Machine Learning Approaches for Predicting Occupancy Patterns and its Influence on Indoor Air Quality in Office Environments

Amir Hamzah Mohd Shaberi[1], Sumayyah Dzulkifly[2]*, Wang Shir Li[3], Yona Falinie A. Gaus[4]
Faculty of Computing and Meta-Technology, Sultan Idris Education University, Perak, Malaysia[1, 2, 3]
Centre of Embedded Education Green Technology (EduGreen@UPSI), Sultan Idris Education University, Perak, Malaysia[2]
Department of Computer Science, Durham University, Durham, United Kingdom[4]

***Abstract*—It is normal for the modern population to spend 12 hours or more daily indoors where the level of comfort can be moderated. Yet, indoor occupants are similarly exposed to various air pollutants just as outdoors. Indoor air pollution could be detrimental toward the occupant's health noted by the United Nation Environment Programme (UNEP) in the Pollution Action Note, published on 7th of September 2021. According to the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) standards, occupancy patterns could influence indoor air quality. Hence, this paper investigates the utilisation of machine learning algorithms in predicting occupancy patterns against indoor air quality (IAQ) variables such as humidity, temperature, light, and carbon dioxide (CO_2). This study compares the performance of selected machine learning approaches, namely deep learning (LSTM, CNN), regression (ANN) and (SVR) models. In addition, it explores the diverse range of evaluation metrics utilized to evaluate the performance of machine learning in the specific context of Mean Squared Error (MSE) and Mean Absolute Error (MAE). In the training phase, the SVR model achieved the lowest MAE of 0.0826 and MSE of 0.0280 as compared to the other algorithms. The ANN model demonstrated slightly better generalization capabilities in the testing phase, while the LSTM model demonstrated robust performance in the test phase. Overall, the results highlighted the significant impact of occupancy behaviour on Indoor Air Quality (IAQ) variables and underscored the importance of advanced modelling techniques in IAQ monitoring and management, emphasizing the need for tailored approaches to address the complex relationship between occupancy patterns and IAQ variables.**

Keywords—Indoor air quality; occupancy patterns; machine learning; deep learning; regression models; Mean Squared Error; Mean Absolute Error; IAQ monitoring; IAQ management

I. INTRODUCTION

Indoor air quality (IAQ) is a critical component in maintaining occupants' health, impacting the well-being of both the humans and the interior ecosystems. Prolonged exposure to harmful substances in the air could lead to persistent discomfort, severe illnesses, and could lead to respiratory attributed deaths annually [1]. Compromised air quality encompasses of various factors, which includes, but not limited to the concentration of pollutants, such as the cleaning supplies and the building materials [2], the indoor humidity levels [3], the temperature control [4] and the adequacy of the ventilation systems [5]. Unhealthy working space for indoor occupants due to poor IAQ can lead to a range of health issues, such as respiratory problems [6-7], asthma [8] and fatigue [9], which can impact their productivity and overall well-being.

The ability to monitor and control IAQ is essential for building managers to identify potential issues and implement corrective measures [9]. Yet, to some, comprehensive installation of indoor sensing system could incur a hefty cost which could discourage the motivation to maintain healthy indoor air quality. The variation of indoor air quality with respect to the occupancy pattern could lead to complex data analysis which requires the use of machine learning to study their linear or non-linear relationships. In recent years, the integration of machine learning techniques has emerged as a promising avenue for predicting and optimizing air quality, in general. Additionally, as societies grapple with the consequences of air pollution, understanding the effectiveness of machine learning models in this context is imperative [10].

As indoor air pollution (IAP) poses a major risk to human health and is responsible for millions of deaths annually, preserving a good IAQ is significant for the health sector [11]. The intersection studies of IAQ and machine learning have garnered significant attention in recent years as researchers explore innovative ways to leverage machine learning techniques to enhance indoor air quality monitoring and management. For example, authors in the study [11], state that machine learning technologies are highly capable of providing real-time indoor air quality monitoring, which is essential for determining and managing indoor air pollutants. In addition, [12] also states that sets of algorithms are utilized to extract and filter general principles from massive datasets, allowing for the automated learning of user preferences in relation to the IAQ.

Despite of this, there is a gap in current literatures, specifically in terms of the employment of machine learning techniques to predict occupancy patterns within the context of indoor air quality management. While previous studies have highlighted the potential of machine learning for real-time monitoring and general data analysis, few have focused on its application to predict occupancy patterns, which can be crucial for understanding indoor air quality dynamics. Therefore, this study aims to fill this gap by investigating the effectiveness of machine learning approaches in predicting occupancy patterns based on variables such as humidity, temperature, light, and carbon dioxide (CO_2). In addition, this study will determine which algorithms performed better and using selected evaluation metrics, namely, the Mean Squared Error (MSE) and Mean Absolute Error (MAE).

