in real-world building.

Table 1: Reconstructed loss scores for the model

Dataset	Reconstruction loss
Training set	0.0554
Testing set	0.0341
Removed temperature sensor	0.0385
Removed CO2 sensor	0.0437
Removed humidity sensor	0.0421
Removed light sensor	0.0383
Removed PIR sensor	0.0482
Removed window arm sensor	0.0351

Clustering and physical interpretation

Upon completion of training and testing phases, the model exports the discrete latent space along with the K-means clustering results. As shown in Fig. 5, the similar distribution of cluster assignments across both training and testing sets indicates that the model generalises well beyond the training data.

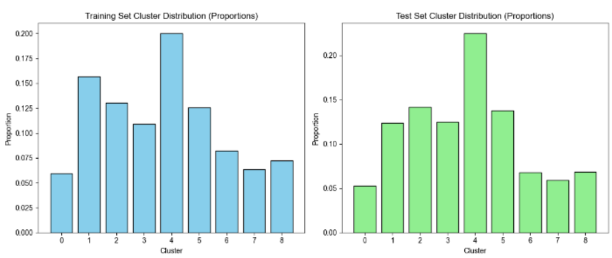

Figure 5: Clustering ratio of training and testing sets

Figure 6 presents the colour-coded clustering results from the testing set, overlaid with the corresponding raw sensor data. The result illustrates various patterns over time that reflect the combined influence of occupant activity, window status, and particularly, ambient rooms. Table 2 provides a detailed physical interpretation for each cluster, derived from the periodicity of cluster occurrences, variations between weekdays and weekends, specific times of the day, and other knowledge about the building's usage and operational contexts. Behaviours such as opening windows or doors introduce substantial variability into the sensor readings, often disrupting the clarity of the clustering outcomes. This highlights the high degree of environmental interconnectedness within the building, where open conditions increase the influence of adjacent spaces on the local sensor data.

Table 2: Physical interpretation of each cluster

Cluster	Physical interpretation
1	Window closed, located in the afternoon when the building is closed to closing
2	Daytime with windows open
3	The behaviour of opening windows and leaving room doors open
4	Unoccupied, possibly meaning that the intensity of activity in other spaces within the building is beginning to rise
5	Closed windows, unoccupied nights after the building closes
6	Window closed, Occupied
7	The Night of the Open Window
8	Window closed, unoccupied late night
9	Noon with windows closed and unoccupied

Discussion

This paper proposes a self-supervised approach based on Ti-MAE. Leveraging unlabelled BMS data, the approach demonstrates satisfactory performance in recognising occupant activity without manual label annotation or collecting personal occupant data (e.g., from wearables). Additionally, it offers the potential to generate data labels for further supervised learning, enabling more accurate classification and prediction of occupant activities.

Technologically, the approach implements a hybrid representation learning model that integrates the Transformer architecture with an MAE incorporating a Multitasking Block. The transformer efficiently processes time series data, capturing both short-term variations and long-term dependencies, while the MAE enhances robustness by reconstructing missing or masked data, making it resilient to incomplete datasets. This is, in particular, critical for occupant activity recognition in buildings. Data gaps are frequently found in BMS due to battery depletion, communication failures, maintenance issues, and other issues. The proposed architecture mitigates the impact of missing data by leveraging interdependencies among different sensed parameters. Moreover, as window open/close and occupancy sensors are less common in many buildings compared to other indoor environmental sensors, the model allows for the temporary deployment of these sensors, or more likely manual recording of occupancy and window status during non-continuous periods. This flexibility enables the model to learn from intermittent data inputs, maintaining reliable performance even after these data sources are no longer available.

The proposed occupant activity recognition approach has certain limitations. Manually interpreting the physical meaning of clustering labels is subjective, relying on prior knowledge and empirical experience. As demonstrated in the case study, cluster labels not only reflect occupant activities but also incorporate the effects of window status and ambient room conditions. The granularity of occupant activity recognition is inherently tied to the information embedded in the BMS data. For instance, the case study data could distinguish between unoccupied, briefly occupied, and long-time occupied statuses, while in other studies, BMS data may provide insights into the intensity of activities, such as the number of occupants. The case study was conducted in a single meeting room, limiting the diversity of data sources and the variety of occupant activities identified through clustering. While the token embedding employed in this paper makes the model efficient, the lack of feature engineering may also reduce the ability to uncover more information from the sensor data. Future work will extend the application of the approach to a broader range of spaces within the case building and integrate additional sensor types, such as outdoor environmental sensors, to assess whether the additional sensor data can provide finer granularity in activity recognition.

