Energy & Buildings 295 (2023) 113292

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb

When does it pay off to use electricity demand data with rich information about households and their activities? A comparative machine learning approach to demand modelling

Han Wang, Jacek Pawlak, Ahmadreza Faghih Imani, Fangce Guo, Aruna Sivakumar

Urban Systems Laboratory, Centre for Transport Studies, Department of Civil and Environmental Engineering, Imperial College London, United Kingdom

ARTICLE INFO

Keywords:
Energy demand modelling
Machine learning
Meter records
Activities

ABSTRACT

Energy demand modelling has been widely applied in various contexts, including power plant generation, building energy simulation and demand-side management. However, it is still an ongoing research topic in terms of the choice of modelling method, feature engineering for data-driven methods, the application contexts and the type of data used. In the residential sector, survey-based and meter-based approaches are categorised according to the type of input data used, i.e. the activity records from the time use survey and energy consumption from meters respectively. These two paradigms are not necessarily easy to combine, which warrants the questions of when one may be preferred over the other and whether they need to be combined despite the significant data requirements. Other details also have a huge impact on the data structure and performance of the energy demand model, including the choice of influential factors, the historical time window of factors selected, the split between training and test data, and the choice of machine learning (ML) algorithm. There is a lack of comparative research to guide researchers and practitioners in developing energy demand modelling capability, specifically as it pertains to these issues. This study analyses three groups of test scenarios in a multi-household residential context based in the UK. Six ML algorithms (LightGBM, Random forest, ANN, SVM, KNN and LSTM), with eight sets of various influential features, at four different historical time window widths and two train-test splits were compared. An appropriate methodology was designed to capture the temporal impact of activities on energy demand and represent the overlap and interaction of activities. The results show that the combination of meter-based and survey-based energy demand models performs better in terms of modelling accuracy and robustness against sudden load variation. Particularly, integrating energy tariffs, household and individual attributes, appliance usage and general activity features can improve the energy demand model. Among the ML algorithms, LightGBM and ANN perform better than other algorithms while LSTM and SVM may not be suitable in this multi-household short monitoring context.

1. Introduction

The need to model energy demand stems from a number of reasons. Power system operators need an accurate forecast of energy demand to arrange the power plant generation, integrate renewable energy and guarantee grid security. Building owners rely on building energy simulation to evaluate and improve energy efficiency. The electricity market and demand-side management (DSM) projects require energy demand modelling to design the proper price signals or incentives for supply–demand balance. Even single customers are willing to know more about their energy consumption patterns due to rising electricity prices and the growing awareness of energy conservation. Despite various applications and years of research, how to model energy demand is still an ongoing research topic in terms of model categories, feature engineering, modelling algorithms and application contexts.

The role of activities and time use data in energy demand modelling is getting increasing attention. This 'survey-based' approach stems from the notion that energy demand is derived when energy-consuming activities are performed. Then activity-to-power conversion factors were estimated to generate energy demand profiles based on time use activity records [1,2]. Generated results were validated against meter records and showed that these conversion models can reproduce realistic end-

E-mail addresses: han.wang19@imperial.ac.uk (H. Wang), jacek.pawlak@imperial.ac.uk (J. Pawlak), s.faghih-imani@imperial.ac.uk (A. Faghih Imani), fangce.guo@imperial.ac.uk (F. Guo), a.sivakumar@imperial.ac.uk (A. Sivakumar).

https://doi.org/10.1016/j.enbuild.2023.113292
Received 15 February 2023; Received in revised form 12 June 2023; Accepted 19 June 2023
Available online 22 June 2023
0378-7788/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

use composition and hourly demand variation. This conversion model is widely used in DSM since a direct linear relationship between activities and energy demand is established, which can show the change in energy demand when households alter activities in response to incentives or energy prices. However, these conversion models ignore different activity demand patterns across households and simplify the impact of contextual factors such as weather, household socio-demographics, simultaneous activities and energy prices. An alternative approach models energy demand by exploring 'thinner' meter datasets: which are large in terms of data points (i.e., sample size and time scale) but 'thin' without supplementary features. Such data has become increasingly available from various smart metering technologies, providing an opportunity to collect sizeable datasets [3–5]. Machine learning (ML) algorithms are usually utilised in this approach to capture nonlinear relationships and achieve high accuracy, which is related to the consistency of the energy demand patterns and regular activity schedules [6]. This approach also has other advantages such as non-intrusive data collection requirements and simple input features, which make it common in load forecasting and building energy simulation. However, its application in DSM is limited since the activity information is not integrated The data collection requirements for survey-based and meter-based energy demand models are different. Therefore, these two paradigms are not necessarily easy to combine, which warrants the question of when one may be preferred over the other and whether they need to be combined despite the heavy data collection work. A review paper [7] points out that simultaneous activity records and energy consumption measures were necessary for survey-based models. Activity data is relatively sensitive and privacy issues are still a concern, which increases the difficulty of collecting data over a long period. In the residential context, activity data are usually recorded through surveys or recorders (e.g., smartphones) among a large number of households but usually covers a short period (e.g. several days) to avoid participant fatigue and discomfort. In contrast, energy consumption data can be recorded by smart meters and sensors over a longer time (e.g., months or years) with less disturbance to the participants compared with surveys. However, meter data collection is usually time-consuming to guarantee the accuracy of the meter-based approach.

In addition to the activities that drive energy demand, various other influential features affect energy demand patterns and there is no 'one size fits all' method to select the most appropriate features. These influential features include socioeconomics, calendar features, weather and energy price. Socioeconomic features (e.g., age and income) are usually used in the unified predictive model for multiple households as they can reflect the general energy consumption level and activity habits [6]. Calendar features (e.g., the hour of day, the day of week and seasons) are related to space usage schedules and activity profiles that impact energy demand [8]. Weather features tend to have a huge impact on heating and cooling energy consumption. Therefore, short-term weather forecasts and long-term climate change in energy demand modelling were popular topics [8–10]. Energy price has become an influencing factor of energy consumption as multiple energy tariffs (e.g, dynamic time-of-use, dToU) have been used in the real world to reshape load profiles [11,12]. Although each of these features has been studied in the literature, their relative contributions to energy demand modelling have not been compared in the same context. Therefore, an overall understanding of their contribution is required to guide efficient data collection and feature engineering.

As an increasing number of ML algorithms have been introduced into energy demand modelling, the discussion of determining the suitable algorithms, training inputs and train-test splits remain. Popular ML algorithms include Linear Regression (LR), Tree-based Algorithms, Support Vector Machine (SVM), Neural Networks (NN) and Deep Learning (DL) [13]. Although the performance of these methods depends on the data quality and sample size, the general agreement is that machine learning methods such as SVM, NN and DL perform better than LR when a large amount of data is available [14–16]. Despite the intensive study of ML algorithms, research about how to build a proper input dataset is rare. For example, historical data is widely used as input but the width of the historical time window for features is usually arbitrarily decided. Another less discussed issue is how the train-test split impacts the model performance. Two common train-test splits in energy demand modelling correspond to two application contexts. When historical data of modelling targets are available, ML algorithms can train and test in the same context and the train-test split is "within data source". On the contrary, when the modelling target has limited or no historical data, ML algorithms need to be trained in one context but tested in another context. Then the train-test split is "between data source". These different train-test splits should be compared in one study to accurately evaluate the performance in different application contexts.

To address the gaps above, this study designed three groups of test scenarios and compared the performance of six ML algorithms to guide the selection of models. In addition, the impact of eight sets of influential features, four different widths of historical time windows, and two train-test splits on energy demand modelling were analysed in a multi-household residential context to summarise insight for the detailed modelling process. The pros and cons of survey-based and meter-based approaches were also discussed based on the comparison results. ML approaches compared include LightGBM, Random forest, ANN, SVM, KNN and LSTM. Influential factors considered include historical load records, household and individual attributes, appliance ownership, activities, weather and energy policy [17]. A UK dataset was used and contains simultaneous electricity consumption and time use records to fulfil the requirement of both survey-based and meter-based energy demand modelling (see section 2).

The remainder of this paper is structured as follows. Section 2 covers the dataset used in this study and the processing methods. Section 3 describes the modelling framework including ML models and the experimental setup. Section 4 presents and discusses test results. Section 5 summarises the limitation of this study. Section 6 concludes the paper and proposes future work directions.

2. Data

Project METER (https://www.energy-use.org/) provides novel simultaneous activity records and electricity consumption measurements for this study to build and validate energy demand models. This section briefly reviews the dataset collected from 2016 to 2018 and more details were described in [18,19]. Project METER recruited over 200 UK households via radio, the internet, and through campaigns at selected community events. In the data sample, high-income households with solar photovoltaics (PV) and electric vehicles (EV) were over-represented since the reward was the chance to win the cash equivalent of a year's worth of electricity [18]. Recruited households provided socio-demographic information during the registration and received a parcel with a self-administered current clamp, a purpose-built device with an activity recording app and an instruction booklet. The survey date was chosen by the participant from 3 randomly assigned dates. To record typical energy-intensive periods between 5 pm and 7 pm, the electricity monitoring and activity recording started at 5 pm and lasted about 30 h. Each household member ages 8 and above was encouraged to record activities, which created the possibility to explore the impact of collective activities on household-level load profiles.

