Applied Soft Computing 142 (2023) 110335

Contents lists available at ScienceDirect

Applied Soft Computing

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

Optimized Seq2Seq model based on multiple methods for short-term power load forecasting

Yeming Dai [a,*], Xinyu Yang [a], Mingming Leng [b]

[a] School of Business, Qingdao University, Qingdao 200071, China
[b] Faculty of Business, Lingnan University, Hong Kong, China

ARTICLE INFO

Article history:
Received 19 October 2022
Received in revised form 7 April 2023
Accepted 12 April 2023
Available online 24 April 2023

Keywords:
Power load forecasting
Convolutional neural network
Attention mechanism
Sequence to Sequence
Bidirectional long-short term memory network
Bayesian optimization

ABSTRACT

Accurate power load prediction plays a key role in reducing resource waste and ensuring stable and safe operations of power systems. To address the problems of poor stability and unsatisfactory prediction accuracy of existing prediction methods, in this paper, we propose a novel approach for short-term power load prediction by improving the sequence to sequence (Seq2Seq) model based on bidirectional long-short term memory (Bi-LSTM) network. Different from existing prediction models, we apply convolutional neural network, attention mechanism, and Bayesian optimization for the improvement of the Seq2Seq model. Moreover, in the data processing stage, we use the random forest algorithm for feature selection, and also adopt the weighted grey relational projection algorithm for holiday load processing to process the data and thereby overcome the difficulty of holiday load prediction. To validate our model, we choose the power load dataset in Singapore and Switzerland as experimental data and compare our prediction results with those by other models to show that our method can generate a higher prediction accuracy.

1. Introduction

Electricity is significantly important to the sustained, stable, and healthy development of human society as a cost-effective energy source. However, the current technology cannot realize the large-scale storage of electricity, since the electricity production, transmission, distribution, and consumption occur almost simultaneously. This behooves power suppliers to make reasonable schedules for smooth operations of power systems so as to achieve a balance between power supply and demand [1,2]. Accurate power load forecasting can optimize the scheduling control of power grid and achieve smooth operations of power systems, which can then result in higher economic and social benefits. However, a decrease in thermal power generation, an increase in renewable energy generation, and a massive increase in electric vehicle usage have put the supply and demand balance of power system to a new test and made it more difficult to predict the power load. It thus follows that we need more sophisticated power load forecasting technologies to improve the accuracy of load forecasting and then better cope with the increasingly variable power consumption environment nowadays [3].

In practice, power load forecasting can be divided into long-, medium-, and short-term predictions [4,5]. The short-term power load prediction is more important to efficient operations of power systems and the reduction of power wastes, which serves as an important decision basis for stable and sustainable developments of power systems. The short-term power load forecasting approaches mainly belong to two categories: traditional methods and machine learning methods [6]. Traditional short-term methods mainly include time series analysis [7], Kalman filtering algorithm [8], gray models [9], etc. Those methods can process linear data effectively; but, their prediction results are not good if data is nonlinear. Machine learning methods mainly include random forest algorithm, support vector machine (SVM), gradient boosting decision tree, and artificial neural network (ANN), etc., which are more suitable for prediction of nonlinear data [10–12]. As ANN has a strong self-learning ability and a generalization ability, it has been regarded a highly-effective machine learning method and thus has been widely used for power load forecasting [13,14]. Nonetheless, ANN method often cannot obtain stable prediction results when dealing with time series problems, because it has disadvantages such as overfitting, slow learning rate, and tendency to fall into local minima [15].

As an extension of ANN, recurrent neural network (RNN) can deliver better prediction results when processing time series data because of its circular structure. However, some issues still exist in vanishing gradient and gradient explosion in the training of RNN. To overcome these issues, Atef and Eltawil [16] proposed a variant of RNN, which was called "long short-term memory network" (LSTM). Compared with RNN, LSTM has unique

* Corresponding author.
E-mail addresses: yemingdai@163.com (Y. Dai), 865050851@qq.com (X. Yang), mmleng@ln.edu.hk (M. Leng).

https://doi.org/10.1016/j.asoc.2023.110335
1568-4946/

gate structure that can solve the vanishing gradient and gradient explosion problems. Nevertheless, the time series prediction analysis based on LSTM still cannot achieve satisfactory prediction results, because it only considers historical information and ignores future information [17]. To better apply LSTM for power load forecasting, Wang et al. [18] applied the Bi-LSTM for power load prediction and obtained better prediction results. The Bi-LSTM is a combination of two LSTMs i.e., forward LSTM and backward LSTM, and this structure can make the most of historical and future information to present more accurate prediction results [19]. In addition to make the LSTM better capture the information of temporal and spatial features of the input data when making predictions, Song et al. [20] carried out hourly heating load prediction by combining convolutional neural network (CNN) and LSTM, and the results showed that spatial feature information extraction of input data by CNN can improve the accuracy of model prediction.

In recent years, the Seq2Seq model with high flexibility in dealing with time series has been proposed. As the Seq2Seq model can effectively solve the relationship between different length sequences and dynamically determine the steps of network structure, many researchers have used the model to work on the power load prediction-related issues [21]. In addition, the Seq2Seq model can deal with multi-variable predictions well, because it has an Encoder–Decoder structure network where the encoding and decoding layers are composed of deep neural network (such as RNN or LSTM) [22,23]. Moreover, since LSTM has an outstanding performance in solving the gradient explosion problem, it has been used as the encoder and decoder of the Seq2Seq model [24]. Particularly, a Bi-LSTM-based Seq2Seq model for day-ahead peak load predictions has been proposed to effectively improve the prediction accuracy [25]. To improve the prediction performance on the data sequences of the Seq2Seq model, researchers have applied the attention mechanism to solve the shortcomings of the Seq2Seq model since the core idea of attention mechanism is to dynamically adjust the weights between different factors and highlight data sequences [26].

In addition to applying more advanced models for power load forecasting, the proper processing of original data and the hyperparameter optimization of prediction model can further improve the prediction accuracy [27]. The power load dataset is a kind of high latitude time series data which is affected by multiple factors such as electricity price, temperature, time, holidays, etc. To eliminate the influence of redundant features on load prediction, we consider the commonly used feature extraction methods at this stage that include minimum redundancy maximum correlation (mRMR) [28], principal component analysis (PCA) [29], independent component analysis (ICA) [30], and random forest. However, the methods above such as PCA and ICA have better results for feature extraction for linear data but often fail to achieve the desired results when processing nonlinear power load data [31]. Although mRMR can effectively process nonlinear data, it is usually used for feature extraction on datasets with small data volumes. In contrast, as a decision tree based nonlinear method, the random forest algorithm not only captures the nonlinear relationship between features and target variables and indicates the dependencies between features, but also automatically determines the importance of features as well as being able to effectively handle complex datasets with high latitude [32]. Moreover, as the power load data in holidays is rather different from regular datasets, the load data in holidays need to be processed prior to the prediction. The weighted grey relational projection (WGRP) algorithm has been used for processing holiday datasets in many recent studies [33].

The choice of hyperparameters is critical in many machine learning based prediction models, and assigning appropriate hyperparameters can improve the predictive performance and the model accuracy [34]. However, many researchers rely on their experience and repeated experiments to perform hyperparameterization, which requires significant computational resources. Therefore, many optimization methods such as, particle swarm optimization (PSO) algorithm, grid search, random search and genetic algorithm (GA), are used for parametric optimization [35,36]. Nonetheless, those traditional optimization methods are often inefficient in optimizing multiple hyperparameters, which is not only time-consuming but also less effective. As an emerging hyperparameter optimization method, the Bayesian optimization (BO) algorithm allows for a more extensive exploration of the space of hyperparameters, consumes short running time, and presents good optimization results when optimizing multiple hyperparameters [37,38].

