IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 September 2023, accepted 2 October 2023, date of publication 5 October 2023, date of current version 16 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3322167

RESEARCH ARTICLE

Short-Term Load Forecasting Based on Multi-Scale Ensemble Deep Learning Neural Network

QIN SHEN, LI MO, GUANJUN LIU, JIANZHONG ZHOU, YONGCHUAN ZHANG, AND PINAN REN
School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
Hubei Provincial Key Laboratory of Digital Watershed Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
Institute of Water Resources and Hydropower, Huazhong University of Science and Technology, Wuhan 430074, China

Corresponding author: Li Mo (moli@hust.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 52379011 and Grant 51979114.

ABSTRACT High-precision load forecasting is crucial for the power system planning and electricity market transactions. Recently, deep learning models have been widely used due to their powerful data mining capabilities. However, the existing research mainly focus on model structure adjustment and input feature selection, ignoring the influence of model ensemble on prediction. A single deep learning model is not yet able to address the various complex challenges that arise in short-term load forecasting. To improve the accuracy of short-term load forecasting, this paper proposes a novel multi-scale ensemble method and multi-scale ensemble neural network. This neural network uses long short-term memory, gate recurrent units, and temporal convolutional network as the basic model. By coupling the stochastic weight averaging ensemble method and differential evolution ensemble method, these deep learning networks were assembled from single-model scale and multi-model scale, respectively, thereby effectively improving the model prediction accuracy. For predicting the power load of Hubei Province in China, meteorological features and time features were in consideration. The proposed model was trained and compared with eleven intelligent short-term load forecasting models, including machine learning, deep learning and ensemble deep learning models. Simulations show that the proposed model has the best comprehensive prediction performance. This study highlights the power of ensemble deep learning models coupled with multiple ensemble techniques and the promising prospect of our proposed model in short-term load forecasting.

INDEX TERMS Short-term load forecasting, ensemble model, stochastic weight averaging, deep learning, differential evolution algorithm.

I. INTRODUCTION

Accurate load forecasting is an indispensable component in power system planning and electricity market transactions [1]. With the proposed emission reduction target, it is an inevitable trend to construct a new power system with new energy as the main body [2]. According to the forecasting period, load forecasting can be segmented into four categories: ultra-short-term, short-term, medium-term, and long-term forecasting [3], the most widely studied is short-term load forecasting (STLF), which generally refers to predicting the load of the next day and week [4]. Three research methods are mainly used for STLF: statistical, data-driven, and ensemble models [5].

The statistical model determines the potential rules from historical load samples according to statistical formulas or known relationships [6], including multiple linear regression (LR) [7], exponential smoothing [8], and autoregressive integrated moving normal algorithms (ARIMA) [9]. In [10], the paper integrates ARIMA and the nonlinear grey model (MNGM) into a combined MNGM-ARIMA model. The datasets from the three countries demonstrate the proposed

The associate editor coordinating the review of this manuscript and approving it for publication was Fabio Mottola.

model has shown sound performance. In [11], ARIMA and seasonal ARIMA methods were combined to measure future primary energy demands, and this combined model is being used with success in a real-time control system. However, due to the nonlinear identities of the power load time series, statistical methods perform fast in the calculation but perform poorly in STLF owing to their limitations. In contrast, data-driven models can better handle the nonlinear characteristics of load sequences and are widely used in STLF [12].

The data-driven model generates prediction by establishing a mapping model between input and output [13]. Typical data-driven includes decision tree [14], support vector machine (SVM) [15], and artificial neural network (ANN) [16]. The most commonly used are ANNs, including reinforcement [17], feedforward [18], regression [19], generalized [20] and wavelet [21] neural networks. A method that combines the ANN and wavelet denoising algorithm has been demonstrated to significantly improve the accuracy of STLF [22]. Due to the rapid development of computer science and technology, deep learning networks on the strength of ANN have gained extensive attention [23]. Deep learning neural networks can handle data non-stationarity and long-range dependencies owing to their powerful deep learning framework with multiple hidden layers [24]. Representative algorithms include long short-term memory (LSTM) [25], gate recurrent units (GRU) [26], and temporal convolutional networks (TCN) [27]. In the STLF tasks, the improved LSTM model considering relevant factors was adopted, and hyper-parameter optimization was implemented on the Bayesian optimization algorithm (BOA), the proposed method is validated by seven benchmark methods [28]. In [29], TCN and attention mechanism (AM) are combined to enhance the forecast performance. Simultaneously, fuzzy c-means (FCM) joined with dynamic time warping (DTW) are for data processing. The experiment results illustrate that the improved TCN model is more effective than the contrast models. Data-driven models have captured widespread concern from researchers in load forecasting, owing to their excellent generalization ability, global optimal solution, and fast calculation property [30]. However, there are still limitations in prediction accuracy for a single model. Existing studies have demonstrated that ensemble models have more accuracy and robustness than single models [31].

Ensemble models can obtain superior generalization performance by using specific ensemble learning techniques. According to different ensemble learning techniques, ensemble models can be further divided into bagging-based ensemble models, boosting-based ensemble models, stacking-based ensemble models, and ensemble deep learning models [32]. Bagging-based ensemble models [33], which use bootstrap aggregation ensemble technology, are one of the first proposed ensemble learning models. The core idea of them is to independently train multiple weak learners and connect them in parallel. Random forest [34] is a widely used bagging-based ensemble model. In contrast, boosting-based ensemble models [35] are a special model that uses the boosting ensemble technology. The main idea behind them involves iteratively applying the weak learners to transform them into strong learners through a series of connections. Typical boosting-based ensemble models applied to load forecasting include extreme gradient boosting (XGBoost) [36] and light gradient boosting machine (lightGBM) [37], etc. Besides, models with stacked generalization technology are called stacking-based ensemble models [38]. Unlike the above ensemble models, stacking-based ensemble models usually consider heterogeneous weak learners (different learning algorithms are combined) and adopt meta-models to combine base models. In [39], a novel load forecasting method combining SVR and stacked generalization technology is proposed for STLF and the results validate the effectiveness of the stacking-based ensemble. With the development of deep learning, the idea of the ensemble was gradually applied to deep learning, and the ensemble deep learning model came into being. Lai et al. [40] paralleled multiple radial basis function neural networks into a deep ensemble model and verified its effectiveness on three different load datasets. Niu et al. [41] proposed a new ensemble deep learning model for STLF by concatenating a convolutional neural network (CNN) and bidirectional recurrent unit (BiGRU). The results show that the ensemble model has a higher generalization ability. The above studies initially demonstrate the powerful potential of ensemble deep learning models in dealing with STLF problems. However, in general, the use of ensemble for deep learning models is not nearly as widespread as it is for other models.

The above research all belong to the classical multi-model scale ensemble, which ensembles multiple models in series or in parallel. This approach offers a more competitive model but inevitably leads to an increase in the amount of computation. In recent years, an advanced single-model scale ensemble technique stochastic weight averaging (SWA) [42] has emerged in the deep learning field. Unlike traditional multi-model scale ensembles, this technique can ensemble multiple same models without incurring any additional computational cost. In summary, considering that the existing research in the STLF field mostly focuses on multi-model scale ensembles, the application of single-model scale ensemble is ignored, let alone the coupling of multiple-scale ensemble. Thus, this study pays close attention to the application of the multi-scale ensemble technique in the STLF field. The contributions of this article include:

(1) By introducing the advanced SWA technique and differential evolution (DE) ensemble approach from different scales, a patent multi-scale ensemble method (MSEM) for deep learning models is proposed in this study, which can enhance the generalization ability of models efficiently. This approach can provide a general paradigm for the ensemble of deep learning models.