Electricity consumption was recorded every second by the current clamp installed under the main electricity meter. Then the electricity records were downsampled to mean values every 1 min and 10 min. Activities were reported through the recording app by selecting choices in six steps: activity, time, location, other people, enjoyment and edit [19]. The last step "edit" allows participants to add more details about activities such as the particular nature of the activity (e.g., hot or cold meal) and a prompt for appliances that may have been in use [18].

2.1. Processing

The data from Project METER underwent 8 steps of processing in [17] to enrich the information and remove incomplete observations (Fig. 1).

This paper adopted the same processing method and adjusted the following steps to enable its use for the ML models. In the METER dataset, the start time of activities rather than the activity status in each time slot was recorded at the 1 min interval. This causes one substantial shortcoming the end time of activities is absent and most time slots have no activity records. These sparse start times provide little information for the ML predictive models to capture the impact of activities on energy consumption. In [17] the time slots were imputed with missing activity records with the most recently reported activity. For example, the recorded start time of activity A was set as the missing end time of the former activity B. However, this imputation method may generate inaccurate estimations of end times. Some activities, such as the usage of appliances, usually end when the appliance operation cycles end. In other words, the end time of the appliance operation cycle is not necessarily the time that the next activity starts. Therefore, the activity duration can be overestimated or underestimated and also neglect the overlap of activities.

Instead of the imputation method, this study extracted and designed two features, "Peopleinvolved" and "TimeCredit", for each activity type. "Peopleinvoved" was recorded by participants and shows how many people were involved in the activity. "TimeCredit" was designed to offset the sparsity of activity records, represent the temporal impact of activities and explore the interactions between activities. The idea for "TimeCredit" is that an activity is more likely to happen and last, and affect electricity consumption in the time slot closer to the recorded start or end time. Since the duration of this activity was not recorded, it is reasonable to assume the "survival" possibility of the activity, as well as the probability of affecting electricity consumption, decreases when the time moves away from the recorded time. Then for each activity type in a household, "TimeCredit" first assigned a constant value to the time slot when the activity is recorded. In most cases, the activity record means the start of an activity and the "TimeCredit" value will decrease for the maximum duration of the corresponding activity. While sporadically the activity record meant the end of an activity (e.g., activities with the suffix "(end)") and then the "TimeCredit" decreased backwards along the time for the same period. Without loss of generality, this constant value at the recorded activity time was set as 180 (mins) as the longest duration among the maximum durations of activities [3,20]. An example is shown in Fig. 2. In summary, "TimeCredit", denoted by $TC_{t,A}$, can be seen as the extent to which activity A is believed to happen or last at time t based on the distance between the recorded start/end time $\hat{t}$ of activity A and time t. In addition, each activity type has its own "TimeCredit" feature and a time slot may have several non-zero "TimeCredit" values of different activities which represents the simultaneous presence and interactions of multiple activities.

This study also adopts the moving average method from previous research based on Project METER. It was found that the accuracy of individual activity records with 10 min intervals in the METER dataset is about 67%, which suggested the need to smoothen the records [18]. Therefore, researchers proposed to average the energy consumption records in the vicinity of the reported time (S periods before and after, S = 20 min) [17]. After all the data processing steps and the averaging of the energy consumption, 164 households remained and the 1 min interval energy consumption records ranged from 1.8 Watt to 8302.5 Watt, with the observed mean of 659.6 Watt and the standard deviation (std) of 784.1 Watt. Fig. 3 shows the histogram and density of energy consumption records after all the data processing is completed.

To facilitate the modelling with ML algorithms, categorical features (e.g., household type and income range) were processed into numeric features with Pre-processing scheme Target Encoder [21]. Cyclical features (e.g., the minute of day) were encoded by their sine and cosine values Eq. (1) [22]. By doing this, the time stamp 23:59 is very close to 00:00, as it should be.

$$sin_{minute} = \sin\left(\frac{2\pi T_{minute}}{24 \times 60}\right), cos_{minute} = \cos\left(\frac{2\pi T_{minute}}{24 \times 60}\right) \quad (1)$$

After the processing methods above, the input features were summarised in Appendix 1 and can be categorised into six main types: smart meter records, weather, energy tariff, household and individual attributes, appliance usage and general activity categories.

3. Research methodology

3.1. Experimental setup

Energy demand models in the literature usually generate daily load profiles based on estimated daily activity schedules at 30 mins or 60 min intervals. This kind of generation model fits the purpose of generating typical or random load profiles well but is not suitable for real-time

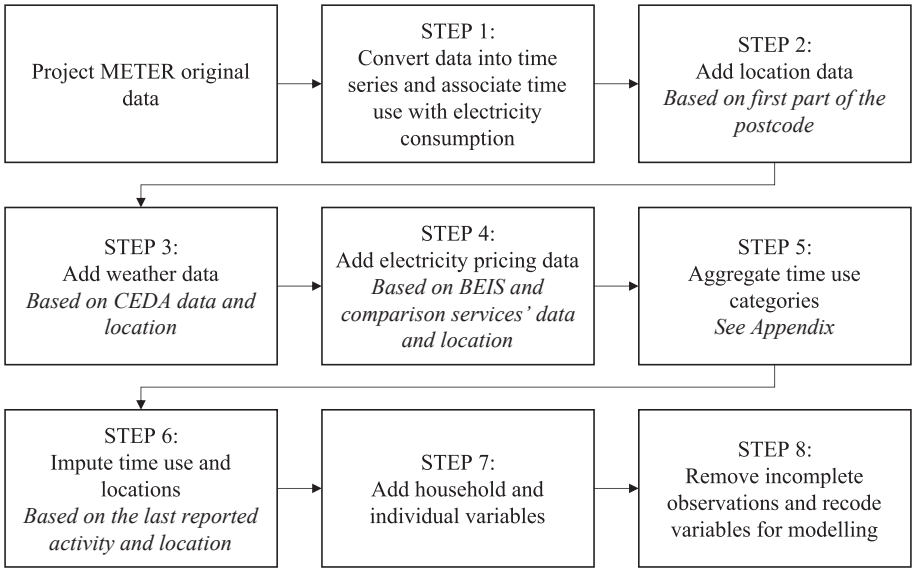


Fig. 1. Data processing steps [17].

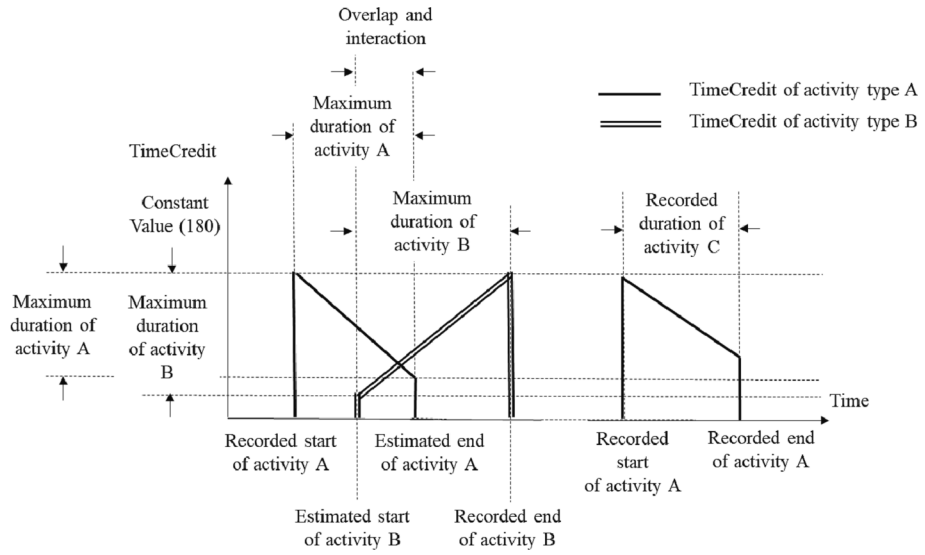


Fig. 2. An example of "TimeCredit".

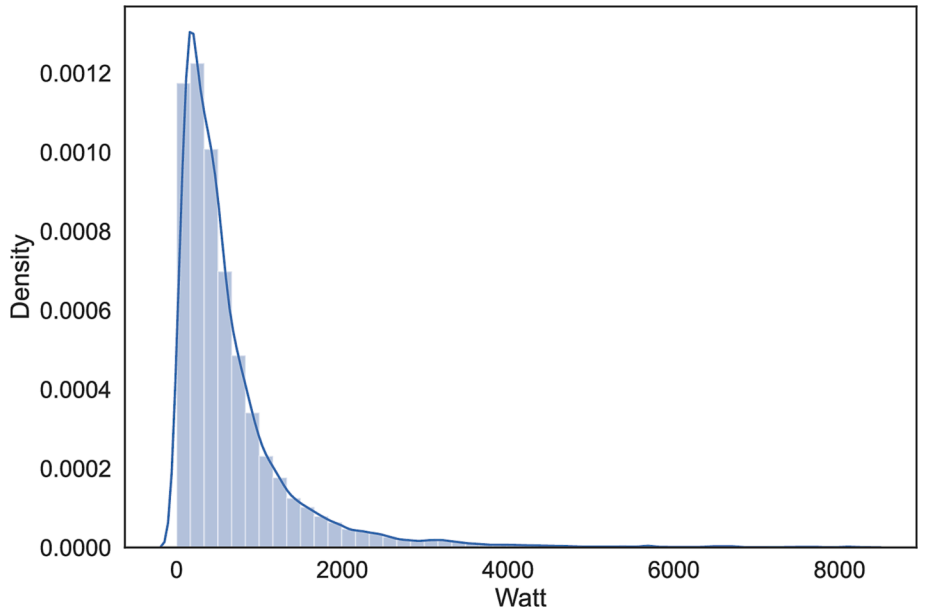


Fig. 3. Histogram of electricity demand after processing.

application of energy management systems and DSM [23–25], where only activities before the control time are known. Regarding the relatively large interval of half-hour, one reason, from a data perspective, is the potentially high degree of variability, noise and missing values that accompany data of finer granularity. Another reason is that the current applications of energy demand modelling, such as the calculation of energy prices, are also based on hour and half-hour intervals. On the contrary, smart or real-time energy control and energy safety tend to have more precise requirements for the load profiles.

