Deep autoencoder with localized stochastic sensitivity for short-term load forecasting

Ting Wang [a], Chun Sing Lai [b,c], Wing W. Y. Ng [a], Keda Pan [c], Mingyang Zhang [a], Alfredo Vaccaro [d], Loi Lei Lai [c]

[a] Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information, School of Computer Science and Engineering, South China University of Technology, Guangzhou 510630, China
[b] Brunel Interdisciplinary Power Systems Research Centre, Brunel University London, London, UB8 3PH, UK
[c] Department of Electrical Engineering, School of Automation, Guangdong University of Technology, Guangzhou 510006, China
[d] Engineering Department, University of Sannio, Benevento 82100, Italy
* Corresponding authors: l.l.lai@ieee.org

***Abstract*—This paper presents a short-term electric load forecasting model based on deep autoencoder with localized stochastic sensitivity (D-LiSSA). D-LiSSA can learn informative hidden representations from unseen samples by minimizing the perturbed error (including the training error and stochastic sensitivity) from historical load data. Specifically, this general deep autoencoder network as a deep learning model improves prediction accuracy and reliability. Moreover, a nonlinear fully connected feedforward neural network as a regression layer is applied to forecast the short-term load, with the generalization capability of the proposed model using hidden representations learned by D-LiSSA. The performance of D-LiSSA is evaluated using real-world public electric load markets of France (FR), Germany (GR), Romania (RO), and Spain (ES) from ENTSO-E. Extensive experimental results and comparisons with the classical and state-of-the-art models show that D-LiSSA yields accurate load forecasting results and achieves desired reliable capability. For instance, with the French case, D-LiSSA yields the lowest mean absolute error, mean absolute percentage error, root mean squared error; providing up to 61.89%, 63.20%, and 56.40% forecasting accuracy improvements as compared to the benchmark model for forecasting hourly horizon, respectively.**

***Index Terms*—Short-term load forecasting, deep autoencoder, deep learning, stochastic sensitivity**

1. INTRODUCTION

With the rapid development of social economy and increasing of global warming [1][2], the electric load demand shows a trend of increasing year by year in industry, commercial activities, offices, communication, and transportation sectors [3]. Whether an electric power system can ensure a continuous supply of electricity without a blackout for these events is vitally important. Therefore, an accurate and stable load forecasting model has the potential to avoid blackout incidents in the city's operations [4]. Moreover, the accuracy of electric load forecasting accounts for the financial performance of electric utility companies, as well as the resilience of power grids [5]. Electric load forecasting has been widely researched in the past several years. However, there are prediction challenges due to the variability of the load as a consequence of seasonality and holidays. Other factors such as weather, population characteristics, electricity prices, geographical conditions, the natural environment also complicate forecasting task because of their non-linear relationships. Nowadays, short-term load forecasting (STLF) mainly focuses on predicting the loads in the next few minutes to a week-ahead [6]. Considering the significance of load forecasting, various forecasting methods have been explored and generally classified as: 1) classical statistical methods, 2) artificial intelligence methods, and 3) hybrid forecasting methods [7].

Classical statistical methods based on mathematical statistics approximate the relationship between explicable variables, such as online measured data and future load values. Usually, these approaches treat historical data as an input to make the short-term forecasting. Various conventional statistical methods are accessible, including autoregressive (AR) [8], auto regressive moving average (ARMA) [9], autoregressive integrated moving average (ARIMA) [10] [11], the regression analysis methods [12][13]. The nonlinear autoregressive model (NARM) [14] has been successfully used for the time series prediction patterns, where it works as a time series algorithm which is a kind of progressive neural interfaces. The random forest and nonlinear autoregressive approach [15] is introduced to forecasting electric load for utility energy management systems. Based on actual environmental and energy consumption data, this model uses NARM, stepwise regression, and least square boosting to estimate the energy. However, these methods are based on the linear analysis and are not suitable for the non-linear load series forecasting [16].

