Energy Proceedings
Vol 54, 2025
ISSN 2004-2965

Predicting long-term building energy consumption using multiple feature clustering and machine learning: applications in Shanghai, China[#]

Yukai Jin, Ayyoob Sharifi*

Urban Environmental Science Lab (URBES), Graduate School of Innovation and Practice for Smart Society, Hiroshima University, Higashi-Hiroshima, 739-8529, Japan
(*Corresponding Author: sharifi@hiroshima-u.ac.jp)

ABSTRACT

As urbanization progresses, global building energy consumption is on the rise, emphasizing the need for a dependable energy consumption prediction model. This study presents a multi-stage machine learning approach comprising a clustering decomposition model (GMM), a prediction model (XGBoost), and an optimization model (PSO). Prior to clustering, the RF model evaluates the significance of various features influencing building energy consumption. GMM partitions the data into distinct clusters, while the PSO model fine-tunes the initial parameters of XGBoost. Validation is conducted using a dataset comprising 458,836 hourly records spanning three years from 20 office buildings in Shanghai, China. The average hourly energy consumption for all buildings is 79.2 kWh, but there is significant variation, with a standard deviation of 126.3 kWh. The prediction results indicate that the proposed model consistently achieves an R² exceeding 0.85 across diverse test sets, demonstrating robust accuracy and generalization capabilities. These findings offer valuable insights for future building design and energy management strategies.

Keywords: Energy consumption, Machine learning, Prediction, HVAC, Clustering

NONMENCLATURE

Abbreviations	
HVAC	Heating Ventilation Air Conditioning
GMM	Gaussian Mixture Model
XGBoost	eXtreme Gradient Boosting
PSO	Particle Swarm Optimization
GAN	Generative Adversarial Network
LightGBM	Light Gradient Boosting Machine
MLP	Multilayer Perceptron
RF	Random Forest
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
MAPE	Mean Absolute Percentage Error
R^2	Coefficient of Determination
Temp.	Temperature
DP Temp.	Dew Point Temperature
SC	Silhouette Coefficient
GBDT	Gradient Boosting Decision Tree
VRV	Varied Refrigerant Volume

1. INTRODUCTION

Amid global urbanization, buildings are playing an increasingly important role[1]. According to the China Building Energy Efficiency Annual Development Research Report[2], energy consumption in China's construction industry grew from nearly 400 million tce in 2004 to 1.35 billion tce in 2020, a growth rate exceeding 200%. Office buildings are a key component, with the average annual electricity consumption per unit area in Shanghai office buildings reaching 100.2 kWh/m² in 2021, and total carbon emissions exceeding 4 million tons. The Chinese government has set energy-saving targets[3], aiming to retrofit over 350 million square meters of existing buildings for energy efficiency by 2025. The area of newly constructed ultra-low energy buildings is expected to increase by more than 20 million square meters compared to 2023[4]. Therefore, accurately assessing and predicting

[#] This is a paper for the 16th International Conference on Applied Energy (ICAE2024), Sep. 1-5, 2024, Niigata, Japan.

the energy consumption of buildings, especially office buildings, is crucial for achieving energy-saving and low-carbon goals[5].

In recent years, energy consumption prediction models have become an essential component of Building Energy Management Systems[6]. Predicting energy consumption provides building managers with timely and intuitive insights into building energy use and simplifies the calculation of total electricity consumption and related billing costs, enabling them to implement effective energy-saving measures[7]. Currently, energy consumption prediction models are categorized into three types: the first type uses engineering methods based on physical and thermodynamic laws. For example, Tian et al.[8] proposed a hybrid physical model combining Generative Adversarial Network (GAN) and EnergyPlus to predict the 24-hour power demand of buildings, achieving an error rate of 5%. However, despite significantly improving computation speed compared to traditional EnergyPlus, the time for a single computation still exceeds two hours. The second type is statistical methods. Zeng et al.[9] used Gaussian Process Regression to predict the energy consumption of office buildings, achieving balanced prediction accuracy but with a relatively high error rate of about 15%. The third type is machine learning methods, including typical methods like XGBoost[10], Light Gradient Boosting Machine (LightGBM)[11], and Multilayer Perceptron (MLP)[12]. Compared to the first two methods, machine learning methods offer higher prediction accuracy and faster computation speeds when predicting long-term building energy consumption across large datasets, which has led to their widespread use in recent years[6]. Additionally, clustering models, as a type of unsupervised learning in machine learning, have been proven to enhance the accuracy of prediction models and are widely used in large-scale dataset predictions[13].

