Applied Soft Computing 130 (2022) 109632

Contents lists available at ScienceDirect

Applied Soft Computing

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

Improving the Bi-LSTM model with XGBoost and attention mechanism: A combined approach for short-term power load prediction

Yeming Dai [a,*], Qiong Zhou [a], Mingming Leng [b], Xinyu Yang [a], Yanxin Wang [a]

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

ARTICLE INFO

Article history:
Received 31 March 2022
Received in revised form 15 September 2022
Accepted 15 September 2022
Available online 22 September 2022

Keywords:
Power load forecasting
Attention mechanism
Bidirectional long–short term memory network
Extreme gradient boosting
Weighted grey relational projection algorithm

ABSTRACT

Short term power load forecasting plays an important role in the management and development of power systems with a focus on the reduction in power wastes and economic losses. In this paper, we construct a novel, short-term power load forecasting method by improving the bidirectional long short-term memory (Bi-LSTM) model with Extreme Gradient Boosting (XGBoost) and Attention mechanism. Our model differs from existing methods in the following three aspects. First, we use the weighted grey relational projection algorithm to distinguish the holidays and non-holidays in the data preprocessing. Secondly, we add the Attention mechanism to the Bi-LSTM model to improve the validity and accuracy of prediction. Thirdly, XGBoost is a newly-developed, well-performing prediction model, which is used together with the Attention mechanism to optimize the Bi-LSTM model. Therefore, we develop a novel, combined power load prediction model "Attention-Bi-LSTM + XGBoost" with the weight determination theory-error reciprocal method. Using two power market datasets, we evaluate our prediction method by comparing it with two benchmark models and four other models. With our prediction method, the MAPE, MAE, and RMSE for the Singapore's power market are 0.387, 43.206, and 54.357, respectively; and those for the Norway's power market are 0.682, 96.278, and 125.343, respectively. The test results are smaller than the results for six other models. This indicates that our prediction method outperforms the LSTM, Bi-LSTM, Attention-RNN, Attention-LSTM, Attention-Bi-LSTM, and XGBoost in effectiveness, accuracy, and practicability.

1. Introduction

1.1. Background and motivation

With the continuous growth of power demand, traditional power grid faces the challenges in centralized distribution, manual monitoring and recovery, and two-way communication [1,2]. Smart grid acts as an effective solution to the above issues, since it is conducive to monitoring power production, transmission, and consumption as well as balancing the relationship [3]. However, the power load often fluctuates greatly due to the influence of uncertain factors such as climate, economy, and environment [4]. It is thereby difficult to estimate the future trend of power demand. Furthermore, the overestimation or underestimation of power load is detrimental to power grid strategic decision-makings. Therefore, accurate and precise power load prediction is of help to the power consumption management, reasonable arrangement of power grid operation mode, and improvement of economic and social benefits of power systems. Today, power load forecasting has become one of the important contents to realize the modernization of power system management and the transformation of power retail companies to the spot market, which ensures the safe operations of power systems as well as the balance between power supply and demand.

1.2. Literature reviews

There are a number of power load prediction methods in existing literatures, which can be classified as four categories: (i) Classical prediction methods, (ii) Modern prediction methods, (iii) Hybrid prediction methods, and (iv) Combined prediction methods. Among them, classical prediction methods include time series analysis [5,6], regression analysis [7], and other statistical methods, which all perform well in solving simple linear problems by using time series methods to estimate the future power

* Corresponding author.
E-mail addresses: yemingdai@163.com (Y. Dai), zq17806262874@163.com (Q. Zhou), mmleng@ln.edu.hk (M. Leng), 865050851@qq.com (X. Yang), 457825811@qq.com (Y. Wang).

https://doi.org/10.1016/j.asoc.2022.109632
1568-4946/

Nomenclature

Abbreviations

LSTM	Long short-term memory
Bi-LSTM	Bidirectional long short-term memory
XGBoost	Extreme Gradient Boosting
ANN	Artificial neural network
mFFO	Modified fire-fly optimization algorithm
PSO	Particle swarm optimization
BOA	Bayesian optimization
HSVR	Hybrid support vector regression
STA	State transition algorithm
SVM	Support vector machine
GWO	Grey wolf optimizer
EEMD	Ensemble empirical mode decomposition
VMD	Variational modal decomposition
MMI	Modified mutual information
RF	Random forest
WGRP	Weighted grey relational projection algorithm
MOGWO	Multi-objective grey wolf algorithm
RBF	Radial basis function network
GRNN	Generalized regression neural network
ELM	Extreme learning machine
RNN	Recurrent neural network
MAPE	Mean absolute percent error
MAE	Mean absolute error
RMSE	Root mean square error

Functions and variables

n_1	Selected sample data
$\boldsymbol{A}$	Gray relationship matrix
$\boldsymbol{A}'$	Weighted grey incidence matrix
$\boldsymbol{\gamma}$	The weight of each influencing factor
$\boldsymbol{A}'_0$	The row vector of the sample
$\boldsymbol{A}'_{\boldsymbol{i}}$	The row vector of other historical samples
$\cos\theta_i$	Cosine value of gray projection angle
B_i	Weighted grey correlation projection value
$\boldsymbol{X_t}$	The current time step input
H_{t-1}	The previous time step
I_t	Input gate of LSTM
F_T	Forgetting gate of LSTM
O_t	Output gate of LSTM
h	The number of hidden units
H_t	The hidden state of the current time step
σ	Sigmoid function
$\boldsymbol{W_{xi}}$, $\boldsymbol{W_{xo}}$, $\boldsymbol{W_{xf}}$	The weight matrix
b_i, b_o, b_f	The offset term
$\widetilde{C}_t$	Candidate memory cells
C_t	The cell state of the current time step
C_{t-1}	The cell state at the previous time
$\hat{y}_i$	The predicted value
w_j	The weight
x_{ij}	Sample data
$\hat{y}_i^{(t)}$	The model after training t round
$\hat{y}_i^{(t-1)}$	The reserved function added in the previous round
$Obj^{(t)}$	The objective function of XGBoost
$\Omega f\ (t)$	The regular term in the objective function
T	The number of leaf nodes
γ	Control the number of leaf nodes
$\tau^{(t)}$, $\widetilde{\tau}^{(t)}$	Objective function simplified by second-order Taylor expansion
ω_j^*	Optimal solution of objective function
x	The power load data value before normalization
x^n	The power load data value after normalization
f_t	The formula of error reciprocal method
ω_1	The weight value of Attention-Bi-LSTM
ω_2	The weight value of and XGBoost
ε_1	The error values of Attention-Bi-LSTM
ε_2	The error values of XGBoost
y_t	The real power load data
$\hat{y}_t$	The predicted power load data

load. However, these methods are challenged in dealing with nonlinear problems. In order to better predict those nonlinear problems, the nonlinear mapping-based prediction technologies have been proposed. The input data is embedded into high-dimensional space, which can transform the nonlinear problems into linear problems. Modern prediction methods mainly include fuzzy logic, gray system [8], and machine learning algorithms [9], etc. Especially, the main two categories of machine learning algorithms are non-supervised and supervised learning [10]. Non-supervised learning does not have any training data samples. It is thus necessary to model the data directly with clustering and dimensionality reduction [11]. The supervised learning is trained by existing training samples to obtain an optimal model. Then, this model is used to map all new data samples to the corresponding output results. It is worth noting that above prediction methods have inherent limitations such as complex calculation [12–14], poor generalization ability [15], and over fitting [16–18], which all challenge power load predictions.

To overcome the weaknesses of above prediction methods, the hybrid prediction models have been developed by various optimization algorithms used to optimize the prediction performance, which include the modified fire-fly optimization (mFFO) algorithm [19], particle swarm optimization (PSO) [20], and Bayesian optimization (BOA) [21]. For example, Wang et al. [9] used the hybrid support vector regression (HSVR) method to predict the medium and long-term loads, and applied the hierarchical method based on nested strategy and state transition algorithm (STA) to optimize the parameters of prediction models. Barman and Choudhury [22] optimized the parameters of support vector machine (SVM) using the grey wolf optimizer (GWO) and predicted the power demand that is significantly affected by social factors such as culture or religious rituals. Moreover, in general, the hybrid prediction methods consist of data preprocessing and forecasting parts. Preprocessing data through different technologies can help eliminate outliers, correct data errors, and improve data quality. The relevant technologies include (i) data decomposition technologies such as ensemble empirical mode decomposition (EEMD) [5,23] and variational modal

decomposition (VMD) [3], and (ii) feature selection technologies such as modified mutual information (MMI) [24], random forest (RF) [25], and weighted grey relational projection algorithm (WGRP) [26]. In summary, the hybrid prediction methods can significantly reduce prediction error and improve prediction accuracy by performing parameter optimization, data decomposition or feature selection. However, the inherent disadvantages of the single prediction model still cannot be solved.

To further improve and optimize the prediction model and overcome the inherent defects of various single models in classical, modern, and hybrid prediction methods, the combined prediction methods began to be proposed by combining two or more different prediction models with a specific weighting method. Bates and Granger [27] put forward the idea of combined forecasting for the first time. They proposed a seminal combined prediction model, which combines two independent airline datasets with a weighting system. The results show that the combined prediction set can produce a lower error than the original prediction. Nie et al. [28], Deng et al. [29], and Chu et al. [30] also proved that the performance of combined prediction models is better than that of single models. Chen et al. [31] used the LSTM and XGBoost models to predict the power load, respectively, they then assign weights to the two models according to the error reciprocal method. For a better weighting method, the error should be reduced because a smaller error implies a higher prediction accuracy. Zhuang et al. [32] set an initial weight of model combination to search for the best weight combined with the MAPE-RW algorithm, and then constructed the CNN-LSTM-XGBoost combined prediction model, which significantly reduced the error index compared with the single prediction models. Nie et al. [33] used the multi-objective grey wolf algorithm (MOGWO) to determine the weights to the radial basis function network (RBF), generalized regression neural network (GRNN), and extreme learning machine (ELM). They established a combined prediction model based on the swarm intelligence optimization, which can effectively reduce the adverse effects of weak adaptability of single models and better grasp the characteristics of power load, thus significantly improve the prediction accuracy and adaptability.