After reviewing the existing power load forecasting models, data processing methods, and optimization methods, we propose a new power load forecasting method different from previous studies. First, in the data process stage, we use the random forest algorithm for feature extraction which can eliminate redundant features, and then use the WGRP algorithm to select similar data for holidays so that the holiday data can be generalized. Secondly, we apply the CNN, attention mechanism and BO algorithm to improve the Bi-LSTM-based Seq2Seq power load forecasting model. The contributions of this paper are summarized in the following points:

1. We develop a novel power load forecasting approach by improving the Seq2Seq model with various advanced algorithms and neural networks.
2. We use the Bi-LSTM in the encoder and decoder of Seq2Seq model to enhance the information utilization. By combination with the Seq2Seq model, the CNN is used to extract the input data features and BO algorithm is applied to optimize the hyperparameters of the model.
3. We introduce the attention mechanism in the Seq2Seq model to help the decoder focus on key sequence information that influences prediction results.
4. For redundant features and holiday data that are not general, we use random forest algorithm and WGRP algorithm for data processing, respectively.
5. We perform a comparative error analysis by comparing the forecast results of six forecasting models in two electricity markets. The experimental results expose the effectiveness and reliability of our model.

The remaining chapters of this paper are organized as follows. We present the methodology related to the paper in Section 2. Section 3 proposes our novel power load prediction model. In Section 4, we introduce the experimental setup and indicate the effectiveness and reliability of our model by comparing our prediction results with others from different models. This paper ends with a summary in Section 5.

2. Methodologies

2.1. Random forest

Random forest algorithm is an ensemble learning method based on bagging algorithm, which is composed of multiple decision trees in combination. The training set of all decision trees in the random forest is repeatedly and randomly selected from original samples with put-back, this process is also called bootstrap resampling. The random forest algorithm as a representative method for integrated learning is widely used in various fields, such as data classification, load prediction, and feature selection [39]. When we use the random forest algorithm for

feature importance measurement, we need to choose the appropriate method for measuring feature importance. We apply the Gini Index to measure the importance of features in this paper. For ease of expression, we use *Gini* and *VIM* to denote the Gini index and the feature importance score, respectively. We assume that there are m features, which are denoted as $X_1, X_2, \ldots, X_m$, l decision trees, and K categories. The specific steps are shown below.

(1) The *Gini* of feature $X_j, j = 1, \ldots, m$ at the *z-th* tree node q is calculated as

$$Gini_q^z = 1 - \sum_{k=1}^{K} p_{qk}^2 \tag{1}$$

where p_{qk} means the proportion of category k in node q.

(2) The importance at node q, i.e., the change of *Gini* at node q after branching, is calculated as follows:

$$VIM_{jq}^z = Gini_q^z - Gini_q^i - Gini_q^l \tag{2}$$

where $Gini_q^i$ and $Gini_q^l$ are the *Gini* of the new nodes after branching.

(3) Suppose W is the set of all nodes of the occurrence features X_j in the *z-th* tree. Then we can obtain the total feature importance of X_j in *z-th* tree as

$$VIM_j^z = \sum_{q \in W} VIM_{jq}^z \tag{3}$$

(4) As there are l decision trees, the importance of feature X_j is

$$VIM_j = \sum_{z=1}^{l} VIM_j^z \tag{4}$$

(5) We calculate the importance scores of the remaining features by repeating the above process, normalize all the obtained importance scores, and then rank them to select the best feature set.

2.2. WGRP

The WGRP algorithm is a comprehensive evaluation method that is based on grey system theory. This method addresses the limitations of the grey relational analysis method by introducing the concept of weighted sum projection. First, the key factors that have a great influence on target variables are obtained by a weighting method, and the relationship between the historical dataset and the dataset to be predicted is obtained by constructing the weighted grey correlation matrix. As a result, we can obtain the historical dataset that is similar to the forecast dataset [40]. The specific steps are as follows:

(1) We let $\mathbf{Y_0}$ and $\mathbf{Y_i}$ denote the feature vector of the data to be predicted and the feature vector of the data at day i, respectively, i.e.,

$$\mathbf{Y_0} = [y_{01}, y_{02}, \ldots, y_{0n}] \tag{5}$$

$$\mathbf{Y_i} = [y_{i1}, y_{i2}, \ldots, y_{in}], i = 1, 2, \ldots, m \tag{6}$$

where m is the total number of influencing factors, and y_{in} is the nth influencing factor for the ith sample.

(2) We take $\mathbf{Y_0}$ as the subsequence, $\mathbf{Y_i}$ as the parent sequence, and construct the gray correlation matrix $\mathbf{F}$ by calculating the coefficients of the relationship between $\mathbf{Y_0}$ and $\mathbf{Y_i}$. Then, we calculate the weight φ of each influencing factor according to the entropy weighting method, and weigh the gray correlation matrix to obtain the weighted gray correlation matrix $\mathbf{F}'$.

$$\mathbf{F} = \begin{bmatrix} F_{01} & \cdots & F_{0n} \\ \vdots & \ddots & \vdots \\ F_{m1} & \cdots & F_{mn} \end{bmatrix} \tag{7}$$

$$\boldsymbol{\lambda} = [\varphi_1, \varphi_2, \ldots, \varphi_n] \tag{8}$$

$$\mathbf{F}' = \mathbf{F}\boldsymbol{\lambda}^T = \begin{bmatrix} \varphi_1 & \cdots & \varphi_n \\ \vdots & \ddots & \vdots \\ \varphi_1 F_{m1} & \cdots & \varphi_n F_{mn} \end{bmatrix} \tag{9}$$

where F_{mn} represents the gray correlation value corresponding to the nth factor of the mth sample.

(3) Based on what is found in formula (9), we can obtain the weighted gray correlation projection value between the feature vector of the data to be predicted and the feature vector of the data on day i is

$$D_i = \frac{\sum_{j=1}^{n} \varphi_j F_{ij} \varphi_j}{\sqrt{\sum_{j=1}^{n} \varphi_j^2}} \tag{10}$$

(4) The weighted gray correlation projection values of the obtained historical data are sorted from the largest to the smallest, and the one with the larger projection value is selected as the similar dataset.

2.3. BO

When we use a model for prediction, tuning the hyperparameters of the model is essential to obtain the best prediction performance. Determining the hyperparameters of the model through experience or multiple attempts is time-consuming and often does not allow the model to perform to its full potential. Accordingly, we need to optimize the hyperparameter screening process of the model so that the model can achieve the best prediction performance. Compared to other traditional optimization methods, the BO algorithm has better optimization results, faster convergence, and less time consuming. The BO algorithm is an optimization algorithm based on probability distribution, and there are two main core processes—i.e., the prior function and the acquisition function. As usual, the prior function uses the Gaussian process regression (GP), and the acquisition function uses the expected improvement (EI) function. We illuminate the specific optimization process as in Fig. 1.

2.4. CNN

CNN is a deep neural network containing multilayer convolutional structure, which has been widely used in information retrieval, image processing and face recognition image [41]. The CNN model adopts the method of local connection and weight sharing, so it can also be used for data processing, i.e., extracting spatially informative features of the data [42]. In general, CNN is composed of convolutional layers, pooling layers, and a fully connected layer. The convolutional layer extracts the effective features of the input data through convolution, and the pooling layer optimizes the network by selecting the most representative features, thereby reducing the dimensionality [43]. In addition, the number of convolutional layers and pooling layers can be freely matched. We learn from a large number of relevant publications that CNN has been used to process time series data [44].

2.5. Bi-LSTM

As a variant of RNN, LSTM greatly solves the problems of gradient disappearance and gradient explosion when using RNN by changing its structure, namely adding multiple “gate” structures (forget gate, input gate, and output gate), and effectively controls the flow of information, which has been widely used in the prediction and classification of time series, as described by Li

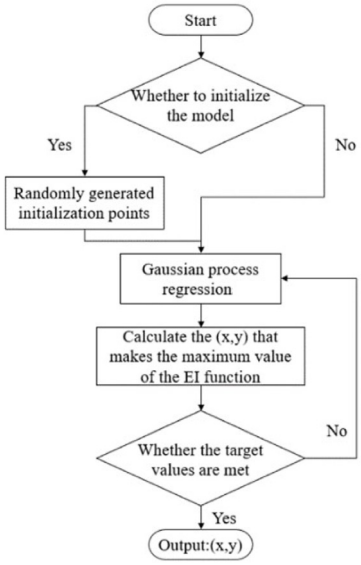

Fig. 1. The optimization process of BO algorithm.

