INTERNATIONAL JOURNAL OF INTEGRATED ENGINEERING
ISSN: 2229-838X e-ISSN: 2600-7916

IJIE

Vol. 17 No. 1 (2025) 14-29
https://publisher.uthm.edu.my/ojs/index.php/ijie

Enhancing Energy Consumption Prediction by Integrating Occupant Activity with Machine Learning Models

Farzana Sharmin Nila[1], Wooi-Haw Tan[1]*, Chee-Pun Ooi[1], Muhammad Umair[1], Yi-Fei Tan[1], Soon-Nyean Cheong[2]

[1] *Faculty of Engineering,*
Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, MALAYSIA
[2] *Faculty of Creative Multimedia,*
Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, MALAYSIA

*Corresponding Author: twhaw@mmu.edu.my
DOI: https://doi.org/10.30880/ijie.2025.17.02.002

Article Info

Received: 10 November 2024
Accepted: 28 June 2025
Available online: 18 July 2025

Keywords

Synchronization, data collection, machine learning models, prediction, occupant activity

Abstract

The precision of the forecast of the power consumption of buildings is essential for big constructions in the present day. However, many of the models in use fail to consider the effect of people's activities within the building on energy consumption. To overcome this limitation, this paper uses a synchronized data collection approach to collect data from different sensors about occupancy activity and power consumption. Several machine learning models are employed with this coordinated data, and the effects of occupant behaviour on power usage are explored. By analyzing the results of the models generated by the two algorithms, the best ways of reaching behaviour-sensitive power consumption prediction are determined. Therefore, the findings establish that the additional data concerning occupant activity provides more accurate assessments of energy usage that can be quite beneficial for enhancing the further development of better adaptive and more efficient building management systems. This work also helps to fill the existing gap in energy prediction literature wherein, unlike other fields, the human factor is considered in machine learning models that can lead to more accurate and less distortion-prone energy forecasting.

1. Introduction

According to reports, about 30 to 40 percent of the world's energy consumption is within the building sector, while more than 80 percent of the used energy in a building happens in a building's entire lifespan, that is, the operational phase, which is inclusive of heating, cooling, ventilation, and lighting [1]. It, therefore, becomes very important to have predictions regarding a building's use of energy to make energy-efficient decisions. Use of data analytics: Lately, data analytics has made huge strides in development, such that new and advanced machine learning models are in place to be developed and deployed in the field of energy prediction [2]. As much as data analytics is improved, machine learning models have evolved to handle big data sets that capture several factors that could affect energy use—for example, environmental conditions, weather patterns, and equipment performances [3].

An artificial intelligence system employs machine learning, which is a guideline or procedure that enables the system to locate beneficial patterns and solutions within a predetermined amount of data, or it can be employed to predict output values depending on a certain amount of input values [4]. Machine learning, in turn, needs algorithms to learn [5]. One needs a set of data and then investigate the relationship between them, define patterns

and use algorithms, which allow one to take a sum of input data and, according to definite patterns, produce definite outputs [5], [6].

Sensors are fundamental components of the IoT, as they allow information to be gathered from several physical spaces [6]. Keeping data consistent in time is important for accuracy when building machine learning models. Therefore, if data responses are asynchronous, then the simulations are inaccurate, predictions wrong, and decisions unhelpful to the model, hence making a model ineffective. Such high accuracy is essential in real-time applications and processes, including monitoring and controlling industrial processes, managing resources, providing smart solutions in the environment, and ensuring the security and efficiency of crucial infrastructure.

Nevertheless, for high-trust and authentic Machine Learning models to be developed, synchronized data is required to be collected properly and accurately from different sensors. Using traditional methods of data acquisition, the issue of time synchronization in the distributed nodes of the sensor complicates the problem and creates a time asynchrony in the data. This has developed first as a distinct mode for increasing models of energy forecasting, such as the use of synchronized data collection systems. The use of a synchronized data collection system provides the mixed model with an opportunity to gather a broader spectrum of the drivers to energy use comprising ambient data, including temperature, humidity, and airflow, as well as data concerning occupants in terms of presence, activity, and behaviour. Such synchronizing of data makes it easy to correlate occupant behaviour with energy use in all its variables to arrive at better and more responsive predictions. If incorporated into the ML models, these behavioural factors could give scaly accuracy in predicting consumption and energy efficiency that building operators and energy managers can use to give users meaningful insights for cutting down consumption and enhancing efficiency [7].

All the developed models following this methodology of prediction do not develop models that consider occupants' behaviour [8]. Occupants, through their very many and varying actions and preferences, critically contribute to the real necessity of having a building. Not addressed to the best possible degree, the behaviour of occupants may lead to erroneous predictions and improper energy management.

In addition to the problems of highly synchronized data acquisition, there is the question of how to quantify the occupants' behaviour, which is random and, therefore, difficult to predict. Individuals' actions, decisions and use of spaces are unique to a particular individual or a group of people. As a result, there is a need to obtain the right and sufficient information on how the building occupants influence and manage the building systems such as the heating, ventilation and air conditioning (HVAC) and lighting systems. Modern artificial intelligence systems are now being used to address this issue and gradually integrate variability into the learning models. When used in conjunction with real-time data collection systems that are in phase with the control systems, the control systems become more dynamic, and energy use can be regulated according to the environment and the people occupying the building. These improved models, which assist in the management of the interconnectivity of the building systems and the human activities, also enable the efficient use of resources and correct use of energy.

In this research, data is obtained from the different sensors implanted in the building, as well as the questionnaires completed by the occupants, which are used in training and testing different machine learning systems. The models used are Support Vector Regression (SVR), k-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGB), Random Forest (RF), and LIGHTGBM Regressor. As a conclusion of the models' performance, R^2, MAE, and MAPE indices are used, which reflect the accuracy and the degree of proximity of the models. This multiple-criteria approach provides a better way of analyzing the performance of each of the models and their abilities to make predictions.

2. Literature Review

The increasing concern for energy conservation across the world has led to increased research on energy use in buildings with special attention to modelling. Most of the research work in this field was done in an earlier period of time and focused mainly on the conventional engineering methods of estimating the energy demand with the help of statistical models, which are mainly historical. Nevertheless, in the present time, the focus of the research has been changed towards dynamic machine learning models that can handle real-time data getting collected through the IoT sensors. Studies reveal that occupant behaviour is an important factor in energy use but is often neglected by most prediction models, thus giving imprecise estimates. Moreover, the increasing application of synchronized data collection systems has appeared as the main solution to improving these models, as it can provide better chances to obtain environmental and occupant-related factors. This review focuses on the development of machine learning methods applied to building energy prediction and on the importance of time-synchronized data acquisition and the integration of occupant behaviour into these models.

Data collection is an essential factor that is used in simulating and conducting research on occupant behaviour. Scholars have come up with several models based on data and data gathering techniques that are different from one another. Such methods include direct collection of occupancy data using sensors such as motion [9], [10], [11], cameras [12][13], and questionnaires such as the Time-Use Survey [14]. There are no real-time clocks accessible in affordable sensors. As a result, it is not possible to synchronize the continuous streams of data directly using

them. Shih et al. [15] provided a special data alignment method for the IoT and CPS in order to address this. By putting in place a Network Gateway of Multiple Protocols' Transport Network, this objective was accomplished. In their design, Hadjidemetriou et al.

L. Hadjidemetriou et al. [16] suggested a design for smart buildings that imitates comfort, air quality, as well as electricity consumption of occupants in their paper.

Another challenge tackled by Craciunas et al. [17] is that, due to the loss of synchronization frequently incurred, there is a temporal misalignment which disturbs the determinism of transmission in Time-Sensitive Networks (TSN). Sakaguchi et al. [18] proposed a detailed hardware implementation plan for the IEEE 802.1AS-2020 precision time protocol to enhance the time synchronization accuracy of Industrial IoT devices with very high time synchronization accuracy. The proposed hardware approach is observed to have a time synchronization accuracy of the order of 50 ns, while the variation should not be more than 20 ns; this standard is admissible for industrial use according to IEC 61850-9-5.

