 https://doi.org/10.24191/jeesr.v28i1.001

Best Subset Selection Multiple Linear Regression Model for Energy Consumption Prediction in Educational Buildings

Rijalul Fahmi Mustapa*, Atiqah Hamizah Mohd Nordin, Mohd Ezwan Mahadan and Muhammad Asraf Hairuddin

***Abstract*— Baseline Energy Models developed using Multiple Linear Regression (MLR) have been widely adopted due to their simplicity in representing energy consumption and its related independent variables. Moreover, MLR provides easily interpretable insights into the possible independent variables that govern energy consumption. Nonetheless, the risk of overfitting or underfitting may occur when unnecessary or correlated independent variables are included in the model. The objective of this paper is to enhance the selection of independent variables through the Best Subset Selection (BSS) method applied to baseline energy models developed using MLR. Two educational buildings were used as case studies. Baseline energy models were developed using both MLR and MLR enhanced with BSS, and the results of model development and prediction accuracy were compared. In case building 1, the MLR model enhanced with BSS showed improved prediction accuracy of Mean Square Error and, Root Mean Square Error of 255.0059kWh and 15.9689 kWh respectively. However, in case building 2, the enhanced model did not improve prediction accuracy. Nonetheless, in case building 2, the differences in accuracy between the enhanced and non-enhanced models were minimal, indicating that building owners may adopt either the standard or enhanced model with confidence when developing baseline energy models.**

***Index Terms*— Baseline Energy Model, Multiple Linear Regression, Best Subset, and Prediciton.**

I. INTRODUCTION

Energy consumption is a critical global concern, with the demand for electricity continuing to rise due to rapid industrialization, population growth, and digital transformation. According to the International Energy Agency (IEA) [1], global electricity consumption has increased by nearly 4.3% in 2024: with buildings accounting for over 30% of total energy use [2]. Educational buildings, in particular, contribute significantly to energy demand, as they operate for extended hours and accommodate various activities such as lectures, research, and administrative tasks. With rising energy costs and environmental concerns, optimizing energy consumption in these buildings has become an urgent necessity.

In order to achieve optimized energy consumption, the first essential step is to model it in relation to its influencing independent variables. This model, commonly referred to as the baseline energy model, serves as a fundamental tool for evaluating and tracking changes in these variables and their impact on energy use. Predicting energy consumption through this model is crucial, as it enables effective assessment and continuous monitoring of how fluctuations in the variables affect overall energy performance. The success of predicting energy consumption in buildings relies heavily on the development of a reliable baseline energy model. A baseline energy model is a mathematical representation that relates energy consumption to its influencing independent variables. These variables represent environmental and operational factors that significantly affect energy use within a building.

One commonly used approach for developing such models is multiple linear regression. This due to its simplicity and interpretability in establishing relationships between independent variables and energy consumption [3, 4]. Accurate energy consumption prediction is becoming increasingly important for effective energy planning and management. Before implementing any energy conservation measures (ECMs), building owners must identify the key variables that significantly impact energy usage. If these variables are correctly identified, the baseline model can predict future energy consumption with higher accuracy. This enables building owners to strategically plan their energy-saving activities, set realistic consumption targets, and evaluate the potential benefits of ECMs hence ultimately leading to more efficient and cost-effective energy use.

Current research on Baseline Energy Model (BEM) development [5-8] often selects independent variables arbitrarily, This approach risks including irrelevant variables, leading to suboptimal models that misrepresent energy consumption patterns. Careful variable selection is crucial for improving model accuracy, interpretability, and efficiency. Eliminating redundant variables reduces noise, enhances predictive reliability, and ensures the model captures key energy drivers. A well-optimized BEM enables accurate forecasting, supporting effective energy management in educational buildings. Neglecting proper variable selection leads to excess overfitting complexity, poor generalization or underfitting that includes incomplete representation, or inaccurate predictions [9]. Both scenarios compromise model performance, limiting its usefulness for energy efficiency planning and decision-making.

This manuscript is submitted on August 8, 2025, revised on December 1, 2025, accepted on December 2, 2025, and published on April 30, 2026.

All authors are with the Faculty of Electrical Engineering Universiti Teknologi MARA Johor Branch Pasir Gudang campus.

*Corresponding author
Email address: rijalulfahmi@uitm.edu.my

Thus, the primary objective of this paper is to identify the most significant independent variables influencing energy consumption in educational buildings by employing the Best Subset Selection method. This method ensures that only the most relevant predictors are retained for the development of an accurate BEM while simultaneously eliminating irrelevant or redundant variables that may lead to overfitting. The selected subset of variables is then used to construct a multiple linear regression-based BEM, which is compared against a model developed without the aid of Best Subset Selection (BSS).

This work is a continuation of previous studies conducted in [10] and [11]. In [10], the focus was on investigating the correlation of independent variables mainly occupancy using both single linear regression and multiple linear regression, without conducting any energy consumption prediction. In [11], energy consumption was predicted using a baseline energy model developed through multiple linear regression, where the independent variables were combined incrementally from the minimum to the maximum. The main distinction of the present work from [10] and [11] lies in its objective, which is to enhance variable selection using the best subset selection method. Furthermore, two educational buildings are examined as case studies to evaluate the effectiveness, reliability, and predictive performance of the proposed modeling approach.

II. Literature Review

Baseline energy models using Multiple Linear Regression (MLR) have been widely applied to predict energy consumption in buildings by analyzing various independent variables. These variables include weather conditions such as temperature, humidity, and solar radiation, as well as building-specific factors like lighting levels and internal temperatures. Additionally, occupancy patterns, working hours, and seasonal variations play a significant role in influencing energy usage [12]. MLR models have been implemented in diverse building types, including educational institutions, laboratories, residential apartments, and commercial facilities. For example, university buildings have been extensively studied due to their variable occupancy and operational schedules [13], while laboratory buildings have been analyzed for the impact of lighting and temperature control on energy demand [14]. A regression-based model [15] was developed to predict energy consumption in schools, one of Saudi Arabia's major energy-consuming facility types. Using 350 data points from schools in the eastern province, the model identified AC capacity and building age as the most significant factors. A study [16] applied regression analysis to predict long-term energy demand trends across residential, transport, and commercial sectors based on population and GDP as key factors. Both linear and nonlinear models were validated through statistical tests. The geographical distribution of case studies reveals that most research has been conducted in Asia, particularly in Malaysia and China, followed by European countries such as Norway and Spain, where energy efficiency is a key focus [17].