We can learn from the above literatures review that the recent improvement of prediction models with the combination methods not only needs more than one single prediction model but also combines a variety of different algorithms to calculate the weights for each model [34]. However, the existing combined prediction models just combine some existing mature prediction models without emphasizing the importance of data preprocessing. In this context, we use the WGRP algorithm to preprocess the data and eliminate the impact of holidays. As for the prediction process, the Bi-LSTM model had been widely viewed as one with an excellent forecasting effect since it can fully consider the hidden information and obtain better prediction results. Moreover, since the Attention mechanism has the advantages of large-scale parallel processing, distributed information storage, and acceptable self-organization and self-learning ability, we add the Attention mechanism to the Bi-LSTM model [35] to eliminate the unreasonable impact and emphasize the impact of key input data. This makes the prediction results more comprehensive and is thus called "Attention-Bi-LSTM model" [36]. Although the Bi-LSTM model has been improved and optimized, its inherent defects still cannot be avoided. Hence, to further avoid the defects of a single prediction model, we should introduce the idea of combination and develop another prediction model. For neutralize the error as much as possible, we need a model with excellent prediction ability. After screening, we select the XGBoost [37] model, with specific attribution to the following performances [38]: (i) strong generalization ability; (ii) more flexible by using GART as the base classifier; (iii) controllable complexity by adding regular terms to prevent over-fitting; and (iv) fast computing speed because it only depends on the input data value and does not choose the specific form of loss function to perform leaf splitting. As a result, we develop the "Attention-Bi-LSTM + XGBoost" combination model to further improve the prediction accuracy of "Attention-Bi-LSTM" model. In order to verify the effectiveness of developed prediction method, we apply two power market cases in Singapore and Norway.

1.3. Our contributions

The novelty and major technical contributions are as follows:

(1) We develop a novel "Attention-Bi-LSTM + XGBoost" combined prediction model.
(2) We use the weighted grey relational projection algorithm to distinguish the holiday and non-holiday data.
(3) We consider the Attention mechanism in the Bi-LSTM model to improve the prediction accuracy.
(4) We use the XGBoost model to further improve the performance of Bi-LSTM model in a combined manner.
(5) We use two power market datasets and six power load prediction models to verify the effectiveness and reliability of our model.

The organization of this paper is as follows: The basic methods and algorithms used in this paper are introduced in Section 2. Section 3 introduces the weight method of combining the models. Section 4 presents the specific steps of our proposed prediction method. We consider two practical case to verify the prediction accuracy and stability of our method in Section 5. This paper ends with concluding remarks and possible future directions in Section 6.

2. Methodologies

2.1. Weighted grey relational projection algorithm

The WGRP algorithm [26] is a method for measuring the degree of similarity or difference between the development trends of various factors, i.e., "grey relational degree". This method is not limited by the sample size. For the data with small sample size and discreteness, it can avoid the one-way deviation caused by comparing the index values of single factors of each scheme, and also can comprehensively analyze the relationship between the indexes while the size of the module and the cosine of the included angle are combined. The proximity between each decision scheme and the ideal scheme is fully and accurately reflected. It is also applicable to the regular sample size with small amount of calculation. Therefore, this paper uses the WGRP algorithm to sort the factors that affect power load, and assigns weights according to the importance of these factors. Thus, the prediction results are general. The details regarding the calculation steps are as follows.

Firstly, select the data of the preceding n_1 samples and the data of the samples to be predicted, calculate the relationship coefficient between them, and construct the following grey relationship matrix.

$$\boldsymbol{A} = \begin{bmatrix} A_{01} \cdots A_{0m_1} \\ \vdots \quad \ddots \quad \vdots \\ A_{n_1 1} \cdots A_{n_1 m_1} \end{bmatrix} \tag{1}$$

where $A_{n_1 m_1}$ represents grey correlation coefficient whose m_1th factor in the n_1th sample.

Then, the weight of each influencing factor is calculated by entropy weight method, and the weighted grey relation matrix is

obtained by weighting the grey relation matrix, as shown below:

$$\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{\gamma}^{\boldsymbol{T}} = \begin{bmatrix} \gamma_1 \cdots \gamma_{m_1} \\ \vdots \ddots \vdots \\ \gamma_1 A_{n_1 1} \cdots \gamma_{m_1} A_{n_1 m_1} \end{bmatrix} \tag{2}$$

where, $\boldsymbol{\gamma}$ represents the weight of each influencing factor, the first row in the matrix is expressed as the row vector of the sample to be predicted of $\boldsymbol{A}_0'$, the row vector of other historical samples is expressed as $\boldsymbol{A}_i'$, and the included angle between each $\boldsymbol{A}_0'$ and $\boldsymbol{A}_i'$ is the gray projection angle of the sample, expressed as θ_i, and calculate the $\cos\theta_i$.

$$\cos\theta_i = \frac{\sum_{j=1}^{m_1} \gamma_j A_{ij} \gamma_j}{\sqrt{\sum_{j=1}^{m_1} (\gamma_j A_{ij})^2} \sqrt{\sum_{j=1}^{m_1} \gamma_j^2}} \tag{3}$$

Thus, the weighted grey correlation projection value B_i is

$$B_i = \frac{\sum_{j=1}^{m_1} \gamma_j A_{ij} \gamma_j}{\sqrt{\sum_{j=1}^{m_1} \gamma_j^2}} \tag{4}$$

In summary, with the appropriate weighting method, the WGRP algorithm can determine the key factors that have a great impact on the target variables, and combines the projection of each historical sample with the samples to be predicted to obtain the weighted grey correlation projection value B_i. Finally, we sort the obtained projection values from large to small, and select the samples with large projection values as similar samples for replacement, in order to find the similar sample set from the samples to the predictions. Then, the adverse impact of holiday samples with large deviation on the prediction results can be reduced and the historical load data is more general, which improve the prediction accuracy.

2.2. Bi-LSTM forecasting model

In the gradient algorithm of recurrent neural network, when the time steps are too small or too large, the gradient of recurrent neural network can easily explode and disappear [39]. Therefore, in order to solve this problem, LSTM uses gating mechanism to control information, as shown in Fig. 1 [40], and introduces input gate, forgetting gate and output gate to remove some contents that are not important to the current situation [34], thereby prolonging the storage time of information and save some older information. The input of LSTM gate is the hidden state between the current time step input $\boldsymbol{X_t}$ and the previous time step H_{t-1}. The output is calculated by the full connection layer.

$$\text{Input gate:} \quad I_t = \sigma(\boldsymbol{X_t}\boldsymbol{W}_{\text{xi}} + H_{t-1}\boldsymbol{W}_{\text{hi}} + b_{\text{i}}) \tag{5}$$

$$\text{Forgetting gate:} \quad F_T = \sigma(\boldsymbol{X_t}\boldsymbol{W}_{\text{xf}} + H_{t-1}\boldsymbol{W}_{\text{hf}} + b_{\text{f}}) \tag{6}$$

$$\text{Gated unit:} \quad \widetilde{C}_t = \tanh(\boldsymbol{X_t}\boldsymbol{W}_{\text{xc}} + H_{t-1}\boldsymbol{W}_{\text{hc}} + b_{\text{c}}) \tag{7}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \widetilde{C}_t \tag{8}$$

$$\text{Output gate:} \quad O_t = \sigma(\boldsymbol{X_t}\boldsymbol{W}_{\text{xo}} + H_{t-1}\boldsymbol{W}_{\text{ho}} + b_{\text{o}}) \tag{9}$$

where h is the number of hidden units, $\boldsymbol{X_t}$ is the small batch input of a given time step t, H_{t-1} is the hidden state of the previous time step, σ is sigmoid function, $\boldsymbol{W}_{\text{xi}}$ and $\boldsymbol{W}_{\text{hi}}$ are the weight matrix of the input gate, b_{i} is the offset term of the input gate; $\boldsymbol{W}_{\text{xf}}$ and $\boldsymbol{W}_{\text{hf}}$ are the weight matrix of the forgetting gate, b_{f} is the offset term of the forgetting gate; $\widetilde{C}_t$ is the candidate memory cells, $\boldsymbol{W}_{\text{xc}}$ and $\boldsymbol{W}_{\text{hc}}$ are weight matrices of gated unit, b_{c} is the offset term of the gated unit, C_t is the new cell state at the current time, C_{t-1} is the cell state at the previous time; $\boldsymbol{W}_{\text{xo}}$ and $\boldsymbol{W}_{\text{ho}}$ are the weight matrix of the output gate, and b_{o} is the offset term of the output gate. Using the $\tan h$ function with the value range in $[-1,1]$ as the activation function, the information flow in the hidden state is controlled by multiplying by elements $\odot$.