Farhadi *et al.* [3] described how, in general, machine learning methods contribute to the accelerated discovery and optimization of advanced energy materials, such as perovskite, for solar cells, batteries, or phase change materials in the study. The problems addressed are material property prediction, device structure design and enhancement, which is related to this work in the use of machine learning for accurate predictions in big data sets in energy applications. This goes on to depict the role of data in propelling innovations within energy systems as they transition to more complex systems.

Nabavi et al. [19] compared the energy use in Iran's residential and commercial buildings sector through macro-economic and technological indicators such as population, GDP, natural gas and electricity prices, and share of renewable energy. These include the following: multi-variable linear regression (MLR), logarithmic multi-variable linear regression (LMLR) and the nonlinear autoregressive with exogenous input artificial neural networks (NARX) to predict energy consumption to 2040. The NARX model proved to be the most accurate in predicting Iranian energy consumption at 76.97 Mtoe by 2040.

The research work of Bouyakhsaine et al. [20] applied machine learning models for the prediction of residential building occupancy using sensor and survey data, including occupant activities, energy, water consumption, and environmental factors. Random Forest turned out to be the most accurate technique, with 95.96% accuracy in binary prediction, while Bayesian Networks resulted in an accuracy of 73% in active and sleep zone differentiation. Other models, such as XGBoost and Regularized Greedy Forest, gave accuracy ranging from 70% to 74%.

Vafeiadis et al. [21] aimed to explore occupancy detection in residential homes using machine learning algorithms on energy and water consumption from smart meters. The input features include power consumption from the various household appliances, water usage and occupancy, which is measured through an infrared door counter. The accuracy of each model was as follows: Random Forest had the highest accuracy of 80.53% and an F-measure of 83.37%, which is better than Decision Trees and Support Vector Machines.

In the paper, Li et al. [9] presented a new approach for occupancy prediction based on the inhomogeneous Markov chain model with comparison to Probability Sampling, Artificial Neural Networks (ANN), and Support Vector Regression (SVR). Input features include occupancy presence data from sensors in residential buildings, either in five-minute or ten-minute bins. The proposed Markov model is used for short term forecasts, and the results show an improvement of 5% in accuracy compared to the other models for the forecast time of 15 minutes ahead, while the ANN and SVR models are more accurate for the long-term forecast of 24 hours ahead.

This work complements the growing literature on the prediction-based model in energy with a comparative analysis that accounts for the human factor, thereby offering a clear path that can lead to robust and reliable prediction. The dataset used in this study was collected independently as part of our research, ensuring originality and relevance to the study objectives. No existing research utilizes this particular dataset so similar analysis cannot be directly evaluated.

3. Methodology

Data collection, data preprocessing, modelling, cross-validation, and model performance will be discussed in this section. The methodology in this study shows an innovative approach to linking occupant behaviors with energy use patterns but delivers significant results for the research domain. The flowchart in Fig. 1 below demonstrates the overall framework of this study. The flowchart outlines the operational cycle of an IoT-based embedded system used for the formulation and validation of a predictive model. Beginning with the stage of IoT-based Embedded System Design, construction and implementation are subsequent procedures in the process. Data is then collected from three main sources: environmental data, occupant data, and power consumption data. The subsequent step as soon as the data is collected is Data Analysis, and the sub-steps included are Data Preprocessing, Model Selection, Model Building and Model Validation. What is more, the structure described here exhibits logical progression from system construction to data acquisition and from modelling to validation.

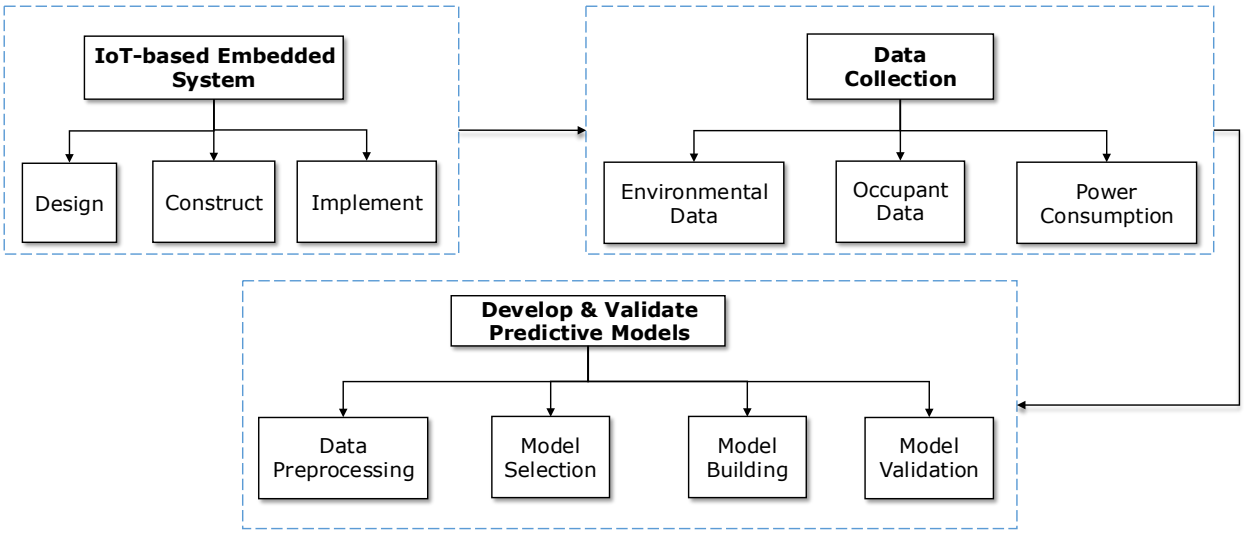

Fig. 1 *Framework of this study*

3.1 Data Collection

Fig. 2 outlines the flowchart of a clear method on how to regulate the flow of data in the IoT sensor networks towards enhancing consumption modelling of occupants' activities with regard to power consumption. The first stage of this methodology is to configure the GPIO pins and connect them to the MQTT broker to guarantee good communication of the sensors. These sensors are set up to record environmental data within a specific area or facility on demand over time. After the initialization, the system subscribes to topics that can be associated with environmental information or power consumption. Upon receiving the sequence number for each set of sensors, the system analyzes the data and makes a decision on what should follow next. If the system does not detect any sequence number, it loops back to the start to ensure continuous data collection without interruption. In the event a sequence number is received, a callback function is triggered, which processes the data by assigning it to the appropriate node number along with the sensor's data and the sequence number. The processed data is then republished by the MQTT broker, thus letting off the current cycle of data transfer. It also starts a new cycle by sending a new sequence number to the sensors at the beginning of transmission. This makes it possible for the system to flow in parallel with the correct arrangements, keeping the environment under check in a continuous manner.

