A Machine Learning Framework for Optimising Indoor Thermal Comfort and Air Quality through Sensor Data Streams

Chun Wai Chiu*
Christos Efstratiou*
m.chiu@kent.ac.uk
c.efstratiou@kent.ac.uk
School of Computing,
University of Kent
Canterbury, Kent, UK

Marialena Nikolopoulou
m.nikolopoulou@kent.ac.uk
School of Architecture, Design and Planning,
University of Kent
Canterbury, Kent, UK

Matthew Barker
Andrew Baldwin
Malcolm Clarke
mbarker@baxallconstruction.co.uk
abaldwin@baxallconstruction.co.uk
mclarke@baxallconstruction.co.uk
Baxall Construction Ltd.
Paddock Wood, Kent, UK

ABSTRACT

Optimising thermal comfort and air quality in indoor environments presents a complex, dynamic challenge that traditional static systems struggle to address effectively. We propose a novel real-time framework that tackles this multifaceted problem by leveraging stream clustering and time-series forecasting techniques. Our system continuously analyses sensor data to summarise comfort conditions and predict future indoor states. Simulations based on the stream clustering model indicate potential for significant improvements in indoor comfort, increasing comfort duration from 6% to 74%. Furthermore, the time-series forecasting model demonstrated strong performance, achieving mean absolute errors of 0.026 and 0.034 on test and demonstration datasets, respectively. This resource-efficient approach demonstrates promise for real-time indoor environment management, effectively balancing thermal comfort and air quality considerations.

CCS CONCEPTS

• **Computing methodologies** → **Online learning settings**; **Mixture modeling**; **Supervised learning by regression**; **Neural networks**; • **Applied computing** → **Multi-criterion optimization and decision-making**.

KEYWORDS

thermal comfort, air quality, built environments, machine learning

ACM Reference Format:
Chun Wai Chiu, Christos Efstratiou, Marialena Nikolopoulou, Matthew Barker, Andrew Baldwin, and Malcolm Clarke. 2024. A Machine Learning Framework for Optimising Indoor Thermal Comfort and Air Quality through Sensor Data Streams. In *The 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '24), November 7–8, 2024, Hangzhou, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3671127.3699531

*Both authors contributed equally to this research.

BuildSys '24, November 7–8, 2024, Hangzhou, China

ACM ISBN 979-8-4007-0706-3/24/11
https://doi.org/10.1145/3671127.3699531

1 INTRODUCTION AND RELATED WORK

The built environment, especially office spaces, plays a major role in our daily lives as people spend a lot of time indoors. Ensuring comfort and well-being in these spaces is becoming more important. A recent survey [10] highlights four key areas of indoor comfort: thermal, visual, acoustic, and air quality. While machine learning for thermal comfort is well-studied [4, 7, 9, 12], the other aspects and in particular their interactions during environmental intervention, remain largely unexplored [10].

Although recent studies [3, 8] have begun addressing this gap, their machine learning models typically operate in offline, non-streaming settings. They cannot incorporate real-time fluctuations in environmental conditions. This is a common issue in the literature, despite the need for real-time modelling given the dynamic nature of indoor conditions, influenced by factors like seasonal changes and occupants' physiological states.

While reinforcement learning approaches [12] offer adaptive solutions, they are often computationally demanding, making them unsuitable for resource-constrained applications. To address this, our study proposes a resource-efficient framework for real-time analysis, learning, and optimisation of thermal comfort and air quality, balancing efficiency and computational resource use.

2 PROBLEM FORMULATION

Optimising office environments often focuses on thermal comfort while overlooking the interplay of factors such as air quality. To address this gap, we formulate the problem as sensor data stream processing, targeting environments with sufficient environmental sensors. We collected sensor data from a small office with five occupants between September 2023 to July 2024, where data from June and July 2024 are reserved for demonstration. The dataset consists of time-stamped records collected every 5 minutes, including indoor, outdoor and A/C temperatures, humidity, CO_2 levels, light intensity, and window status.

Our framework continuously analyse this data stream to identify trends and employs resource-efficient predictive models to forecast short-term changes in thermal comfort and air quality. Based on these predictions, the system suggests real-time interventions, such as adjusting ventilation or cooling, to maintain comfort. Overall, our framework aims to minimise the indoor discomfort $D = PMV + \max(0, CO_2 - 1000)$, where $-0.5 \leq PMV \leq 0.5$ is the Predictive Mean Vote [1] and CO_2 levels should remain below 1000 ppm.

