A novel approach for identifying customer groups for personalized DSM services using household socio-demographic data

Wen, Hanguan; Liu, Xiufeng; Yang, Ming; Lei, Bo; Xu, Cheng; Chen, Zhe

Published in:
Energy

Link to article, DOI:
10.1016/j.energy.2023.129593

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Energy 286 (2024) 129593

Contents lists available at ScienceDirect

Energy

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

A novel approach for identifying customer groups for personalized demand-side management services using household socio-demographic data

Hanguan Wen [a,b], Xiufeng Liu [c,*], Ming Yang [d,*], Bo Lei [e], Xu Cheng [f], Zhe Chen [b]

[a] School of Electric Power Engineering, South China University of Technology, 510006 Guangzhou, China
[b] Department of Energy Technology, Aalborg University, 9000 Aalborg, Denmark
[c] Department of Technology, Management and Economics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
[d] College of Physics and Optoelectronic Engineering, Shenzhen University, 518060 Shenzhen, China
[e] School of Resource Environment and Safety Engineering, University of South China, 410083 Hengyang, China
[f] School of Computer Science and Engineering, Tianjin University of Technology, 300386 Tianjin, China

ARTICLE INFO

Keywords:
Residential energy consumption
Demand-side management
Load patterns
Feature engineering
Deep learning

ABSTRACT

Demand-side management (DSM) is crucial to smart energy systems. This paper presents a data-driven approach for understanding the relationship between energy consumption patterns and household characteristics to better provide DSM services. The proposed method uses a robust learning fuzzy c-Means clustering algorithm to automatically generate the optimal number of customer groups for DSM, and then uses symmetric uncertainty techniques to identify the identified load patterns and socio-demographic characteristics as the features for training a deep learning model. The model obtained can be used to predict the possibility of DSM group membership for a given household. This approach can be applied even in situations where smart meter data are not available, such as when new customers are added to the system or when residents change, or due to privacy concerns. The proposed model is evaluated comprehensively, including prediction accuracy, comparison with other baselines, and case studies for DSM. The results demonstrate the usefulness of weekly energy consumption data and associated household socio-demographic information for distinguishing between different consumer groups, the effectiveness of the proposed model, and the potential for targeted DSM strategies such as time-of-use pricing, energy efficiency measures, and demand response programs.

1. Introduction

Enhancing energy efficiency is pivotal for reducing carbon emissions and facilitating the transition to a low-carbon economy [1]. Ensuring the security and affordability of energy also plays a significant role in achieving climate objectives, such as the European Union's 2030 goals [2]. To meet these goals, targeted actions are required to boost energy efficiency, lower greenhouse gas emissions, and increase the adoption of renewable energy sources. Notably, in 2019, the building sector accounted for a substantial 22% of energy consumption in European countries and 19.7% globally [3]. Household socio-demographic factors are acknowledged as key influencers of energy consumption patterns [4]. In recent years, the widespread deployment of smart meters in residential buildings has generated substantial datasets. These meters provide detailed energy consumption records, typically at 30-minute intervals, enabling utilities to gain deeper insights into household consumption behaviors and offer personalized services [5]. Furthermore, analyzing daily load consumption data can shed light on household activities like cleaning, TV-watching, and cooking. This knowledge can empower households to understand their consumption habits and make changes to conserve energy. As a result, the study of daily electricity consumption can provide valuable insights for devising demand-side management (DSM) strategies, including time-of-use (ToU) pricing, energy efficiency (EE) programs, spinning reserve (SR), and demand response (DR) programs.

Machine learning is the best tool for mining consumer habits, particularly clustering techniques, which are among the most effective and popular methods available. Clustering algorithms are used to identify groups of similar objects or patterns, and can be applied to a wide range of data types, including time-series data such as daily load profiles. By applying clustering to smart meter data, utilities can identify customer groups with similar consumption patterns and offer targeted energy services based on customer needs and preferences. Additionally, clustering can be used to segment customer loads into

* Corresponding authors.
E-mail addresses: hawe@energy.aau.dk (H. Wen), xiuli@dtu.dk (X. Liu), mingyang@szu.edu.cn (M. Yang), hame88@usc.edu.cn (B. Lei), xu.cheng@ieee.org (X. Cheng), zch@energy.aau.dk (Z. Chen).

https://doi.org/10.1016/j.energy.2023.129593
Received 13 December 2022; Received in revised form 22 October 2023; Accepted 4 November 2023
Available online 6 November 2023

different time windows and then group them based on pattern similarity [6]. This can provide a more detailed understanding of each cluster and even individual energy use. However, most existing work in this area focuses on analyzing individual consumers using historical load data, rather than cluster analysis, which does not provide a good understanding of the underlying drivers of consumer behavior. Socio-demographic information refers to characteristics related to the spatial structure of buildings, economic income, household composition, social status, and the age of the house, among other factors [7,8]. From an economic perspective, the consumption behavior is largely determined by one's socio-economic characteristics [9], which means that having a good understanding of the socio-demographic information can help utilities to better understand the underlying reasons for their customers' consumption and make more accurate and tailored decisions about DSM and energy efficiency programs [10,11]. For DSM, smart meter data are commonly used to segment different customer groups based on their consumption patterns or intensity, and then provide targeted DSM services. However, obtaining energy consumption data can be a challenge in situations such as new buildings, changes in household residents, or constraints such as data privacy. In these cases, using socio-demographic information can be an effective way to estimate the groups of DSM services for a new customer. For example, utilities can use this information to recommend services to a customer, e.g., from the first day of moving into a new apartment. However, most existing work has been done to infer household socio-demographic characteristics by analyzing smart meter data, while much less work has been done to infer consumption based on household characteristics, especially in the several cases mentioned above. To bridge this gap, This paper presents a model for studying the relationship between residential energy consumption and household characteristics in order to provide targeted demand-side management (DSM) services. The proposed approach uses the Robust-learning Fuzzy c-Means (RL-FCM) algorithm to identify customer groups based on weekly load patterns, and uses symmetric uncertainty and Pareto analysis to extract significant features for each cluster for training a deep learning network for membership prediction across all the identified clusters. With the trained model, utilities can estimate the DSM membership probabilities for new customers. This study can help utilities to better manage energy demand and design targeted DSM strategies that are tailored to the specific needs of different consumer groups. By using the proposed method, utilities can identify representative load patterns with given household characteristics and predict which DSM services are likely to be effective for each customer. This can improve the efficiency and effectiveness of DSM programs and help utilities to better serve their customers.

In summary, The main contributions of our approach are:

- The challenge of delivering targeted Demand-Side Management (DSM) services in scenarios lacking smart meter data, such as when new customers are introduced or resident changes occur, or due to privacy concerns, has been identified.
- A machine learning model has been introduced to predict the likelihood of DSM membership for households based on their characteristics. This model offers utility providers the capability to provide customized DSM services upon the addition of a new customer to the system, irrespective of the availability of their energy consumption data.
- A comprehensive evaluation of the proposed model has been conducted, encompassing assessment of prediction accuracy, comparison with three baseline models, and a specific case study illustrating DSM application. The results obtained affirm the effectiveness of the proposed model.

The rest of this paper is organized as follows: Section 2 reviews the literature on clustering of load patterns and feature selection. Section 3 describes the proposed method, including the introduction of the RL-FCM algorithm, symmetric uncertainty, and the LSTM prediction model. Section 4 conducts the experiments to evaluate the proposed method. Finally, Section 6 concludes the paper and presents directions for future work.

2. Related work

This section begins with a review of state-of-the-art studies and techniques for identifying household load patterns and exploring customers' socio-demographic features. Subsequently, we delve into the process of feature selection.

2.1. Clustering of consumption profiles

For some years ago when smart meters were not available, customers were typically classified into different groups based on their household characteristics, such as building structure, socio-economic features, consumption habits, and attitudes towards energy use. These characteristics were typically identified through door-to-door surveys. However, the widespread rollout of smart meters has led to a shift in research on household clustering from an attribute-based approach to a consumption pattern-based approach. The availability of fine-grained consumption data has made it possible to perform more precise and innovative household clustering [12]. This has allowed researchers to identify more detailed and accurate load patterns and better understand the underlying drivers of consumer behavior.

Clustering is a commonly used technique in data analysis that involves finding groups in the data that have the highest similarity within the same cluster and the greatest variation between different clusters. The main methods currently used for clustering household electricity consumption include k-means [13], fuzzy k-means and fuzzy c-means (FCM) [14], probabilistic and generative models [15], hierarchical clustering [16], self-organizing maps [17], and DBSCAN [18]. Additionally, since smart meter data are continuously recorded over time, there are also clustering algorithms developed specifically for time series data, such as online [19] and dynamic [20] clustering. A number of relevant studies using clustering-based methods are listed in Table 1 and their use of smart meter data to investigate household consumption is described in detail.

Robust-learning Fuzzy c-Means (RL-FCM) is an improved version of the FCM algorithm that can automatically determine the optimal number of clusters [38]. The FCM algorithm uses fuzzy membership to assign each data sample to multiple clusters with different membership values, allowing for more accurate identification of data samples belonging to different clusters [39]. However, the conventional FCM algorithm requires manual post-processing methods, such as the elbow, silhouette coefficient, and gap statistic, to determine the optimal number of clusters, which can be time-consuming and computationally expensive [40–42]. To address this issue, the RL-FCM algorithm uses an adaptive learning rate and a robust error function to automatically determine the optimal number of clusters [38,43]. The adaptive learning rate allows the algorithm to converge quickly and avoid local minima, while the robust error function ensures that the clusters are well-separated and compact. In our study, the RL-FCM algorithm is employed to model real-world smart meter data and to ascertain the primary representative load patterns and the consumption distribution of each household.

2.2. Household load patterns and socio-demographic information

Predicting the DSM membership of new customers based on their socio-demographic information is a challenging and important task for utilities and retailers to design and implement customized DSM services. DSM membership is the category of load consumption pattern that a household belongs to, which reflects its electricity usage behavior and preferences. In this paper, we address this question by applying a machine learning approach. Previous studies have applied machine

Table 1
Studies on clustering of consumption data using different methods for different purposes.

