Investigator(s) Information

Last name	First name	Student ID
Rayegan	Saeed	40204050
Aliakbarlou	Hamid	40239192

A data mining-based approach to forecast the buildings' load

Abstract *Buildings play a significant role in the global energy consumption and the greenhouse gas (GHG) emissions. Therefore, forecasting, and characterizing the building energy consumption is the first step for improving the energy management to mitigate the impact of GHG emissions on the climate change and maintain sustainability in smart cities. In this regard, sensor-based approaches are practical to infer the complex relationship between building consumption and impactful parameters, such as the outdoor weather, time of day, and the indoor environmental parameters. In the current study, two machine learning models are applied to examine the impact of temporal (i.e., one-min, 15- min, and hourly intervals) and spatial (zone and floor level) granularity on the load predictions of an office building. Artificial Neural Network (ANN) and Multiple Linear Regression (MLR) models are applied to a sensor-based gathered dataset for an office building in Bangkok, Thailand. The results show that the predictions obtained from the ANN model is much closer to the actual values than the MRL model predictions. Moreover, both models forecast the building load much better using the floor level and hourly time intervals data. Our findings have a practical significance for deploying and installing advanced smart metering devices by finding the scales for which data can be extracted and transformed into meaningful information.*

1. Introduction and Background

According to the reports, the building sector is responsible for almost one-third of energy consumption and 15% of direct CO_2 emissions worldwide [1]. Therefore, there is an urgent call to move toward energy-efficient buildings by all means, and any attempt to find a solution can make a huge difference. Many efforts have been made to study the energy performance as well as the energy management in buildings. Data-driven models are robust tools that enable examining the energy operation in buildings, which are not feasible through experiments nor efficient by numerical studies. Data-driven models are the first keys to pinpointing the inefficiencies in different parts of a building where the energy is used abnormally or is wasted.

Load forecasting is the vital step to finding the most significant parameters to reduce the energy consumption of the building. This study aims to forecast the building load based on various building-related data, such as internal loads (air conditioning (AC) units, lighting, and plug), indoor conditions (temperature, humidity, and illumination), as well as outdoor conditions (temperature and radiation) to contribute to the mentioned global alarming problem.

2. Problem Statement and Objective

Building operational (i.e., AC units, lighting, and plug loads) and indoor/outdoor environmental (i.e., indoor temperature, humidity, and illumination and outdoor temperature and radiation) parameters characterize the building load. To the best of our knowledge, a comprehensive evaluation of the load prediction in buildings, specifically at zone level, has not been studied adequately.

The electricity loads for individual AC units, lighting, and plug loads with indoor temperature, humidity, and illumination are available in every zone of a 7-story office building as a case study located in Bangkok, Thailand [2]. Moreover, outdoor temperature and radiation for this location are available, which can be added to the dataset. Our project's primary goal is to implement a data mining approach to predict the building load at the zone and the floor level.

Based on the above problem statement, the objectives of the current work are:

I. Load Prediction- Obtaining a model to predict the building load:

For this object, regression models need to be applied to predict the load of the building based on sensory-gathered data. Moreover, the models should be compared in terms of accuracy and errors to find the best predictions for load forecasting.

II. Spatial and time resolution granularity affects load forecasting:
   - ✓ Spatial resolution: zone and floor level
   - ✓ Time resolution: one, 15, and 60-minute time resolution of the dataset

The dataset selected here contains the building operational and indoor environmental parameters at 1-minute intervals. For the time resolution study, data in attributes should be averaged over different time intervals, for instance, over 5 minutes, 15 minutes, and so on, to prepare separate datasets for modelling. Furthermore, the summation of all zones' loads and the average of indoor parameters should be calculated for the floor level modelling (the details will be discussed in data preparation).

For this objective, regression models need to be applied and the accuracy of models for various spatial/time resolutions should be compared to study the effect of spatial/time granularity of the dataset on predictions.

3. Previous Works

Load forecasting of buildings has been paid attention in past decays as the first step to find new solutions to reduce the building energy demand. Zhang et al. [3] performed a comprehensive review of using machine learning technique for building load predictions. They argued that the building load prediction under demand response and building-grid interaction in smart cities are becoming more complicated and challenging. Li et al. [4] developed a data mining framework to obtain energy-related feedback and identify the energy wastage patterns in the building. Cordeiro-Costas et al. [5] used the machine learning tool to optimize the electricity consumption in the building based on the data of an example office building. They showed that a reduction of 52.72 € in the electricity bills was obtained. K. Jain et al. [6] investigated the impact of time and spatial resolution granularity on the performance of the machine learning models and the accuracy of the building load predictions. They pointed out the hourly data at the floor level yields the best prediction using the machine learning tool.

To the best of our knowledge, there is a lack of information regarding a detailed approach using the machine learning technique to forecast the building load at different time and spatial resolutions. Thus, the aim of current study is to take a deeper look at the problem stated here, and finally, the future works are recommended at the end of the conclusion.

4. Methodology

This section describes the comprehensive CRISP-DM process followed in our study. The CRISP-DM process consists of several consecutive stages with back and force feedbacks as: Understanding Data, Data Preparation, Modeling, and Evaluation which are carefully followed and introduced in the present work.