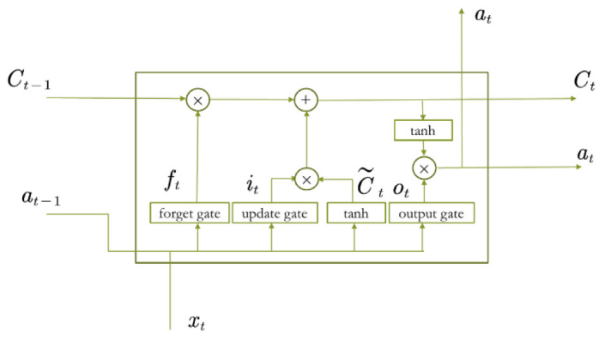

Fig. 2. LSTM unit structure.

et al. [45]. The cell structure diagram of LSTM is shown in Fig. 2, and the calculation process is as follows:

$$f_t = \sigma(W_f\,[a_{t-1}, x_t] + b_f) \tag{11}$$

$$i_t = \sigma(W_i\,[a_{t-1}, x_t] + b_i) \tag{12}$$

$$\widetilde{C}_t = \tanh(W_g\,[a_{t-1}, x_t] + b_g) \tag{13}$$

$$C_t = f_t C_{t-1} + i_t \widetilde{C}_t \tag{14}$$

$$o_t = \sigma(W_o[a_{t-1}, x_t] + b_o) \tag{15}$$

$$a_t = o_t \tanh C_t \tag{16}$$

where a_t denotes the output of the hidden layer at the current time; σ denotes the sigmoid activation function; W and b are the weight coefficient matrix and bias term, respectively; C and tanh refer to state volume and hyperbolic tangent activation function, respectively.

LSTM only considers historical information and ignores future information. Moreover, Bi-LSTM, which combines forward LSTM and backward LSTM, was proposed to effectively involve the future information (ignored by LSTM) by fitting the forward and reverse data of the sequence. Therefore, Bi-LSTM is better for load forecasting than LSTM. The structure of Bi-LSTM is shown in Fig. 3.

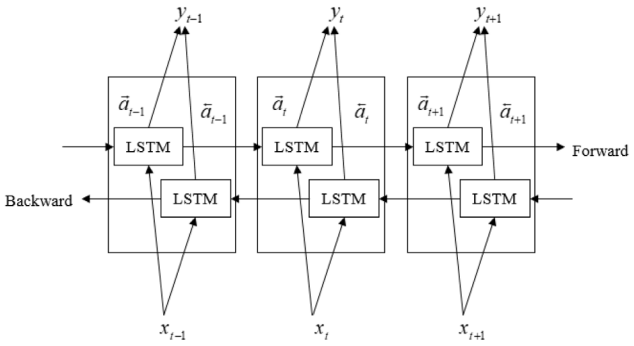

Fig. 3. The structure of Bi-LSTM.

2.6. Attention mechanism

The attention mechanism simulates the brain signal processing mechanism of human visual receiving signals. In certain situations, a person pays much attention to someone or something whereas the person reduces or ignores other things. The attention mechanism in deep learning is proposed by mimicking the brain's attention mechanism, which improves the model accuracy by assigning different probability weights to inputs to highlight important factors. In extant attention mechanisms, the most common ones are additive attention, multiplicative attention, and self-attention mechanisms [46,47]. Among them, multiplicative attention, is also known as Luong attention, uses the dot product of input and query vectors to calculate the weight of attention, which is used effectively for serial data such as time series data because it can focus on the specific parts of series and is also useful for multivariate inputs. Then, multiplicative attention is more expressive than additive attention, as the latter can only apply to pairs of inputs whereas the former can be applied to any number of inputs. Although self-attention has been widely used in transformer, it requires more complex computations and may increase the number of parameters to be learned. Therefore, this paper uses multiplicative attention whose structure and related calculation are by Niu et al. [46].

2.7. Seq2Seq model

The Seq2Seq model is an Encoder–Decoder structured model, where the input and output are sequences. The encoder turns the input sequence into a fixed-length vector expression, and then the decoder decodes the fixed-length vector to obtain the required sequence for output. The Seq2Seq model can be encoded and decoded by LSTM or RNN, and LSTM can effectively solve the problem of RNN. Therefore, both the encoder and decoder of Seq2Seq model are usually constructed by LSTM. The basic structure of Seq2Seq model is shown in Fig. 4. where h represents the state of the hidden layer in the encoder and Z represents the hidden layer state of the decoder. h_t in current state is determined by the current input X_t and h_{t-1} in the previously hidden layer state, expressed as follows:

$$h_t = f\,(h_{t-1}, X_t) \tag{17}$$

Vector $\mathbf{C}$ is determined by the state of each hidden layer and is calculated as follows:

$$\mathbf{C} = r\,(h_1, h_2, \ldots, h_n) \tag{18}$$

In the decoding process, the decoder takes the hidden vector $\mathbf{C}$ generated by the encoder, the previous hidden layer state Z_{t-1}, and the previous output y_t as inputs, and adds a nonlinear

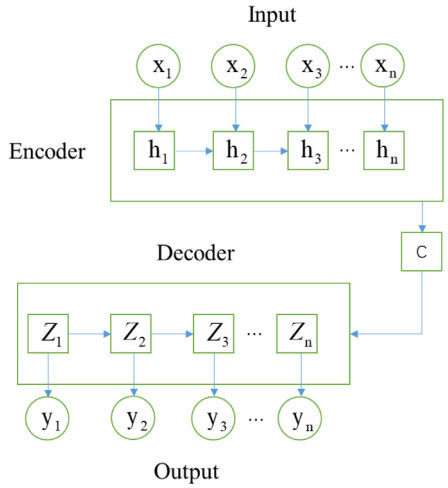

Fig. 4. Seq2Seq model.

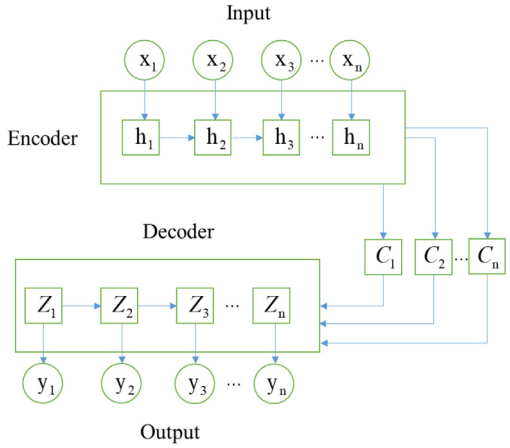

Fig. 5. Attention-Seq2Seq model.

function g to calculate the current state output y_t. The calculation formula is as follows:

$$y_t = g(y_{t-1}, Z_{t-1}, \mathbf{C}) \tag{19}$$

3. The Bi-LSTM-based Seq2Seq power load forecasting model with CNN, attention mechanism, and BO algorithm

3.1. Attention-Seq2Seq model

Although the Seq2Seq model is effective in power load prediction, it cannot focus on the data series that have a key impact on prediction results. The loss of key features is likely to cause the gradient degradation of neural network. Thus, we introduce the attention mechanism to connect each output step of the encoder and each generation step of the decoder. The attention mechanism enables the encoder to encode information of arbitrary length, effectively solving the loss of critical information due to fixed-length encoding. The Attention-Seq2Seq model structure is shown in Fig. 5.

As indicated by Fig. 5, when attention mechanism is introduced, a single vector **C** is the output according to the hidden layer state of the encoder instead of the previous one. Instead, multiple vectors are outputs according to the hidden layer state of each step, which thus ensures the integrity of the output information of the decoder. The vector **C** obtained by attention mechanism is calculated as follows:

$$a_i = \frac{\exp(e_{ij})}{\sum_{k=1}^{t} \exp(e_{ik})} \tag{20}$$

$$e_{ij} = v_a^T \tanh(W_i h_i + U_i Z_i + b_i) \tag{21}$$

$$c_i = \sum_{i=1}^{t} a_i h_i \tag{22}$$

In the formulas (20–22), h_i represents the state information of the encoding layer, Z_i represents the state information of the decoding layer, e_{ij} refers to the correlation coefficient between the ith hidden state of the input sequence and the jth hidden state of the output sequence, and v_a^T, U_i and W_i represent the correlation weight, respectively.