Therefore, considering the issues above, this study tested 30 mins ahead prediction of household-level electricity consumption based on the 1 min interval multi-household METER dataset. Only features in the historical time window are used as inputs to satisfy the requirement of real-time control. The granularity of 1 min interval is used to capture the fine fluctuation of demand. The output is the electricity consumption 30 mins head to match the interval of real-time electricity prices. The developed models can be used to predict electricity consumption, as well as generate typical load profiles, for decision-making in real-time or day ahead DSM and smart electricity management systems. Three groups of scenarios were designed to test these models in terms of (1) the contribution of features, (2) the impact of the width of historical time windows, and (3) the impact of train-test splits, on the prediction accuracy. Fig. 4 shows the flowchart of the tests. The results will be analysed to show what are better methodologies in feature engineering and model selection to obtain accurate electricity consumption profiles in different scenarios.

Group 1 contains eight scenarios of input features and categorises them into meter-based, survey-based and meter-and-survey-based models according to the combination of features used (see Table 1). Scenarios 1 to 4 are meter-based models, where the prediction model uses smart meter records, with other energy-related features (weather, energy tariff, household and individual attributes) introduced one at a time. Scenarios 5 and 6 are Meter-and-survey-based models, which utilise both meter-related and survey-related features. Scenarios 7 and 8 are Survey-based models with all features included except for smart meter records. In addition, all scenarios contain general calendar

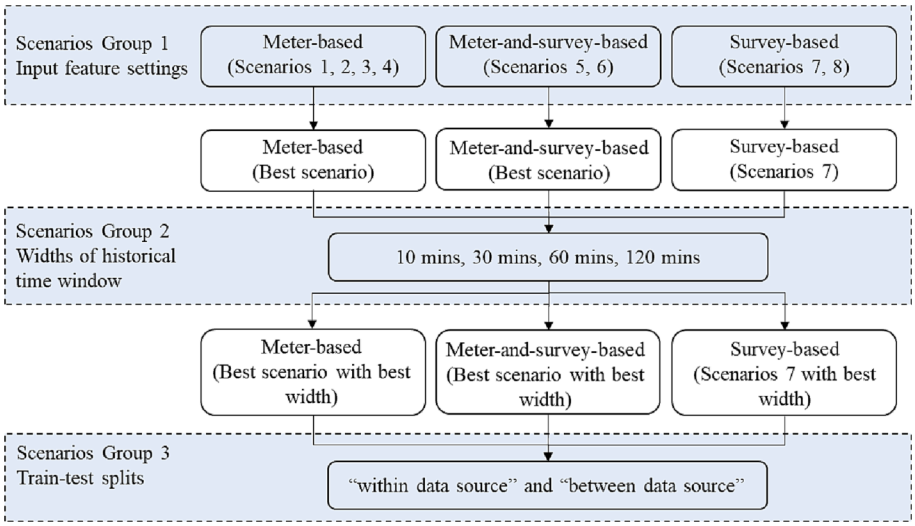


Fig. 4. Flowchat of the tests.

Table 1
Input feature settings in Group 1.

Scenario		Meter-related features		Other energy-related features			Survey-related features	
Category	ID	Smart meter records	Statistics (Smart meter records)	Weather	Energy tariff	Household and individual attributes	Appliance usage	General activity categories
Meter-based	1	•						
	2	•		•				
	3	•		•	•			
	4	•		•	•	•		
Meter-and-survey-based	5	•		•	•	•	•	
	6	•		•	•	•	•	•
Survey-based	7			•	•	•	•	•
	8		•	•	•	•	•	•

features such as the time of day and the day of week. Kolmogorov–Smirnov test (K–S test) was used to assess the statistical significance of the difference between adjacent scenarios. The difference is considered to be significant when the resulting p-value was less than 0.05 (i.e., p less than 0.05).

In Scenario 8, "Statistics (Smart meter records)", which are the statistical features of historical energy consumption records, were added to Scenario 7 to test whether statistical load features can improve the survey-based models when detailed load profiles are not available. Statistical load features are typically used in load forecasting research, in different forms such as the loads at the same time on similar previous days, the average load of the same time on previous days and the average load of the previous day [14]. However, these common statistical features cannot be obtained in the METER dataset since each household was monitored for only 30 h. Then, the statistical features of the whole monitoring period were added based on the assumption that statistical load features do not change greatly between days. These substitutional statistical features include the average, maximum, minimum, median, 25% percentile, 75% percentile and standard deviation of the smart meter records. Although at the risk of data leakage, we expect that Scenario 8 will show the best improvement due to adding the statistical features. Within Group 1, all the features in the past 30 mins were used in the input dataset. The dataset was split within households ("within data source"), and the initial 80% and the remaining 20% of each household's monitoring records were used as training and test sets respectively. For each time slot, the electricity consumption record 30 min after the current time t was the prediction target and features within the historical time window before the time slot were corresponding inputs. As shown in Fig. 5, k is the width of the historical time window and n is the gap between the time of prediction target and current time t.

Group 2 tested four scenarios with different widths of the historical time window (k in Fig. 5): 10 mins, 30 mins, 60 mins and 120 mins. For each scenario in Group 2, three input feature settings from Group 1 were tested. Two of the three are settings with the best performance in each of the meter-based models (Scenarios 1 to 4, Group 1 and the meter-and-survey-based models (Scenarios 5 and 6, Group 1. The last one is the survey-based model Scenario 7. The train-test split remains within households.

Group 3 tested two train-test splits, "within data source" as described in Groups 1 and 2 , and "between data source". In "between data source", the train-test dataset was split across households and a 5-fold cross-validation method was used. For each fold, 80% of households were selected and all of their monitoring data was set as the training dataset. The data from the remaining households was used as the test dataset. These two splits represent different application contexts that reflect whether the historical data of target households is available. The

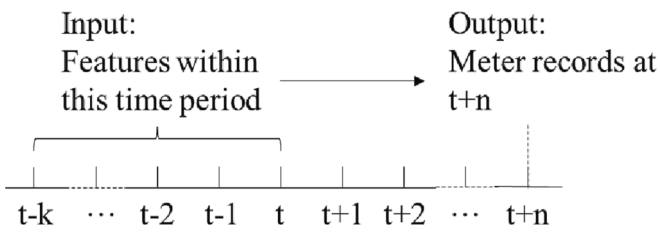


Fig. 5. Historical time window.

same three input feature settings were used as in Group 2 and their corresponding width of historical time windows was the one with the best performance in Group 2.

All the groups and scenarios were trained and tested using Python's statistical and machine learning toolbox.

3.2. Machine learning algorithms and performance metrics

A set of ML algorithms were tested in this study. Zhang et al. (2021) summarised nine main types of ML algorithms that were common in building load prediction: (1) Linear Regression, (2) Support Vector Machine (SVM), (3) Neural Network (NN), (4) Deep Learning, (5) Tree-based Algorithms, (6) Hybrid, (7) AutoRegressive methods, (8) fuzzy time-series model, and (9) Others. This study selected six typical regression models from these categories: Support Vector Regression (SVR), Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), Light Gradient Boosting Machine (LightGBM), Random Forest (RF) and K-Nearest Neighbors (KNN). These models were tuned by cross-fold validation and grid search method to obtain the best performance. Linear regression models were not tested since multiple research shows that linear regression models usually perform worse due to the nonlinearity of the relationship between energy consumption and influential factors [14,16]. Instead, the lazy learning model KNN, which generalizes the data after a query is made, was used as the benchmark model. Features were scaled and then selected by F-value before training.

Root mean square error (RMSE), mean absolute percentage error (MAPE) and coefficient of variation (CV) were used as performance metrics to evaluate the model performance (Eqs. (2)–(4)). Lower RMSE, MAPE and CV mean better prediction performance and greater similarity between the predicted and actual energy consumption records. RMSE was the primary performance metric with the unit: Watt • min or Watt as the interval is 1 min. This smaller unit makes the RMSE in this study more than 1000% larger than that with a common unit kW • h in similar research. In addition, the 1 min interval data may also result in a higher degree of variability and higher RMSE than the hourly data.

$$RMSE(\text{watt} \bullet \text{min}) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}} \tag{2}$$

$$MAPE(\%) = \frac{100}{n} \times \sum_{i=1}^{n} |(y_i - \widehat{y_i})/y_i| \tag{3}$$

$$CV(\%) = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}}}{\frac{1}{n}\sum_{i=1}^{n} y_i} \times 100 \tag{4}$$

Where y_i is the true value, $\widehat{y_i}$ is the predicted value and n is the number of forecasted targets.

