Short-term Residential Load Forecasting Based on *K*-shape Clustering and Domain Adversarial Transfer Network

Jizhong Zhu, *Fellow*, *IEEE*, Yuwang Miao, Hanjiang Dong, Shenglin Li, Ziyu Chen, and Di Zhang

***Abstract*—In recent years, the expansion of the power grid has led to a continuous increase in the number of consumers within the distribution network. However, due to the scarcity of historical data for these new consumers, it has become a complex challenge to accurately forecast their electricity demands through traditional forecasting methods. This paper proposes an innovative short-term residential load forecasting method that harnesses advanced clustering, deep learning, and transfer learning technologies to address this issue. To begin, this paper leverages the domain adversarial transfer network. It employs limited data as target domain data and more abundant data as source domain data, thus enabling the utilization of source domain insights for the forecasting task of the target domain. Moreover, a *K*-shape clustering method is proposed, which effectively identifies source domain data that align optimally with the target domain, and enhances the forecasting accuracy. Subsequently, a composite architecture is devised, amalgamating attention mechanism, long short-term memory network, and seq2seq network. This composite structure is integrated into the domain adversarial transfer network, bolstering the performance of feature extractor and refining the forecasting capabilities. An illustrative analysis is conducted using the residential load dataset of the Independent System Operator to validate the proposed method empirically. In the case study, the relative mean square error of the proposed method is within 30 MW, and the mean absolute percentage error is within 2%. A significant improvement in accuracy, compared with other comparative experimental results, underscores the reliability of the proposed method. The findings unequivocally demonstrate that the proposed method advocated in this paper yields superior forecasting results compared with prevailing mainstream forecasting methods.**

***Index Terms*—Load forecasting, domain adversarial, *K*-shape clustering, long short-term memory network, seq2seq network, attention mechanism.**

Manuscript received: September 6, 2023; revised: December 25, 2023; accepted: March 3, 2024. Date of CrossCheck: March 3, 2024. Date of online publication: March 27, 2024.

This work was supported by the National Natural Science Foundation of China (No. 52177087) and Guangdong Basic and Applied Basic Research Foundation, China (No. 2022B1515250006).

J. Zhu, Y. Miao, H. Dong (corresponding author), S. Li, Z. Chen, and D. Zhang are with the School of Electric Power Engineering, South China University of Technology, Guangzhou 510641, China, and H. Dong is also with the Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: zhujz@scut.edu.cn; miaoyw2021@163.com; hanjiang.dong@foxmail.com; iamlshl@126.com; epchenzy@mail.scut.edu.cn; deezhang_ee@outlook.com).

DOI: 10.35833/MPCE.2023.000646

I. Introduction

SHORT-TERM load forecasting constitutes a fundamental pillar within power system planning, consumption analysis, and operational coordination. It is important to attain precision and rationality in short-term load forecasting, furnishing pivotal data that underpin decisions encompassing optimal dispatching, operational planning, and demand-side administration within the power system [1]. This, in turn, begets heightened economic gains and propels efficient power utilization. Against power grid expansion nowadays, the population of power consumers on the consumption side is on a relentless ascent. This surges users in fresh challenges for load forecasting: the absence of historical load data for emerging power consumers renders their patterns of electricity consumption uncertain [2]. Consequently, conventional forecasting methods struggle to forecast their electricity loads accurately.

Numerous technologies have found applications in load forecasting, with prominent categories namely statistical methods and machine learning methods [3]. Among these, statistical methods encompass the linear regression method [4], the exponential smoothing method [5], and the autoregressive moving average method [6]. On the other hand, machine learning methods encompass those such as support vector machines (SVMs) [7], but it is the domain of deep learning that has witnessed rapid advancement in recent times. Deep learning has gained unparalleled traction due to its prowess in handling time series data.

Deep learning encompasses a range of methods with key players including the convolutional neural networks (CNNs) [8] and the long short-term memory (LSTM) networks. While CNNs primarily shine in fields such as computer vision [8] and natural language processing [9], the contemporary recurrent neural networks (RNNs) typified by LSTM have found their niche in load forecasting. LSTM has emerged as the cornerstone of load forecasting because of its capacity to capture long-term dependencies and surmount the challenge of gradient vanishing. Moreover, it demonstrates an aptitude for abstract feature extraction from load data, curtailing data intricacies. An extensive body of literature attests to the noteworthy achievements of LSTM in the forecasting domain [10]. Reference [11] conceives a forecasting framework grounded in LSTM, thereby effecting short-

term load forecasting for individual residential energy loads with commendable outcomes. Similarly, [12] harnesses LSTM to model temporal intricacies within time series data, crafting a forecasting model for building load that exhibits remarkable accuracy.

Furthermore, load forecasting has witnessed the application of more sophisticated deep learning methods. Notably, methods such as the seq2seq network and attention mechanism have come to the fore, acclaimed for their efficacy in capturing both long-term and short-term dependencies inherent in load data [13]. Illustratively, [14] delves into load forecasting within renewable energy, employing a seq2seq network to excel in the short-term forecasting of office building loads. Concomitantly, [13] innovates a transformer-based seq2seq network that hinges on the attention mechanism, leading to the precise forecasting of reactive loads in residential areas. Noteworthy also is the endeavor of [15] that concentrates on the forecasting of building energy and orchestrates an attention-enhanced seq2seq network, thereby attaining remarkable accuracy in the advanced multi-step forecasting of building energy dynamics.

The passage mentioned above delineates the extensive utilization of deep learning technology in regional residential load forecasting. However, this efficacy is hindered when newly developed residential zones present a scarcity of load data. Conventional deep learning methods grapple with establishing effective forecasting models under such constrained conditions, exacerbating the challenges in load forecasting [16].

Nonetheless, a resolution surfaces in transfer learning, adept at ameliorating the dearth of data by drawing insights from alternative datasets for model training and subsequent target dataset forecasting. Within this context, the domain adversarial transfer network (DATN) emerges as an exemplar, augmenting traditional transfer learning with an adversarial mechanism. This strategic amalgamation of domain adaptation and profound feature learning engenders an efficacious mapping between source and target domains, thereby enhancing transfer efficiency [17]. Illustratively, [18] harnesses domain adaptation to facilitate the seamless migration of building load data, culminating in discernible enhancements in forecasting accuracy. Similarly, [19] introduces an ensemble forecasting model grounded in domain adaptation principles, adeptly circumventing the issue of inadequate training of model attributed to limited samples stemming from concept drift. The convergence of deep learning, transfer learning, and domain adaptation constitutes a pivotal stride toward addressing the intricacies of sparse data scenarios in load forecasting.

In transfer learning endeavors, it becomes a pivotal consideration to select suitable source domain data when multiple load datasets are available. Contemporary practices frequently encompass methodologies such as *K*-means clustering and dynamic time warping (DTW). Notably, *K*-means clustering stands as a conventional method. Reference [20] introduces a forecasting model that harmoniously melds the *K*-means clustering with deep learning. Specifically, the *K*-means clustering is leveraged to distill the resemblance within residential load profiles, thereafter facilitating load data forecasting through deep learning methods. On a similar note, [21] advocates for the utilization of *K*-means to categorize users in load forecasting. Nevertheless, it is acknowledged that the classification efficacy might wane in face of voluminous data, potentially impinging upon the forecasting accuracy. Alternatively, the DTW has recently emerged as a dominant method for gauging correlation. Reference [22] harnesses the DTW distance as a benchmark to gauge similarity between source and target domain data, facilitating the selection of source domain data. However, a drawback inherent to DTW is its elevated calculation complexity.