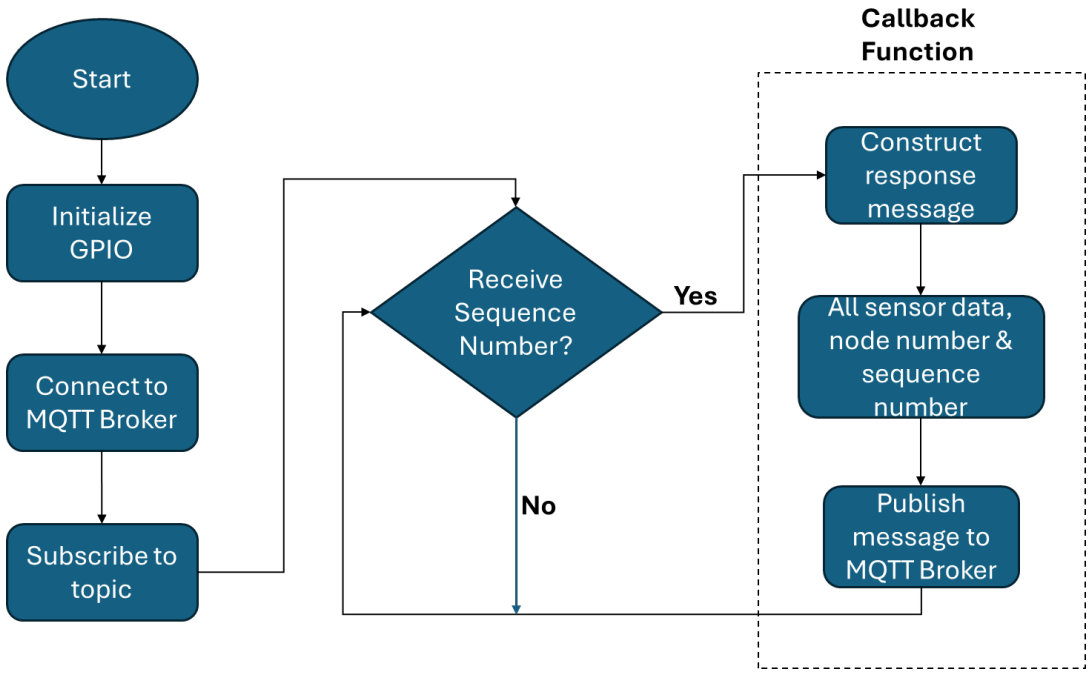

Fig. 2 *Synchronization system flowchart*

Int. Journal of Integrated Engineering Vol. 17 No. 2 (2025) p. 14-29

This systematic approach makes it possible to achieve real-time integration of an IoT-based data acquisition system with a machine learning-based data analysis system in order to monitor and optimize the power consumption in relation to the activities of the occupants. The structure proposed in this paper, represented in Fig. 3, is to fulfil the synchronization of data and to minimize the time synchronization as much as possible. The MQTT server can, at the same time, send the same sequence number to different sensor nodes. Only when the sensor nodes receive the sequence number, these nodes forward their sensory data to the fixed MQTT topic. It is a mechanism that ensures that all data that is received is accurate. Data from the MQTT server is directly written into a CSV file for more data processing as needed.

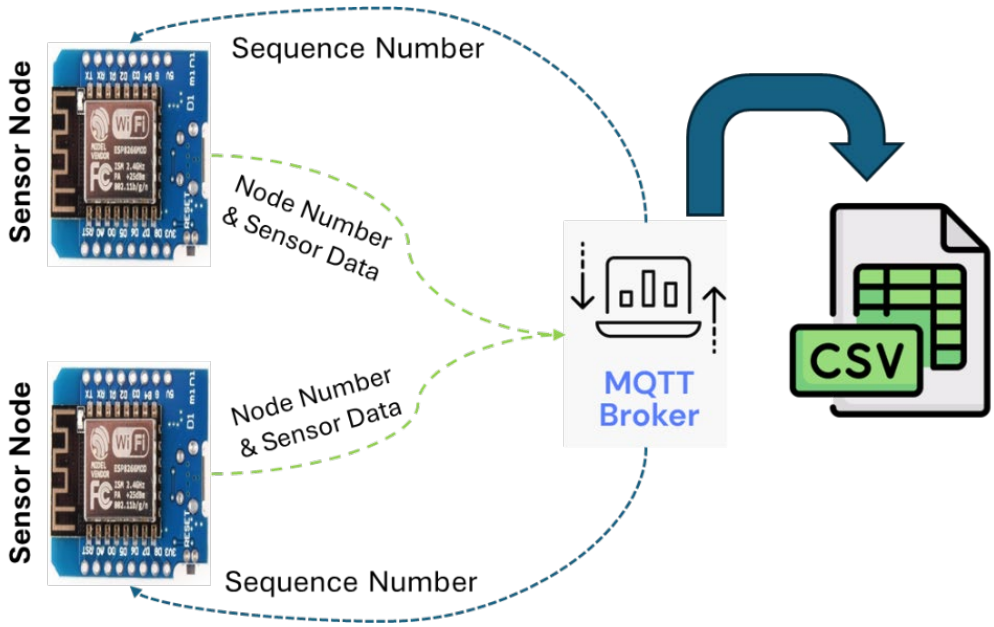

Fig. 3 *Synchronization of the sensor nodes through the MQTT broker*

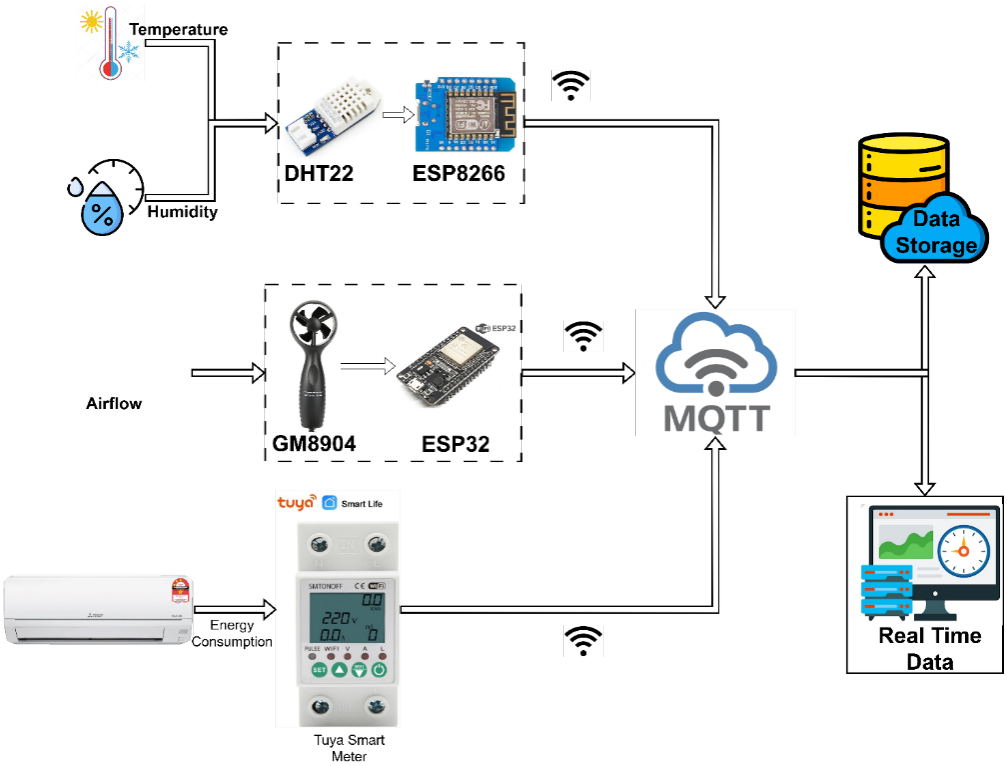

Fig. 4 *Overall experimental setup*

The schematic exhibit illustrated in Fig. 4 depicts the experimental setup proposing the gathering of temperature, humidity, airflow and power consumption data from the Digital Home Lab, which is situated in the Faculty of Engineering, a five-story building at Multimedia University, Cyberjaya, Malaysia. The experimental location was divided into six equal-sized grids where all the hardware setups were implemented. For real-time data synchronization, the synchronization algorithm was uploaded onto the six ESP8266 WeMos D1 Mini and one ESP32S microcontroller through the Arduino IDE software. The lab's temperature and humidity were measured using six DHT22 sensors connected to the WeMos D1 Mini boards. Another sensor, GM8904, was used in the UART connection to the ESP32S for wind speed measurement. To create a data pipeline, the Wi-Fi-connected boards that were interlinked to a private/secure MQTT broker through the publish/subscribe model, which facilitated the transmission of sequential data, were used. The Tuya-powered smart meter, illustrated in Fig. 4, will collect energy consumption data from air conditioners (AC). This valuable data will be seamlessly transmitted to the MQTT cloud, enabling real-time monitoring and storage of data for insightful analysis.