In the past decades, artificial intelligence methods including artificial neural network (ANN) [17][18][19], support vector regression (SVR) [20], fuzzy logic method [21], have been applied in electric load forecasting. Artificial intelligence methods utilize the historical load data to complete model training and can nonlinearly map the inputs to the target set. ANN in deep learning can comprehensively consider various factors to improve forecasting results for developing STLF models. ANN has gained recognition in smart grids, with model varieties including radial basis function (RBF) neural networks [22], deep belief network (DBN) [23], causal Markov Elman network (CMEN) [24], and long-short term memory (LSTM) [25]. Restricted Boltzmann machines (RBM) as a classical neural network has been applied to load forecasting [26]. Researchers have been using deep residual networks (ResNet) [27] for STLF. To improve the robustness of prediction model in commercial buildings with deep learning, the authors in [19] propose a recurrent neural network (RNN) and a convolutional neural network (CNN) for building-level day-ahead multi-step load forecasting model. Reference [22] presents a review of some earlier known ANN approaches for load forecasting and designs an algorithm using typical radial basis function (RBF) networks for a 24-hour electric load forecasting. With the increasing complexity of forecasting environment, traditional forecasting methods difficult to meet management's need for forecasting accuracy. An improved deep belief network [23] is used to solve the short-term load forecasting which considers input data, model, and performance in demand-side management. Load forecasting is becoming increasingly complex, and uncertain, a data-driven deep learning framework based on deep belief network (DBN) method [28] has been proposed to forecast the hourly load of the power system. Causal Markov Elman network (CMEN) characterizes the various interdependence among heterogeneous time series for load forecasting in multi-network systems [23]. This approach analyzes the joint information between electricity and transportation networks. A novel two-layer architecture ensemble neural network framework is developed, called enhanced ELITE (E-ELITE) for STLF [29]. The neural network framework of the E-ELITE is designed based on each neural network forecaster with different optimal weights and structures. The memristor-based echo state network (MESN) adopts Newman and Watts small-world network and uses the online least mean square (LMS) algorithm to train the output weights [30] for STLF. Due to the high uncertainty, residents' activities and volatility, the individual residential short-term load forecasting is facing a serious challenging. Long short-term memory (LSTM) and recurrent neural network (RNN) are the most popular techniques in deep learning. A LSTM-based framework [25] is proposed to address such short-term residential load forecasting problem. It employs a clustering technique for density estimation to evaluate the inconsistency of the residential load distribution. The paper presents an extended deep residual networks (ResNet) model [27] for STLF. This model utilizes domain knowledge and adopts the ensemble strategy by combining multiple individual networks. However, recent research has showed that some representative load forecasting models like SVR and ANN easily fail under data integrity attacks. To address this challenge, two variants of the re-weighted least squares regression models and a L_1-normregression model are proposed to enhance the robustness of load forecasting models [6]. Although artificial intelligence methods have better predict ability, but they need the optimal parameters when establishing them so that the process of optimizing is time-consuming or easily over-fitting on the training set.