Understanding Data

The CU-BEMS [2] dataset (available in Kaggle website) includes the electrical loads of the lighting, the plug, and the AC units together with indoor environmental measurements for every single zone of a 7-story office building with a total area of 11,700 m^2, in Bangkok, Thailand. The CU-BEMS dataset reports these three electrical loads separately with indoor conditions at one-minute intervals. The data are collected over 18 months, from July 1st, 2018, to December 31st, 2019. Table 1 summarizes the information of attributes available in the dataset. The data for floor #6 collected during 2019 are used in our modeling.

Figure 1 shows the schematic view of the building and the layout of floor #6. Moreover, Table 2 tabulates the statistical reports of the attributes in floor #6, zone #1 before starting cleaning or any data processing. It should be noted that to keep the report short, the data in Table 2 is just listed for zone 1 of floor 6, and approximately similar statistical values are observable for other zones in this floor. The sensors used to gather the data of indoor temperature, humidity, and illuminations were off for maintenance for several months. This is why numbers of non-null cells for indoor environmental parameters are smaller than indoor loads. Setting aside the time that the sensors were off, a small portion of examples have missing values. Furthermore, the dataset originally did not include the outdoor weather data. Two new attributes, outdoor temperature and radiation, with one minute resolution are introduced to the original dataset to include the effect of outdoor conditions

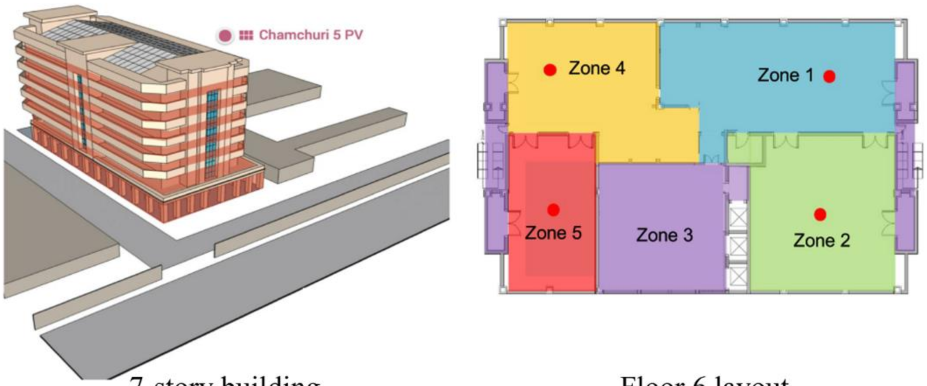
Figure 2 illustrates the variations in zone #1 AC load versus the zone indoor temperature and humidity and outdoor temperature for one week. As it can be seen, when the AC unit starts working, indoor temperature and humidity reduce, and when the AC unit turns off, indoor temperature and humidity increase. Moreover, the operation time of the AC unit is during the daytime. These trends are the technical validation of the dataset.

Table 1: Measured data description in CU-BEMS dataset

[1] *i is the zone number and varies between 1 to 5 (showing zone number)*

Attribute name in dataset	**Attribute exact name**	**Unit**	**Datatype**
Z(i[1])_AC(j[2])	AC unit load	kW	Numerical
Z(i)_Light (j)	Lighting load	kW	Numerical
Z(i)_Plug (j)	Plug load	kW	Numerical
Z(i)_S (j)	Zone temperature	°C	Numerical
Z(i)_S (j)	Zone relative humidity	%	Numerical
Z(i)_S (j)	Zone Ambient light	lux	Numerical

[2] *j is the floor number and varies between 1 to 7 (showing floor number)*



7-story building

Floor 6 layout

Figure 1. Seven-story building with the layout of floor #6 used in our modeling

Table 2: Statistical reports of the attributes in floor #6, zone #1 before cleaning

Attribute	# of non-null cells[1]	Center tendency statistics	Spread Statistics	Independent or Target attribute?

AC unit load (kW)	423560	Min = 0.00
		Max = 51.33
		Mean = 4.27
			Dev = 8.64	Target
				(The summation of 3 loads is the target)

Lighting load (kW)	423525	Min = 0.00
		Max = 7.33
		Mean = 1.38
			Dev = 1.64	

Plug load (kW)	423465	Min = 0.00
		Max = 3.27
		Mean = 0.44
			Dev = 0.34	

Zone temperature (°C)	315353	Min = 18.38
		Max = 32.90
		Mean = 25.26
			Dev = 2.35	Independent

Zone relative humidity (%)	315352	Min = 44.75
		Max = 86.90
		Mean = 64.90
			Dev = 6.58	Independent

Zone Ambient light (lux)	315350	Min = 0.00
		Max = 85.00
		Mean = 19.10
			Dev = 22.90	Independent