Ref	Data size	Methods	Description and results
[4]	265 residential meters/surveys	Hierarchical	Using hierarchical clustering to identify 10 patterns of electricity consumption by the dataset in the city of Évora, Portugal, combined with daily electricity consumption from smart meters, and grouped into four different types of annual consumption curves, including:U-shaped (sharp and soft), W-shaped and flat.
[21]	220 K household data	K-means, Hierarchical	The consumers' lifestyle is captured by typical load shape, and five distinctive segmentation schemes enable selecting certain program development, pricing, and marketing purposes.
[22]	785 households	Hierarchical	Dividing daily consumption into six time intervals with different load shape characteristics as the daily features, and exploring the consuming peaks of each interval in a day. The cluster-specific results enable customers identified as potential targets for DSM, net metering, and hybrid programs.
[23]	269 peak-period load profiles	K-means, Hierarchical agglomeration	The raw data sample was considered with 5 p.m to 9 p.m (28 h) and be normalized. Then, the cumulative load profiles were clustered to capture their temporal variations in consumption patterns. Finally, the clustering results have been applied to several classification algorithms for predicting residential peak demand.
[24]	300 users consumption data	Hierarchical	Using the hierarchy tree to study the whole monthly consumption data samples. Nine abnormal users consuming behaviors were identified, and four representative monthly load patterns were obtained after the clustering.
[25]	103 homesuse data	Optimal k-means	The shape of seasonal profiles, an optimal number of clusters in each season, and the correlation between different profiles of the 103 homes were determined. The results found that the data fell into one of two seasonal groups, and some households use more expensive electricity (from the perspective of the wholesale electricity market) than others. It also suggests that some policies may have a more significant impact on low-income households during peak electricity consumption periods.
[26]	4963 households	Gaussian Mixture Model	Analyze the fine-grained temporal profiles to obtain the behavior features, which were used to identify consumers with homogeneous consumption profiles. After the analysis of Davies–Bouldin score, 15 clusters were identified and then summarized seven regular load profiles and one sparse group that was considered as the abnormal profile.
[27]	A household with 60 days consumption data	Hierarchical, FCM	The consumption patterns of the household appliances were analyzed by both clustering algorithms. The contextual features of an hour of the day, day of the week, and appliances have been extracted, and the household behaviors based on their house characteristics are revealed. The study can help consumers, and power companies understand the relationship between energy demand and support.
[28]	672 customers	FCM	By using the Hausdorff distance, a dynamic clustering algorithm based on the FCM is presented to study dynamic time series clustering, identification, and visualization of temporal load profiles. After quantitative analysis, the FCM-based algorithm provides a well-balanced cluster and load patterns, which can help companies quickly obtain the main consumption patterns and making decision support for consumers.
[29]	15,433 households	K-means	Electricity load patterns are broken down across the Danish region, and polynomial probability regression is used to examine household characteristics that have a significant impact on load patterns. Variations in the timing and size of electricity consumption are then taken into account, pointing to four household groups in Denmark that show similarities in terms of evening peaks in electricity consumption, seasonal variations in electricity demand, and rising demand at weekends.
[30]	3427 records and 235 feature values	Deep learning	Proposed a load forecasting method that uses aggregated smart meter data and a population dataset to predict customer electricity consumption, and the results find that the features significantly impact their consumption profiles.
[31]	845 English households	Regression analysis	By considering the multicollinearity in electricity consumption data, different types of features (socio-demographics, building factors, attitudes, self-reported behaviors, and appliance ownership and use) are tested to study which individual features have the most significant explanatory power in the region of non-heating electricity consumption. Thetextbackslash results show that appliance ownership and usage are the highest explanatory power variables.
[32]	3941 customers	Clustering, multinomial logistic regression	Investigate the load profiles of each day in six months to describe the weekends, intra-daily, and seasonal difference of domestic demand. A set of profile categories (PCs) is then identified, and each PC is linked to a corresponding household characteristic by applying multinomial logistic regression. The results show that it is possible to classify customers according to their personal characteristics in relation to their electricity use without knowing the household's electricity consumption beforehand.
[33]	29,393 buildings; 2,075,259 profiles in a single home	Clustering	A framework based on a deep auto-encoder and an adaptive self-organizing map (SOM) clustering algorithm is proposed to address the statistical analysis of the electricity consumption data and their features. The results show that the energy consumption levels can be plotted on a city map, and the related significant features also have been identified.
[34]	15,797 households	Time series analysis, Emploi du temps, EDT	The first quantifies the variation in the direct energy intensity, and non-energy expenditure intensity of daily activities and examines the extent to which the energy and non-energy intensity of activities are sensitive to household characteristics. Three dimensions of household differences have been explored: income, household composition, and housing type. The results show that income would impact energy consumption, while single-parent households would also become another potential trend for energy intensity consumption.

(*continued on next page*)

Table 1 (*continued*).

Ref	Data size	Methods	Description and results
[35]	4232 households	Classification,Multiple linear regression	A system using supervised machine learning methods for automatically estimating household "characteristics" that is related to socio-economic features based on their load patterns was proposed. The results show that the system achieves an accuracy of over 70% for most of the characteristics assessed in all households and even more than 80% for some characteristics.
[11]	4232 households	Support Vector Machine (SVM), Principal Component Analysis (PCA)	Propose a deep convolutional neural network (CNN) to extract features from the raw profiles, then the household characteristics were identified by a support vector machine model. Finally, the comparison of the method of the study with state-of-the-art machine learning techniques has been conducted. The results show CNN model can promote the accuracy of identifying socio-demographic information based on the load profiles.
[36]	4232 households	Discrete wavelet transform, SVM	Proposed a Time–Frequency Feature combination model which includes the discrete wavelet transform, random forest, and support vector machine to infer the household characteristics. The results show that the proposed method displays a better performance with the incorporation of frequency domain features.
[37]	4232 households	PCA, Deep learning	By using the federated learning technique, a model base-PCA for identifying distributed electricity households characteristics with the consideration of the privacy of retailers was proposed. Based on this, an artificial neural network is trained in a joint manner using three weighted averaging strategies to explore the relationship between consumer data and the corresponding social characteristics of users. The results show that by using PCA to extract features along with consumption load patterns, the recognition model achieves better performance.

learning techniques to explore the relationship between load consumption patterns and socio-demographic information. Some studies predicted household electricity consumption based on socio-demographic features [4]. Others identified different load consumption patterns using clustering algorithms and then inferred the socio-demographic characteristics of households [8,11]. We have summarized some of the related works in Table 1.

Our approach differs from existing works in two aspects. First, we use the Robust-learning Fuzzy c-Means (RL-FCM) clustering algorithm to identify diverse load consumption patterns, which is more robust and flexible than conventional clustering methods. Second, we use these patterns as labels for prediction, taking into account the socio-demographic characteristics of households. This enables us to classify new customers into potential customer groups and offer corresponding DSM services.

2.3. Features selection

Feature selection has been shown to be a useful technique for identifying and removing irrelevant and redundant features, enhancing the efficiency of the learning task, and improving the interpretability of the results [44]. In general, there are four main categories of feature selection methods: embedded, filter, wrapper, and hybrid approaches. The filter-based method takes into account the attributes of the features in the feature selection process. The filter-based feature selection process is independent of any classifier, and the correlation between features plays an important role in the multivariate approach [45]. Filtering approaches have higher computational performance compared to other categories [44]. Wrapper methods are a type of feature selection approach that involves training a model with a subset of features, and then evaluating its performance to determine which features to keep or discard [46]. In the wrapper approach, the feature selection process is embedded in a specific classifier and uses the classification accuracy as a criterion for the feature selection process. This means that the classifier is trained and tested repeatedly to select the most relevant features for the classification task [47]. The wrapper approach is computationally more expensive than the filter approach, but it can provide better results in some cases because it takes into account the interaction between the features and the classifier [48]. The hybrid approaches combine the advantages of different methods and are widely used in the field of feature selection. These approaches include the use of an embedded approach to pre-process the data, followed by a filter or wrapper approach to further refine the selected features. An example of a hybrid approach is the combination of a genetic algorithm with a wrapper method [49]. This approach has been shown to be effective in selecting the most relevant features and improving the performance of the learning task [49]. In the field of feature selection, symmetric uncertainty is a popular method used to measure the statistical independence between two feature values and between features and target classes [50]. This method has been applied in a variety of studies to identify effective features for different tasks and contexts [51–53]. In the context of electricity consumption and socio-demographic features, symmetric uncertainty has been used to identify the most relevant features for predicting household electricity consumption [54]. In this paper, symmetric uncertainty will be employed to identify effective socio-demographic features for different load patterns.

3. Method

Fig. 1 shows an overview of the proposed method. The proposed model uses a four-step approach to unravel the relationship between residential electricity consumption profiles and socio-demographic information. The first step involves obtaining and processing the original data into a suitable format. The second step involves using the Robust-Learning Fuzzy C-Means (RL-FCM) clustering algorithm to identify typical consumption patterns for all of the consumers' data, which can capture the optimal number of clusters for the consumption data. In the third step, feature engineering is used to identify a subset of decisive socio-demographic information for different consumption patterns in each cluster. In the final step, a tailored model is built to evaluate the mapping relationship between household consumption patterns and the selected socio-demographic information. This approach allows for a better understanding of the relationship between household electricity consumption and socio-demographic factors, which can help electricity providers to provide more tailored services and develop more effective policies for demand-side management.

3.1. Data preparation

The raw data contains both energy consumption data and household social background information, but not all of the data is complete. As a result, preprocessing is necessary. The consumption data has strong temporal variations within each day, making it difficult to capture with distinct dynamic conditions. In addition, as noted in a study by Wang et al. [55], the same household may use electricity differently on weekdays and weekends. As a result, it is advisable to analyze consumption characteristics at the weekly level. We denote the number of households as S, and the consuming data is divided

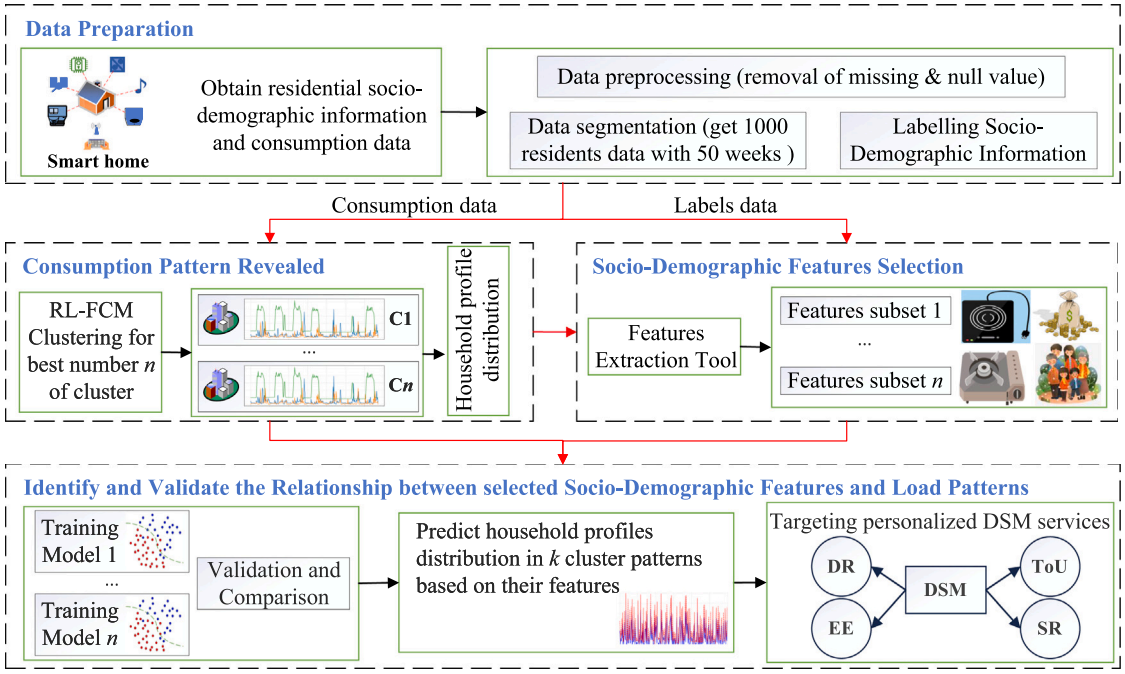

Fig. 1. Overview of the proposed framework.

Table 2
The Socio-Demographic information to be identified.

Feature	Question no.	Socio-Demographic information	F	Question no	Socio-Demographic information
F1	300	Age of the chief earner	F7	450	House type
F2	310	Chief earner has retired or not	F8	452	Rented of owned
F3	401	Social class of chief earner	F9	4531	Age of house
F4	410	Live alone	F10	460	Number of bedrooms
F5	420	Number of adult residents	F11	4704	Cooking facility type
F6	43111	Number of children residents	F12	4905	Energy efficient light bulb proportion

into weekly formats. We denote the total number of weeks in the load data as W, and each week is denoted as w ($w \in 1,2,\ldots,W$). Any weekly data with missing or formatting errors is removed. Finally, each household's weekly consumption data is complete, and each week w is split into 336 semi-hourly intervals (48 × 7), i.e., $t = 1,\ldots,336$. The consumption profile of household $s(s \in S)$ in semi-hourly interval t on week w is defined as $X^s_{w,t}$. Because $X^s_{w,t}$ is a scale of load profiles within 336 dimensions, we can ignore the normalization step and use the real data directly. In this study, twelve questions related to household socio-demographic information were identified and compared with existing literature. These questions are presented in Table 2, alongside the corresponding question numbers from the original questionnaire. The socio-demographic information encompasses both integer variables (e.g., "age of house") and categorical variables (e.g., "social class of chief earner"), which have been assigned integer labels. The twelve questions have been quantified and converted into ordinal classifications, as indicated in Table 3. However, it is worth noting that socio-demographic information is typically gathered in a single snapshot through door-to-door surveys, while consumption data is continuously recorded by smart meters. If the socio-demographic information of a household undergoes changes (e.g., due to alterations in income, household size, or retirement), this may disrupt the mapping between household consumption data and socio-demographic information. To mitigate this issue, consumption data for all households has been selected for up to 50 weeks, commencing from the date when residents completed the survey. It is assumed that socio-demographic information remains relatively stable over this period and is applicable to the selected consumption data.