(2) Performing the proposed MSEM to three typical deep learning models, the paper proposes a novel multi-scale

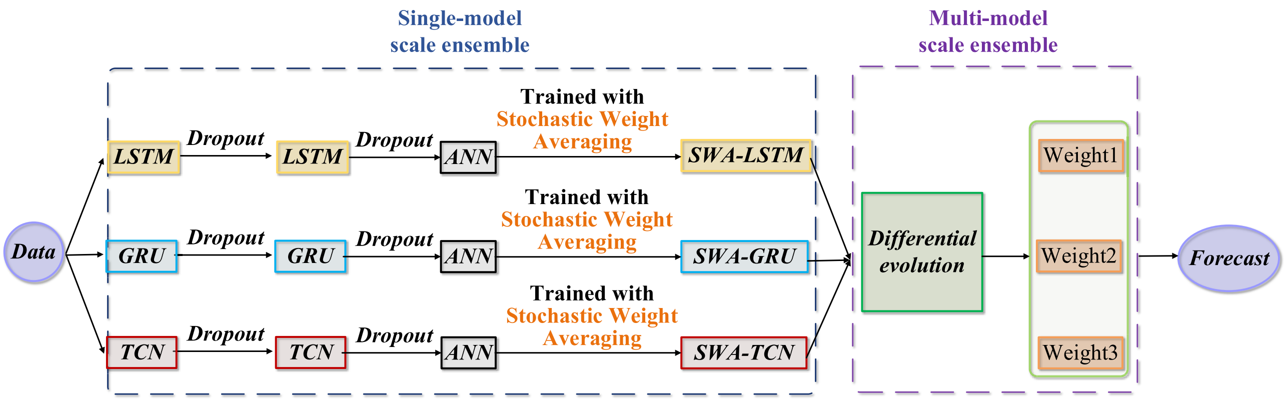

FIGURE 1. The Structure of proposed MSENN model.

ensemble neural network (MSENN) for load forecasting. By reducing the generalization error from multiple scales, the model can achieve high-precision fitting of future unknown samples in power load forecasting. This work can not only provide a competitive alternative model for grid workers but also enrich the diversity of STLF models.

(3) A comprehensive experimental contrast analysis is conducted among twelve models, which contain machine learning models, deep learning models, and ensemble deep learning models. The empirical results tested on three datasets show that the proposed MSENN surpasses other comparative models in comprehensive forecast performance.

The remainder of this study is structured as follows. Section II introduces the background of relevant methods, including deep learning models, the model ensemble method MSEM, and the process of the proposed method MSENN. Section III elaborates on the experiments and analyses of the proposed approach. The model results and discussion in contrast are illustrated in Section IV. Finally, Section V sketches out the conclusions of the paper.

II. METHODOLOGY

The paper proposes a novel multi-scale ensemble neural network MSENN to deal with STLF problems. The MSENN is constructed by ensemble LSTM, GRU and TCN from single-model scale and multi-model scale, respectively. The structure of the proposed MSENN is shown in Fig. 1. The details of the model will be introduced in the following subsections.

A. DEEP LEARNING MODELS

1) LONG SHORT-TERM MEMORY (LSTM)

The LSTM network is a memory-strengthened version of the RNN originally proposed by Hochreiter and Schmidhuber in 1997 [43]. The LSTM module consists of two hidden states and three control gates. The control gates include the forget, input, and output gates [44]. The internal operation of LSTM is presented in Fig. 2.

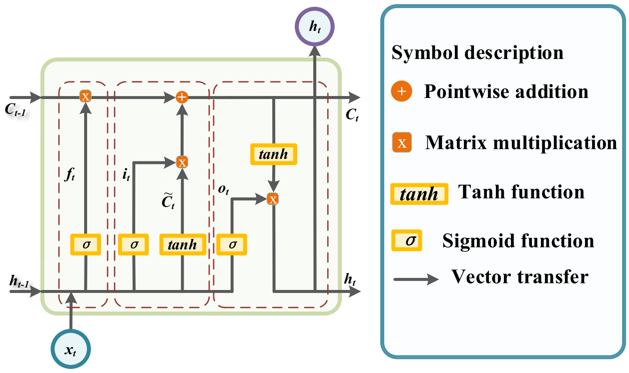

FIGURE 2. Structure of LSTM unit.

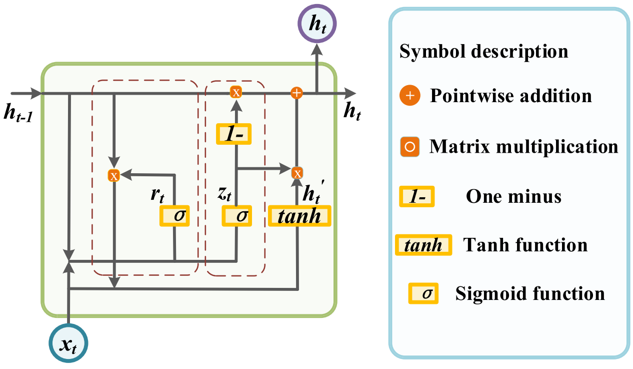

FIGURE 3. Structure of GRU unit.

2) GATED RECURRENT UNIT (GRU)

The GRU only has two gates: the reset and update gates [45]. The internal GRU operation is shown in Fig. 3. The reset gate, update gate, output candidate, and GRU output are computed as follows [46].

3) TEMPORAL CONVOLUTIONAL NETWORK (TCN)

The TCN is a time series data analysis neural network [47]. TCN incorporates causal convolution, dilated convolution and residual connection. Residual block is applied to avoid gradient descent, thus facilitating the gradient propagation

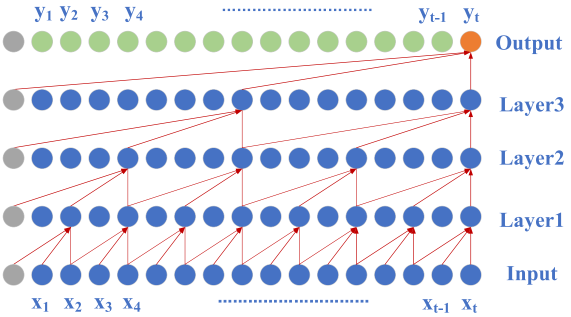

FIGURE 4. Structure of dilated casual convolution.

throughout the neural network [48]. The TCN structure is presented in Fig. 4.

B. MULTI-SCALE ENSEMBLE METHOD

A novel deep learning multi-scale ensemble method (MSEM) for load forecasting is proposed in this section. which combines SWA ensemble method and DE ensemble method.

1) STOCHASTIC WEIGHT AVERAGING (SWA)

In the proposed MSEM, SWA [42] was employed to ensemble same models from a single-model scale. It is a deep learning single-model weight ensemble technique that employs a simulated annealing cyclic learning rate strategy to explore the optimal solution space. The principle of SWA is introduced as follows. First, in the perspective of stochastic convex optimization, each trained model is a point in the weight space. The local minimum generated at the end of single learning rate cycle, inclining to scrape up on the border of the loss surface where the value is small. By averaging these points, obtaining a global optimal solution that tends toward the center of the region and has a small loss function value is extremely probable, as shown in Fig. 5.

For better understanding, Fig. 6 presents the sketch map of simulated annealing learning rate. As can be seen from the figure, simulated annealing learning rate α reduces within the range (α_2, α_1), and the pace of reduction varies from slow to fast to slow. After the minimum learning rate is reached, then start directly from the maximum learning rate, so it is discontinuous. Then repeating the same cycle to acquire multiple local minima. By saving and averaging multiple local optimal model weights, the generalization ability of deep learning models can be effectively improved. In addition, the method can enhance the stability of the training process without incurring additional cost.

In summary, SWA method can permit the model to converge faster and more stable, which can improve the generalization ability and prediction accuracy of model without additional computational cost from single-model scale.

2) DIFFERENTIAL EVOLUTION (DE)

To further integrate the advantages and improve the overall accuracy of multiple models, this study then introduces DE algorithm [49]. The method focuses on model ensemble from multi-model scale. In detail, the core idea of this method is adaptively assigning weights according to the prediction accuracy of each sub-model on the validation set. Instead of manual setting or mean value of weights in general method, DE method is capable of obtaining adaptive weights by its embedding optimization mechanism to focus on the advantages of multiply models. DE receives output from multiple models, which is then optimal weight averaged to acquire comprehensive output. The DE algorithm flowchart is shown in Fig. 7. In summary, DE is capable of assigning different weights according to the performance of sub-models, which can significantly improve the overall prediction accuracy of ensemble from multi-model scale.

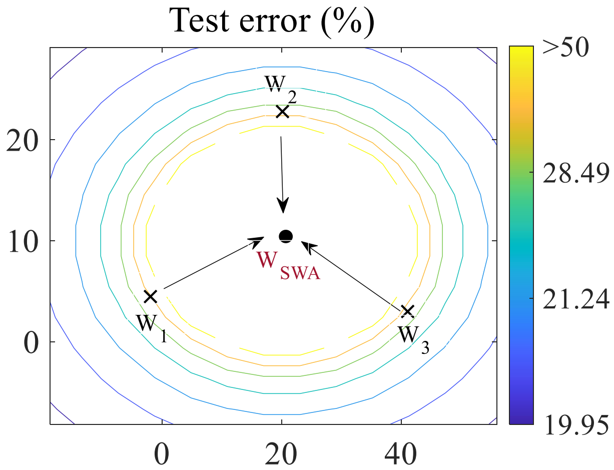

FIGURE 5. Train loss surface of model; (W1, W2, and W3 are the weights of model; WSWA is the average of W1, W2, and W3).