An interesting work has been published where MLR model were used for predicition in energy consumption for electric vehicle. Before modelling were made, the influence factor of energy consumption were clustered using clustering method where the key factors then were used for modelling purpose[18]. A recent study [19] applied a two-stage ensemble multiple linear regression (TSE-MLR) model to predict energy consumption in an educational building. The TSE-MLR demonstrated higher accuracy than other methods by effectively identifying key factors influencing energy use and minimizing prediction error, confirming its suitability for building energy modeling. The accuracy of MLR models varies depending on the quality of the dataset and the selection of independent variables. Some studies have demonstrated high accuracy, such as an MLR model applied to apartment complexes in China, which achieved an R^2 value of 95.77%, indicating strong predictive performance [13]. However, in other applications, such as energy forecasting for smart parks, the accuracy rate was slightly lower at 82.6% [20].

Maintaining the MLR model to remain relevant in predicting energy consumption may pose certain threats due to the increasing adoption of machine learning models in modeling and forecasting applications. Machine learning techniques have demonstrated advanced capabilities in capturing complex and nonlinear relationships between variables, often resulting in superior predictive performance compared to traditional methods [21, 22]. However, these models are frequently criticized for their limited interpretability, high computational requirements, and dependency on large volumes of data [23]. In contrast, MLR continues to be favored for its simplicity, transparency, and ease of use [24]. Its interpretability allows stakeholders such as building managers and policymakers to better understand the influence of key variables on energy consumption. To enhance its predictive reliability without compromising interpretability, this research proposes the use of best subset selection. This method improves model accuracy by identifying only the most relevant predictors while preserving the clarity and accessibility that make linear regression a practical tool for energy modelling [25].

III. Methodology

This section is organized into three main subsections to clearly present the data description, modeling method, and variable selection process. The first and second subsection provides a detailed description of the case study buildings number one and case study building number 2 respectively. The description includes the measurement of electrical energy consumption and the collection of independent variables. The third subsection is a mini summary of the case studies. The fourth subsection discusses multiple linear regression as the selected modeling approach for predicting energy consumption. Finally, the fifth subsection introduces the best subset selection method, which is used to identify the most significant independent variables to enhance model accuracy and reliability.

A. Case Building 1

The first case study involves a six-story academic building located within the College of Engineering, specifically designated for Electrical Engineering Studies. This building is primarily used for teaching and learning activities and contains a variety of spaces, including lecturers' offices, classrooms, and laboratories. The analysis focuses on lecture weeks when the

building is actively occupied from Sunday to Thursday between 8:00 a.m. and 5:00 p.m.

Electrical energy consumption in the building is supplied through a three-phase 0.415 kV line, which originates from the secondary voltage of an 11 kV to 0.415 kV transformer. This transformer is connected to the local utility grid. To monitor the energy usage, a data logger was installed at the main switchboard, capturing hourly energy consumption data from midnight until 11:00 p.m. The data collected during weekends and public holidays is excluded from this study because the energy consumption during these periods is significantly lower and may require a separate model due to differing operational patterns.

Several independent variables were identified that potentially influence the building's energy consumption. These include lecturer occupancy in office rooms, classrooms, and laboratories. Lecturer presence in office rooms was determined using biometric fingerprint attendance data, which records both arrival and departure times. It is assumed that equipment in office rooms such as lights, computers, and air-conditioning units remains in operation throughout office hours, regardless of short-term absence. Lecturer occupancy in classrooms and laboratories was obtained from the official teaching and laboratory schedules. Since these sessions may not occur simultaneously, the occupancy in classrooms and laboratories was treated separately to accurately capture variations in energy usage caused by different activities and equipment.

Student occupancy in both classrooms and laboratories was determined using the academic timetable in combination with enrollment records. The number of students registered for each session was used to estimate occupancy levels. Similar to lecturers, student presence was assumed to be zero during non-operational hours, specifically from midnight to 7:00 a.m. and from 6:00 p.m. to 11:00 p.m. Outdoor temperature was also included as an environmental variable due to its influence on air-conditioning demand. Hourly temperature data was obtained from www.weatherunderground.com, a weather platform that provides reliable satellite-based historical weather data. This variable is essential in modeling fluctuations in energy demand related to climate conditions.

B. Case Building 2

The second case study focuses on a four-story multipurpose academic building that houses 34 lecture rooms, six computer laboratories, and two examination halls. This building is actively used during lecture weeks and experiences full occupancy due to the high frequency of scheduled teaching and laboratory activities. The electrical supply for the building is provided through a three-phase 0.415 kV system, which is derived from the secondary side of an 11 kV to 0.415 kV transformer connected to the local utility. The electrical connection setup is similar to that of Building 1, with a data logger installed at the main switchboard to monitor energy usage at hourly intervals.

The independent variables considered for this building include lecturer occupancy, student occupancy in classrooms and laboratories, and outdoor temperature. Unlike the first building, lecturer presence in classrooms and laboratories is treated as a single combined variable. This is due to the nature of the timetable, where multiple sessions often occur simultaneously across both types of spaces. It is assumed that during these periods, lecturers make use of classroom equipment such as LCD TVs, lighting, and ventilation systems.

Student occupancy in classrooms and laboratories is estimated using the same method as in Building 1. The official academic timetable is used to identify all lecture and laboratory sessions, and the number of students registered for each session is retrieved from the student information system. These data are then used to calculate the number of students present on an hourly basis during operational hours. Outdoor temperature is also included as an external factor influencing energy demand.

Hourly temperature data is obtained from www.weatherunderground.com, providing consistent environmental input across both buildings. This variable is expected to influence the cooling load and overall energy consumption pattern in the building.

