Machine learning occupancy prediction models - a case study

Conference or Workshop Item

Published Version

Machine Learning Occupancy Prediction Models - A Case Study

Bashar Alfalah, P.Eng
Member ASHRAE

Mehdi Shahrestani, PhD
Member ASHRAE

Li Shao, PhD

ABSTRACT

There is a necessity to reduce energy consumption in the building sector to mitigate the associated carbon emissions, which, globally, account for 39% of total emissions. Several factors have to be considered in the effort to reduce energy consumption. Among them, building occupancy is one of the key drivers in the operation of building services and hence leading to energy consumption and its associated carbon emissions. Information regarding building occupancy at different times of the day or year could potentially help to enhance the energy and environmental performance of buildings. This study investigated the performance of two machine learning models (Hidden Markov Model and Artificial Neural Network) in predicting the occupancy numbers of a high-density higher education building in England that was used as the case study. The models were developed using high-resolution actual occupancy data obtained at five-minute intervals for 12 months, covering workdays, weekends, holidays, and vacations. Occupancy data were collected using high-accuracy infrared video image sensors to cover the gap in the paucity of occupancy data in such buildings and to eliminate the uncertainties faced in previous studies. Several statistical analyses, such as principal component analysis and cross-validation, were conducted to ensure that optimal inputs are used for developing and evaluating the models. The results of the two occupancy prediction models developed indicated that the Hidden Markov Model performs better than the Artificial Neural Network in predicting occupancy numbers.

INTRODUCTION

Building operation schedules vary based on the building type, activity, and opening hours. Buildings such as higher education institution libraries have fixed schedules for building operating systems, with the assumption that the building is fully occupied. For instance, HVAC systems for all zones operate from early morning to the end of the working day, which could lead to significant energy wastage when the building is partially occupied. Such practices could be attributed to a lack of knowledge of occupancy numbers, patterns, and schedules in the building during different periods. This is one of the reasons for the growing interest over the last decade in studies on collecting, estimating, and predicting building occupancy numbers. In addition, the occupancy in certain building types, such as educational buildings, has not been sufficiently examined.

Several studies have developed machine learning models, such as the Hidden Markov Model (HMM) and Artificial Neural Network (ANN), to predict the occupancy numbers of different types of buildings (Alam et al., 2017; Amayri et al., 2017; Li and Dong, 2017; Zuraimi et al., 2017). Han, Gao and Fan (2012) developed an HMM to predict occupancy numbers on specific floors of an office building of a higher education institution. Occupancy data were collected during workdays using passive infrared (PIR), carbon dioxide, and humidity sensors over a period of three weeks. The proposed model predicted four days of occupancy numbers with high accuracy of 96%. In another study, an HMM was developed by Ali and Bouguila (2022) to predict the status of occupants in a single office. Occupancy data were collected for two weeks using light, temperature, humidity, and carbon dioxide sensors to train the model. The results of the model predicted the status of occupants in the office with an accuracy of 96%. Amayri et al. (2017) developed an HMM to

Bashar Alfalah is a Lecturer, in college of architectural and planning, Imam Abdulrahman Bin Faisal University, Saudi Arabia. **Mehdi Shahrestani** is a Lecturer, and **Li Shao** is a professor in School of the Built Environment, University of Reading, Reading, UK.

predict occupancy status in an open-space office area. Occupancy data were collected for ten days using motion detection, energy consumption, and acoustic pressure to train the model. The proposed model predicted the occupancy status for ten days with 89% accuracy.

ANN models have also been developed in previous studies. In a study conducted by Das, Swetapadma and Panigrahi (2017), ANN was developed to predict occupancy status in a single office. Occupancy data were collected using temperature, humidity, light, and carbon dioxide sensors. The developed model predicted occupancy status with high accuracy of 95.6%. Conversely, Ashouri et al. (2019) used Wi-Fi signals in an office to collect two days of occupancy data to train the ANN model. The proposed model predicted the occupancy numbers during four workdays at high accuracy of 90%. In another study, Wang, Chen and Hong (2018) developed an ANN model to predict the occupancy numbers in an office area. Occupancy data were collected for nine days using ambient sensors (including temperature, humidity, and carbon dioxide sensors), CCTV cameras, and Wi-Fi signals. The proposed model predicted occupancy numbers during office hours for three days. Despite these successes in predicting the occupancy numbers, there were many uncertainties in the previous studies associated with the prediction of occupancy numbers based on data collection from temperature, humidity, and light sensors. Similarly, using Wi-Fi signals or manual head counting from the recorded CCTV camera could contain human error due to incorrect counting or the failure of the Wi-Fi connection.