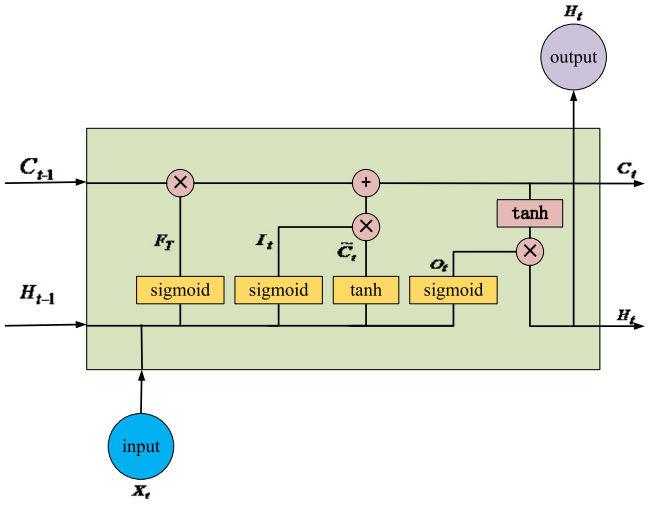

Fig. 1. Gating mechanism of LSTM.

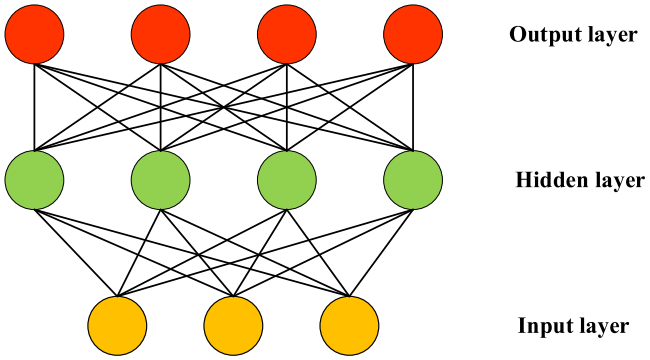

Fig. 2. Component connection topology of LSTM.

The output gate O_t controls the information flow from the memory cell to the hidden state, and the final output H_t is

$$H_t = O_t \odot \tan h(C_t) \tag{10}$$

and its component connection topology is shown as in Fig. 2.

Different from the LSTM, the Bi-LSTM (Bi-directional long short-term memory) method is composed of forward LSTM and backward LSTM. When extracting data features, we take into account the overall information hidden in the data, and extract features from both forward and reverse angles [32]. Then, the results of two-way extraction are combined in a specific way and summarized from two dimensions, which can eliminate the impact of the order of input data in a single LSTM on the final result to a certain extent and make the results more comprehensive.

2.3. Attention mechanism

The core idea of Attention mechanism is to simulate attention ability of people. For the information to be processed, people usually focus on a few key points instead of evenly distributing their attention to all information. Therefore, the introduction of Attention mechanism into the prediction model can assign different weights to the data, eliminate the unreasonable impact of input data on output data, and improve the impact of key input

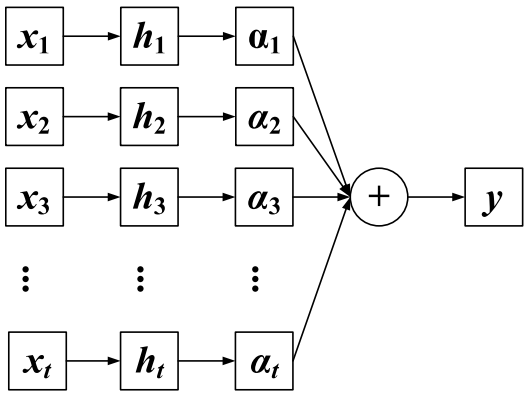

Fig. 3. Structure of attention mechanism.

data. The model structure of attention is shown in Fig. 3. For the specific calculation steps, see, for example, Zheng and Chen [36].

2.4. XGBoost power load forecasting model

Extreme gradient boosting is essentially a gradient boosting decision tree, which can improve the speed and efficiency of prediction. It is an optimization of the boosting algorithm that builds a decision tree by continuously adding trees and continuously splitting features [30]. When we add a tree, a new function $f(x)$ is learned to fit the residual predicted last time. When the training is completed and k trees are obtained, each tree falls to a corresponding leaf node, and each leaf node corresponds to a score. It is only necessary to add up the corresponding scores of each tree to get the predicted value of the sample. The XGBoost model is as follows:

$$\hat{y}_i = \sum_{j=1}^{n} w_j x_{ij} \tag{11}$$

where $\hat{y}_i$ is the predicted value, n is the number of trees, w_j is the weight, and x_{ij} is sample data.

In each iteration, a tree is added on the basis of the existing tree to fit the residual between the predicted results of the previous tree and the real value. The iterative process is as follows:

$$\begin{aligned}
&\hat{y}_i^{(0)} = 0\\
&\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)\\
&\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = y_i{}^{(1)} + f_2(x_i)\\
&\cdots\\
&\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned} \tag{12}$$

where $\hat{y}_i^{(t)}$ is the model after training t round; $\hat{y}_i^{(t-1)}$ is the reserved function added in the previous round; and $f_t(x_i)$ is the newly added function. The objective function of XGBoost is as follows:

$$\begin{aligned}
Obj^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i)\\
&= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)
\end{aligned} \tag{13}$$

$$\Omega f(t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{14}$$

The ultimate goal is to find f_t that minimizes the objective function, the $\sum_{i=1}^{t} \Omega(f_i)$ in formula (13) is a regular term in the objective function, which determines the complexity of the tree. Moreover, a smaller value results in a lower complexity and the stronger generalization ability. In formula (14), T is the number of leaf nodes, ω is the score of leaf node, γ is used to control the number of leaf nodes, and λ ensures that the score of leaf nodes is not too large.

In order to find a f_t to minimize the objective function, Taylor's second-order expansion is carried out at $f_t = 0$, and the objective function obtained is approximately as follows:

$$\tau^{(t)} \simeq \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{15}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ is the first derivative and $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ is the second derivative.

Since the prediction score of the first $t-1$ trees and the residual error of y will not affect the optimization of the objective function, it is directly removed and the objective function is further simplified as:

$$\tilde{\tau}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{16}$$

Formula (16) further simplifies the objective function by superimposing the loss function values of each sample. As each sample eventually falls into a leaf node, we reorganize all samples of the same leaf node on the basis of formula (16), in order to achieve the purpose of simplifying and rewriting the objective function. The process is as follows:

$$\begin{aligned}
Obj^{(t)} &\simeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)\\
&= \sum_{i=1}^{n} \left[g_i \omega_{q_{(x_i)}} + \frac{1}{2} h_i \omega^2_{q_{(x_i)}} \right] + \gamma T + \lambda \frac{1}{2} \sum_{j=1}^{T} \omega_j^2\\
&= \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i)\omega_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda)\omega_j^2 \right] + \gamma T
\end{aligned} \tag{17}$$

Therefore, by rewriting the above formulas, we can rewrite the objective function into a unary quadratic function about the leaf node fraction ω. It becomes simple to obtain the optimal ω and the value of the objective function by using vertex formula directly. For the rewritten univariate quadratic function about the leaf node fraction ω, the optimal ω_j^* and objective function can be obtained as follows:

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{18}$$

$$Obj = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{19}$$

In order to facilitate calculation and meet the requirements of data input, the data shall be normalized in advance. The power load data is normalized according to the following formula. The data is limited to the range of [0,1].

$$x^n = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{20}$$

where x and x^n are the power load data value before and after normalization, respectively. $x_{\min}$ and $x_{\max}$ are the minimum and maximum of the power load data value before normalization, respectively.

3. Attention-Bi-LSTM + XGBoost power load combined prediction model

3.1. Weighting method

The research result of Chen et al. [31] shows that the error reciprocal method is not only easy to operate, but also can significantly optimize the prediction performance of the model. Therefore, in this paper, the reciprocal error method is used to assign weights to the model. According to this method, the weights of Attention-Bi-LSTM and XGBoost are calculated through the error results obtained from the main evaluation index MAPE. Thus, the prediction model with a smaller error in this combined model is given a larger weight. Hence, the overall error of the combined prediction model can be reduced significantly. To confirm the weight coefficient, the formula of error reciprocal method is as follows:

$$f_t = \omega_1 f_{1t} + \omega_2 f_{2t}, t = 1, 2, \ldots, n \tag{21}$$

$$\omega_1 = \frac{\varepsilon_2}{\varepsilon_1 + \varepsilon_2} \tag{22}$$

$$\omega_2 = \frac{\varepsilon_1}{\varepsilon_1 + \varepsilon_2} \tag{23}$$

where ω_1 and ω_2 mean the weight value of Attention-Bi-LSTM and XGBoost respectively; f_{1t} and f_{2t} mean the predicted value obtained by Attention-Bi-LSTM and XGBoost. The weight value is obtained from formulas (22) and (23), where ε_1 and ε_2 are the error values of the prediction models Attention-Bi-LSTM and XGBoost respectively.

3.2. Attention-Bi-LSTM + XGBoost combined prediction model

Different from the existing prediction models, the Attention-Bi-LSTM can not only fully consider the overall information hidden in the input data from two dimensions to obtain more comprehensive results, but also emphasize the impact of key input data. Thus, the use of Attention-Bi-LSTM model can improve the prediction accuracy of results. Moreover, XGBoost, as a newly proposed prediction model with a low complexity, can prevent over fitting and has an excellent prediction performance. We then use Attention-Bi-LSTM and XGBoost methods to forecast the power load, and obtain the corresponding errors. The weights for the above two models are calculated by using the error reciprocal method according to the error results, which gives a greater weight to the model with a smaller error, so as to maximize the advantages of the model and reduce the error as much as possible. Finally, we combine the different prediction results of above two models by using the error reciprocal method, which can overcome various inherent defects of a single prediction model. The framework is shown in Fig. 4.