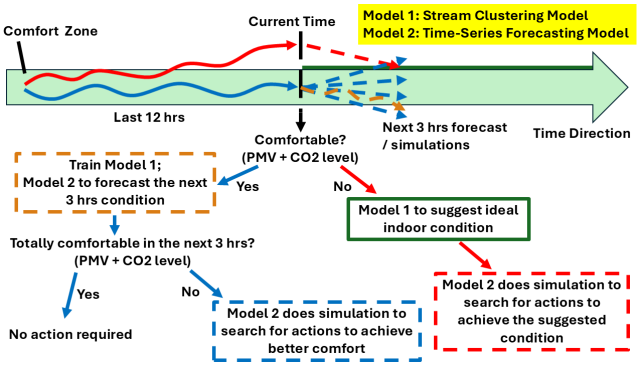


Figure 1: Illustration of the Proposed Framework

3 SYSTEM DESCRIPTION

Our framework monitors an indoor environment (an office room) via sensor data streams at five-minute intervals. Its goals are to: 1) summarise past comfort conditions during office hours (8 am to 6 pm, Monday to Friday, excluding bank holidays), and 2) suggest interventions to maintain thermal comfort and air quality via forecasting simulations. It consists of two models to achieve these goals: a stream clustering model and a time-series forecasting model. Fig. 1 presents an overview of the framework.

The system collects real-time sensor data from the environment and determines the current comfort level based on the Predicted Mean Vote (PMV) model [1] and CO_2 threshold. We adopted the PMV function from *pythermalcomfort* package [11] with assumptions of a typical office environment, i.e., metabolic rate index at 1.1 (typing), clothing index at 0.96 (trousers, long-sleeve shirt), and indoor air velocity at 0.2 m/s. We also assume that the radiant temperature is the same as the dry bulb temperature.

If the environment is comfortable (blue route on Fig. 1), the framework trains the stream clustering model using current indoor conditions (indoor temperature, humidity and CO_2 levels). To anticipate potential changes, the framework employs a pre-trained forecasting model to predict indoor conditions in the next three hours (dotted brown lines on Fig. 1). A comfort ratio is then calculated and compared to a threshold (θ). If the ratio exceeds θ, no intervention is required. Otherwise, the framework suggests actions (e.g., adjusting windows or A/C) to maintain comfort.

When the environment is uncomfortable (red route in Fig. 1), the framework skips training the stream clustering model and searches for the nearest comfort point in the feature space. It then predicts the effects of various interventions and recommends the most effective one to achieve the nearest comfort point.

The following sections detail the core components. Section 3.1 details the stream clustering model and its function in suggesting comfort condition. Section 3.2 covers the time-series model and its function in forecasting and intervention effect estimation.

3.1 Continuous Comfort Conditions Summarisation through Stream Clustering

The main purpose of integrating a stream clustering model into the framework is to establish a target comfort zone for the environment, which guides the search for interventions when the current condition is uncomfortable. The data stream clustering model continuously summarises the comfort conditions of the office through a stream of environmental data (i.e., indoor temperature, humidity, and CO_2 levels) into micro-clusters [2], which are data structures grouping close data points in the feature space. As data streams are potentially infinite, the model generates more micro-clusters over time. To avoid memory issues, it's essential to use a model that limits the maximum number of micro-clusters. Clustream [2] was selected for our framework due to its ability to manage this limitation effectively.

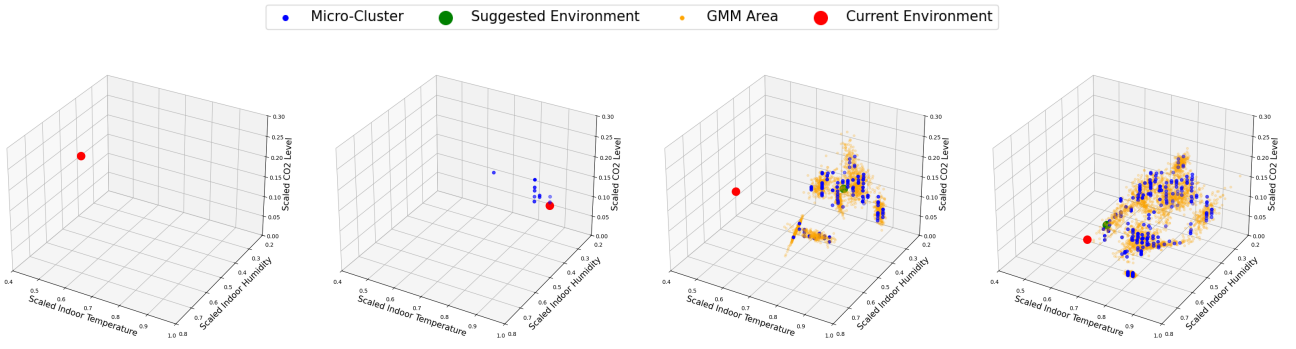
When the system needs to suggest an achievable target comfort condition (temperature, humidity, and CO_2 levels), it finds the nearest comfort point in the feature space based on historical data. The process begins with applying Affinity Propagation [6] to estimate the number of macro-clusters from micro-clusters, which are then modelled using a Gaussian Mixture Model (GMM). GMM provides essential soft clustering for representing comfort zones with varying shapes and sizes influenced by environmental factors, enabling flexible target suggestions based on how close we want to be to the centre of the comfort zone and the associated confidence level. To identify the nearest comfort point, the system draws a line from the current condition to the centre of the nearest macro-cluster and finds the intercept with its boundary, defined as 1.5 standard deviations (z-scores) from the centre, covering 86.64% of the Gaussian's area. This choice, instead of using 3 z-scores (covering 99%), allows for a target closer to the centre while maintaining comfort without significantly increasing intervention efforts. NOte that Macro-clusters are formed only when the system requests a target condition to conserve computational resources.