[1] *The total number of examples for floor #6, year 2019 is 424169*

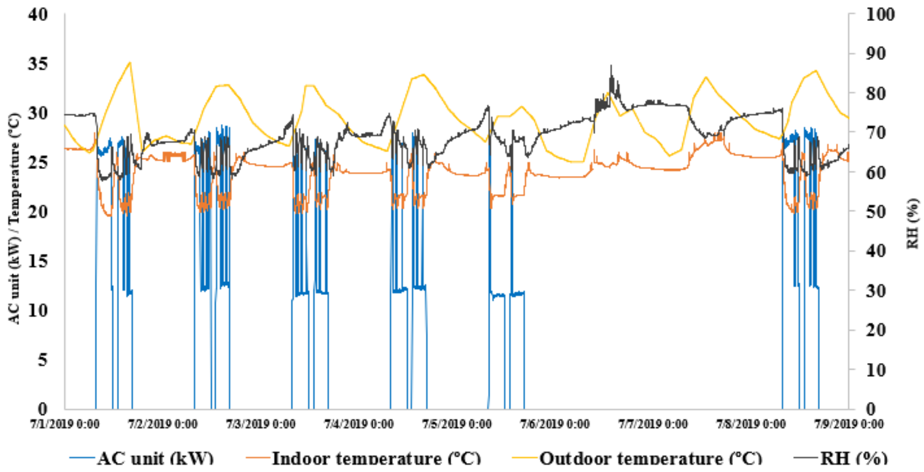


Figure 2. Zone #1 AC unit load versus indoor temperature and humidity and outdoor temperature for one week

Data preparation

a) Outdoor Condition

Outdoor temperature and radiation are added to the dataset as two new attributes, from the following reference [7].

b) Indoor Environmental Sensors Off-time

The time when the indoor environmental sensors were off is omitted from the dataset. Due to this filter, the examples before March 7th are omitted. It should be noted that a datetime attribute is available in the dataset which does not have any missing values. This attribute is used in the dataset preparation but set aside for modeling.

c) Office Hours and Weekdays

Using the datetime attribute, just the examples for office hours (8 AM to 5 PM) and weekdays (Monday to Friday) are kept and others are dropped. The reason is that outside the office hours and weekends, the AC unit load (which represents the largest part of the building loads) is zero.

d) Missing data imputation

The missing data for each attribute is filled with the average of that attribute. The reason is that, after doing previous steps, most of the data in each attribute are available. Therefore, filling the null values with the mean values does not affect the accuracy of results.

e) Label Attribute

In each zone, the label attribute is the summation of the three loads, lighting, plug, and AC loads. In other words, one new attribute which is the summation of these three loads is created and these three individual loads are omitted.

f) Floor Level Dataset

To study the building load at floor level, 4 new attributes are created in a separate script as follows: label attribute: summation of all loads in all zones; indoor temperature: the average of the temperature in all zones; indoor humidity: the average of the humidity in all zones; indoor illumination: the average of the illumination in all zones. In addition to the above 4 attributes, outdoor temperature and radiation are added to the dataset.

g) Outliers

Figure 3 shows the box plot diagram for all the attributes of floor 6 after the previous pre-processes. To the best of our knowledge, we concluded that all the attributes are in the normal range and no outlier can be found in the dataset although the diagram alarms the possibility of outliers.

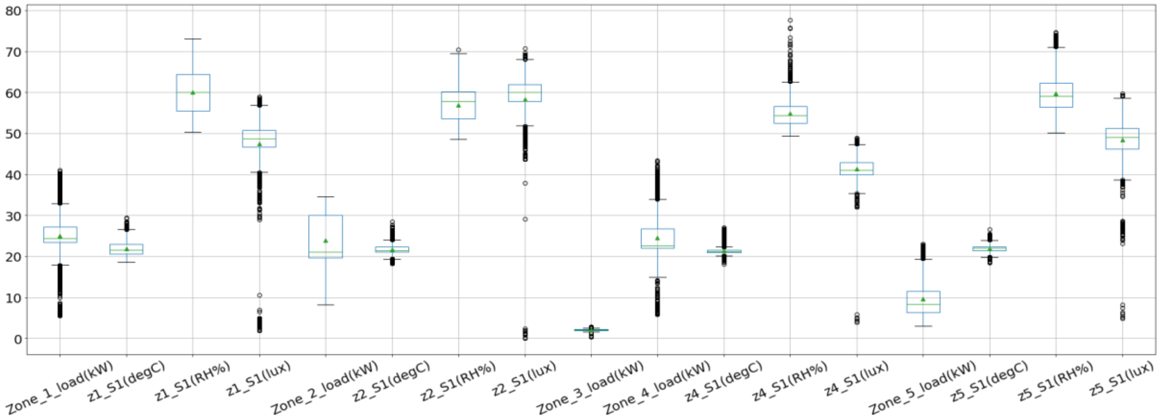


Figure 3. The box plot diagram of the floor #6 attributes

h) Dataset Time Resolution

The dataset time, originally is 1 minute interval, need to be prepared with different time intervals. In other words, different datasets with 5-min, 10-min, …, 60-min time resolutions need to be created. Using the Python function which changes the frequency of the data based on the date, easily adjustable datasets with different time resolutions are created. All the data are averaged over the frequency (5 minutes, 10 minutes, …, 60 minutes, and so on) specified by the user.

i) Normalization

Since all the attributes (both label and predictors) are numerical type and have different unit of measure, both label and predictors are normalized between 0 and 1 before starting the modeling section. After doing all the previous data processing, the Pearson correlation matrix is created containing all the attributes. This figure is not shown here to keep the report short.; however, the figure can be generated by running the script of the codes developed for the ANN. Based on the observed information in this figure, there is a weak linear collinearity between the load of each zone and its indoor environmental parameters and outdoor data. Hence, we expect that the linear regression model (LRM) cannot predict the loads with appropriate precision.

j) Shuffling the Dataset

Before starting the modeling part, the dataset is shuffled to ensure that the data from different months will be used in training and final testing of the models trained.