4. Prediction method

Based on our newly proposed combination forecasting model, the power load prediction process in our method is shown in Fig. 5. The prediction steps of the full text are mainly divided into four stages:

Stage 1: Data preprocessing. First, we select several influencing features with the greatest correlation, such as time, day type, holiday type, real-time price. Then, we use the WGRP algorithm to process the data of holidays to distinguish holiday and non-holiday data, making the data more general. Finally, we normalize the data.

Stage 2: Prediction using single models. The Attention mechanism is used to optimize the LSTM model, so that the influence of unreasonable factors can be eliminated and then the influence of key input data can be emphasized to make the results more comprehensive. Single Attention-Bi-LSTM and XGBoost model are used to predict the same dataset and prepare for the combination of the two models according to the prediction results in Stage 3.

Stage 3: Weight the models. After the power load data are forecasted with Attention-Bi-LSTM and XGBoost methods, we use the error reciprocal method to obtain the weights according to the error predicted, which means that single models are weighted. Then, the Attention-Bi-LSTM + XGBoost combined prediction model forms.

Stage 4: Evaluation of prediction results. By comparing the prediction errors for two benchmark models and four other models with two power markets, we show whether this method can improve the accuracy of power load forecasting or not.

5. Evaluation and analysis

In this section, we evaluate and discuss the performance of developed prediction method based on two cases. Simulation is carried out in Python to verify the effectiveness of the proposed method. Since Attention-Bi-LSTM and XGBoost are relevant to our model in this paper, we regard them as two benchmark models and also select LSTM, Bi-LSTM, Attention-LSTM, and Attention-RNN as other models for comparison.

5.1. Datasets and experimental environment

After screening the data of several national power markets, we find that the data of Singapore and Norway power markets embrace all the influencing factors we need. The relevant data is complete and comprehensive consideration is more suitable for the model proposed in this paper. Therefore, we use the data to evaluate the proposed method and consider the factors such as day type, time, holiday type, real-time price, etc. These data include Singapore's data from January 1, 2019 to December 31, 2020 and Norway's data from January 1, 2020 to December 31, 2021. The sampling period of historical power load data is 1 h; and, 80% of the prepared data is used for training and 20% for testing [41–43]. Finally, we compare the results with the real value and investigate the errors.

The hardware platform of this experiment is equipped with Intel i5-1035G1 processor, with 8 GB memory, 477 GB solid state disk capacity and MX230 CPU graphics card. The method proposed in this paper is implemented based on Python language. The Attention-Bi-LSTM model uses Keras deep learning framework and XGBoost uses py-xgboost framework.

5.2. Evaluation criteria

We consider the prediction accuracy as our objective to test the proposed combined model's efficiency. To evaluate this objective, we use three standard statistical indicator models. Specifically, we select mean absolute percent error (MAPE) as the main evaluation index of each prediction model, and choose mean absolute error (MAE) and root mean square error (RMSE) as the auxiliary evaluation indices. These statistical indicator models are used to measure the accuracy rate of the proposed model. The calculation formulas of MAPE, MAE and RMSE are as follows:

$$X_{\mathrm{MAPE}} = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|y_t - \hat{y}_t\right|}{y_t} \tag{24}$$

$$X_{\mathrm{MAE}} = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \hat{y}_t\right| \tag{25}$$

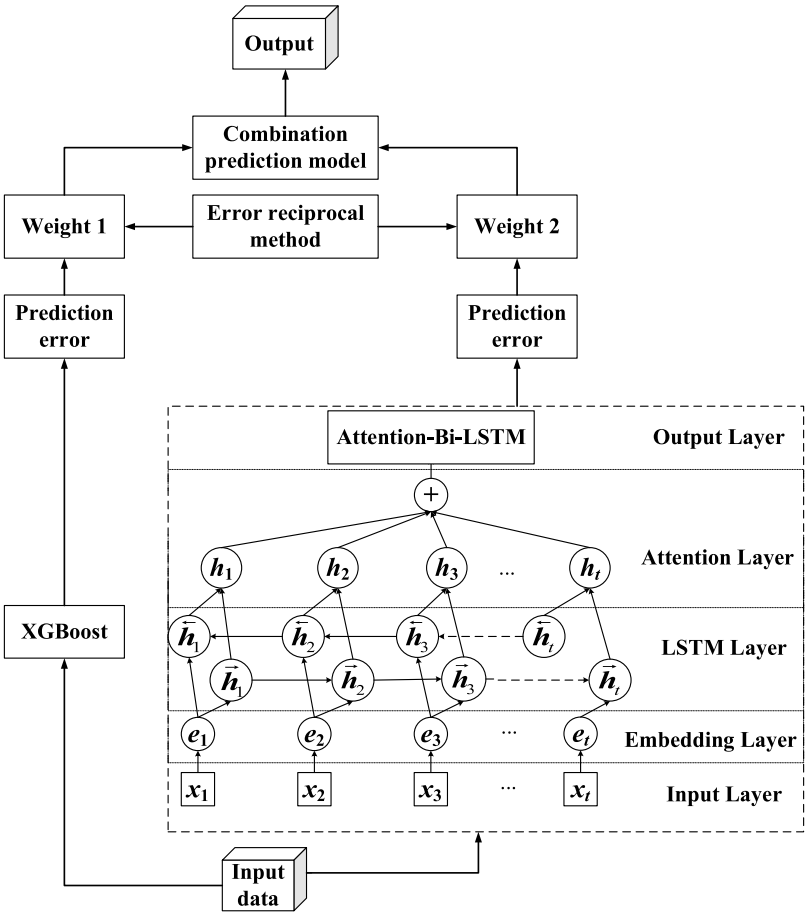

Fig. 4. The framework of Attention-Bi-LSTM + XGBoost combined prediction model.

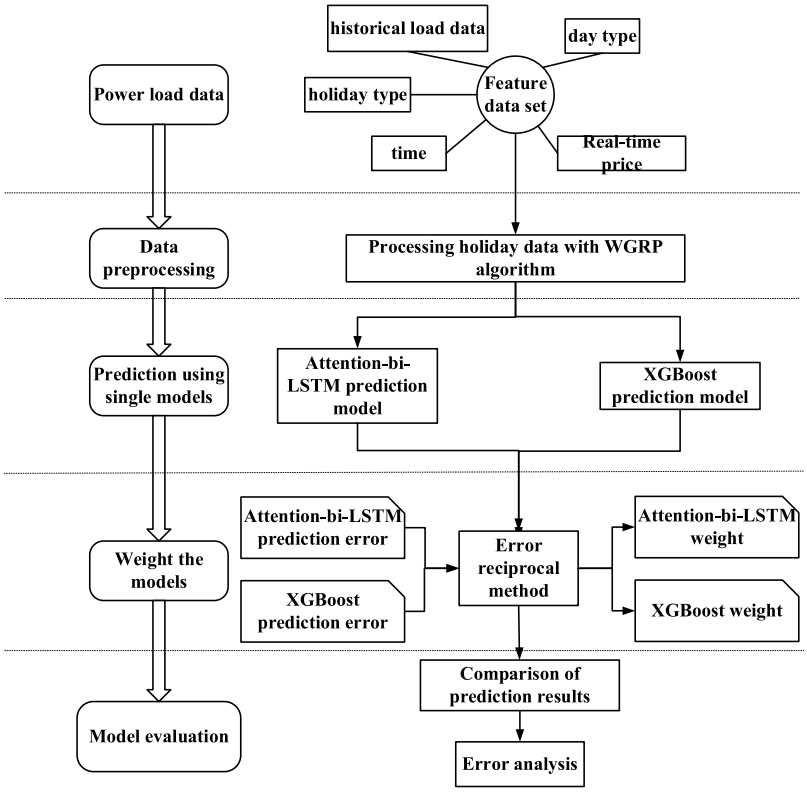

Fig. 5. Power load prediction process in our novel method.

Table 1
Parameter settings of Attention-Bi-LSTM.

Algorithm	Parameter meaning	Parameter	The parameter value	Setting basis
Attention-Bi-LSTM	The number of units	Unit	128	It is the key parameter that affects the accuracy of the model and has the optimal quantity.
	Time step	time_step	24	Determine whether each input data is related to the previous number of successively input data.
	The number of unit layers	num_layers	2	The default value is 1 layer. If it is set to 2, the second layer receives the calculation results of the first layer.
	Hidden layer width	batch_size	256	The amount of data entered at one time, through the setting of this parameter, it can distinguish whether the input data is the same batch of data.
	Iteration times	Epochs	40	It depends on the computing power of the computer to determine the optimal number of iterations.

Table 2
Parameter settings of XGBoost.

Algorithm	Parameter meaning	Parameter	The parameter value	Setting basis
XGBoost	The number of decision trees	n_estimators	70	This parameter is very powerful and can adjust the model to the limit at one time.
	The maximum depth of decision tree	max_depth	7	The common value range is 10–100, and when the sample size and characteristic quantity are large, it can be modified appropriately.
	Training progress of the model	silent	1	When the data is huge and the algorithm speed is slow, this parameter can be used to monitor the training progress.
	Sample size of random sampling place with return	subsample	1	Control the sample size of sampling. The default is 1, which means 100% of the data is extracted at a time, and 0.1 means 10% of the data is extracted at a time.
	Iterative decision tree	eta	0.1	eta is the step size of the iterative decision tree, also known as the learning rate, which is used to ensure that each new tree has the best prediction effect.
	Selection of weak evaluator	booster	gbtree	Some trees are discarded in the process of tree building, which has better over fitting function than gradient lifting tree.
	Parameters of regular terms	alpha	10	When alpha and lambda are larger and the penalty is heavier, the proportion of regular terms is larger and the complexity of the model is lower.
	Control of model complexity	gamma	2	Important parameters to prevent over fitting.

$$X_{\mathrm{RMSE}} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{26}$$

where n means the quantity of power load data, y_t means the real power load data, and $\hat{y}_t$ means the predicted power load data.