A recent addition to this landscape is the *K*-shape clustering, which incorporates a refined model grounded in the principles of *K*-means. This method augments operational efficiency and bolsters the robustness of data classification, thereby achieving calculation efficacy akin to DTW [23]. Regrettably, its application in load forecasting remains limited as of now. The efficacy of the *K*-shape clustering has been demonstrated in specific application domains. For instance, [24] proposes a convolutional RNN predicated upon *K*-shape clustering and an attention mechanism, culminating in proficient short-term wind speed forecasting. Likewise, [25] advances the *K*-shape clustering to classify building energy consumption data. Empirical outcomes highlight the commendable clustering proficiency of this method across various energy consumption data granularities.

Consequently, this paper presents a solution for the challenges posed by the scarcity of data from newly established residential users to underpin load forecasting. It addresses the conundrum of source domain data selection for the transfer process. The proposed method for short-term load forecasting within this context is poised to enhance forecasting accuracy effectively.

The contributions of this paper can be outlined as follows.

1) A forecasting model based on the LSTM network and an attention mechanism is devised, thereby enhancing the encoding process of load data.

2) The integration of the domain adversarial mechanism into the seq2seq network gives rise to the DATN, facilitating the seamless transition of data from the source domain to the target domain.

3) A novel method grounded in the *K*-shape clustering is introduced. This method quantifies load data similarities, subsequently serving as a benchmark for selecting pertinent source domain data.

The organization of this paper unfolds as follows. Section II delineates the model structures for this paper. Subsequently, the Section III presents the overall framework advanced within this paper. Section IV presents illustrative case studies to corroborate the viability of the proposed method. Finally, Section V encapsulates the findings and conclusions.

II. Model Structure for This Paper

This paper uses the *K*-shape clustering to sieve the source domain data, while the seq2seq network, LSTM network, and attention mechanism collaboratively enhance the perfor-

mance of conventional DATN method. Subsequent subsections will expound on each of these components individually.

A. K-shape Clustering

Clustering is a widely employed unsupervised data mining method applicable to time series data processing [26]. Through clustering, insights into regularities and correlations among data can be gleaned, enabling the categorization of time series data according to their affinities [27]. A prominent conventional unsupervised clustering method is *K*-means, which calculates similarity between time series via Euclidean distance (ED). This method boasts a straightforward principle and notable calculation efficiency [28]. Nevertheless, *K*-means disregards potential horizontal scaling and data translation effects, rendering it less effective in scenarios where dissimilar time series lengths or phase disparities are involved.

The *K*-shape clustering represents an enhanced clustering method derived from *K*-means, showcasing notable advancements across two key dimensions [23].

1) Distance calculation: the distance is calculated utilizing the cross-correlation metric, and a statistical method is employed to ascertain the similarity between two time series, even when their lengths differ and they lack precise alignment.

The time series are denoted as $\vec{x}=(x_1,x_2,\ldots,x_n)$ and $\vec{y}=(y_1,y_2,\ldots,y_n)$, respectively, where n is the number of time series. The sequence $\vec{x}$ is subject to transformation by (1) to attain displacement invariance:

$$\vec{x}_{(s)}=\begin{cases}(\overbrace{0,\ldots,0}^{|s|},x_1,x_2,\ldots,x_n-s) & s\geq 0\\ (x_{1-s},\ldots,x_n-1,x_n,\underbrace{0,\ldots,0}_{|s|}) & s<0\end{cases} \tag{1}$$

where s represents the displacement of the sequence. Considering a range of distinct s values, we incrementally derive the dot product between $\vec{x}_{(s)}$ and $\vec{y}$ as visually depicted in Fig. 1.

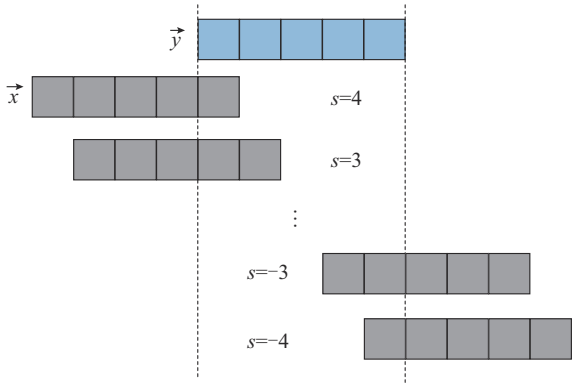

Fig. 1. Calculation of dot product between sequences.

In Fig. 1, considering the sequence length as an illustrative instance with $n=5$, the values for s ranges within $[-4,4]$. Diverse values of s prompt a systematic calculation of the dot product using (2), resulting in a cumulative count of obtained dot products totaling $2n-1$, which equates to 9 in this scenario.

Subsequently, the correlation coefficient between the two series is obtained:

$$R_s(\vec{x},\vec{y})=\begin{cases}\sum_{i=1}^{n-s}y_i x_i+s & s\geq 0\\ R-s(\vec{y},\vec{x}) & s<0\end{cases} \tag{2}$$

where R_s is the inner product between sequences.

Ultimately, the shape-based distance (SBD) is evaluated employing the normalized correlation coefficient, represented by:

$$SBD(\vec{x},\vec{y})=1-\max\frac{R_s(\vec{x},\vec{y})}{\sqrt{R(\vec{x},\vec{x})R(\vec{y},\vec{y})}} \tag{3}$$

where R is the inner product within the identical sequences. Proximity to 0 in the SBD value indicates a heightened correlation between the two sequences.

2) Centroid calculation: in instances where clusters encompass multiple time series, the entire cluster can be succinctly encapsulated by a single series termed the centroid. A straightforward method to deriving the centroid from a series collection entails calculating the arithmetic mean of corresponding coordinates across all series, thus determining the centroid coordinates. Nevertheless, the efficacy of the proposed method is compromised when dealing with temporally shifted time series. As depicted in (4), the *K*-shape clustering calculates the centroid using the SBD, thereby reformulating the centroid calculation into an optimization paradigm.

$$\vec{\mu}_k^{*}=\arg\max_{\vec{\mu}_k^{*}}\sum_{\vec{x}\in D}\left(\frac{\max R_s(\vec{x},\vec{\mu}_k)}{\sqrt{R(\vec{x},\vec{x})}\sqrt{R(\vec{\mu}_k,\vec{\mu}_k)}}\right)^2 \tag{4}$$

where $\vec{\mu}_k$ represents the centroid that becomes imperative to determine the optimal value $\vec{\mu}_k^{*}$ for the centroid to ensure the maximal similarity between the centroid and each sequence within cluster D.

B. Seq2seq Network

The seq2seq network was initially introduced in 2014, gaining prominence for its capacity to handle tasks characterized by uncertain output lengths, particularly in machine translation [29]. Unlike conventional neural networks, seq2seq networks are constructed upon an encoder-decoder architecture. In the forecasting context, they process variable-length sequences as input for the encoder, conducting feature extraction to yield a fixed-shape hidden state. This hidden state is then transformed into a variable-length sequence through the decoder, culminating in the eventual generation of the final forecasting outcome. Differing from deep learning methods typified by LSTM, the seq2seq network architecture embodies an autoregressive model and generates time series transformations from one domain to another, addressing the challenge of inadequate data in traditional deep learning methods. Consequently, it finds its forte in forecasting scenarios marked by limited availability of historical data.

Nonetheless, traditional seq2seq networks confront specific concerns as follows.

1) Given that the encoder-decoder structure in the seq2seq network is founded on a fundamental feedforward neural network (FNN), the amplification of input sequence length poses challenges in effectively encompassing all the information embedded within extended sequences into the hidden state. Consequently, this can impede the forecasting accuracy [30].

2) The inherent limitation of a fixed-length hidden state might compromise its capacity to comprehensively retain the entirety of input sequence information, potentially leading to the overwriting of earlier data input by subsequently arriving data [14].

