Load prediction with an improved feature selection method for building energy management of an office park

Caiyu Li[1], *Yang* Geng[1*], *Hao* Tang[1], *Dejan* Mumovic[2], *Ivan* Korolija[2], *Zihui* Lv[3], *Tianxiang* Cao[3], *Xiaobin* Gu[4], and *Borong* Lin[1†]

[1]Department of Building Science, School of Architecture, Tsinghua University, Beijing, China
[2]University College London, Institute of Environmental Design and Engineering, United Kingdom
[3]Alibaba Cloud Intelligence, Alibaba Group, Hangzhou, China
[4]Corp Admin & Workspace Management, Alibaba Group, Hangzhou, China

Abstract. Load prediction plays a significant role in building energy management. An accurate HVAC load prediction model highly depends on the feature selection and the quality of training data. In previous work on load prediction, the input features are majorly manually selected by expertise, which is relatively subjective and lacks theoretical supports. Using the real building operational data collected from an office park located in Hangzhou, this paper developed a short-term cooling load prediction model, in which the input features are selected based on an analysis on the heat transfer process. Combined with qualitative analysis of the real data, several features such as outdoor air enthalpy and indoor black-bulb temperatures from different orientations are introduced into the model. The proposed model was then applied to the HVAC control system of the office park. Compared to the load prediction model with commonly used features, the proposed model reduced CRVMSE by 21% and MAPE by 30% during the operation period of the system. Furthermore, the impacts of training dataset size and prediction time range on model's accuracy and training time were discussed.

1 Introduction

Building energy accounts for a large part of total energy consumption. The energy consumption of public buildings has been almost doubled during the past decade in China, with its total amount and increase rate exceeding other building types [1]. In order to ensure a healthy, productive and energy-efficient office environment, the air-conditioning system is widely and extensively used in office buildings, which makes HVAC energy consumption in office buildings account for over 40% of the total energy consumption [2]. Most of the HVAC control systems in office buildings adopts a feedback-control based operation approach, i.e. only taking action after indoor environment problems have already occurred (such as increasing the cooling supply when indoor temperature has exceeded the upper bound of the comfort range in summer), which leads to the mismatch between supply and demand of air-conditioning, causing unnecessary energy waste. HVAC load forecasting can help buildings predict and cope with the changing load in advance. This predictive-based control strategy can solve the mismatch problem, bring energy saving effect and improve thermal comfort.

Building HVAC load prediction models can be classified into two major types: physical models and data-driven models. Physical models require detailed and large amounts of building information, including shape of the building, envelope property, facilities' capability, and local climate information to simulate future loads based on thermodynamic principles, while data-driven models mainly rely on historical building operational data and meteorological data to develop machine learning models and predict future loads. With the advancement of building operation data acquisition methods, data-driven models become simpler and easier to use. Generally, it is necessary to extract several features from the vast amount of historical data as the inputs of load prediction model. On one hand, feature extraction can reduce the dimension of data and save computation cost; on the other hand, the features as the input of the model can directly affect the result of load prediction.

In existing studies, certain features are usually selected based on data quality and building operation characteristics, and then the final model inputs are generated through feature engineering techniques such as time sequence translation or statistical description. To select suitable features, there are major two ways: manual selection and selection based on data analysis. Manual selection of features greatly depends on domain knowledge and experience, and can include physical mechanism of building operational data. In different studies, the commonly used feature types include outdoor air temperature and humidity, time label, past load, etc. However, these features contain both the

* Corresponding author: gengy@tsinghua.edu.cn
† Corresponding author: linbr@tsinghua.edu.cn

influence factors of load such as outdoor air temperature, and the consequences caused by load changes, like historical loads. Using the consequences of historical load variation to predict load in the future can lead to the lack of mechanism basis. On the other hand, the methods based on data analysis focus on the characteristics of the data itself and uses statistical analysis or machine learning methods to automatically generate several features which are most relevant to the load. Common methods include partial autocorrelation coefficient (PACF) analysis [7] and the use of autoencoder [14]. This kind of methods usually can improve the accuracy of load prediction models, but it cannot reflect the physical principle, and as a result the model interpretability is weak. Moreover, when the data characteristics change (such as extreme weather), the selected features will no longer be applicable. An extended and more detailed list of the features used in previous work on load prediction is summarized in Table1.