The paper is organized into five main sections, namely the Introduction in Section I**,** where the overall background of the research is elaborated; the Literature Review in Section II, which provides extensive reviews on deep learning, regression, classification model, and the relationship between IAQ and occupancy behaviour; the Methodology in Section III, which elaborates on the data acquisition, data training, data testing and selected machine learning algorithms' evaluation approaches; the Results and Analysis in Section IV, which provides in depth evaluations and discussions on the overall analysis of this study; and Conclusion in Section V, which sums up the investigation and highlights the key topics of this study.

II. Literature Review

A. Comparison between Deep Learning, Regression and Classification Models.

IAQ monitoring is a critical aspect of ensuring healthy and comfortable indoor environments, particularly in settings such as homes, offices, and schools [13]. With the increasing prevalence of indoor air pollutants and their impact on human health, there is a growing need for advanced predictive models to accurately monitor and forecast occupancy patterns because occupancy patterns are closely related to the variables of IAQ. In this comparison models, this study delves into the field of deep learning, regression, and classification models, exploring their method and capabilities in the context of occupancy patterns prediction.

1) Deep learning model. Deep learning, a subset of machine learning, has emerged as a powerful tool for processing complex data and extracting meaningful patterns [14]. By leveraging deep neural networks, deep learning models can effectively analyze large volumes of IAQ data, including particulate matter (PM) [15–18] volatile organic compounds (VOCs) [17], CO_2 [17], and sulfur dioxide (SO_2) [19, 20], to provide real-time insights into IAQ levels. These models can learn intricate relationships within the data, enabling them to make accurate predictions and identify potential IAQ issues before they escalate. However, for specific occupancy patterns [21] mentioned that predictions for occupancy were carried out using various deep learning architectures, such as Deep Neural Network (DNN), Long Short-Term Memory (LSTM), Bi-directional LSTM (Bi-LSTM), Gated Recurrent Unit (GRU), and Bi-directional GRU (Bi-GRU), in different settings like an office, library and lecture room. The results demonstrated that the feature selection algorithm proposed performed better than a commonly used one, leading to higher model performance while requiring fewer sensors.

Other than that, from 11 studies [15–20, 22–26] employing various architectures such as long short-term memory (LSTM), Convolutional Neural Network (CNN), unique combinations like Combined Self-Attention (SA) mechanism, Empirical Mode Decomposition (EMD) algorithm and LSTM network (SA–EMD–LSTM) with Ensemble Empirical Mode Decomposition-Sparrow Search Algorithm (EEMD-SSA-LSTM). The performance metrics for these studies include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Symmetric Mean Absolute Percentage Error (SMAPE), Mean Absolute Percentage Error (MAPE), Index of Agreement (IA), Theil Inequality Coefficient (TIC), coefficient of determination (R2) and Absolute Average Deviation. This diversity in metrics underscores the complexity and multifaceted nature of evaluating deep learning models in the context of air quality prediction, considering both spatial and temporal aspects.

2) Regression model. Regression models, on the other hand, are well-suited for predicting continuous IAQ variables, such as PM concentration [27], temperature [28] or occupancy pattern [29] based on historical data and other relevant variables. By fitting a regression model to IAQ data, regression model can create a mathematical equation that describes the relationship between the input variables and the output [30], allowing to make predictions with a high degree of accuracy.

For example, six studies [27, 28, 31–34] employ diverse techniques such as Artificial Neural Network (ANN), Extra Trees Regressor, Wavelet Artificial Neural Network, and Support Vector Regression (SVR). Evaluation metrics for these models include MAE, RMSE, SMAPE, MAPE, and Pearson correlation coefficient (R). The inclusion of regression models adds a valuable dimension to the literature, as these models provide insights into continuous air quality prediction. In addition, the study in [29] also states linear and logistic regression models were created using the variables to forecast occupant activity schedules and the probability of occupant presence.

3) Classification models. Classification models are particularly useful for identifying discrete IAQ states, such as air quality categories (e.g., good, moderate, poor) or the presence of specific pollutants [35]. These models classify IAQ data into different classes based on their characteristics, enabling us to categorize IAQ levels and take appropriate actions to improve indoor air quality. Previous study primarily leverages random forest [31, 36–38] and decision tree [32] while employing a range of evaluation metrics such as specificity, precision, recall, accuracy, F1 score, area under the ROC Curve (AUC), and sensitivity. This reflects a robust assessment strategy to measure the performance of these models in predicting air quality.

While it is possible for a classification model to estimate the occupancy patterns, there are some important considerations and limitations. In a typical classification problem, the goal is to predict the class label of an observation based on its features. So, this study will not include classification models, as the focus is on utilizing deep learning and regression models, which are deemed more important for the study.