This paper advocates the substitution of conventional FNN within the encoder-decoder framework with LSTM to address the first concern. This transition aims to enhance the treatment of extended sequences by affording improved long-term sequence processing.

In response to the second concern, this paper introduces an attention mechanism into the conventional seq2seq network. This augmentation assigns greater weights to salient information within the input sequence. Consequently, the limited-length hidden state is empowered to capture a more comprehensive spectrum of sequence information.

C. LSTM Network

The LSTM network model constitutes an advancement over the conventional RNN [30]. Drawing inspiration from computer logic gates, it offers a solution to the issue of gradient vanishing or explosion that arises from ongoing matrix multiplications in RNN [31]. At present, it stands as the preeminent contemporary model in the realm of RNNs. LSTM infuses memory components into the RNN in its evolution, simulating human cognitive processes [32]. This augmentation facilitates enhanced processing of historical information. State gates are introduced to regulate the output of each memory cell, which governs these memory cells [33]. The structural depiction of the LSTM network is presented in Fig. 2.

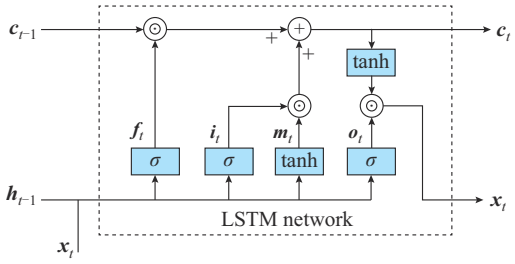

Fig. 2. Structural depiction of LSTM network.

The calculation formula corresponding to each component within Fig. 2 is given by:

$$\boldsymbol{i}_t = f(\boldsymbol{x}_t \boldsymbol{w}_{\mathrm{xi}} + \boldsymbol{h}_{t-1} \boldsymbol{w}_{\mathrm{hi}} + \boldsymbol{b}_{\mathrm{i}}) \tag{5}$$

$$\boldsymbol{f}_t = f(\boldsymbol{x}_t \boldsymbol{w}_{\mathrm{xf}} + \boldsymbol{h}_{t-1} \boldsymbol{w}_{\mathrm{hf}} + \boldsymbol{b}_{\mathrm{f}}) \tag{6}$$

$$\boldsymbol{o}_t = f(\boldsymbol{x}_t \boldsymbol{w}_{\mathrm{xo}} + \boldsymbol{h}_{t-1} \boldsymbol{w}_{\mathrm{ho}} + \boldsymbol{b}_{\mathrm{o}}) \tag{7}$$

$$\boldsymbol{m}_t = \tanh(\boldsymbol{x}_t \boldsymbol{w}_{\mathrm{xm}} + \boldsymbol{h}_{t-1} \boldsymbol{w}_{\mathrm{hm}} + \boldsymbol{b}_{\mathrm{m}}) \tag{8}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{m}_t \tag{9}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh \boldsymbol{c}_t \tag{10}$$

where $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, $\boldsymbol{o}_t$, and $\boldsymbol{m}_t$ are the input gate, forget gate, output gate, and candidate memory units, respectively; $\boldsymbol{x}_t$ is the input of the memory elements; $\boldsymbol{c}_t$ is the output of the memory elements; $\boldsymbol{h}_t$ is the hidden state at time t; $\boldsymbol{w}_{\mathrm{xi}}$, $\boldsymbol{w}_{\mathrm{xf}}$, $\boldsymbol{w}_{\mathrm{xo}}$, and $\boldsymbol{w}_{\mathrm{xm}}$ are the weight matrices multiplied by $\boldsymbol{x}_t$; $\boldsymbol{w}_{\mathrm{hi}}$, $\boldsymbol{w}_{\mathrm{hf}}$, $\boldsymbol{w}_{\mathrm{ho}}$, and $\boldsymbol{w}_{\mathrm{hm}}$ are the weight matrices multiplied by $\boldsymbol{h}_{t-1}$; $\boldsymbol{b}_{\mathrm{i}}$, $\boldsymbol{b}_{\mathrm{f}}$, $\boldsymbol{b}_{\mathrm{o}}$, and $\boldsymbol{b}_{\mathrm{m}}$ are the weight biases; f is the employed sigmoid function σ; and $\odot$ is the Hadamard product.

As illustrated in (9), the output of the memory cell $\boldsymbol{c}_t$ at time t is primarily shaped by two components, i.e., $\boldsymbol{c}_{t-1}$ and $\boldsymbol{m}_t$, corresponding to the previous memory cell state and the candidate memory unit, respectively. The input gate $\boldsymbol{i}_t$ governs the incorporation of fresh information from $\boldsymbol{m}_t$, while the forget gate $\boldsymbol{f}_t$ dictates the preservation of historical data from the prior memory cell state $\boldsymbol{c}_{t-1}$. This dual interplay facilitates integrating current and antecedent data for a comprehensive understanding.

D. Attention Mechanism

The attention mechanism emulates the cognitive information processing observed in the human brain. Initially proposed in 2014, it has found diverse applications in domains like machine translation, speech recognition, and image processing [34]. In recent years, the utilization of the attention mechanism has progressively extended to the realm of deep learning. This mechanism facilitates the focal engagement of the model with pivotal feature information, amplifying its weight during training while diminishing the weight attributed to less significant ancillary data [34]. Consequently, this mechanism curtails sensitivity to secondary information, augmenting the data processing efficacy, especially when dealing with restricted datasets.

The attention mechanism can be categorized into hard attention and soft attention. A binary weight coefficient (0 or 1) is assigned to each input element in hard attention, indicating selective focus on specific input data. Conversely, weight coefficients ranging from 0 to 1 are allocated across all input data in soft attention, signifying a more comprehensive engagement with the input [35]. Hence, this paper employs soft attention exemplified by the calculation procedure shown in Fig. 3.

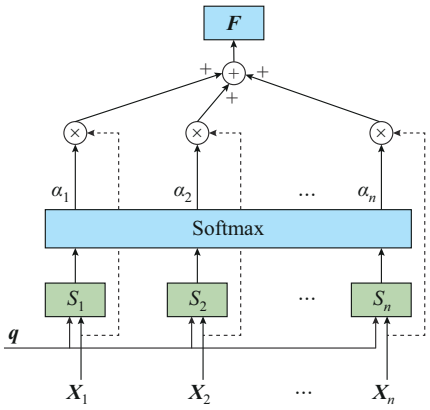

Fig. 3. Calculation procedure of attention mechanism.

The calculation procedure is divisible into three distinct steps.

1) Calculate the attention score S_i. In Fig. 3, the input vector is denoted as $\boldsymbol{X}=[\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n]$, while $\boldsymbol{q}$ represents the query vector. The assessment of the significance of each input vector concerning the query vector can be executed using an attention-scoring function. Among the prevalent scoring functions, the dot product function is widely employed. Its formulation is succinctly captured by:

$$S_i=\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{q} \tag{11}$$

2) Calculate the attention distribution weight α_i. Utilize the softmax function f^{softmax} to normalize the attention score S_i within the range of [0, 1]:

$$\alpha_i=f^{\mathrm{softmax}}(S_i)=\frac{\exp(S_i)}{\sum_{j=1}^{n}\exp(S_j)} \tag{12}$$

3) Calculate the final attention output vector $\boldsymbol{F}$. Based on the attention distribution weight α_i, the input data are subjected to a weighted averaging process, yielding the ultimate output:

$$\boldsymbol{F}=\sum_{i=1}^{n}\alpha_i\boldsymbol{X}_i \tag{13}$$

This mechanism directs the attention of the model towards pivotal feature information, enabling the allocation of heightened weight ratios to such aspects during model training while allocating lesser weights to comparatively less significant secondary data [34]. Consequently, it mitigates sensitivity to secondary data, enhancing the data processing capability of the model mainly when working with limited datasets.