3.2. The Bi-LSTM-based Seq2Seq model with CNN and attention mechanism

The proposed model consists of CNN, attention mechanism, and Bi-LSTM-based Seq2Seq model. The structure of our model is shown in Fig. 6. First, the data goes through CNN for spatial feature extraction. The CNN used in this paper is composed of two one-dimensional convolutional (Conv1D) layers and two Pooling layers, each following a Conv1D layer. The Conv1D layer extracts the effective features of the input data through convolution, and the Pooling layer filters these features to reduce the complexity of the features. Then, the CNN processed data is passed to the encoder of the Seq2Seq model for encoding, and the encoder can better extract the hidden information of the input data and improve the utilization of information by using Bi-LSTM-based encoding. After that, the encoded data is passed to the attention mechanism, which enables the encoder to encode information of arbitrary length, so as to effectively solve the problem of key information loss caused by fixed-length encoding. Finally, the data after a series of processing is entered into the decoder for decoding and prediction; and, by using the Bi-LSTM-based decoder, we can better analyze the information to make the power load prediction more accurate.

3.3. The specific stages for our prediction method

Based on the prediction model developed in Section 3.2, the specific forecasting process in this paper is shown in Fig. 7, and includes into three main stages:

Stage 1: Data processing. First, we rank the importance of features using the random forest algorithm approach and eliminate redundant features by considering various factors affecting power load, such as electricity price, temperature, holidays, etc. Secondly, the load data is distinguished between holidays and non-holidays. The holiday data is processed by the WGRP algorithm, which can help generalize the data. Finally, we normalize the data.

Stage 2: Model training. The data processed in the first stage is divided into training set and test set. The training set is brought into our model for training. Then, we use the test set to verify our model. Meanwhile, in order to reduce the training time of the model as well as to fully utilize the performance of the prediction model, we optimize the hyperparameters of the model by the BO algorithm.

Stage 3: Evaluation of prediction results. Evaluate the forecasting performance of the model by using various error analysis methods and compare our prediction results with various advanced models. The results show that our method has the smallest prediction error as well as the best prediction results.

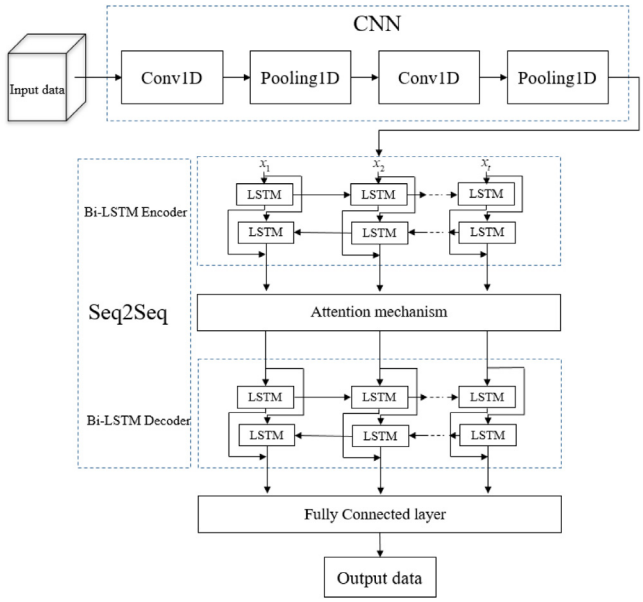

Fig. 6. The structure of CNN and attention mechanism for the improved, Bi-LSTM-based Seq2Seq model.

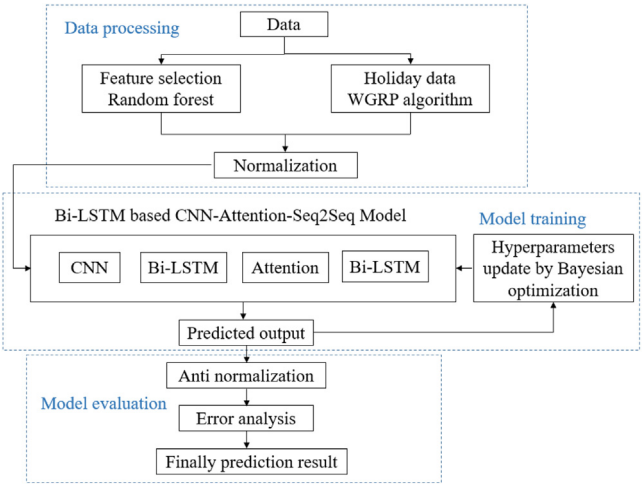

Fig. 7. Power load prediction process.

4. Case study

4.1. Experimental setup

4.1.1. Datasets

We use the electricity market datasets of Singapore and Switzerland as the experimental datasets. The electricity data of Singapore is collected from Energy Market Company Pte Ltd (EMC) and these data are available from the official website of EMC. The dataset is collected for the period from January 1, 2019 to June 30, 2021, and the dataset is sampled for 1 h. The Swiss electricity load data is collected from the ENTSO-E Transparency Platform (Entsoe), also designed for a sampling interval of 1 h. In addition, we need to collect various factors that affect the electricity load, which include real-time prices, holiday type, hourly, day type, dew point temperature, wind speed and temperature. The weather information for Singapore and Switzerland comes from the Nasa weather database, and the real-time electricity prices for Singapore and Switzerland come from EMC and Entsoe, respectively. Moreover, since the sampling interval of the collected data is one hour, we perform the hourly load prediction in this paper.

4.1.2. Data preprocess

After obtaining the data through the public platform, we first check the dataset to find possible problems such as missing values, outliers and duplicate data, and the missing values that are filled by linear interpolation. Secondly, to facilitate the training of the model as well as to reduce the computation time, we restrict the dataset to [0,1] using the normalization method. The data from Singapore from January 1, 2019 to December 31, 2020 is divided into training and validation sets in a ratio of 9:1, and data from January 1 to June 30, 2021 is used as the test set. The data from January 1, 2020 to December 31, 2021 in Switzerland is divided into training and validation sets in the ratio of 9:1, and the data from January 1 to May 30, 2022 is used as the test set. In addition, to better represent the relevant time characteristics, we set the holiday type as D = [0,1], where 0 and 1 indicate holidays and non-holidays, respectively. We also set the day type as T = [1,2,3,4,5,6,7], where the numbers correspond to the weekdays, and the 24 h in a day are H = [1,2,…,24].

4.1.3. Evaluation criteria

To effectively compare and analyze our prediction models, root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE), and determination coefficient (R^2) are commonly used as evaluation indices for each prediction model. The specific formulas for the four indices are given as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - Y_i'\right|^2} \tag{23}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|Y_i - Y_i'\right|}{|Y_i|} \times 100\% \tag{24}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - Y_i'\right| \tag{25}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - Y_i')^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}^2)^2} \tag{26}$$

In the formulas (23–26), n represents the predicted times; Y_i and Y_i' are the actual value and predicted value of power load at time slot i, respectively; and $\overline{Y}$ indicates the average value calculated from actual value Y_i.

4.1.4. Model hyperparameters setting

We consider our prediction model and other comparison models in Python using TensorFlow and Keras. For each model, the hyperparameters choice of model affects prediction results. Accordingly, we use the BO algorithm to determine optimal hyperparameters. Table 1 summarizes the hyperparameters setting of developed model in this paper. In addition, in this paper, the lag number of Bi-LSTM is set to 24.

4.2. Results and comparative analysis

To highlight the advantages of our model, we select several advanced power load forecasting models and compare prediction results generated by our models and others. For ease of reference, we denote (1) our new model by “BO-BCA-Seq2Seq”; (2) the Bi-LSTM-based Seq2Seq model with CNN and attention mechanism by “BCA-Seq2Seq”; (3) the LSTM-based Seq2Seq model with CNN and attention mechanism by “CA-Seq2Seq”; (4) the Bi-LSTM-based Seq2Seq model by “B-Seq2Seq”.

Table 1
Hyperparameters of proposed model.