Fig. 2 shows the evolution of micro and macro clusters over time. The red dot is the current condition, the green dot represents the target, blue dots are micro-clusters, and the orange areas are macro-clusters (Gaussians).

At *Time Step 0* (Fig. 2(a)), only the red dot is visible, as the system has just started and has not yet learnt from the environment. As the system continues gathering data, micro-clusters begin to form, revealing patterns in the feature space. By *Time Step 1360* (Fig. 2(b)), several micro-clusters emerge, indicating initial insights into temperature, humidity, and CO_2 levels. As more data is collected, the system applies a Gaussian Mixture Model (GMM) to generalise the micro-clusters into macro-clusters, representing broader trends, as seen in *Time Step 3810* (Fig. 2(c)). The GMM is applied only when the framework considers the current indoor condition as uncomfortable and is queried to suggest a new comfort condition. Finally, at *Time Step 5165* (Fig. 2(d)), the system has summarised sufficient past data, defining the comfort zone more clearly. This clustering process enables the framework to effectively identify interventions to maintain optimal thermal comfort and air quality.

3.2 Intervention Simulation based on Time-Series Forecasting

The proposed framework proactively maintain comfort by utilising a time-series forecasting model. Fig. 3 shows the model prediction on indoor temperature, humidity, and CO_2 levels. The blue lines represent the last 12 hours of data, recorded every 5 minutes, while



(a) Time Step: 0 (b) Time Step: 1360 (c) Time Step: 3810 (d) Time Step: 5165

Figure 2: Micro and Marco clusters Evolution over a Sensor Data Stream of an Office Room

the green lines are forecasts for the next three hours. The orange lines show actual conditions.

The time-series forecasting model is a bidirectional Gated Recurrent Unit (GRU) deep neural network [5]. It processes the past 12 hours of sensor data, including indoor temperature, humidity, CO_2 levels, light levels, temperature and humidity near windows and A/C, and window states, along with temporal factors such as the day of the week and change rates of temperature and humidity (blue line). Based on this input, the model predicts the next 3 hours of indoor CO_2, light levels, temperature, humidity, and window temperature and humidity (green line).

The model was trained, validated, and tested using data from September 2023 to May 2024, as detailed in Section 4. Therefore, its predictions reflect the average office room conditions, given the input data. This explains why predictions for certain features, like temperature and CO_2 levels, may deviate from the actual values (orange lines) in the example shown in Fig. 3.

Besides forecasting, the model simulates the effects of different interventions, helping the system recommend the most effective strategies. Fig. 3 (dotted lines) shows how the system simulates actions like adjusting windows or air-conditioning (A/C) and predicts their effect. For examples, opening windows reduces CO_2 levels but has little effect on temperature as in that scenario outdoor temperature is similar to indoor, adjusting the A/C affects temperature but may not improve air quality. By forecasting these effects, the system identifies the best interventions to maintain both thermal comfort and air quality.

4 EXPERIMENTAL SETUP

The experimental setup involved key steps to download, preprocess, and utilise sensor data for both developing a time-series forecasting model and conducting data stream simulations to demonstrate the entire framework. Data was collected from a set of sensors, covering September 2023 to July 2024. Data from September 2023 to May 2024 was used for model training, validation, and testing the time-series forecasting model, while data from June to July 2024 was reserved for framework demonstration.

To evaluate the stream clustering model, we iterated through the demonstration dataset to simulate a data stream and recorded the target conditions suggested by the model at each time step. Assuming the immediate implementation of these suggested conditions, we compared the density distributions of CO_2 levels and PMV (Predicted Mean Vote) before and after optimisation. This comparison allowed us to assess improvements in indoor comfort across the demonstration set. The results were then visualised to showcase the clustering model's performance in a real-time scenario.