Table 1. Commonly used features in load prediction in previous work

Reference	Year	Outdoor environment					Time label				Past load	Occupancy
		T	RH	DP	SR	WS	M	D	H	Holiday		
Gao Z. et al.[3]	2022	√	√		√						√	√
Rana M. et al.[4]	2022	√									√	
Liu R. et al. [5]	2022	√				√					√	
Kang X. et al. [6]	2022	√	√		√	√		√	√		√	
Ghenai C. et al. [7]	2022	√	√				√	√	√		√	
Ahmad.T et al.[8]	2020	√	√	√	√	√					√	
Wang Z. et al. [9]	2020	√	√					√	√	√		
Zhang C. et al. [10]	2020	√	√						√		√	
Fan C. et al.[11]	2019	√	√		√							√
Zhu G. et al.[12]	2018	√	√		√	√			√		√	
Ahmad M. et al. [13]	2017	√	√	√		√		√	√		√	√
Fan C. et al.[14]	2017	√	√				√	√	√	√	√	

*T: temperature, RH: relative humidity, DP: Dewpoint, SR: solar radiation, WS: wind speed, M: month of the year, D: day of the week, H: hour of the day.

In order to improve the interpretability and accuracy of HVAC load prediction, this paper proposes a feature selection method which combines the influence factors of load generation and real data characteristics. Then, the proposed approach was applied in a cooling load prediction case study to verify the effectiveness of this method. In addition, there is always a trade-off between the accuracy and the computation cost of the model in practical applications, so the impacts of the amount of training data and the prediction window on the effect of the model is further discussed. We conclude this paper by pointing out some potential future research directions.

2 Methodology

2.1 Feature selection

The method of feature selection can be divided into two steps. First, according to the influence factors of load generation and variation, the feature types that require further particularized are determined. Building HVAC load mainly includes the heat loss through building envelope, the cooling loads required to handle fresh air, and internal heat gains (occupancy and electrical appliances inside the building), thus it is affected by outdoor weather conditions, occupancy, the type and operation of devices, and the property of building envelope.

Among these factors, the building envelope seldom changes after the building is constructed. In office buildings, the operation schedule of appliances is usually highly consistent, depending on the schedule of occupancy, so the number of people in the room is a good indicator of internal heat gains. In addition, the operation schedule of office buildings is relatively fixed and cyclical, so time-related features such as the hour of the day or the day of the week can well represent building operation schedule. The heat loss through building envelope mainly comes from the heat transfer from outdoor environment and ambient solar radiation, so the outdoor air temperature and solar radiation intensity are also included in the feature types. The ventilation load is affected by the enthalpy difference between indoor and outdoor air. However, the indoor environment is usually stable in air-conditioned period, so the humidity and enthalpy of outdoor air can be important features.

After determining the preliminary feature types, the final input features of the load prediction model can be decided according to the number of buildings in the HVAC system, the data quality and the load characteristics by analysing the real data and identify the representativeness of each feature. Fig. 1 shows the flow chart of the feature selection process.

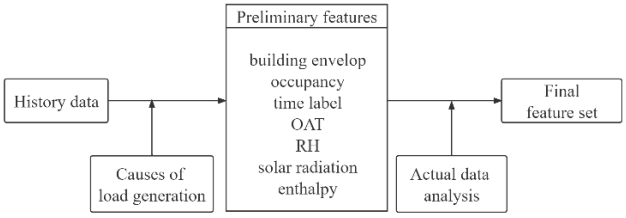


Fig. 1. Flow chart of proposed feature selection process for load prediction

2.2 Machine learning algorithm

In this paper, a case of cooling load prediction in an office park is used to validate the proposed feature selection method. A machine learning algorithm named Light Gradient Boosting Machine (LightGBM) is used in the modelling process.