C. Case Building Summary

The summary of the independent variables is shown in Table I. In total, case building 1 (CB1) will have six independent variables and case building 2 (CB2) will have a total of four independent variables. In CB1, lecturer occupancy was segmented into three categories: office rooms, classrooms, and laboratories. This separation was adopted to account for the differences in energy usage patterns across various functional spaces. Lecture sessions and laboratory sessions typically do not occur simultaneously, and modeling these spaces independently allows the baseline energy model to capture their respective energy contributions more accurately. Additionally, lecturers are assumed to leave their office equipment powered on even when temporarily away for classes or laboratory sessions. As such, energy loads in office rooms may remain active despite the absence of occupants. This assumption justifies the inclusion of lecturer occupancy in office rooms as a distinct variable, enabling the model to better reflect continuous background energy consumption. In contrast, CB2 combines lecturer occupancy in classrooms and laboratories into a single variable due to the scheduling structure. Lectures and labs are often conducted simultaneously across different spaces, making it practical to aggregate occupancy data. Office occupancy is not considered for CB2 as there are no personal space (office room) in the building.

TABLE I. INDEPENDENT VARIABLES SUMMARY

Independent Variable	Building 1	Building 2
Lecturer occupancy in office rooms	Included	Not included
Lecturer occupancy in classrooms	Included (separate from labs)	Combined with lab occupancy
Lecturer occupancy in laboratories	Included (separate from classrooms)	Combined with classroom occupancy
Student occupancy in classrooms	Included	Included
Student occupancy in laboratories	Included	Included
Outdoor temperature	Included	Included

D. Multiple Linear Regression Model

The MLR model is a statistical technique used to quantify the relationship between a dependent variable and two or more independent variables. This model assumes a linear association and is commonly applied to predict an outcome based on several influencing factors. The general form of the MLR model is shown in (1):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathsf{K} + \beta_n X_n + \varepsilon \quad (1)$$

In this equation Y represents the dependent variable, β_0 is the intercept, β_1 , β_2 , and β_n are the coefficients of the independent variables X_1 , X_2, and X_n , respectively while and ε is the error term representing unexplained variation. The coefficients indicate the magnitude and direction of the effect each independent variable has on the dependent variable. The methodology for developing the multiple linear regression model in this study is shown in the flowchart in Fig. 1. The flow chart begins with the collection of independent variables data. These variables, which include occupancy patterns and outdoor temperature, are recorded based on the operational activities of the buildings during lecture weeks.

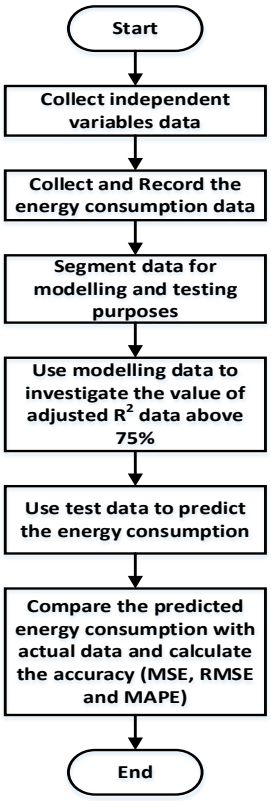

Fig. 1. Flowchart of Multiple Linear Regression Baseline Energy Model Development and Energy Prediction

Simultaneously, electrical energy consumption data is measured using data loggers installed at the main switch room, capturing hourly readings to reflect the building's daily operational profile. After data collection, the dataset is segmented into two parts, with one portion designated for model development and the remaining portion reserved for testing. The modelling data is used to build the multiple linear regression model, and its performance is assessed through the adjusted coefficient of determination to ensure that the model explains a significant proportion of the variance in energy consumption. A threshold of at least 75 percent adjusted R^2 is considered acceptable before proceeding further to prediction process.

Once the model is developed, the independent variables from the testing dataset are inserted into the model to predict energy consumption. The predicted values are then compared with the actual measured energy consumption from the testing dataset to evaluate the model's predictive capability. The accuracy of the predictions is assessed using three error metrics Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) and is shown in (2) – (4) respectively.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \quad (3)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \quad (4)$$

In Equation (2)-(4) Yi is the actual energy consumption, $\hat{Y}_i$ is the predicted energy consumption, and n is the number of observations in the testing dataset. These metrics were used to assess the predictive accuracy of the developed models for both case study buildings. These metrics quantify how close the predicted values are to the actual consumption, providing a clear measure of the model's precision.

Lower values of MSE and RMSE indicate that the predicted energy consumption closely approximates the measured data. The MAPE value represents the percentage of prediction error, where a smaller MAPE corresponds to higher prediction accuracy. In this study, no explicit threshold is defined to distinguish between good and poor model performance. Instead, the reliability of the developed baseline energy model is assessed in accordance with the IPMVP guideline, whereby a coefficient of determination (R^2) of 75% or higher is considered acceptable for prediction purposes [26].

In addition to the MSE, RMSE and MAPE, Akaike Information Criterion (AIC) will be utilized in order to evaluate how the model balances between the quality of fit with the simplicity of structure. In other words AIC evaluates the quality of a regression model by inspecting how the model fits the data while also considering the number of predictors included. The AIC will be calculated as shown in (5) :

$$AIC = n\ln\left(\frac{RSS}{n}\right) + 2k \quad (5)$$

In (5) the RSS is the Residual Sum of Squares, n is the total number of data points, and k is the number of model

parameters. The number of model parameters includes all the regression coefficeints as well as the intercept term of the regression model. The RSS were calculated based on the squared value of the summation residual between the actual energy consumption and predicted energy consumption as in (6).

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (6)$$

In (6) y_i is the actual energy consumption and $\hat{y}_i$ is the predicted energy consumption. The difference of $y_i - \hat{y}_i$ is the residual of the energy consumption prediction. In general, lower AIC value indicates a more efficient model because it reflects both lower residual errors and fewer predictors. For this reason, AIC is highly suitable for best subset selection where the objective is to choose the best combination of predictors without unnecessary complexity.

E. Best Subset Selection

The best subset selection (BSS) method is employed in this study to identify the most significant combination of independent variables that influence energy consumption. BSS evaluates every possible combination of independent variables and selects the subset that yields the best model performance based on adjusted coefficient of determination. This approach ensures that only the most relevant variables are included in the final model, avoiding the inclusion of unnecessary predictors that could lead to overfitting.