Modeling

Based on the studies, some machine learning models are trained for load forecasting. Multi-Linear Regression (MLR) and Artificial Neural Networks (ANN) models are implemented as previously stated in the literature.

Multiple Linear Regression (MLR)

MLR model is the most straightforward tool for regression problems and is used as the first attempt when all attributes are numerical type, and it is easy to implement. We expect that the MLR cannot generate satisfactory predictions as the physics of the problem here is highly non-linear. The MLR looks for the best linear combination of the attributes to predict the target.

Artificial Neural Network (ANN)

The feedforward and backpropagation network are adopted for ANN as the second model. ANN is often implemented because of its robustness to predict multivariate and nonlinear problems [8]. An ANN structure is an interconnected array of processors called neurons [9]. An ANN architecture consists of input, hidden, and output layers. Each layer has several neurons, and in each layer, they are connected to the neurons in the adjacent layer with different weights. Signals as inputs enter the first layer (input layer), pass the hidden layers, and finally reach the output stage. The main parameters of an ANN model are the number of hidden layers and neurons and the learning rate. Usually, a trial-and-error process should be followed to obtain a suitable number of hidden neurons, as done in our study. Based on our findings, the best number of neurons is strongly related to the time resolution of the dataset. Finally, we considered 30 hidden neurons for a 1-minute time resolution and 10-15 hidden neurons for a 60-minute time interval to obtain the most optimal results. Except for the 1-minute time interval, all other simulations are done using two hidden layers. Increasing the hidden layers to three has provided better results for 1-minute temporal resolution.

Cross-validation is used to evaluate the ANN model and monitor the overfitting problem. The dataset is divided into training, validation, and testing sets which will be discussed in the following sections. It is worth noting that the second one is used to avoid the overfitting problem.

Spatial & Temporal Granularity Analysis (Modeling Approach):

To investigate the impact of spatial and temporal granularity effects, we examined the effects of spatial levels, including floor and zone levels, as well as three temporal levels, 1-, 15-, and 60-minutes intervals, on predictions. For temporal granularity analysis, the load of each zone at three different time intervals are predicted. For the floor level analysis, the model predicts the summation of loads of individual zones.

Performance Evaluation

For the performance evaluation of the models, three statistical indices are calculated: Coefficient of Distribution (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). For spatial and temporal granularity analysis, R^2 and RMSE are used as metrics to compare the scenarios explained above.

5. Results and Discussion

In this section, the results of each model in all temporal and spatial cases are presented and discussed.

MLR: Model Performance

The MLR is implemented for 3 different time intervals and 4 zones of which the data are available. The results for the model with 60-min time interval and for different zones and the floor level are presented here, which have the highest R^2 among all other cases predictions. Figure 4 shows the normal distribution of the residuals for the predictions with 60-min time resolution. Based on this figure, it can be inferred that the residuals obtained from the MLR predictions are not good enough and significant numbers of residuals are far from the zero.

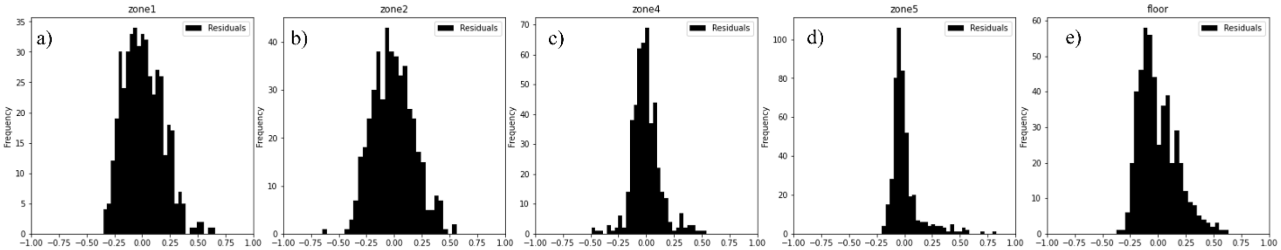


Figure 4. Residual plots for 60-min time resolution and different spatial level: a) Z_1, b) Z_2, c) Z_4, d) Z_5, and e) Floor (Z_1 is zone 1, and so on)

MLR: Regression Fitting

As mentioned above, there are three main time intervals for each modeling. The predictions of the floor level using MLR model will result in the following regressions:

1 minute interval:

$Y = 0.541 – 0.788 X_1 + 0.019 X_2 + 0.130 X_3 – 0.052 X_4 + 0.051 X_5$

15 minutes interval:

$Y = 0.529 – 0.918 X_1 + 0.141 X_2 + 0.143 X_3 – 0.044 X_4 + 0.154 X_5$

60 minutes interval:

$Y = 0.479 – 1.083 X_1 + 0.320 X_2 + 0.154 X_3 – 0.024 X_4 + 0.226 X_5$

In the above equations, X_1, X_2 and X_3 are respectively the floor temperature, humidity, and the illumination. X_4 and X_5 respectively show the outdoor temperature and radiation. The floor temperature has the highest impact on the regressions fitted. Table 3 compares the importance of each attribute based on different time resolutions of the dataset. It can be argued that the floor temperature has the highest impact on the modeling. In addition, the importance of other attributes for different time intervals is not similar and change. To improve the MLR predictions, feature selection is done with the number of selected features of 3; however, no major changes are observed.