3.1.1 Occupant

To gather data from the occupants, questionnaires were administered. Tables 1 and 2 below present selected sample inquiries about their thermal experiences during the four assigned activities based on ASHRAE (2013). The air conditioning settings were adjusted according to their activities and thermal comfort preferences to enhance their comfort during the tasks, while power consumption was simultaneously recorded.

Table 1 *Occupant questionnaire on their thermal sensation*

Question	Scale	Thermal Sensation	Vote
How would you describe thermal comfort in this environment?	0	Cool	
	1	Slightly Cool	
	2	Neutral	
	3	Slightly Warm	
	4	Warm	

Table 2 *Occupant questionnaire on activity*

Question	Activity	Activity Level
What is your Activity Level right now?	Seat and relax	1
	Seat and working	2
	Stand and relax	3
	Light Activity while Standing	4

The sample consisted of 268 data points, including environmental, energy, and occupant measurements collected from three occupants. Fig. 5 is a pie chart presenting the distribution of the "Feeling Scale," which reflects the thermal comfort that occupants of the building felt during their activities. All the occupants were academician students. The largest portion of responses belongs to "Neutral" at 54.3%, meaning that more than half of the occupants did not feel either too warm or too cold. This is followed by 16.7% of occupants who felt "Slightly warm" and 14.1% who described the temperature as "Slightly cool." A total of 11.5% of occupants identified the building's temperature as "Warm," while fewer than 3.3% experienced "Cool" sensations.

Int. Journal of Integrated Engineering Vol. 17 No. 2 (2025) p. 14-29

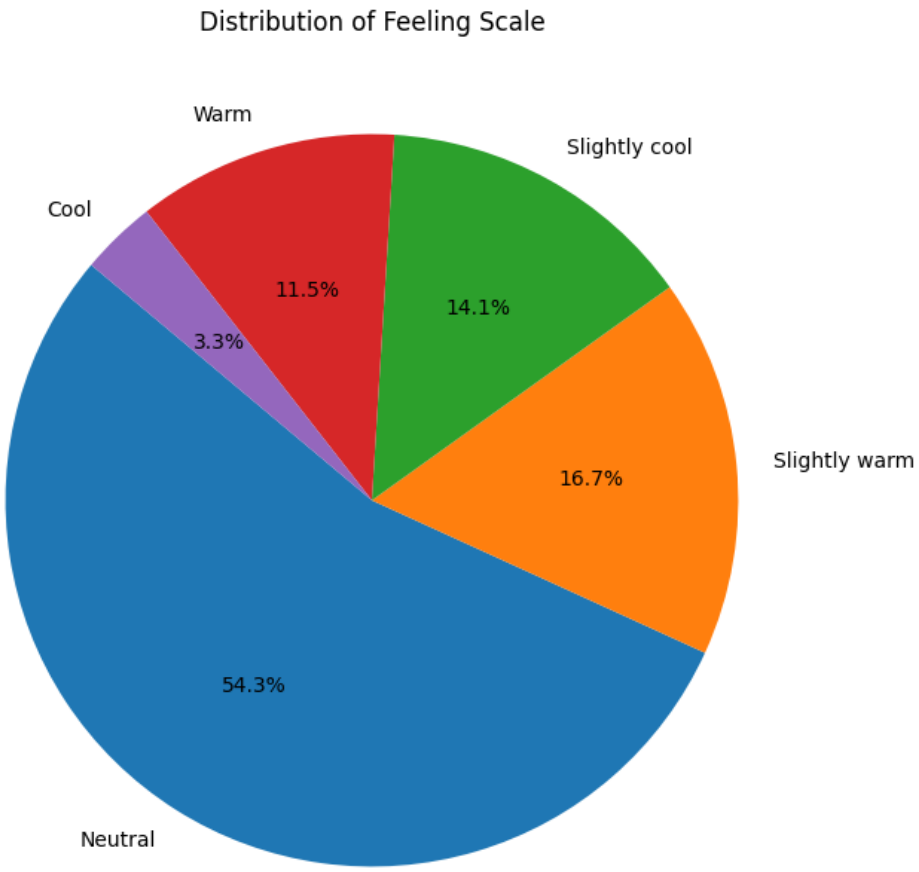

Fig. 5 *Data distribution of occupants' feeling scales*

3.2 Data Preparation

Once data collection is synchronized, the methodology further incorporates machine learning algorithms to analyze the relationship between occupant activity and power consumption. The gathered data is fed into machine learning models that identify patterns and trends, enabling the system to predict energy usage based on specific occupant behaviour. The feedback from these models helps optimize energy management strategies, contributing to more efficient power consumption in smart buildings.

Fig. 6 illustrates the Machine Learning pipeline, demonstrating the essential stages from data collection to model evaluation and final model selection. The pipeline begins with the collected dataset, which serves as the primary input for the machine learning process. These datasets often come from various sources and may contain raw or unstructured data.

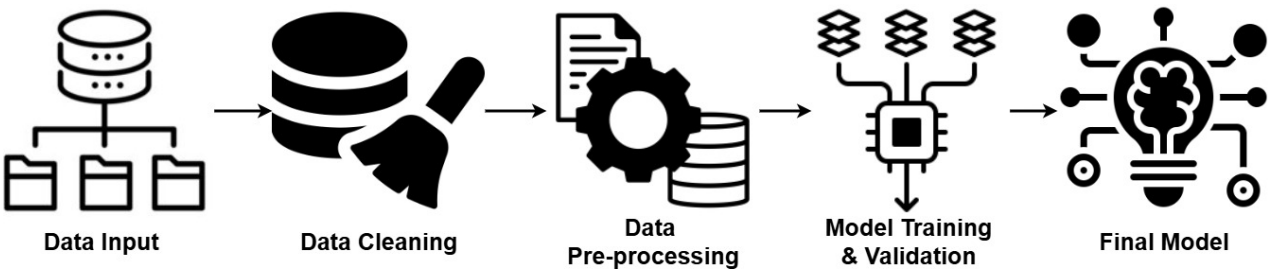

Fig. 6 *Machine learning pipeline*

3.2.1 Data Input

The data input for the model is environmental measurement and sensor data, which include Air flow from the Air flow sensor (m/s), Temperatures from the following sensors Sensor1_Temp (C) to Sensor6_Temp (C) and the humidity of those sensors are Sensor1_Humidity (%) to Sensor6_Humidity. Other input features are the Feeling Scale, Activity Level, and State of Heater Grid, which also includes feedback from human participants and organizational factors. The dependent variable (output) is the air conditioning power consumption expressed in terms of AC Power (kW). This vast set of data will allow the model to predict the consumption of energy depending on the environmental and behavioural factors.

3.2.2 Data Cleaning

The first critical stage in the pipeline is Data Cleaning. This phase involves refining the dataset to ensure that it is accurate, complete, and suitable for analysis [22]. Data cleaning techniques include removing duplicate records, handling missing or incomplete data, correcting data type mismatches, and addressing inconsistencies. This stage is crucial as it lays the foundation for a reliable dataset that can be effectively utilized in later stages.

3.2.3 Data Pre-processing

Preprocessing differs from data cleaning although it is also a process of data transformation, it is of a higher level and aims at preparing the data for modelling. Some of the techniques that may be employed during this step include scaling of features, converting categorical variable couple to numerical form, normalization of the data distribution, handling of outliers and feature construction from the existing features [23]. These steps operate to qualify and quantify the data to the greatest extent possible and to align them to the needs of the machine learning algorithms to be employed.

3.2.4 Model Selection

The subsequent step to data preprocessing is termed as model selection. In this experiment, five different types of regression models were used. Model selection assists in comparing the number of algorithms to determine which of them is suitable for the tasks that are intended and with regard to the features of the data set and the problem under consideration. These models could be simple regression models, such as linear regression, or more complex models, such as decision trees, random forests or support vector machines, depending on the nature of the problem or the character of the data to be used.