4. Results and discussion

4.1. Group 1: input feature settings

Table 2 summarises the prediction results of eight scenarios in Group 1. The best performance was achieved by LightGBM in meter-based Scenario 4 (RMSE: 463.36) and LightGBM in meter-and-survey-based Scenario 5 and 6 also reached close performance levels (RMSE: 464.83 and RMSE: 463.71 respectively). Although load statistical features in Scenario 8 can slightly improve the survey-based Scenario 7, these two survey-based models do not perform as well as other models. The differences in results between adjacent scenarios caused by adding new features are significant as p-values in K–S test were less than 0.05.

Fig. 6 shows the RMSE of predicted results in Group 1 and illustrates the impact of the change in feature settings on the RMSE per algorithm. In the y-axis of Fig. 6, the note below the scenario id represents features or the difference of features compared with the scenario on the left of it. Fig. 6 shows that weather features (in Scenario 2) had little impact on the RMSE except for deteriorating the performance of LSTM. Energy tariff, household and individual attributes (in Scenarios 3 and 4) can improve the performance of ANN and LightGBM but do the opposite for other algorithms. In the meter-and-survey-based scenarios, appliance usage (in Scenario 5) improved the performance of LSTM, RF and KNN but slightly deteriorated the performance of SVM. On the contrary, general activity features (in Scenario 6) only improved the performance of RF. Due to their best performances in three categories respectively, Scenario 4 from meter-based models, Scenario 5 from meter-and-survey-based models and Scenario 7 as the survey-based model with no data leakage were chosen to be tested in Groups 2 and 3. More discussion about meter-based, survey-based and meter-and-survey-based models will be shown in Section 4.4.

4.2. Group 2: Historical time window

Table 3, Table 4 and Table 5 present the results of scenarios in Group 1 with the input features of Scenarios 4, 5 and 7 in Group 1 respectively. For meter-based (see Table 3) and meter-and-survey-based (see Table 4) models, their best performances were both achieved by LightGBM models with features in the past 10 mins (RMSE 458.09 and 458.92 respectively). For survey-based models (See Table 5), the best performance was achieved by the LightGBM with features in the past 30 mins (RMSE: 567.00). This result suggests the consistency of electricity consumption patterns in a short period is so strong that adding features in the earlier period may not improve the prediction accuracy. On the other hand, the impact of activities on electricity consumption may at least last 30 mins, so the activity features in wider historical time windows are helpful in survey-based energy demand models.

These impacts are shown in Figs. 7, 8 and 9 as the RMSE (the y-axis) of most algorithms rises with the increment of the width of historical windows (the x-axis). Therefore, Scenario 4 and Scenario 5 in Group 1 with features in the past 10 min, and Scenario 7 in Group 1 with features in the past 30 min were tested with different train-test splits in Group 3.

4.3. Group 3: Train-test splits

The performances of scenarios in Group 3 are shown in Tables 6, 7 and 8. LightGBM achieved the best performance when the data was split within households ("within data source") in meter-based, meter-and-survey-based and survey-based models. When the train-test split changed from within households to across households ("between data source"), the average RMSE increment was 13.42%, 11.14% and 21.65% for meter-based, meter-and-survey-based and survey-based models respectively. The meter-and-survey-based models have the smallest change of RMSE and the reason may be the rich information from combining both meter and activity records. The comparison result of the two train-test splits comes as no surprise that when the dataset was split within households, historical data in the same context was available and can provide more useful information about the load patterns. On the contrary, when the dataset was split across households, the source and target domain of ML algorithms were different, which lead to worse modelling performance. The prediction with different sources and target domains involves another branch of the ML research field: transfer learning [26] which is beyond the scope of this paper.

In all three groups of tests, the best performances were achieved by LightGBM with train-test split "within data source" in three scenarios, Meter-based Scenario 4 with features in the past 10 min, Meter-and-survey-based Scenario 5 with features in the past 10 min and Survey-based Scenario 7 with features in the past 30 min. Then, the comparisons between records and predicted loads in test datasets in these three scenarios are shown in Fig. 10. These records are at 1 min intervals from different households during different times so the ticks in the x-axis are omitted. This figure shows that the developed models predict the lower records more accurately than the sudden peak records. Model

Table 2
Result of input feature setting test (Group 1)

Scenario		Metrics	LightGBM	Random forest	ANN	SVM	KNN	LSTM
Meter-based	1	RMSE	500.35	512.17	500.67	581.21	549.90	549.72
		MAPE	136.89%	381.17%	273.94%	142.58%	406.79%	452.56%
		CV	64.79%	66.32%	64.83%	75.39%	71.21%	70.88%
	2	RMSE	502.05	513.42	501.96	582.48	554.15	585.46
		MAPE	150.06%	411.40%	268.66%	142.64%	435.06%	363.93%
		CV	65.01%	66.48%	65.00%	75.42%	71.75%	75.49%
		K-S	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$
	3	RMSE	473.63	519.54	490.53	591.00	531.42	645.82
		MAPE	154.23%	401.30%	307.02%	151.84%	409.85%	346.24%
		CV	61.33%	67.27%	63.52%	76.53%	68.81%	83.27%
		K-S	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$
	4	RMSE	463.36	568.52	478.56	596.76	543.27	638.60
		MAPE	148.06%	1096.75%	312.70%	244.69%	658.68%	428.17%
		CV	60.00%	73.62%	61.97%	77.27%	70.35%	82.34%
		K-S	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$
Meter-and-survey-based	5	RMSE	464.83	559.40	480.99	607.60	529.20	606.03
		MAPE	141.18%	1091.10%	319.59%	246.77%	648.87 %	482.99%
		CV	60.19%	72.43%	62.28%	78.68%	68.52%	78.14%
		K-S	$P > 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$
	6	RMSE	463.71	552.28	480.66	612.64	535.06	649.09
		MAPE	136.83%	971.83%	295.40%	248.17%	610.60%	435.52%
		CV	60.04%	71.51%	62.24%	79.33%	69.28%	83.70%
		K-S	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$
Survey-based	7	RMSE	567.00	616.20	573.48	707.76	607.38	678.23
		MAPE	784.87%	983.66%	841.59%	654.82%	724.86%	834.78%
		CV	73.42%	79.79%	74.26%	91.65%	78.65%	87.45%
		K-S	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$
	8	RMSE	554.31	611.49	549.68	687.56	600.08	654.68
		MAPE	773.16%	925.88%	793.87%	548.82%	722.38%	710.25%
		CV	71.78%	79.18%	71.18%	89.03%	77.70%	84.42%
		K-S	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P < 0.05$	$P > 0.05$	$P < 0.05$

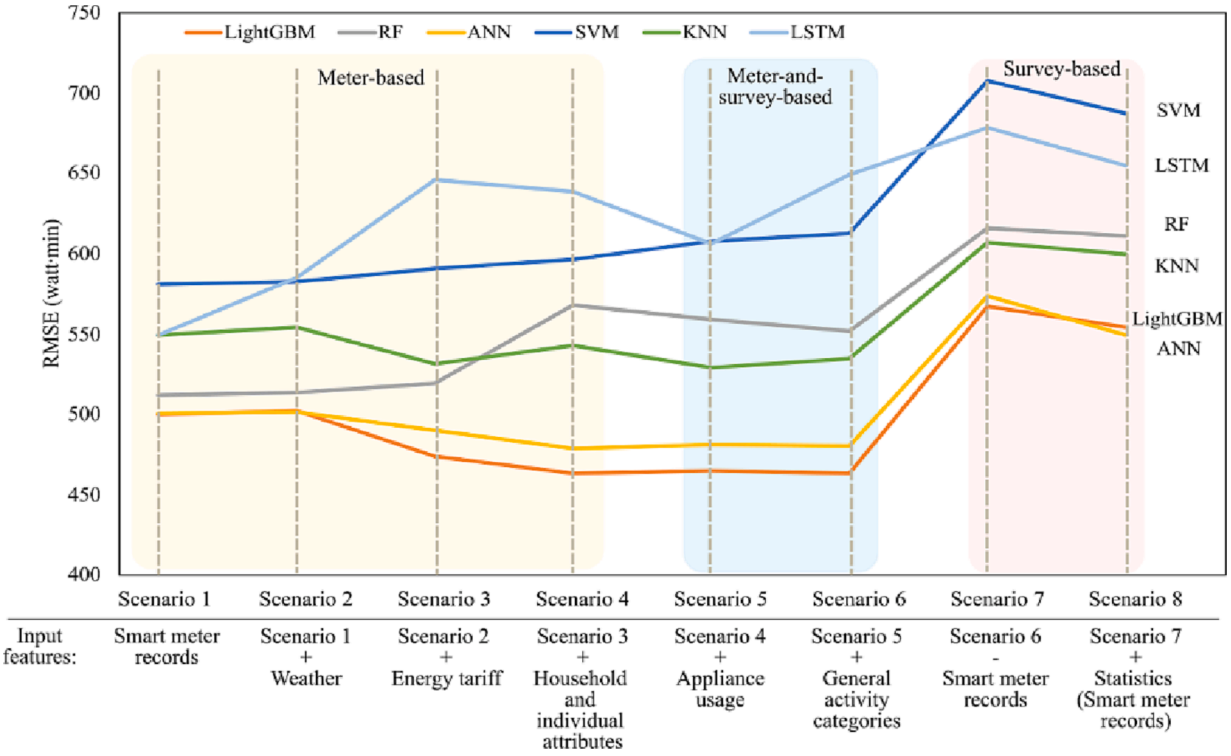


Fig. 6. Result of input feature test Group 1

Table3
The result of the historical time window test (Group 2 with meter-based Scenario 4)