Table 3: t-stars and p-values for different time resolutions of the floor level predictions

Attributes	**Time resolution**					
	1-min		15-min		60-min	
	t_star	p_value	t_star	p_value	t_star	p_value
Floor temperature	136.776	0.000	40.480	0.000	26.029	0.000
Outdoor Radiation	21.471	0.000	10.873	0.000	9.142	0.000
Floor illumination	20.076	0.000	5.598	0.000	3.295	0.001
Outdoor temperature	11.227	0.000	2.453	0.014	1.216	0.224
Floor relative humidity	2.712	0.000	4.919	0.000	5.224	0.000

MLR: Spatial & Temporal Granularity Analysis

The results of temporal and spatial analysis for MLR predictions are reported. In Figure 5, the vertical axis shows the spatial level and the horizontal axis presents temporal resolutions. In order to calculate the RMSE and R^2 as a single value representing all zones, the metrics are calculated by summing the predictions of all zones and are compared with the summation of actual values. Based on the results, the predictions using the MLR is not accurate enough (the maximum R^2 obtained after the feature selection was 68%); hence, the ANN method is implemented for better predictions.

ANN: Model Performance

The loss (or learning) curve is used to identify the overfitting issues related to the training and asses if the training and validation datasets are sufficiently representative. As shown in Figure 6, a point of stability is reached in the plot of training and validation loss, and there is a small gap between them, which confirms that no overfitting has been occurred. The number of epochs has been obtained with trial and error. After about 300 iterations, both loss values become constant and are in an acceptable range. Moreover, Figure 7 shows the distribution of residual errors for the hourly time resolution ANN predictions at the both zone and floor levels. The residual plots show normal distribution around zero, which confirm the reliability of predictions.

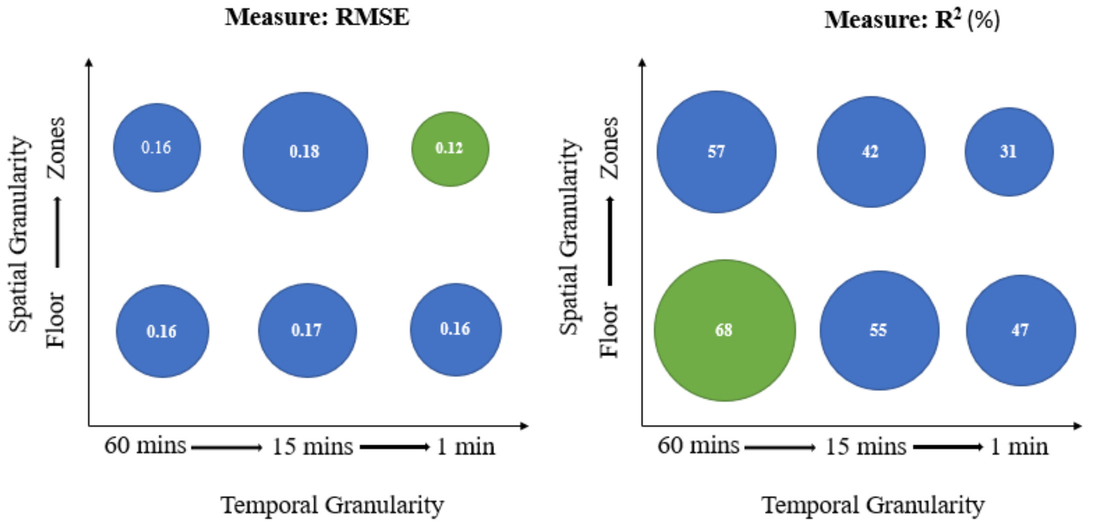


Figure 5. A comparison of zones and the floor level MLR predictions for different time resolutions based on RMSE and R^2 metrics