5.3. Experimental setup

(1) Parameter settings of Attention-Bi-LSTM

The accuracy of the Attention-Bi-LSTM is mainly determined by the number of units, dimension of input feature, dimension of hidden layer state, the number of unit layers, hidden layer width and iteration times. The parameter settings of Attention-Bi-LSTM are given in Table 1.

(2) Parameter settings of XGBoost

The accuracy of the XGBoost is mainly determined by the number of decision trees, training progress of the model, sample size of random sampling place with return, iterative decision tree, selection of weak evaluator, objective function of XGBoost, parameters of regular terms, and control of model complexity. The parameter settings of XGBoost are provided in Table 2.

5.4. Influence of data preprocessing on prediction results

To verify the significance of the WGRP algorithm, we extract the short holiday data from May 1, 2020 to May 5, 2020 in the dataset of Singapore power market from January 1, 2019 to December 31, 2020. According to the local conditions of Singapore, the characteristics of historical load series, hour, day type, and holiday type are determined. Then, the prediction accuracy of benchmark models, other models and our prediction model are compared before and after data preprocessing. Specifically, as shown in Table 1, Model 1, Model 3, Model 5, Model 7, Model 9, Model 11 and Model 13 are not preprocessed by WGRP algorithm. For holidays, Model 2, Model 4, Model 6, Model 8, Model 10, Model 12 and Model 14 select the historical data with a high similarity to this holiday with the WGRP algorithm, to make the overall historical data general. The detailed description of the 14 models in seven groups is provided in Table 3.

The WGRP performance experimental results of group 1 to 6 are shown as in Fig. 6, and the comparison results of group 7 is indicated as in Fig. 7. As the curve trend shows, it is not difficult to reveal that in the seven groups of experiments, the predicted results of the model pretreated by WGRP are more consistent with the actual value curve in terms of proximity and trend. Thus, it follows that WGRP algorithm can make the data more universal, and using it to preprocess the data has great advantages in improving the prediction accuracy of the model.

The error analysis of seven groups is shown in Table 4. In the light of evaluation indicators, the prediction results of Model 2, Model 4, Model 6, Model 8, Model 10, Model 12 and Model 14 pretreated by the WGRP algorithm are all better than those without the WGRP algorithm in the group. Thus, for the two benchmark models, four other models, and our prediction model, the WGRP algorithm can help improve the prediction accuracy. Therefore, according to all error results of seven groups, we can

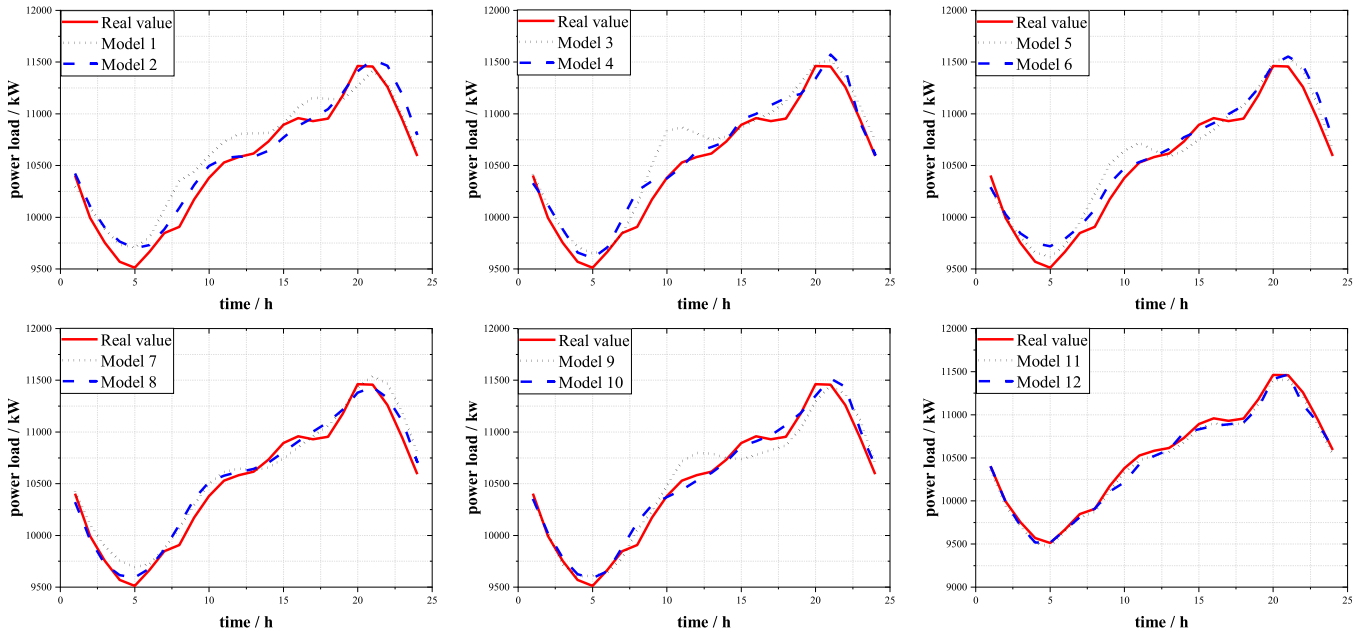

Fig. 6. Comparative experiment of WGRP algorithm of group 1 to 6.

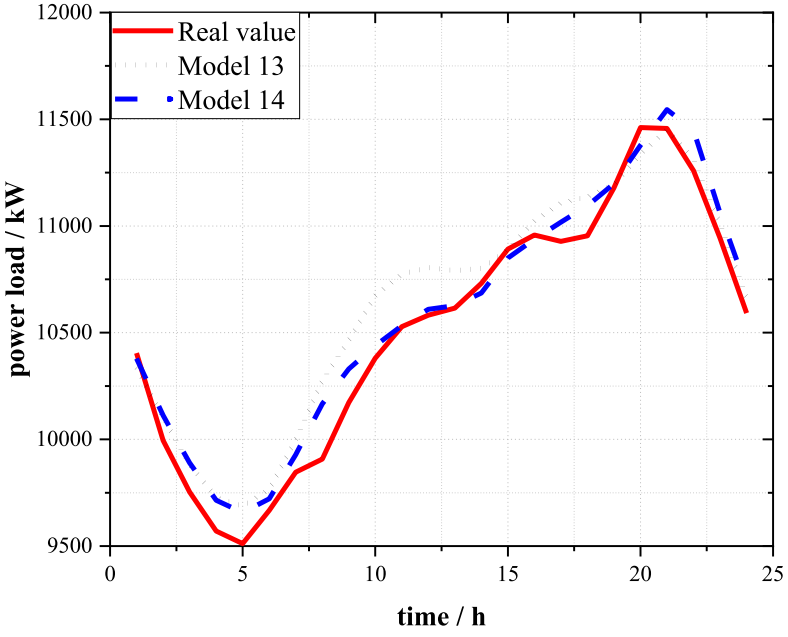

Fig. 7. Comparative experiment of WGRP algorithm of group 7.

conclude that it is necessary to use WGRP algorithm to preprocess holiday data, which proves the superiority of our prediction method.

When the WGRP algorithm is not considered, the predicted results are basically consistent with the change trend of the actual value; but, the differences between them are large. After we use the WGRP algorithm, the gap between the prediction results and real value can be reduced. Consequently, to decrease the prediction error, we first use the WGRP algorithm to process the holiday data in the historical data. Then, in order to verify the advantages of the algorithm in improving prediction accuracy and reducing error, we employ the same model to predict the processed and unprocessed data. According to the experiment results of the seven groups, we find that the prediction results of data processing are closer to the true value, and the accuracy of prediction can be further improved. Therefore, using the WGRP algorithm to process the holiday data can convincingly make the data more general and improve the prediction accuracy.

5.5. Analysis of prediction results

5.5.1. Experiment I: Singapore electricity market

Taking Singapore power load data with hourly resolution as an example, we do this experiment to verify the applicability of our proposed power load forecasting approach. First, we use the Attention-Bi-LSTM and XGBoost models to predict the same set of data, and obtain the prediction errors of the two models. Secondly, according to the errors, we assign weights to the above two prediction models and combine them with the error reciprocal method. The model with a smaller error is given a higher weight.

Table 3
Model description.

Group	Model	Description
Group 1	Model 1	LSTM
	Model 2	WGRP-LSTM
Group 2	Model 3	Bi-LSTM
	Model 4	WGRP-Bi-LSTM
Group 3	Model 5	Attention-RNN
	Model 6	WGRP-Attention-RNN
Group 4	Model 7	Attention-LSTM
	Model 8	WGRP-Attention-LSTM
Group 5	Model 9	Attention-Bi-LSTM
	Model 10	WGRP-Attention-Bi-LSTM
Group 6	Model 11	XGBoost
	Model 12	WGRP-XGBoost
Group 7	Model 13	Attention-Bi-LSTM+XGBoost combined model
	Model 14	WGRP-Attention-Bi-LSTM+XGBoost combined model

Table 4
Error analysis of 7 groups of WGRP algorithm comparison experiments.