To investigate the energy consumption of office buildings, this study will concentrate on hourly predictions of long-term energy usage in such buildings. It will employ a combination of models including the feature selection model Random Forest (RF), clustering model GMM, prediction model XGBoost, and optimization model PSO to build a comprehensive machine learning framework. Subsequently, a training dataset containing over 400,000 energy consumption data points will be utilized for validation, examining the hourly energy consumption variations of three buildings with three distinct HVAC systems over a three-year period. The objectives of this study are as follows:

i). To clarify the influence of factors such as weather elements, time elements, building area, and HVAC mode on building energy consumption and identify the most significant influencing factors.

ii). To use Gaussian Mixture Clustering Model to classify energy consumption data with different features.

iii). To establish a GMM-PSO-XGBoost model, validate it in a real dataset, calculate accuracy and computation time, and compare it with base models and similar models.

2. METHODOLOGY

2.1 Random forest

RF is an ensemble learning algorithm that improves model performance and stability by constructing multiple decision trees and integrating their results, and is often used for feature selection[14]. During training, Random Forest builds multiple decision trees, each from different subsets of the training data and random subsets of features, to enhance model generalization and prevent overfitting. It measures feature importance by calculating the information gain or Gini impurity reduction brought by each feature during node splitting in each decision tree. The cumulative impurity difference before and after splitting is added to the feature's importance score. The scores from all decision trees are accumulated, averaged, and used to rank the features for selection.

2.2 GMM

The GMM is a probabilistic model used to represent a mixture distribution composed of multiple Gaussian distributions[15]. GMM assumes that data is generated by a linear combination of multiple Gaussian distributions, where each Gaussian distribution represents a cluster in the data. The probability of each data point belonging to a particular cluster is determined by the Gaussian distribution of that cluster. The goal of GMM is to estimate the parameters of these Gaussian distributions by maximizing the likelihood function of the data. A d-dimensional Gaussian distribution is described by a mean vector μ and a covariance matrix Σ, with its probability density function given by：

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \tag{1}$$

Where: μ represents the mean vector of the k-th Gaussian distribution, serving as the center of the cluster. Σ denotes the covariance matrix of the k-th Gaussian distribution, determining the shape and orientation of the cluster.

2.3 XGBoost

XGBoost is an ensemble learning method that adopts the gradient boosting tree model. It constructs the ensemble model by combining multiple decision trees[10]. Each tree is trained based on the residuals of the previous tree to gradually reduce the model's error. The model's structure resembles that of traditional decision trees, including nodes, splitting rules, and leaf nodes. Each leaf node has a prediction value representing the predicted output for the corresponding sample, which is determined by minimizing the loss function and regularization term. During the prediction phase, for regression problems, the predicted result is the weighted sum of the leaf node values of all trees. In each iteration, XGBoost attempts to minimize the objective function L(θ), defined as:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

Where l is the loss function, measuring the difference between the predicted value $\hat{y}_i$ and the true value y_i. $\Omega(f_k)$ is the regularization term used to control the model's complexity and prevent overfitting.

2.4 PSO

PSO is a metaheuristics optimization algorithm based on the foraging behavior of birds[16]. PSO can be broken down into three main components: position update, velocity update, and fitness evaluation. In the algorithm, each particle (solution) in the search space has a position and velocity. The position represents the current solution, while the velocity indicates the direction and step size of the solution's search. Fitness evaluation involves computing the objective function value based on the solution's position to guide the search process.