The significant aspects of this study are reflected in the development of the two machine learning models with high accuracy of real-world occupancy data input for the prediction of the number of occupants in an entire high-density higher education building for four months. There were insufficient studies on higher education buildings in the existing literature as the focus was on residential and office buildings. In addition, the focus on higher education buildings was on shorter periods of weeks in a selected space, such as a classroom, lecture room, or specific floor(s) in buildings, resulting in a lower number of occupants in the study area. The occupancy data used in this study were collected using high-accuracy infrared video image sensors for an extended period of 12 months for the entire high-density library building without excluding any periods, such as weekends or vacations, compared to the previous studies that focused on short data collection during working hours only (Tekler et al., 2020). In addition, using infrared video image sensors provides an opportunity to address the issue of uncertainty reported in the previous studies (Mora et al., 2019; Huchuk, Sannera and O'Brien, 2019). The mitigation of the uncertainties associated with occupancy prediction in this study can potentially contribute to mitigating the performance gaps by evaluating the impact of occupants on energy consumption, as higher educational buildings have a high-performance gap of up to 80% (Mahdavi et al., 2021).

MATERIALS AND METHODS

In this study, the case study building of a higher education library building was selected for the collection of occupancy data. Several occupancy attributes were identified from the collected data, which were used together with the occupancy numbers as inputs for the training of the HMM and feed-forward ANN models to predict the occupancy numbers in the building.

Case Study Building

The Urban Region Studies (URS) building is used as a library building located at the University of Reading Whiteknights Campus in England, was selected as the case study building to illustrate the proposed approach of the study. As a new library is being constructed, the URS building has temporarily served as a study area for students. The building was built in 1970 and categorised as a heritage building in 2016. The building comprises seven levels, which are divided between the students and staff. The students have access to the first three levels, while the remaining levels are reserved for staff members and serve as their working spaces. The general specification of the URS building is presented in Table 1. The building is open 24 hours during workdays, and specific hours only during weekends, depending on the academic term.

Table 1. General specification of the URS library building

Location	Reading, England

Size	7200 m2
No. of floors	7
Occupied year	1972
Building Capacity	830

Data Collection

In order to eliminate uncertainties in collecting the number of occupants entering and leaving the building, infrared video image sensors were used to collect the data. The sensors were installed at the main entrance of the building since the remaining entrances were closed during the ongoing refurbishment. The sensors recorded data on the number of occupants at five-minute intervals in real-time for an extended period of 12 months, comprising over 105,000 data points. Compared with previous studies that eliminated some days from the data, the data collection in this study covered the entire academic year, which comprised three consecutive academic terms, without eliminating any occasions, such as vacations, holidays, weekends, or periods with few occupants.

Occupancy Attributes

Several occupancy attributes were identified from the data collected during the 12 months period. As shown in Table 2, values were assigned to each attribute, including day, week, month, and term, which were used as inputs for training the prediction models. The day attribute used values from 1 to 7, representing the seven days of the week. The week attribute used values from 1 to 5, with week five assigned to the extra day not belonging to any of the other four weeks. The month attribute values were from 1 to 12 to represent the 12 months in the year. The academic term attribute represented the three academic terms and three vacations.