B. Relationship between IAQ and Occupancy Behavior

The relationship between IAQ and occupancy behavior is complex and symbiotic. The way in which occupants use spaces, their activities, and their preferences all have an impact on the IAQ, which includes temperature, humidity, and pollutants [39]. Occupancy behavior plays a crucial role in determining pollutant emissions, ventilation needs and desired comfort levels. For instance, spaces with more occupants engaged in activities that generate pollutants may necessitate higher ventilation rates [40]. Managing IAQ effectively involves comprehending these interconnections and devising strategies that could harmonise occupants' behavior while maintaining optimal indoor air quality.

To encourage a healthier indoor environment, a comprehensive strategy that considers both IAQ variables and occupancy behavior is imperative. According to [41], architectural designs should cater to varied activities and occupant densities, optimizing factors like temperature control to curb discomfort and air stagnation. Incorporating intelligent building technologies to monitor and regulate IAQ variables in real time based on occupancy patterns can further amplify the interplay between IAQ and occupant behavior [42]. Ultimately, cultivating an atmosphere where occupants are conscientious about their actions and where IAQ commands shared attention can culminate in an enhanced IAQ and overall well-being.

Given the intricate relationship between IAQ and occupancy behavior, this study aims to predict occupancy patterns in relation to selected IAQ variables. Understanding how occupants utilize spaces and engage in activities that affect IAQ variables like temperature, humidity, and pollutant levels is crucial for effective IAQ management. By leveraging the predictive modelling, this study seeks to develop insights into how occupancy patterns influence the IAQ dynamics, ultimately contributing to strategies that enhance the indoor environmental quality and occupant well-being.

III. Methodology

In the context of predicting occupancy patterns, the model training for deep learning and regression models can be similar, as both approaches aim to predict occupancy based on input variables such as temperature, humidity, light and CO_2 levels. The phase would involve several steps as shown in the Fig. 1 below. However, the specific implementation details and the model architectures would differ between the deep learning and regression, depending on the algorithms used. To choose a better algorithm that suited the aim of this study, data collection, data pre-processing, model training and testing, and model evaluation will be conducted. Fig. 1 below shows the phases involved in this study to predict occupancy pattern in relation to the IAQ variables.

A. Phase 1 (Data Collection)

For the data collection stage, the historical data for the variables such as temperature, humidity, light, CO_2, and occupancy will be gathered. It is crucial to ensure that the data is representative and covers a wide range of values to capture the variability of the IAQ. This comprehensive dataset will serve as the foundation for training and evaluating the deep learning and regression models for predicting occupancy patterns. This study addresses key variables present in office room environment, with a focus on identifying an acceptable range for these variables that may pose health risks to occupants if exceeded.

B. Phase 2 (Data Pre-Processing)

In the data pre-processing phase, the input features were normalized to ensure data consistency. This normalization step is crucial for deep learning and regression models, as it helps to prevent features with larger scales from dominating the training process. Additionally, the data were split into the training (70%) and the testing sets (30%). The training set was used to train the models, while the testing set was used to evaluate their performance. This split is essential to assess how well the generalization of the models were towards unseen data and to prevent overfitting. Overall, these preprocessing steps help to ensure the models can effectively learn from the data.

C. Phase 3 (Model Training and Testing)

This stage implements a custom callback, 'MetricsCallback', to calculate and print MAE and MSE during the training and testing phases of LSTM, CNN, and ANN models. This callback was used to monitor the model's performance on the training and the validation sets. However, the SVR model is not a neural network and therefore does not use the same training process. For such case, the MAE and the MSE were calculated manually after the training completed.

The training process involves splitting the dataset into the training and the testing sets, scaling the features, and reshaping the data for the LSTM and CNN models. The models are then compiled and trained using the fit method, with the callback used to print the loss, the MAE, and the MSE at the end of each epoch. The detailed machine learning parameters used for each model are as follows:

LSTM Model

- Sequential model with an LSTM layer (64 units) and a dense output layer.
- Compiled with Mean Squared Error (MSE) loss and Adam optimizer.
- Trained for 50 epochs with a batch size of 32.
- Validation data is specified for monitoring performance during training.
- Callback is used to monitor and print MAE and MSE during training and validation.

CNN Model

- Sequential model with a 2D convolutional layer (32 filters, 2x2 kernel size, ReLU activation), a flattening layer, and a dense output layer.
- Compiled and trained similarly to the LSTM model.