LightGBM is a gradient boosting framework that uses tree-based learning algorithms, which is usually used in time-series data prediction. Different from commonly used tree models, LightGBM grows trees leaf-wise in each decision tree [documentation]. It will choose the leaf with max delta loss to grow instead of growing all leaves (as shown in Fig. 2), which makes the algorithm more accurate. It uses algorthms to find the best split point within each tree, which greatly reduces the time cost of model training while keeping the model accuracy basically unchanged.

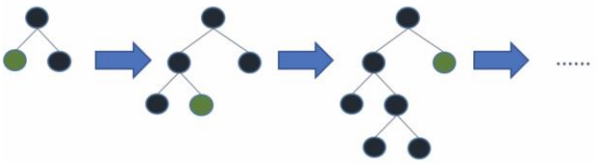


Fig. 2. Illustration of leaf-wise tree growth of LightGBM (credit to LightGBM documentation)

The parameters of the model are critical to model performance. One model can have several parameters, and for different models, the best combinations of parameters vary a lot. Hence, the parameters should be tuned in each scenario based on historical data. In this paper, an auto-tune tool named Optuna is used for searching the hyper-parameters in the model. It tries different hyper-parameter combinations within the specified parameter ranges to build and train models respectively, among which the model with the highest accuracy is selected as the best model, and the parameters of this best model are the optimal hyper-parameters.

2.3 Modelling

The method of online learning is used in this paper for model training. Different from offline training which has fixed training set and testing set, online learning trains the model using the most recent data (as shown in Fig. 3), which means that the model needs to be regularly retrained. Although this approach leads to higher computational costs, it allows the model to always use the latest data and therefore being able to respond to exceptional situations in time such as abrupt weather changes.

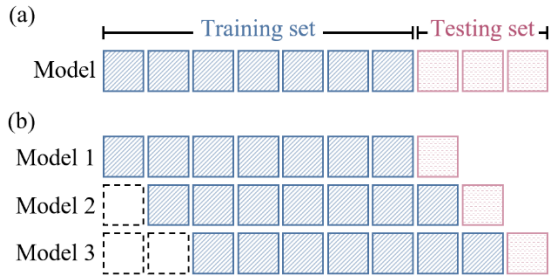


Fig. 3. Illustration of the training sets and test sets of traditional prediction model (a) and rolling prediction model (b)

2.4 Metrics

In the process of parameter tuning and model training, coefficient of the variation of the root mean square error (CVRMSE) and mean absolute percentage error (MAPE) are used as the evaluation metrics of model accuracy. The two errors are calculated as follows:

$$CVRMSE = \frac{\sqrt{\frac{1}{n}\cdot\Sigma(y_i-\widehat{y_i})^2}}{\overline{y}} \quad (1)$$

$$MAPE = \frac{1}{n}\cdot\Sigma\frac{|y_i-\widehat{y_i}|}{\widehat{y_i}} \quad (2)$$

The meanings of the variables are as follows:
n: the number of data
y_i: the i^{th} term of the predicted data
$\widehat{y_i}$: the i^{th} term of the ground truth data
$\overline{y}$: the average of the ground truth data

3 Case study

3.1 Case introduction

The selected case study building is an office park located in Hangzhou, Zhejiang Province, China. This office park has 8 office buildings and includes 4 HVAC systems, and this paper chooses one HVAC system for the case study. The selected HVAC system provides the cooling for building 6, 7, 8 in the park, with a total air conditioning area of more than 100,000 square meters. The building operational data collection platform and equipment management system of the office park are well-developed, which can obtain several types of historical data including equipment operation state, HVAC system parameters (such as chilled water flow rates and temperatures), occupant counts and outdoor weather parameters. The data sampling rate is 5 minutes. The building envelope is mainly glass curtain wall, so solar radiation has a significant impact on the building cooling load.

The data is collected between June 1, 2022 and August 10, 2022. The prediction target at each

prediction point is the cooling load in the next hour, and the history data is resampled every hour. Cooling load prediction was conducted out every hour, and the model was also retrained per hour. In every training process, the historical data within one week before the prediction point was used for training. The model randomly divides 70% of the training data into training set and the remaining 30% into the validation set. The test period was from August 2, 2022 to August 9, 2022.