Table 2. Identified attributes and their assigned values

Attribute	Value
Day	Monday 1, Tuesday 2, Wednesday 3, Thursday 4, Friday 5, Saturday 6, Sunday 7
Week	Week 1, Week 2, Week 3, Week 4, Week 5
Month	January 1, February 2, March 3, April 4, …, August 8, September 9, October 10, November 11, December 12
Term	Christmas Vacation 1, Spring Term 2, Easter Vacation 3, Summer Term 4, Summer Vacation 5, Autumn Term 6

Statistical Indicators

The performance of the prediction model was evaluated by comparing the prediction results with the actual occupancy data collected from the case study building. Two statistical indicators, namely the root mean square error (RMSE) and mean absolute error (MAE), were used to measure the differences between the two datasets, with a smaller value indicating similarity between the predicted and measured data. The RMSE is calculated using Equation (1), where O is the actual occupancy number, S represents the predicted occupancy number, and n is the total number of data points (Chai and Draxler, 2014). The second method employed to evaluate the accuracy of the prediction results is MAE, which is calculated using Equation (2). The difference between these two indicators is that the MAE does not assign much weight to a high value (Willmott and Matsuura, 2006).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2} \quad (1)$$

$$MAE = \frac{\sum_{i=1}^{n}|S_i - O_i|}{n} \quad (2)$$

MODEL DEVELOPMENT

In this study, two machine learning models (HMM and feed-forward ANN) were developed to predict the occupancy numbers in the case study building at any given period, including academic terms and vacations. The models were developed in MATLAB R2020b environment. In this section, principal component analysis (PCA) (Abdi and Wiliams, 2010) was used to identify the key occupancy attributes for high prediction accuracy, and the cross-validation (CV) method (Bro et al., 2008) was used to determine the optimal split of the datasets for training and validation to ensure highly accurate predictions.

The HMM and feed-forward ANN models were developed using 12 months of occupancy data, along with four of the identified occupancy attributes (shown in Table 2). A PCA analysis is carried out to evaluate the impact of attributes on the accurate prediction of occupancy and to identify the key attributes. Subsequently, CV was conducted to determine the optimal dataset split for training and validation of the prediction models. The dataset was split into 12 folds, and each split was examined to find the best split that could provide an accurate prediction. Figure 1 illustrates the steps of developing the HMM using 12 months of occupancy data as input. Two permutations were used to develop the model; the first permutation is the day, week, month, and term, which are all the identified occupancy attributes from the collected data. The second permutation is day, week, and month, which are the results of performing PCA. From the selected permutations, the domain of the attributes is defined as observable variables (transition), and the hidden variable represents the occupancy numbers (emission). The prediction period was related to the size of the validation dataset, which was defined by performing cross-validation to determine the optimal split of the data for training and validation sets. Based on the inputs, the model predicts the occupancy numbers, followed by an evaluation of the prediction results. The feed-forward neural network proceeded in a forward direction, starting from the input data (occupancy attributes), then through the hidden layer and ending up with the output layer (occupancy numbers) without returning to any of the layers (no loops). Similarly, in developing feed-forward ANN in Figure 2, the inputs of occupancy attributes for both permutations and determining the split of training and validation are the same as HMM. In contrast, the ANN model requires determining the numbers of hidden layers, neurons, and learning rates, which the model used to predict the occupancy numbers. Finally, the results were compared to the actual occupancy data.

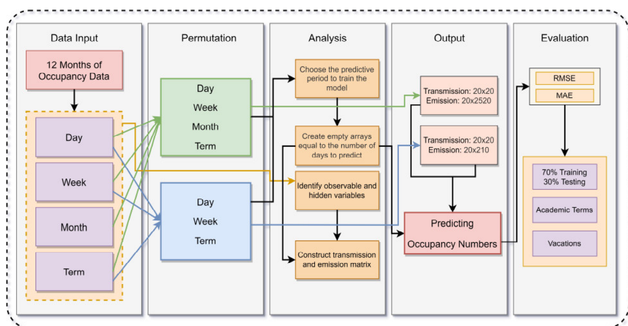

Figure 1 Process of developing HMM in two different permutations.