2.5 Framework

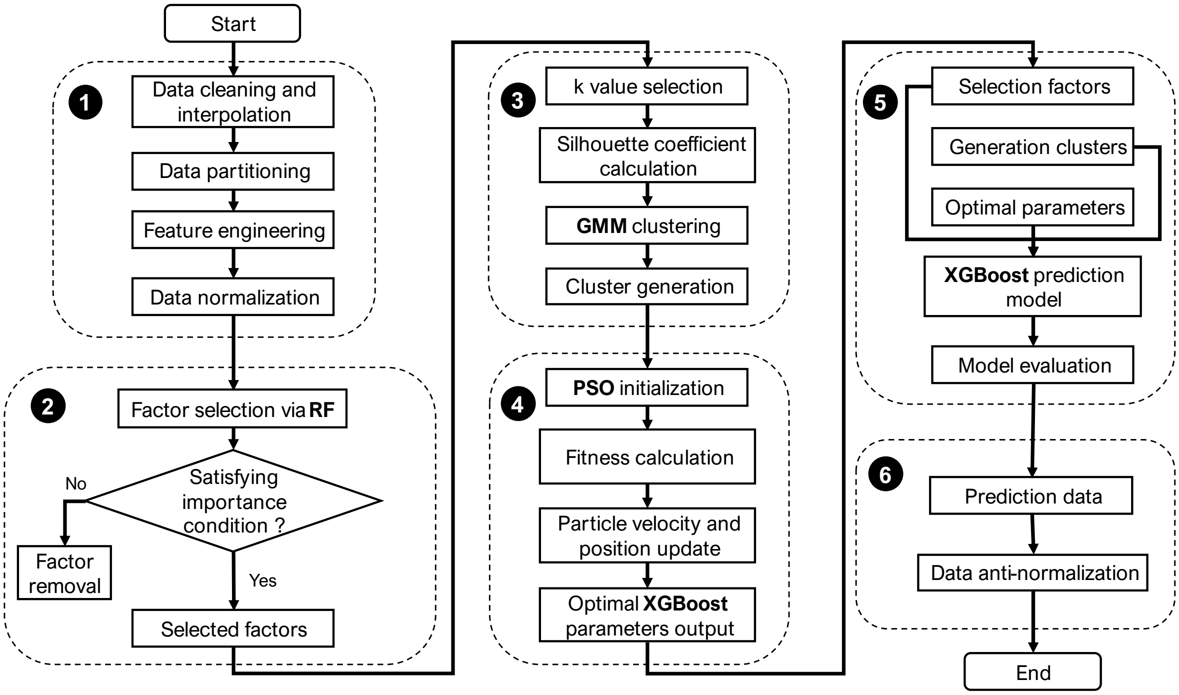

Fig. 1 The experimental framework of this study. ①is data preprocessing, ② is feature selection, ③ is clustering, ④ is parameter optimization, ⑤ is prediction, and ⑥ is model evaluation

2.5.1 Data Preprocessing

The framework of this work is shown in Fig. 1. The data preprocessing stage involves data cleaning, which entails removing data points with missing values. In the data partitioning section, we number 20 buildings from 0 to 19. Buildings 0 (23077), 2 (20660), and 9 (23506) are designated as the test set, representing centralized all-air systems, fan coil units with fresh air systems, and split air conditioners or Varied Refrigerant Volume (VRV) local unit systems, respectively. The remaining data (391593) is used as the training set.

2.5.2 Model Evaluation

To reasonably evaluate the overall performance of the model, this study constructs four metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Coefficient of determination (R²). MAE better reflects the actual situation of the observed error, while MSE measures the deviation between observed values and true values[17]. The research purposes of these two metrics differ, and their formulas are defined as follows::

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

$$MAPE = \sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot \frac{100}{n} \tag{5}$$

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_{ti} - \bar{y}_t)\cdot(y_{pi} - \bar{y}_p)\right)^2}{\sum_{i=1}^{n}(y_{ti} - \bar{y}_t)^2 \cdot \sum_{i=1}^{n}(y_{pi} - \bar{y}_p)^2} \tag{6}$$

Where y_i represents the actual value, $\hat{y}_i$ represents the predicted value, and n is the number of observations. y_{ti} represents the true value, y_{pi} represents the predicted value, y_{p} and y_{t} are the mean values of the predictions and true observations respectively, and *n* is the number of observations.