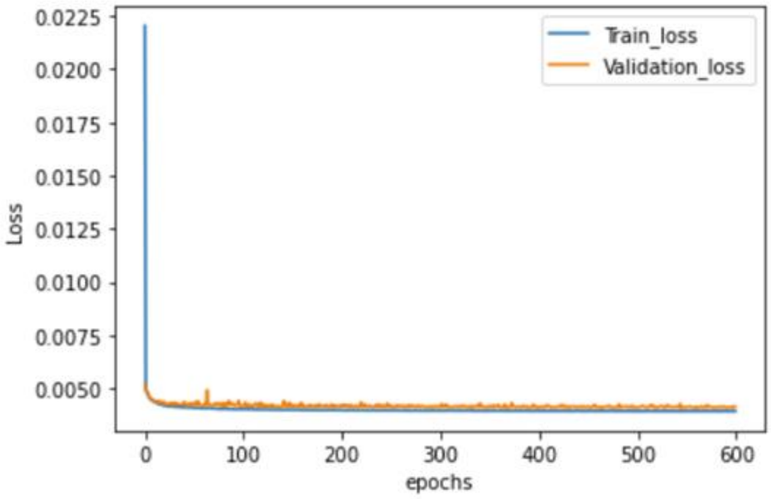


Figure 6. Loss plot for the ANN model

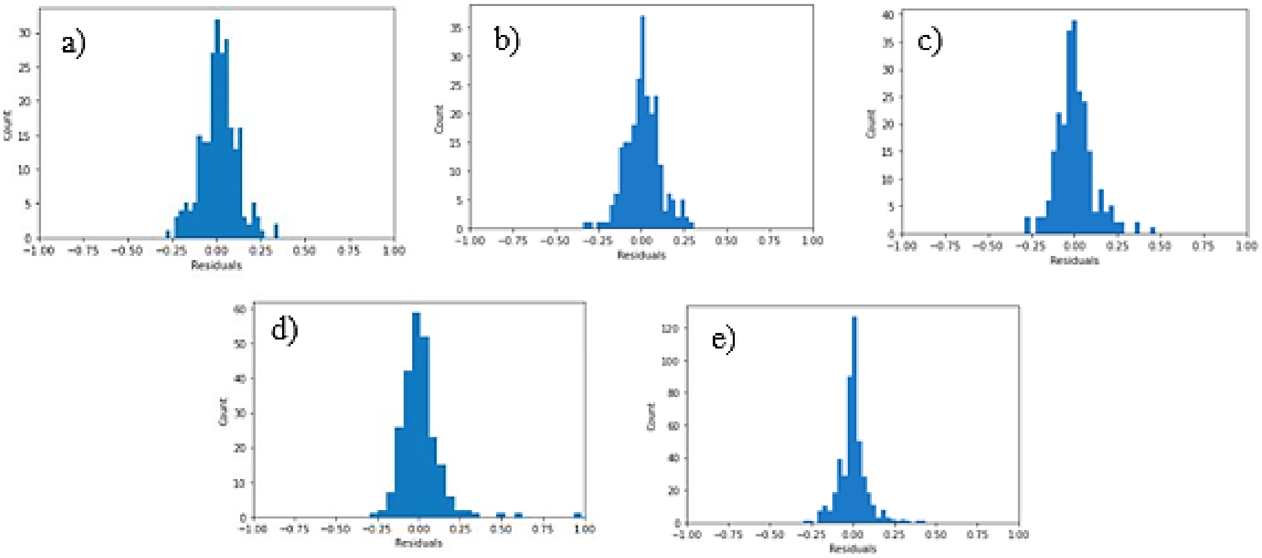


Figure 7. Residuals for 60-min time resolution: a) Z_1, b) Z_2, c) Z_4, d) Z_5, and e) Floor

ANN: Spatial & Temporal granularity analysis

Three measures (R^2, RMSE, and MAE) related to the ANN predictions are calculated for zone and floor level considering different time resolutions, as listed in Table 4. Based on literature and considering our project nature, these three metrics are in the acceptable range in our project.

Table 4: Evaluation parameters in each zone

Spatial Level	Time Interval (min)	R^2 (%)	RMSE (%)	MAE (%)
Z_1	1	65.1	6.3	3.6
	15	56.2	9	5.71
	60	62.5	10	7.9
Z_2	1	68.6	6.3	3.9
	15	71.9	10.6	7.1
	60	79.4	10.1	7.5
Z_4	1	61.5	8.5	6.1
	15	58.4	9.8	5.4
	60	63	10.5	7.6
Z_5	1	64.2	11	8.1
	15	67.7	13.1	9.3
	60	66	12.2	7.8
Floor	1	86.52	10.04	5.93
	15	81.62	12.64	7.46
	60	92.56	8.44	5.60

The R^2 and RMSE metrics are calculated for zone and floor level predictions at different time resolutions, as shown in Figure 8. Based on the figure, no significant changes are noticeable in RMSE; however, the R^2 values for floor level predictions are better, which means that the floor level granularity yields more accurate predictions.