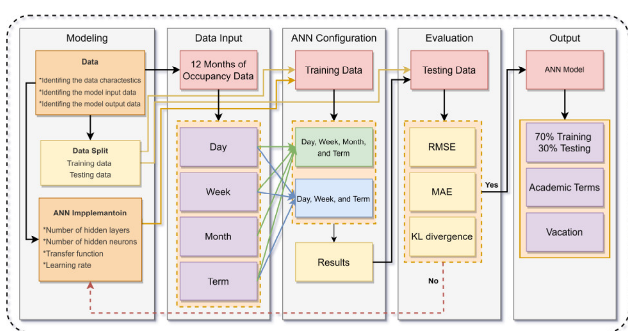

Figure 2 Process of developing the feed-forward ANN model in two different permutations.

RESULTS

This section presents the results of the analyses performed on the collected occupancy data that determined the model inputs and the results of the developed models. Figure 3 shows the results of performing PCA to evaluate the level of influence of each set of occupancy attributes. The darker the colour, the more influence of the attribute on the prediction results, where the white colour means no influence and the crossed box means a specific attribute not included in the permutation. The result of PCA in Figure 3 shows that the permutation of the first, second and fourth attributes of the occupancy data [1 2 4] representing day, week, and academic terms (Table 2) is the optimal input for training the model as all three attributes have an influence on the prediction results compared to other permutations. The results of performing CV to identify the optimal dataset split for training and validation (shown in Figure 4) illustrate the sensitivity of the model accuracy to the increase in the percentage of data used for training the model. The trend shows that 70%/30% and 80%/20% are the optimal training/ validation splits for the dataset collected, which provide the most accurate prediction over longer periods. The selected split of 70%/30% allowed the model to be trained and validated with adequate data.

The results of PCA and CV were used first as inputs to train the models (HMM and ANN) and predict the occupancy number, using 70% of the data for training and 30% for validation. Secondly, one month of occupancy data from each academic term and vacation period was used for training to predict occupancy for the rest of the academic terms and vacation periods. The results of training the HMM and ANN models using the occupancy attributes day, week, month, and the academic terms [1 2 3 4] to predict the occupancy numbers are shown in Figures 5 and 6 consequently. Figures 7 and 8 illustrate the outcomes of HMM and ANN, respectively, using the optimal occupancy attributes day, week, and the academic term [1 2 4].

As an explanation of Figures 5–8, four plots are shown for each figure. The red line in the plots represents the predicted data, and the blue line represents the actual occupancy data. Plot (a) shows the actual and predicted occupancy numbers using 70% of the data for training and 30% for validation. Plots (b), (c), and (d) demonstrate the results of occupancy prediction for Autumn and Spring terms as well as Summer Vacation, where the models are trained using one month of occupancy data to predict the rest of the terms/summer vacations. It should be noted that Autumn and Spring terms are 11 weeks, whereas Summer Vacation is 15 weeks. Tables 3 and 4 present the results of evaluating the prediction performance of ANN and HMM, respectively. The prediction of occupancy numbers was evaluated under two sets of occupancy attribute inputs (day, week, month, and academic term [1 2 3 4]) and (day, week, and academic term [1 2 4]) using RMSE and MAE.

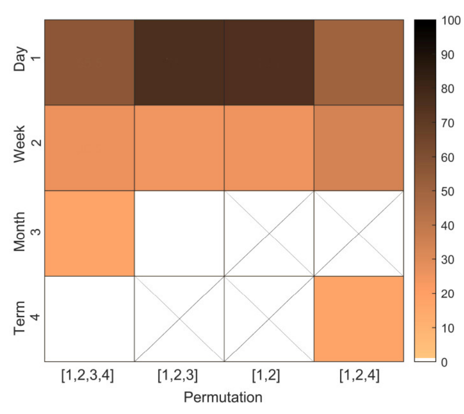

Figure 3 Principal component analysis results considering different permutations of the occupancy attributes, (1) associated with the attribute day of the week, (2) week of the month, (3) month of the year, and (4) the academic term. The cross box indicates the attribute is not used in the permutation.