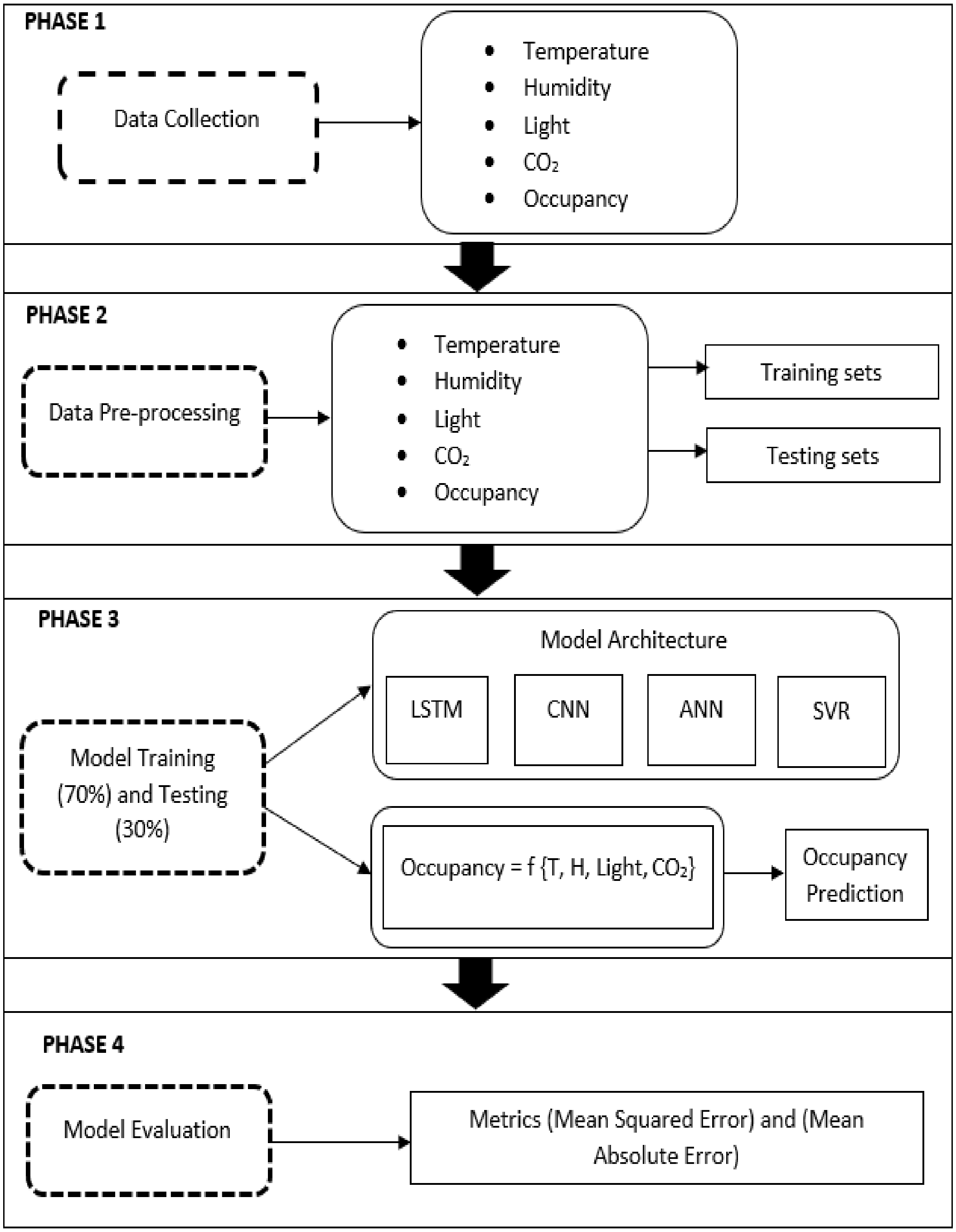

Fig. 1. Phases of predicting occupancy pattern.

ANN Model (Regression)

- Sequential model with two dense layers (64 units, ReLU activation) and a dense output layer.
- Compiled and trained similarly to the LSTM and CNN models.

SVR Model

- SVR model trained separately using the SVR class from scikit-learn.
- Does not use the custom callback as it does not follow the same training process as neural networks.

Once the models completed the training and testing stage, they were evaluated using the MAE and the MSE to observe the difference in term of the performance. This approach provides a comprehensive analysis of the models' performance, allowing a comparison for their ability to predict occupancy pattern based on the given IAQ variables.

D. Phase 4 (Model Evaluation)

For model evaluation, the trained model is assessed using the testing dataset to gauge its performance in predicting occupancy. Metrics such as the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) are used. The MSE provides a measure of the average squared difference between the predicted occupancy values and the actual values, offering insight into the model's overall accuracy.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (yi - \bar{y}i)^2 \tag{1}$$

where:

- n is the number of observations.
- yi is the actual value of the target variable for the i-th observation.
- $\bar{y}i$ is the predicted value of the target variable for the i-th observation.

On the other hand, the MAE measures the average absolute difference between the predicted and the actual values, providing an indication of the model's precision. These metrics collectively offer a comprehensive assessment of the model's performance in predicting occupancy values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |yi - \bar{y}i| \tag{2}$$

where:

- n is the number of observations.
- yi is the actual value of the target variable for the i-th observation.
- $\bar{y}i$ is the predicted value of the target variable for the i-th observation.