3.2. Clustering consumption pattern by RL-FCM algorithm

The real world data often contains outliers and irregular data. As discussed earlier, the traditional Fuzzy C-Means (FCM) clustering algorithm is sensitive to noise, which can seriously affect the performance of clustering. In order to overcome this issue, manual techniques such as the elbow method, silhouette coefficient, and gap statistic are often used to identify the optimal number of clusters. However, these techniques can be challenging to implement and require significant technical expertise. In order to address this issue, a robust-learning FCM (RL-FCM) clustering algorithm was proposed in [38] that can automatically identify the optimal number of clusters without requiring parameter selection or initialization. In this study, we apply the RL-FCM algorithm to uncover typical consumption patterns in the preprocessed data from step 3.1. The RL-FCM algorithm is also free of the fuzziness index m, allowing for more flexible and efficient clustering.

In our study, we have defined the $X^s_{w,t}$ as the consumption profile of the household $s(s \in S)$ in semi-hourly t on week w, and the number n of all the consumption data weekly is determined as $n = 50 \times S$. We let $X = \{x_1,\ldots,x_n\}$ be the n consumption sample data in a t-dimensional Euclidean space $\mathbb{R}^d$, and $V = \{v_1,\ldots,v_c\}$ be the c cluster centers. The Euclidean distance measure is used to calculate the distance between every pair of load profiles, with $d_{ik} = |x_i - v_k|2 = \sqrt{\sum j = 1^d \left(x_{ij} - v_{kj}\right)^2}$.

Compare with the conventional fuzzy c-means (FCM) algorithm [56], the objective function of the RL-FCM adds several entropy terms to free the fuzziness index m, adjust bias, and find the best number of clusters. Specifically, the entropy term of membership with

Table 3
The quantified ordinal classification of the features.

F1	Label value	F2	Label value	F3	Label value	F4	Label value
18 – 25	1	An employee	1	AB	1	Yes	1
26 – 35	2	Self-employed (with employees)	2	C1	2	No	2
36 – 45	3	Self-employed (with no employees)	3	C2	3		
46 – 55	4	Unemployed (actively seeking work)	4	DE	4		
56 – 65	5	Unemployed (passively seeking work)	5	F [RECORD ALL FARMERS]	5		
65+	6	Retired	6				
F5	**Label value**	**F6**	**Label value**	**F7**	**Label value**	**F8**	**Label value**
1	1	0	1	Apartment	1	ent (from a private landlord)	1
2	2	1	2	Semi-detached house	2	Rent (from a local authority)	2
3	3	2	3	Detached house	3	Own Outright (not mortgaged)	3
4	4	3	4	Terraced house	4	Own with mortgage etc	4
5	5	4	5	Bungalow	5	Other	5
6	6	5	6				
7 or more	7	6 or more	7				
F9	**Label value**	**F10**	**Label value**	**F11**	**Label value**	**F12**	**Label value**
< 5 years	1	1	1	Electrical	1	None	1
5 – 10 years	2	2	2	Not electrical	2	About a quarter	2
10 – 30 years	3	3	3			About half	3
30 – 75 years	4	4	4			About three quarters	4
> 75 years	5	5+	5			All	5

$\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}\ln\mu_{ik}$ is used to replace the fuzziness index m and adjust bias. Next, the mixing proportion $\alpha=(\alpha_1,\ldots,\alpha_c)$ of clusters is applied, where α_k is the probability that one sample point belongs to the kth cluster under the given binding $\sum_{k=1}^{c}\alpha_k=1$. Thus, $-\ln\alpha_k$ represents the information in the event that a data sample point belongs to the kth cluster, and the entropy term $\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}\ln\alpha_k$ is used to summarize the average information under each data sample point belonging to the corresponding cluster. Furthermore, the entropy term $\sum_{k=1}^{c}\alpha_k\ln\alpha_k$ is introduced to represent the average information for the occurrence of each data point belonging to the corresponding cluster. Finally, the updated RL-FCM objective function is as follows:

$$J(\mathbf{U},\alpha,\mathbf{V})=\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}d_{ik}^2-r_1\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}\ln\alpha_k+r_2\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}\ln\mu_{ik}-r_3n\sum_{k=1}^{c}\alpha_k\ln\alpha_k \tag{1}$$

where $r_1,r_2,r_3\geq 0$ for adjusting bias. As for the question of how to study the values of the parameters r_1, r_2, and r_3 for the three entropy penalty terms $\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}\ln\alpha_k$, $\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}\ln\mu_{ik}$ and $\sum_{k=1}^{c}\alpha_k\ln\alpha_k$, respectively, the original paper [38] had rigorous assumptions and derivations, so particularly interested readers can view the detail process. The results of r_1, r_2, and r_3 are used as follows:

$$r_1^{(t)}=e^{-t/10} \tag{2}$$

$$r_2^{(t)}=e^{-t/100} \tag{3}$$

$$r_3=\min\left(\frac{\sum_{k=1}^{c}\exp\left(-\eta n\left|\alpha_k^{(\text{new})}-\alpha_k^{(old)}\right|\right)}{c},\frac{1-\max_{1\leq k\leq c}\left(\frac{1}{n}\sum_{i=1}^{n}\mu_{ik}\right)}{\left(-\max_{1\leq k\leq c}\alpha_k^{(old)}\sum_{t=1}^{c}\alpha_t^{(old)}\ln\alpha_t^{(old)}\right)}\right) \tag{4}$$

where $\eta=\min\left\{1,2/t^{\lfloor t/2-1\rfloor}\right\}$, the t is the dimensions of the data and the notation $\lfloor g\rfloor$ denotes the maximum integer that is less than or equal to g. The updating equation for the RL-FCM objective function in (1) with respect to v_k and μ_{ik} are as Eqs. (5) and (6), respectively:

$$v_k=\frac{\sum_{i=1}^{n}\mu_{ik}x_i}{\sum_{i=1}^{n}\mu_{ik}} \tag{5}$$

$$\mu_{ik}=\exp\left(\frac{-d_{ik}^2+r_1\ln\alpha_k}{r_2}\right)\Big/\sum_{t=1}^{c}\exp\left(\frac{-d_{ik}^2+r_1\ln\alpha_t}{r_2}\right) \tag{6}$$

Then the updating equation for α_k can be obtained as follows:

$$\alpha_k^{(\text{new})}=\frac{1}{n}\sum_{i=1}^{n}\mu_{ik}+\frac{r_3}{r_1}\alpha_k^{(\text{old})}\left(\ln\alpha_k^{(\text{old})}-\sum_{t=1}^{c}\alpha_t^{(\text{old})}\ln\alpha_t^{(\text{old})}\right). \tag{7}$$

To address the initialization problem, we assign all the data points as initial clusters for the first iteration. Specifically, $c^{(0)}=n$ and $\alpha_k^{(0)}=1/c=1/n, k=1,\ldots,c$. Obviously, there is a comparison and competition between the mixing proportions as in Eq. (7), which can drive the iteration to proceed. In the course of the iteration, the RL-FCM algorithm can discard illegitimate mixing proportion $\alpha_k^{(\text{new})}$ if $\alpha_k^{(\text{new})}<1/n$. Then, we can obtain the updated cluster $c^{(\text{new})}$ as

$$c^{(\text{new})}=c^{(\text{old})}-\left|\left\{\alpha_k{}^{(\text{new})}\mid\alpha_k{}^{(\text{new})}<1/n,k=1,2,\ldots,c^{(\text{old})}\right\}\right| \tag{8}$$

where $|\{\}|$ represents the cardinality of the set$\{\}$. After updating the number of clusters c, the remaining mixing proportion α_k^* and the corresponding μ_{ik}^* is normalized by

$$\alpha_k^*=\frac{\alpha_k^*}{\sum_{t=1}^{c^{\text{new}}}\alpha_t^*} \tag{9}$$

$$\mu_{ik}^*=\frac{\mu_{ik}^*}{\sum_{t=1}^{c^{\text{new}}}\mu_{it}^*} \tag{10}$$

where are subject to $\sum_{t=1}^{c^{\text{new}}}\alpha_t^*=1$ and $\sum_{t=1}^{c^{\text{new}}}\mu_{it}^*=1$. Therefore, the concept of Eq. (8) can be used to evaluate the best number of clusters c^*.

Based on the above analysis, we can know that the total computational complexity for the RL-FCM algorithm is the same as that of the conventional FCM, with $O\left(nc^2t\right)$, where n is the total number of sample data points, c is the number of clusters, and t is the dimensionality of the data. However, we observed that even though the RL-FCM algorithm uses the number of data points n as the initial number of clusters c (i.e., $c=n$) in the early stages of iterations, clusters with $\alpha_k\leq 1/c=1/n$ can be ignored during subsequent iterations. This allows for a significant reduction in the time per iteration after a few iterations, providing an advantage over the traditional FCM algorithm.

Algorithm 1: Robust-learning Fuzzy c-Means clustering

Input : The total weekly consumption profile $\boldsymbol{X} \leftarrow \{x_1, \cdots, x_n\}$, the households Id set, threshold $\epsilon \leftarrow 10^{-6}$

Output: k representative load patterns sets $\hat{\boldsymbol{X}}_k$ and corresponding centroid sets $\hat{\boldsymbol{v}}_k$, the distribution matrix $\boldsymbol{p_{sk}}$

Initialize: $c^{(0)} \leftarrow n$, $\alpha_k^{(0)} \leftarrow 1/n$, $r_1^{(0)} = r_2^{(0)} = r_3^{(0)} \leftarrow 1$, $\alpha_k^{(0)} \leftarrow 1/n$, $v_k^{(0)} \leftarrow x_i$, t$\leftarrow$ 1, iteration times $T \leftarrow 100$
Calculate the $\mu_{ik}^{(t)}$ by using the $c^{(t-1)}$, $r_1^{(t-1)}$, $r_2^{(t-1)}$, $v_k^{(t-1)}$, $\alpha_k^{(t-1)}$ as Eq. (6)
Update $r_1^{(t)}$ and $r_2^{(t)}$ according to Eqs. (2) and (3), respectively
Update $\alpha_k^{(t)}$ with $\mu_{ik}^{(t)}$ and $\alpha_k^{(t-1)}$ by Eq. (7)
Update $r_3^{(t)}$ with $\alpha_k^{(t)}$ and $\alpha_k^{(t-1)}$ according to Eq. (4)
Update $c^{(t-1)}$ to $c^{(t)}$ by ignoring those clusters under $\alpha_k^{(t)} \le 1/n$, then normalize $\alpha_k^{(t)}$ and $\mu_{ik}^{(t)}$ by the Eqs. (9) and (10), respectively.
if $t \ge T$ *and* $c^{(t-T)} - c^{(t)} \leftarrow 0$ **then**
    $r_3^{(t)} \leftarrow 0$
Update $v_k^{(t)}$ with $c^{(t)}$ and $\mu_{ik}^{(t)}$ by Eq. (5)
Compare $v_k^{(t)}$ and $v_k^{(t-1)}$
if $\max_{1 \le k \le c^{(t-1)}} \left\| v_k^{(t)} - v_k^{(t-1)} \right\| < \varepsilon$ **then**
    Record each profiles x_i to corresponding $\hat{\boldsymbol{X}}_k$;
    $\hat{\boldsymbol{v}}_k = v_k^{(t)}$;
    Compute the distribution of the number of each household s profiles in kth cluster by corresponding Id as $\boldsymbol{p_{sk}}$;
    Break;
else
    t++, repeat to **Step 1** for calculating the $\mu_{ik}^{(t)}$
return $\hat{\boldsymbol{X}}_k$, $\hat{\boldsymbol{v}}_k$, $\boldsymbol{p_{sk}}$

According to algorithm 1, the p_{sk} represents the proportion of consumption profiles of household s in the kth representative load pattern, as expressed in Eq. (11). This value can provide insight into the electricity consumption habits of household s across different load patterns. Once the RL-FCM clustering is finalized, feature engineering techniques will be implemented to discern subsets of features corresponding to each clustering pattern.

$$p_{sk} = \frac{All_{s,k}}{50}, \tag{11}$$

where $All_{s,k}$ presents the total of load profiles of customer s in the kth cluster.