3.2 Real data characteristics

Based on a theoretical analysis on the impact factors of cooling load, the preliminarily selected feature types include outdoor air temperature and humidity, outdoor air enthalpy, solar radiation, occupant counts and time-related features. On this basis, we analysed the collected building operational data to identify the load characteristics and determine the input of load prediction model. The analysis mainly focuses on the factors which are related to indoor environment, namely the cooling load, indoor black-bulb temperature, and the number of people.

The cooling load of the HVAC system can be calculated by the temperature difference and flow rate of chilled water. As shown in Fig. 4, most of the cooling load in a working day are non-zero between 9:00 and 21:00, corresponding to the occupancy schedule of the office building. On the contrary, on weekends, the cooling load of the system usually are non-zero at random time between 9:00 and 18:00, and meanwhile the cooling load is smaller than that of weekdays.

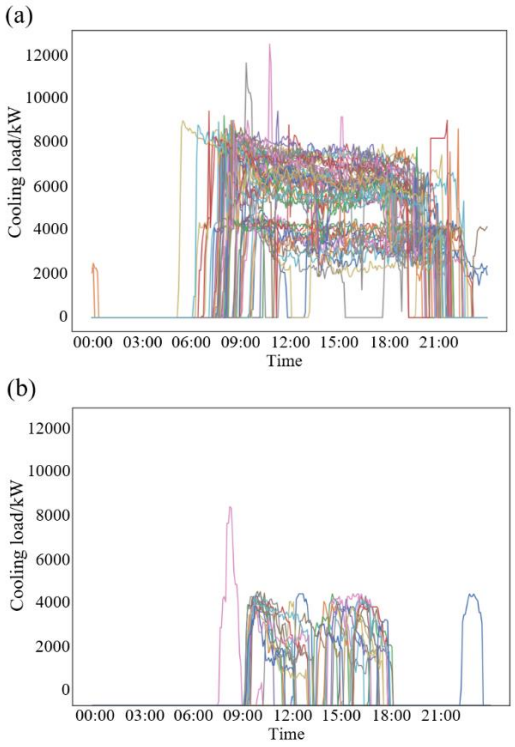


Fig. 4. Daily cooling load curves of the case buildings on weekdays (a) and weekends (b)

Since there is no solar radiometer installed in the target building, black-bulb temperatures are used to measure solar radiation intensity in this paper. In the three buildings of this case, five black-bulb thermometers were installed in different orientations on different floors in the buildings.

Fig. 5 shows how the black-bulb temperatures varies within a typical day on different orientations and floors. It indicates that with the influence of the sun's position, the black-bulb temperature peaks at 8:00-9:00 on the east side and at 16:00 on the west side. Moreover, affected by the shielding of surrounding buildings, the indoor black-bulb temperature of the second floor is lower than that of the fourth floor, that is, the solar radiation is less received. The black-bulb temperature on the south and north sides is similar to the variation trend of outdoor air temperature, which cannot indicate the impact of solar radiance.