IV. Results and Discussion

The discussion for this study focused on comparisons between the prediction results of various machine learning algorithms for occupancy patterns based on IAQ variables including an analysis of their time and memory complexities. In addition, this study also explores the impact of occupancy behavior on IAQ variables. Key findings included the effectiveness of certain algorithms in predicting occupancy patterns and how changes in IAQ variables were influenced by occupancy.

A. Prediction Results for Training and Testing Data Across Different Machine Learning Models

Comparing the training and the testing results is essential in machine learning to assess how well a model performs in analyzing the data. Study by [43] mentioned that during the training, a model learns to map input features to output labels using the provided data. However, this process can lead to overfitting, where the model memorizes the training data instead of learning the underlying patterns. By evaluating the models' performance on a separate testing dataset, one can gauge its ability to generalize. Discrepancies between training and testing results indicate potential overfitting, highlighting the need for adjustments such as hyperparameter tuning [43]. Additionally, comparing results helps in model selection, as the best-performing model on testing data is typically chosen for deployment.

Tables 1 and 2 compare the performance of four different models (LSTM, CNN, ANN, SVR) in the training and the testing phases using the MAE and the MSE metrics. In the training phase (see Table 1), the SVR model achieved the lowest MAE, namely 0.0826 and MSE of 0.0280, indicating better performance in predicting occupancy pattern based on the input features (CO_2, Light, Temperature, Humidity) compared to the other models. The ANN model also performed well, with a MAE of 0.0940 and MSE of 0.0353, followed by the CNN and LSTM models.

In the testing phase (see Table 2), the ANN model demonstrated the best performance, achieving the lowest MAE 0.0834 and MSE 0.0364. This indicates that the ANN model was more accurate in predicting occupancy pattern on unseen data compared to the other models. The CNN model also performed well in the testing phase, with an MAE of 0.0866 and MSE of 0.0385, followed by the LSTM and SVR models.

TABLE. I. Comparison of MAE and MSE for Training Data

Model	MAE	MSE
LSTM	0.1153	0.0474
CNN	0.1024	0.0419
ANN	0.0940	0.0353
SVR	0.0826	0.0280

TABLE. II. Comparison of MAE and MSE for Testing Data

Model	MAE	MSE
LSTM	0.0977	0.0420
CNN	0.0866	0.0385
ANN	0.0834	0.0364
SVR	0.0968	0.0411

Overall, the ANN and CNN models demonstrated robust performance in predicting the occupancy patterns, with the ANN model showing slightly better generalization capabilities in the testing phase.

B. Complexity Analysis: Time and Memory

Table 3 below provides the time and memory complexities for each of the models used in this study. For the time complexity, it is measured in seconds and represents the duration taken by each model to complete the training process. The LSTM model took the longest time at 0.2044 seconds, followed closely by the CNN model at 0.1921 seconds. The ANN model was next, at 0.1342 seconds, and the SVR model was the fastest at only 0.0156 seconds. These time complexities give an indication of how efficiently each model can process and learn from the training data.

In terms of memory complexity, measured in MiB (Mebibytes), it represents the peak memory usage during the training process. Interestingly, all three models (CNN, ANN, and SVR) exhibited very similar memory usage, ranging from approximately 723 MiB to 738 MiB. This suggests that these models require a similar amount of memory to store and process the data during training. The LSTM model, on the other hand, showed a slightly higher memory usage of 738.836 MiB, indicating that it may require a bit more memory compared to the other models.

Overall, these complexities provide insights into the efficiency and resource requirements of each model, which can be valuable for selecting the most suitable model for a given application based on computational resources and time constraints.

TABLE. III. TIME AND MEMORY COMPLEXITY COMPARISON FOR EACH MODEL

Model	Time Complexity (s)	Memory Complexity (MiB)
LSTM	0.2044	738.836
CNN	0.1921	723.012
ANN	0.1342	723.246
SVR	0.0156	723.086

C. Exploring the Impact of Occupancy Behavior on Indoor Air Quality Variables

The comparison of machine learning algorithms for predicting occupancy patterns based on IAQ variables provided valuable insights into their effectiveness and efficiency. The study focused on four models: LSTM, CNN, ANN, and SVR, evaluating their performance using MAE and MSE metrics in both training and testing phases.

In the training phase, the SVR model exhibited the lowest MAE and MSE, indicating superior performance in predicting occupancy patterns. The ANN model also performed well, followed by the CNN and LSTM models. However, in the testing phase, the ANN model demonstrated the best performance, achieving the lowest MAE and MSE. This suggests that the ANN model was more accurate in predicting occupancy patterns on unseen data, highlighting its superior generalization capabilities compared to the other models.

Furthermore, the study analyzed the time and memory complexities of each model. The LSTM and CNN models exhibited longer training times compared to the ANN and SVR models. In terms of memory complexity, all models (CNN, ANN, and SVR) showed similar memory usage, while the LSTM model required slightly more memory. These complexities provide insights into the computational efficiency and resource requirements of each model, which are crucial considerations for real-world applications.