3.3. Feature selection process

In our study, we have collected socio-demographic information on households from the survey. However, the smart meter data contain outliers and irregularities, and the corresponding feature records have low accuracy and contain redundant and irrelevant information. To tackle this issue, feature selection techniques based on feature engineering will be employed to eliminate redundant features. Using the clustering load patterns obtained via RL-FCM, the application of symmetric uncertainty methods will facilitate the selection of valuable features from Table 2. This process will result in the extraction of precise feature sets corresponding to each cluster profile $\hat{\boldsymbol{X}}_k$. These feature sets will subsequently play a pivotal role in predicting the consumption distribution for each household, a critical indicator for DSM. First, we introduce Mutual information (MI) [57]. Mutual information $MI(F, P)$ is a measure of how well two factors F and P are correlated. It is defined as:

$$MI(F; M) = \sum_{f \in F} \sum_{p \in P} P(f, p) \log \frac{P(f, p)}{P(f) P(p)} \tag{12}$$

where $P(f)$ and $P(p)$ represent the marginal probabilities of the variables $f(f \in F)$ and $p(p \in P)$, respectively. $P(f, p)$ is the probability of simultaneous occurrence of f and p. The value of $MI(F, P)$ is non-negative and the larger the value of $MI(F, P)$, the stronger the dependency between f and p. However, using the results of MI directly for feature selection can introduce bias [58]. Therefore, it is necessary to normalize the results of $MI(F, P)$ to allow for comparison of the correlation between f and p under unbiased conditions. One of the common normalization methods for mutual information is symmetric uncertainty (SU) [59], which allows the results of MI to be normalized within the range [0, 1] and is defined as:

$$SU(F, P) = \frac{2 \times MI(F; P)}{H(F) + H(P)} \tag{13}$$

where $H(X)$ is the Shannon entropy for quantifying the information degree of X and is calculated using the marginal probabilities $p(x)$ as:

$$H(X) = -\sum_{x \in X} P(x) \log P(x) \tag{14}$$

A value of 1 in $SU(F, P)$ indicates that f and p are perfectly correlated, i.e., the value of one variable fully predicts the value of the other variable. However, a value of 0 indicates that the two features are completely independent.

In this study, we use the results $\boldsymbol{p_{sk}}$ from Algorithm 1 and all quantified socio-demographic feature sets $Fi(i \in 1, \ldots, 12)$ from Table 3 to compute the $SU(F, p_{sk})$ for each set Fi and p_{sk} in the kth cluster. We apply Pareto analysis, a multi-objective optimization technique, to select the most important features for each cluster. For example, in the jth cluster, we rank all of the computed results of the twelve Fi and count a cumulative distribution function for the top 80%. Finally, these factors are considered as elements for the selected subsets F^j. We summarize all of the steps in this subsection in Algorithm 2. The k selected feature subsets are obtained as F^j ($F^j \in$ FF, and $j \in 1, \ldots,$ k) and will be used as vital indicators for predicting the consumption pattern distribution of all households within a given cluster.

Algorithm 2: The feature selection process

Input : The $\mathbf{p_{sk}}$ from Algorithm 1, total qualified feature set **F** from Table 3, a null set **FF**

Output: The K selected feature subsets F^j

for *cluster* $j \leftarrow 1$ *to* K **do**
    Calculate the MI(F, p_{sj}) by Eq. (12)
    Calculate the SU$^{\text{normalized}}(F, p_{sj})$:
    for $i \leftarrow 1$ *to* 12 **do**
        $SU^{\text{normalized}}(F_i, p_{sj}) \leftarrow \frac{SU(F_i, p_{sj})}{\sum_{i=1}^{12} SU(F_i, p_{sj})}$; `// Normalize 12 socio-demographic features within a cluster`
    Pareto analysis:
    $F^j \leftarrow \arg\max \left\{ CDF_{80\%} \text{ ranking set} \left\{ SU^{\text{normalized}}(F_i, p_{sj}) \mid i = 1 \text{ to } 12 \right\} \right\}$; `// CDF depicts the weight distribution`
    Record F^j to the **FF**
Return selected feature subsets **FF** for the k clusters

3.4. Analyzing the relationship between consumption patterns and selected features

We have developed an analytical model based on Long Short-Term Memory (LSTM) to predict households' consumption distributions using the selected feature subset F^j. These feature subsets F^j for each cluster load profile were derived from the previously mentioned section on symmetric uncertainty. Subsequently, we construct k distinct LSTM models with F^j as the input training data, and employ $\boldsymbol{p_{sj}}$ within the j cluster as the predicted results for consumption distribution. The LSTM analytical model comprises an LSTM layer, a fully connected layer, and a softmax layer. The softmax layer serves for classification, mapping the network output to the (0,1) range to predict the probability of the consumption distribution for all users within each cluster. Taking the jth LSTM as an example, in this model, our goal is to map the

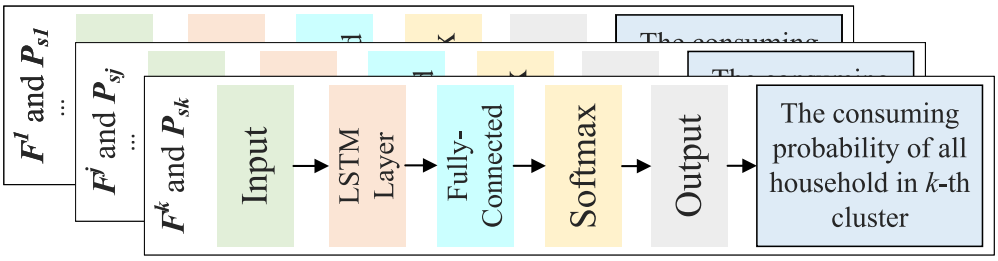

Fig. 2. The structure of LSTM model for forecasting the consumption distribution of each household in kth cluster.

labeled values in subset F^j with several selected features Fi to the consumption distribution p_{sj}. The LSTM neural network at time t is defined as follows: $\mathbf{h}t = \text{LSTM}(\mathbf{h}t-1, \mathbf{x}_t)$

$\mathbf{o}_t = \text{ReLU}(\mathbf{W}_o\mathbf{h}_t + \mathbf{b}_o)$

$\mathbf{y}_t = \text{Softmax}(\mathbf{W}_y\mathbf{o}_t + \mathbf{b}_y)$ where $\mathbf{h}_t$ is the hidden state of the LSTM at moment t, $\mathbf{x}_t$ is the input at moment t, $\mathbf{o}_t$ is the output of the fully connected layer, and $\mathbf{y}_t$ is the predicted probability of the consumption distribution of all users in the jth cluster. $\mathbf{W}_o$ and $\mathbf{W}_y$ are the weight matrices, and $\mathbf{b}_o$ and $\mathbf{b}_y$ are the bias vectors. The ReLU and Softmax functions are applied element-wise.

In our study, we employ the softmax function, a popular normalization algorithm, to ensure that the prediction probabilities of each household in the k clusters sum to 1. The softmax function is defined as:

$$p^*_{sj} = \frac{e^{p_{sj}}}{\sum_{j=1}^{k} e^{p_{sj}}}$$

This equation maps the predicted probability of household s's consumption profile in the jth cluster to a nonlinear, non-negative value within the interval [0, 1]. The resulting p^*_{sj} values are used to calculate the loss function for each LSTM model. To evaluate the performance of our proposed model, we conducted a comparative analysis with three baselines that represent different approaches to customer group membership prediction:

- Baseline 1 is a supervised learning method that used the same LSTM structures as our proposed model (see Fig. 2). The input for this baseline was the full set of socio-demographic features from Table 3. This baseline was used to evaluate the effects of feature selection on prediction accuracy.
- Baseline 2 is an unsupervised learning method that used Extreme Learning Machine (ELM) structures instead of LSTM structures. The input for this baseline was also a subset of socio-demographic features selected by our proposed model. This baseline was used to compare ELM and LSTM as methods for customer group membership prediction.
- Baseline 3 is also an unsupervised learning method that used ELM structures. However, unlike Baseline 2, this baseline used the full set of socio-demographic features from Table 3 as input. This baseline was used to evaluate the effects of feature selection on ELM-based prediction.

The Root Mean Square Error (RMSE) is used as the performance metric. This metric quantifies the deviation between the predicted probabilities p^*_{sj} and the actual probabilities p_{sj} for each household s in cluster j. The RMSE is calculated as follows:

$$RMSEj = \sqrt{\frac{1}{S'}\sum_{s=1}^{S'}\left(p_{sj} - p^*_{sj}\right)^2}, \tag{15}$$

where S' is the number of households used for training the model, and is typically 30% of the total number of households S. The training process stops when the RMSE falls below a certain threshold or the number of iterations exceeds a certain threshold.

Overall, our proposed model and the baselines will be trained and evaluated on the same data to provide a fair comparison of their performance. The results of this comparison will help us to understand the effectiveness of LSTM, ELM, and feature selection in predicting the consumption distribution of households in different clusters.

4. Experiments

The experimental results of this study were obtained using the Commission for Energy Regulation (CER) dataset, which contains data on 4232 residential households in Ireland. The dataset includes 536 daily consumption records at a semi-hourly interval, as well as pre- and post-trial surveys that provide socio-demographic information about the households. To prepare the data for our experiments, we first reconstructed the consumption data in weekly format. Any null values or consecutive zero values within a week's time were deleted, along with the corresponding socio-demographic information. In addition, households with incorrect or missing socio-demographic information were also removed from the dataset. After this preprocessing step, we were left with 50 weeks of consumption data from 1000 residential households, along with their socio-demographic information. This data was used to evaluate the proposed model.

4.1. Representative pattern identification

We use Algorithm 1 to cluster the weekly load profiles of households and obtain eight clusters shown in Fig. 3. Each cluster has a distinctive distribution of consumption data and a unique consumption pattern. These patterns show different dynamics between weekdays and weekends, which reveals the importance of analyzing household electricity consumption on a weekly basis. In this paper, we use the term load profile to refer to the electricity consumption data of a household over a period of time. We also use the term demand-side management (DSM) to describe the actions taken by consumers or utilities to modify the electricity demand. The data we used for clustering is from the Commission for Energy Regulation (CER) smart metering project in Ireland.