E. DATN

Domain adaptation is a pivotal facet of transfer learning that involves mapping data from source and target domains characterized by disparate distributions into a unified feature space. The minimization of the distance between the data points from the two domains within this feature space under a defined metric, which acquires domain-invariant features [35], diminishes the divergence between domain distributions. This endeavor thereby equips the model with enhanced prowess in addressing novel tasks within the target domain. The conceptual incorporation of confrontation into domain adaptation gives rise to the formulation of DATN.

The traditional architecture of transfer learning is comprised of a feature extractor and a predictor. The DATN integrates a domain classifier to achieve an adversarial effect to expand upon this foundation [36], as illustrated in Fig. 4. In this method, the feature extractor maps data into the feature space, facilitating the ability of predictor to discern categories within the source domain data. However, the domain discriminator struggles to distinguish the originating domain of data. Subsequently, the predictor employs the output of feature extractor to forecast the domain-specific class label of the input data. During the training phase, the role of the domain classifier involves categorizing data within the feature space and striving to differentiate the originating domains maximally. This complex interaction aims to enhance the domain adaptation capabilities of the DATN. Within this framework, the interplay between the feature extractor and the domain classifier gives rise to an antagonistic effect. The introduction of a gradient inversion layer during training serves to minimize classification errors. Meanwhile, the integration of the feature extractor and predictor contributes to the reduction of training errors. The collective training error E encompassing the feature extractor, predictor, and domain classifier is depicted in (14).

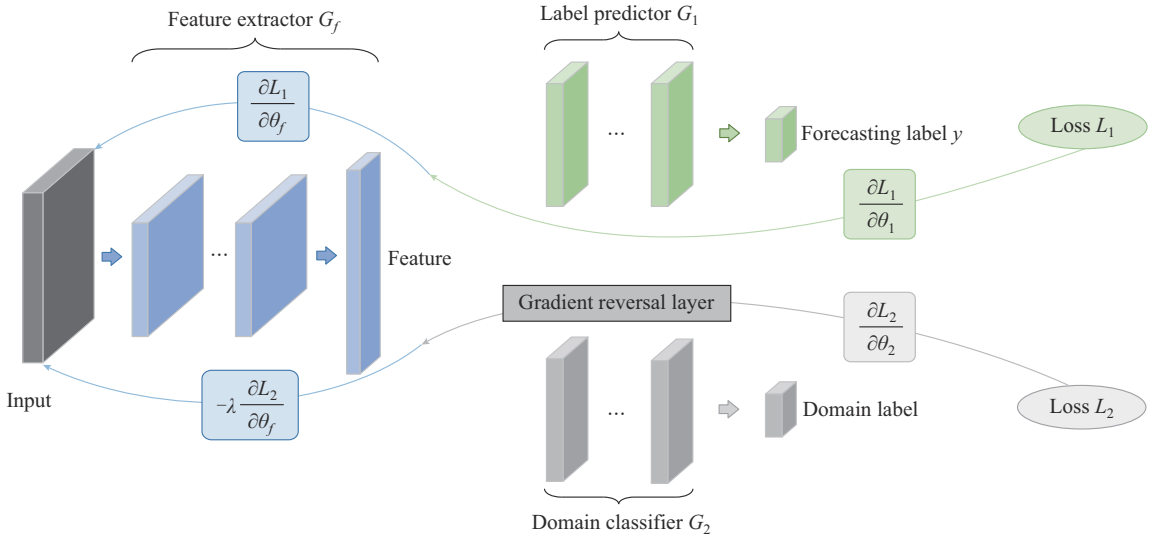

Fig. 4. DATN structure diagram.

$$E(\theta_f,\theta_1,\theta_2)=\sum_{i=1}^{N}L_1(G_1(G_f(x_i,\theta_f),\theta_1),y)-\lambda\sum_{i=1}^{N}L_2(G_2(G_f(x_i,\theta_f),\theta_2),y)=\sum_{i=1}^{N}L_1(\theta_f,\theta_1)-\lambda\sum_{i=1}^{N}L_2(\theta_f,\theta_2) \tag{14}$$

where G_1, θ_1, and L_1 are the label predictor, its corresponding parameters, and the associated loss function, respectively; G_2, θ_2, and L_2 are the domain classifier, its parameters, and its specific loss function, respectively; G_f and θ_f are the

feature extractor and its parameters, respectively; λ is the trade-off coefficient of the loss function L_2; and N is the number of training data. Throughout the training process, iterative optimization of θ_1, θ_2, and θ_f is undertaken to ascertain their saddle points, which is illustrated in (15) and (16).

$$(\theta_f', \theta_1') = \arg\min_{\theta_f, \theta_1} E(\theta_f, \theta_1, \theta_2') \tag{15}$$

$$\theta_2' = \arg\min_{\theta_2} E(\theta_f', \theta_1', \theta_2) \tag{16}$$

At the saddle point, the parameter θ_1 of the label predictor minimizes the forecasting loss, while the parameter θ_2 of the domain classifier maximizes the classification loss. Upon integrating the feature inversion layer, the parameter θ_f of the feature extractor simultaneously minimizes both the forecasting and classification losses.

III. Overall Framework

Overall, the functions of each part of the proposed method in this paper are summarized as shown in Fig. 5. The proposed method integrates the advantages of its components, making it more suitable for the forecasting scenarios introduced in this paper.

K-shape clustering	Utilizing an enhanced distance calculation method and centroid computation method, it exhibits superior clustering performance compared with conventional clustering methods.
seq2seq	Advantage: establishing a mapping between two domains, it is more suitable for application scenarios lacking historical data. Disadvantage: the basic architecture comprises FNN, and as the length of the input sequence increases, the forecasting performance deteriorates.
LSTM	Advantage: it can consider both current and past data information, providing better handling of time series. Disadvantage: in situations with insufficient historical data, the forecasting performance is suboptimal, hence it is combined with the seq2seq network.
Attention mechanism	Simulating the way the human brain processes information enables the method to focus more on crucial information during training.
DATN	Advantage: introducing the domain adversarial mechanism on top of the traditional transfer learning framework leads to improved transfer effectiveness. Disadvantages: the network architecture of the feature extractor, label predictor, and domain classifier is overly simplistic, resulting in limited capability to handle time series data.

Fig. 5. Summary of functions of each part of proposed method.

The overall framework of the proposed method can be divided into three main components, as illustrated in Fig. 6. Each of these components will be introduced sequentially.

A. K-shape Clustering

Initially, it is imperative to identify the dataset slated for forecasting as the target domain data. Subsequently, a dataset that optimally aligns with the target domain is chosen from the pool of various datasets earmarked for selection as the source domain data. Through the implementation of *K*-shape clustering, all load data are partitioned into distinct clusters. Subsequently, the SBD between the target domain data and the data within the same cluster is calculated. The dataset exhibiting the minimum SBD value is then selected as the source domain data.