3.2.5 Model Training & Validation

After the models are chosen, the next process is Training and Validation. In this stage, the data set is divided between the training data set and the cross-validation data set, where the model is trained on the data set and tested on the unknown data set. Here, cross-validation methods are used to deal with the problem of memorization rather than generalization of the model to unseen data. Hyperparameters may also be tuned at this stage so as to fine tune the model for better performance. After training and validation, the models thereafter go through evaluation. This includes evaluating each of the models on the size of errors such as mean square error, the coefficient of determination (R^2), or any other evaluation indicators [24]. Thus, by comparing the results of the different regression models, one can understand the strengths and drawbacks of each when it comes to the specific problem and dataset at hand and which of the models should be used.

3.2.6 Final Model

In the end, the model with the highest performance will be selected as the Final Model. This final will be used for making prediction on new data summarizing insights or for solving other problem based on the learning that has taken place in this pipeline. By doing so, the data is then moved from a state that is often referred to as 'raw' or 'unstructured,' and becomes meaningful insights empowered by a machine learning model that has been vetted and fine-tuned.

3.3 Machine Learning Models

Some of the approaches of machine learning for the air conditioning power consumption prediction using the recorded sensor, environmental data, and AC power consumption are described in this section. In order to compare the ability of each introduced model to model the relationships between the inputs and the output all of them are trained and validated.

Int. Journal of Integrated Engineering Vol. 17 No. 2 (2025) p. 14-29

3.3.1 Random Forest Regressor (RF)

An ensemble learning model is the Random Forest Regressor, which builds many decision trees during training and then averages their predictions to enhance accuracy and reduce overfitting [25]. It works well on both linear and nonlinear data; hence, it will always be a versatile model for regression tasks. It tends to provide robust predictions with lower variance compared to individual decision trees by leveraging the "wisdom of crowds" coming from many trees [26]. In contrast, random forest is immune to outliers and missing data; it can handle them on its own and becomes one of the most trustworthy and used regression models on complex datasets [27].

3.3.2 XGBoost Regressor

It is mainly a gradient boosting algorithm in which, at a time, ensembles of decision trees are built and each tree corrects previous errors [28]. Known for its speed and high predictive accuracy, XGBoost uses advanced regularization techniques to prevent overfitting. Hence, it becomes equally efficient with large and complex datasets. The model is very famous for flexibility in bias-variance control, handling missing values, and computation optimization. For these reasons, this model has become one of the most preferred models used in competitive machine learning tasks and regression problems [29].

3.3.3 Support Vector Regressor (SVR)

The Support Vector Regressor is a machine learning model based on support vectors in regard to forecasting continuous outcomes by finding the best line within a margin of tolerance [30]. SVR would fit best with data with a non-linear relationship as it can use different functions to map higher dimensions, hence allowing linear separation functions. On one hand, even though SVR has the capability of providing accurate predictions, it is sensitive to outliers, and sometimes its performance can be affected by the choice of hyperparameters, making it less reliable when compared to other models such as Random Forest or XGBoost, especially in scenarios with extreme values or large datasets [31].

3.3.4 K-Nearest Neighbors (KNN)

Another simple, non-parametric model is the K-Nearest Neighbors regressor, which does its prediction by averaging the values of the K closest data points—referred to as neighbors—in the feature space [32]. It is easy to understand and easy to implement, hence useful in the case of smaller datasets in which the relationships among points may be local. However, KNN can be sensitive to the choice of K, and its performance may degrade with high-dimensional data or large datasets due to increased computational cost [33]. That is, if the data are very well trended and less noisy, KNN can give reliable predictions [34].

3.3.5 Light Gradient Boosting Machine (LIGHTGBM)

The LIGHTGBM regressor is an XGBoost-like, highly efficient and scalable implementation of gradient boosting designed for speed and performance [35]. The approach for the construction of a decision tree in this is sequential, making every attempt at accuracy without turning to overfitting, with advanced techniques like leaf-wise splitting and handling categorical features natively. LIGHTGBM is particularly suited for large datasets and realizes fast training time compared to other boosting methods, like XGBoost, with high predictive accuracy [36]. It can handle complex data and hyperparameter tuning very effectively; thus, LIGHTGBM has become very popular for regression tasks in machine learning competitions and in practical applications [37].

3.4 Evaluation Parameters

The full dataset was randomly divided into training and test sets. The majority, 80%, was allocated for training and validation, while the remaining 20% was set aside for testing. During the training phase, the models were evaluated using cross-validation, where 20% of the training data was withheld for validation purposes. To measure the prediction error rates and evaluate the model's performance in Regression Analysis, several statistical indicators, including R^2, Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were utilized.

3.4.1 Coefficient of Determination (R^2 or R-squared)

R-squared is the coefficient of determination that indicates how well those values fit in relation to the original chosen values [38]. It varies between 0 and 1 and it is expressed in percentage. The higher the score of the index it is, the better the model.

3.4.2 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is basically the difference between the original and the predicted value extracted by taking the average of the absolute difference over the data set [39].

3.4.3 Mean Absolute Percentage Error (MAPE)

One of these metrics is the Mean Absolute Percentage Error which is, on average, the percentage of difference between the forecasted and actual values [40]. This is reckoned by just taking the arithmetic mean of the absolute percentage errors across the dataset, which is logical as it quantifies prediction error in terms of percentage.

3.5 Equations

In this study, the performance of the machine learning models for predicting air conditioning power consumption is evaluated using three statistical metrics: Such measures are R^2, MAE, and MAPE. These are given by equations (1)– (3) and are normally used to ascertain the efficiency of the models by comparing the computed values with the observed ones.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3}$$

where y_i is the actual observed value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the actual observed values, and n is the total number of data points.

4. Results and Discussion

Synchronizing the datasets is very important, especially with regard to enhancing the level of accuracy in energy control and the efficiency of the predictions. It creates a dataset with equal time steps every hour, which is very useful for researchers who require accurate data to calibrate and predict the demand as well as consumption of energy. This type of parallelism makes the data properly standardized for general use and gives a stable ground on which the subsequent enhancements of the existing, more complex prediction algorithms can be made. These algorithms are rather useful for system supervising with a view to 'saving' energy in smart structures and complexes, which consume a large amount of energy in order to work.

Int. Journal of Integrated Engineering Vol. 17 No. 2 (2025) p. 14-29

Table 3 *Sensor, occupant and energy data from MQTT stored in CSV file*

Date Time	Sequence Number	Airflow (m/s)	Sensor Temperature (°C)	Sensor Humidity (%)	Occupant Activity	Occupant Feeling	Power (kW)
11:45:46	0	3.22	28.60	73.80	Seat and relax	Normal	0.0857
11:46:46	1	3.22	28.60	73.70	Seat and relax	Slightly Cool	0.0854
11:47:46	2	3.32	28.60	73.90	Seat and working	Normal	1.004
11:48:46	3	3.12	28.50	73.90	Seat and working	Normal	1.004
11:49:46	4	3.32	28.50	74.20	Seat and working	Slightly Warm	1.083
11:50:46	5	3.12	28.50	74.30	Seat and working	Slightly Warm	1.085
11:51:46	6	3.22	28.40	74.30	Seat and working	Slightly Warm	1.083
11:52:46	7	3.12	28.50	74.30	Stand and relax	Normal	1.101
11:53:46	8	3.32	28.50	74.00	Stand and relax	Normal	1.114
11:54:46	9	3.32	28.50	73.90	Stand and relax	Slightly Warm	1.126
11:55:46	10	3.12	28.50	74.00	Stand and relax	Slightly Warm	1.107
11:56:46	11	3.12	28.50	74.10	Stand and relax	Slightly Warm	1.109
11:57:46	12	3.12	28.50	74.20	Stand and working	Slightly Warm	1.249
11:58:46	13	3.22	28.50	74.30	Stand and working	Slightly Warm	1.312
11:59:46	14	3.12	28.60	74.10	Stand and working	Warm	1.339
11:45:46	15	3.22	28.60	73.80	Stand and working	Warm	1.346

This correlation matrix in the Fig. 7 present various Pearson correlation coefficients that ranges from -1 to 1. Positive Correlation is depicted by a value of 1 in that both variables go up simultaneously while -1 depicts a perfect negative correlation in that when one variable goes up the other goes down. When the value is 0, it means that there is no linear relationship between the two variables. The diagonal values are all of 1, which implies that each variable is perfectly correlated with itself.