Overall, these findings underscore the importance of selecting the right machine learning model for predicting occupancy patterns based on IAQ variables. The study's results can guide future research in optimizing IAQ monitoring systems and prediction algorithms, ultimately leading to improved indoor air quality and occupant comfort.

V. CONCLUSION AND FUTURE ENHANCEMENT

This study demonstrates the importance of considering occupancy behavior in predicting IAQ patterns. The results highlight the effectiveness of machine learning algorithms, particularly ANN and CNN, in accurately predicting occupancy patterns based on IAQ variables. ANN emerged as the most accurate algorithm, followed by CNN, LSTM and SVR. These findings underscore the significance of advanced modeling techniques in IAQ monitoring and management, emphasizing the need for tailored approaches to address the complex relationship between occupancy behavior and IAQ variables. In addition, integrating machine learning models into IAQ management strategies can lead to improved indoor environmental quality and occupant well-being.

This study also highlights the promising potential of relationship between IAQ variables and occupancy pattern. This study uses machine learning algorithms in predicting occupancy patterns within office room environments. This research aims to address this deficiency by examining how machine learning methods can predict occupancy patterns using factors like humidity, temperature, light, and CO_2 levels. If the IAQ variables increase, it aims to determine if there's a corresponding increase in occupancy in room environments. It will also identify the most effective algorithms for this task and explore the various evaluation metrics, particularly focusing on MSE and MAE.

Despite the successes observed, it's essential to acknowledge the limitations inherent in this study, particularly the reliance on data collected from external sources rather than proprietary datasets specific to the office building under investigation. The occupancy value is binary, either 0 or 1, which means the prediction can only predict these two states. Secondly, the data size and variables may not be sufficient to provide highly accurate results. Lastly, in this study, there is no available measurement of the room space, which is imperative when studying the impact of IAQ against occupancy pattern.

Future enhancements could involve obtaining proprietary datasets specific to the office building under investigation. This can provide more accurate and relevant data for analysis. Then, increase the dataset size by collecting more data over a longer period. Additionally, consideration could be made to add more variables that may impact indoor air quality and occupancy patterns. This study must consider using indirect methods to estimate occupancy, such as analyzing patterns of other variables that are correlated with occupancy, like motion sensors. Lastly, to calculate the volume of the space, typically need the dimensions of the room (length, width, and height). If there have access to the physical space, it can measure these dimensions directly.

ACKNOWLEDGMENT

This research is funded by the Malaysian Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS), grant number 2022-0092-104-02 (FRGS/1/2022/TK08/UPSI/02/1).

REFERENCES

[1] 1. Almetwally, A.A., Bin-Jumah, M., Allam, A.A.: Ambient air pollution and its influence on human health and welfare: an overview, (2020). https://doi.org/10.1007/s11356-020-09042-2.

[2] Wall, D., McCullagh, P., Cleland, I., Bond, R.: Development of an Internet of Things solution to monitor and analyse indoor air quality. Internet of Things (Netherlands). 14, (2021). https://doi.org/10.1016/j.iot.2021.100392.

[3] Majdi, A., Alrubaie, A.J., Al-Wardy, A.H., Baili, J., Panchal, H.: A novel method for Indoor Air Quality Control of Smart Homes using a Machine learning model. Advances in Engineering Software. 173, (2022). https://doi.org/10.1016/j.advengsoft.2022.103253.

[4] Sulzer, M., Christen, A., Matzarakis, A.: Predicting indoor air temperature and thermal comfort in occupational settings using weather forecasts, indoor sensors, and artificial neural networks. Build Environ. 234, (2023). https://doi.org/10.1016/j.buildenv.2023.110077.

[5] Taheri, S., Hosseini, P., Razban, A.: Model predictive control of heating, ventilation, and air conditioning (HVAC) systems: A state-of-the-art review. Journal of Building Engineering. 60, (2022). https://doi.org/10.1016/j.jobe.2022.105067.

[6] Jain, N., Burman, E., Stamp, S., Shrubsole, C., Bunn, R., Oberman, T., Barrett, E., Aletta, F., Kang, J., Raynham, P., Mumovic, D., Davies, M.: Building performance evaluation of a new hospital building in the uk: Balancing indoor environmental quality and energy performance. Atmosphere (Basel). 12, (2021). https://doi.org/10.3390/ATMOS12010115.

[7] Felgueiras, F., Mourão, Z., Moreira, A., Gabriel, M.F.: Indoor environmental quality in offices and risk of health and productivity complaints at work: A literature review. Journal of Hazardous Materials Advances. 10, (2023). https://doi.org/10.1016/j.hazadv.2023.100314.

[8] Paleologos, K.E., Selim, M.Y.E., Mohamed, A.M.O.: Indoor air quality. In: Pollution Assessment for Sustainable Practices in Applied Sciences and Engineering. pp. 405–489. Elsevier (2020). https://doi.org/10.1016/B978-0-12-809582-9.00008-6.