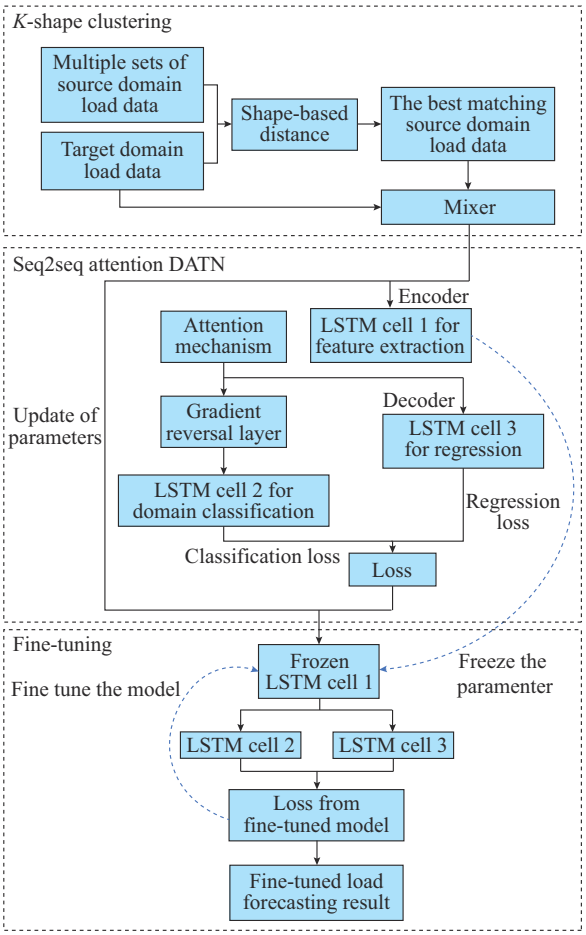

Fig. 6. Overall framework of proposed method.

Upon selecting the most suitable source domain data, the domain classification labels are appended to both datasets using one-hot encoding. Specifically, a domain label of 0 is assigned to the source domain data, while a label of 1 is assigned to the target domain data. Subsequently, a comprehensive shuffling and blending process is applied to the two datasets, resulting in an amalgamated dataset that serves as the input data $\boldsymbol{x}_t$. This process is denoted as:

$$\boldsymbol{x}_t = [\boldsymbol{x}_t^{load}, \boldsymbol{x}_t^{lable}] \tag{17}$$

where $\boldsymbol{x}_t^{load}$ represents the t^{th} load data; and $\boldsymbol{x}_t^{label}$ represents the t^{th} domain classification label data.

B. Seq2seq Attention DATN

The overall framework of the proposed method in this paper is grounded in the seq2seq network. This framework comprises an LSTM encoder, an LSTM decoder, and an LSTM classifier, integrating an attention mechanism. The un-

derlying principles are delineated as follows.

The data are input into LSTM cell 1 to commence, wherein the attention-LSTM network facilitates feature extraction from the input data. The calculation process is outlined in (18) and (19).

$$\boldsymbol{h}_t^{\text{enc}} = f^{LSTM}(\boldsymbol{h}_{t-1}^{\text{enc}}, \boldsymbol{x}_t, \theta_f) \tag{18}$$

$$\boldsymbol{c}^{\text{enc}} = \sum_{t=\tau}^{n} \boldsymbol{h}_t^{\text{enc}} \odot f^{\text{softmax}}(\boldsymbol{h}_t^{\text{enc}} \boldsymbol{h}_\tau^{\text{enc}}) \tag{19}$$

where f^{LSTM} is the calculation formula of the LSTM; $\boldsymbol{h}_t^{\text{enc}}$ is the hidden state calculated by the LSTM encoder at time t; and $\boldsymbol{c}^{\text{enc}}$ is the local context determined by the attention mechanism with a length of $T-\tau$, T is the length of the input sequence, and τ is the modulated length of the local context.

On the one hand, the feature-extracted data are fed into LSTM cell 2 to facilitate domain classification. The goal is to accurately discern whether the input data originate from the source or target domains. Subsequently, the minimization of classification errors is achieved by utilizing a gradient inversion layer.

$$\boldsymbol{h}_t^{\text{class}} = f^{LSTM}(\boldsymbol{h}_{t-1}^{\text{class}}, \boldsymbol{c}^{\text{enc}}, \theta_2) \tag{20}$$

where $\boldsymbol{h}_t^{\text{class}}$ is the hidden layer output of the predictor.

On the other hand, the feature-extracted data are channeled into LSTM cell 3 to undergo regression analysis, with the primary objective of reducing regression errors. In order to concurrently mitigate both regression and classification errors, an iterative process is pursued until the stipulated accuracy criteria are satisfied.

$$\boldsymbol{h}_t^{\text{dec}} = f^{LSTM}(\boldsymbol{h}_{t-1}^{\text{dec}}, \boldsymbol{c}^{\text{enc}}, \theta_1) \tag{21}$$

where $\boldsymbol{h}_t^{\text{dec}}$ is the hidden layer output of the classifier.

The outputs of LSTM cells 2 and 3 can be calculated using (5)-(10). Ultimately, the loss functions of these two outputs are amalgamated through the utilization of a gradient reversal layer.

C. Fine-tuning

Fine-tuning constitutes a crucial research method within the domain of transfer learning. It entails additional training for a specific task leveraging an already trained model, enhancing the alignment of the model with the task requirements. Fine-tuning strategically situates the model parameters through pre-training for more favorable initialization, resulting in time savings during subsequent training phases. Consequently, it has garnered extensive adoption in supervised learning [37].

Fine-tuning is imperative for the proposed method in this paper to refine the mapping between the source and target domains and enhance the forecasting accuracy. In this regard, the parameters of the LSTM encoder, having undergone the initial training round, are frozen and adopted as the foundational parameters for the subsequent round of fine-tuning training. The calculation formula for this process is provided by:

$$\boldsymbol{h}_t^{\text{enc}\prime} = f^{LSTM}(\boldsymbol{h}_{t-1}^{\text{enc}\prime}, \boldsymbol{x}_t, \theta_f') \tag{22}$$

where θ_f' represents the feature extractor parameters subsequent to the initial round of training; and $\boldsymbol{h}_t^{\text{enc}\prime}$ is the output of the hidden layer following the process of fine-tuning.

IV. Case Study

The case study in this paper employs residential load data sourced from the New England region of the United States, spanning from January 2022 to June 2023, as supplied by Independent System Operator (ISO) New England. This dataset encompasses six states and includes eight sets of temporal and load data [38], each recorded with a time resolution of 1 hour. Focusing on New Hampshire as the target domain, the data are partitioned into training, validation, and testing subsets with a ratio of 6:2:1. To emulate the scarcity of load data within the newly established power consumption region, the entirety of load data in the dataset is utilized as the source domain data, encompassing the timeframe from January 2022 to June 2023. For the target domain data, records from the initial half of 2023 are selected, thus facilitating further exploration through example analysis.

The experimental hardware platform comprises an Intel Core i7-11700 CPU and a NVIDIA GeForce GTX 1650 GPU. The implementation is carried out in Python 3.9, while the essential functionalities are realized by utilizing the Keras deep learning library within the TensorFlow framework.

A. Evaluation Metrics

Before embarking on experimental analysis, it is essential to choose evaluation metrics to assess the experimental outcomes. Among various evaluation metrics, the relative mean square error (RMSE) and mean absolute percentage error (MAPE) [39] stand out as widely employed indicators in the forecasting domain, delivering optimal evaluation results. Among these, MAPE offers insight into the general average performance of the forecasting models, while RMSE quantifies the spread of deviation between the projected load and the observed load [40]. They effectively capture the disparity between forecasted and actual values. Consequently, this paper opts for these two criteria to evaluate the forecasting results, which are calculated as:

$$y_{\text{RMSE}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{23}$$

$$y_{\text{MAPE}} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{24}$$

where $\hat{y}_i$ is the forecasted value; y_i is the actual value; and m is the number of forecasted data.

B. K-shape Clustering

Initially, the load data exhibiting the highest similarity must undergo screening via cluster analysis. This paper conducts a comparative analysis with *K*-means clustering to validate the effectiveness of the previously proposed *K*-shape clustering. Specifically, the procedure of *K*-shape clustering is outlined as follows.