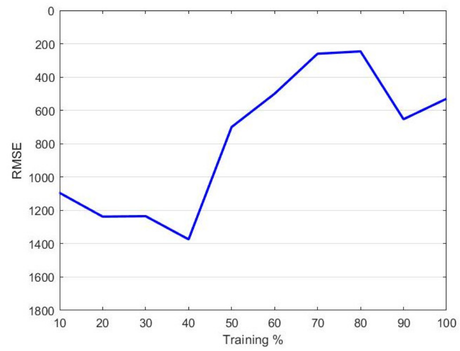

Figure 4 Trend of training the model at different input percentages of the dataset. Showing 70% and 80% got the highest accuracy indicated by RMSE.

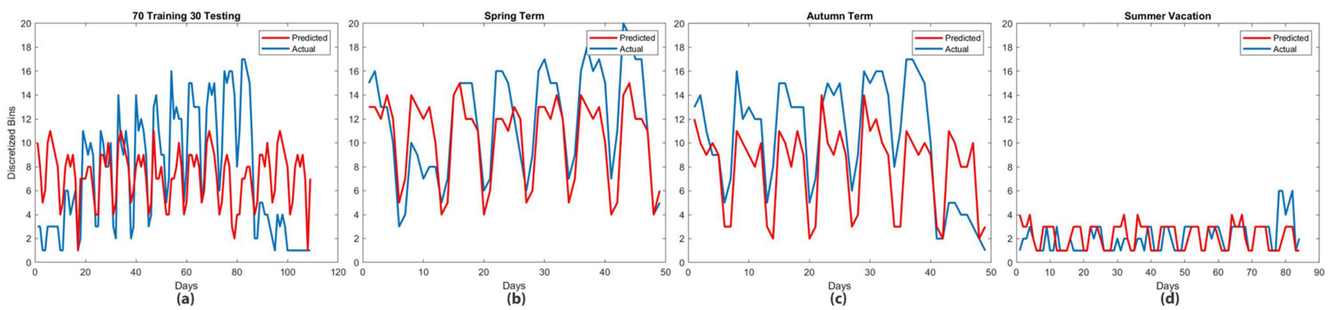

Figure 5 HMM prediction of occupancy numbers using the permutations [day, week, month, and term]; plot (a) predicts the occupancy in 30% of days of the year based on the model trained using the occupancy data for 70% of days in the year, plot (b) outcome of occupancy prediction in the Spring Term where the model is trained for one month and predict the rest of days (49 days), (c) outcome of occupancy prediction in the Autumn Term where the model is trained for one month and predict the rest of days(49 days), and plot (d) outcome of occupancy prediction in the summer vacation where the model is trained for one month and predict the rest of days (77 days).

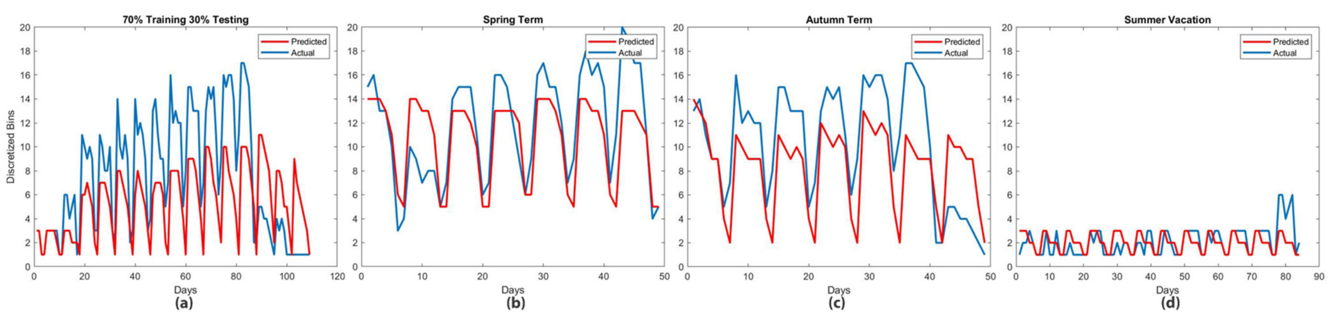

Figure 6 ANN prediction of occupancy number using the permutations [day, week, month, and term], plot (a) predicts the occupancy in 30% of days of the year based on the model trained using the occupancy data for 70% of days in the year, plot (b) outcome of occupancy prediction in the Spring Term where the model is trained for one month and predict the rest of days (49 days), (c) outcome of occupancy prediction in the Autumn Term where the model is trained for one month and predict the rest of days(49 days), and plot (d) outcome of occupancy prediction in the summer vacation where the model is trained for one month and predict the rest of days (77 days).