3. RESULTS AND DISCUSSION

3.1 Data description

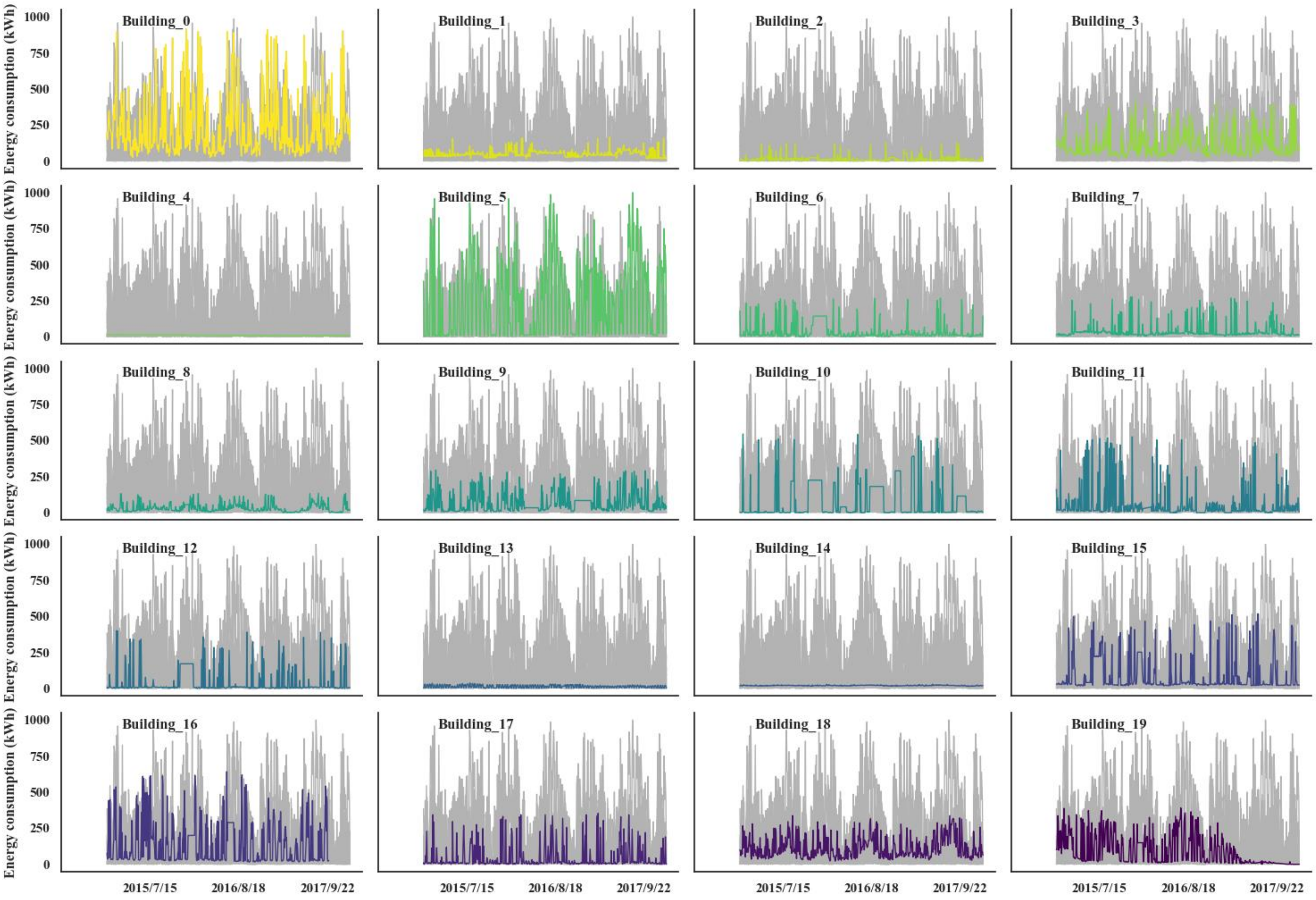

Fig. 2 Hourly energy consumption distribution of 20 office buildings in Shanghai, China from 2015/1/1 to 2017/12/31. The colored curves represent the current building's energy consumption, while the gray curves indicate the combinition of the 20 buildings.

The data used in this study comes from 20 office buildings in Shanghai, China. The data spans from January 1, 2015, to December 31, 2017, with hourly precision. After data preprocessing, there are a total of 458,836 records. The hourly energy consumption trends As shown in the Fig. 2, the energy consumption of most buildings exhibits clear periodic patterns. For example, Building 0, Building 5, and Building 17 have low energy consumption levels in spring and autumn, while consumption spikes sharply in summer and winter.Additionally, Building 0 and Building 5 have the largest fluctuations in energy consumption, with peak values exceeding 900 kWh. Buildings 4, 13, and 14 have smaller fluctuations. Table 1 displays the building area, number of floors, and HVAC type for all buildings.

Table 1 Total area and HVAC types of the 20 office buildings

Building ID	Area (m^2)	HVACType
0	77563.54	Centralized all-air system
1	34275.6	Centralized all-air system
2	38232.3	Fan coil units & fresh air system
3	47830	Fan coil units & fresh air system
4	50091	Others
5	31809	Centralized all-air system
6	34335	Fan coil units & fresh air system
7	35000	Fan coil units & fresh air system
8	27000	Fan coil units & fresh air system
9	21627	Split air conditioning or VRV local unit system
10	49838	Fan coil units & fresh air system
11	87251.53	Fan coil units & fresh air system
12	42266	Fan coil units & fresh air system
13	43000	Fan coil units & fresh air system
14	38000	Fan coil units & fresh air system
15	52800	Fan coil units & fresh air system
16	30000	Fan coil units & fresh air system
17	50000	Fan coil units & fresh air system
18	50928	Fan coil units & fresh air system
19	35000	Centralized all-air system