1) Determine the number of clusters via the elbow method. The unsupervised clustering method delineates distinct categories based on the magnitude of mean square error (MSE) [41]. The elbow method is employed to ascertain the optimal cluster count by observing the trajectory of the in-

crease of MSE. Starting from a single cluster, the cluster count progressively rises. When the number of clusters reaches the optimum, the MSE experiences a pronounced decline, eventually reaching a plateau.

Cluster analysis is conducted on the dataset corering the first half 2023 from the set of eight load data sequences. The elbow method is employed to ascertain the optimal cluster count. The test results are depicted in Fig. 7.

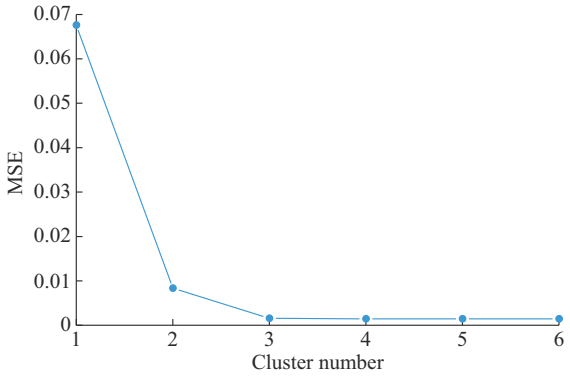

Fig. 7. Test results of elbow method.

As depicted in Fig. 7, the MSE maintains a consistent level after reaching a cluster count of 3. Consequently, this paper opts for a cluster count of 3. Within this arrangement, the data of Vermont and the Midwest of Massachusetts each belong to distinct categories, while the data from the remaining regions are allocated to the third category.

2) Utilize the SBD metric for determining the optimal matching source domain data. This entails conducting additional assessment on source domain data within the same cluster as the target domain data. Subsequently, calculate the SBD distance between each potential candidate data point and the target domain data, as defined by (3).

The *K*-means clustering employs the ED to assess the effectiveness, facilitating the selection of optimal source domain data [42]. The SBD and ED between each dataset and the New Hampshire data can be found in Table I.

TABLE I
RESULTS OF SBD AND ED CALCULATION

Cluster number	Area	SBD (10^{-4})	ED
1	Vermont	201.91	29.90
2	Midwestern Massachusetts	140.49	38.04
3	Connecticut	19.29	7.71
3	Rhode Island	13.19	8.26
3	Northeast Massachusetts	10.30	6.90
3	Southeastern Massachusetts	28.93	10.19
3	Maine	22.07	6.58

It can be observed that the SBD between the load data in Northeast Massachusetts and the target data is minimized, while the ED between the load data in Maine and the target data is minimized. Further experimental analysis is required.

3) Confirm the accuracy of clustering results. Various datasets are chosen as source domain data to validate the forecasting efficacy to substantiate the accuracy and efficacy of the *K*-shape, as mentioned earlier.

As mentioned earlier, we have conducted *K*-shape verification experiments on data from different regions, and the results are presented in Table II.

TABLE II
RESULTS OF *K*-SHAPE VERIFICATION EXPERIMENT

Area	RMSE (MW)	MAPE (%)
Vermont	58.84	3.27
Midwestern Massachusetts	57.14	3.34
Connecticut	34.88	2.16
Rhode Island	42.35	2.59
Northeast Massachusetts	31.04	1.93
Southeastern Massachusetts	43.55	2.60
Maine	31.21	2.50

It can be observed from Table II that the experimental groups corresponding to clusters 1 and 2 exhibit the most substantial forecasting errors, characterized by RMSE values exceeding 57 MW and MAPE values surpassing 3%. Conversely, the experimental group associated with cluster 3 showcases notably diminished forecasting errors compared with the initial two clusters. Furthermore, a discernible trend emerges in which lower SBD values coincide with reduced forecasting errors. Notably, the SBD in the Northeast Massachusetts is the smallest, indicating the lowest error. Meanwhile, the ED in Maine is the smallest. However, its error is slightly larger than the former, highlighting the greater accuracy and effectiveness of *K*-shape clustering compared with *K*-means clustering. To sum up, the present experiment leverages the dataset from Northeastern Massachusetts as the source domain data and the dataset from New Hampshire as the designated target domain data for the conducted study.

C. Comparative Experiment

In order to validate the enhanced efficacy of the proposed method relative to alternative forecasting methods, a comparative experiment is undertaken. Specifically, the comparative experiments encompass a range of contemporary forecasting models including SVM, feedforward neural network (FNN), extreme gradient boosting (XGBoost), gradient boosting decision tree (GBDT), random forest (RF), and gated recurrent unit (GRU). Additionally, to comprehensively assess the effectiveness of all the refined components within the method, ablation experiments are conducted utilizing LSTM, seq2seq, and DATN. Among them, SVM and FNN are prevalent traditional machine learning models with simple principles, which can achieve favorable forecasting results in straightforward forecasting scenarios [43]. XGBoost, GBDT, and RF are representative ensemble learning models based on the integration training of multiple decision trees, allowing to exhibit strong generalization capabilities [44]. GRU, LSTM, and seq2seq are widely applied deep learning models nowadays, known for their capability to capture intricate features in time series data [13]. DATN is a transfer learning model based on domain adversarial training. All these models are

widely used in contemporary forecasting. Using them as comparative experiments can effectively validate the superiority of the proposed method in this paper. The comparative experiment results are presented in Table III, while the comparison of forecasting results of a consecutive week are visually depicted in Fig. 8.

TABLE III
COMPARATIVE EXPERIMENT RESULTS

Experiment type	Number	Model	RMSE (MW)	MAPE (%)
Proposed method	1	Improved DATN	27.97	1.92
Comparative experiment	2	SVM	42.72	3.04
	3	FNN	81.26	5.23
	4	XGBoost	51.96	3.83
	5	GBDT	50.44	3.65
	6	RF	50.96	3.50
	7	GRU	35.67	2.23
Ablation experiment	8	LSTM	50.15	3.27
	9	seq2seq	41.97	2.69
	10	DATN	58.90	3.78

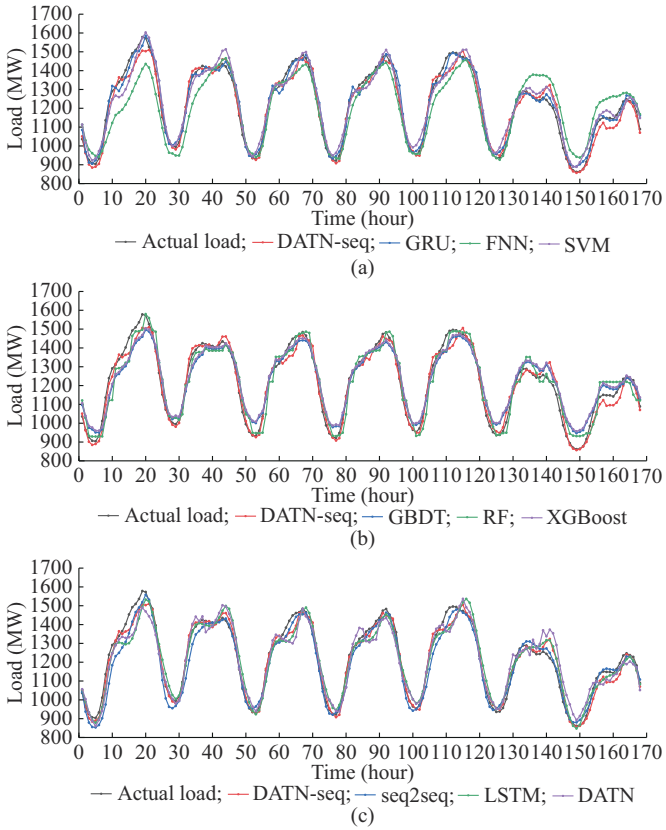

Fig. 8. Comparison of forecasting results of a consecutive week. (a) Forecasting results of the first set of comparative experiments. (b) Forecasting results of the second set of comparative experiments. (c) Forecasting results of ablation experiments.