The correlation matrix reveals that Activity Level strongly drives AC Power (kW), where the correlation is 0.76, indicating a high consumption of AC power with respect to activity level. Therefore, logically, activity levels highly influence the demand for cooling, probably due to heat generation from the movement. Other factors, such as temperature and humidity readings from various sensors, tend to relate to power consumption, although to a lesser degree. Overall, it is the Activity Level that turns out to be a major determinant of the power consumption on AC, thus indicating its role in the estimation or control of power usage.

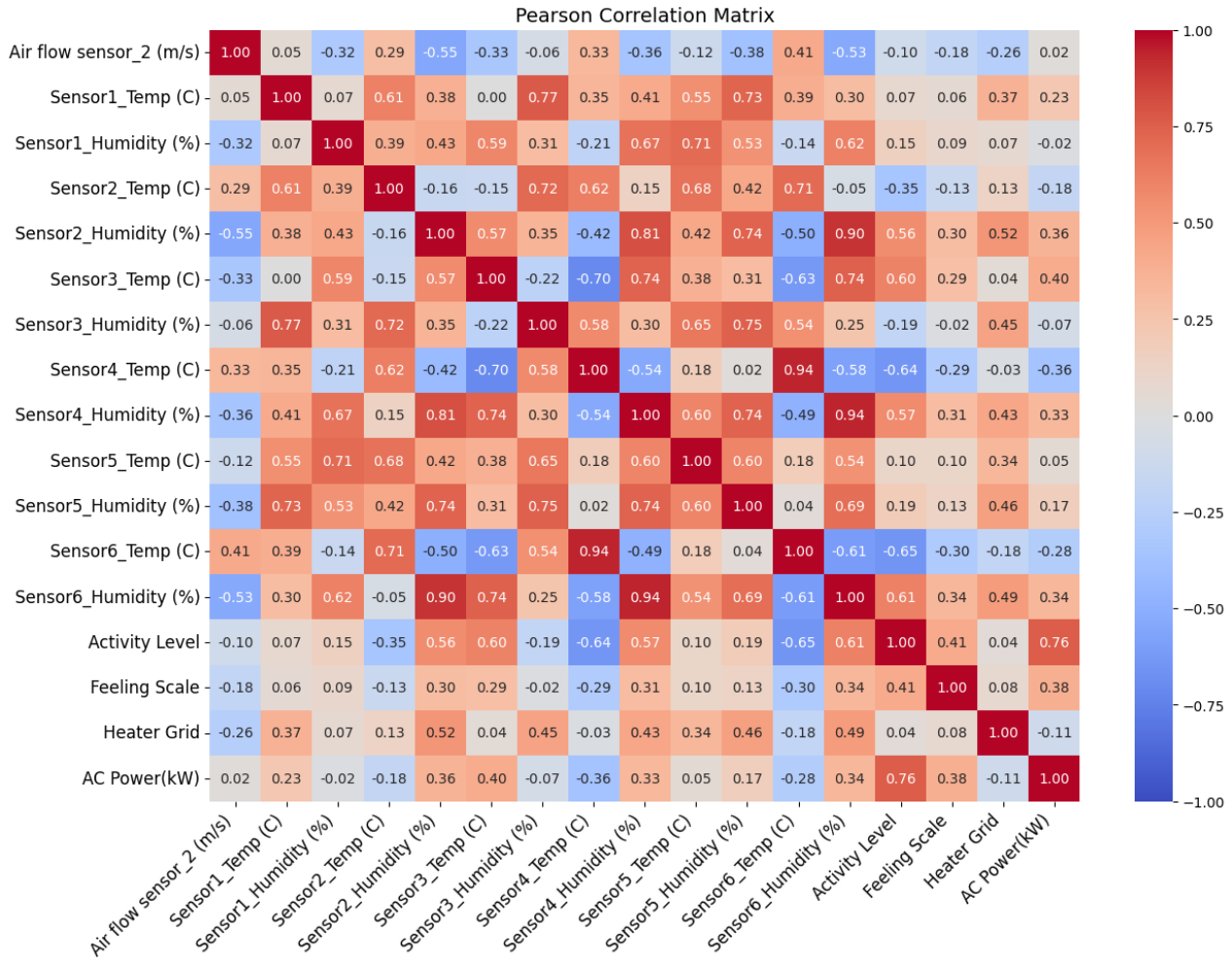

Fig. 7 *Pearson correlation matrix*

In this study, five machine learning models are employed. Among the models discussed and evaluated for their feasibility were SVR, KNN, XGBoost Regressor, RF, and LightGBM Regressor, as presented in Table 4. In the case of synchronized data sets, all the models are found to be efficient for estimation of energy demand and this paper shows, how these models are useful for formulating effective energy saving operating strategies in various fields.

Table 4 *Values of R^2, MAE, and MAPE for the five models*

Model	Set	R^2	MAE	MAPE
SVR	Train	0.95	0.09	52.08%
	Test	0.92	0.11	68.10%
KNN	Train	0.94	0.05	19.52%
	Test	0.94	0.05	15.63%
XGBoost	Train	1.00	0.00	0.76%
	Test	0.99	0.02	4.61%
RF	Train	1.00	0.01	1.86%
	Test	0.99	0.02	4.93%
LightGBM Regressor	Train	1.00	0.02	5.42%
	Test	0.98	0.04	10.31%

Int. Journal of Integrated Engineering Vol. 17 No. 2 (2025) p. 14-29

The models show high average accuracy, while RF Regressor provides an almost perfect line fit both on the training set and test one. Analysing the results, the model has the lowest values of absolute errors and percentage errors, which only slightly differ from zero, meaning negligible deviations from the ideal values and great generalization to the unseen data.

The smoothness of the fits achieved by SVR is good, but its stability is somewhat lower than that of other models. That could be said because the obtained R^2 values are quite good, at least for training, equal to 0.95, slightly lower for the test set (0.92). This may indicate that SVR is more sensitive to extreme values which will make it more variable and unpredicted at times.

KNN is relatively stable between the training and test datasets, with a coefficient of determination R^2 of 0.94. Thai KNN has low errors comparatively to SVR, and the percentage errors are held constant, thus showing good overall generalization and accuracy, thus making KNN a competent tool in this respect.

Of all the models shown, XGBoost has the closest to ideal observation with R^2 values of 1.00 on training and 0.99 on test data. The model has extremely low absolute and percentage errors, making it highly accurate and capable of exceptional predictive performance.

A similar excellent performance is observed with the LIGHTGBM Regressor model, with an R^2 of 1.00 on the training set and 0.98 on the test set. Its errors are marginally higher than those of the XGBoost, but at the same time it also has good prediction capability and considerably low overfitting which actually places LightGBM as the strong competitor to XGBoost in terms of performance.

Based on the results presented in this comparison, XGBoost outperforms the other algorithms and has the lowest errors and highest R^2, which suggests high accuracy and predictive capability. Random Forest and LightGBM Regressor are again not very far from perfect accuracy; thus, they are good contenders. KNN is always accurate and stable; thus, it is suitable to be used for models that require high levels of accuracy. In contrast, although the SVR has excellent performance as a predictive method, it is less reliable because of its extreme sensitivity to outliers and larger percentage errors. All in all, XGBoost and Random Forest are the best models, whereas LightGBM Regressor is also not very behind; KNN can be used as the next most efficient model. Nevertheless, SVR has higher error rates than other models.