3.2 Statistical analysis

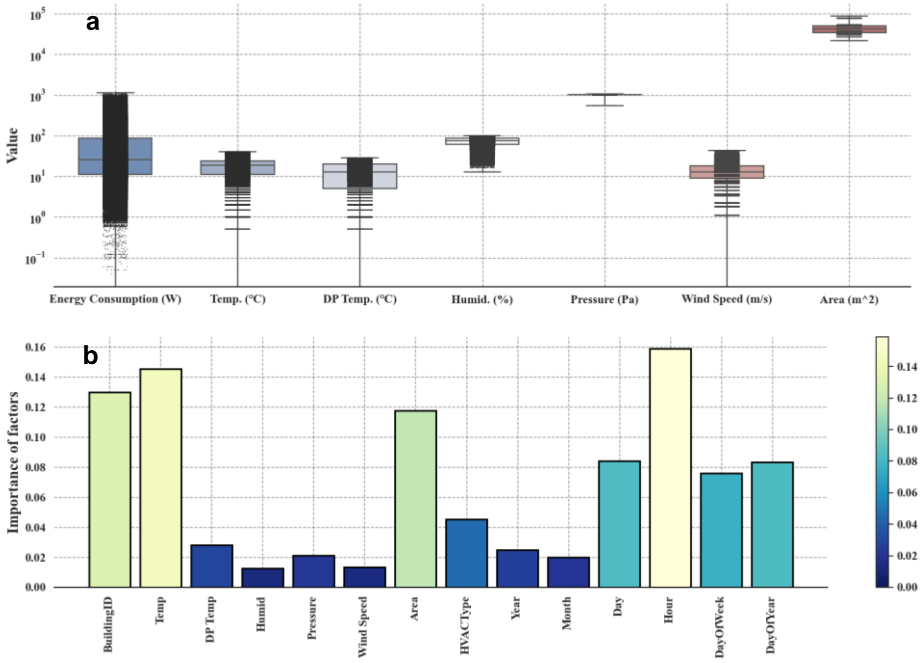

Fig. 3 (a) Box plot distribution of building energy consumption, total area, and meteorological factors for the 20 office buildings, with the median indicated by the line inside the box; (b)Importance of each feature with building energy consumption, with lighter colors indicating a stronger importance.

We conducted a statistical analysis on seven factors: energy consumption, temperature, dew point temperature, humidity, air pressure, wind speed, and building area. The results are shown in Table 1. The average values of these seven factors are 79.2 kWh, 17.7°C, 12.2°C, 72.5%, 1016.3 kPa, 13.2 m/s, and 43,696.5 m², respectively. As shown in Fig 3 (a) and Table 2, the energy consumption data has a wide range of distribution, resulting in a large standard deviation. In contrast, air pressure has a small standard deviation, reflecting its more concentrated distribution in the figure.

Table 2 Statistical analysis of building energy consumption, total area, and meteorological factors for the 20 office buildings.

Factors	mean	std	50%
Energy consumption (kWh)	79.2	126.3	25.4
Temprature (˚C)	17.7	8.5	18.5
DP Temp (˚C)	12.2	9.2	13.0
Humidity (%)	72.5	17.5	75.5
Pressure (kPa)	1016.3	9.4	1016.0
Wind Speed (m/s)	13.2	6.4	12.6
Area (m^2)	43696.5	15357.2	42266.0

3.3 Feature selection

The Gini coefficient of the RF model was used to evaluate the importance between different features and building energy consumption. A larger value indicates a stronger importance with energy consumption. As shown in Fig. 3 (b), temperature and the hour of the day are the most importance with building energy consumption, with values of 0.145 and 0.159, respectively. Therefore, these two features were selected for the next stage of clustering.