Group	Model	X_{MAPE}/%	X_{MAE}/kW	X_{RMSE}/kW
Group 1	Model 1	1.453	149.915	178.589
	Model 2	**1.092**	**113.619**	**141.969**
Group 2	Model 3	1.279	133.021	174.571
	Model 4	**1.050**	**109.842**	**135.554**
Group 3	Model 5	1.128	117.693	144.475
	Model 6	**1.015**	**105.492**	**134.449**
Group 4	Model 7	1.028	107.183	125.881
	Model 8	**0.855**	**90.262**	**117.171**
Group 5	Model 9	0.971	103.063	119.726
	Model 10	**0.613**	**64.476**	**82.995**
Group 6	Model 11	0.475	50.053	50.932
	Model 12	**0.531**	**56.312**	**72.032**
Group 7	Model 13	0.359	38.341	48.076
	Model 14	**0.345**	**36.515**	**47.714**

Specifically, according to the main evaluation index MAPE, the results are substituted into the error reciprocal formula. Thus, the weights of Attention-Bi-LSTM and XGBoost are calculated as 0.4252 and 0.5748, respectively. Thirdly, to verify the effectiveness of the combination model proposed in this paper, we use the LSTM, Bi-LSTM, Attention-RNN, Attention-LSTM, Attention-Bi-LSTM, XGBoost and "Attention-Bi-LSTM + XGBoost" combined forecasting model to predict the data, and compare the seven prediction results to show that our proposed model is most effective. From our results we observe the following issues.

(1) Trend comparison between prediction results and real value.

(a) Fig. 8 shows the comparison between the real values and the prediction results of LSTM, Bi-LSTM, Attention-RNN, Attention-LSTM, Attention-Bi-LSTM, XGBoost models, and "Attention-Bi-LSTM + XGBoost" combined prediction model. We learn that the prediction accuracy of "Attention-Bi-LSTM + XGBoost" combined prediction model is the highest, and the Bi-LSTM model has the lowest prediction accuracy.

(b) The predicted value curve of the "Attention-Bi-LSTM + XGBoost" combined prediction model is the closest to the real value curve. That is, the fitting effect of our proposed model is the best, and the change trend is roughly the same.

(c) Overall, the "Attention-Bi-LSTM + XGBoost" combined prediction model shows the performance of the optimal and behaves is better than others in prediction accuracy and sensitivity to proportionality changes. In order to more clearly observe the trend and proximity between the prediction results of each model selected in this paper and the actual value, we locally enlarge Fig. 8. The locally-enlarged drawing of the results of 24 h on January 7, 2020 in Singapore is shown in Fig. 9. From Fig. 9, we can observe that the trend of the prediction result curve of our combined prediction model is more consistent with the real value curve than the benchmark and other models. Thus, we conclude that the prediction accuracy of our proposed method is the highest according to both the overall trend diagram and the locally enlarged diagram.

(2) Error comparison and analysis

The MAPE, MAE, and RMSE values of the above six models are shown in Table 5. By comparing the error values in Table 5, we can draw the following conclusions.

(a) Since LSTM is an improvement of recurrent neural network (RNN), the error of Attention-LSTM is less than that of Attention-RNN, which shows the necessity of selecting the LSTM model in this paper.

(b) According to the comparison of LSTM, Attention-LSTM, and Attention-RNN models, the comprehensive comparison of the three prediction methods indicates the significant effect of attention mechanism on prediction accuracy.

(c) Among all the accuracy test standards for benchmark and other models, the error value of XGBoost is the smallest, which means that XGBoost has an extremely excellent prediction performance. We use XGBoost to optimize the Attention-Bi-LSTM model, which can significantly improve the accuracy of the Attention-Bi-LSTM model.

(d) Table 5 exposes that the values of MAPE, MAE and RMSE of the "Attention-Bi-LSTM + XGBoost" combined prediction model is 0.387, 43.206, and 54.357, respectively, which are the smallest from a holistic perspective. Therefore, the test results show that the combined prediction method of the two models can reduce the prediction error as a whole, thus being better than the single prediction model and having the highest prediction accuracy.

5.5.2. Experiment II: Norway electricity market

The hourly load data of Norway is used as another test data in this experiment. The purpose is to further verify and evaluate the effectiveness of the proposed method by using a different dataset to compare our results with those from other prediction approaches. Similarly, according to the main evaluation index MAPE, we find that, with the reciprocal error formula, the calculated weights of Attention-Bi-LSTM and XGBoost are 0.4273 and 0.5727, respectively. Figs. 10 and 11 as well as Table 6 describe the graphics and error results of experiment II of the proposed combined prediction model compared to the existing prediction models. The results reveal the insights below.

(1) Trend comparison between prediction results and real value

Through the comparison of the seven prediction models in Figs. 10 and 11, we learn that, compared to the benchmark and other models, the line of the combined prediction model proposed in this paper is the closest to the trend and value of the real value. This indicates it is feasible to optimize the model with the combined method. Particularly, we learn from Fig. 10 that the predicted result curve and the real value curve of the "Attention-Bi-LSTM + XGBoost" combined model are the closest in both trend and proximity. Except for our prediction method, all the curves for the prediction results of the remaining benchmark and other models have significantly large deviation to the curves of the real value. That is, from the perspective of graphic trend,

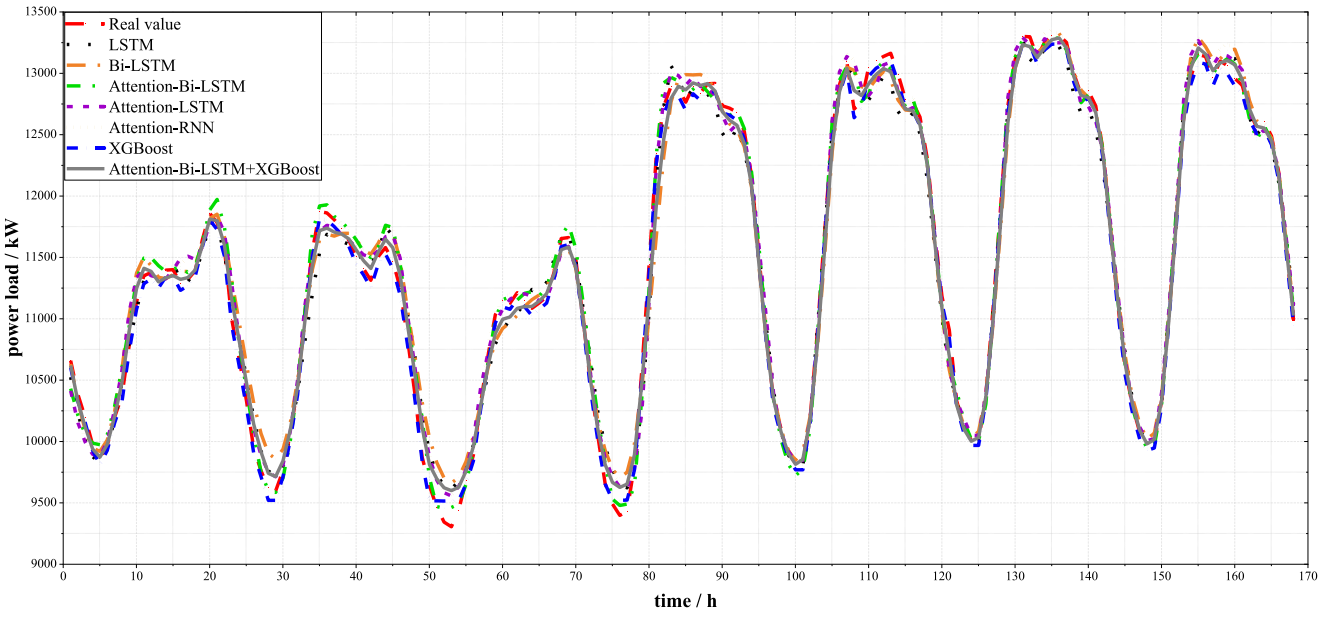

Fig. 8. Comparison between LSTM, Bi-LSTM, Attention-RNN, Attention-LSTM, Attention-Bi-LSTM, XGBoost and Attention-Bi-LSTM + XGBoost prediction results with real values of Singapore power market.

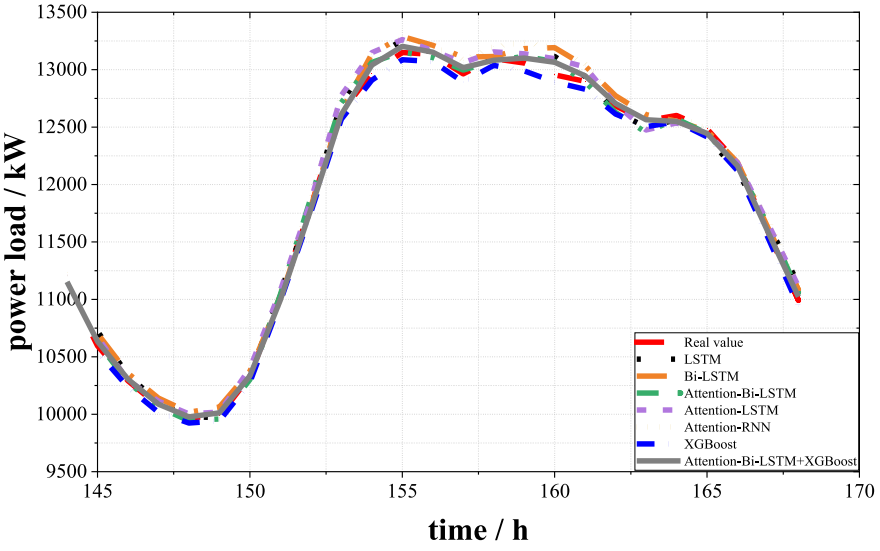

Fig. 9. The local enlarged drawing of Singapore's results on January 7, 2020 (24 h).

Table 5
Error analysis between LSTM, Bi-LSTM, Attention-RNN, Attention-LSTM, Attention-Bi-LSTM, XGBoost and Attention-Bi-LSTM+XGBoost Singapore power market.