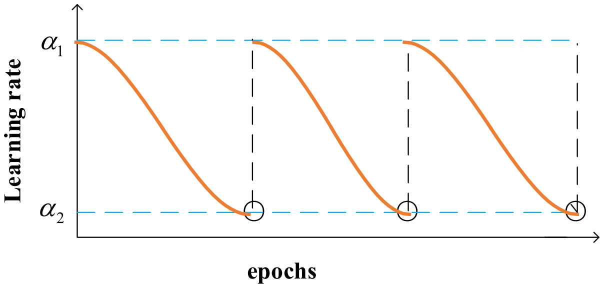

FIGURE 6. Simulated annealing learning rate of SWA.

C. DEEP LEARNING MULTI-SCALE ENSEMBLE MODEL

To enhance the prediction performance and calculation speed of model, a novel MSENN model on the strength of the MSEM method is brought forward. The specific implementation procedure of the proposed model is presented as follows.

1) ENSEMBLE MODEL SELECTION

The first step of constructing a novel ensemble model is to select the suitable fundamental neural network. The choice of model structure and number of sub-models has no restrictions. LSTM, GRU and TCN are sorted as basis of the

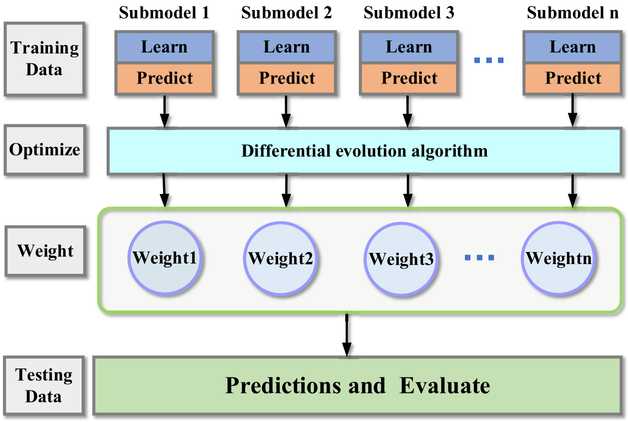

FIGURE 7. Sketch map of differential evolution ensemble method.

ensemble model. After selecting the base models, the proposed approach ensembles them at two different scales: single-model and multi-model ensembles. Details are presented in the following subsections.

2) SINGLE-MODEL SCALE ENSEMBLE

In our proposed MSENN, SWA is adopted to ensemble the model from a single-model scale. The core of SWA method is to adopt a special learning rate on the loss surface to visit several local minima, and converge to expected optimal solution by averaging the local minima weights. Currently, SWA supports two types of annealing learning rate strategies, including cosine and linear learning rate. In this study, a discontinuous cosine annealing learning rate strategy was adopted [50], which has been demonstrated to perform well in global optimization. Cosine annealing can urge the model to converge towards the direction of local minima after several epochs. The learning rate α of the epoch i, is calculated as follows:

$$\begin{aligned}\alpha(i) &= (1-t(i))\alpha_1 + t(i)\alpha_2, \alpha_1 \geq \alpha_2\\ t(i) &= \frac{1}{c}(\bmod(i-1,c)+1)\end{aligned} \tag{1}$$

where $\alpha(i)$ decreases from α_1 to α_2; c represents the cycle length; and i denotes epoch.

Raising the cosine annealing learning rate after several epochs, the weights of model will be stored for yielding average "SWA" weights. The model weights are denoted as $w = \{w_0, w_1, w_2, \cdots, w_n\}$, n is the total number.

The calculation formula is as follows:

$$w^{SWA} = \frac{1}{T}\sum_{t=1}^{T} w_t^{noc} \tag{2}$$

where w^{SWA} represents the average weight of the model; T denotes the number of cycles of the cosine annealing strategy; w_t^{noc} is the weight value corresponding to the minimum learning rate at the end of t-th cycle.

3) MULTI-MODEL SCALE ENSEMBLE

After the single-model ensemble is completed using the SWA, the next step is to ensemble model from a multi-model scale by DE. In this study, the multi-model scale refers to the ensemble within several models, i.e., SWA-LSTM, SWA-GRU and SWA-TCN, which is ensembled between the output of several models. The number of models is defined as N. The objective function F_{obj} is as follows:

$$\min F_{obj} = \left(\sum_{n}^{N}\left(W_n \cdot L_n^{loss}\right)\right) \tag{3}$$

where F_{obj} represents the minimum value of the right-hand side, W_n and L_n^{loss} represent the weight and loss value of the n-th model, respectively.

After the best weight W_n^{best} between models through the objective function F_{obj} is settled, the final output can be obtained:

$$Y^{final} = \sum_{n}^{N}\left(W_n^{best} \cdot Y_n\right) \tag{4}$$

where Y^{final} denotes the final output after the implementation of multi-model scale ensemble. $Y_n = \{y_0, y_1, y_2, \cdots, y_i\}$, Y_n is the predicted value of n-th model, i is the number index.

D. PERFORMANCE EVALUATION METRICS

To evaluate the model performance comprehensively, four typically used metrics for prediction accuracy are selected: root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE), coefficient of determination (R^2), maximum error (ME) and calculation time (CT) [51].

$$RMSE = \sqrt{\frac{1}{I}\sum_{i=1}^{I}\left(y_i - y_i^{obs}\right)^2} \tag{5}$$

$$MAE = \frac{1}{I}\sum_{i=1}^{I}\left|y_i - y_i^{obs}\right|. \tag{6}$$

$$\mathrm{MAPE} = \frac{\sum_{i=1}^{I}\frac{\left|y_i^{obs}-y_i\right|}{\left|y_i^{obs}\right|}}{I} \tag{7}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{I}\left(y_i^{obs}-y_i\right)^2}{\sum_{i=1}^{I}\left(y_i^{obs}-\bar{y}\right)^2} \tag{8}$$

$$ME = \max(\left|y_i^{obs} - y_i\right|) \tag{9}$$

where y_i, y_i^{obs}, and I represent the predictand, observations, and number of observations, respectively. $\bar{y}$ denotes the mean of the observations. CT denotes the calculation time of models.

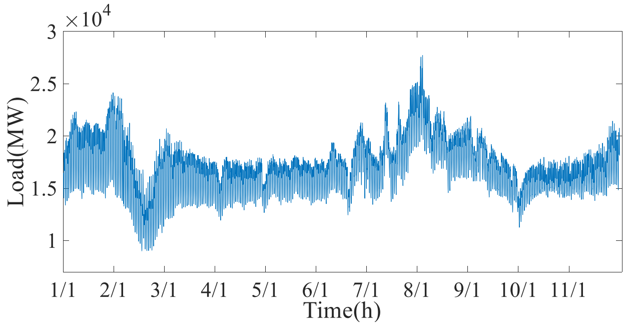

FIGURE 8. Hourly load of Hubei in 2015.

III. EXPERIMENTAL STUDY

A. STUDY AREA AND DATA DESCRIPTION

The data used in this study are the hourly loads of Hubei Province from January 1, 2015 to December 1, 2015; the data are from the Central China Power Grid. The hourly load fluctuations are shown in Fig. 8 for visual display. Moreover, the meteorological data of Hubei Province were obtained from Tianhe weather station in Wuhan.

B. DATA PREPROCESSING

1) DIVISION OF DATASETS

To increase the validity of this study, the sliding window method was adopted to divide the original data into three new datasets. The length of the sliding window accounted for 80% of the original data, and 10% of these data slid each time. Furthermore, in the sliding window, the first 70% and the next 10% data are adopted as the training set and validation set respectively, the last 20% are as the test set. A schematic of the dataset partition is presented in Fig. 9. As illustrated in the figure, the original data are classified into three datasets by the data sliding window method, namely datasets A, B, and C.