The data in Table III underscore a notable contrast in performance between the proposed and the comparative experiment methods. Notably, the proposed method exhibits a reduction in RMSE of 14.75 MW, 53.29 MW, 23.99 MW, 22.47 MW, 22.99 MW, 7.7 MW, 22.18 MW, 14 MW, and 30.93 MW across the comparative models. A corresponding decrease in MAPE of 1.12%, 3.31%, 1.91%, 1.73%, 1.58%, 0.31%, 1.35%, 0.77%, and 1.86% is observed. Figure 8 visually illustrates that the proposed method in this paper excels in forecasting peak and trough values, thus yielding a more accurate overall forecasting outcome.

In the context of ablation experiments, a comparative analysis reveals that exclusively employing the LSTM or seq2seq networks in the first two groups yields suboptimal forecasting outcomes due to the limitations of the domain adaptation. This underscores the shortcomings of deep learning when data availability is insufficient. Furthermore, juxtaposed with the final group, it becomes evident that leveraging the traditional DATN for forecasting facilitates the data transfer from the source domain to the target domain. However, this method has a diminished capability to comprehend time series data, leading to inferior forecasting outcomes compared with the proposed method in this paper. The proposed method, on the other hand, effectively capitalizes on both data migration advantages and deep learning, thereby achieving superior forecasting results.

To delve deeper into the forecasting efficacy of the proposed method in this paper, Fig. 9 presents a violin plot of the forecasting results. This plot effectively visualizes the distribution characteristics and probability density across multiple datasets. The medians, depicted as white points within the plots, and interquartile ranges, symbolized by central black bars, are distinctly represented. Additionally, the width of the violin indicates the number of data distributions [13].

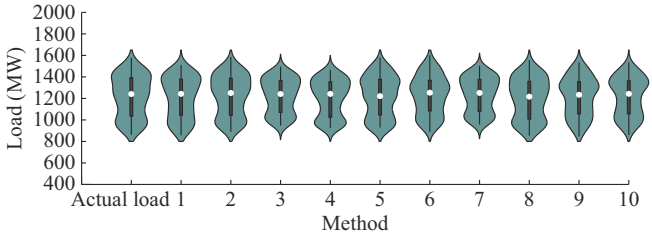

Fig. 9. Violin plot of forecasting results.

As depicted in Fig. 9, the data distribution of the forecasting outcomes generated by the proposed method in this paper closely resembles the actual load data distribution. Notably, the superiority of the forecasting capabilities of proposed method is discernible through the aspect of data distribution.

V. CONCLUSION

This paper introduces an enhanced short-term residential load forecasting method based on *K*-shape clustering and DATN to address the challenge of load forecasting stemming from limited data availability among new power users. Initially, the load data undergo screening via *K*-shape clustering. Subsequently, a seq2seq network is formed by employing an LSTM encoder and decoder architecture integrated with an attention mechanism. This amalgamation facilitates the roles of feature extraction, domain classification, and la-

bel forecasting. A fusion is performed between the source domain data identified through clustering and the target domain data to realize the migration network function within the DATN. Subsequent fine-tuning of the ultimate forecasting outcomes yields the final target domain forecasting results.

In case study, varying datasets are initially employed as source domain data, thereby validating the efficacy of *K*-shape clustering. Subsequently, a comparative evaluation is conducted between the proposed method in this paper and diverse forecasting methods. The findings underscore the superior forecasting performance of the proposed method. Moreover, the efficacy of each constituent is substantiated through ablation experiments. Compared with alternative methods, the proposed method achieves a notable reduction in RMSE ranging from 7.7 MW to 53.29 MW and a corresponding reduction in MAPE from 0.31% to 3.31%.

However, this paper still has certain limitations. For instance, in practical engineering scenarios, the source domain data are often substantial, while the source domain data used in this paper are relatively limited. In future research, we will contemplate employing more extensive source domain data for experimentation. Additionally, this paper focuses on forecasting the load for large regional areas. In contrast, newly constructed residences are more often represented in smaller clusters such as residential communities or individual buildings. Future endeavors will involve electrical load forecasting for smaller-scale entities, aiming to enhance the forecasting accuracy.

REFERENCES

[1] S. H. Rafi, S. R. Deeba, and E. Hossain, "A short-term load forecasting method using integrated CNN and LSTM network," *IEEE Access*, vol. 9, pp. 32436-32448, Feb. 2021.

[2] C. Peng, Y. Tao, Z. Chen *et al.*, "Multi-source transfer learning guided ensemble LSTM for building multi-load forecasting," *Expert Systems with Applications*, vol. 202, p. 117194, Apr. 2022.

[3] Y. Dong, S. Ma, H. Zhang *et al.*, "Wind power prediction based on multi-class autoregressive moving average model with logistic function," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 5, pp. 1184-1193, Sept. 2022.

[4] X. Zhou, Y. Gao, W. Yao *et al.*, "A robust segmented mixed effect regression model for baseline electricity consumption forecasting," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 1, pp. 71-80, Jan. 2022.

[5] M. Miranbeigi and H. Iman-Eini, "Hybrid modulation technique for grid-connected cascaded photovoltaic systems," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 12, pp. 7843-7853, Dec. 2016.

[6] Y. Cao, Q. Li, Y. Tan *et al.*, "A comprehensive review of energy internet: basic concept, operation and planning methods, and research prospects," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 3, pp. 399-411, May 2018.

[7] W. van Deventer, E. Jamei, G. S. Thirunavukkarasu *et al.*, "Short-term PV power forecasting using hybrid GASVM technique," *Renewable Energy*, vol. 140, pp. 367-379, Sept. 2019.

[8] D. Bhatt, C. Patel, H. Talsania *et al.*, "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics*, vol. 10, no. 20, p. 2470, Sept. 2021.

[9] Z. Li, F. Liu, W. Yang *et al.*, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999-7019, Dec. 2021.

[10] J. Zhu, H. Dong, W. Zheng *et al*., "Review and prospect of data-driven techniques for load forecasting in integrated energy systems," *Applied Energy*, vol. 321, p. 119269, Sept. 2022.

[11] W. Kong, Z. Y. Dong, Y. Jia *et al*., "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841-851, Jan. 2019.

[12] N. Somu, G. Raman, and K. Ramamritham, "A deep learning framework for building energy consumption forecast," *Renewable and Sustainable Energy Reviews*, vol. 137, p. 110591, Mar. 2021.

[13] H. Dong, J. Zhu, S. Li *et al*., "Short-term residential household reactive power forecasting considering active power demand via deep transformer sequence-to-sequence networks," *Applied Energy*, vol. 329, p. 120281, Jan. 2023.

[14] E. Skomski, J. Y. Lee, W. Kim *et al.*, "Sequence-to-sequence neural networks for short-term electrical load forecasting in commercial office buildings," *Energy and Buildings*, vol. 226, p. 110350, Nov. 2020.

[15] G. Li, F. Li, T. Ahmad *et al*., "Performance evaluation of sequence-to-sequence-attention model for short-term multi-step ahead building energy predictions," *Energy*, vol. 259, p. 124915, Nov. 2022.

[16] D. Wu, B. Wang, D. Precup *et al*., "Multiple kernel learning-based transfer regression for electric load forecasting," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1183-1192, Mar. 2020.