Hyperparameter	Value
Conv1D	filters=64 kernel_size=1
MaxPooling1D	pool_size=1
Conv1D	filters=128 kernel_size=1
MaxPooling1D	pool_size=1
Bi-LSTM-based encoder	units=128
Bi-LSTM-based decoder	units=256
Batch size	128
Epoch	50

Table 2
Error comparison of different models (before feature extraction).

Model	MAPE	MAE	RMSE	R^2
BO-BCA-Seq2Seq	**3.6974**	**423.2976**	**642.8244**	**0.6993**
BCA-Seq2Seq	3.8520	449.1344	664.8877	0.6783
CA-Seq2Seq	4.0414	461.9368	689.3017	0.6543
B-Seq2Seq	4.0213	471.0035	692.8483	0.6507
RNN	4.3191	536.8521	722.3408	0.6203
XGBoost	4.6113	512.0717	734.8093	0.6071

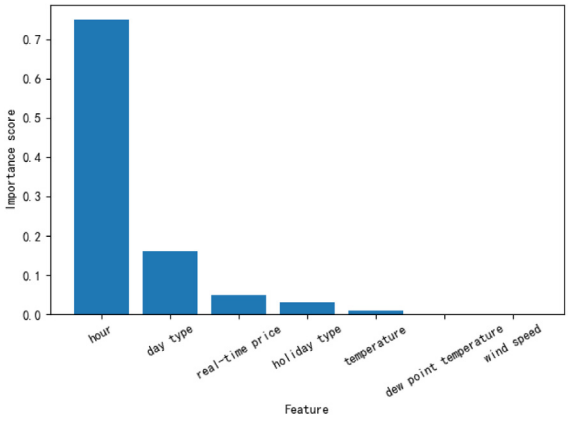

Fig. 8. Feature importance ranking table.

4.2.1. Influence of feature selection on prediction results

When performing feature selection analysis, we first select temperature, real-time price, holiday type, hour, dew point temperature,wind speed, and day type in Singapore as candidate features. Then, we use the random forest algorithm (extracted from large to small permutations) to sort these candidates according to their importance. When we order feature importance with random forests, the number of decision trees and the selection of the maximum number of features affect the accuracy of random forest in ranking the importance of features. Therefore, we determine the selection of related parameters by comparing MAPE and total accuracy through several experiments. Finally, when the number of final decision tree is set to 100 and the maximum feature number is set to 4, the MAPE is 1.72% and the total accuracy is 98.27%. The ranking table of feature importance is obtained as shown in Fig. 8.

Fig. 8 indicates that the characteristic importance score of dew point temperature and wind speed is 0, which means that these two features can be removed. To show that our novel method can better predict the power load through feature extraction, we performed load forecasting for Singapore from January 18 to January 24, 2021. We begin by making the load prediction with the unfeatured, extracted data, and obtain the model error comparison as shown in Table 2.

Table 3
Error comparison of different models (after feature extraction).

Model	MAPE	MAE	RMSE	R^2
BO-BCA-Seq2Seq	**3.6627**	**417.9481**	**639.5653**	**0.7024**
BCA-Seq2Seq	3.6843	426.8203	654.8040	0.6880
CA-Seq2Seq	3.8642	449.8587	659.8390	0.6832
B-Seq2Seq	3.9450	463.7143	652.9624	0.6898
RNN	3.9931	469.6667	657.9415	0.6850
XGBoost	4.1565	484.4464	705.4769	0.6379

Table 4
Comparison of holiday load forecast errors.

Model	MAPE	MAE	RMSE	R^2
BO-BCA-Seq2Seq	**1.8978**	**219.1910**	**254.5412**	**0.8613**
BCA-Seq2Seq	2.1130	245.7120	281.4285	0.8305
CA-Seq2Seq	2.1338	248.2894	298.1857	0.8097
B-Seq2Seq	2.2635	259.2483	327.0669	0.7710
RNN	2.3764	272.6471	330.7234	0.7659
XGBoost	2.5200	293.8400	351.2400	0.6803

Table 5
Comparison of holiday load forecast errors.

Model	MAPE	MAE	RMSE	R^2
BO-BCA-Seq2Seq	**1.7550**	**211.2582**	**241.2118**	**0.9116**
BCA-Seq2Seq	1.7951	219.4800	245.4491	0.9085
CA-Seq2Seq	1.9266	229.6772	269.3944	0.8897
B-Seq2Seq	2.0959	251.8861	287.0132	0.8748
RNN	2.3629	282.3021	328.6590	0.8359
XGBoost	2.5416	305.7432	332.8487	0.8317

We then use the data after feature extraction to make the model error comparison (after the feature extraction), as shown in Table 3. The comparative analysis of error results in Tables 2 and 3 show that after feature extraction, all error indicators of the models mentioned in this paper are reduced, and the accuracy of load prediction is improved.

4.2.2. Influences of holiday data on forecasting results

To improve the accuracy of holiday load forecasting, this paper uses the WGRP algorithm to process the holiday data before forecasting. In Singapore, April 2, 2021 is its official holiday, and this paper takes the data of April 2 as the data to be predicted, and the data from January 1 to April 1, 2021 as the sample data, while putting the data of the same moment of the sample data together as a sample set. In this paper, the data on April 2 is to be projected, and the data from January 1 to April 1, 2021 is the sample data, while the data at the same moment of the sample data is a sample set. Then, the weighted gray correlation projection values of all samples in each sample set are obtained according to the relevant formula of the WGRP algorithm, and then the projection values of each sample set are ranked, and then the data corresponding to the large projection values of each sample set are selected and composed for April 2 in Singapore. Figs. 9 and 10 are plotted to show prediction results before and after the holiday data is processed, respectively. Tables 4 and 5 are the error comparison diagrams before and after the holiday data processing, respectively. Through the comparative analysis of the results in these two tables, we learn that processing holiday data with the WGRP algorithm can effectively solve the problem of low accuracy of holiday load forecasting, and is suitable to all models considered in this paper.

4.2.3. Comparative analyses of prediction results

We now compare and analyze six prediction models (i.e., BO-BCA-Seq2Seq, BCA-Seq2Seq, CA-Seq2Seq, B-Seq2Seq, RNN and XGBoost). In the preceding section, we predict the monthly power load and holiday load, and also compute the errors for various

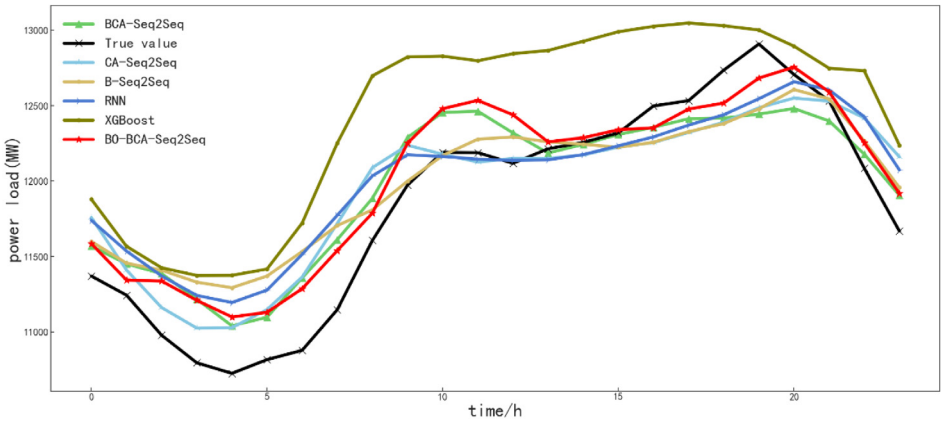

Fig. 9. Load forecasting results on holidays before the holiday data is processed.