2) FEATURE FACTOR PROCESSING AND SELECTION

Load variations are closely related to meteorological conditions. This study introduces a combination of meteorological elements to increase the performance of load forecasting. The combination of meteorological factors includes average temperature, dew point, and relative humidity. In addition, the time feature was strongly correlated with load changes. To reflect the influence of time feature on load, weekdays were mapped as 1, and weekends were mapped as 0.8 in this study. In this way, all times are converted into corresponding time features, for example, the features of one full week are [1, 1, 1, 1, 1, 0.8, 0.8] and the features are the same within a day.

3) NORMALIZED PROCESSING

To exclude the influence of variable dimension and variation range, the data were min-max normalized. All values are mapped to the [0,1] interval by min-max normalization. The normalized processing is applied to the training, validation and test set separately. The specific calculation formula is given by Eq. (10):

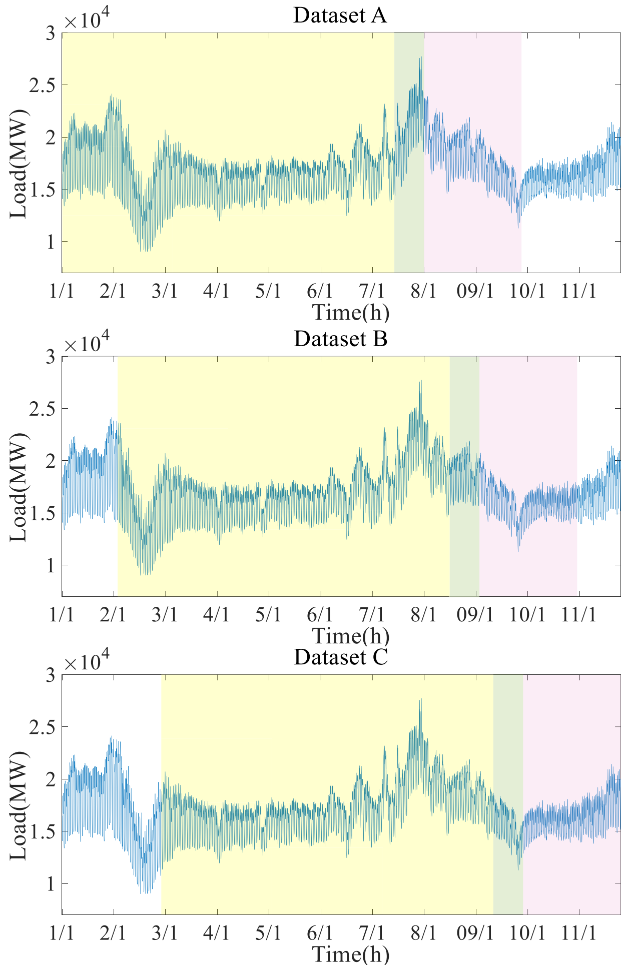

FIGURE 9. *Schematic of dataset partition (The yellow, green and pink blocks represent the training set, validation set and test set, respectively).*

$$x' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{10}$$

where x' represents the normalized values; x_i denotes the i-th real values; $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values, respectively.

C. MODEL DEVELOPMENT AND PARAMETER SETTINGS

To verify the performance of the MSENN model, eleven models of three types are selected for comparison. 1) Machine learning model, such as ridge regression (RR). 2) Deep learning networks (such as LSTM, GRU and TCN). 3) Ensemble deep learning models. The first is single-model scale ensemble of deep learning models LSTM, GRU and TCN combined with the SWA method (namely SWA-LSTM, SWA-GRU and SWA-TCN). The second is ensemble of LSTM, GRU and TCN model based on the DE algorithm, which is named weight ensemble neural network (WENN). Additionally, we have selected two state-of-the-art ensemble models. One

TABLE 1. Hyper-parameter optimization results of TWELVE MODELS.

Model	Model structure and parameters	Ensemble method	Optimizer/ Epoch
RR	Regularization strength = 1.0	Non-ensemble	——
LSTM	16 (LSTM)-16 (LSTM)-1(ANN)	Non-ensemble	Adam/200
GRU	16 (GRU)-16 (GRU)-1(ANN)	Non-ensemble	Adam/200
TCN	16 (TCN)-16 (TCN)-1(ANN)	Non-ensemble	Adam/200
SWA–LSTM	16 (LSTM)-16 (LSTM)-1(ANN)	SWA Ensemble	Adam/200
SWA–GRU	16 (GRU)-16 (GRU)-1(ANN)	SWA Ensemble	Adam/200
SWA–TCN	16 (TCN)-16 (TCN)-1(ANN)	SWA Ensemble	Adam/200
WENN	[TCN, LSTM, GRU]- Weighted ensemble based on DE	DE Ensemble	Adam/200, DE/100
CEM-1	[TCN, LSTM, GRU]- Weighted ensemble based on Grid Search Method	Grid Search Ensemble	Adam/200
CEM-2	[NN, LSTM, RF, Evtree]- XGBoost	Stacking Ensemble	——
MSENN-D	[SWA-TCN, SWA-LSTM, SWA-GRU]- Weighted ensemble based on DE	SWA Ensemble, DE Ensemble	Adam/200, DE/100
MSENN	[SWA-TCN, SWA-LSTM, SWA-GRU]- Weighted ensemble based on DE	SWA Ensemble, DE Ensemble	Adam/200, DE/100

N (Network): Neural network layer with *N* neurons.
[Models 1, 2, and 3]: Models 1, 2, and 3 are connected in parallel.

TABLE 2. Comprehensive performance results of TWELVE MODELS on dataset A.

Model	RMSE (MW)	MAE (MW)	MAPE (%)	R^2	ME (GW)	CT (s)
RR	476.901	354.332	1.89	0.9719	2.61	7.32
LSTM	415.995	296.159	1.55	0.9783	2.79	101.08
GRU	439.605	317.081	1.64	0.9758	2.82	83.18
TCN	467.826	309.778	1.57	0.9726	2.93	201.25
SWA–LSTM	377.694	270.225	1.43	0.9821	2.17	101.42
SWA–GRU	409.587	292.167	1.53	0.9790	2.28	83.33
SWA–TCN	432.124	307.671	1.59	0.9766	2.66	201.31
WENN	371.243	257.917	1.34	0.9822	2.76	310.51
CEM-1	374.980	261.085	1.36	0.9817	2.86	350.87
CEM-2	357.482	256.924	1.32	0.9837	2.24	531.72
MSENN-D	426.628	312.751	1.67	0.9772	2.81	273.41
MSENN	348.005	249.158	1.31	0.9848	2.09	311.84

is ensemble LSTM, GRU and TCN based on grid search method [52], which is named contrast ensemble model 1 (CEM-1) in this study. The other is a stacking ensemble model based on random forests (RF), LSTM, ANN and evolutionary trees (Evtree) [53], which is named contrast ensemble model 2 (CEM-2) in this study. At last, non-meteorological data (days and consumption only) was as input to MSENN to evaluate the effect of meteorological data, the reorganized model is called MSENN-D in this study.

The daily load fails to embody the accumulated influence of the past few days, so one week (168h) is taken as the input and 1h as the output in this study. The Bayesian optimization algorithm was used to optimize the hyperparameters of all models. In SWA mechanism, the bound limits of the learning rate are 0.005 and 0.0005, respectively. The total number of iterations was 200, the cosine annealing learning rate was set to be executed from 80% of the total number of generations (i.e., the 160th generation). Every cycle epoch was five generations. In DE algorithm, the scaling factor was set to 0.5, the crossover probability was set to 0.7 and the number of iterations is set to 100. The detailed parameter settings of all the models are listed in Table 1.

IV. RESULTS AND DISCUSSIONS

In this section, the predictions based on the three datasets are presented in the form of tables and figures. The average forecast results after running 10 times of the 12 models on dataset A, dataset B and dataset C are listed in Table 2, Table 3

TABLE 3. Comprehensive performance results of TWELVE MODELS on dataset B.