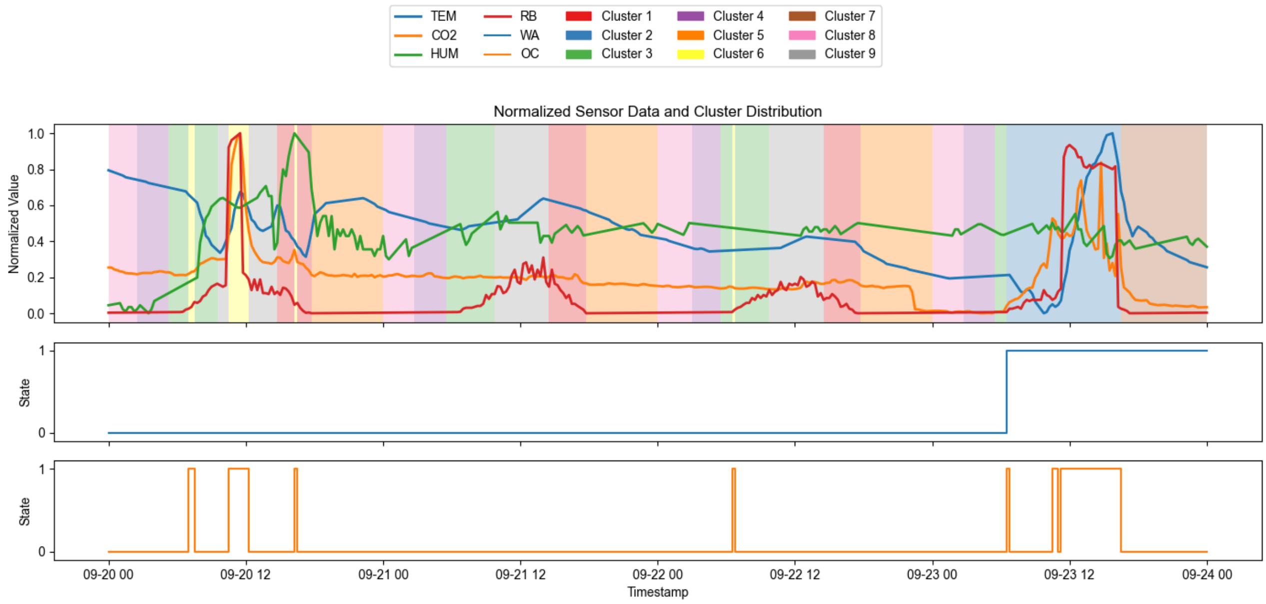

Figure 6: Clustering results for sensor data in the testing set

Conclusions

This paper proposes a self-supervised learning approach for occupant activity recognition, which extracts implicit occupant activity patterns from BMS sensor data using a Transformer-based Masked Autoencoder (Ti-MAE) and K-means clustering. The case study demonstrates that the method effectively identifies occupant activities in a university meeting room using unlabelled BMS data and remains robust even when some data are missing. Compared to supervised learning methods, the proposed approach reduces reliance on manual labelling and wearable devices, offering a privacy-preserving and resource-efficient solution for occupant behaviour recognition in building environments. While the granularity of recognised occupant activities is inherently tied to the 'enrichment' of information embedded in the BMS data, future work will incorporate additional outdoor environmental sensors and apply the approach to multifunctional spaces to enhance its versatility and accuracy.

This approach opens new avenues for modelling spatiotemporal occupant dynamics within building environments. In specific, recognising occupant activities in individual rooms, such as faculty leaving their offices to gather in the common room for coffee breaks at particular times of the day, enables the association of activities across multiple spaces over time, providing deeper insights into how individuals collectively utilise building environments.

References

Abedin, A., Rezatofighi, H., & Ranasinghe, D. C. (2021). Guided-GAN: Adversarial representation learning for activity recognition with wearables. arXiv preprint arXiv:2110.05732.

Ahmed, N., Rafiq, J.I. & Islam, M.R. (2020) Enhanced Human Activity Recognition Based on Smartphone Sensor Data Using Hybrid Feature Selection Model. Sensors, 20 (317).

Al Horr, Y., Arif, M., Kaushik, A., Mazroei, A., Katafygiotou, M. & Elsarrag, E. (2016) Occupant productivity and office indoor environment quality: A review of the literature. Building and Environment, 105, pp.369–389.