[17] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, Feb. 2015, pp. 1180-1189.

[18] M. Huang and J. Yin, "Research on adversarial domain adaptation method and its application in power load forecasting," *Mathematics*, vol. 10, no. 18, p. 3223, Aug. 2022.

[19] S. Li, Y. Zhong, and J. Lin, "AWS-DAIE: incremental ensemble short-term electricity load forecasting based on sample domain adaptation," *Sustainability*, vol. 14, no. 21, p. 14205, Oct. 2022.

[20] F. Han, T. Pu, M. Li *et al*., "Short-term forecasting of individual residential load based on deep learning and *k*-means clustering," *CSEE Journal of Power and Energy Systems*, vol. 7, no. 2, pp. 261-269, Dec. 2020.

[21] X. Wang, W. J. Lee, H. Huang *et al.*, "Factors that impact the accuracy of clustering-based load forecasting," *IEEE Transactions on Industry Applications*, vol. 52, no. 5, pp. 3625-3630, Apr. 2016.

[22] C. Peng, Y. Tao, Z. Chen *et al.*, "Multi-source transfer learning guided ensemble LSTM for building multi-load forecasting," *Expert Systems with Applications*, vol. 202, p. 117194, Sept. 2022.

[23] J. Paparrizos and L. Gravano, "*K*-shape: efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, Melbourne, Australia, Jan. 2015, pp. 1855-1870.

[24] L. Yang and Z. Zhang, "A deep attention convolutional recurrent network assisted by *k*-shape clustering and enhanced memory for short term wind speed predictions," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 2, pp. 856-867, Dec. 2021.

[25] J. Yang, C. Ning, C. Deb *et al*., "*K*-shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement," *Energy and Buildings*, vol. 146, pp. 27-37, Jul. 2017.

[26] C. Si, S. Xu, C. Wan *et al*., "Electric load clustering in smart grid: methodologies, applications, and future trends," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 2, pp. 237-252, Mar. 2021.

[27] Y. Li, D. Han, and Z. Yan, "Long-term system load forecasting based on data-driven linear clustering method," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 2, pp. 306-316, Jan. 2017.

[28] K. P. Sinaga and M. S. Yang, "Unsupervised *k*-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716-80727, May 2020.

[29] K. Cho, B. V. Merrienboer, C. Gulcehre *et al*. (2014, Sept.). Learning phrase representations using RNN encoder-decoder for statistical machine translation. [Online]. Available: http://arxiv.org/abs/1406.1078

[30] K. Cho, B. V. Merrienboer, D. Bahdanau *et al*. (2014, Oct.). On the properties of neural machine translation: encoder-decoder approaches. [Online]. Available: http://arxiv.org/abs/1409.1259

[31] H. Zheng, J. Yuan, and L. Chen, "Short-term load forecasting using EMD-LSTM neural networks with a XGboost algorithm for feature importance evaluation," *Energies*, vol. 10, no. 8, pp. 1168, Jul. 2017.

[32] C. Li, Z. Dong, L. Ding *et al*., "Interpretable memristive LSTM network design for probabilistic residential load forecasting," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 6, pp. 2297-2310, Jun. 2022.

[33] X. Zhang, S. Kuenzel, and N. Colombo, "Hybrid short-term load forecasting method based on empirical wavelet transform and bidirectional long short-term memory neural networks," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 5, p. 1216-1228, Sept. 2022.

[34] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48-62, Sept. 2021.

[35] X. Yang, C. Deng, T. Liu *et al.*, "Heterogeneous graph attention network for unsupervised multiple-target domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1992-2003, Apr. 2020.
[36] M. Wang and W. Deng, "Deep visual domain adaptation: a survey," *Neurocomputing*, vol. 312, pp. 135-153, Oct. 2018.
[37] Z. Cao, M. Kwon, and D. Sadigh, "Transfer reinforcement learning across homotopy classes," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2706-2713, Apr. 2021.
[38] ISO New England Inc. (2023, Jun.). Energy, load, and demand reports. [Online]. Available: https://www.iso-ne.com/
[39] Y. Guo, Y. Li, and X. Qiao, "BiLSTM multitask learning-based combined load forecasting considering the loads coupling relationship for multienergy system," *IEEE Transactions on Smart Grid*, vol. 13, no. 5, pp. 3481-3492, Sept. 2022.
[40] Q. Cui, J. Zhu, J. Shu *et al*., "Comprehensive evaluation of electric power prediction models based on D-S evidence theory combined with multiple accuracy indicators," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 3, pp. 597-605, May 2022.
[41] J. Qin, W. Fu, H. Gao *et al.*, "Distributed *k*-means algorithm and fuzzy *c*-means algorithm for sensor networks based on multiagent consensus theory," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 772-783, Mar. 2017.
[42] K. P. Sinaga and M. Yang, "Unsupervised *k*-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716-80727, Jul. 2020.
[43] A. A. Mamun, M. Sohel, N. Mohammad *et al.*, "A comprehensive review of the load forecasting techniques using single and hybrid predictive models," *IEEE Access*, vol. 8, pp. 134911-134939. Aug. 2020.
[44] W. You, D. Guo, Y. Wu *et al*., "Multiple load forecasting of integrated energy system based on sequential-parallel hybrid ensemble learning," *Energies*, vol. 16, no. 7, p. 3268, Apr. 2023.

Jizhong Zhu received the B.S., M.S., and Ph.D. degrees in electrical engineering from Chongqing University, Chongqing, China, in 1985, 1987, and 1990, respectively. He is a Fellow of IEEE and a Professor at the School of Electric Power Engineering, South China University of Technology, Guangzhou, China. His research interests include power system operation and control, smart grid, microgrid, virtual power plant, electric vehicle, renewable energy application, and integrated smart energy system.

Yuwang Miao received the B.E. degree in electrical engineering and automation from Xiamen University, Xiamen, China, in 2021. He is currently working towards the M.S. degree of electrical engineering in South China University of Technology, Guangzhou, China. His research interest includes load forecasting of power system based on artificial intelligence.

Hanjiang Dong received the B.S. degree in electrical engineering and its automation from Jinan University, Guangzhou, China, in 2020. He is currently working toward the joint Ph.D. degree with the School of Electric Power Engineering, South China University of Technology, Guangzhou, China, and the Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong, China. His current research interests include applications of artificial intelligence in power systems such as in energy forecasting and economic dispatch.

Shenglin Li received the B.S. and M.S. degrees from Shanghai University of Electric Power, Shanghai, China, in 2016 and 2019, respectively, and the Ph.D. degree in electrical engineering from South China University of Technology, Guangzhou, China, in 2023. He is currently working as a Post-doctoral Fellow in South China University of Technology. His research interests include demand-side energy management, optimal dispatch in microgrids and distribution networks.

Ziyu Chen received the B.S. degree in electrical engineering from Guangdong University of Technology, Guangzhou, China, in 2017, and the Ph.D. degree in electrical engineering from South China University of Technology, Guangzhou, China, in 2022. She is currently a Postdoctoral and Research Assistant in electrical engineering from South China University of Technology. Her current research interests include cyber-physical-social system, operation and control of smart energy system, and cyber-attack detection.

Di Zhang received the B.S. degree from Henan Normal University, Xinxiang, China, in 2014, and the Ph.D. degree in electrical engineering from Hunan University, Changsha, China, in 2022. From 2018 to 2020, she was a Visiting Ph.D. Student with the Robert W. Galvin Center for Electricity Innovation, Illinois Institute of Technology, Chicago, USA. She is currently working as a Postdoctoral Researcher in South China University of Technology, Guangzhou, China. Her major research interests include power system operation and control, and application of artificial intelligence in power systems.