Model	RMSE (MW)	MAE (MW)	MAPE (%)	R^2	ME (GW)	CT (s)
RR	346.146	270.482	1.65	0.9642	1.52	7.33
LSTM	334.615	259.347	1.60	0.9666	1.63	102.71
GRU	339.496	264.469	1.63	0.9659	1.66	84.56
TCN	329.980	258.403	1.57	0.9675	1.42	203.84
SWA–LSTM	314.163	241.100	1.47	0.9709	1.41	103.01
SWA–GRU	326.744	253.437	1.55	0.9681	1.56	84.42
SWA–TCN	312.889	245.307	1.50	0.9708	1.39	203.61
WENN	314.470	248.787	1.52	0.9705	1.41	314.08
CEM-1	315.433	247.735	1.53	0.9701	1.43	533.08
CEM-2	284.084	221.862	1.35	0.9758	1.36	356.71
MSENN-D	346.050	275.155	1.62	0.9648	1.63	274.60
MSENN	292.510	226.876	1.38	0.9745	1.33	313.98

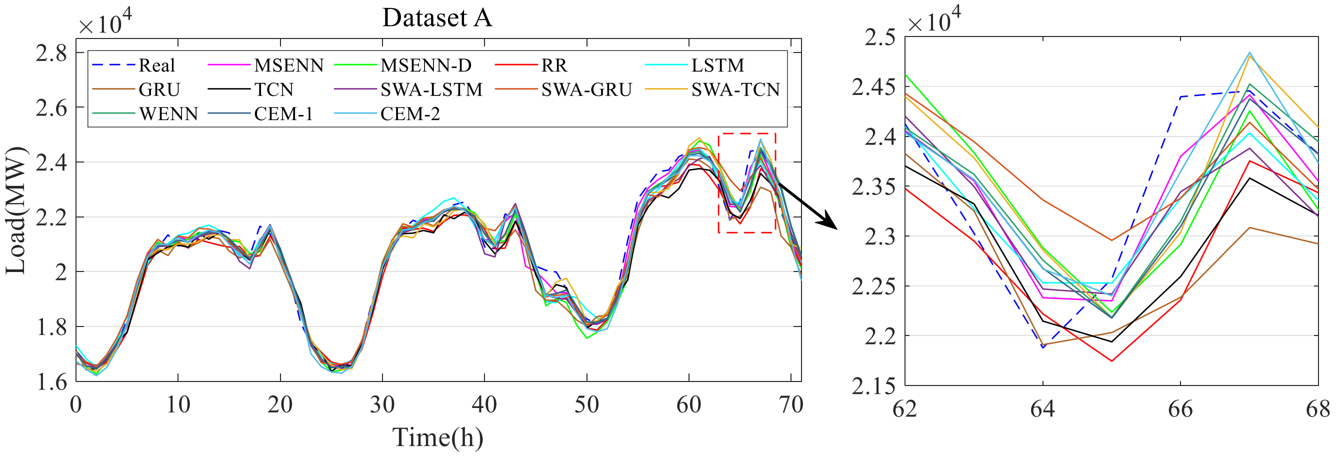

FIGURE 10. Comparison of multi-model prediction performance for three-day on dataset A.

and Table 4. Where, dark gray and light gray represent the best and sub-best results among these models.

In general, in Table 2, Table 3 and Table 4, the proposed MSENN has the best performance on the dataset A and dataset C, and CEM-2 has the best performance on the dataset B. In contrast, the RR model has the worst performance on all three datasets. The above results preliminarily demonstrate the superiority of the ensemble deep learning model. To visually demonstrate the forecasting effect of 12 models, the predicted results of 12 models for a time horizon of 72h (three-day) forecasting are presented in Fig. 10, Fig. 11 and Fig. 12. As can be seen from the figures, MSENN effectively tracks multiple inflection points of true values on dataset A. Almost all models have captured the fluctuation trend of the real load, but the predicted value of MSENN is closer to the real load as a whole. Furthermore, Fig. 13 presents predicted and observed hourly load scatter plots on three datasets of seven representative models, including MSENN, WENN, RR, LSTM, SWA-LSTM, CEM-1, CEM-2.

It is clearly visible that the results of the MSENN exhibit the closest fit with the ideal line on the dataset A and dataset C. These results all illustrate that the MSENN has strong nonlinear fitting ability and excellent predictive performance. Overall, the above results preliminarily demonstrate the superiority of the proposed model in dealing with STLF problems.

Note that all models, both for dataset A, dataset B and dataset C, yielded almost identical results. Therefore, we take dataset A as an example to carry out further detailed analysis. Start by comparing the performance of non-ensemble model. It can be observed from Table 2 that the machine learning model RR performs the best in CT, but performs the worst in other accuracy indicators. Compared with RR, the RMSE of deep learning models, LSTM, GRU and TCN decreased by 12.77%, 7.82% and 1.90%, ME increased by 12.77%,

TABLE 4. Comprehensive performance results of TWELVE MODELS on dataset C.

Model	RMSE (MW)	MAE (MW)	MAPE (%)	R^2	ME (GW)	CT (s)
RR	344.508	275.487	1.66	0.9602	1.33	7.32
LSTM	338.069	263.842	1.63	0.9662	1.75	101.25
GRU	332.426	265.893	1.62	0.9638	1.44	83.87
TCN	340.772	261.838	1.61	0.9653	1.49	200.77
SWA–LSTM	313.164	239.934	1.45	0.9679	1.48	101.13
SWA–GRU	315.988	246.141	1.48	0.9666	1.25	83.08
SWA–TCN	307.556	237.624	1.42	0.9683	1.14	201.74
WENN	292.560	223.698	1.33	0.9713	1.23	309.73
CEM-1	294.638	221.704	1.32	0.9710	1.23	349.99
CEM-2	288.971	219.742	1.29	0.9729	1.10	530.92
MSENN-D	341.488	259.095	1.59	0.9648	1.44	273.23
MSENN	281.220	212.980	1.27	0.9735	1.11	311.88

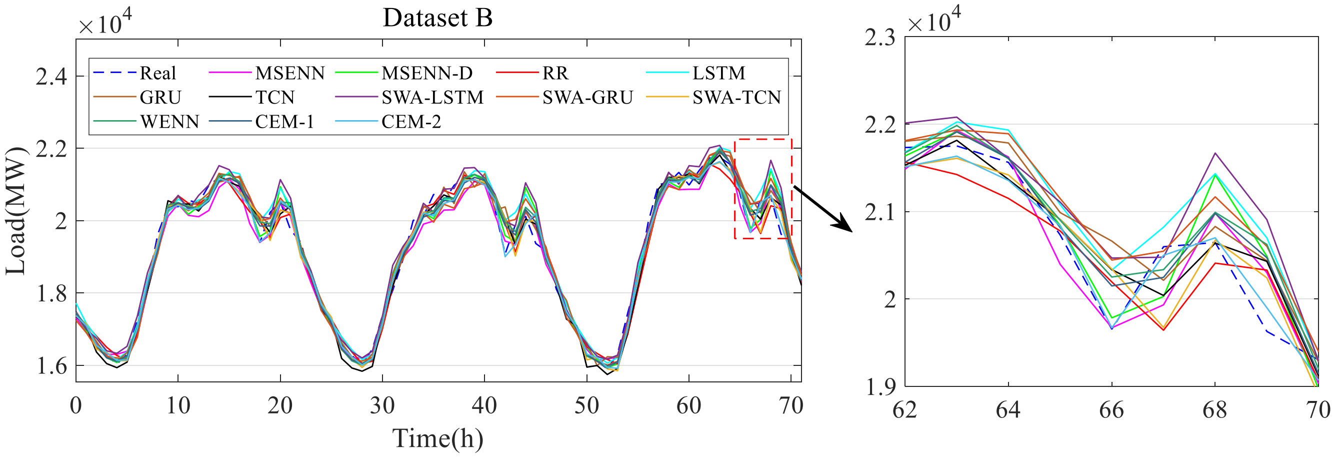

FIGURE 11. Comparison of multi-model prediction performance for three-day on dataset B.

7.82% and 1.90%, respectively. These data shows that LSTM, GRU and TCN have higher load prediction accuracy than RR model in this study.

Secondly, the effect of the SWA ensemble method on deep learning models is analyzed. It can be clearly seen from the table that the model with SWA has higher accuracy than its original model. Compared with TCN, LSTM and GRU, the RMSE of the single-model scale ensemble models (i.e., SWA–TCN, SWA–LSTM and SWA–GRU) decreased by 7.63%, 9.21% and 6.83%, respectively. ME decreased by 22.22%, 19.15% and 9.22%. R^2 increased by 0.39%, 0.33% and 0.41%. Moreover, the calculation time of the model with SWA is almost the same as its original model. This result shows that SWA mechanism can improve the prediction accuracy of a single model without adding additional calculation time. Furthermore, to visualize the effect of SWA on model ensemble, detailed loss curve is presented in Fig. 14. Evidently, the loss curve rises and falls periodically starting from the 160th epoch. This phenomenon is due to the special learning rate mechanism of SWA, which can make a single model converge to multiple local optima in one training session. This trait effectively expands the insight of a single model on the search area, thereby enhancing the generalization ability of the model. Moreover, the models with SWA method have lower loss curve, indicating that the SWA method is able to enhance the searching ability of the model. In summary, SWA technology can effectively promoting the forecast performance from the single-model scale.