Scenario 4	Metrics	LightGBM	Random forest	ANN	SVM	KNN	LSTM
Past 10 mins	RMSE	458.09	551.93	475.75	596.61	542.71	601.42
	MAPE	165.90%	1090.54%	300.56%	243.70%	831.18%	496.40%
	CV	59.42%	71.59%	61.71%	77.39%	70.39%	77.70%
Past 30 mins	RMSE	463.36	568.52	478.56	596.76	543.27	638.60
	MAPE	148.06%	1096.75%	312.70%	244.69%	658.68%	428.17%
	CV	60.00%	73.62%	61.97%	77.27%	70.35%	82.34%
Past 60 mins	RMSE	461.46	538.98	479.18	600.28	544.95	614.5
	MAPE	145.57%	584.75%	273.53%	227.18%	612.38%	391.18%
	CV	59.59%	69.60%	61.88%	77.51%	70.37%	79.01%
Past 120 mins	RMSE	480.58	534.41	492.06	607.86	559.93	658.40
	MAPE	142.39%	464.40%	335.76%	230.89%	577.75%	388.31%
	CV	61.70%	68.61%	63.18%	78.05%	71.89%	84.18%

Table 4
The result of the historical time window test (Group 2 with meter-and-survey-based Scenario 5)

Scenario 5	Metrics	LightGBM	Random forest	ANN	SVM	KNN	LSTM
Past 10 mins	RMSE	458.92	546.87	477.22	606.19	529.74	614.13
	MAPE	159.30%	1079.68%	276.18%	244.98%	819.93%	607.69%
	CV	59.53%	70.93%	61.90%	78.63%	68.71%	79.34%
Past 30 mins	RMSE	461.83	559.40	480.99	607.60	529.20	606.03
	MAPE	141.18%	1091.10%	319.59%	246.77%	648.87 %	482.99%
	CV	60.19%	72.43%	62.28%	78.68%	68.52%	78.14%
Past 60 mins	RMSE	461.75	532.44	482.12	614.53	533.08	607.72
	MAPE	140.53%	574.14%	266.57%	235.00%	599.51%	369.15%
	CV	59.62%	68.75%	62.26%	79.35%	68.84%	78.13%
Past 120 mins	RMSE	481.80	525.50	498.47	626.49	564.89	609.20
	MAPE	136.71%	457.21%	336.94%	251.69%	557.47%	358.78%
	CV	61.86%	67.47%	64.00%	80.44%	72.53%	77.89%

Table 5
The result of the historical time window test (Group 2 with survey-based Scenario 7)

Scenario 7	Metrics	LightGBM	Random forest	ANN	SVM	KNN	LSTM
Past 10 mins	RMSE	570.13	627.34	570.44	706.58	628.64	672.37
	MAPE	814.48%	1046.12%	854.61%	652.42%	803.02%	732.83%
	CV	73.95%	81.37%	73.99%	91.65%	81.54%	86.86%
Past 30 mins	RMSE	567.00	616.20	573.48	707.76	607.38	678.23
	MAPE	784.87%	983.66%	841.59%	654.82%	724.86%	834.78%
	CV	73.42%	79.79%	74.26%	91.65%	78.65%	87.45%
Past 60 mins	RMSE	578.00	612.74	573.76	720.54	610.72	705.95
	MAPE	792.54%	1080.46%	833.37%	667.93%	648.13%	781.49%
	CV	74.64%	79.12%	74.09%	93.04%	78.86%	90.76%
Past 120 mins	RMSE	600.14	661.93	588.12	741.15	645.63	750.43
	MAPE	852.62%	979.01	809.14%	691.73%	712.88%	1278.74%
	CV	77.05%	84.99%	75.51%	95.16%	82.89%	95.95%

performance and predicted errors are discussed in the next sections.

4.4. Meter-based, meter-and-survey-based and survey-based models

In this study, meter-based models and meter-and-survey-based models achieved better performance than survey-based models. The reasons that survey-based models do not perform well in the METER dataset are as follows:

(1) Interval of 1 minute. The fine granularity of data strengthens the consistency between adjacent energy consumption patterns and probably magnifies the mismatch between activity records and energy consumption records. This mismatch is related to the second reason below.
(2) Misrecorded activities and absent end time. It is nearly impossible for the participants to manually record their activities at the exact time. Therefore, load peaks caused by misrecorded activities are hard to predict with only activity records. One example is shown in Fig. 11 where two peak loads and no activities were recorded around 23:00 Jan 13, 2017, and around 12:00 Jan 14, 2017. In addition, the end time of activities was not recorded in the survey, which creates difficulties in determining the temporal boundary of activity impact. This, however, is an artefact specifically of the METER survey.
(3) The energy-related heterogeneity of households. The METER dataset was collected from more than 200 households. These households are likely to have different 'always-on' loads (e.g. lighting, refrigerators and freezers), activity patterns and power rates of appliances. Although the household and individual attributes were included in the METER dataset, the energy-related heterogeneity above was not recorded or fully represented. Therefore, survey-based ML models probably fail to learn these differences between households and are unable to predict accurately in this study.

To better understand the model performance in different household contexts, the prediction results in Group 1 were analysed per household. An intuitive assumption is that survey-based models can achieve better

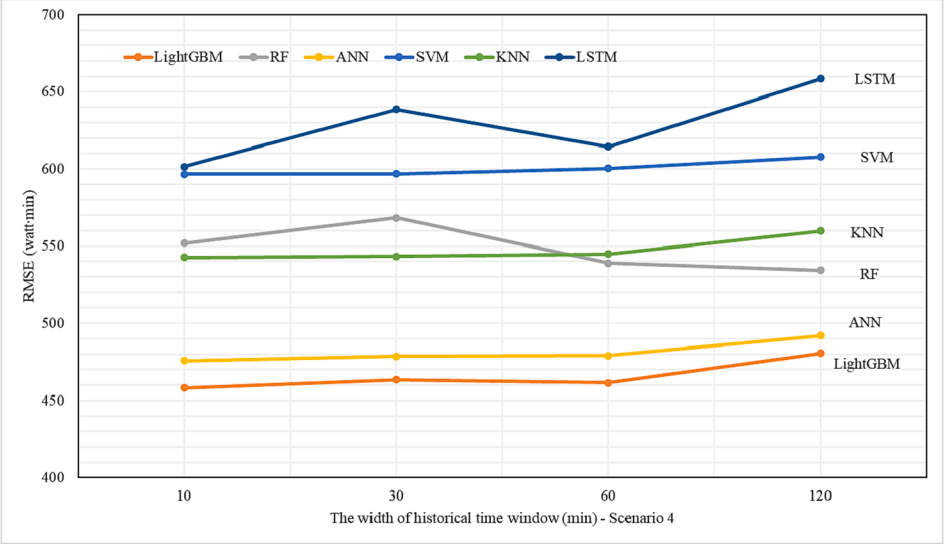


Fig. 7. The result of the historical time window test (Group 2 with meter-based Scenario 4)

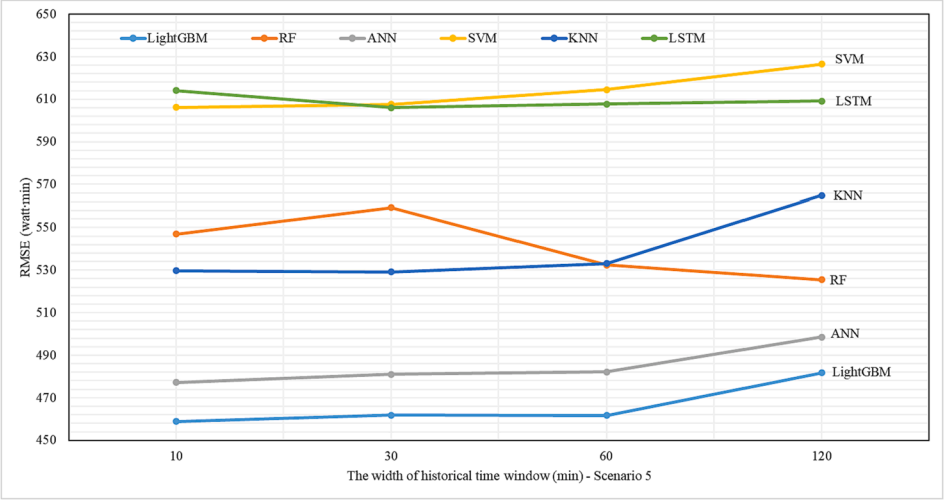


Fig. 8. The result of the historical time window test (Group 2 with meter-and-survey-based Scenario 5)

prediction performance than meter-based models when sudden load variation appears due to activities, which weakens the consistency of electricity consumption patterns. Therefore, the average of the standard deviation of energy consumption records over every 120 mins was calculated as the "average std" for each household to represent the extent of sudden load variation. Each spot in Fig. 12 represents an algorithm's prediction result of a household in one scenario in Group 1. The y-axis is RMSE and the x-axis is "average std". Three best-fit linear trendlines were drawn based on all results of meter-based (Scenario 1 to 4), meter-and-survey-based (Scenario 5 to 6) and survey-based (Scenario 7 to 8) scenarios respectively. This figure shows that the larger "average std" of energy consumption records, representing more sudden load variation, will result in worse prediction performance. The slope of the meter-and-survey-based scenarios' trendline (red) is smaller than the slope of the meter-based scenarios' trendline (blue). Therefore, this figure also shows that although the prediction performance of survey-based models is the most sensitive one, activity features can help improve the robustness against "average std".