The dataset was split into training, validation, and test sets: 80% of the data was allocated for training and validation (with an 80/20 split), and the remaining 20% was reserved for testing. All datasets were normalised based on the training set, with the normaliser saved for future use.

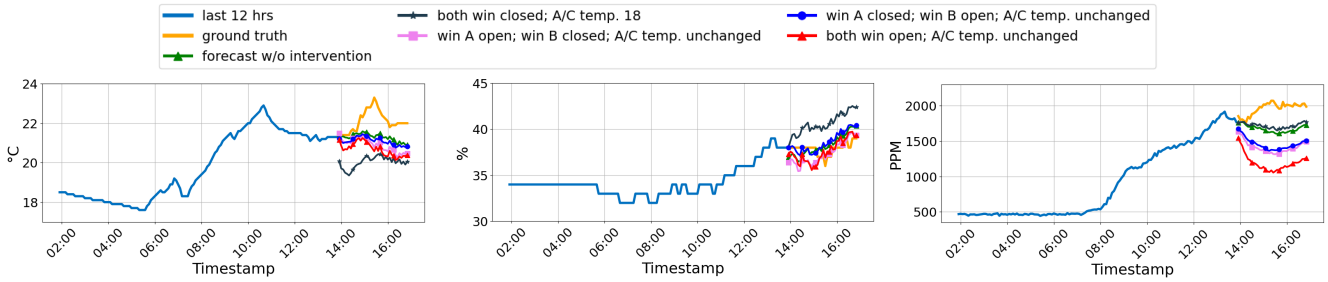
Data was then prepared as time-series sequences, with input sequences of 144 time steps (12 hours) and output sequences of 36 time steps (3 hours). Each dataset was shuffled to ensure robust learning and evaluation. The model, with an input dimensionality of 22 and an output dimensionality of 6, was trained for 500 epochs with a batch size of 8. Early stopping was applied, restoring the best-performing weights from the 6th epoch. The model was then tested and saved to be used in the framework.

5 RESULTS

Our framework shows significant improvements in maintaining indoor comfort. The stream clustering model effectively summarises comfort levels, as shown in Fig. 4. Assuming instant implementation of suggested conditions, the density distributions of CO_2 levels and PMV shift significantly towards more comfortable conditions after optimisation. Initially, only 6% of the time was spent in comfortable conditions, but this is increased by 68% after optimisation.

The time-series forecasting model also provides accurate predictions of future indoor states. Evaluated on both test and demonstration datasets, Fig. 5 and 6 present the model's loss and Mean Absolute Error (MAE) over time steps on each dataset. The model achieved stable performance, with average MAEs of 0.026 and 0.034 on the test and demonstration sets, respectively. These results demonstrate the model's robustness in predicting future indoor conditions with new data.

Overall, these results show that our framework effectively summarises past comfort conditions and forecasts future ones to guide intervention search. The substantial 68% increase in comfortable periods demonstrates its potential to improve indoor thermal comfort and air quality in office environments, while the low MAE values affirm the accuracy of the forecasting model.



(a) Indoor Temperature (b) Indoor Humidity (c) Indoor CO_2 level

Figure 3: Simulated Effect of Window and A/C Settings on Indoor Temperature, Humidity, and CO_2 Levels over the Next 3 Hours of the Environment

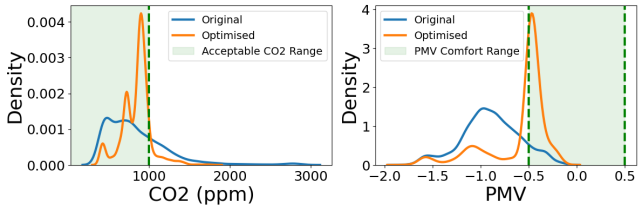


Figure 4: Comparison of CO_2 and PMV Data Density Before and After Optimisation on Demonstration Set

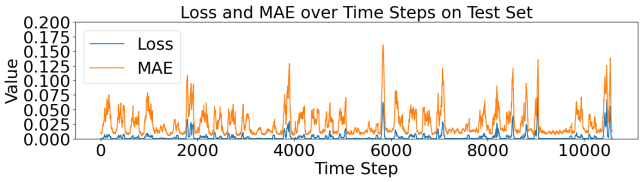


Figure 5: Loss and Mean Absolute Error (MAE) over Time Steps on Test Set

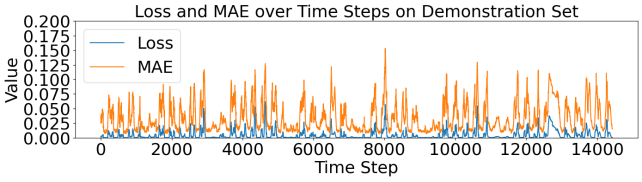


Figure 6: Loss and Mean Absolute Error (MAE) over Time Steps on Demonstration Set