The methodology of the BSS is shown in flowchart in Fig. 2. It begins with the collection of independent variables data, which includes parameters such as occupancy patterns and outdoor temperature. These variables are recorded based on the operational activities of the building during lecture weeks. Alongside this, energy consumption data is measured and recorded using data loggers installed at the main switchboards, capturing hourly consumption values.

Following data collection, the dataset is segmented into two parts. One portion is used for model development (modelling data), while the remaining portion is reserved for model testing (testing data). The modelling data is used to develop various regression models, each representing a unique combination of independent variables. The total number of possible combinations is determined using the combination in (5):

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \quad (5)$$

where n is the total number of independent variables, and r is the number of variables selected in a combination. This ensures that all subsets, from single-variable models to models containing all variables, are evaluated systematically.

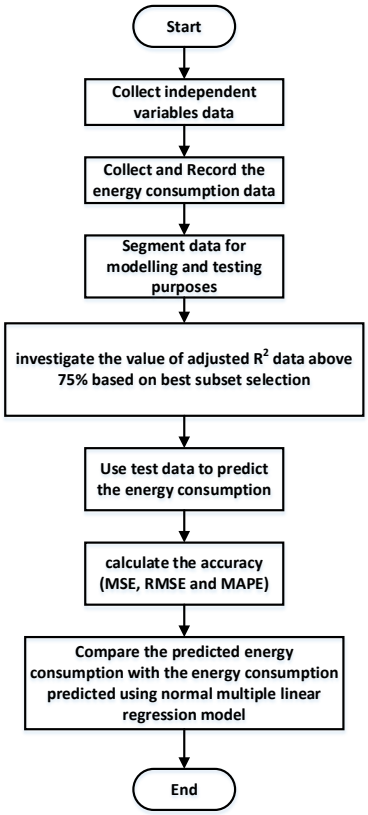

Fig.2. Best Subset Selection for Multiple Linear Regression Baseline Energy Model Development and Energy Prediction

For each combination, a multiple linear regression model is developed, and its performance is assessed using the adjusted coefficient of determination, calculated using (6):

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)(n_{data}-1)}{n_{data}-p-1} \quad (6)$$

where R^2 is the coefficient of determination, n_{data} is the number of observations, and p is the number of independent variables in the model. Only models with adjusted R^2 values exceeding 75 percent are considered acceptable. Once the best subset combination is identified based on the highest adjusted R^2, the independent variables from the testing dataset are inserted into the developed model to predict energy consumption. These predicted values are then compared to the actual energy consumption measurements from the testing data to evaluate the model's accuracy. The precision of the predictions is quantified using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

IV. RESULTS AND ANALYSIS

The results and analysis are presented in three main sections. The first section focuses on the results for Case Building 1, followed by the second section which discusses the results for Case Building 2. In both the Case Building 1 and Case Building 2 sections, the baseline energy model is developed using the Multiple Linear Regression (MLR) method and further

enhanced through the application of the best subset selection technique. These sections also present the energy consumption prediction results obtained from both models. The third section provides a comparative discussion of the results from Case Building 1 and Case Building 2. This comparison aims to offer insights and interpretations regarding the outcomes, highlighting the similarities, differences, and implications of the modelling approaches used in the study. The discussion that follows builds upon these findings to evaluate the suitability of each modelling approach for practical energy management applications in academic buildings.

A. Case Building 1

For the first case study building, the MLR model was developed using six independent variables which is the independent variables were identified as staff occupancies (X_1), students in classrooms (X_2), students in laboratories (X_3), outdoor temperature (X_4), lecturers in classrooms (X_5), and lecturers in laboratories (X_6). The dependent variable, denoted as Y, represents the energy consumption of the building in kilowatt-hours (kWh). The total dataset consisted of 1,464 data points, which 840 were allocated for model development and 624 were used for testing the model's predictive performance. The 840 data points is an hourly interval data points from 35 days. Subsequently, the 624 data points is an hourly interval data point from 26 days.

Using the 840 data points for model development, the MLR model coefficients for X_1, X_2, X_3, X_4, X_5, and X_6 were determined to be 1.18090, 0.05311, 0.002762, 5.18745, -0.1837, and -0.293256 respectively, with an intercept value of -117.13646 and is shown in (7):

$$Y = 1.18090X_1 + 0.05311X_2 + 0.002762X_3 + 5.18745X_4 - 0.1837X_5 - 0.293256X_6 - 117.13646 \quad (7)$$

The best subset selection identified three variables which is staff occupancies (X_1), students in classrooms (X_2), and outdoor temperature (X_4) out of 63 possible combination based on (5). For these variables, the model coefficients were calculated as 1.1786, 0.04946, and 5.1526 respectively, with an intercept of -116.19 and is shown in (8):

$$Y = 1.1786 + 0.04946X_2 + 5.1526X_4 - 116.19 \quad (8)$$

Table II shows the results for model development in CB1. The MLR model using all independent variables achieved an R^2 value of 0.84259, an adjusted R^2 value of 0.84146, and an Akaike Information Criterion (AIC) value of 6974.8. The enhance model i.e. the best subset selection method, resulting model achieved the same R^2 value of 0.84259 and adjusted R^2 value of 0.84146, with a slightly lower AIC value of 6969.6, indicating a marginal improvement in model simplicity.

TABLE II. CASE BUILDING 1 MODEL DEVELOPMENT RESULT

Model Type	Independent Variables Used	R^2	Adjusted R^2	AIC
MLR (All Variables)	X_1, X_2, X_3, X_4, X_5, X_6	0.84259	0.84146	6974.8
MLR with Best Subset Selection	X_1, X_2, X_4	0.84259	0.84146	6969.6

When tested with the 624 data points in the testing dataset, the model using all independent variables achieved a MSE of 255.6228 kWh, RMSE of 15.9882 kWh, and MAPE of 40.91 percent. The model derived from the best subset selection achieved an MSE of 255.0059 kWh, an RMSE of 15.9689 kWh, and a MAPE of 40.83 percent. The corresponding prediction results is shown in Table III and the plot of actual energy consumption with the predicted energy consumption for MLR model and MLR with BSS is shown in Fig. 3 and Fig. 4 respectively. In Fig. 3, the predicted energy consumption (dashed line) is closely resembles the actual energy consumption. Nontheless the predicted energy cosnsumption has some deviation between the lower peak and upper peak of the graph plot. In Fig. 4 similar pattern was observed between the predicted energy consumption and the actual consumption but the deviation between the lower peak and upper peak of the plotted consumption was reduced.