The results in Tables 2-8 show that the RMSE of the developed models after combining meter-related and survey-related features is still large. The main source of the remaining errors could be the fine granularity of the dataset as stated before and the accompanying sudden load variations. Electricity records at the minute level can capture more variations compared to hourly and half-hourly records. Fig. 10 shows the failure of models to predict sudden peak loads and Fig. 12 illustrates the positive correlation between RMSE and sudden load variation "average std". Therefore, one direction for future work to improve the model performance can be training models for periods or households with and without sudden load variations separately as well as designing models to detect the sudden variation.

4.5. ML model performance

All the test results and Figs. 6 to 9 show that LightGBM and ANN usually perform better than other algorithms in this multi-household dataset. Particularly, Tables 6, 7 and 8 show that LightGBM performs better when the train-test split is within households while ANN performs better when the train-test split is across households. Following the same per households analysis in the last section, each spot in Fig. 13 represents an algorithm's prediction result of a household in one scenario in Groups 1 and 2. The axes are RMSE and "average std". The best fit linear trendlines of algorithms also indicate that LightGBM and ANN are robust

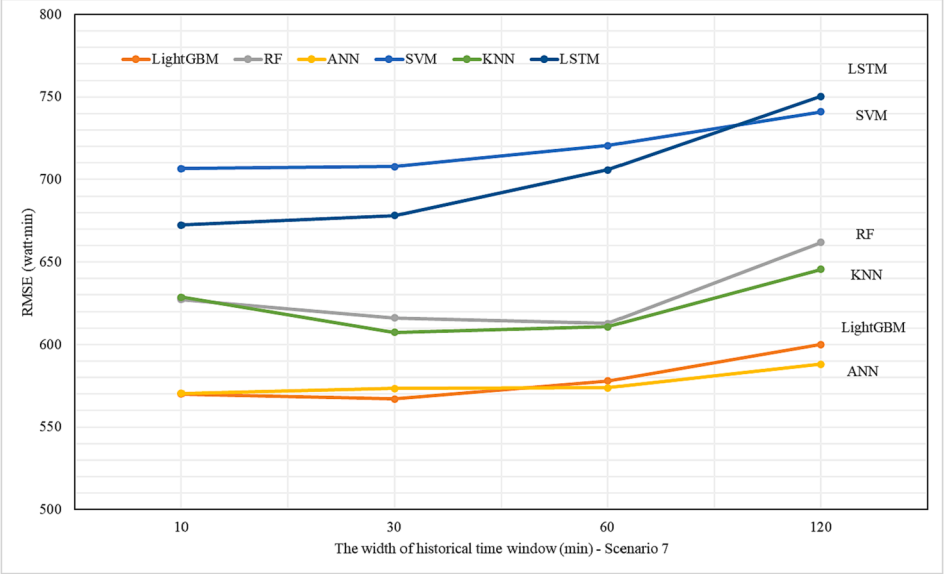


Fig. 9. The result of the historical time window test (Group 2 with survey-based Scenario 7)

Table 6
Result of the train-test split test (Group 3 with Scenario 4, past 10 mins)

Scenario 4	Metrics	LightGBM	Random forest	ANN	SVM	KNN	LSTM
Split within household	RMSE	458.09	551.93	475.75	596.61	542.71	601.42
	MAPE	165.90%	1090.54%	300.56%	243.70%	831.18%	496.40%
	CV	59.42%	71.59%	61.71%	77.39%	70.39%	77.70%
Split across household	RMSE	564.32	589.95	517.23	627.58	612.62	743.55
	MAPE	67.04%	460.41%	187.75%	345.49%	385.05%	420.29%
	CV	84.25%	91.90%	79.22%	93.25%	94.32%	117.42%
	RMSE increase	23.19%	6.89%	8.72%	5.19%	12.88%	23.63%
	Average Increase	13.42%					

Table 7
Result of the train-test split test (Group 3 with Scenario 5, past 10 mins)

Scenario 5	Metrics	LightGBM	Random forest	ANN	SVM	KNN	LSTM
Split within household	RMSE	458.92	546.87	477.22	606.19	529.74	614.13
	MAPE	159.30%	1079.68%	276.18%	244.98%	819.93%	607.69%
	CV	59.53%	70.93%	61.90%	78.63%	68.71%	79.34%
Split across household	RMSE	563.45	595.80	518.26	624.34	609.98	665.53
	MAPE	79.58%	457.08%	214.86%	337.92%	406.53%	401.80%
	CV	84.12%	93.02%	79.45%	92.76%	93.93%	102.30%
	RMSE increase	22.78%	8.95%	8.60%	2.99%	15.15%	8.37%
	Average Increase	11.14%					

to the sudden load variation. On the other hand, LSTM and SVM do not perform well in the METER dataset, which is not consistent with previous research.

The different performances of algorithms may be explained by their particular application contexts and the characteristics of the dataset. The METER dataset contains short monitoring periods of multiple households. Various data sources in the dataset are not suitable for LSTM which is well-suited for long-term time series data from a single source. The basic idea behind SVM is to use kernel functions to transform data and find hyperplanes to separate data. However, multiple data sources in the METER dataset increase the difficulty to find proper kernel functions and hyperplanes. KNN takes the advantage of its simple structure and the similarity between the prediction target and the observation in the short past. RF is the ensemble model with a bagging technique and multiple simple decision trees. KNN and RF are famous for their easy deployment rather than their ability to address complex relationships.

On the contrary, ANN is probably one of the most fundamental ML models which can contain multiple layers of various numbers of neurons. The complex connections between neurons make ANN more robust to the noise and capable to study nonlinear relationships. Therefore, ANN may be more suitable for the METER dataset which has multiple data sources. LightGBM is a free and open-source distributed gradient boosting framework for machine learning originally developed by

Table 8
Result of the train-test split test (Group 3 with Scenario 7, past 30 mins)

Scenario 7	Metrics	LightGBM	Random forest	ANN	SVM	KNN	LSTM
Split within household	RMSE	567.00	616.20	573.48	707.76	607.38	678.23
	MAPE	784.87%	983.66%	841.59%	654.82%	724.86%	834.78%
	CV	73.42%	79.79%	74.26%	91.65%	78.65%	87.45%
Split across household	RMSE	735.41	800.29	715.35	712.61	788.87	780.14
	MAPE	617.91%	687.50%	583.80%	546.42%	620.71%	478.82%
	CV	110.28%	123.17%	110.58%	106.68%	120.97%	120.94%
	RMSE increase	29.70%	29.88%	24.74%	0.69%	29.88%	15.03%
	Average Increase	21.65%					

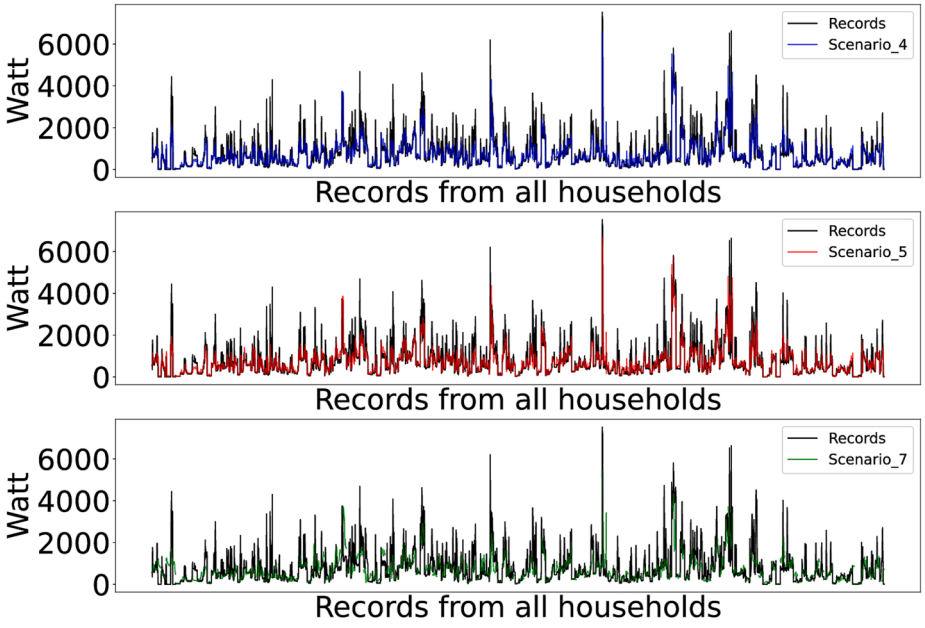


Fig. 10. Prediction results (LightGBM in Group 3)