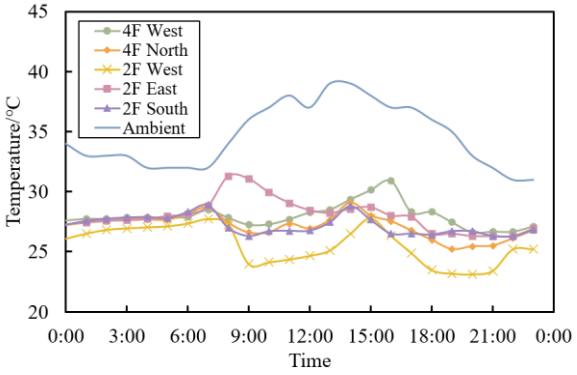
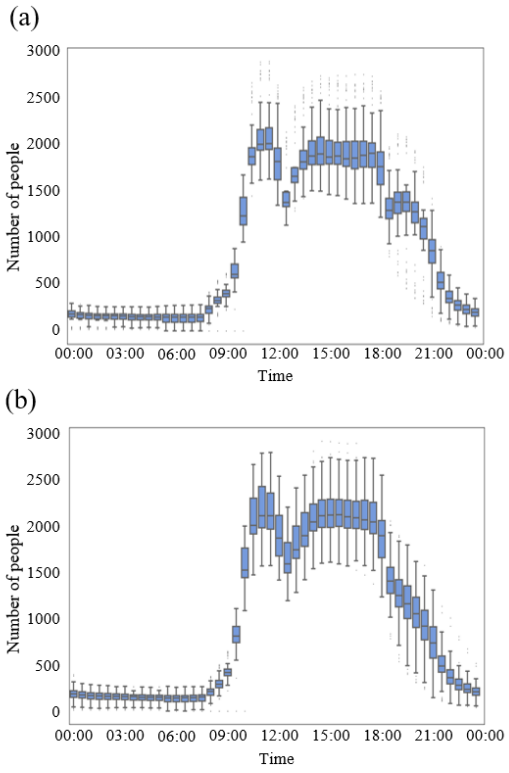


Fig. 5. Black-bulb temperature in different orientations and floors of the target buildings



(c)

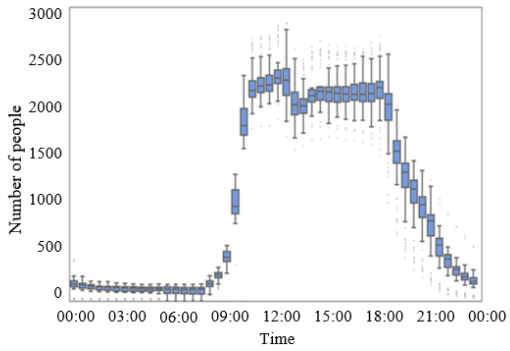


Fig.6. Boxplot of the hourly number of people in building No. 6 (a), building No. 7 (b) and building No. 8 (c)

In Fig. 6, the daily variation of occupant counts in the three target buildings is plotted. It shows that there is a small peak in the number of people in building No. 6 between 18:00 and 20:00, indicating that a few people may remain in the office after dinner in this building. There is also a gentle rise in the occupant count in building No.8 before lunch break and off-duty time. It can be discovered that the schedule patterns in the three buildings are different from each other.

4 Results

4.1 Feature generation

Based on the analysis of real data in the previous section, the feature types that will eventually be inputted to the load prediction model are identified.

Firstly, the load profile of the office park shows that cooling load has clear difference between weekdays and weekends, and between working periods and non-working periods. This indicates that day of week and hour of day should be important features for cooling load prediction. It is worth mentioning that some of the Chinese holidays fall in weekdays, so whether the day is holiday or not is regarded as another time-related feature.

Secondly, through analyzing the data of the black-bulb thermometers, several representative black-bulb temperatures were selected as input features of the model. The black-bulb temperatures on the west and east sides are more representative according to section 3.2, thus, they are selected as features of load prediction.

Besides, each of the three buildings has its own pattern of occupancy schedule. If only the total number of occupants is used as a single feature, the discrepancy between the buildings cannot be perceived, which has an influence on the load prediction accuracy. Therefore, instead of the total occupant counts in all three buildings, it is necessary to take the number of people in each building separately as input features.

The selected features of the cooling load prediction model in this case are identified as feature set 1.

Additionally, in order to validate the effectiveness of the proposed feature selection method, we conducted a literature review to identify the features which are most commonly used in existing studies for building HVAC load prediction. Those features are named as feature set 2, which was modeled and trained on the same dataset using the same method as feature set 1. The specific feature types in the two feature sets are shown in table 2.