Thirdly, the effect of the DE ensemble method on the deep learning models is elaborated. It can be clearly seen from the table 2 that the model with DE ensemble method has higher accuracy than its sub-models. Compared with LSTM, GRU and TCN, the RMSE values of WENN decreased by 10.76%, 15.55% and 20.65%, respectively. Similarly, compared with SWA–LSTM, SWA–GRU and SWA–TCN, the RMSE of MSENN significantly decreased by 7.86%, 15.04%

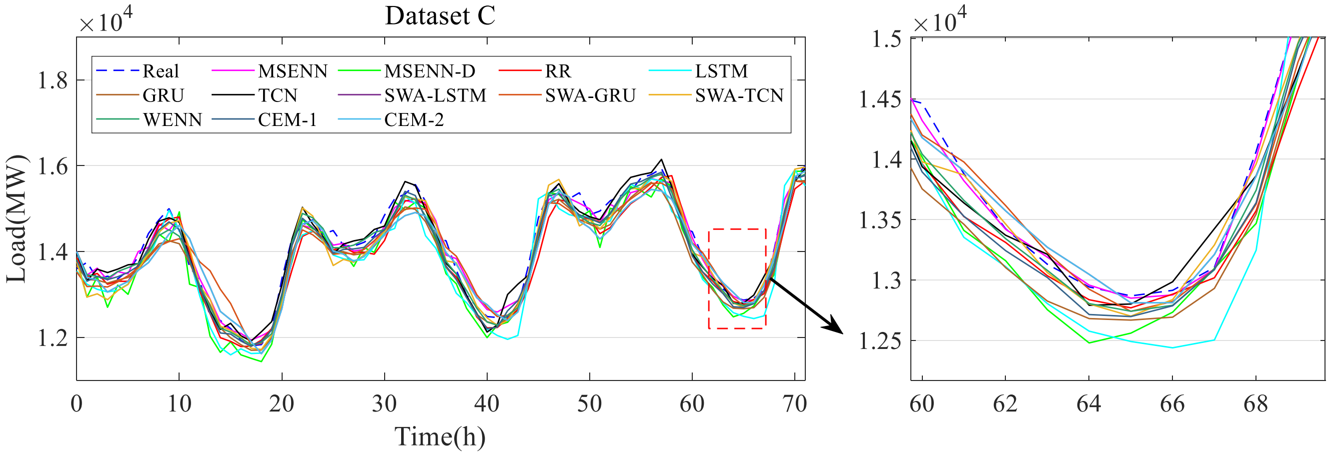

FIGURE 12. Comparison of multi-model prediction performance for three-day on dataset C.

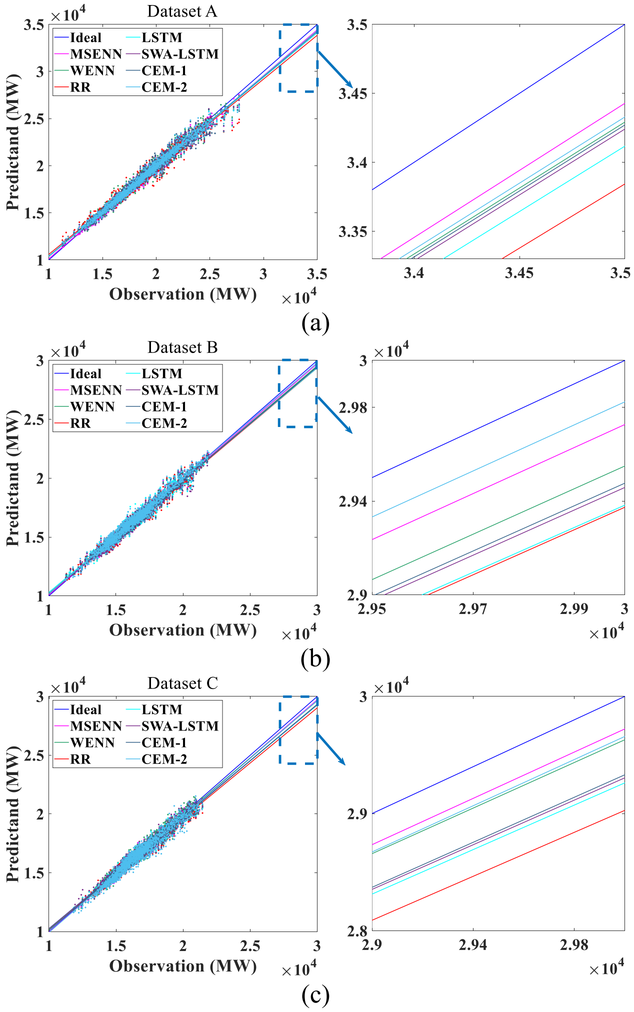

FIGURE 13. Load predictand by seven models against observation:(a) dataset A; (b) dataset B; (c) dataset C.

and 19.47% separately. Fig. 15 presents the intuitive results for weights allocation. As presented in the figure, the allocated weights of the sub-models are negatively correlated with their RMSE scores; that is, sub-models with low RMSE scores tended to be assigned higher weights. The data analysis shows that the DE ensemble method is extremely scientific and rigorous in terms of weight distribution. It focuses on the strengths and avoids the weaknesses of each sub-model. This gives full play to the advantages of each sub-model, thereby obtaining better prediction results. Additionally, the above results also reveal that the simultaneous use of SWA and DE does not conflict, which verifies the rationality of the MSEM proposed in this paper as well.

Fourthly, a comparison with state-of-the-art ensemble deep learning models is demonstrated. In general, MSENN has a more outstanding comprehensive performance than CEM-1 and CEM-2. In more detail, although CEM-1 and MSENN have the same sub-model, MSENN has a 26.92% and 11.12% improvement over CEM-1 in terms of ME and CT, respectively. This phenomenon can be explained from two aspects. On the one hand, MSENN adopts the SWA ensemble mechanism, which improves the prediction accuracy of the sub-model from the single-model scale, so MSENN has better prediction performance. On the other hand, CEM-1 adopts an ensemble strategy based on grid search method instead of DE ensemble. The grid search method is essentially an exhaustive search method, which needs to traverse all possible weight combination results, resulting in poor computational efficiency. So, the computational efficiency of CEM-1 is worse than that of MSENN. As for CEM-2, its prediction accuracy is slightly lower than that of MSENN, but it is far behind MSENN in terms of computational efficiency. Compared with CEM-2, MSENN has an improvement of 6.70% and 41.35% on ME and CT, respectively. This result shows that stacking ensemble is an effective ensemble strategy to reduce prediction error, but the complicated implementation and low computational efficiency make it not competitive.

Fifthly, the effect of meteorological factors on the model is discussed. Compared with MSENN-D, the RMSE and ME of MSENN decreased by 18.43% and 25.62%. The experimental results found that the model using only date and

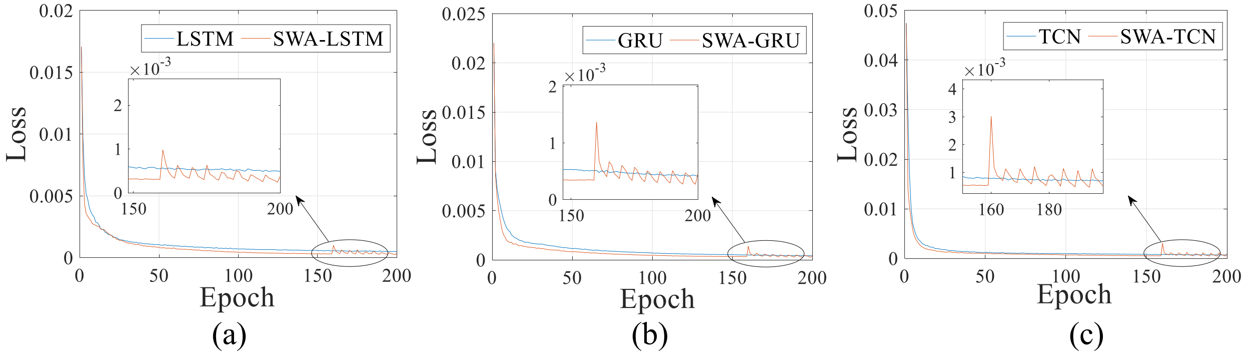

FIGURE 14. Comparison loss function curves among representative models: (a) comparison between LSTM and SWA-LSTM; (b) comparison between GRU and SWA-GRU; (c) comparison between TCN and SWA-TCN.