6 CONCLUSION

This paper proposes a novel real-time framework that combines stream clustering and time-series forecasting to optimise indoor thermal comfort and air quality. The stream clustering model summaries the comfort zone and suggests adjustments as needed, showing a potential increase in comfort duration from 6% to 74% through experiments. The time-series forecasting model predicts indoor conditions for the next three hours which supports proactive intervention suggestions. It performed well in tests, with a mean absolute error (MAE) of 0.026 on test data and 0.034 on demonstration data.

However, the framework's performance may vary across different environments. Future work will investigate the necessary adaptations to ensure effectiveness in diverse contexts, including its deployment in two primary schools in the UK equipped with sensing infrastructure. Besides, we plan to incorporate more environmental variables and aspects of indoor comfort, expanding intervention options, and adapting the time-series model for continuous data stream learning. Furthermore, we will explore simultaneous clustering of both comfortable and uncomfortable zones. This dual approach could accelerate the modelling process, particularly during the initial deployment of the framework.

ACKNOWLEDGMENTS

This work was supported by Innovate UK through KTP funding, Project No. 13060. Special thanks to Dr Ismail Alarab for his contributions to the initial investigation and sensor cloud API integration.

REFERENCES

[1] 2005. ISO 7730 - Ergonomics of the thermal environment — Analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria.

[2] Charu C. Aggarwal, Jiawei Han, Jianyong Wang, and Philip S. Yu. 2003. A Framework for Clustering Evolving Data Streams. *IEEE VLDB* (2003), 81–92.

[3] T. Al Mindeel, E. Spentzou, and M. Eftekhari. 2024. Energy, thermal comfort, and indoor air quality: Multi-objective optimization review. *Renewable and Sustainable Energy Reviews* 202 (2024), 114682. https://doi.org/10.1016/j.rser.2024.114682

[4] Pablo Aparicio-Ruiz, Elena Barbadilla-Martín, José Guadix, and Julio Nevado. 2023. Analysis of Variables Affecting Indoor Thermal Comfort in Mediterranean Climates Using Machine Learning. *Buildings* 13, 9 (2023). https://doi.org/10.3390/buildings13092215

[5] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[6] Brendan J. Frey and Delbert Dueck. 2007. Clustering by Passing Messages Between Data Points. *Science (New York, N.Y.)* 315 (03 2007), 972–6. https://doi.org/10.1126/science.1136800

[7] Tseng-Fung Ho, Hsin-Han Tsai, Chi-Chih Chuang, Dasheng Lee, Xi-Wei Huang, Hsiang Chen, Chin-Chi Cheng, Yaw-Wen Kuo, Hsin-Hung Chou, Wei-Han Hsiao, Ching Hsu Yang, and Yung-Hui Li. 2024. Thermal Comfort Model Established by Using Machine Learning Strategies Based on Physiological Parameters in Hot and Cold Environments. *Indoor Air* (2024).

[8] Qiwen Jiang, Jialu Liu, and Xian Yang. 2024. Optimization of indoor quality and thermal comfort for university classrooms using data-based machine learning. *E3S Web of Conferences* (2024).

[9] Zahra Qavidel Fard, Zahra Sadat Zomorodian, and Sepideh Sadat Korsavi. 2022. Application of machine learning in thermal comfort studies: A review of methods, performance and challenges. *Energy and Buildings* 256 (2022), 111771. https://doi.org/10.1016/j.enbuild.2021.111771

[10] Ying Song, Fubing Mao, and Qing Liu. 2019. Human Comfort in Indoor Environment: A Review on Assessment Criteria, Data Collection and Data Analysis Methods. *IEEE Access* 7 (2019), 119774–119786. https://doi.org/10.1109/ACCESS.2019.2937320

[11] Federico Tartarini and Stefano Schiavon. 2020. pythermalcomfort: A Python package for thermal comfort research. *SoftwareX* 12 (2020), 100578. https://doi.org/10.1016/j.softx.2020.100578

[12] S.L. Zhou, A.A. Shah, P.K. Leung, X. Zhu, and Q. Liao. 2023. A comprehensive review of the applications of machine learning for HVAC. *DeCarbon* 2 (2023), 100023. https://doi.org/10.1016/j.decarb.2023.100023