3.4 Clustering

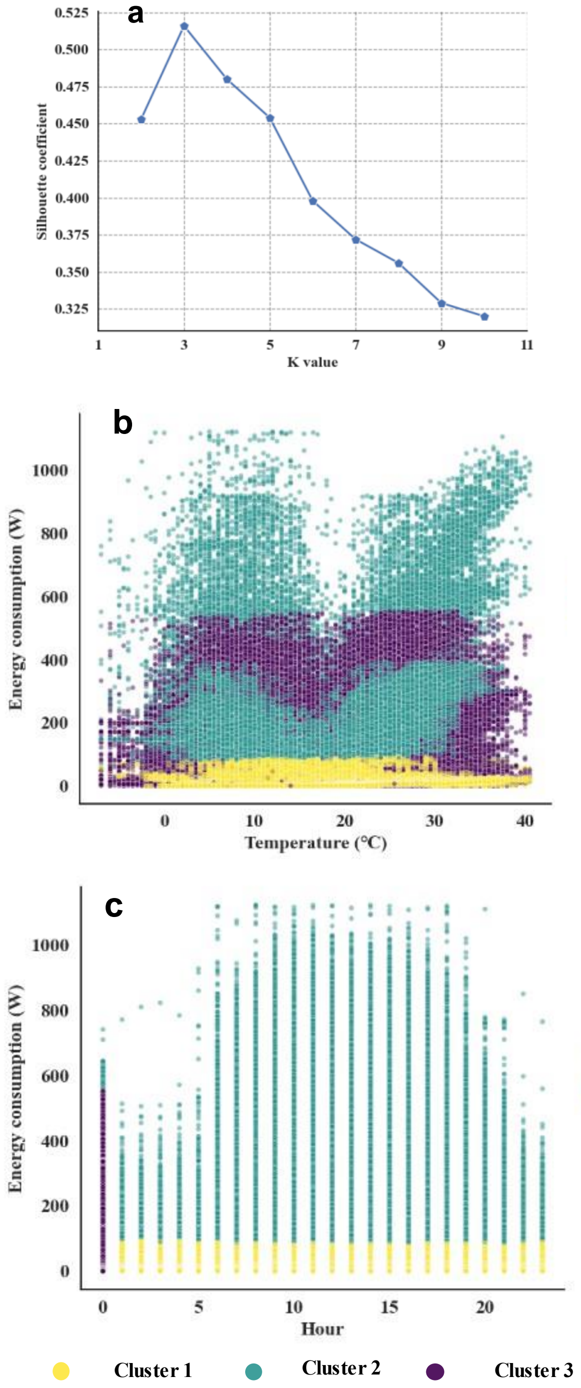

Fig. 4 (a) Silhouette coefficient for different k values; (b) results of GMM for energy consumption and temprature; (c) results of GMM for energy consumption and Hour.

After feature selection, the original dataset was transformed into an N × 3 sample space. A three-dimensional GMM was used to divide it into several similar clusters. The silhouette coefficient (SC) was used to select the optimal value of k[13]. Fig. 4 (a) shows the different SC values for different k values, ranging from 2 to 10. The SC value is highest when k = 3 and then gradually decreases, so the final k value is 3. Consequently, the data was divided into three clusters. As shown in Fig. 4 (b) and 4 (c), the first cluster has lower energy consumption values, and the third cluster's data is concentrated around hour 0 with moderate energy consumption values.

3.5 Prediction

In this study, four base models—XGBoost, LightGBM, Gradient Boosting Decision Tree (GBDT), and MLP—along with three combination models—GMM-XGBoost, GMM-LightGBM, and GMM-PSO-LightGBM—were used as comparison models to verify the accuracy of the proposed GMM-PSO-XGBoost model. The experimental results of the prediction models were analyzed using MAPE, MAE, RMSE, and R^2. The energy consumption prediction results for Building 0, Building 2, and Building 9 are shown in Table 3. The computations were implemented in a Python 3.7 environment on a computer with an Intel i5-9400 CPU and 8 GB of RAM.

As shown in Table 3, the proposed GMM-PSO-XGBoost model outperformed the other models in terms of error and accuracy when predicting the energy consumption of the three buildings with different HVAC systems. This indicates that the proposed model has better predictive performance and generalization ability. Additionally, models optimized with PSO showed a significant improvement in accuracy compared to the base models, suggesting that the optimal parameters selected by the PSO algorithm help enhance the predictive accuracy of XGBoost and LightGBM models. GMM-XGBoost's predictive accuracy was better than that of XGBoost alone, indicating that adding GMM can improve model accuracy. After clustering, the data within each group had higher similarity, which can improve the predictive accuracy of the original model.

Taking Building 0 as an example, as shown in Fig. 5, the proposed model had the highest R^2 value of 0.87 among all models, followed by GMM-PSO-LightGBM and GMM-XGBoost, both with an R^2 of 0.82. MLP had the lowest R^2 value at 0.19. Among all models, LightGBM had the shortest runtime at 128.26 seconds, with XGBoost close behind at 130.21 seconds. MLP had the longest runtime, taking over 600 seconds per run. This suggests that XGBoost and LightGBM are more suitable for running on large datasets compared to MLP.