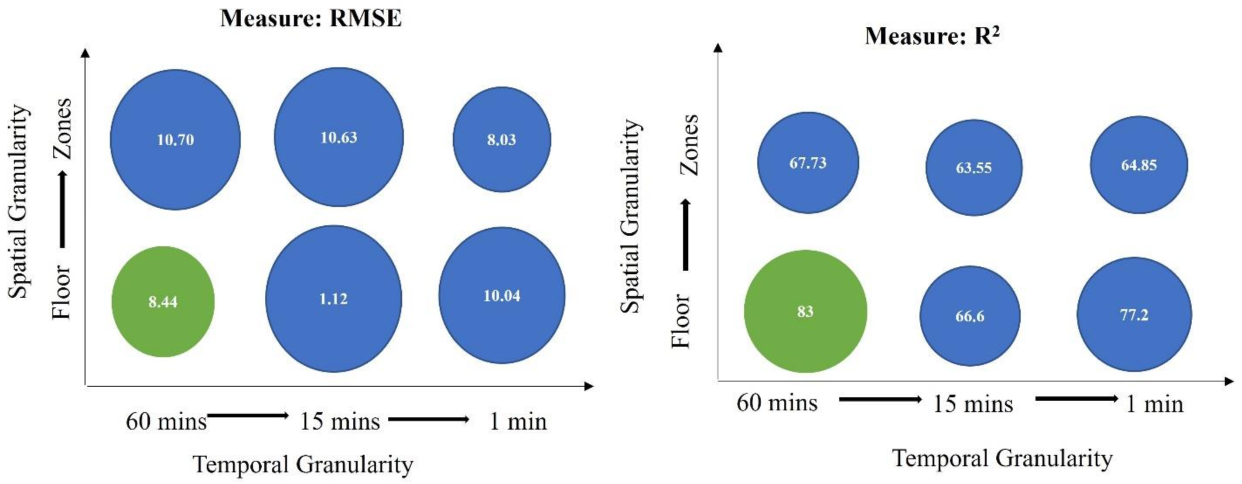


Figure 8. A comparison of zones and the floor level ANN predictions for different time resolutions based on RMSE and R^2 metrics

For further comparison, hourly regression results of ANN models in zone and floor level are shown in Figure 9. According to this Figure, the predicted values are almost evenly distributed for the floor level granularity, which means the ANN model works better using this spatial level.

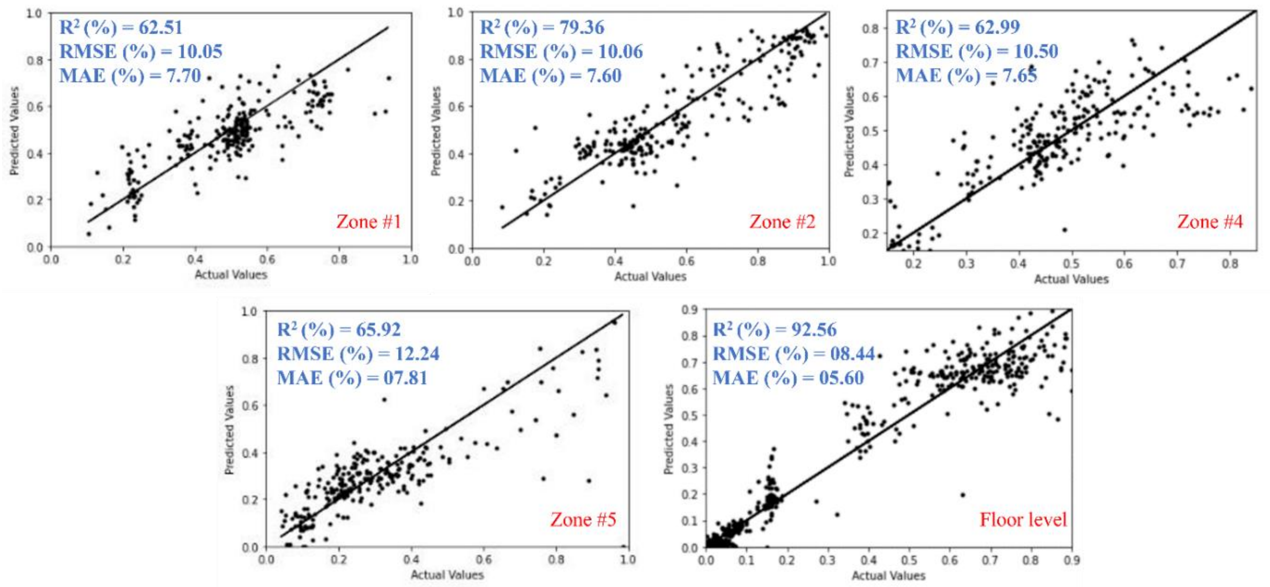


Figure 9. ANN predictions for zone and floor level with 60-min time resolution along with their performance measures

Best Prediction Results (hourly floor level)-ANN

A one-week load prediction based on the hourly time resolution of the floor level is compared with the actual data in Figure 10. Based on the results, the model exhibits good predictions for off-peak hours and is not able to forecast the load of peak hours as perfect as off-peak hours. In addition, Figure 11 shows the residual distribution for the best predictions.