Almaslukh, B., AlMuhtadi, J. & Artoli, A. (2017) An effective deep autoencoder approach for online smartphone-based human activity recognition. International Journal of Computer Science and Network Security, 17, pp.160–165.

Anguita-Molina, M.Á., Cardoso, P.J.S., Rodrigues, J.M.F., Medina-Quero, J. & Polo-Rodríguez, A. (2025) Multi-Occupancy Activity Recognition Based on Deep Learning Models Fusing UWB Localisation Heatmaps and Nearby-Sensor Interaction. IEEE Internet of Things Journal, pp.1–1.

Ariza-Colpas, P.P., Vicario, E., Oviedo-Carrascal, A.I., Butt Aziz, S., Piñeres-Melo, M.A., Quintero-Linero, A. & Patara, F. (2022) Human Activity Recognition Data Analysis: History, Evolutions, and New Trends. Sensors, 22 (3401).

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z. & Liu, Y. (2021) Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges, and Opportunities. ACM Computing Surveys, 54 (77), pp.1–40.

Dirgová Luptáková, I., Kubovčík, M. & Pospíchal, J. (2022) Wearable Sensor-Based Human Activity Recognition with Transformer Model. Sensors, 22 (1911).

Ek, S., Presotto, R., Civitarese, G., Portet, F., Lalanda, P., & Bettini, C. (2025). Comparing self-supervised learning techniques for wearable human activity recognition. CCF Transactions on Pervasive Computing and Interaction, 1-18.

Fernández, J.E. (2007) Resource Consumption of New Urban Construction in China. Journal of Industrial Ecology, 11, pp.99–115.

Geng, Y., Ji, W., Lin, B. & Zhu, Y. (2017) The impact of thermal environment on occupant IEQ perception and productivity. Building and Environment, 121, pp.158–167.

Haresamudram, H., Essa, I., & Plötz, T. (2022). Assessing the state of self-supervised human activity recognition using wearables. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 6(3), 1-47.

Hu, H., Baldassarre, F., Azizpour, H., 2023. Learnable Masked Tokens for Improved Transferability of Self-supervised Vision Transformers, in: Amini, M.-R., Canu, S., Fischer, A., Guns, T., Kralj Novak, P., Tsoumakas, G. (Eds.), Machine Learning and Knowledge Discovery in Databases. Springer Nature Switzerland, Cham, pp. 409–426. https://doi.org/10.1007/978-3-031-26409-2_25

Ito, C., Cao, X., Shuzo, M. & Maeda, E. (2018) Application of CNN for Human Activity Recognition with FFT Spectrogram of Acceleration and Gyro Sensors. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp' 18). New York, NY, USA, Association for Computing Machinery, pp.1503–1510.

Kanthila, C., Boodi, A., Beddiar, K., Amirat, Y. & Benbouzid, M. (2021) Building Occupancy Behavior and Prediction Methods: A Critical Review and Challenging Locks. IEEE Access, 9, pp.79353–79372.

Kämäräinen, J. K. (2025). Minimal Time Series Transformer. arXiv preprint arXiv:2503.09791.

Li, Z., Rao, Z., Pan, L., Wang, P., & Xu, Z. (2023). Ti-mae: Self-supervised masked time series autoencoders. arXiv preprint arXiv:2301.08871.

Naylor, S., Gillott, M., & Lau, T. (2018). A review of occupant-centric building control strategies to reduce building energy use. Renewable and Sustainable Energy Reviews, 96, 1-10.

Nesa, N. & Banerjee, I. (2017) IoT-Based Sensor Data Fusion for Occupancy Sensing Using Dempster–Shafer Evidence Theory for Smart Buildings. IEEE Internet of Things Journal, 4, pp.1563–1570.

Ordóñez, F.J. & Roggen, D. (2016) Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. Sensors, 16 (115).

Shoemaker, S. (2023) NREL researchers reveal how buildings across United States do—and Could—Use Energy. NREL Transform. Energy.

Wang, J. & Biljecki, F. (2022) Unsupervised machine learning in urban studies: A systematic review of applications. Cities, 129, p.103925.

Zhang, X., Zhou, T., Kokogiannakis, G., Xia, L. & Wang, C. (2023) Estimating the number of occupants and activity intensity in large spaces with environmental sensors. Building and Environment, 243, p.110714.