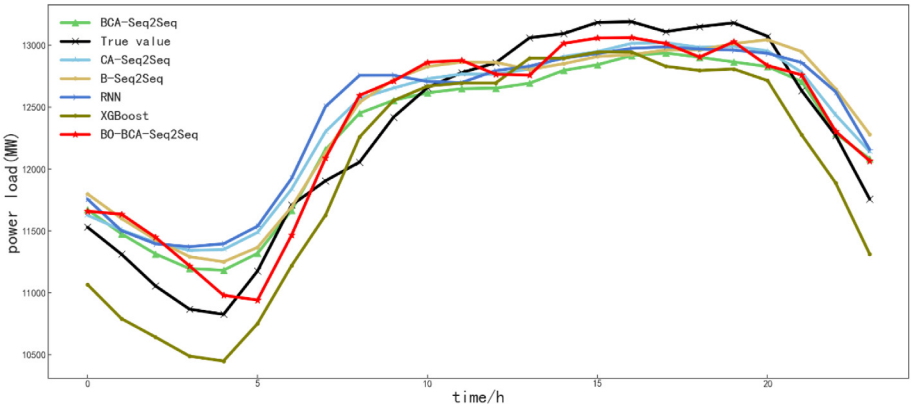

Fig. 10. Load forecasting results on holidays after the holiday data is processed.

models. The prediction results reveal that all error indicators for the BO-BCA-Seq2Seq model are the lowest. To further show that our model outperforms others in the power load prediction, we also conduct weekly load forecasting and non-holiday load forecasting, using electricity data from Singapore. According to the load prediction for Singapore from April 11 to April 17 in 2021 as shown in Fig. 11 and the non-holiday load forecast for Singapore in 2021 as shown in Fig. 13, we can find that the predicted values of all models are close to the true values in terms of trend. However, one can note that the predicted values of our model are closer to the true values than those of the other models, which indicates the superiority of our model in this paper. Nonetheless, it may not be accurate to conclude which model is optimal only in accordance with the observation results. Thus, we further analyze the evaluation by comparing the MAPE, MAE, RMSE, and R^2 of each model. We find from Table 6 and Fig. 14 that the use of the BO algorithm is effective in increasing the accuracy of load prediction and improving the predictive performance. By comparing the BCA-Seq2Seq model and the B-Seq2Seq model, we expose that introducing CNN and attention mechanism in the Seq2Seq model can improve the accuracy of prediction. In addition, the prediction accuracy of Seq2Seq model is higher compared with the traditional single prediction model.

To avoid the problem of one-sidedness caused by using a single dataset, we also perform weekly load and non-holiday load forecasting for Switzerland. See Fig. 12 which indicates the load prediction for Switzerland from February 6 to February 13 in 2022, and Fig. 15 which shows the non-holiday load forecast for Switzerland in 2022. Although the volatility of Switzerland electricity data is higher than that of Singapore, the predicted values of our model are still closer to the true values than those of other models. The weekly load forecasting errors in Table 7 and the non-holiday load forecasting errors shown in Fig. 16 reveal the conclusions similar to those previously made for load forecasting in Singapore. That is, the proposed model in this paper has the highest prediction accuracy and the best prediction performance.

Table 6
Comparison of weekly load forecasting errors (Singapore).

Model	MAPE	MAE	RMSE	R^2
BO-BCA-Seq2Seq	**2.9587**	**373.0500**	**494.5731**	**0.8145**
BCA-Seq2Seq	3.0110	381.5123	510.6277	0.8022
CA-Seq2Seq	3.1442	400.9776	547.2400	0.7729
B-Seq2Seq	3.1808	404.5096	593.5294	0.7328
RNN	3.2067	405.4656	563.0845	0.7595
XGBoost	3.2604	411.8957	569.3239	0.7542

Table 7
Comparison of weekly load forecasting errors (Switzerland).

Model	MAPE	MAE	RMSE	R^2
BO-BCA-Seq2Seq	**4.8232**	**394.0333**	**537.5872**	**0.3806**
BCA-Seq2Seq	4.9055	398.7077	551.7313	0.3476
CA-Seq2Seq	5.0380	409.7991	555.5442	0.3385
B-Seq2Seq	5.2604	434.2994	558.4443	0.3316
RNN	5.3421	435.6984	573.6323	0.2947
XGBoost	5.4634	442.0810	581.2108	0.2760

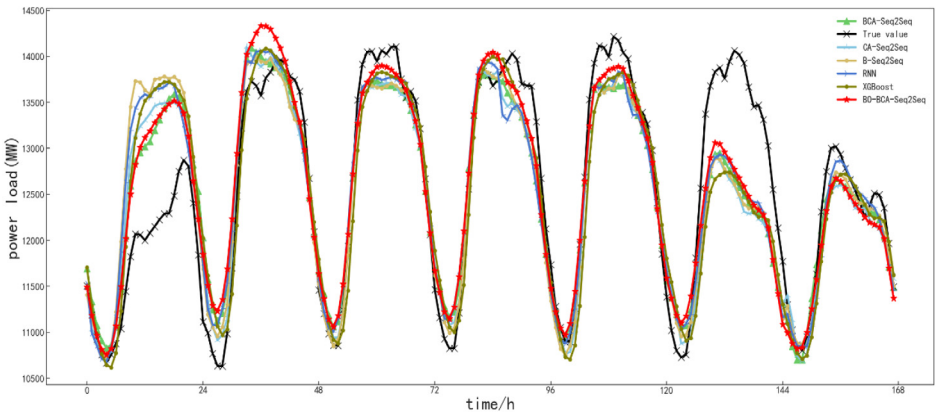

Fig. 11. Weekly load forecasting results (Singapore).

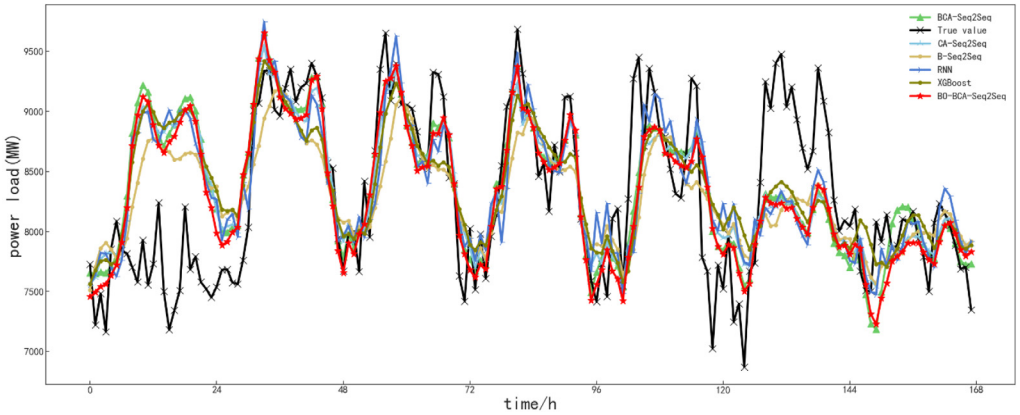

Fig. 12. Weekly load forecasting results (Switzerland).

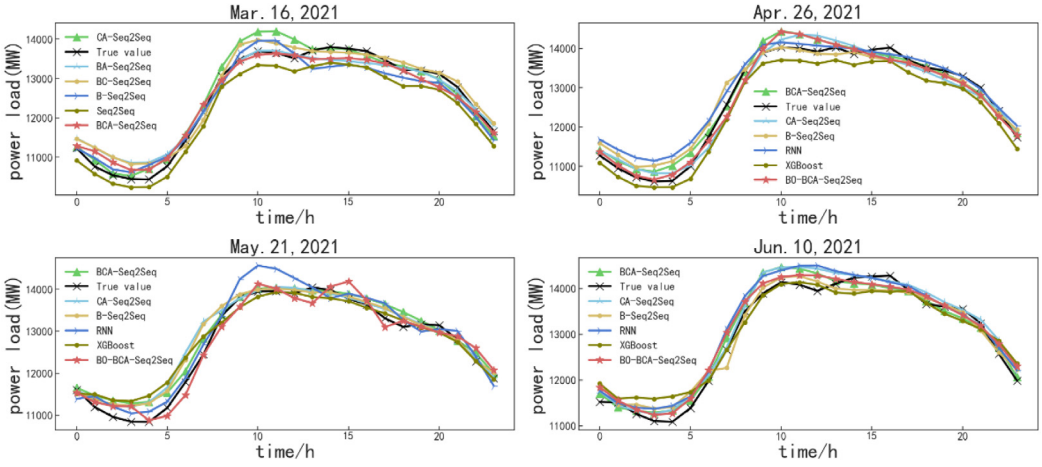

Fig. 13. Non-holidays load forecasting results (Singapore).