Table 2. The features selected as model input in feature set 1 (proposed) and feature set 2 (common in literatures)

Features		**Feature set 1**	**Feature set 2**
Time label	Day of week	√	√
	Hour of day	√	√
	Whether is holiday	√	√
Black-bulb temperature	2nd floor on the west	√	
	2nd floor on the east	√	
	4th floor on the west	√	
Occupancy	Number of people in building No.6	√	√
	Number of people in building No.7	√	√
	Number of people in building No.8	√	√
	Total number of people in all three buildings	√	√
Outdoor environment	Outdoor air temperature	√	√
	Outdoor air relative humidity	√	√
	Outdoor air enthalpy	√	
Past load	Past 1 hour cooling load		√

4.2 Cooling load prediction

Fig. 7 shows the results of using two feature sets to train the model respectively and predict the cooling load during the test period. Because the historical cooling load in the previous hour is one input feature in the feature set 2, and the cooling load in the previously hour is usually close to the load in the next hour, thus adding the feature of previous load into the model can significantly improve the model's prediction ability especially when cooling load is high. However, in this circumstance, the model tends to place too much weight on this feature (load in the previous hour), ignoring the significance of other features, then the predicted load is always close to the load in the previous hour. When the cooling load changes rapidly, that is to say, the cooling load in the next hour is no longer similar to that in the past hour, the model fails to respond in time, resulting in large prediction errors. This can usually happen when the chiller was just turned on. However, similar problem doesn't appear in the prediction results given by the model using the feature set 1.

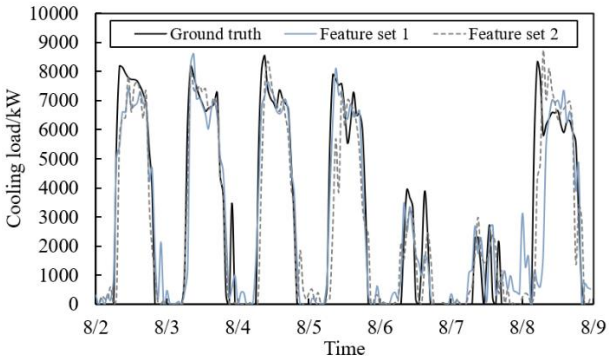


Fig. 7. Cooling load prediction results in test period using two feature sets

Since the load prediction model is used to inform the optimal operation of the HVAC system, in addition to the errors during the whole test period, the errors during the system operating period should also be concerned. The evaluation metrics during the whole test periods (CVRMSE-all, MAPE-all) and system operating periods (CVRMSE-on, MAPE-on) on the models are shown in table 4.

Table 4. Evaluation metrics on models with the two feature sets

	Feature set 1	**Feature set 2**
CVRMSE-all	0.392	0.415
MAPE-all	0.280	0.309
CVRMSE-on	0.220	0.280
MAPE-on	0.137	0.197

It can be found that the CVRMSE-all of model with the feature set 1 is 6% lower than that of model with the feature set 2, while the CVRMSE-on is 21% lower. The MAPE-all and MAPE-on also decreased by 9% and 30%. This indicates that the model using the proposed feature set outperforms the model using common features, especially for the load prediction in operation periods, demonstrating the rationality and effectiveness of the proposed feature selection method.

5 Discussion

The purpose of load prediction is to optimize the operation of HVAC system. Therefore, applicability in actual buildings is also an important criterion of load prediction model. Taking into account the data storage cost and computing performance of real building automation system, the load prediction model needs to ensure certain accuracy while minimizing the model computational time. Thus, it may be difficult to retrain the model every hour or to acquire all the historical data within the past week in practice. Therefore, this section discusses the influence of training set data size and prediction window on model's accuracy and computational cost.

5.1 Impact of training set size on accuracy

To ensure the accuracy of the model, the training set of the prediction model should include as many samples as possible. If the training set size is too small, the model will be more likely to encounter situations that is not covered in the testing set, which will affect the prediction accuracy. On the contrary, a too large training set will lead to the increase of data storage cost. Same as the prediction goal mentioned in Section3, the influence of the size of training set on the accuracy of the model can be revealed by using historical data of different sizes to predict the cooling load in the next hour.

In this section, seven load prediction models are trained using historical data ranging from 1 to 7 days before the prediction time step, and the test period and modelling method are the same as that described in Section3. To avoid the randomness of model training, each model was trained for five times, and the average prediction errors were compared.