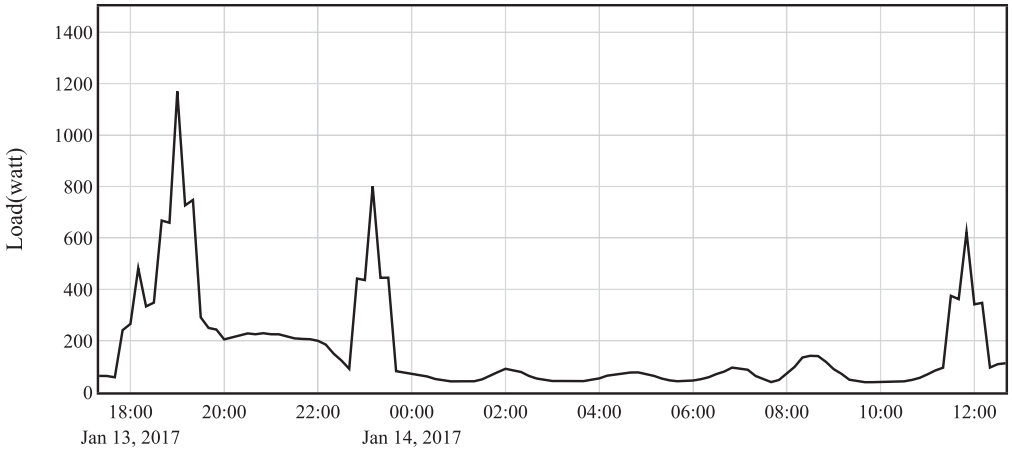


Fig. 11. Load profile of a household.

Microsoft. It is not a basic model like other ML algorithms tested in this study but a well-tuned and improved decision-tree-based model. LightGBM uses plenty of techniques to speed up the training and optimise the accuracy of results. More details can be found in the official document.[1] Therefore, LightGBM can outperform other basic ML algorithms.

[1] https://lightgbm.readthedocs.io/en/latest/Features.html.

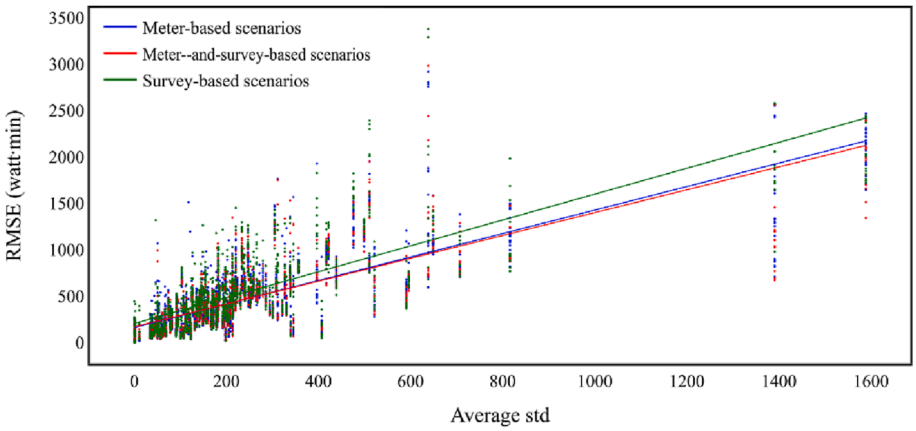


Fig. 12. Per household analysis of test results in Group 1.

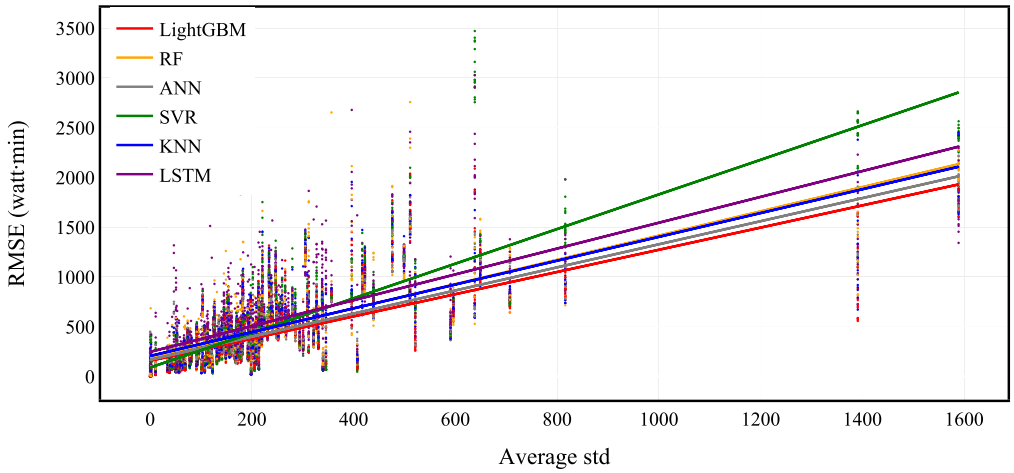


Fig. 13. Per household analysis of algorithm performance.

5. Limitations

This section summarises three main limitations of this study. First, the METER dataset lacks several important features and the sample size is small. The end time of activities and the power rate of appliances were not recorded, which deteriorates the performance of survey-based models. What's more, only 30 h of novel simultaneous records of activities and energy consumption were recorded per household. These short periods weaken the models' ability to learn the trends and daily cycles of time-series electricity consumption. Despite the short period, the data collection of different households covers all months and seasons, which were represented by the calendar features (e.g., the time of day, the day of week, month and seasons). The integration of calendar features and the potential mixed use of gas and electricity for heating in the UK may make the contribution of weather features to electricity consumption insignificant.

Second, this research built a novel survey-related feature "TimeCredit" (Section 2.1) to overcome the shortcoming of the missing activity end times and the sparsity of activity records in the METER dataset. This feature also attempts to capture the temporal impact of activities on energy consumption and represent the overlap and interaction of activities. However, the assumption that ML methods can learn the boundary of activity impact and the interaction of activities based on this feature needs to be validated with more datasets. The first limitation of multiple households with short monitoring periods and missing information about appliance also increase the difficulty of validation in this dataset.

Third, this study tested the performance of six popular ML algorithms. Although the grid search method was used to tune hyperparameters (e.g., k in KNN, number of neural units in neural network and regularization parameters) in ML algorithms, this tuning process in ML is tricky, especially when the METER dataset is unique and novel. Therefore, these results do not guarantee the same level of performance in other contexts. Testing different algorithms for different datasets and different contexts is essential [16].

6. Conclusions

This study designed three groups of test scenarios to compare the performance of energy demand models. The first group of scenarios analysed the contribution of energy-related features one by one and compared the prediction performance of meter-based, meter-and-survey-based and survey-based models. The second group of scenarios tested the impact of four widths of historical time windows of features. The third group of scenarios applied the energy demand modelling in two application contexts in terms of different train-test splits. Six popular ML algorithms were compared in all three test groups based on a dataset built on the Project METER survey data in the UK, which contained simultaneous meter and activity records, weather, energy tariff, household and individual attributes, appliance usage and general activity categories.

The main conclusions of this study are summarised below:

- Integrating energy tariffs, household and individual attributes, appliance usage and general activity features into energy demand modelling can improve the prediction performance.
- Survey-based models were outperformed by meter-and-survey-based and meter-based models due to the fine granularity, misrecorded activities, absent activity end time, and heterogeneity between households in the METER dataset. Despite the worse performance of survey-based models, survey-related features can make the model performance robust against sudden load variation.
- Meter-based and meter-and-survey-based models with features in a short period (past 10 min) can achieve good performance while survey-based models require features in earlier and longer periods (past 30 min).
- All models performed worse when the train-test split changed from within households to across households and the performance of meter-and-survey-based models was less deteriorated than that of survey-based and meter-based models. This shows that the survey data plays a role in generalising the energy patterns learnt from the meter data.
- Generally, LightGBM and ANN performed better than other ML algorithms. LSTM and SVM may not be suitable in this multi-household short monitoring context. Particularly, LightGBM and ANN show superiority when the train-test split was within households and across households respectively.

In summary, as the answer to the question in the title, there are three contexts where it is worth using electricity demand data with rich information about households and their activities in demand modelling: (A) The output of demand modelling is used for DSM with indoor activities; (B) Sudden load variations exist in the load profiles; (C) Energy demand models are used for targets with less historical data or are different from the source of training data. In these three contexts, rich information about households and their activities can help establish the relationship between activities and energy demand, robust the model performance and generalise the energy patterns.

This study contributes to the knowledge of modelling energy demand in fine granularity using ML algorithms with various influential factors, different widths of historical time windows and various contexts. In addition, the overlap and interactions of activities were represented by a designed feature "TimeCredit". Those influential factors and model details were rarely studied in the literature. Future work could be testing these ML models in a larger dataset with simultaneous activity and meter records for validation. In addition, models can be specifically trained to detect and predict sudden load variations to improve accuracy. Another direction is applying these ML models in DSM projects to explore the potential of integrating the interactions of activities.