Types of models	Model	X_{MAPE}/%	X_{MAE}/kW	X_{RMSE}/kW
Other models	LSTM	1.011	113.480	147.083
	Bi-LSTM	1.155	127.812	168.841
	Attention-RNN	0.886	98.16	130.704
	Attention-LSTM	0.877	97.691	125.574
Benchmark models	Attention-Bi-LSTM	0.711	79.418	105.023
	XGBoost	0.526	59.463	62.585
Our proposed model	Attention-Bi-LSTM+XGBoost combined model	**0.387**	**43.206**	**54.357**

our proposed model has the highest fitting degree and excellent prediction performance.

(2) Error comparison and analysis.

Table 6 lists the numerical results of three accuracy tests, which show that in the accuracy test, the method proposed in this paper achieves the minimum results of MAPE, MAE, and RMSE (i.e., 0.682, 96.278, and 125.343, respectively). According to the evaluation, we conclude that the proposed model is more accurate than those benchmark and other frameworks. In addition, we further ensure the effectiveness of the attention mechanism and combination optimization model that were proved in Experiment I.

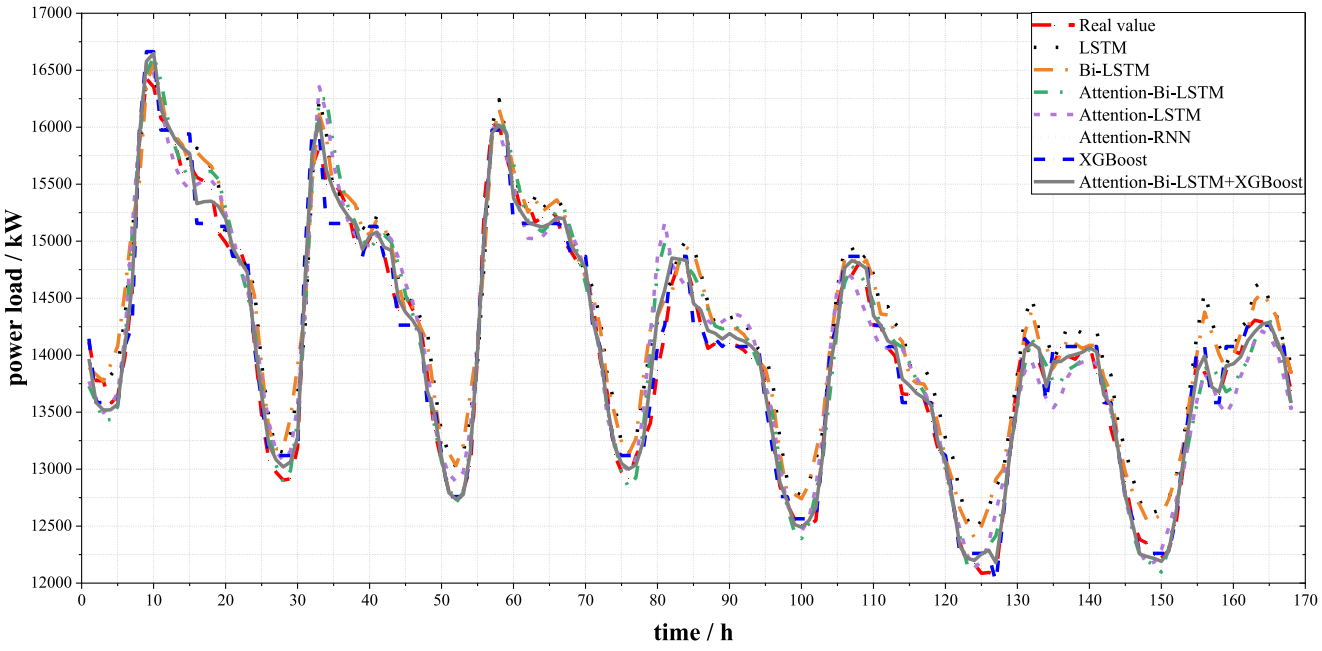

Fig. 10. Comparison between LSTM, Bi-LSTM, Attention-RNN, Attention-LSTM, Attention-Bi-LSTM, XGBoost and Attention-Bi-LSTM + XGBoost prediction results with real values of Norway power market.

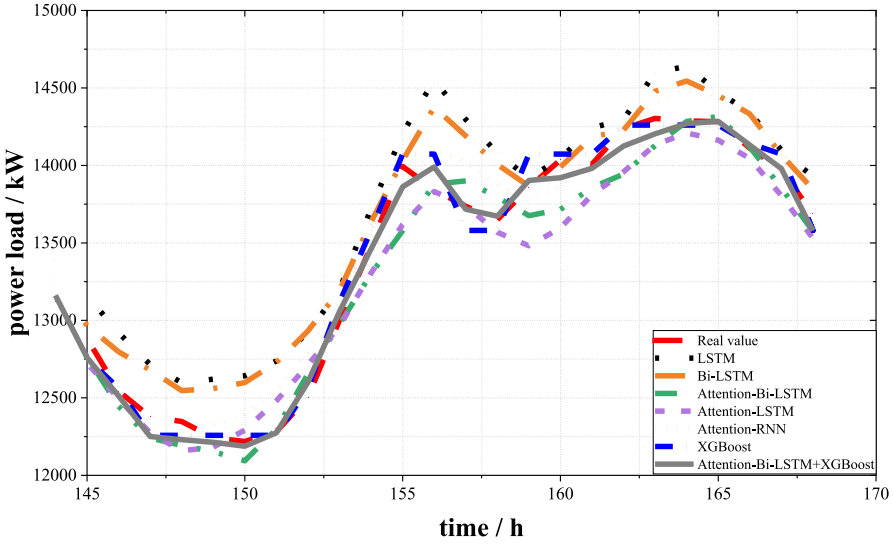

Fig. 11. The local enlarged drawing of Norway's results on May 16, 2021 (24 h).

Table 6
Error analysis between LSTM, Bi-LSTM, Attention-RNN, Attention-LSTM, Attention-Bi-LSTM, XGBoost and Attention-Bi-LSTM+XGBoost of Norway power market.

Types of models	Model	X_{MAPE}/%	X_{MAE}/kW	X_{RMSE}/kW
Other models	LSTM	1.886	257.791	301.141
	Bi-LSTM	1.648	225.233	279.765
	Attention-RNN	1.230	182.122	238.673
	Attention-LSTM	1.229	171.672	233.651
Benchmark models	Attention-Bi-LSTM	1.091	153.595	210.720
	XGBoost	0.814	115.892	148.007
Our proposed model	Attention-Bi-LSTM+XGBoost combined model	**0.682**	**96.278**	**125.343**

By comparing with the local enlarged drawing of Singapore' results, that is, Fig. 9, we find that the accuracy of the seven models used in this paper for the prediction of the Singapore electricity market is higher than that of the Norway electricity market. This exposes that, as the verification of two markets indicates, our method has good prediction performance. Moreover, our combined prediction method is more suitable for the data of Singapore power market. That is, according to the unique characteristics of different market data, we still need to constantly develop better and suitable prediction models. Then,

by comparing Figs. 9 and 11, we can find that the trend of the local enlarged drawing of the Singapore power market is more stable, which proves the prediction results of the Singapore power market have a better fitting effect. Therefore, we can draw the conclusion that different power markets have different characteristics. In order to evaluate the performance of some prediction method, it is necessary to use this method to conduct experimental evaluation on multiple data sets.

6. Conclusions

Power load forecasting plays an important role in balancing energy distribution, economy, and the safe and reliable operation of power systems. An accurate load forecasting can reduce the cost and risk of power operations, and can improve the environmental and economic benefits of power grids. Thus, in this paper, with an aim to enhance the accuracy and stability of power load predictions, we propose a novel hybrid Attention-Bi-LSTM + XGBoost power load combined forecasting method based on WGRP algorithm. On the phase of data preprocessing, the historical load series of holidays are selected by WGRP algorithm, and better prediction results are obtained. In addition, for the accuracy comparison between our combined forecasting model and the benchmark and other models such as Attention-Bi-LSTM, XGBoost, LSTM, Bi-LSTM, Attention-LSTM, and Attention-RNN, we perform two case studies using the datasets of Singapore and Norway power markets. We can draw the following conclusions.

(1) Using the WGRP algorithm to preprocess holiday data can effectively improve the prediction accuracy of the model.
(2) Attention mechanism allows the Bi-LSTM model to emphasize the influence of important factors, which can eliminate redundancy and improve prediction performance.
(3) Adding regular terms to XGBoost can effectively prevent over fitting and reduce calculation, so as to greatly improve the efficiency of the algorithm. Therefore, using XGBoost model for optimization can greatly reduce the error of the model.
(4) According to the comparison of prediction results compared to two benchmark models and four other models, we find that XGBoost model has the best prediction performance and the smallest error. Then, compared to the optimal benchmark model XGBoost, the MAPE of our prediction method is 0.387 and 0.682, which are reduced by 26.43% and 16.22%, respectively; the MAE of our prediction method is 43.206 and 96.278 which are reduced by 27.34% and 16.92%, respectively; the RMSE of our prediction method is 54.357 and 125.343 which are reduced by 13.15% and 15.31%, respectively. Hence, we can certainly draw the conclusion that the prediction result of the "Attention-Bi-LSTM + XGBoost" combined model has the lowest errors and is the closest to the actual value than those of the single models, and the trend of the proposed model is roughly the same as the real values.

In conclusion, the combined forecasting method proposed in this paper is more effective than the single classical and modern prediction methods, hybrid prediction methods, and other existing combined prediction methods. Our model can reduce the error and obtain a higher power load prediction accuracy to reduce the unnecessary waste in power markets and improve the reliability and safety of power system operations.