The results show that the prediction errors decrease with the increase of training dataset size (as Fig. 8 shows). The prediction error will stabilize when the training dataset is adequately large. For different scenarios, the minimum required size of training dataset may vary, for our case, this required dataset size is 5 days.

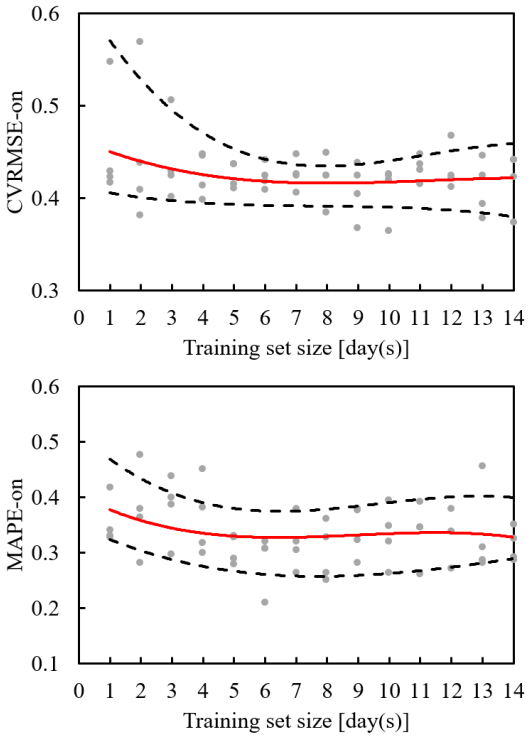


Fig. 8. Errors for models with different training set size

In terms of the computational demand, the training time does not increase a lot when the training dataset is less than seven days.

5.2 Impact of prediction window on accuracy and time cost

Prediction window means the duration of predicted cooling load between every model retraining. A longer

prediction window means a lower frequency of model retraining, which can significantly reduce the computational cost of model development. At the same time, it also means that the model cannot always utilize the latest data, so the longer prediction window may affect the prediction accuracy of the model. In order to probe into this impact, the same size of training set as in Section3 (7days) was used to predict future cooling load in different durations, and the test period remains the same. When the prediction window is more than one day, it has been found that model error would increase as the prediction window became longer [4]. Therefore, the prediction windows in this paper have been set within one day. Different prediction windows and corresponding model update frequency are shown in table 5.

Table 5. Prediction window lengths and corresponding model update frequency

Prediction window lengths(hours)	Model update frequency	Model(s) trained per day
1	Per hour	24
2	Every 2 hours	12
3	Every 3 hours	8
4	Every 4 hours	6
6	Every 6 hours	4
8	Every 8 hours	3
12	Every 12 hours	2
24	Every 24 hours	1

The evaluation metrics of operating period on different prediction windows is shown in table 6. As the length of the prediction window increases, the prediction error of the model does not change significantly. This indicates that the model has collected enough data samples through one week's history data, and can be used to predict the cooling load in the next 24 hours.

Table 6. Evaluation metrics on different prediction windows

Prediction window lengths(hours)	CVRMSE-on	MAPE-on
1	0.220	0.137
2	0.229	0.150
3	0.242	0.145
4	0.287	0.200
6	0.273	0.162
8	0.239	0.150
12	0.230	0.131
24	0.198	0.132

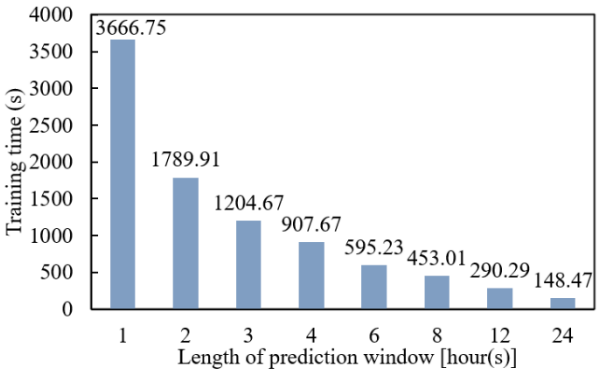


Fig. 9. Training time of models with different prediction windows

Fig.9 shows different model training times under each prediction window. The data experiment was carried out on an 8-core Intel i7-10700K computer with 16GB RAM. The total amount of time a model spends on training highly depends on the frequency of model updating. The shorter the prediction window is, the more frequently the model updates, which means more models are trained during the test period, leading to a longer training time. Therefore, to balance the trade-off between computational costs and prediction accuracy, it is suggested to use one week's history data to predict the load in the next 24 hours in practical application, since the accuracy is similar when the prediction window is 1 hour and 24 hours. But for different history data sizes and quality, the suitable prediction window will vary.