This research used a dataset specifically developed for this study which makes direct comparison against prior work difficult to perform. The analysis methods and machine learning models used in this research can be compared despite the exclusive dataset. The present research selects machine learning models Random Forest, XGBoost and LightGBM for prediction because these algorithms have shown excellent performance in previous work. A thorough model evaluation was achieved through the combination of R^2 and MAE and MAPE performance metrics for complete accuracy and reliability assessment.

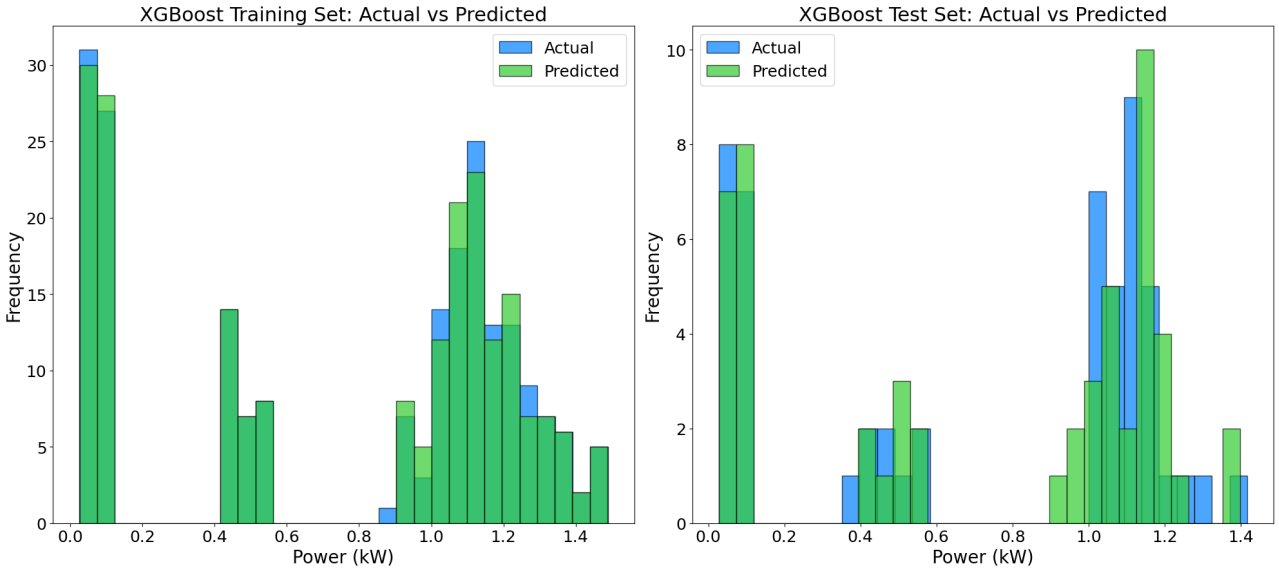

Fig. 8 *XGBoost model performance on training and test sets: actual vs. predicted power consumption*

The XGBoost model performance can be evaluated by comparing actual and predicted AC Power consumption values through the histograms displayed in Fig. 8. The green bars in the left plot of the training set demonstrate an almost perfect correspondence with the blue bars representing actual values which indicates an extraordinarily strong model performance. The right plot of the predictions on the test set does, in general, hold some resemblance to the actual values; however, observable deviations appear in certain frequency bins,

particularly at the higher and lower ends of the distribution. This indicates that, generally, the model generalizes well but may have minor discrepancies when used on unseen data, mostly at extreme values in the range.

5. Conclusion

Buildings are one of the main sources of the world's energy consumption, but integrating data about occupant behaviour may be the most effective way to customize machine learning-based energy prediction models in particular. Only when it is possible to train models which more fully represent the complexities of energy use, including the human interactions with their environment to which these are in turn coupled, is there any realistic opportunity for global energy-efficiency goals to be met. The paper proposes a machine-learning technique to support energy-use prediction in buildings but in a manner that is sensitive to occupant behaviour. The evidence from the research proposes having developed machine learning models in conjunction with subsequent data acquisition systems to carry out residential building power prediction. In the predictive model, combining environmental data, occupant behaviour, and power usage, XGBoost performed slightly better than Random Forest models because its error rate was low. More emphasis was given to acquiring same-time data, especially on occupant activity, which increased the power of prediction of the model. Although this study is on power consumption, this is indeed directly related to energy consumption since power is the representation of energy consumed per unit of time. Therefore, more accurate forecasts for power consumption mean better understanding and control of total energy consumption. Consequently, this work distinguishes itself in the literature of prognosticating energy use, as it includes occupant behaviour as an element and thereby makes it possible to give important insights into the design of adaptive building management systems that could manage the use of energy judiciously, based on real-time data.

Acknowledgement

The authors would like to express their appreciation for all the support received from Multimedia University, Malaysia, and TM R&D, and acknowledge The TM R&D Research Grant (MMUE/230043).

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

Author Contribution

The authors are responsible for the study conception, research design, data collection, data analysis, result interpretation and manuscript drafting.

References

[1] D. K. Abideen, A. Yunusa-Kaltungo, P. Manu, and C. Cheung, "A Systematic Review of the Extent to Which BIM Is Integrated into Operation and Maintenance," *Sustainability*, vol. 14, no. 14, 2022, doi: 10.3390/su14148692.

[2] M. K. M. Shapi, N. A. Ramli, and L. J. Awalin, "Energy consumption prediction by using machine learning for smart building: Case study in Malaysia," *Dev. Built Environ.*, vol. 5, p. 100037, 2021, doi: https://doi.org/10.1016/j.dibe.2020.100037.

[3] B. Farhadi *et al.*, "Machine learning for fast development of advanced energy materials," *Next Mater.*, vol. 1, no. 3, p. 100025, 2023, doi: https://doi.org/10.1016/j.nxmate.2023.100025.

[4] S. K. Singh, A. K. Tiwari, and H. K. Paliwal, "A state-of-the-art review on the utilization of machine learning in nanofluids, solar energy generation, and the prognosis of solar power," *Eng. Anal. Bound. Elem.*, vol. 155, pp. 62–86, 2023, doi: https://doi.org/10.1016/j.enganabound.2023.06.003.

[5] R. P. França, A. C. Borges Monteiro, R. Arthur, and Y. Iano, "Chapter 3 - An overview of deep learning in big data, image, and signal processing in the modern digital age," in *Trends in Deep Learning Methodologies*, V. Piuri, S. Raj, A. Genovese, and R. Srivastava, Eds., in Hybrid Computational Intelligence for Pattern Analysis. Academic Press, 2021, pp. 63–87. doi: https://doi.org/10.1016/B978-0-12-822226-3.00003-9.

[6] F. S. Nila, W.-H. Tan, C. P. Ooi, and Y.-F. Tan, "IoT-Based Embedded System for Streamlined Thermal Comfort Data Collection in Buildings," *Int. J. Integr. Eng.*, vol. 16, no. 3, pp. 78–91, 2024, [Online]. Available: https://penerbit.uthm.edu.my/ojs/index.php/ijie/article/view/15892

[7] M. Bourdeau, X.-Q. Zhai, E. Nefzaoui, X. Guo, and P. Chatellier, "Modeling and forecasting building energy consumption: A review of data-driven techniques," *Sustain. Cities Soc.*, vol. 48, 2019, doi: 10.1016/j.scs.2019.101533.

[8] K. Amasyali and N. El-Gohary, "Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings," *Renew. Sustain. Energy Rev.*, vol. 142, p. 110714, 2021, doi: https://doi.org/10.1016/j.rser.2021.110714.

[9] Z. Li and B. Dong, "A new modeling approach for short-term prediction of occupancy in residential buildings," *Build. Environ.*, vol. 121, pp. 277–290, 2017, doi: 10.1016/j.buildenv.2017.05.005.

[10] B. Yang, F. Haghighat, B. C. M. Fung, and K. Panchabikesan, "Season-Based Occupancy Prediction in Residential Buildings Using Machine Learning Models," *e-Prime - Adv. Electr. Eng. Electron. Energy*, vol. 1, no. August, p. 100003, 2021, doi: 10.1016/j.prime.2021.100003.