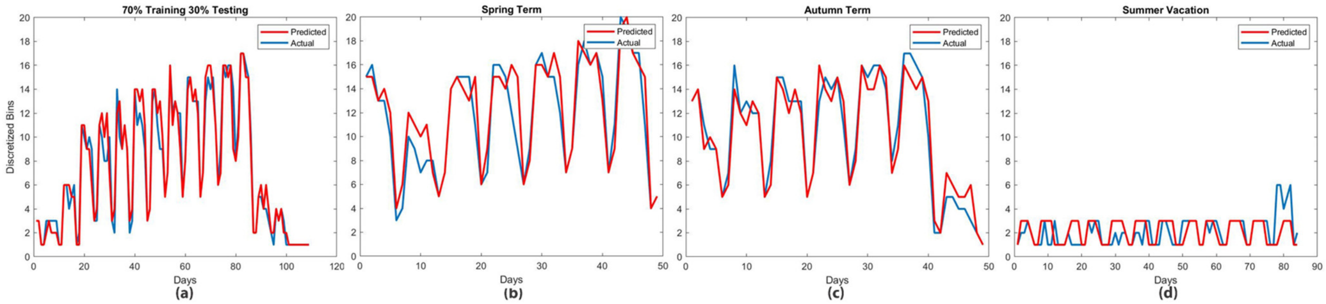

Figure 7 HMM prediction of occupancy number using the permutations [day, week, and term], plot (a) predicts the occupancy in 30% of days of the year based on the model trained using the occupancy data for 70% of days in the year, plot (b) outcome of occupancy prediction in the Spring Term where the model is trained for one month and predict the rest of days (49 days), (c) outcome of occupancy prediction in the Autumn Term where the model is trained for one month and predict the rest of days(49 days), and plot (d) outcome of occupancy prediction in the summer vacation where the model is trained for one month and predict the rest of days (77 days).

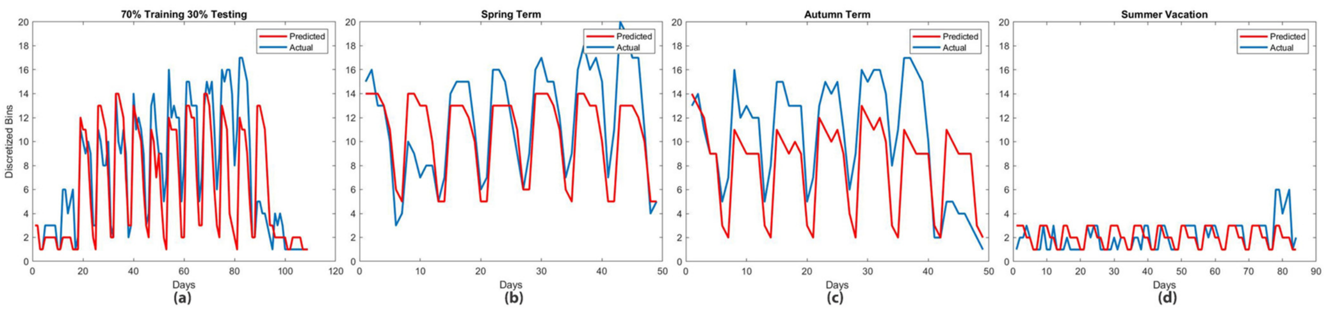

Figure 8 ANN prediction of occupancy number using the permutations [day, week, and term], plot (a) predicts the occupancy in 30% of days of the year based on the model trained using the occupancy data for 70% of days in the year, plot (b) outcome of occupancy prediction in the Spring Term where the model is trained for one month and predict the rest of days (49 days), (c) outcome of occupancy prediction in the Autumn Term where the model is trained for one month and predict the rest of days(49 days), and plot (d) outcome of occupancy prediction in the summer vacation where the model is trained for one month and predict the rest of days (77 days).