TABLE III. CASE BUILDING 1 PREDICTION ACCURACY RESULTS

Model Type	MSE (kWh)	RMSE (kWh)	MAPE (%)
MLR (All Variables)	255.6228	15.9882	40.91
MLR with Best Subset Selection	255.0059	15.9689	40.83

B. Case Building 2

For the second case study building, the MLR model was developed using four independent variables. These variables were outdoor temperature (X_1), students in classrooms (X_2), students in laboratories (X_3), and staff occupancies (X_4). The dependent variable, Y, represents the energy consumption of the building in kilowatt-hours (kWh). The dataset consisted of 1,128 data points, with 408 points allocated for model development and 720 points used for testing the model's predictive capability. The 408 data point is an hourly data point for 17 days while the 720 data point is an hourly interval for 30 days.

Using the 408 data points for model development, the MLR model coefficients for X_1, X_2, X_3, and X_4 were determined to be 2.5039, 0.01671, 0.10225, and -0.30372 respectively, with an intercept value of -48.786. The equation of the MLR model is shown in (9):

$$Y = 2.5039X_1 + 0.01671X_2 + 0.10225X_3 - 0.30372X_4 - 48.786 \quad (9)$$

To enhance the model, the best subset selection method was applied to the same 408 data points to identify the most relevant variables for predicting energy consumption. The best subset selection identified three variables i.e. outdoor temperature (X_1), students in classrooms (X_2), and students in laboratories (X_3) out of 15 possible combinations. For these variables, the model coefficients were calculated as 2.4795, 0.029534, and 0.11129 respectively, with an intercept of -48.114. The equation of the MLR best subset model is shown in (10).

$$Y = 2.4795X_1 + 0.029534X_2 + 0.11129X_3 - 48.114 \quad (10)$$

The MLR model using all independent variables achieved an R^2 value of 0.78446, an adjusted R^2 value of 0.78232, and an Akaike Information Criterion (AIC) value of 2971.3. The resulting model enhanced by the best subset selection achieved an R^2 value of 0.78401 and an adjusted R^2 value of 0.78241, with an AIC value of 2970.1 the results of this model performances is shown in Table IV.

TABLE IV. CASE BUILDING 2 MODEL DEVELOPMENT RESULT

Model Type	Independent Variables Used	R^2	Adjusted R^2	AIC
MLR (All Variables)	X_1, X_2, X_3, X_4	0.78446	0.78232	2971.3
MLR with Best Subset Selection	X_1, X_2, X_3	0.78401	0.78241	2970.1

When tested using the 720 data points in the testing dataset, the model with all independent variables achieved a MSE of 118.7784 kWh, RMSE of 10.9896 kWh, and MAPE of 24.92 percent. The model developed from the best subset selection achieved an MSE of 119.2479 kWh, an RMSE of 10.9201 kWh, and a MAPE of 24.96 percent. The related energy consumption prediction results is shown in Table V. the plot of actual energy consumption with the predicted energy consumption for MLR model and MLR with BSS is shown in Fig. 5 and Fig. 6 respectively. Fig. 5 and Fig. 6 shows a similar behaviour as in Fig. 3 and Fig. 4. In Fig.5 the predicted energy consumption (dashed line) using MLR resembles similar pattern to the actual energy consumption but the is a majority difference for the upper peak and lower peak of the plotted energy consumption. In Fig 6. the predicted energy consumption using MLR BSS depict that certain upper peak and lower peak is almost similar with the actual energy consumption.

TABLE IV. CASE BUILDING 2 PREDICTION ACCURACY RESULTS

Model Type	MSE (kWh)	RMSE (kWh)	MAPE (%)
MLR (All Variables)	118.7784	10.9896	24.92
MLR with Best Subset Selection	119.2479	10.9201	24.96

C. Comparative Discussions Between MLR Baseline Energy model and MLR Best Subset Selection Baseline Energy Model

The results from both case studies were examined to compare the performance of the MLR models using all independent variables with the MLR-BSS method. The comparison focuses on two key aspects, which are model performance during development and prediction accuracy when applied to testing datasets. The prediction accuracy comparison and percentage difference is shown in Table V.