Declaration of Competing Interest

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.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work is supported by the Singapore Ministry of National Development and the National Research Foundation, Prime Minister's Office under the Land and Liveability National Innovation Challenge (L2NIC) Research Programme (L2 NIC Award No. L2NICTDF1-2017-3). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the reviews of the Singapore Ministry of National Development and National Research Foundation, Prime Minister's Office, Singapore. The authors also acknowledge the support from Nanyang Technological University-Imperial Collaboration Fund (G14045) and the Integrated Development of Low-Carbon Energy Systems (IDLES) research programme. IDLES at Imperial College London is funded by the Engineering and Physical Sciences Research Council (EPSRC) Programme Grants (EPSRC Grant No. EP/R045518/1).

Appendix 1:. Features in the METER dataset after processing

Feature Category		Feature	Description
Calendar		Weekday	–
		Month	–
		BankHoliday	–
		Year	–
		Minute(sin)	–
		Minute(cos)	–
Meter-related	Smart meter record	Watt	–
	Smart meter record-statistics statistics	Watt (average)	–
		Watt (max)	–
		Watt (min)	–
		Watt (median)	–
		Watt (25% percentile)	–
		Watt (75% percentile)	–
		Watt (standard deviation)	–
Other energy-related	Weather	Air temperature	–
		Dewpoint	–
		Relative humidity	–
	Energy tariff	EnergyType	The type of energy tariff
		EnergyTypePeak	Indicator of peak time
		StandingEnergyTariff	Fixed cost of the tariff (in GBP)
		EnergyConsumptionTariff	kWh cost of the tariff (in 0.01 GBP, i.e. pence)
	Household and individual attributes	People	How many people live in the household?
		Age_group_i(i = 1, 2, 3, 4, 5,6)	How many of you are in each age group?
		House_type	What type of house do you live in?
		Rooms	How many rooms does your home have?
		Own	Do you own or rent your property?
		Income	What is your household income?

(*continued on next page*)

(*continued*)

Feature Category		Feature	Description
		Student	How many of you are in each occupation group?
		Employed	
		Self-employed	
		Volunteer	
		Unemployed	
		Electric car	Appliance ownership
		Solar thermal	
		PV panels	
		Electricity display	
		Electric hob	
		Gas boiler	
		Heatpump	
		Underfloor heating	
		Washer dryer	
		Tumble dryer	
		Washing machine	
		Dehumidifier	
		Airconditioner	
		Portable heaters	
		Night storage heaters	
		Power shower	
		TVcomputerscreens	
Survey-related	Appliance usage	Kettle	"Peopleinvolved" and "TimeCredit" for each appliance and activity category
		Making breakfast	
		Making lunch	
		Making dinner	
		Hob	
		Oven	
		TV	
		Information and communication technology (ICT) products	
		Screen time (e.g., tablet, laptop)	
		Shower	
		Microwave	
		Toaster	
		Other	
	General activity categories	Care_self	
		Food	
		Recreation	
		Work	
		Care_other	
		Care_house	
		Other_category	

References

[1] J. Widén, M. Lundh, I. Vassileva, E. Dahlquist, K. Ellegård, E. Wäckelgård, Constructing load profiles for household electricity and hot water from time-use data-Modelling approach and validation, Energ. Buildings 41 (2009) 753–768, https://doi.org/10.1016/j.enbuild.2009.02.013.

[2] J. Widén, E. Wäckelgård, A high-resolution stochastic model of domestic activity patterns and electricity demand, Appl. Energy 87 (2010) 1880–1892, https://doi.org/10.1016/j.apenergy.2009.11.006.

[3] S. Bissey, S. Jacques, J.-C. Le Bunetel, The fuzzy logic method to efficiently optimize electricity consumption in individual housing, Energies 10 (11) (2017) 1701.

[4] S. Kim, Y. Sung, Y. Sung, D. Seo, Development of a consecutive occupancy estimation framework for improving the energy demand prediction performance of building energy modeling tools, Energies 12 (3) (2019) 433.

[5] N. Somu, G.R. M r, K. Ramamritham, A hybrid model for building energy consumption forecasting using long short term memory networks, Appl. Energy 261 (2020) 114131.

[6] P.A. Schirmer, C. Geiger, I. Mporas, Residential energy consumption prediction using inter-household energy data and socioeconomic information, European Signal Processing Conference. 2021-Janua (2021) 1595–1599. https://doi.org/10.23919/Eusipco47968.2020.9287395.

[7] A. Grandjean, J. Adnot, G. Binet, A review and an analysis of the residential electric load curve models, Renew. Sustain. Energy Rev. 16 (2012) 6539–6565, https://doi.org/10.1016/j.rser.2012.08.013.

[8] L. Xu, S. Wang, R. Tang, Probabilistic load forecasting for buildings considering weather forecasting uncertainty and uncertain peak load, Appl. Energy 237 (2019) 180–195, https://doi.org/10.1016/j.apenergy.2019.01.022.

[9] E. Lucas Segarra, G. Ramos Ruiz, C. Fernández Bandera, Probabilistic load forecasting optimization for building energy models via day characterization, Sensors 21 (9) (2021) 3299.

[10] Y.X. Ma, C. Yu, Impact of meteorological factors on high-rise office building energy consumption in Hong Kong: From a spatiotemporal perspective, Energ. Buildings 228 (2020), 110468, https://doi.org/10.1016/j.enbuild.2020.110468.

[11] Y. Dai, P. Zhao, A hybrid load forecasting model based on support vector machine with intelligent methods for feature selection and parameter optimization, Appl. Energy 279 (2020), 115332, https://doi.org/10.1016/j.apenergy.2020.115332.

[12] J. Schofield, R. Carmichael, S. Tindemans, M. Woolf, M. Bilton, G. Strbac, Report A3 Residential consumer responsiveness to time-varying pricing., 2014.

[13] L. Zhang, J. Wen, Y. Li, J. Chen, Y. Ye, Y. Fu, W. Livingood, A review of machine learning in building load prediction, Appl. Energy 285 (2021), 116452, https://doi.org/10.1016/j.apenergy.2021.116452.

[14] P. Lusis, K. Rajab, L. Andrew, A. Liebman, Short-term residential load forecasting: impact of calendar e ff ects and forecast granularity, Appl. Energy 205 (2017) 654–669, https://doi.org/10.1016/j.apenergy.2017.07.114.

[15] N. Shabbir, R. Ahmadiahangar, L. Kutt, A. Rosin, Comparison of machine learning based methods for residential load forecasting, 2019 Electric Power Quality and Supply Reliability Conference and 2019 Symposium on Electrical Engineering and Mechatronics, PQ SEEM 2019 (2019) 2019–2022, https://doi.org/10.1109/PQ.2019.8818267.

[16] K. Amasyali, N.M. El-Gohary, Real data-driven occupant-behavior optimization for reduced energy consumption and improved comfort, Appl. Energy 302 (2021), 117276, https://doi.org/10.1016/j.apenergy.2021.117276.

[17] J. Pawlak, A.F. Imani, A. Sivakumar, How do household activities drive electricity demand ? Applying activity-based modelling in the context of the United Kingdom, Energy Res. Soc. Sci. 82 (2021), 102318, https://doi.org/10.1016/j.erss.2021.102318.

[18] P. Grünewald, M. Diakonova, The electricity footprint of household activities - implications for demand models, Energ. Buildings 174 (2018) 635–641, https://doi.org/10.1016/j.enbuild.2018.06.034.

[19] P. Grünewald, M. Diakonova, D. Zilli, J. Bernard, A. Matousek, What we do matters-a time-use app to capture energy relevant activities, Europ. Council Energy Effic. Economy Summ. Study Proceed. (2017) 2085–2093.

[20] S.D. Nazemi, M.A. Jafari, E. Zaidan, An incentive-based optimization approach for load scheduling problem in smart building communities, Buildings 11 (6) (2021) 237.
[21] D. Micci-Barreca, A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems, ACM SIGKDD Explorations Newsletter. 3 (2001) 27–32, https://doi.org/10.1145/507533.507538.
[22] J. Tian, K. Li, W. Xue, An adaptive ensemble predictive strategy for multiple scale electrical energy usages forecasting, Sustain. Cities Soc. 66 (2021), 102654, https://doi.org/10.1016/j.scs.2020.102654.
[23] W. Shi, N. Li, C.C. Chu, R. Gadh, Real-Time Energy Management in Microgrids, IEEE Trans. Smart Grid 8 (2017) 228–238, https://doi.org/10.1109/TSG.2015.2462294.
[24] M. Elkazaz, M. Sumner, E. Naghiyev, S. Pholboon, R. Davies, D. Thomas, A hierarchical two-stage energy management for a home microgrid using model predictive and real-time controllers, Appl. Energy 269 (2020), 115118, https://doi.org/10.1016/j.apenergy.2020.115118.
[25] W. Vermeer, G.R. Chandra Mouli, P. Bauer, Real-time building smart charging system based on PV forecast and Li-Ion battery degradation, Energies 13 (13) (2020) 3415.
[26] S.J. Pan, Q. Yang, A Survey on Transfer Learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345–1359, https://doi.org/10.1109/TKDE.2009.191.