Table 3. Evaluation of the prediction results of HMM

Prediction Period	[1 2 3 4]		[1 2 4]	
	RMSE	MAE	RMSE	MAE
70% Training	991	840	260	162
Spring Term	610	541	367	249
Autumn Term	851	743	263	192
Summer Vacation	265	162	249	143

Table 4. Evaluation of the prediction results of ANN

Prediction Period	[1 2 3 4]		[1 2 4]	
	RMSE	MAE	RMSE	MAE
70% Training	893	743	664	472
Spring Term	611	498	562	461

Autumn Term	786	657	539	539
Summer Vacation	227	145	224	143

DISCUSSION

The results of predicting the occupancy data of a high-density building using the HMM and ANN are presented in Figures 5–8. The results of the HMM prediction of the weekly variations using the permutation [1 2 3 4], which is the day, week, month and term, are shown in Figure 5a. However, a discrepancy was found between the measured and predicted data in the first and last 20 days. Conversely, the model developed using the permutation of day, week, and term [1 2 4] shows higher prediction accuracy. Table 3 shows the results of an RMSE of 260 and MAE of 162, indicating a high prediction accuracy using the permutation of day, week, and the term [1 2 4] compared with the model developed using the permutation of day, week, month, and term [1 2 3 4]. In addition, the HMM developed using the day, week, and term attributes of occupancy data predicted the academic terms (Spring and Autumn Terms) and vacation (Summer Vacation) with high accuracy compared with the model trained using the permutation day, week, month, and term [1 2 3 4], as shown in Table 3. Comparably, the ANN results using the permutations of day, week, month, and term [1 2 3 4], and day, week, and the term [1 2 4] as inputs for training the model are similar to the HMM results where the permutations day, week, and term [1 2 4] obtained high prediction accuracy. As shown in Table 4, it confirmed that the model trained using the permutation of day, week, and term [1 2 4] obtained optimised accuracy compared with that of the permutation of day, week, month, and the term [1 2 3 4] in all the prediction periods (academic terms and vacation) and models (HMM and ANN). Both models were able to predict the occupancy numbers of the Autumn Term more accurately compared to the Spring Term, with an RMSE of 611 and MAE of 498.

The results of evaluating the performance of the developed models in predicting occupancy numbers (Tables 3 and 4) show that the performance of HMM is better compared to the ANN model. Several points can be drawn from the evaluation results, such that the optimal permutation (day, week, and term [1 2 4]), determined by performing PCA, obtained the highest accuracy in both models (HMM and ANN). In addition, it was demonstrated that not all the occupancy attributes (day, week, month, and term) were required for training the model to obtain accurate prediction results, as the performance of the prediction models increased when the month attribute was eliminated. This highlights the importance of the identification of principal components in the dataset to be used for training the prediction models and to avoid overfitting the model using dependent attributes.

CONCLUSION

In this study, two machine learning models (HMM and ANN) were developed to predict the occupancy numbers in a high-density higher education building. The optimal split of the data for training and validation of the occupancy prediction models was identified by performing CV. The results of CV indicated 70%/30% as the optimal split of data collected to be used for training/prediction. The developed models were trained based on a set of occupancy attributes, and the best set (permutation of the occupancy attributes) was identified using the PCA technique. The first permutation included all the occupancy attributes (day, week, month, and term) from the data presented in Table 2. The second permutation of occupancy attributes (day, week, and term) was the optimal one that was identified by applying PCA to identify the principal components of the occupancy attributes. The results of predicting the occupancy numbers show that HMM predicted the occupancy numbers more accurately compared to the ANN.