In the future, we may further improve our model from the following aspects:

(1) We shall use more evaluation indicators of regression prediction model to verify the accuracy and further improve the reliability of our proposed model.
(2) The weight of our proposed model is only calculated according to the error, which is one-sided. Next, we shall find an optimization method to weight assignment method, so that the weight can be calculated by the data, and the weight changes according to the selected data.
(3) We shall find a suitable method to further optimize the parameters of our prediction model, and try to use more advanced models for combination, in order to continuously improve the prediction performance.
(4) XGBoost model is a part of our prediction method. We use its basic model but do not optimize its hyper-parameters. Therefore, in the future work, we should use methods such as Bayesian optimization to optimize the hyper-parameters of XGBoost to further improve its prediction performance.

We expect that, in the future, we can consider the above issues, and also strive to apply the method to more fields.

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, China (No. 20YJA630009), and Social Science Planning Project of Shandong Province, China (No. 20CSDJ15). This work is also financially supported by the Faculty Research Grant (FRG) of Lingnan University, China (No. DB21B1).

References

[1] M. Sobhani, T. Hong, C. Martin, Temperature anomaly detection for electric load forecasting, Int. J. Forecast. 36 (2) (2020) 324–333.
[2] A. Rosato, M. Panella, A. Andreotti, O.A. Mohammed, R. Araneo, Two-stage dynamic management in energy communities using a decision system based on elastic net regularization, Appl. Energy 291 (2021) 116852.
[3] P. Jiang, R. Li, N. Liu, Y. Gao, A novel composite electricity demand forecasting framework by data processing and optimized support vector machine, Appl. Energy 260 (2020) 114243.
[4] P. Nystrup, E. Lindström, J.K. Møller, H. Madsen, Dimensionality reduction in forecasting with temporal hierarchies, Int. J. Forecast. 37 (3) (2021) 1127–1146.
[5] X. Qiu, Y. Ren, P.N. Suganthan, G.A. Amaratunga, Empirical mode decomposition based ensemble deep learning for load demand time series forecasting, Appl. Soft Comput. 54 (2017) 246–255.
[6] X. Qiu, L. Zhang, P.N. Suganthan, G.A. Amaratunga, Oblique random forest ensemble via least square estimation for time series forecasting, Inform. Sci. 420 (2017) 249–262.
[7] F. Wu, C. Cattani, W. Song, E. Zio, Fractional ARIMA with an improved cuckoo search optimization for the efficient short-term power load forecasting, Alexandria Eng. J. 59 (5) (2020) 3111–3118.
[8] H. Zhao, S. Guo, An optimized grey model for annual power load forecasting, Energy 107 (2016) 272–286.
[9] Z. Wang, X. Zhou, J. Tian, T. Huang, Hierarchical parameter optimization based support vector regression for power load forecasting, Sustainable Cities Soc. 71 (2021) 102937.
[10] M.L.M. Lopes, C.R. Minussi, A.D.P. Lotufo, Electric load forecasting using a fuzzy ART & ARTMAP neural network, Appl. Soft Comput. 5 (2) (2005) 235–244.
[11] R. Gordillo-Orquera, L.M. Lopez-Ramos, S. Muñoz-Romero, P. Iglesias-Casarrubios, D. Arcos-Avilés, A.G. Marques, J.L. Rojo-Álvarez, Analyzing and forecasting electrical load consumption in healthcare buildings, Energies 11 (3) (2018) 493.

[12] Y. Yang, J. Che, C. Deng, L. Li, Sequential grid approach based support vector regression for short-term electric load forecasting, Appl. Energy 238 (2019) 1010–1021.
[13] Z. Guo, K. Zhou, X. Zhang, S. Yang, A deep learning model for short-term power load and probability density forecasting, Energy 160 (2018) 1186–1200.
[14] Y. Wang, L. Wang, F. Yang, W. Di, Q. Chang, Advantages of direct input-to-output connections in neural networks: The elman network for stock index forecasting, Inform. Sci. 547 (2021) 1066–1079.
[15] Y. Liang, D. Niu, W.C. Hong, Short term load forecasting based on feature extraction and improved general regression neural network model, Energy 166 (2019) 653–663.
[16] K. Wang, C. Xu, Y. Zhang, S. Guo, A.Y. Zomaya, Robust big data analytics for electricity price forecasting in the smart grid, IEEE Trans. Big Data 5 (1) (2017) 34–45.
[17] M. Brégère, M. Huard, Online hierarchical forecasting for power consumption data, Int. J. Forecast. 38 (1) (2022) 339–351.
[18] G.T. Ribeiro, V.C. Mariani, L. dos Santos Coelho, Enhanced ensemble structures using wavelet neural networks applied to short-term load forecasting, Eng. Appl. Artif. Intell. 82 (2019) 272–281.
[19] G. Hafeez, I. Khan, S. Jan, I.A. Shah, F.A. Khan, A. Derhab, A novel hybrid load forecasting framework with intelligent feature engineering and optimization algorithm in smart grid, Appl. Energy 299 (2021) 117178.
[20] E.O.N. Jnr, Y.Y. Ziggah, S. Relvas, Hybrid ensemble intelligent model based on wavelet transform, swarm intelligence and artificial neural network for electricity demand forecasting, Sustainable Cities Soc. 66 (2021) 102679.
[21] S.R. Polamuri, K. Srinivas, A.K. Mohan, Multi-model generative adversarial network hybrid prediction algorithm (MMGAN-HPA) for stock market prices prediction, J. King Saud Univ.-Comput. Inf. Sci. (2021).
[22] M. Barman, N.B.D. Choudhury, A similarity based hybrid GWO-SVM method of power system load forecasting for regional special event days in anomalous load situations in assam, India, Sustainable Cities Soc. 61 (2020) 102311.
[23] Z. Wu, X. Zhao, Y. Ma, X. Zhao, A hybrid model based on modified multi-objective cuckoo search algorithm for short-term load forecasting, Appl. Energy 237 (2019) 896–909.
[24] 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.
[25] A. Lahouar, J.B.H. Slama, Day-ahead load forecast using random forest and expert input selection, Energy Convers. Manage. 103 (2015) 1040–1051.
[26] 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.
[27] J.M. Bates, C.W. Granger, The combination of forecasts, J. Oper. Res. Soc. 20 (4) (1969) 451–468.
[28] H. Nie, G. Liu, X. Liu, Y. Wang, Hybrid of ARIMA and SVMs for short-term load forecasting, Energy Procedia 16 (2012) 1455–1460.
[29] D. Deng, J. Li, Z. Zhang, Y. Teng, Q. Huang, Short-term electric load forecasting based on EEMD-GRU-MLR, Power Syst. Technol. 44 (2) (2020) 593–602.
[30] Y. Chu, P. Xu, M. Li, Z. Chen, Z. Chen, Y. Chen, W. Li, Short-term metropolitan-scale electric load forecasting based on load decomposition and ensemble algorithms, Energy Build. 225 (2020) 110343.
[31] Z. Chen, J. Liu, C. Li, X. Ji, D. Li, Y. Huang, F. Di, Ultra short-term power load forecasting based on combined LSTM-XGBoost model, Power Syst. Technol. 44 (2) (2020) 614–620.
[32] J. Zhuang, G. Yang, H. Zheng, CNN-LSTM-xgboost short-term power load forecasting method based on multi model fusion, Electr. Power 54 (5) (2021) 46–55.
[33] Y. Nie, P. Jiang, H. Zhang, A novel hybrid model based on combined preprocessing method and advanced optimization algorithm for power load forecasting, Appl. Soft Comput. 97 (2020) 106809.
[34] R.C. Zheng, J. Gu, Z.J. Jin, Research on short-term load forecasting variable selection based on fusion of data driven method and forecast error driven method, Proc. CSEE 40 (2) (2020) 487–499.
[35] X.M. Yu, W.Z. Feng, H. Wang, Q. Chu, Q. Chen, An attention mechanism and multi-granularity-based Bi-LSTM model for Chinese Q & A system, Soft Comput. 24 (8) (2020) 5831–5845.
[36] X. Zheng, W. Chen, An attention-based bi-LSTM method for visual object classification via EEG, Biomed. Signal Process. Control 63 (2021) 102174.
[37] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, Xgboost: extreme gradient boosting, 2015, pp. 1–4, R package version 0.4-2, 1(4).
[38] P. Trizoglou, X. Liu, Z. Lin, Fault detection by an ensemble framework of extreme gradient boosting (xgboost) in the operation of offshore wind turbines, Renew. Energy 179 (2021) 945–962.
[39] W.H. Chung, Y.H. Gu, S.J. Yoo, District heater load forecasting based on machine learning and parallel CNN-LSTM attention, Energy 246 (2022) 123350.
[40] X. Tong, J. Wang, C. Zhang, T. Wu, H. Wang, Y. Wang, LS-LSTM-AE: Power load forecasting via long-short series features and LSTM-autoencoder, Energy Rep. 8 (2022) 596–603.
[41] W. Zha, J. Liu, Y. Li, Y. Liang, Ultra-short-term power forecast method for the wind farm based on feature selection and temporal convolution network, ISA Trans. (2022).
[42] S. Goyal, G.K. Goyal, Cascade and feedforward backpropagation artificial neural networks models for prediction of sensory quality of instant coffee flavoured sterilized drink, Canad. J. Artif. Intell. Mach. Learn. Pattern Recognit. 2 (6) (2011) 78–82.
[43] T. Jayalakshmi, A. Santhakumaran, Statistical normalization and back propagation for classification, Int. J. Comput. Theory Eng. 3 (1) (2011) 1793–8201.