The eight clusters (C1-C8) show different consumption patterns, which indicate that it is better to analyze household electricity on a weekly basis rather than daily. For example, C1 has totally flat consumption profiles, which means those households in C1 do not vary their electricity consumption according to time. C2 has dual peaks on weekend mornings and evenings, while only having a relatively higher consumption on weekday evenings. C3 has a peak on weekday evenings and dual peaks on Saturday mornings and evenings. C4 has a stable consumption during weekday mornings and afternoons, but rapidly increases to a peak in the evening, similar to Saturdays. However, Sundays have lower consumption in the afternoon and a peak in the evening. C5 has a relatively high level of consumption on weekdays and Saturdays, with obvious dual peaks in the morning and evening on Saturdays. Sundays have the highest peak of consumption within the cluster. C6 has a similar consumption pattern to C4, but with some differences, such as a small peak in the morning on weekdays and a

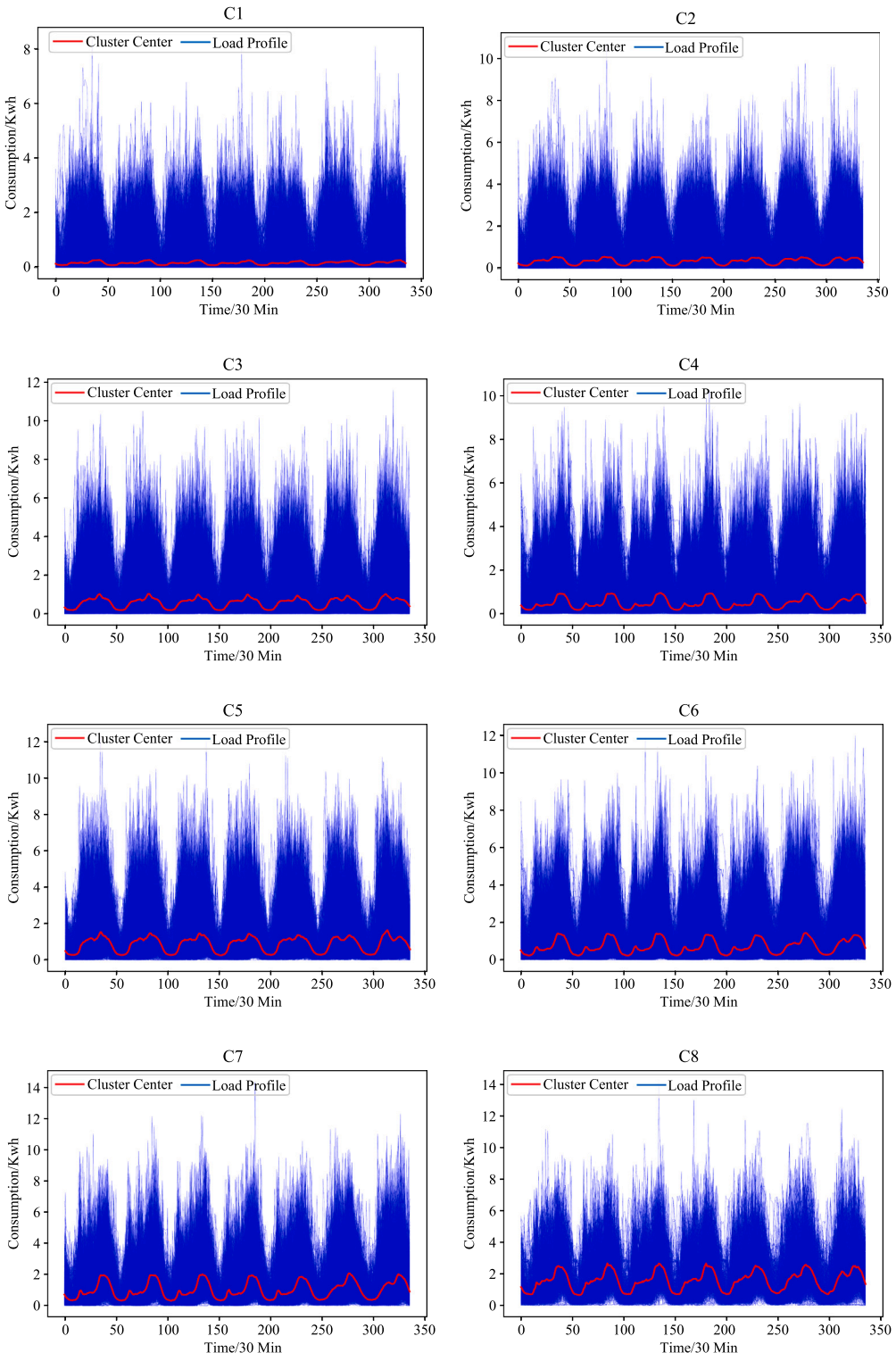

Fig. 3. Eight typical load patterns.

relatively sharp consumption curve. C7 has a small peak in the morning and a higher peak in the evening on weekdays. However, there is only a single peak on weekends. C8 has relatively sharp consumption dynamics and obvious dual peaks on Sunday mornings and evenings, but the weekdays and Saturdays are not significant. The analysis of the eight clusters reveals that the RL-FCM algorithm performs well in clustering the CER data and can be used to make individual DSM decisions based on the results of the clustering.

Fig. 4 shows the distribution of load profiles in each cluster, which reflects the prevalence of each consumption pattern within the dataset. The x-axis enumerates the eight clusters (C1 through C8), while the y-axis indicates the proportion (%) of load profiles that fall into each cluster. Each cluster represents a group of households that share similar electricity consumption patterns. For instance, Cluster C2, which has the highest proportion of 21.42%, represents the most common consumption pattern among the households in the dataset, while Cluster

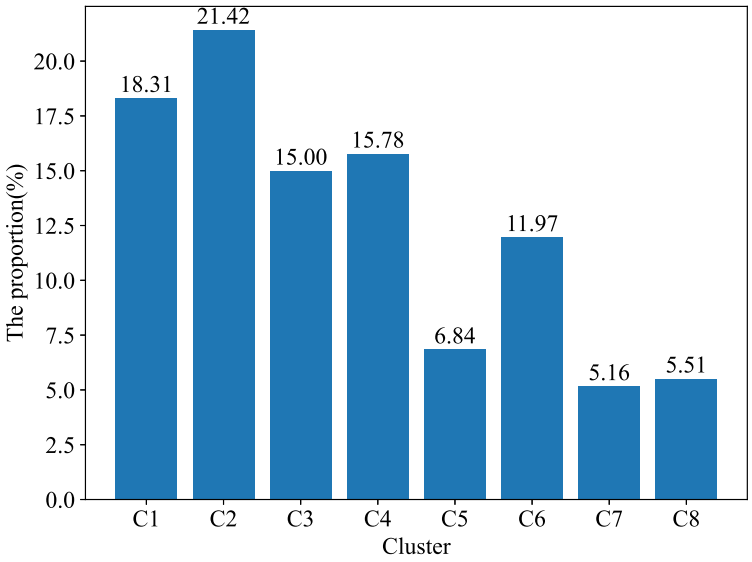

Fig. 4. The load profiles distribution of each cluster.

C7, which has the lowest proportion of 5.51%, represents the least common one. This clustering result is significant as it enables individualized DSM decisions. By understanding the distinct consumption patterns in each cluster, we can tailor DSM strategies to effectively manage energy use within each group. These clusters provide a solid foundation for further exploration. The next step involves applying a Long Short-Term Memory (LSTM) model to delve into the non-linear relationship between socio-demographic information and household consumption patterns. This approach promises to reveal nuanced insights into how different socio-demographic factors influence energy consumption within each cluster.

4.2. Features comparison and selection

In this subsection, we identify highly relevant socio-demographic characteristics for each cluster (each representative loading pattern). We perform a Pareto analysis of the eight patterns obtained using Algorithm 2 to obtain the results in Fig. 5. We plot the cumulative percentage lines and mark the features that exceed 80% with a red vertical dashed line. According to the results, $F5$ has the highest symmetric uncertainty score in all loading models except C8, where $F10$ is the highest. $F5$ represents the "number of adult residents", while $F10$ represents the "number of bedrooms" (see Table 2). Thus, the number of adults is the most important sociodemographic characteristic. In contrast, $F11$, $F6$ and $F8$ contribute the least to the eight representative loading patterns, which are 'type of cooking facilities', 'number of child residents' and 'renting or owning', respectively. According to the results, $F2$ is a highly influential characteristic, namely the retirement status of the primary income earners. It can be seen that some sociodemographic information plays an important role and can be used to predict the distribution of loading patterns of households. In addition, Pearson correlation analysis was used to gain insight into the pairwise correlations between each socio-demographic characteristic, as well as the correlations between these characteristics and the clustering loading patterns. The scale of the coefficients indicates the strength of the relationship, while the symbols indicate positive and negative correlations. The results are shown in Fig. 6, where the more red, the more positive the correlation, while the more purple, the more negative the correlation. For example, $F5$, "number of adult residents", has a significant negative correlation with $F1 - F3$, C1, and C2, indicating that it is difficult to conveniently identify adults with more complex resident conditions, given their increasing age, retirement, and income earner status. Furthermore, we can see that $F11$ is neutral for all socio-demographic information as well as for all load patterns, suggesting that there is no significant relationship between the type of cooking facilities and households' electricity consumption patterns. This may suggest that electricity use for cooking is not a major part of their electricity use, reflecting the diversity. The results obtained from Pearson correlation analysis serve as a robust validation of the symmetric uncertainty and Pareto analysis techniques. Nevertheless, some variation is observed within our feature selection process. For instance, $F4$ pertains to households living in the neighborhood, which logically aligns with $F5$ and $F6$, as indicated by the Pearson correlation analysis, given their clear association with resident population. However, as outlined in our approach in Section 3.3, our objective is to minimize the presence of redundant features while retaining the utmost relevant ones. Hence, a feature selection method based on symmetric uncertainty remains an appropriate and prudent technique for our purposes.

4.3. Prediction performance and compare with baselines

Based on the correlations in Fig. 5, we determine the socio-demographics as features and the corresponding consumption patterns as the labels. We randomly select 70% of the households as training and the remaining 30% as testing. The model uses the RMSE in Eq. (15) as the cost function with optimized hyperparameters. Fig. 7 shows the prediction results compared to the ground truth with 8 clusters, which shows that our model can achieve favorable prediction performance.

To evaluate our method, a comparison is made against three baselines, as discussed in Section 3.4, and Eq. (15) is employed to quantify the error for all methods. The results are presented in Table 4. Baseline 1 exhibits relatively strong performance using our predictions without the feature selection process. In the prediction models corresponding to various cluster loads, such as C4, C5, and C7, our model demonstrates similar performance. However, baseline 2, which incorporates the ELM prediction model along with our proposed feature selection process, performs notably worse than baseline 1. Moreover, baseline 3, which relies on the ELM prediction model without taking feature selection into account, displays inferior performance. Our model achieves a substantial reduction in the average RMSE of 52.5287%, 85.5859%, and 86.0828% when compared to baselines 1, 2, and 3, respectively. Across the eight clusters, it is evident that our model more effectively captures the nonlinear relationship between cluster patterns and the corresponding socio-demographic characteristics.