[9] Cheng, J.C.P., Kwok, H.H.L., Li, A.T.Y., Tong, J.C.K., Lau, A.K.H.: BIM-supported sensor placement optimization based on genetic algorithm for multi-zone thermal comfort and IAQ monitoring. Build Environ. 216, (2022). https://doi.org/10.1016/j.buildenv.2022.108997.

[10] Karimi, S., Asghari, M., Rabie, R., Emami Niri, M.: Machine learning-based white-box prediction and correlation analysis of air pollutants in proximity to industrial zones. Process Safety and Environmental Protection. 178, 1009–1025 (2023). https://doi.org/10.1016/j.psep.2023.08.096.

[11] Van Tran, V., Park, D., Lee, Y.C.: Indoor air pollution, related human diseases, and recent trends in the control and improvement of indoor air quality, (2020). https://doi.org/10.3390/ijerph17082927.

[12] Salih Hasan, B.M., Abdulazeez, A.M.: A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. Journal of Soft Computing and Data Mining. 02, (2021). https://doi.org/10.30880/jscdm.2021.02.01.003.

[13] Saini, J., Dutta, M., Marques, G.: A comprehensive review on indoor air quality monitoring systems for enhanced public health, (2020). https://doi.org/10.1186/s42834-020-0047-y.

[14] Al-Amri, R., Murugesan, R.K., Man, M., Abdulateef, A.F., Al-Sharafi, M.A., Alkahtani, A.A.: A review of machine learning and deep learning techniques for anomaly detection in iot data, (2021). https://doi.org/10.3390/app11125320.

[15] Mengara Mengara, A.G., Park, E., Jang, J., Yoo, Y.: Attention-Based Distributed Deep Learning Model for Air Quality Forecasting. Sustainability (Switzerland). 14, (2022). https://doi.org/10.3390/su14063269.

[16] Zhang, K., Yang, X., Cao, H., Thé, J., Tan, Z., Yu, H.: Multi-step forecast of PM2.5 and PM10 concentrations using convolutional neural network integrated with spatial–temporal attention and residual learning. Environ Int. 171, (2023). https://doi.org/10.1016/j.envint.2022.107691.

[17] Gabriel, M., Auer, T.: LSTM Deep Learning Models for Virtual Sensing of Indoor Air Pollutants: A Feasible Alternative to Physical Sensors. Buildings. 13, (2023). https://doi.org/10.3390/buildings13071684.

[18] Abirami, S., Chitra, P.: Regional air quality forecasting using spatiotemporal deep learning. J Clean Prod. 283, (2021). https://doi.org/10.1016/j.jclepro.2020.125341.

[19] Prado-Rujas, I.I., García-Dopico, A., Serrano, E., Córdoba, M.L., Pérez, M.S.: A multivariable sensor-agnostic framework for spatio-temporal air quality forecasting based on Deep Learning. Eng Appl Artif Intell. 127, (2024). https://doi.org/10.1016/j.engappai.2023.107271.

[20] Zaini, N., Ahmed, A.N., Ean, L.W., Chow, M.F., Malek, M.A.: Forecasting of fine particulate matter based on LSTM and optimization algorithm. J Clean Prod. 427, (2023). https://doi.org/10.1016/j.jclepro.2023.139233.

[21] Tekler, Z.D., Chong, A.: Occupancy prediction using deep learning approaches across multiple space types: A minimum sensing strategy. Build Environ. 226, (2022). https://doi.org/10.1016/j.buildenv.2022.109689.

[22] Zeng, Y., Chen, J., Jin, N., Jin, X., Du, Y.: Air quality forecasting with hybrid LSTM and extended stationary wavelet transform. Build Environ. 213, (2022). https://doi.org/10.1016/j.buildenv.2022.108822.

[23] Yuan, E., Yang, G.: SA–EMD–LSTM: A novel hybrid method for long-term prediction of classroom PM2.5 concentration. Expert Syst Appl. 230, (2023). https://doi.org/10.1016/j.eswa.2023.120670.

[24] Sarkar, N., Gupta, R., Keserwani, P.K., Govil, M.C.: Air Quality Index prediction using an effective hybrid deep learning model. Environmental Pollution. 315, (2022). https://doi.org/10.1016/j.envpol.2022.120404.

[25] Gokul, P.R., Mathew, A., Bhosale, A., Nair, A.T.: Spatio-temporal air quality analysis and PM2.5 prediction over Hyderabad City, India using artificial intelligence techniques. Ecol Inform. 76, (2023). https://doi.org/10.1016/j.ecoinf.2023.102067.

[26] Sun, X., Tian, Z., Zhang, Z.: A new decomposition-integrated air quality index prediction model. Earth Sci Inform. 16, 2307–2321 (2023). https://doi.org/10.1007/s12145-023-01028-1.