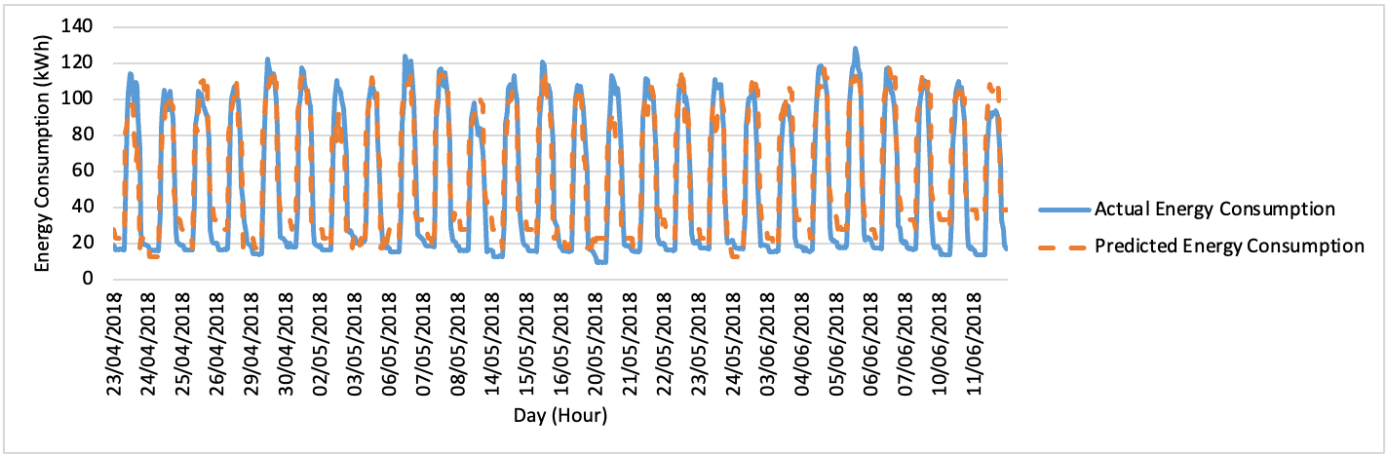

Fig. 3 Case Building 1 Predicition Results Using MLR Baseline Energy Model

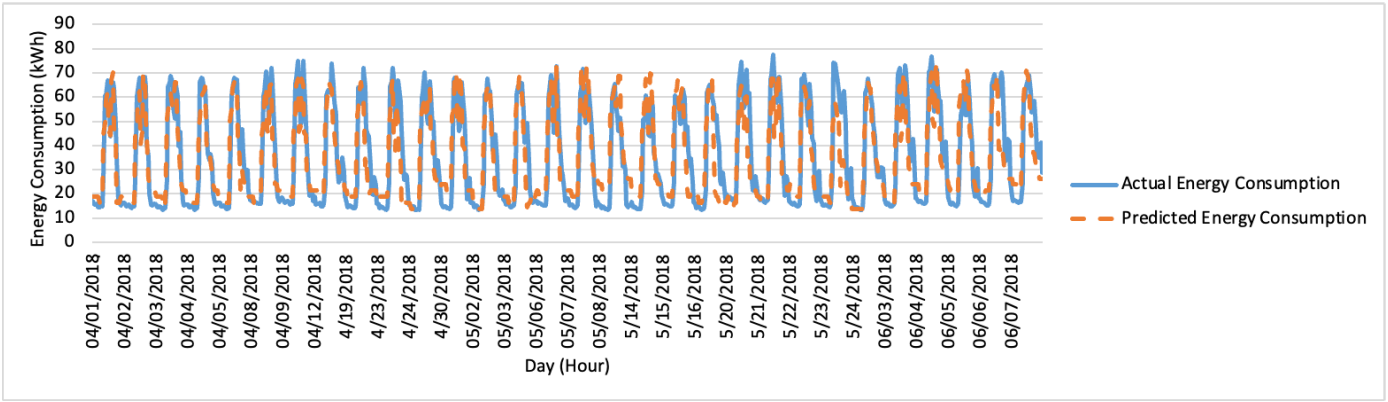

Fig. 4 Case Building 1 Prediction Results using MLR BSS enhanced Baseline Energy Model

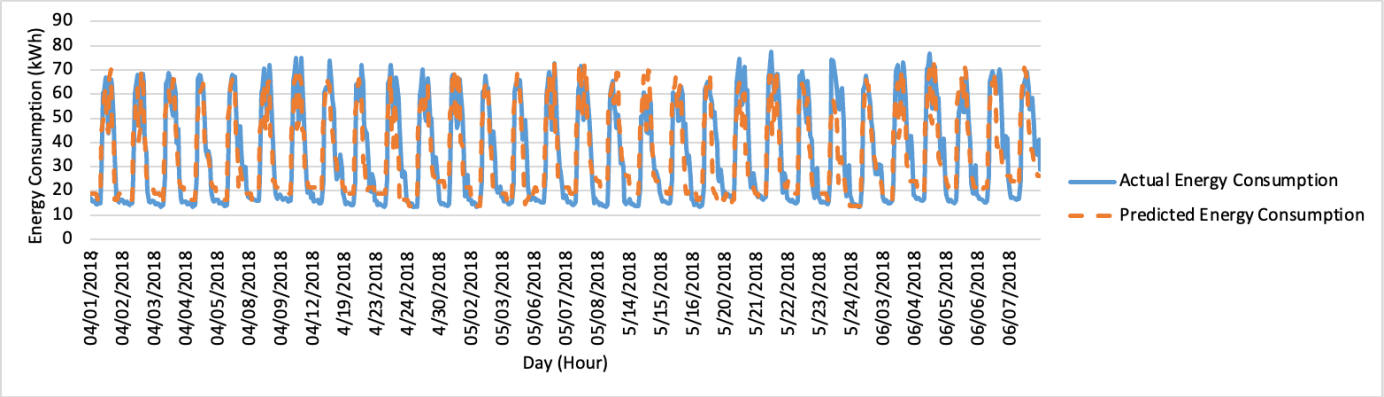

Fig. 5. Case Building 2 Predicition Results Using MLR Baseline Energy Model

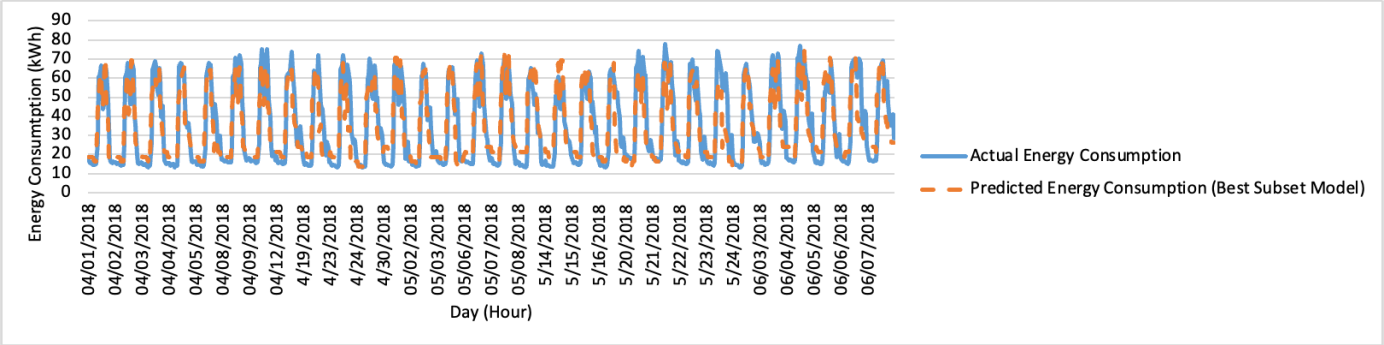

Fig. 6 Case Building 2 Prediction Results using MLR BSS enhanced Baseline Energy Model