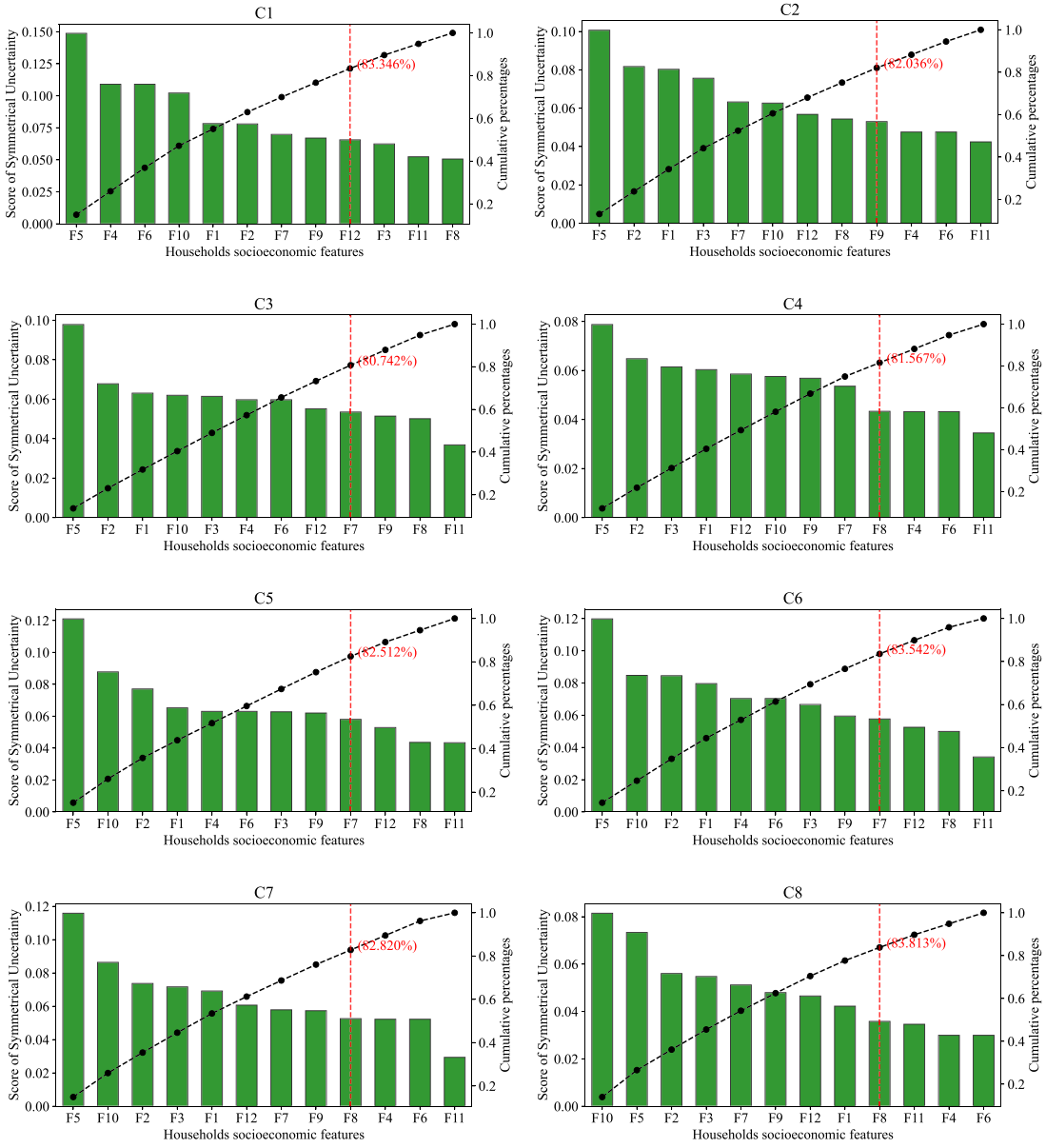

Fig. 5. The eight selected feature subsets.

Table 4
The RMSE comparison of our model and other baselines.

Cluster	RMSE comparison				Improvement of proposed model (%)		
	Our model	Baseline1	Baseline2	Baseline3	Baseline1	Baseline2	Baseline3
C1	0.0684	0.1865	0.3270	0.3255	63.3480	79.0948	79.0034
C2	0.0476	0.2518	0.3476	0.2864	81.1006	86.3107	83.3874
C3	0.0676	0.1510	0.6250	0.4491	55.2295	89.1845	84.9487
C4	0.1415	0.1645	0.3599	0.4954	13.9609	60.6783	71.4335
C5	0.0228	0.0657	0.6447	0.6656	65.2937	96.4618	96.5727
C6	0.0799	0.1224	0.5243	0.4882	34.7272	84.7640	83.6363
C7	0.0537	0.0817	0.6026	0.6749	34.3389	91.0946	92.0485
C8	0.0176	0.0634	0.6068	0.7435	72.2305	97.0986	97.6321
Average	0.0624	0.1359	0.5047	0.5161	52.5287	85.5859	86.0828

4.4. Customer group membership prediction for DSM

The provided household characteristics have been employed as input variables for the prediction of group membership within the context of Demand-Side Management (DSM). In this evaluation, a random selection of four households, specifically labeled as #2093, #2270, #2333, and #2491, has been utilized as the test dataset. The socio-demographic attributes of these selected households are comprehensively outlined in Table 5 for reference.

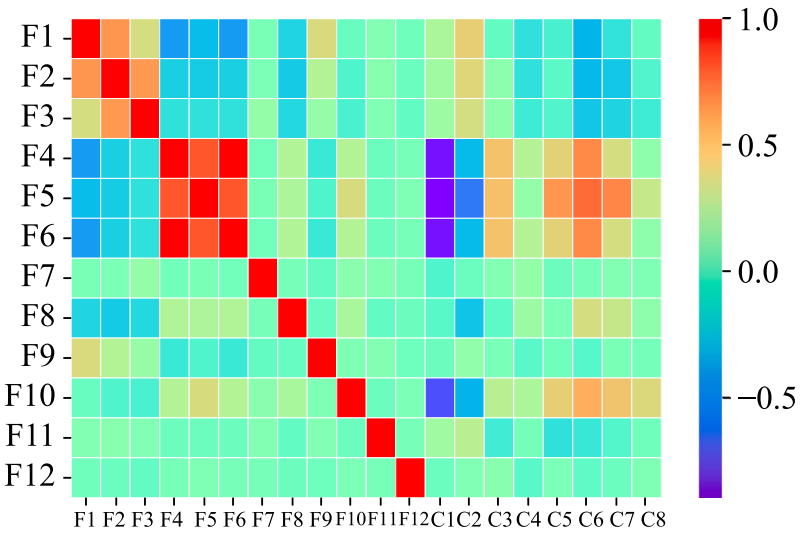

Fig. 6. The Pearson correlation coefficient between the features and patterns.

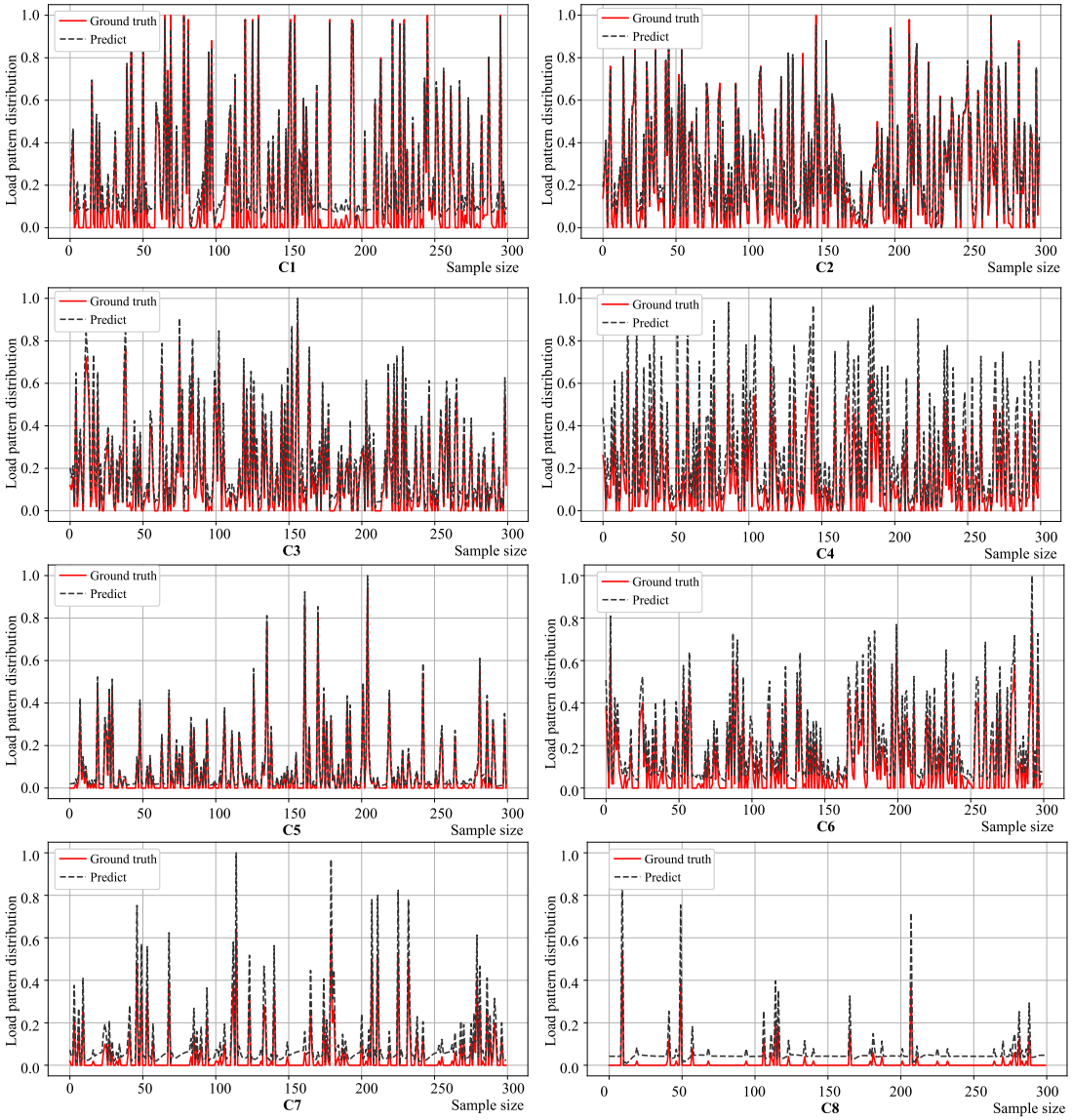

Fig. 7. The eight test results for validating our model.

Table 5
The socio-demographic information of four selected households.