5. Conclusions

In this paper, we develop a novel hybrid power load forecasting method. Our major conclusions are summarized as follows:

1. We find that the random forest algorithm for eliminating redundant features and the WGRP algorithm for reselecting holiday load data can improve forecasting accuracy.
2. When the Bi-LSTM applies to the encoder and decoder of the Seq2Seq model, the ability of the encoder to extract information can be improved and the decoder can better use past and future information for power load prediction. The prediction performance of Seq2Seq model is effectively improved.
3. By introducing CNN, attention mechanism and BO algorithm, the Seq2Seq model load prediction performance and prediction accuracy can be significantly improved. Among them, CNN can effectively extract the spatial features of the input data, attention mechanism can better focus on the sequence information that has an impact on the prediction results, and BO algorithm can obtain the best hyperparameters for the model when making predictions.

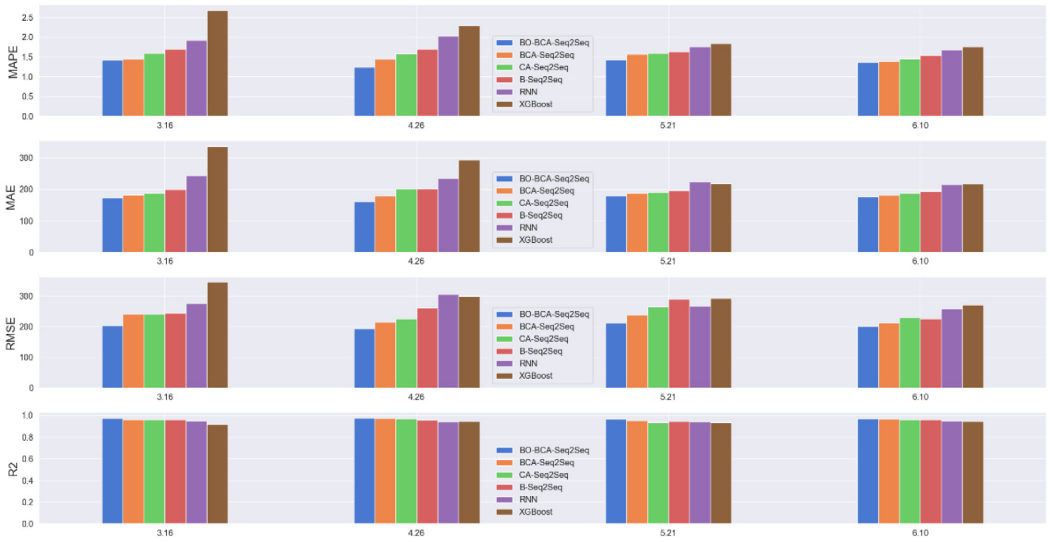

Fig. 14. Comparison of non-holidays load forecasting errors (Singapore).

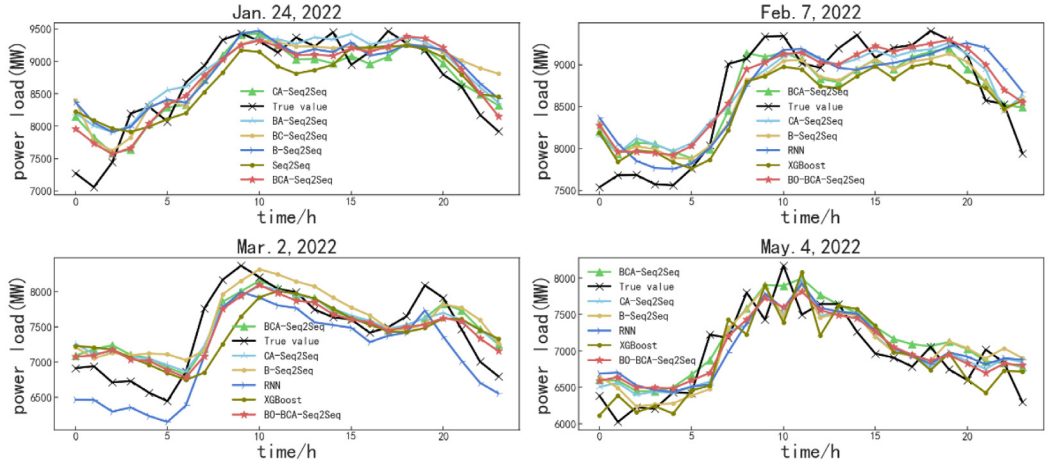

Fig. 15. Non-holidays load forecasting results (Switzerland).

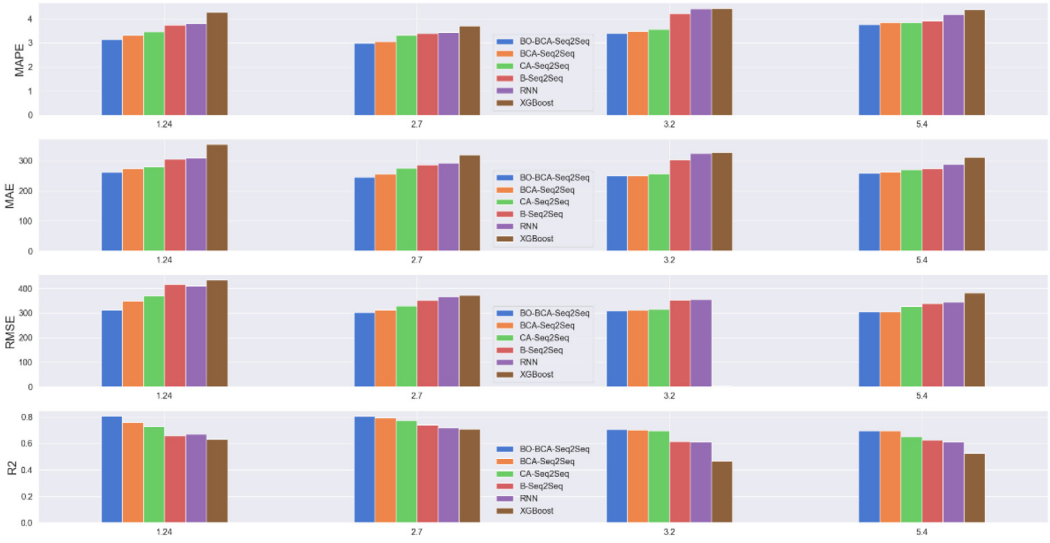

Fig. 16. Comparison of non-holidays load forecasting errors (Switzerland).

4. By comparing the prediction errors for all prediction models in two electricity markets with four evaluation metrics, we find that our prediction model possesses the lowest mean square error, average absolute percentage error, and average absolute error, which effectively shows that our model can present better power load forecasting results.

Although the prediction performance is relatively high, our prediction model could be further extended for future research. Some more advanced intelligent optimization algorithms may be added to our model for a further improvement of prediction efficiency. In addition, to show prediction effectiveness, we may apply our model to other energy prediction fields.

CRediT authorship contribution statement

Yeming Dai: Conceptualization, Methodology, Writing – review & editing. **Xinyu Yang:** Data curation, Comparative analysis, Writing – original draft. **Mingming Leng:** Writing – review & editing.

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 data that has been used is confidential.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 72171126), Ministry of Education Project of Humanities and Social Science (No. 20YJA630009). Natural Science Foundation of Shandong Province (No. ZR2022MG002). This work is also financially supported by the Faculty Researcrant (FRG) of Lingnan University (No. DB21B1).