The developed models predicted the occupancy numbers with high accuracy in the entire building, which can be used by different stakeholders, including building space management, sustainability team, and Estates and Facilities Management team. The predicted occupancy numbers provided opportunities for building facility managers to use the space more efficiently, operate the building based on the predicted occupancy numbers, improve energy efficiency, and reduce energy consumption. Moreover, the predicted occupancy numbers could be essential input for developing an energy simulation model. For future studies, it is recommended to use more occupancy attributes, such as the time of the day and location of occupants, to obtain a more detailed prediction of occupancy numbers and use several occupancy datasets from different academic buildings to evaluate the performance of the developed occupancy prediction models.

REFERENCES

Abdi, H. and Williams, L.J., 2010. Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), pp.433-459.

Alam, A.G., Rahman, H., Kim, J.K. and Han, H. 2017. Uncertainties in neural network model based on carbon dioxide concentration for occupancy estimation. Journal of Mechanical Science and Technology 31(5), pp. 2573–2580. doi: 10.1007/s12206-017-0455-z.

Ali, S. and Bouguila, N. 2022. Towards scalable deployment of Hidden Markov models in occupancy estimation: A novel methodology applied to the study case of occupancy detection. Energy and Buildings 254, p. 111594.

Amayri, M., Ngo, Q.D. and Ploix, S., 2017, September. Bayesian network and hidden Markov model for estimating occupancy from measurements and knowledge. In 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS) (Vol. 2, pp. 690-695). IEEE.

Ashouri, A., Newsham, G.R., Shi, Z. and Gunay, H.B. 2019. Day-ahead Prediction of Building Occupancy using WiFi Signals. IEEE International Conference on Automation Science and Engineering 2019-Augus, pp. 1237–1242. doi: 10.1109/COASE.2019.8843224.

Bro, R., Kjeldahl, K., Smilde, A.K. and Kiers, H.A.L., 2008. Cross-validation of component models: a critical look at current methods. Analytical and bioanalytical chemistry, 390(5), pp.1241-1251.

Chai, T. and Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE). Geoscientific Model Development Discussions, 7(1), pp.1525-1534.

Das, S., Swetapadma, A. and Panigrahi, C. 2018. Building Occupancy Detection Using Feed Forward Back-Propagation Neural Networks. Proceedings - 2017 International Conference on Computational Intelligence and Networks, CINE 2017, pp. 63–67. doi: 10.1109/CINE.2017.12.

Han, Z., Gao, R.X. and Fan, Z., 2012, May. Occupancy and indoor environment quality sensing for smart buildings. In 2012 IEEE international instrumentation and measurement technology conference proceedings (pp. 882-887). IEEE.

Huchuk, B., Sanner, S. and O'Brien, W. 2019. Comparison of machine learning models for occupancy prediction in residential buildings using connected thermostat data. Building and Environment 160, p. 106177.

Li, Z. and Dong, B. 2017. A new modeling approach for short-term prediction of occupancy in residential buildings. Building and Environment 121, pp. 277–290.

Mahdavi, A., Berger, C., Amin, H., Ampatzi, E., Andersen, R.K., Azar, E., Barthelmes, V.M., Favero, M., Hahn, J., Khovalyg, D. and Knudsen, H.N., 2021. The role of occupants in buildings' energy performance gap: myth or reality?. Sustainability, 13(6), p.3146.

Mora, D., Fajilla, G., Austin, M.C. and De Simone, M. 2019. Occupancy patterns obtained by heuristic approaches: cluster analysis and logical flowcharts. A case study in a university office. Energy and Buildings 186, pp. 147–168.

Tekler, Z.D., Low, R., Gunay, B., Andersen, R.K. and Blessing, L. 2020. A scalable Bluetooth Low Energy approach to identify occupancy patterns and profiles in office spaces. Building and Environment 171(December 2019), p. 106681.

Wang, W., Chen, J. and Hong, T. 2018. Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings. Automation in Construction 94(March), pp. 233–243.

Willmott, C.J. and Matsuura, K., 2006. On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. International Journal of Geographical Information Science, 20(1), pp.89-102.

Zuraimi, M.S., Pantazaras, A., Chaturvedi, K.A., Yang, J.J., Tham, K.W. and Lee, S.E., 2017. Predicting occupancy counts using physical and statistical Co2-based modeling methodologies. Building and Environment, 123, pp.517-528.