ID	F1	F2	F3	F4	F5	F6	F7	F8	F10	F9	F11	F12
2093	2	1	1	2	2	2	4	4	2	1	2	4
2270	6	3	2	2	2	2	2	3	4	3	2	5
2333	6	6	4	1	1	1	4	2	3	3	2	3
2491	3	1	3	2	3	3	2	4	4	3	1	4

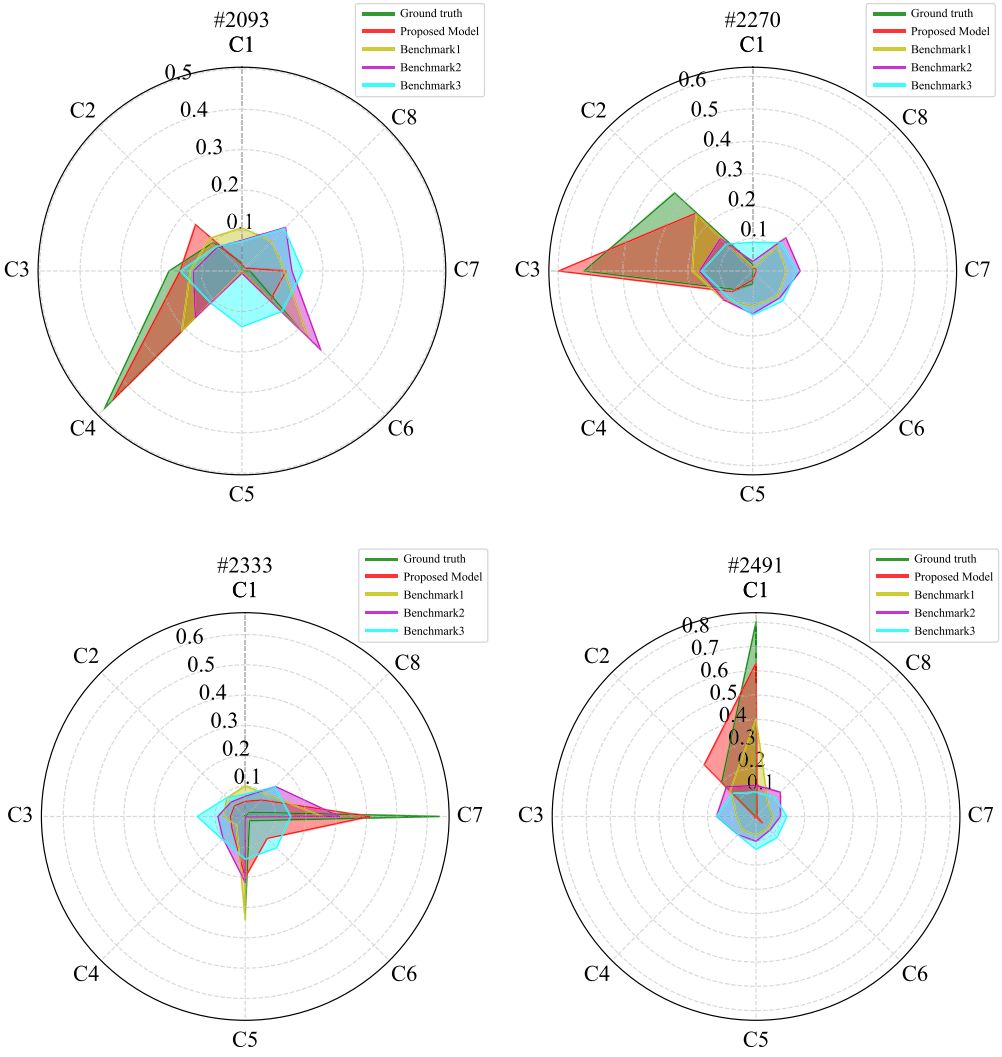

Fig. 8. Comparison of our model and other baselines in four real households consumption.

In general, there is a difference in twelve sociodemographic characteristics among the four households. Next, we performed the customer group membership prediction using our proposed model and compared it with the other three baselines. The prediction accuracy is presented in the form of radar plots in Fig. 8. It can be observed that our proposed model achieved the highest prediction accuracy for the four selected households, i.e., the red dashed line (our proposed model) is closest to the green one (ground truth), compared to the yellow, purple, and cyan dashed lines (baselines 1–3) in each radar plot. Baselines 2 and 3, however, showed poor prediction accuracy, as evidenced by the large gaps between their lines and the ground truth in the four radar plots. After conducting an exploration of the mapping relationship from household socio-demographic information to the primary load pattern, insights and recommendations can be offered to utilities for the development of more refined DSM services.

4.5. Insights load patterns-based for DSM

As mentioned earlier, DSM strategies can be classified into EE, ToU, SR and DR. By carefully planning and implementing these activities, these strategies can help smooth the load demand curve, making it easier for utilities to meet their customers' energy needs without straining their networks. By sublimating customer behavior and promoting energy-efficient technologies, DSM can also help reduce the environmental impact of electricity generation and consumption. As various power system and communication system infrastructure components are concerned in the implementation of DSM principles to enable fast and efficient marketing operation in addition to flexibility within their framework of operation [60], several programmatic strategies have been developed in our case study based on the perspectives of the load patterns of households in the clustering. In addition, according to the resuls of Table 5 and the selected feature sets in Fig. 5, we give

Table 6
The potential DSM strategies for individual cases.

Id	Patterns	Strategies	Categorization
#2093	C4	Install solar PV; Increase tariffs during peak periods, i.e., weekday evenings	EE and ToU, Technical and pricing levels
#2270	C3, C2	Decorate rooms with warm and insulating materials; Use energy storage, e.g., storage electric boilers	EE, Technical level
#2333	C7	Replace low-efficiency appliances with efficient ones., e.g., light bulb; Use IBR for the varying consumption	EE, DR and ToU, Technical and incentive levels
#2491	C1	Install solar PV; Use highly energy-efficiency appliances; Use dynamic RTP for household's tariffs	EE and pricing time-based levels

explanations for the potential strategies and the recommendation for the four studied households in Table 6:

- The #2270 household's main membership was found to be C4, and the corresponding symmetric uncertainty features in decreasing order are $F5$, $F2$, $F3$, $F1$, $F12$, $F10$, $F9$, $F7$, and $F8$. The #2093 has two adults aged 18-25, a house built within the last five years, with two bedrooms and about three-quarters of a light bulb. On the other hand, as the shape of Fig. 3 (C4), we can see a small spike in consumption in the morning and the consumption on weekdays shows relatively steady, in contrast to weekend which is higher. For the purpose of peak shaving, we recommend that the household installs PV projects to implement net metering and reduce demand on the power grid during peak periods. Additionally, since there is an obvious peak in consumption during weekday evenings, we suggest considering pricing-based tariffs to make the household more aware of reducing unnecessary power consumption during those hours.
- The #2270 household consists of two adults and one child. The chief earner is a retired person over the age of 65, and the household lives in a four-bedroom home with energy-efficient light bulbs. The age of the house is over 10 years. Based on the consumption patterns of the household, which are shown in Fig. 3 (C2, C3), there is a peak on weekday evenings and dual peaks on Saturday mornings and evenings. This indicates that the house may be poorly insulated and that the chief earner may not be using modern energy storage tools, such as storage electric boilers. Therefore, we strongly recommend that this customer replaces the insulation with new materials where available and uses energy storage appliances to reduce peak consumption.
- The household #2333 is a retired elderly person living alone in a three-bedroom home that uses half energy-efficient light bulbs. As shown in Fig. 3 (C7), there is a rapid increase in consumption after the afternoon every day. This indicates that the household may have high and low consumption levels. Furthermore, it is likely that the retired person does not use energy-saving electrical devices. Therefore, we propose a series of personalized services for this household, including increasing the proportion of energy-efficient appliances used and using an inclining block rate (IBR) [61] for varying consumption. The IBR grants incentives to customers based on distributing their usage towards other periods of the day to avoid higher tariffs, eventually reducing the grid system's peak-to-average ratio (PAR). We also suggest using energy storage tools such as electric boilers.
- The #2491 household, which consists of three adults and two children, lives in a four-bedroom house that is over 10 years old. The chief earner of the household is self-employed and middle-aged. Their consumption patterns, Fig. 3 (C1), are relatively cluttered throughout the day and are not regular, which may be attributed to the number of household members. Therefore, we recommend that this household should install PV projects and use more energy-saving appliances. In addition, we recommend using a dynamic real-time pricing strategy (RTP) [62], also known as a dynamic pricing strategy, which predetermines different tariffs based on hourly or daily usage patterns. This will help reduce the household's irregular consumption and establish good electricity consumption habits.

5. Discussion

In this paper, we have presented a novel approach for identifying customer groups for personalized DSM services using household socio-demographic data. Our approach aims to address the challenge of providing targeted DSM services in scenarios where smart meter data are not available or reliable, which is a common situation in many regions. By using a machine learning model that can predict the likelihood of DSM membership for households based on their characteristics, our approach can help utilities to offer customized DSM services from the first day of customer engagement, regardless of the availability of their energy consumption data. We have evaluated our approach on a real-world dataset from Ireland and demonstrated its effectiveness and potential for targeted DSM strategies.

Our approach has several implications for both research and practice. For research, our approach can provide a new perspective for understanding the relationship between residential energy consumption and household characteristics, and can inspire further studies on how to leverage socio-demographic information for DSM purposes. For practice, our approach can enable utilities to better manage energy demand and design tailored DSM strategies that are aligned with the specific needs and preferences of different customer groups. However, our approach also depends on several factors that may affect the energy consumption and load patterns of households, which we have not considered in this paper. These factors include the technical and environmental constraints that may influence the actual energy usage and demand of customers, as well as their willingness and ability to participate in DSM programs. In this section, we briefly discuss these factors and their implications for our approach.

- Technical constraints: The type and efficiency of appliances, lighting, heating, cooling, and ventilation systems in households can have a significant impact on their energy consumption and load patterns. For instance, households with smart thermostats or energy management systems may be able to adjust their temperature settings or switch off unnecessary devices to reduce their peak demand or participate in DSM programs. Therefore, it is important to consider the technical characteristics and capabilities of households when designing and implementing DSM strategies.
- Environmental constraints: The weather conditions, such as temperature, humidity, wind, and solar radiation, can also influence the energy consumption and load patterns of households. For example, households may consume more electricity for heating or cooling during extreme weather events or they may generate more electricity from rooftop solar panels during sunny days. Therefore it is essential to take into account the environmental factors and variations when planning and evaluating DSM strategies.

There are also some limitations and challenges that need to be addressed in future work. First, our approach relies on the availability and quality of household socio-demographic data, which may not be easily obtained or updated in some cases. Therefore, more efforts are needed to explore alternative ways of collecting and maintaining such

data. Second, our approach assumes that household characteristics remain relatively stable over time, which may not hold true in some situations. Therefore, more robust methods are needed to handle the dynamics and uncertainties of household characteristics and consumption patterns. Third, our approach does not consider the feedback effects of DSM interventions on household behavior and consumption, which may affect the performance and validity of our model. Therefore, more sophisticated methods are needed to incorporate the feedback mechanisms and evaluate the long-term impacts of DSM programs.

Moreover, our approach relies on hypothetical queries based on household characteristics, such as age, income, and cooking facilities, which may not reflect the actual energy usage patterns of customers, as these factors may change over time due to technological advancements or other reasons. However, we argue that our approach is still useful and applicable in many scenarios where the data-driven strategy is desirable. First, our approach does not rely on the exact values of the household characteristics, but rather on their relative importance and correlation with the DSM membership. Therefore, as long as the underlying relationship between the household characteristics and the DSM membership remains stable, our approach can still provide accurate predictions. Second, our approach can be easily updated and retrained with new data as they become available, which can improve the performance and adaptability of our model over time. Third, our approach can be combined with other methods that can capture the temporal dynamics and variations of the residential customer data, such as time series analysis or recurrent neural networks, which can further enhance the robustness and generalizability of our model.

Additionally, while the data in this study were acquired in the European Ireland region, the methodology presented herein holds potential applicability to any region equipped with a smart meters network. The significance of smart meters in the future is discernible, with benefits accruing to both households and stakeholders in the electricity service supply chain. It is, however, essential to acknowledge that the availability of data encompassing smart meter usage alongside comprehensive sociodemographic information about households remains a rare commodity and introduces concerns regarding the protection of personal privacy. Furthermore, the implementation of DSM strategies must take into account various factors, such as market tariff rules, geographic distribution of households, the level of competition in the electricity market, and the presence of taxes or other policies that may affect the cost of electricity. In the context of our case study, we have delineated several programmatic strategies based on market dynamics and household load patterns, which can contribute to optimizing the application of DSM principles. However, it is imperative that these recommendations undergo rigorous evaluation by the pertinent utility company to ascertain their suitability and feasibility within the specific operational context.

In summary, the results of our analyses can serve as a first step for utilities to explore electricity demand shifts resulting from intra-household peaking and market segmentation. These insights, based on a socio-demographic perspective, are essential for developing enhanced electricity services.

6. Conclusion and future work

In this paper, we have addressed the problem of delivering targeted Demand-Side Management (DSM) services to new customers within smart energy systems. Motivated by the lack or unreliability of smart meter data in many scenarios, we have proposed a novel data-driven approach that leverages household socio-demographic data to identify customer groups for personalized DSM services. Our approach consists of three main steps: (1) applying a robust and flexible clustering algorithm to identify representative electricity consumption patterns from historical smart meter data; (2) employing a comprehensive feature selection method to select highly correlated subsets of household characteristics for each clustering load pattern; (3) developing a state-of-the-art forecasting model to estimate customer groups eligible for DSM services based on their household characteristics.