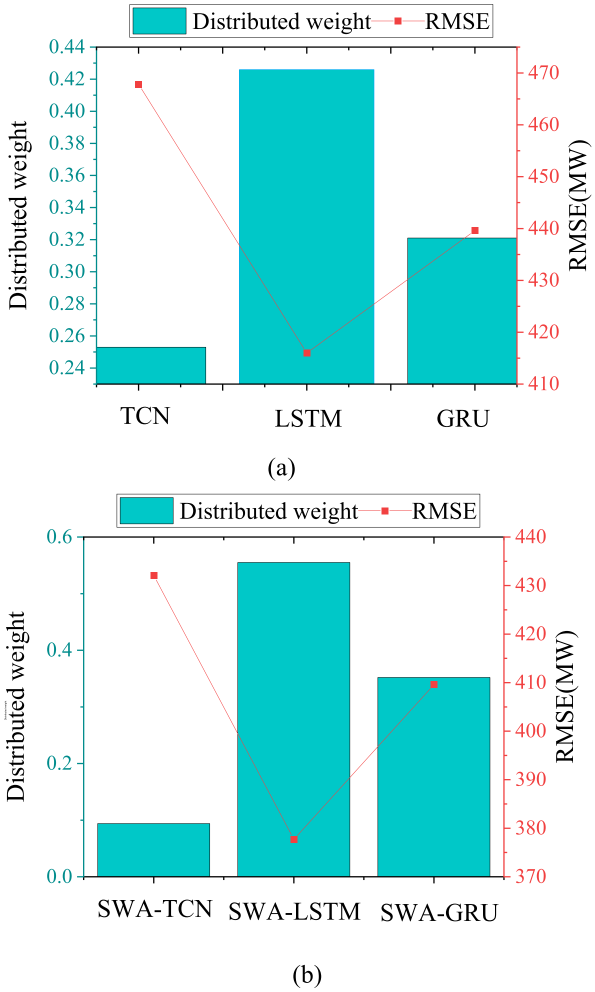

FIGURE 15. Visualization results of DE weights allocation. (a) WENN; (b) MSENN.

consumption is less effective, which indicates that meteorological data can improve the accuracy of load forecasting to a certain extent.

In summary, the above analysis on the one hand illustrates the effectiveness of the MSEM, and on the other hand also verifies the superiority of the proposed MSENN.

V. CONCLUSION

This study proposes a novel deep learning ensemble technique, called multi-scale ensemble method MSEM, to enhance the forecast performance of deep learning models in STLF. By coupling SWA and DE ensemble methods, the method enhances the generalization ability of forecasting model from single-model scale and multi-model scale, respectively. Furthermore, performing the proposed MSEM to three typical deep learning models, the paper proposes a novel multi-scale ensemble neural network (MSENN) to deal with STLF problems. Finally, by considering the hourly load of Hubei Province in China in 2015 for instance, eleven models covering machine learning, deep learning, and ensemble deep learning models (RR, LSTM, GRU, TCN, SWA–LSTM, SWA–GRU, SWA–TCN, WENN, CEM-1, CEM-2 and MSENN-D) and three datasets are employed for comparison. Simulations show that the proposed MSENN has the best comprehensive prediction performance, which also verifies the effectiveness of the MSEM. Moreover, the following conclusions are found: 1) Meteorological factors have a certain positive effect on the accuracy of STLF. 2) Model ensembled from any scale can improve the model prediction performance. 3) Simultaneous model ensemble from multiple scales does not conflict; rather, it can promote the overall performance of the model in various ways.

Notably, STLF is also influenced by other factors, such as human behavior. In future research, we will consider more influencing factors and adopt different methods or networks to enable the forecast accuracy. Moreover, it may be worth investigating to implement this technique to the power generation photovoltaic power and wind power.

ACKNOWLEDGMENT

The authors express their gratitude to the editors and reviewers for their constructive comments.

REFERENCES

[1] G. Hafeez, K. S. Alimgeer, and I. Khan, "Electric load forecasting based on deep learning and optimized by heuristic algorithm in smart grid," *Appl. Energy*, vol. 269, Jul. 2020, Art. no. 114915.

[2] G. Wen, G. Hu, J. Hu, X. Shi, and G. Chen, "Frequency regulation of source-grid-load systems: A compound control strategy," *IEEE Trans. Ind. Informat.*, vol. 12, no. 1, pp. 69–78, Feb. 2016.

[3] G. Gross and F. D. Galiana, "Short-term load forecasting," *Proc. IEEE*, vol. 75, no. 12, pp. 1558–1573, Dec. 1987.

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

[5] M. Zulfiqar, M. Kamran, M. B. Rasheed, T. Alquthami, and A. H. Milyani, "A hybrid framework for short term load forecasting with a navel feature engineering and adaptive grasshopper optimization in smart grid," *Appl. Energy*, vol. 338, May 2023, Art. no. 120829.

[6] A. Tascikaraoglu, B. M. Sanandaji, K. Poolla, and P. Varaiya, "Exploiting sparsity of interconnections in spatio-temporal wind speed forecasting using wavelet transform," *Appl. Energy*, vol. 165, pp. 735–747, Mar. 2016.

[7] K.-B. Song, Y.-S. Baek, D. Hun Hong, and G. Jang, "Short-term load forecasting for the holidays using fuzzy linear regression method," *IEEE Trans. Power Syst.*, vol. 20, no. 1, pp. 96–101, Feb. 2005.

[8] W. R. Christiaanse, "Short-term load forecasting using general exponential smoothing," *IEEE Trans. Power App. Syst.*, vol. PAS-90, no. 2, pp. 900–911, Mar. 1971.

[9] C.-M. Lee and C.-N. Ko, "Short-term load forecasting using lifting scheme and ARIMA models," *Exp. Syst. Appl.*, vol. 38, no. 5, pp. 5902–5911, May 2011.

[10] Q. Wang, S. Li, and Z. Pisarenko, "Modeling carbon emission trajectory of China, U.S. and India," *J. Cleaner Prod.*, vol. 258, Jun. 2020, Art. no. 120723.

[11] G. Juberias, R. Yunta, J. Garcia Moreno, and C. Mendivil, "A new ARIMA model for hourly load forecasting," in *Proc. IEEE Transmiss. Distribution Conf.*, 1999, pp. 314–319.

[12] C. Tian, J. Ma, C. Zhang, and P. Zhan, "A deep neural network model for short-term load forecast based on long short-term memory network and convolutional neural network," *Energies*, vol. 11, no. 12, p. 3493, Dec. 2018.

[13] R. W. Ng, K. M. Begam, R. K. Rajkumar, Y. W. Wong, and L. W. Chong, "An improved self-organizing incremental neural network model for short-term time-series load prediction," *Appl. Energy*, vol. 292, Jun. 2021, Art. no. 116912.

[14] P. Matrenin, M. Safaraliev, S. Dmitriev, S. Kokin, A. Ghulomzoda, and S. Mitrofanov, "Medium-term load forecasting in isolated power systems based on ensemble machine learning models," *Energy Rep.*, vol. 8, pp. 612–618, Apr. 2022.

[15] M. Barman and N. B. Dev Choudhury, "Season specific approach for short-term load forecasting based on hybrid FA-SVM and similarity concept," *Energy*, vol. 174, pp. 886–896, May 2019.

[16] A. G. Abdullah, W. Sopian, W. Arasid, A. Nandiyanto, A. Danuwijaya, and C. Abdullah, "Short-term peak load forecasting using PSO-ANN methods: The case of Indonesia," *J. Eng. Sci. Technol.*, vol. 13, no. 8, pp. 2395–2404, 2018.

[17] R.-J. Park, K.-B. Song, and B.-S. Kwon, "Short-term load forecasting algorithm using a similar day selection method based on reinforcement learning," *Energies*, vol. 13, no. 10, p. 2640, May 2020.

[18] A. Zeng, S. Liu, and Y. Yu, "Comparative study of data driven methods in building electricity use prediction," *Energy Buildings*, vol. 194, pp. 289–300, Jul. 2019.