Table 3 Predict results of different models for building 0, building 2 and building 9

BuildingID	Model	MAE (kWh)	RMSE (kWh)	MAPE (%)	R^2	Time (s)
0	XGBoost	65.14	97.03	0.41	0.77	128.26
	LingtGBM	73.70	108.90	0.46	0.70	114.3
	GBDT	105.72	151.29	0.67	0.43	286.22
	MLP	132.26	180.56	0.86	0.19	673.89
	GMM-XGBoost	57.54	85.63	0.36	0.82	149.9
	GMM-LightGBM	65.82	96.80	0.42	0.77	137.24
	Proposed model	51.27	76.02	0.32	0.86	209.03
	GMM-PSO-LightGBM	57.74	84.97	0.36	0.82	167.19
2	XGBoost	5.43	9.94	0.68	0.72	121.19
	LingtGBM	6.10	11.25	0.79	0.64	109.42
	GBDT	8.07	14.76	1.00	0.38	269.92
	MLP	9.84	17.32	1.25	0.14	601.86
	GMM-XGBoost	4.42	8.07	0.57	0.81	149.14
	GMM-LightGBM	5.44	10.29	0.70	0.70	127.9
	Proposed model	3.40	5.96	0.45	0.90	200.34
	GMM-PSO-LightGBM	4.81	9.13	0.61	0.76	164.41

9	XGBoost	15.24	22.67	0.42	0.89	140.23
	LingtGBM	18.63	27.52	0.51	0.84	123.73
	GBDT	28.35	40.16	0.77	0.66	278.76
	MLP	42.34	57.68	1.26	0.30	671.69
	GMM-XGBoost	11.94	17.70	0.34	0.93	167.11
	GMM-LightGBM	15.70	23.39	0.42	0.89	133.74
	Proposed model	8.93	13.11	0.27	0.96	201.46
	GMM-PSO-LightGBM	13.10	19.65	0.35	0.92	165.87

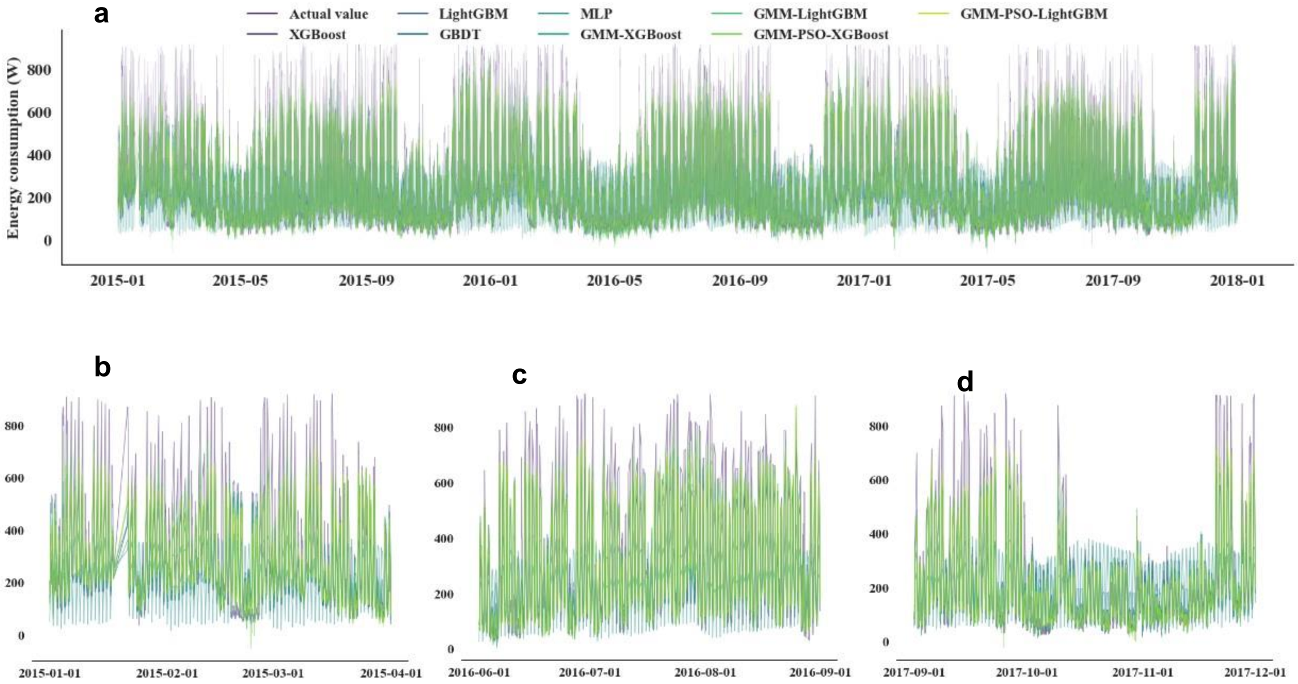

Fig. 5 (a) Prediction results for Building 0 by the proposed model and comparison models; (b) Prediction results distribution from January to March 2015; (c) Prediction results distribution from June to September 2016; (d) Prediction results distribution from September to December 2017.