We have evaluated our approach on a real-world dataset from Ireland and demonstrated its effectiveness and potential for targeted DSM strategies. The results have shown that our approach can achieve high prediction accuracy and outperform three baseline methods in uncovering nonlinear relationships between household characteristics and DSM membership. Moreover, we have conducted a specific case study to illustrate how our approach can help utilities to design customized DSM strategies for four individual households, factoring in their electricity consumption patterns and socio-demographic information.

In our future endeavors, we intend to delve into factors such as seasonal fluctuations, holiday-related patterns, and the influence of tiered electricity tariff structures and time band charging. Additionally, our focus will encompass privacy preservation for occupants through the implementation of federated learning techniques.

CRediT authorship contribution statement

Hanguan Wen: Conceptualization, Methodology, Investigation, Resources, Writing – original draft, Writing – review & editing. **Xiufeng Liu:** Conceptualization, Methodology, Writing – review & editing. **Ming Yang:** Conceptualization, Supervision. **Bo Lei:** Conceptualization, Writing – review & editing. **Xu Cheng:** Conceptualization, Methodology, Writing – review & editing. **Zhe Chen:** Conceptualization, Supervision.

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 was supported by the China Scholarship Council under Grant No. 202106150041.

References

[1] Zhang Y, Wen H, Dai X, Liang J, Xu Z, Xue K, et al. Research on modeling in operator mental workload based on VACP method. Qual Reliab Eng Int 2022.
[2] Dupont C, Kulovesi K, van Asselt H. Editorial: Governing the EU's climate and energy transition through the 2030 framework. Rev Eur Comp Int Environ Law 2020;29(2):147–50.
[3] IEA. Key world energy statistics, Paris. 2021, https://www.iea.org/reports/key-world-energy-statistics-2021. [Last accessed 25 October 2022].
[4] Gouveia JP, Seixas J. Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys. Energy Build 2016;116:666–76.
[5] Li J, Chen Z, Cheng L, Liu X. Energy data generation with wasserstein deep convolutional generative adversarial networks. Energy 2022;257:124694.
[6] Rajabi A, Eskandari M, Jabbari Ghadi M, Ghavidel S, Li L, Zhang J, et al. A pattern recognition methodology for analyzing residential customers load data and targeting demand response applications. Energy Build 2019;203:109455.
[7] Anderson B, Lin S, Newing A, Bahaj A, James P. Electricity consumption and household characteristics: Implications for census-taking in a smart metered future. Comput Environ Urban Syst 2017;63:58–67.
[8] Susanti L, Fithri P, Bestarina K. Demographic characteristics in correlation with household electricity use. In: Industrial engineering, management science and applications 2015. Springer; 2015, p. 959–68.
[9] Guo Z, Zhou K, Zhang C, Lu X, Chen W, Yang S. Residential electricity consumption behavior: Influencing factors, related theories and intervention strategies. Renew Sustain Energy Rev 2018;81:399–412.
[10] Keerthisinghe C, Verbič G, Chapman AC. A fast technique for smart home management: ADP with temporal difference learning. IEEE Trans Smart Grid 2016;9(4):3291–303.

[11] Wang Y, Chen Q, Gan D, Yang J, Kirschen DS, Kang C. Deep learning-based socio-demographic information identification from smart meter data. IEEE Trans Smart Grid 2019;10(3):2593–602.
[12] Fu X, Zeng X-J, Feng P, Cai X. Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China. Energy 2018;165:76–89.
[13] Chicco G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy 2012;42(1):68–80.
[14] Pereira R, Fagundes A, Melício R, Mendes V, Figueiredo J, Martins J, et al. A fuzzy clustering approach to a demand response model. Int J Electr Power Energy Syst 2016;81:184–92.
[15] Crow M, et al. Clustering-based methodology for optimal residential time of use design structure. In: 2014 North American power symposium. IEEE; 2014, p. 1–6.
[16] Biscarri F, Monedero I, Garcia A, Guerrero JI, Leon C. Electricity clustering framework for automatic classification of customer loads. Expert Syst Appl 2017;86:54–63.
[17] Räsänen T, Ruuskanen J, Kolehmainen M. Reducing energy consumption by using self-organizing maps to create more personalized electricity use information. Appl Energy 2008;85(9):830–40.
[18] Yang J, Zhao J, Wen F, Dong Z. A model of customizing electricity retail prices based on load profile clustering analysis. IEEE Trans Smart Grid 2018;10(3):3374–86.
[19] Varga ED, Beretka SF, Noce C, Sapienza G. Robust real-time load profile encoding and classification framework for efficient power systems operation. IEEE Trans Power Syst 2014;30(4):1897–904.
[20] Sun M, Wang Y, Teng F, Ye Y, Strbac G, Kang C. Clustering-based residential baseline estimation: A probabilistic perspective. IEEE Trans Smart Grid 2019;10(6):6014–28.
[21] Kwac J, Flora J, Rajagopal R. Household energy consumption segmentation using hourly data. IEEE Trans Smart Grid 2014;5(1):420–30.
[22] Ahir RK, Chakraborty B. A novel cluster-specific analysis framework for demand-side management and net metering using smart meter data. Sustain Energy Grids Netw 2022;31:100771.
[23] Satre-Meloy A, Diakonova M, Grünewald P. Cluster analysis and prediction of residential peak demand profiles using occupant activity data. Appl Energy 2020;260:114246.
[24] Yang T, Ren M, Zhou K. Identifying household electricity consumption patterns: A case study of Kunshan, China. Renew Sustain Energy Rev 2018;91:861–8.
[25] Rhodes JD, Cole WJ, Upshaw CR, Edgar TF, Webber ME. Clustering analysis of residential electricity demand profiles. Appl Energy 2014;135:461–71.
[26] Kaur R, Gabrijelčič D. Behavior segmentation of electricity consumption patterns: A cluster analytical approach. Knowl-Based Syst 2022;251:109236.
[27] Gajowniczek K, Ząbkowski T. Data mining techniques for detecting household characteristics based on smart meter data. Energies 2015;8(7):7407–27.
[28] Benítez I, Díez J-L, Quijano A, Delgado I. Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance. Electr Power Syst Res 2016;140:517–26.
[29] Trotta G. An empirical analysis of domestic electricity load profiles: Who consumes how much and when? Appl Energy 2020;275:115399.
[30] Jeyaranjani J, Devaraj D. Deep learning based smart meter data analytics for electricity load prediction. In: 2019 IEEE international conference on clean energy and energy efficient electronics circuit for sustainable development. IEEE; 2019, p. 1–5.
[31] Huebner G, Shipworth D, Hamilton I, Chalabi Z, Oreszczyn T. Understanding electricity consumption: A comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes. Appl Energy 2016;177:692–702.
[32] McLoughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl Energy 2015;141:190–9.
[33] Ullah A, Haydarov K, Ul Haq I, Muhammad K, Rho S, Lee M, et al. Deep learning assisted buildings energy consumption profiling using smart meter data. Sensors 2020;20(3):873.
[34] De Lauretis S, Ghersi F, Cayla J-M. Energy consumption and activity patterns: An analysis extended to total time and energy use for French households. Appl Energy 2017;206:634–48.
[35] Beckel C, Sadamori L, Staake T, Santini S. Revealing household characteristics from smart meter data. Energy 2014;78:397–410.
[36] Yan S, Li K, Wang F, Ge X, Lu X, Mi Z, et al. Time–frequency feature combination based household characteristic identification approach using smart meter data. IEEE Trans Ind Appl 2020;56(3):2251–62.
[37] Wang Y, Bennani IL, Liu X, Sun M, Zhou Y. Electricity consumer characteristics identification: A federated learning approach. IEEE Trans Smart Grid 2021;12(4):3637–47.
[38] Yang M-S, Nataliani Y. Robust-learning fuzzy c-means clustering algorithm with unknown number of clusters. Pattern Recognit 2017;71:45–59.
[39] Ruspini EH. A new approach to clustering. Inf Control 1969;15(1):22–32.
[40] Bezdek JC. Numerical taxonomy with fuzzy sets. J Math Biol 1974;1(1):57–71.
[41] Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science 2014;344(6191):1492–6.
[42] Fazendeiro P, de Oliveira JV. Observer-biased fuzzy clustering. IEEE Trans Fuzzy Syst 2014;23(1):85–97.
[43] Lin Y, Chen S. A centroid auto-fused hierarchical fuzzy c-means clustering. IEEE Trans Fuzzy Syst 2021;29(7):2006–17.
[44] Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning. 2003, p. 856–63.
[45] Bakhshandeh S, Azmi R, Teshnehlab M. Symmetric uncertainty class-feature association map for feature selection in microarray dataset. Int J Mach Learn Cybern 2020;11.
[46] Canayaz M. Classification of diabetic retinopathy with feature selection over deep features using nature-inspired wrapper methods. Appl Soft Comput 2022;128:109462.
[47] Agrawal U, Rohatgi V, Katarya R. Normalized mutual information-based equilibrium optimizer with chaotic maps for wrapper-filter feature selection. Expert Syst Appl 2022;207:118107.
[48] Unler A, Murat A. A discrete particle swarm optimization method for feature selection in binary classification problems. European J Oper Res 2010;206(3):528–39.
[49] Got A, Moussaoui A, Zouache D. Hybrid filter-wrapper feature selection using whale optimization algorithm: A multi-objective approach. Expert Syst Appl 2021;183:115312.
[50] Witten IH, Frank E. Data mining: practical machine learning tools and techniques with java implementations. Acm Sigmod Rec 2002;31(1):76–7.
[51] Wang Z, Gao S, Zhang Y, Guo L. Symmetric uncertainty-incorporated probabilistic sequence-based ant colony optimization for feature selection in classification. Knowl-Based Syst 2022;256:109874.
[52] Rahmanian M, Mansoori E. Unsupervised fuzzy multivariate symmetric uncertainty feature selection based on constructing virtual cluster representative. Fuzzy Sets and Systems 2022;438:148–63.
[53] Dai J, Chen J, Liu Y, Hu H. Novel multi-label feature selection via label symmetric uncertainty correlation learning and feature redundancy evaluation. Knowl-Based Syst 2020;207:106342.
[54] Tang W, Wang H, Lee X-L, Yang H-T. Machine learning approach to uncovering residential energy consumption patterns based on socioeconomic and smart meter data. Energy 2022;240:122500.
[55] Wang Y, Chen Q, Kang C, Zhang M, Wang K, Zhao Y. Load profiling and its application to demand response: A review. Tsinghua Sci Technol 2015;20(2):117–29.
[56] Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. Comput Geosci 1984;10(2–3):191–203.
[57] Entropy, relative entropy, and mutual information. In: Elements of information theory. John Wiley & Sons, Ltd; 2005, p. 13–55.
[58] Sarhrouni E, Hammouch A, Aboutajdine D. Application of symmetric uncertainty and mutual information to dimensionality reduction and classification of hyperspectral images. 2012, arXiv preprint arXiv:1211.0613.
[59] Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Rec 2002;31(1):76–7.
[60] Panda S, Mohanty S, Rout PK, Sahu BK, Bajaj M, Zawbaa HM, et al. Residential demand side management model, optimization and future perspective: A review. Energy Rep 2022;8:3727–66.
[61] Borenstein S. Equity effects of increasing-block electricity pricing. 2008.
[62] Edward J, Policy P. Assessment of customer response to real time pricing. New Jersey: Edward J, Bloustein School of Planning and Public Policy, State University of New Jersey; 2005.