TABLE V. PREDICTION ACCURACY COMPARISON AND PERCENTAGE DIFFERENCE

Case Building	Model Type	MSE (kWh)	RMSE (kWh)	MAPE (%)	ΔMSE (%)	ΔRMSE (%)	ΔMAPE (%)
1	Full MLR	255.6228	15.9882	40.91	–	–	–
	Best Subset (X1, X2, X4)	255.0059	15.9689	40.83	–0.24%	–0.12%	–0.20%
2	Full MLR	118.7784	10.8986	24.92	–	–	–
	Best Subset (X1, X2, X3)	119.2479	10.9201	24.96	+0.40%	+0.20%	+0.16%

The analysis of both case studies shows how MLR and their enhancement through BSS perform in developing baseline energy models and predicting energy consumption. In CB1, the application of BSS simplified the model by reducing the number of independent variables without disturbing the model perfomances. Both the coefficient of R and R^2 value were almost identical for the full and enhanced models, but the best subset model achieved a slightly lower AIC value. This indicates that the BSS method improved the simplicity of the model while maintaining the same prediction accuracy. In addition, the prediction results in CB1 displayed a small improvements. The MSE decreased (negative value in Table V) by 0.24 percent, the RMSE decreased by 0.12 percent, and the MAPE decreased by 0.20 percent. Although these changes are minuscule, it shows demonstrate that simplifying the model by discarding the less influential variables led to better prediction performance.

In CB2 meanwhile different outcome was observed. The BSS MLR model produced a slightly lower AIC compared to the full MLR model. However, the prediction accuracy did not improve. The MSE increased by 0.40 percent, the RMSE increased by 0.20 percent, and the MAPE increased by 0.16 percent. These differences can be considered very small which is below one percent implying that the impact on prediction is not significant. Even though the exclusion of staff occupancy (X_4) reduced the ability of the model to replicate the test data performance, building owners would still be able to use the simpler model with reasonable confidence. Furtrhermore, the the MAPE results was not directly intended to compared between CB1 and CB2. The MAPE is was used to evaluate how the prediction accuracy changes within each building after BSS method was applied to the MLR model.

Comparing both case studies, the same observation was presence that BSS reduces the AIC and therefore decreases model complexity. However, its effect on prediction accuracy depends on the context of the building. For CB1, the reduction in complexity improved prediction slightly, while for CB2, it caused a small reduction in accuracy. Overall, the results suggest that BSS can be a useful method for guiding the choice of independent variables. It provides flexibility in deciding whether the priority should be achieving maximum accuracy or a more efficient and simplified model structure.

V. CONCLUSIONS

The objectives of this study were successfully achieved by developing MLR models for both case study buildings and applying the BSS method to identify the most significant independent variables. The key findings show that in case CB 1 the BSS slightly reduced the prediction accuracy MSE (255.0059 kWh), RMSE (15.9689) and MAPE (40.83%) compared to using all independent variables i.e. MSE (255.6228 kWh), RMSE (15.9882) and MAPE (40.91%). In CB 2, the prediction accuracy i.e. MSE (119.2479 kWh),RMSE (10.9201 kWh), MAPE (20.96%) from the BSS is higher compared to the MLR model without using BSS i.e. MSE (118.7784 kWh),RMSE (10.8986 kWh), MAPE (20.92%) but in a very minor value. Nonetheless, it suggest that in CB2 the absence of the staff occupancies variable (X_4) did not cause a significant drop in accuracy and therefore building owners may not need to be concerned about its exclusion. The BSS method provides valuable insights for building owners and energy managers. It can guide decisions on whether to use a simpler model with fewer variables or retain all variables when developing baseline energy models. This is particularly useful when planning energy conservation measures and estimating potential energy savings as it offers flexibility in determining which variables are most important for accurate modelling and decision making. The independent variable selection from the BSS method can be used in other modelling method such as Machine Learning Model and Deep Learning Model to aid in increasing the prediction accuracy where it is helpful in modelling certain non-linear relation that may exist between the energy consumption with the respective independent varaibles.

ACKNOWLEDGMENT

Authors would like to thank all personel in Universiti Teknologi MARA that have contribute directly or indirectly towards the work that have been conducted pertaining to the title of this paper.

REFERENCES

[1] IEA, " Electricity 2025, IEA, Paris," in "CC BY 4.0," 2025. [Online]. Available: https://www.iea.org/reports/electricity-2025

[2] IEA, "Energy Efficiency Policy Toolkit 2025 IEA, Paris," 2025, vol. CC BY 4.0. [Online]. Available: https://www.iea.org/reports/energy-efficiency-policy-toolkit-2025,

[3] B. Dong, S. E. Lee, and M. H. Sapar, "A holistic utility bill analysis method for baselining whole commercial building energy consumption in Singapore," *Energy and Buildings,* vol. 37, no. 2, pp. 167-174, 2005/02/01/ 2005, doi: https://doi.org/10.1016/j.enbuild.2004.06.011.

[4] M. Maaouane, S. Zouggar, G. Krajačić, and H. Zahboune, "Modelling industry energy demand using multiple linear regression analysis based on consumed quantity of goods," *Energy,* vol. 225, p. 120270, 2021/06/15/ 2021, doi: https://doi.org/10.1016/j.energy.2021.120270.

[5] E. Sarmas, A. Forouli, V. Marinakis, and H. Doukas, "Baseline energy modeling for improved measurement and verification through the use of ensemble artificial intelligence models," *Information Sciences,* vol. 654, p. 119879, 2024/01/01/ 2024, doi: https://doi.org/10.1016/j.ins.2023.119879.

[6] Z. Afroz, H. Burak Gunay, W. O'Brien, G. Newsham, and I. Wilton, "An inquiry into the capabilities of baseline building energy modelling approaches to estimate energy savings," *Energy and Buildings,* vol. 244, p. 111054, 2021/08/01/ 2021, doi: https://doi.org/10.1016/j.enbuild.2021.111054.

[7] H. Kuivjõgi, S. Vasman, E. Petlenkov, M. Thalfeldt, and J. Kurnitski, "Data-driven baseline generation for post-retrofit energy saving assessment, a comparison of statistical and machine learning methods," *Journal of Building Engineering,* vol. 98, p. 111016, 2024/12/01/ 2024, doi: https://doi.org/10.1016/j.jobe.2024.111016.