4. CONCLUSION

This study proposes a long-term energy consumption prediction model suitable for large-scale data, GMM-PSO-XGBoost. We used hourly energy consumption data from 17 office buildings over three years, totaling 391593 records, as the training set to predict the hourly energy consumption over three years for three buildings with different HVAC systems. The proposed model performed well, with R² values exceeding 0.85 on all three test sets. The main conclusions of this study are as follows:

i) The Gini coefficient method of the Random Forest showed that temperature and the hour of the day are the two features most strongly importance with building energy consumption, with Gini coefficient values of 0.145 and 0.159, respectively.

ii) Using a three-dimensional Gaussian Mixture Model (GMM), all data were divided into three clusters. Predictions made using the clustered data had higher accuracy than those made using unclustered data.

iii) Compared with four base models (XGBoost, LightGBM, GBDT, MLP) and three hybrid models (GMM-XGBoost, GMM-LightGBM, GMM-PSO-LightGBM), the proposed model achieved the highest accuracy while keeping the runtime under four minutes.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support of SmaSo-X Challenge Project Graduate Students Research Grant, Hiroshima University.

DECLARATION OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. All authors read and approved the final manuscript.

REFERENCE

[1] Debnath K B, Mourshed M.Forecasting methods in energy planning models[J].Renewable and Sustainable Energy Reviews,2018, 88: 297-325.

[2] Tsinghua U C B E E. Annual Development Research Report 2022: China Architecture & Building Press,2022.

[3] Ma M, Ma X, Cai W, Cai W.Low carbon roadmap of residential building sector in China: Historical mitigation and prospective peak[J].Applied Energy,2020, 273: 115247.

[4] Zhang L L, Zhang J R, Ren P P, et al.Analysis of energy consumption prediction for office buildings based on GA-BP and BP algorithm[J].Case Studies in Thermal Engineering,2023, 50.

[5] Jiang B, Li Y, Rezgui Y, et al.Multi-source domain generalization deep neural network model for predicting energy consumption in multiple office buildings[J].Energy,2024, 299.

[6] Somu N, Raman M R G, Ramamritham K.A deep learning framework for building energy consumption forecast[J].Renewable and Sustainable Energy Reviews,2021, 137: 110591.

[7] Chou J-S, Nguyen H-M.Simulating long-term energy consumption prediction in campus buildings through enhanced data augmentation and metaheuristic-optimized artificial intelligence[J].Energy and Buildings,2024, 312: 114191.

[8] Tian C L, Ye Y Y, Lou Y L, et al.Daily power demand prediction for buildings at a large scale using a hybrid of physics-based model and generative adversarial network[J].Building Simulation,2022, 15 (9): 1685-1701.

[9] Zeng A R, Ho H D, Yu Y.Prediction of building electricity usage using Gaussian Process Regression[J].Journal of Building Engineering,2020, 28.

[10] Pan H Z, Wu C J.Bayesian optimization plus XGBoost based life cycle carbon emission prediction for residential buildings-An example from Chengdu, China[J].Building Simulation,2023, 16 (8): 1451-1466.

[11] Yang G, Du S H, Duan Q L, Su J.A Novel Data-Driven Method for Medium-Term Power Consumption Forecasting Based on Transformer-LightGBM[J].Mobile Information Systems,2022, 2022.

[12] Lee H Y A, Kim D, Gu J H.Prediction of Food Factory Energy Consumption Using MLP and SVR Algorithms[J].Energies,2023, 16 (3).

[13] Liu H, Long Z, Duan Z, Shi H.A New Model Using Multiple Feature Clustering and Neural Networks for Forecasting Hourly PM2.5 Concentrations, and Its Applications in China[J].Engineering,2020, 6 (8): 944-956.

[14] Niu D, Wang K, Wu J, et al.Can China achieve its 2030 carbon emissions commitment? Scenario analysis based on an improved general regression neural network[J].Journal of Cleaner Production,2020, 243: 118558.

[15] Kock L, Klein N, Nott D J.Variational inference and sparsity in high-dimensional deep Gaussian mixture models[J].Statistics and Computing,2022, 32 (5).

[16] Shami T M, El-Saleh A A, Alswaitti M, et al.Particle Swarm Optimization: A Comprehensive Survey[J].Ieee Access,2022, 10: 10031-10061.

[17] Cabaneros S M, Calautit J K, Hughes B R.A review of artificial neural network models for ambient air pollution prediction[J].Environmental Modelling & Software,2019, 119: 285-304.