[27] Pozo-Luyo, C.A., Cruz-Duarte, J.M., Amaya, I., Ortiz-Bayliss, J.C.: Forecasting PM2.5 concentration levels using shallow machine learning models on the Monterrey Metropolitan Area in Mexico. Atmos Pollut Res. 14, (2023). https://doi.org/10.1016/j.apr.2023.101898.

[28] Kim, M.H., Park, H.J.: Application of artificial neural networks using sequential prediction approach in indoor airflow prediction. Journal of Building Engineering. 69, (2023). https://doi.org/10.1016/j.jobe.2023.106319.

[29] Yan, B., Yang, W., He, F., Zeng, W.: Occupant behavior impact in buildings and the artificial intelligence-based techniques and data-driven approach solutions, (2023). https://doi.org/10.1016/j.rser.2023.113372.

[30] Korsavi, S.S., Montazami, A., Mumovic, D.: The impact of indoor environment quality (IEQ) on school children's overall comfort in the UK; a regression approach. Build Environ. 185, (2020). https://doi.org/10.1016/j.buildenv.2020.107309.

[31] Cheng, C.H., Tsai, M.C.: An Intelligent Time Series Model Based on Hybrid Methodology for Forecasting Concentrations of Significant Air Pollutants. Atmosphere (Basel). 13, (2022). https://doi.org/10.3390/atmos13071055.

[32] Sethi, J.K., Mittal, M.: Ambient Air Quality Estimation Using Supervised Learning Techniques. EAI Endorsed Transactions on Scalable Information Systems. 6, 1–10 (2019). https://doi.org/10.4108/eai.13-7-2018.159406.

[33] Guo, Q., He, Z., Wang, Z.: Simulating daily PM2.5 concentrations using wavelet analysis and artificial neural network with remote sensing and surface observation data. Chemosphere. 340, (2023). https://doi.org/10.1016/j.chemosphere.2023.139886.

[34] Koo, J.W., Wong, S.W., Selvachandran, G., Long, H.V., Son, L.H.: Prediction of Air Pollution Index in Kuala Lumpur using fuzzy time series and statistical models. Air Qual Atmos Health. 13, 77–88 (2020). https://doi.org/10.1007/s11869-019-00772-y.

[35] Rastogi, K., Lohani, D.: Context-aware IoT-enabled framework to analyse and predict indoor air quality. Intelligent Systems with Applications. 16, (2022). https://doi.org/10.1016/j.iswa.2022.200132.

[36] Tella, A., Balogun, A.L., Adebisi, N., Abdullah, S.: Spatial assessment of PM10 hotspots using Random Forest, K-Nearest Neighbour and Naïve Bayes. Atmos Pollut Res. 12, (2021). https://doi.org/10.1016/j.apr.2021.101202.

[37] Alhathloul, S.H., Mishra, A.K., Khan, A.A.: Low visibility event prediction using random forest and K-nearest neighbor methods. Theor Appl Climatol. (2023). https://doi.org/10.1007/s00704-023-04697-6.

[38] Mahmud, S., Ridi, T.B.I., Miah, M.S., Sarower, F., Elahee, S.: Implementing Machine Learning Algorithms to Predict Particulate Matter (PM2.5): A Case Study in the Paso del Norte Region. Atmosphere (Basel). 13, (2022). https://doi.org/10.3390/atmos13122100.

[39] Kim, J., Bang, J. Il, Choi, A., Moon, H.J., Sung, M.: Estimation of Occupancy Using IoT Sensors and a Carbon Dioxide-Based Machine Learning Model with Ventilation System and Differential Pressure Data. Sensors. 23, (2023). https://doi.org/10.3390/s23020585.

[40] Vassella, C.C., Koch, J., Henzi, A., Jordan, A., Waeber, R., Iannaccone, R., Charrière, R.: From spontaneous to strategic natural window ventilation: Improving indoor air quality in Swiss schools. Int J Hyg Environ Health. 234, (2021). https://doi.org/10.1016/j.ijheh.2021.113746.

[41] Ngarambe, J., Yun, G.Y., Santamouris, M.: The use of artificial intelligence (AI) methods in the prediction of thermal comfort in buildings: energy implications of AI-based thermal comfort controls, (2020). https://doi.org/10.1016/j.enbuild.2020.109807.

[42] Kim, J., Zhou, Y., Schiavon, S., Raftery, P., Brager, G.: Personal comfort models: Predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning. Build Environ. 129, 96–106 (2018). https://doi.org/10.1016/j.buildenv.2017.12.011.

[43] Garbin, C., Zhu, X., Marques, O.: Dropout vs. batch normalization: an empirical study of their impact to deep learning. Multimed Tools Appl. 79, 12777–12815 (2020). https://doi.org/10.1007/s11042-019-08453-9.