[11] S. Akbari and F. Haghighat, "Occupancy and occupant activity drivers of energy consumption in residential buildings," *Energy Build.*, vol. 250, p. 111303, 2021, doi: 10.1016/j.enbuild.2021.111303.

[12] P. W. Tien, S. Wei, J. K. Calautit, J. Darkwa, and C. Wood, "A vision-based deep learning approach for the detection and prediction of occupancy heat emissions for demand-driven control solutions," *Energy Build.*, vol. 226, p. 110386, 2020, doi: https://doi.org/10.1016/j.enbuild.2020.110386.

[13] I. Mutis, A. Ambekar, and V. Joshi, "Real-time space occupancy sensing and human motion analysis using deep learning for indoor air quality control," *Autom. Constr.*, vol. 116, p. 103237, 2020, doi: https://doi.org/10.1016/j.autcon.2020.103237.

[14] S. Hu, D. Yan, J. An, S. Guo, and M. Qian, "Investigation and analysis of Chinese residential building occupancy with large-scale questionnaire surveys," *Energy Build.*, vol. 193, pp. 289–304, 2019, doi: 10.1016/j.enbuild.2019.04.007.

[15] C. S. Shih, C. M. Yang, and Y. C. Cheng, "Data Alignment for Multiple Temporal Data Streams without Synchronized Clocks on IoT Fusion Gateway," *Proc. - 2015 IEEE Int. Conf. Data Sci. Data Intensive Syst. 8th IEEE Int. Conf. Cyber, Phys. Soc. Comput. 11th IEEE Int. Conf. Green Comput. Commun. 8th IEEE Inte*, pp. 667–674, 2015, doi: 10.1109/DSDIS.2015.117.

[16] L. Hadjidemetriou *et al.*, "A digital twin architecture for real-time and offline high granularity analysis in smart buildings," *Sustain. Cities Soc.*, vol. 98, no. January, p. 104795, 2023, doi: 10.1016/j.scs.2023.104795.

[17] S. S. Craciunas and R. S. Oliver, "Out-of-sync Schedule Robustness for Time-sensitive Networks," *IEEE Int. Work. Fact. Commun. Syst. - Proceedings, WFCS*, vol. 2021-June, pp. 75–82, 2021, doi: 10.1109/WFCS46889.2021.9483602.

[18] N. Sakaguchi, R. Kawate, and Y. Nagai, "IEEE 802.1AS Precision Time Protocol Full Hardware Prototyping for Industrial IoT," *Proc. - IEEE Consum. Commun. Netw. Conf. CCNC*, vol. 2023-Janua, pp. 943–944, 2023, doi: 10.1109/CCNC51644.2023.10059981.

[19] S. A. Nabavi, A. Aslani, M. A. Zaidan, M. Zandi, S. Mohammadi, and N. Hossein Motlagh, "Machine Learning Modeling for Energy Consumption of Residential and Commercial Sectors," *Energies*, vol. 13, no. 19, 2020, doi: 10.3390/en13195171.

[20] K. Bouyakhsaine, A. Brakez, and M. Draou, "Prediction of residential building occupancy using Machine learning with integrated sensor and survey Data: Insights from a living lab in Morocco," *Energy Build.*, vol. 319, no. March, p. 114519, 2024, doi: 10.1016/j.enbuild.2024.114519.

[21] T. Vafeiadis *et al.*, "Machine Learning Based Occupancy Detection via the Use of Smart Meters," *Proc. - 2017 Int. Symp. Comput. Sci. Intell. Control. ISCSIC 2017*, vol. 2018-Febru, no. June 2018, pp. 6–12, 2017, doi: 10.1109/ISCSIC.2017.15.

[22] A. F. Y. Mohammed, S. M. Sultan, J. Lee, and S. Lim, "Deep-Reinforcement-Learning-Based IoT Sensor Data Cleaning Framework for Enhanced Data Analytics," *Sensors*, vol. 23, no. 4, 2023, doi: 10.3390/s23041791.

[23] Y.-Y. Hong and R. A. Pula, "Methods of photovoltaic fault detection and classification: A review," *Energy Reports*, vol. 8, pp. 5898–5929, 2022, doi: https://doi.org/10.1016/j.egyr.2022.04.043.

[24] D. Chicco, M. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, p. e623, 2021, doi: 10.7717/peerj-cs.623.

[25] B. Grillone, S. Danov, A. Sumper, J. Cipriano, and G. Mor, "A review of deterministic and data-driven methods to quantify energy efficiency savings and to predict retrofitting scenarios in buildings," *Renew. Sustain. Energy Rev.*, vol. 131, p. 110027, 2020, doi: https://doi.org/10.1016/j.rser.2020.110027.

[26] X. Wang, W. Yuan, C.-J. Lin, L. Zhang, H. Zhang, and X. Feng, "Climate and Vegetation As Primary Drivers for Global Mercury Storage in Surface Soil," *Environ. Sci. Technol.*, vol. 53, no. 18, pp. 10665–10675, Sep. 2019, doi: 10.1021/acs.est.9b02386.

[27] F. Tang and H. Ishwaran, "Random forest missing data algorithms: TANG AND ISHWARAN," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 10, 2017, doi: 10.1002/sam.11348.

[28] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[29] Z. Ali, Z. Abduljabbar, H. Tahir, A. Sallow, and S. Almufti, "Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review," vol. 12, pp. 320–334, 2023, doi: 10.25007/ajnu.v12n2a1612.

[30] O. Montesinos-López, A. Montesinos, and J. Crossa, "Support Vector Machines and Support Vector Regression," 2022, pp. 337–378. doi: 10.1007/978-3-030-89010-0_9.

[31] M. Elhadad, "What is the difference between the three Machine Learning models?" 2023.

[32] S. Huang, M. Huang, and Y. Lyu, "A novel approach for sand liquefaction prediction via local mean-based pseudo nearest neighbor algorithm and its engineering application," *Adv. Eng. Informatics*, vol. 41, p. 100918, 2019, doi: https://doi.org/10.1016/j.aei.2019.04.008.

[33] S. Zhang, "Challenges in KNN Classification," *IEEE Trans. Knowl. Data Eng.*, vol. PP, p. 1, 2021, doi: 10.1109/TKDE.2021.3049250.

[34] H. Liu and S. Zhang, "Noisy data elimination using mutual k-nearest neighbor for classification mining," *J. Syst. Softw.*, vol. 85, no. 5, pp. 1067–1074, 2012, doi: https://doi.org/10.1016/j.jss.2011.12.019.

[35] M. R. Machado, S. Karray, and I. T. De Sousa, "LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry," *14th Int. Conf. Comput. Sci. Educ. ICCSE 2019*, no. Nips, pp. 1111–1116, 2019, doi: 10.1109/ICCSE.2019.8845529.

[36] J. Zhang, D. Mucs, U. Norinder, and F. Svensson, "LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity-Application to the Tox21 and Mutagenicity Data Sets," *J. Chem. Inf. Model.*, vol. 2019, 2019, doi: 10.1021/acs.jcim.9b00633.

[37] "Regression using LightGBM - GeeksforGeeks." [Online]. Available: https://www.geeksforgeeks.org/regression-using-lightgbm/

[38] G. Romeo, "Chapter 13 - Data analysis for business and economics," in *Elements of Numerical Mathematical Economics with Excel*, G. Romeo, Ed., Academic Press, 2020, pp. 695–761. doi: https://doi.org/10.1016/B978-0-12-817648-1.00013-X.

[39] P. Schneider and F. Xhafa, "Chapter 3 - Anomaly detection: Concepts and methods," in *Anomaly Detection and Complex Event Processing over IoT Data Streams*, P. Schneider and F. Xhafa, Eds., Academic Press, 2022, pp. 49–66. doi: https://doi.org/10.1016/B978-0-12-823818-9.00013-4.

[40] A. Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean Absolute Percentage Error for regression models," 2016.