6 Conclusions

This work proposed a new feature selection method for short-term HVAC load prediction based on an analysis on the influence factors of the building load. Different from the commonly used features, indoor black-bulb temperature from different orientations and outdoor air enthalpy are introduced into the model, while historical load is no longer selected, built upon which an empirical analysis is conducted to determine the suitable features for building load prediction. The effectiveness of this method is verified through a cooling load prediction case study of an office park. Compared with commonly used feature set, the proposed method reduces the prediction CVRMSE-on by 21% and MAPE-on by 30%. In addition, by increasing the training dataset size, the model prediction error will decrease first before it will stabilize when the dataset size is adequately large. However, the model accuracy is not significantly affected by the change of prediction window within 24 hours. Our empirical analysis suggests to use one week history data to predict the cooling load in the next 24 hours, which can well balance the trade-off between prediction accuracy and computational costs.

The above results can help develop smart control system to establish accurate and computational-efficient load prediction model, which can inform the optimal control of HVAC systems. Future research may explore whether a different set of features are needed for heating load prediction, and the combination of data analysis methods in feature selection as mentioned in section 1.

Acknowledgement

This study is supported by the National Natural Science Foundation of China (Grant No. 52130803, 52161135201), and Tsinghua University Initiative Scientific Research Program.

References

1. Building Energy Efficiency Research Center, Tsinghua University, *Annual Development Research Report of China's Building energy Efficiency*, China Architecture and Building Press, China (2018)
2. Building Energy Efficiency Research Center, Tsinghua University, *Annual Development Research Report of China's Building energy Efficiency*, China Architecture and Building Press, China (2022)
3. Z. Gao, J. Yu, A. Zhao, Q. Hu, S. Yang, Energy, **238**, 122073 (2022).
4. M. Rana, S. Sethuvenkatraman, M. Goldsworthy, Sustain. Cities Soc., **76**, 103511 (2022)
5. R. Liu, T. Chen, G. Sun, S. M. Muyeen, S. Lin, Y. Mi, Electr. Power Syst. Res., **206**, 107802, (2022)
6. X. Kang, X. Wang, J. An, D. Yan, Energy Build., **275**, 112478, (2022)
7. C. Ghenai, O. A. A. Al-Mufti, O. A. M. Al-Isawi, L. H. L. Amirah, A. Merabet, J. Build. Eng., **52**, 104323, (2022)
8. T. Ahmad, H. Zhang, Energy, **209**, 118477, (2020)
9. Z. Wang, T. Hong, M. A. Piette, Appl. Energy, **263**, 114683, (2020)
10. C. Zhang, J. Li, Y. Zhao, T. Li, Q. Chen, X. Zhang, Energy Build., **225**, 110301, (2020)
11. C. Fan, Y. Ding, Energy Build., **197**, 7–17, (2019)
12. G. Zhu, T. T. Chow, N. Tse, Build. Serv. Eng. Res. Technol., **39**, no. 3, 310–327, (2018)
13. M. W. Ahmad, M. Mourshed, Y. Rezgui, Energy Build., **147**, 77–89, (2017)
14. C. Fan, F. Xiao, Y. Zhao, Appl. Energy, **195**, 222–233, (2017)
15. M. Bourdeau, X. Zhai, E. Nefzaoui, X. Guo, P. Chatellier, Sustainable Cities and Society, **48**, 101533, (2019)
16. J. Zhu, H. Dong, W. Zheng, S. Li, Y. Huang, L. Xi, Appl. Energy, **321**, 119269, (2022)