References

[1] S.K. Rathor, D. Saxena, Energy management system for smart grid: An overview and key issues, Int. J. Energy Res. 44 (6) (2020) 4067–4109.
[2] Z. Shi, W. Yao, Z. Li, et al., Artificial intelligence techniques for stability analysis and control in smart grids: Methodologies, applications, challenges and future directions, Appl. Energy 278 (2020) 115733.
[3] Y. Dai, X. Yang, M. Leng, Forecasting power load: A hybrid forecasting method with intelligent data processing and optimized artificial intelligence, Technol. Forecast. Soc. Change 182 (2022) 121858.
[4] N. Abu-Shikhah, F. Elkarmi, Medium-term electric load forecasting using singular value decomposition, Energy 36 (7) (2011) 4259–4271.
[5] Z. Wen, L. Xie, Q. Fan, et al., Long term electric load forecasting based on TS-type recurrent fuzzy neural network model, Electr. Power Syst. Res. 179 (2020) 106106.
[6] O.M. Butt, M. Zulqarnain, T.M. Butt, Recent advancement in smart grid technology: Future prospects in the electrical power network, Ain Shams Eng. J. 12 (1) (2021) 687–695.
[7] V. Cerqueira, L. Torgo, I. Mozetič, Evaluating time series forecasting models: An empirical study on performance estimation methods, Mach. Learn. 109 (11) (2020) 1997–2028.
[8] H. Takeda, Y. Tamura, S. Sato, Using the ensemble Kalman filter for electricity load forecasting and analysis, Energy 104 (2016) 184–198.
[9] S. Sreekumar, K.C. Sharma, R. Bhakar, Grey system theory based net load forecasting for high renewable penetrated power systems, Technol. Econ. Smart Grids Sustain. Energy 5 (1) (2020) 1–14.
[10] I.K. Nti, M. Teimeh, O. Nyarko-Boateng, et al., Electricity load forecasting: A systematic review, J. Electr. Syst. Inform. Technol. 7 (1) (2020) 1–19.
[11] M.S. Ibrahim, W. Dong, Q. Yang, Machine learning driven smart electric power systems: Current trends and new perspectives, Appl. Energy 272 (2020) 115237.
[12] B. Zhu, S. Ye, P. Wang, et al., Forecasting carbon price using a multi-objective least squares support vector machine with mixture kernels, J. Forecast. 41 (1) (2022) 100–117.
[13] G. Hafeez, K.S. Alimgeer, I. Khan, Electric load forecasting based on deep learning and optimized by heuristic algorithm in smart grid, Appl. Energy 269 (2020) 114915.
[14] J. Zhang, Z. Tan, Y. Wei, An adaptive hybrid model for short term electricity price forecasting, Appl. Energy 258 (2020) 114087.
[15] L. Zhang, J. Wen, Y. Li, et al., A review of machine learning in building load prediction, Appl. Energy 285 (2021) 116452.
[16] S. Atef, A.B. Eltawil, Assessment of stacked unidirectional and bidirectional long short-term memory networks for electricity load forecasting, Electr. Power Syst. Res. 187 (2020) 106489.
[17] P. Kumari, D. Toshniwal, Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting, Appl. Energy 295 (2021) 117061.
[18] S. Wang, X. Wang, S. Wang, et al., Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting, Int. J. Electr. Power Energy Syst. 109 (2019) 470–479.
[19] K. Wu, J. Wu, L. Feng, et al., An attention-based CNN-LSTM-BiLSTM model for short-term electric load forecasting in integrated energy system, Int. Trans. Electr. Energy Syst. 31 (1) (2021) e12637.
[20] J. Song, L. Zhang, G. Xue, et al., Predicting hourly heating load in a district heating system based on a hybrid CNN-LSTM model, Energy Build. 243 (2021) 110998.
[21] Y. Zhang, Y. Li, G. Zhang, Short-term wind power forecasting approach based on Seq2Seq model using NWP data, Energy 213 (2020) 118371.
[22] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, Adv. Neural Inf. Process. Syst. (2014) 27.
[23] Y. Li, Z. Tong, S. Tong, et al., A data-driven interval forecasting model for building energy prediction using attention-based LSTM and fuzzy information granulation, Sustainable Cities Soc. 76 (2022) 103481.
[24] P. Jia, N. Cao, S. Yang, Real-time hourly ozone prediction system for Yangtze River Delta area using attention based on a sequence to sequence model, Atmos. Environ. 244 (2021) 117917.
[25] N. Mughees, S.A. Mohsin, A. Mughees, et al., Deep sequence to sequence Bi-LSTM neural networks for day-ahead peak load forecasting, Expert Syst. Appl. 175 (2021) 114844.
[26] S. Du, T. Li, Y. Yang, et al., Multivariate time series forecasting via attention-based encoder–decoder framework, Neurocomputing 388 (2020) 269–279.
[27] G. Manikandan, S. Abirami, An efficient feature selection framework based on information theory for high dimensional data, Appl. Soft Comput. 111 (2021) 107729.
[28] M. Radovic, M. Ghalwash, N. Filipovic, et al., Minimum redundancy maximum relevance feature selection approach for temporal gene expression data, BMC Bioinformatics 18 (2017) 9.
[29] Y. Huang, L. Shen, Liu H. Grey relational analysis, Principal component analysis and forecasting of carbon emissions based on long short-term memory in China, J. Clean. Prod. 209 (2019) 415–423.
[30] J. Li, Z. Meng, N. Yin, et al., Multi-source feature extraction of rolling bearing compression measurement signal based on independent component analysis, Measurement 172 (2021) 108908.
[31] R. Zebari, A. Abdulazeez, D. Zeebaree, et al., A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction, J. Appl. Sci. Technol. Trends 1 (2) (2020) 56–70.
[32] X.K. Li, W. Chen, Q. Zhang, et al., Building auto-encoder intrusion detection system based on random forest feature selection, Comput. Secur. 95 (2020) 101851.
[33] L. Ruxue, L. Shumin, Y. Miaona, et al., Load forecasting based on weighted grey relational degree and improved ABC-SVM, J. Electr. Eng. Technol. 16 (2021) 2191–2200.
[34] W. Yue, Q. Liu, Y. Ruan, et al., A prediction approach with mode decomposition-recombination technique for short-term load forecasting, Sustainable Cities Soc. 85 (2022) 104034.
[35] Y. Xie, C. Li, G. Tang, et al., A novel deep interval prediction model with adaptive interval construction strategy and automatic hyperparameter tuning for wind speed forecasting, Energy 216 (2021) 119179.
[36] W. Qiao, H. Lu, G. Zhou, et al., A hybrid algorithm for carbon dioxide emissions forecasting based on improved lion swarm optimizer, J. Clean. Prod. 244 (2020) 118612.
[37] F. He, J. Zhou, Z. Feng, et al., A hybrid short-term load forecasting model based on variational mode decomposition and long short-term memory networks considering relevant factors with Bayesian optimization algorithm, Appl. Energy 237 (2019) 103–116.
[38] A.H. Victoria, G. Maragatham, Automatic tuning of hyperparameters using Bayesian optimization, Evol. Syst. 12 (1) (2021) 217–223.

[39] J. Liu, Y. Li, Study on environment-concerned short-term load forecasting model for wind power based on feature extraction and tree regression, J. Clean. Prod. 264 (2020) 121505.
[40] Y. Dai, P. Zhao, A hybrid load forecasting model based on support vector machine with intelligent methods for feature selection and parameter optimization, Appl. Energy 279 (2020) 115332.
[41] I.E. Livieris, E. Pintelas, P. Pintelas, A CNN–LSTM model for gold price time-series forecasting, Neural Comput. Appl. 32 (23) (2020) 17351–17360.
[42] N.M.M. Bendaoud, N. Farah, Using deep learning for short-term load forecasting, Neural Comput. Appl. 32 (18) (2020) 15029–15041.
[43] D. Cannizzaro, A. Aliberti, L. Bottaccioli, et al., Solar radiation forecasting based on convolutional neural network and ensemble learning, Expert Syst. Appl. 181 (2021) 115167.
[44] M. Imani, Electrical load-temperature CNN for residential load forecasting, Energy 227 (2021) 120480.
[45] Y. Li, Z. Zhu, D. Kong, et al., EA-LSTM: Evolutionary attention-based LSTM for time series prediction, Knowl.-Based Syst. 181 (2019) 104785.
[46] Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, Neurocomputing 452 (2021) 48–62.
[47] G. Brauwers, F. Frasincar, A general survey on attention mechanisms in deep learning, IEEE Trans. Knowl. Data Eng. 35 (4) (2021) 3279–3298.