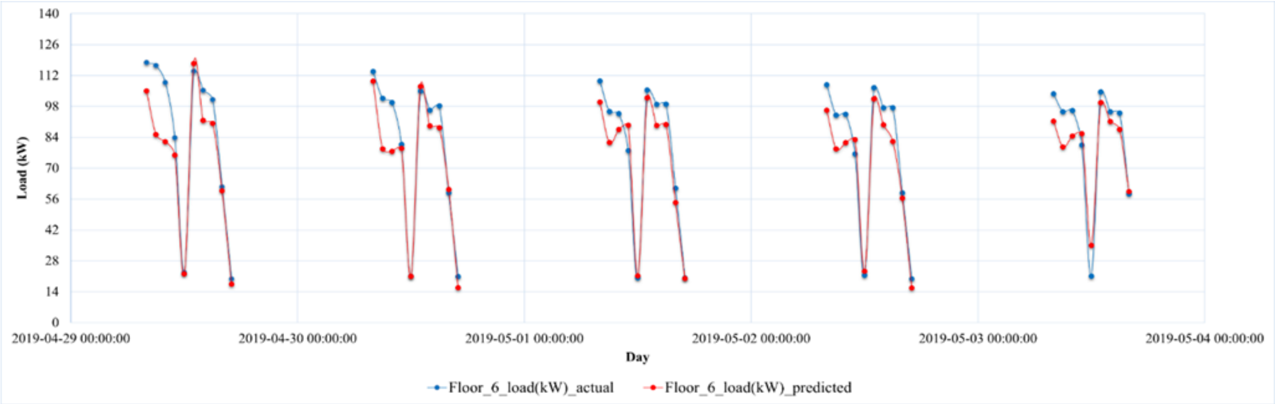


Figure 10. One week load prediction for the floor level with hourly resolution time

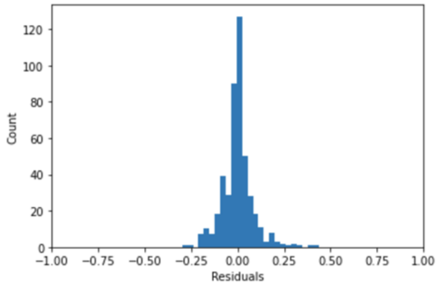


Figure 11. Residual distribution for the floor level with hourly time resolution

Based on Figure 11, the residual plot of the hourly predictions of the floor level display normal distribution around zero, which represents the well-predictions of off-peak hours. However, relatively high residual values in this figure are related to the peak hours, which shows that the ANN model does not show a good performance, and the prediction are lower than the actual values.

6. Concluding Remarks

In our study, all steps of the CRISP-DM process are followed to finally forecast the building load at zone and floor level. After understanding and visualizing the dataset, detailed steps are undertaken to cleanse the dataset, such as imputing the missing values and setting aside the holidays and non-working hours. Outdoor temperature and radiation are added to the dataset to forecast the building load more precisely. The main scope of our study is to examine the effect of spatial and temporal granularity on predictions of the models. Thus, for zone and floor level analysis, the simulations are carried out at 1-min, 15-min, and 60-min time intervals. The MLR and ANN models are trained since all the attributes are numerical types. As preliminary findings of the ANN predictions, the temporal resolution strongly influences the best number of neurons and hidden layers. Due to the nonlinear correlation between attributes, it is not surprising that achieving optimal results using MLR is not possible. Comparing MLR and ANN predictions, it can be inferred that the ANN predictions are more reliable; to the point that R-squared values are higher and RMSE and MAE errors are lower than MLR outcomes. The results of temporal granularity simulations reveal that the most accurate predictions occur at hourly time resolution for the floor level. This is due to the fact that the data for the floor level is smoother and more independent of noise than zone-level data. To sum up, increasing the resolution of the dataset does not necessarily generate more accurate results. As another consequence, the prediction of the ANN for off-peak loads is better than for peak loads. It was expected, especially at hourly simulations, due to the small number of examples at the floor level for training. For future works, comprehensive characteristics of the building, including the envelope, window to wall ratio, and shading along with the occupant's behavior should be taken into account in the training models.

7. References

[1] "Buildings – Topics - IEA." https://www.iea.org/topics/buildings (accessed Feb. 03, 2022).

[2] M. Pipattanasomporn *et al.*, "CU-BEMS, smart building electricity consumption and indoor environmental sensor datasets," *Scientific Data*, vol. 7, no. 1, pp. 1–14, 2020.

[3] L. Zhang *et al.*, "A review of machine learning in building load prediction," *Applied Energy*, vol. 285, p. 116452, 2021.

[4] J. Li, K. Panchabikesan, Z. Yu, F. Haghighat, M. el Mankibi, and D. Corgier, "Systematic data mining-based framework to discover potential energy waste patterns in residential buildings," *Energy and Buildings*, vol. 199, pp. 562–578, 2019.

[5] M. Cordeiro-Costas, D. Villanueva, and P. Eguía-Oller, "Optimization of the electrical demand of an existing building with storage management through machine learning techniques," *Applied Sciences*, vol. 11, no. 17, p. 7991, 2021.

[6] R. K. Jain, K. M. Smith, P. J. Culligan, and J. E. Taylor, "Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy," *Applied Energy*, vol. 123, pp. 168–178, 2014.

[7] "epwmap." https://www.ladybug.tools/epwmap/ (accessed Apr. 17, 2022).

[8] J. Hu *et al.*, "Thermal load prediction and operation optimization of office building with a zone-level artificial neural network and rule-based control," *Applied Energy*, vol. 300, p. 117429, 2021.

[9] D. C. Park, M. A. El-Sharkawi, R. J. Marks, L. E. Atlas, and M. J. Damborg, "Electric load forecasting using an artificial neural network," *IEEE transactions on Power Systems*, vol. 6, no. 2, pp. 442–449, 1991.