[19] S. Bouktif, A. Fiaz, A. Ouni, and M. Serhani, "Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches," *Energies*, vol. 11, no. 7, p. 1636, Jun. 2018.

[20] R. D. Rathor and A. Bharagava, "Short term load forecasting of a region of India using generalized regression neural network," *Global J. Res. Eng.*, vol. 17, no. 7, pp. 1–9, 2017.

[21] Y. Chen, P. B. Luh, C. Guan, Y. Zhao, L. D. Michel, M. A. Coolbeth, P. B. Friedland, and S. J. Rourke, "Short-term load forecasting: Similar day-based wavelet neural networks," *IEEE Trans. Power Syst.*, vol. 25, no. 1, pp. 322–330, Feb. 2010.

[22] L. Ekonomou, C. Christodoulou, and V. Mladenov, "A short-term load forecasting method using artificial neural networks and wavelet analysis," *Int. J. Power Syst.*, vol. 1, pp. 64–68, 2016.

[23] C. Fan, F. Xiao, and Y. Zhao, "A short-term building cooling load prediction method using deep learning algorithms," *Appl. Energy*, vol. 195, pp. 222–233, Jun. 2017.

[24] H. Shi, M. Xu, Q. Ma, C. Zhang, R. Li, and F. Li, "A whole system assessment of novel deep learning approach on short-term load forecasting," *Energy Proc.*, vol. 142, pp. 2791–2796, Dec. 2017.

[25] S. Muzaffar and A. Afshari, "Short-term load forecasts using LSTM networks," *Energy Proc.*, vol. 158, pp. 2922–2927, Feb. 2019.

[26] K. Ke, S. Hongbin, Z. Chengkang, and C. Brown, "Short-term electrical load forecasting method based on stacked auto-encoding and GRU neural network," *Evol. Intell.*, vol. 12, no. 3, pp. 385–394, Sep. 2019.

[27] S. Bai, J. Zico Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.

[28] F. He, J. Zhou, Z.-K. Feng, G. Liu, and Y. Yang, "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*, vol. 237, pp. 103–116, Mar. 2019.

[29] X. Tang, H. Chen, W. Xiang, J. Yang, and M. Zou, "Short-term load forecasting using channel and temporal attention based temporal convolutional network," *Electric Power Syst. Res.*, vol. 205, Apr. 2022, Art. no. 107761.

[30] H. Nie, G. Liu, X. Liu, and Y. Wang, "Hybrid of ARIMA and SVMs for short-term load forecasting," *Energy Proc.*, vol. 16, pp. 1455–1460, 2012.

[31] S. Li, L. Goel, and P. Wang, "An ensemble approach for short-term load forecasting by extreme learning machine," *Appl. Energy*, vol. 170, pp. 22–29, May 2016.

[32] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022.

[33] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[34] G.-F. Fan, L.-Z. Zhang, M. Yu, W.-C. Hong, and S.-Q. Dong, "Applications of random forest in multivariable response surface for short-term load forecasting," *Int. J. Electr. Power Energy Syst.*, vol. 139, Jul. 2022, Art. no. 108073.

[35] R. E. Schapire, "The strength of weak learnability," *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, Jun. 1990.

[36] Y. Wang, S. Sun, X. Chen, X. Zeng, Y. Kong, J. Chen, Y. Guo, and T. Wang, "Short-term load forecasting of industrial customers based on SVMD and XGBoost," *Int. J. Electr. Power Energy Syst.*, vol. 129, Jul. 2021, Art. no. 106830.

[37] Y. Wang, J. Chen, X. Chen, X. Zeng, Y. Kong, S. Sun, Y. Guo, and Y. Liu, "Short-term load forecasting for industrial customers based on TCN-LightGBM," *IEEE Trans. Power Syst.*, vol. 36, no. 3, pp. 1984–1997, May 2021.

[38] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, Jan. 1992.

[39] Z. Tan, J. Zhang, Y. He, Y. Zhang, G. Xiong, and Y. Liu, "Short-term load forecasting based on integration of SVR and stacking," *IEEE Access*, vol. 8, pp. 227719–227728, 2020.

[40] C. S. Lai, Y. Yang, K. Pan, J. Zhang, H. Yuan, W. W. Y. Ng, Y. Gao, Z. Zhao, T. Wang, M. Shahidehpour, and L. L. Lai, "Multi-view neural network ensemble for short and mid-term load forecasting," *IEEE Trans. Power Syst.*, vol. 36, no. 4, pp. 2992–3003, Jul. 2021.

[41] D. Niu, M. Yu, L. Sun, T. Gao, and K. Wang, "Short-term multi-energy load forecasting for integrated energy systems based on CNN-BiGRU optimized by attention mechanism," *Appl. Energy*, vol. 313, May 2022, Art. no. 118801.

[42] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. Gordon Wilson, "Averaging weights leads to wider optima and better generalization," 2018, *arXiv:1803.05407*.

[43] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[44] L. An, Y. Hao, T.-C.-J. Yeh, Y. Liu, W. Liu, and B. Zhang, "Simulation of Karst spring discharge using a combination of time–frequency analysis methods and long short-term memory neural networks," *J. Hydrol.*, vol. 589, Oct. 2020, Art. no. 125320.

[45] Y.-G. Zhang, J. Tang, Z.-Y. He, J. Tan, and C. Li, "A novel displacement prediction method using gated recurrent unit model with time series analysis in the erdaohe landslide," *Natural Hazards*, vol. 105, no. 1, pp. 783–813, Jan. 2021.

[46] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[47] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 156–165.

[48] H. Shi, L. Wang, R. Scherer, M. Wozniak, P. Zhang, and W. Wei, "Short-term load forecasting based on adabelief optimized temporal convolutional network and gated recurrent unit hybrid neural network," *IEEE Access*, vol. 9, pp. 66965–66981, 2021.

[49] Bilal, M. Pant, H. Zaheer, L. Garcia-Hernandez, and A. Abraham, "Differential evolution: A review of more than two decades of research," *Eng. Appl. Artif. Intell.*, vol. 90, Apr. 2020, Art. no. 103479.

[50] G. Liu, Z. Tang, H. Qin, S. Liu, Q. Shen, Y. Qu, and J. Zhou, "Short-term runoff prediction using deep learning multi-dimensional ensemble method," *J. Hydrol.*, vol. 609, Jun. 2022, Art. no. 127762.

[51] L. Yin and J. Xie, "Multi-temporal-spatial-scale temporal convolution network for short-term load forecasting of power systems," *Appl. Energy*, vol. 283, Feb. 2021, Art. no. 116328.

[52] D. Hadjout, J. F. Torres, A. Troncoso, A. Sebaa, and F. Martínez-Álvarez, "Electricity consumption forecasting based on ensemble deep learning with application to the Algerian market," *Energy*, vol. 243, Mar. 2022, Art. no. 123060, doi: 10.1016/j.energy.2021.123060.

[53] A. S. Reddy, S. Akashdeep, and R. Harshvardhan, "Stacking deep learning and machine learning models for short-term energy consumption forecasting," *Adv. Eng. Informat.*, vol. 52, Apr. 2022, Art. no. 101542, doi: 10.1016/j.aei.2022.101542.

QIN SHEN is currently pursuing the Ph.D. degree with the School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology (HUST), Wuhan, China. Her research interests include modeling and operation theory in water resources management and power system optimal dispatching.

LI MO received the B.S. and Ph.D. degrees from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2006 and 2011, respectively. She is currently an Associate Professor with the School of Civil and Hydraulic Engineering, HUST. Her research interest includes optimal operation and control of hydropower energy systems.

GUANJUN LIU received the master's degree from the Huazhong University of Science and Technology (HUST), Wuhan, Hubei, China, in 2020, where he is currently pursuing the Ph.D. degree. His research interest includes optimal operation and control of hydropower energy systems.

JIANZHONG ZHOU received the B.S. degree in automatic control from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1982. His research interest includes optimal operation and control of hydropower energy systems.

YONGCHUAN ZHANG was an Academician with the Chinese Academy of Engineering and a Professor with the School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology. His major research interest includes digital cascaded hydro-electric systems and its optimal dispatch.

PINAN REN received the master's degree from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2015, where he is currently pursuing the Ph.D. degree with the School of Civil and Hydraulic Engineering. His research interests include reservoir (group) optimal dispatch and power system optimal dispatch.