[8] Y. Lu, X. Peng, C. Li, Z. Tian, J. Niu, and C. Liang, "A baseline model combining physics and data-driven approach for operation evaluation of district heating substation," *Energy and Buildings,* vol. 321, p. 114582, 2024/10/15/ 2024, doi: https://doi.org/10.1016/j.enbuild.2024.114582.

[9] A. Aranda, G. Ferreira, M. D. Mainar-Toledo, S. Scarpellini, and E. Llera Sastresa, "Multiple regression models to predict the annual energy consumption in the Spanish banking sector," *Energy and Buildings,* vol. 49, pp. 380-387, 2012/06/01/ 2012, doi: https://doi.org/10.1016/j.enbuild.2012.02.040.

[10] R. F. Mustapa *et al.*, "Educational Building's Energy Consumption Independent Variables Analysis using Linear Regression Model: A Comparative Study," in *2023 IEEE 3rd International Conference in Power Engineering Applications (ICPEA)*, 6-7 March 2023 2023, pp. 202-207, doi: 10.1109/ICPEA56918.2023.10093222.

[11] R. F. Mustapa, A. H. M. Nordin, M. A. Hairuddin, M. E. Mahadan, N. Y. Dahlan, and I. M. Yassin, "The Effect of Occupant in Energy Consumption Prediction via Multiple Linear Regression Model in an Educational Building," in *Smart Grid and Renewable Energy Systems*, Singapore, M. L. Kolhe, Ed., 2024// 2024: Springer Nature Singapore, pp. 95-101.

[12] N. H. Hussin, R. M. S. Aditya, and N. Ishak, "Modelling and predicting energy consumption in laboratory buildings using multiple linear regression," *Jurnal Ilmiah Matematika,* vol. 8, no. 1, 2021, doi: 10.26555/konvergensi.v8i1.21459.

[13] H. Yang, M. Ran, and C. Zhuang, "Prediction of Building Electricity Consumption Based on Joinpoint−Multiple Linear Regression," *Energies*, vol. 15, no. 22, doi: 10.3390/en15228543.

[14] J. A. Potschka, M. O. Oliveira, M. A. Mazzoletti, and R. C. Brazzola, "Electric Energy Consumption Prediction in Technological Education Buildings Using Linear Regression Method," in *2022 IEEE Biennial Congress of Argentina (ARGENCON)*, 7-9 Sept. 2022 2022, pp. 1-6, doi: 10.1109/ARGENCON55245.2022.9939900.

[15] A. Mohammed, A. Alshibani, O. Alshamrani, and M. Hassanain, "A regression-based model for estimating the energy consumption of school facilities in Saudi Arabia," *Energy and Buildings,* vol. 237, p. 110809, 2021/04/15/ 2021, doi: https://doi.org/10.1016/j.enbuild.2021.110809.

[16] S. Di Leo, P. Caramuta, P. Curci, and C. Cosmi, "Regression analysis for energy demand projection: An application to TIMES-Basilicata and TIMES-Italy energy models," *Energy,* vol. 196, p. 117058, 2020/04/01/ 2020, doi: https://doi.org/10.1016/j.energy.2020.117058.

[17] O. Ø. Smedegård, T. Jonsson, B. Aas, J. Stene, L. Georges, and S. Carlucci, "The Implementation of Multiple Linear Regression for Swimming Pool Facilities: Case Study at Jøa, Norway," *Energies*, vol. 14, no. 16, doi: 10.3390/en14164825.

[18] Y. Chen, M. Huang, and Y. Tao, "Density-based clustering multiple linear regression model of energy consumption for electric vehicles," *Sustainable Energy Technologies and Assessments,* vol. 53, p. 102614, 2022/10/01/ 2022, doi: https://doi.org/10.1016/j.seta.2022.102614.

[19] X. Chen, X. Peng, Y. Li, and B. He, "Based on the improved fuzzy analytic hierarchy and the TSE-MLR model energy consumption prediction of university: A case study," *Heliyon,* vol. 10, no. 17, p. e36979, 2024/09/15/ 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e36979.

[20] Z. Zhao, Y. Peng, X. Zhu, X. Wei, X. Wang, and J. Zuo, "Research On Prediction Of Electricity Consumption In Smart Parks Based On Multiple Linear Regression," in *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, 11-13 Dec. 2020 2020, vol. 9, pp. 812-816, doi: 10.1109/ITAIC49862.2020.9338976.

[21] H.-x. Zhao and F. Magoulès, "A review on the prediction of building energy consumption," *Renewable and Sustainable Energy Reviews,* vol. 16, no. 6, pp. 3586-3592, 2012/08/01/ 2012, doi: https://doi.org/10.1016/j.rser.2012.02.049.

[22] A. T. Nguyen, Y. Ahn, S. Park, S. Park, and D. H. Pham, "Meta learning regression framework for energy consumption prediction in retrofitted buildings: A case study of South Korea," *Journal of Building Engineering,* vol. 96, p. 110403, 2024/11/01/ 2024, doi: https://doi.org/10.1016/j.jobe.2024.110403.

[23] Y. Chen, W. Gong, C. Obrecht, and F. Kuznik, "A review of machine learning techniques for building electrical energy consumption

prediction," *Energy and AI,* vol. 21, p. 100518, 2025/09/01/ 2025, doi: https://doi.org/10.1016/j.egyai.2025.100518.

[24] N. Fumo, "A review on the basics of building energy estimation," *Renewable and Sustainable Energy Reviews,* vol. 31, pp. 53-60, 2014/03/01/ 2014, doi: https://doi.org/10.1016/j.rser.2013.11.040.

[25] H. Fang, H. Tan, N. Dai, Z. Liu, and R. Kosonen, "Hourly Building Energy Consumption Prediction Using a Training Sample Selection Method Based on Key Feature Search," *Sustainability*, vol. 15, no. 9, doi: 10.3390/su15097458.

[26] *Efficiency Valuation Organization (EVO), International Performance Measurement and Verification Protocol: Core Concepts,* E. 10000-1:2016, 2016.