Most of the above-mentioned works focus only on either reducing load forecasting error or improving prediction accuracy and stability for load forecasting. However, load patterns have complex behavior, which brings challenges to optimize both independent objectives simultaneously within the same period. Also, according to the former literatures, each forecasting method has its strength or weakness, and they cannot always satisfy all requirements of forecasting accuracy and stability. Thus, hybrid methods combine the strengths of different methods to improve the traditional methods. In a hybrid model for load forecasting, it usually has original data pre-processing, forecasting, and optimization phases [7][31][32][33]. For original pre-processing phase, data decomposition algorithms such as empirical mode decomposition (EMD) [34] are usually used to decompose original load sequence to several subsets. The forecasting methods in this framework are the ANN or SVR models. Some hybrid methods [35][36][37] consider short-term load forecasting with stochasticity by combining various regression and ANN models such as ARIMA, bi-square kernel (BSK) regression, wavelet neural network (WNN), and RNNs to improve the forecasting accuracy. To obtain high accuracy and stability in load forecasting simultaneously, multiple objective optimization algorithms are applied to load forecasting so that the model guarantees accuracy and stability at the same time. Optimization is a key technique in dealing with renewable energy forecasting fields. An accurate (i.e. small error) and reliable (i.e. consistently small error) multi-objective method [38] is presented for daily STLF in European countries. In this model, several parameters are optimally tuned according to a multi-objective strategy that minimized both the prediction error and the variance of the error. In general, load variations may impact the control parameters in power systems. Load-Oriented control parameters are optimized using adaptive particle swarm optimization strategy [39] based on ANNs to forecast loads on a day-ahead time horizon. The optimization process of static synchronous compensator using particle swarm optimization algorithm can effectively mitigate the low-frequency oscillation damping (LFOD) of the power system and improve the robustness of the power system under external disturbances. For load-serving entities, an accurate load forecasting generally requires high computational cost and is a tradeoff to determine an accurate cost for power purchase. Beneficial correlated regularization (BCR) term for day-ahead load forecasting based on a neural network (NN) is presented [40]. This work includes studies for both accuracies on load forecasting and cost-benefit for electricity in the training of NN. Although the accurate short-term load and price forecasting are important for obtaining maximum profits in competitive electricity markets, however, most of the existing literature in short-term load or price forecasting focus on the prediction and lack of simultaneously considering the nonlinearities and interacting features in the forecast processes. A novel ensemble framework [41] is proposed to treat each NN as the individual predictor, including Elman neural network (ELM), feedforward neural network (FNN), and radial basis function neural network (RBFNN) to improve the accuracy load forecasting. The three predictors are trained by global particle swarm optimization (GPSO) and then used a trim aggregation step to combine the outputs of individual predictors. In [42], an ensemble of RBFNN is trained by minimizing the localized generalization error (LGE) for short-term and mid-term load forecasting. This ensemble method uses weighted fusion technique to enhance the generalization capability of the model. An ensemble approach based on information theory and causality [43] merges with an individual network, which can simultaneously characterize the interrelationship between electricity and traffic network patterns for short-term load forecasting. A hybrid neural network forecasting model based on deep belief network (DBN) and bidirectional recurrent neural network (Bi-RNN) is proposed [44]. The method adopts unsupervised pre-training and supervised adjustment training methods which is verified on two different datasets. A hybrid model [45] is proposed by combining multiple LSTM and back-propagation neural network (BPNN) for hour-ahead load forecasting on a building-level. To capture the nonlinear and complex pattern in yearly peak load, a hybrid long term forecasting method based on data mining technique and time series is proposed

[41]. In the model, the SVR as a forecasting algorithm and the parameters of the SVR are optimized using a particle swarm optimization (PSO) method. A novel hybrid electric load forecasting model based on modified mutual information and restricted Boltzmann machine is developed for the decision making of a smart grid [46]. The hybrid ensemble deep learning (HEDL) [47] approach uses deep belief network (DBN) for deterministic and probabilistic low-voltage load forecasting. Actually, if a forecasting model can solve both independent accuracy and stability at the same time, which increases the complex of the load pattern. So that most of the previous research focused only on either increasing load forecast accuracy or enhancing the stability, very few studies focused on these two issues simultaneously. Thus, a hybrid model [48] achieves two objectives simultaneously by combing ANN and multi-objective optimization algorithm (MOFTL). The MOFTL is based on follow the leader algorithm. However, multi-objective optimization strategies require a lot of computing resources and time consuming in practical applications due to generating a series of optimal pareto-optimal solutions in each iteration.

To guarantee the reliability and economic benefits of power grid, stability also plays a vital role in electric load or price forecasting models. The most above aforementioned literatures only consider either the accuracy or stability, with difficulty achieving the accuracy and stability simultaneously except for some hybrid methods. However, social and external natural factors like seasonality, weather, electricity prices, geographical conditions, industrial manufacture, and human activities can influence the susceptibility of model to STLF, which makes load forecasting more difficult. The recent work based on autoencoder [49] mainly considers the input with small perturbations that effectively improve the performance of the model. But it is an unsupervised learning method with a single hidden layer for image classification which limits its extraction of informative learned features. Generally, with more hidden layers, a deep neural network will produce a better performance [50]. The above various factors motivate us to build an ANN model for STLF by adopting a deep localized stochastic sensitivity autoencoder (D-LiSSA): to reduce load forecasting error and produce reliable prediction on real world load data in European countries. The following summarizes the main contributions of this paper:

1) The social and external natural factors such as calendar, holidays, electricity prices, weather, and seasonality have a significant impact on future electric load forecasting. The nonlinear, nonstationary, and variable behaviors of these factors pose a big challenge to load forecasting model. We consider minor disturbances in these factors by theoretical model analysis. The proposed predictive model promoting the forecast accuracy and stability can potentially contribute to power system operation and management. Thus, an accurate and stable forecasting model is key aspect for ensuring maximum benefit in the grid market.

2) Modeling the deep neural network structure using a stacked autoencoders with stochastic sensitivity to extract informative hidden representations. It is a supervised learning model which considers unseen samples in a Q-neighborhood surrounding historical training samples. This is the key strategy that enhances the informative of learned features and making effective predictions for STLF.

3)Specifically, the model trained by the minimization of perturbation error (PE) is insusceptible to small perturbations of inputs, and the generalization ability of D-LiSSA is enhanced. The PE represents the sensitivity of the model to unseen samples that are similar to training samples so that the model is still sensitive to large perturbations. Thus, the proposed model attains a high accuracy in STLF. Additionally, D-LiSSA can as a general framework, be applied to other energy forecasting tasks such as wind speed forecasting and solar irradiance forecasting. These various applications demonstrate D-LiSSA has a good generalization ability in the energy forecasting.

4) Furthermore, the proposed forecasting method has been evaluated based on well-known and reliable electricity market from the ENTSO-E [51] dataset. The proposed method yields state-of-the-art performance compared with other five forecasting models including ARIMA [11], LSTM [25], ResNet [27], DBN [28], and binary decision tree (BDT) [52] on four real-world electricity markets of France (FR), Germany (GR), Romania (RO), and Spain (ES) in Europe.

The remainder of the paper is organized as follows. Section 2 formulates the proposed model. The results and discussions of STLF by the proposed model are presented in Section 3. Conclusions and future work for this paper are provided in Section 4.

2. D-LISSA BASED-STLF

2.1. Model Input Data

The time series data $V \in \mathbb{R}^{T\times 2}$ is a matrix that includes load and temperature characteristics across T time stamps, where v_t is the measurement of the load and temperature recorded at the t time stamp. To fully explore the temporal characteristics of the data, we resample the whole time series into a series of samples using sliding windows according to a time window size w. The sliding window resamples the load and temperature time-series data with a sliding the step size of 1 until finishing the whole time-series data. Therefore, these

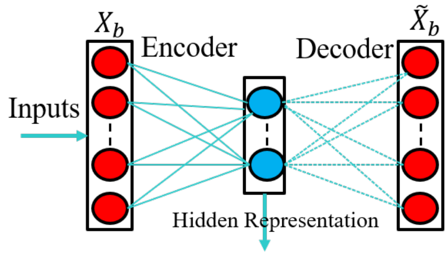


Fig. 1. Illustration of AE training.

windows have different data from time segment. The $b^{th}(b = 1,2,\cdots, M)$ segment being denoted as $X_b = (v_b, v_{b+1}, \dots, v_{b+w-1}) \in R^{1\times n}$, where M denotes the total number of segments and $n = w \times 2$. The subset of a timescale is then reformulated as $D = [X_1; X_2; \cdots; X_M]$. The samples are fed into D-LiSSA as the inputs to make a predicted value for a further load.

2.2. Deep Localized Stochastic Sensitivity Autoencoder (D-LiSSA)

Autoencoder (AE), as an ANN aims to find a set of optimal connection weights by minimizing the reconstruct error between original inputs and outputs of AE. For AE training problem, a training dataset D with M samples $\{X_b\}$ is given from the problem domain where X_b denotes the n-dimensional input vector of the b^{th} training sample. Generally, an AE consists of an input layer, an encoding layer, and a decoding layer as shown in Fig. 1. The encoding layer first maps X_b onto a hidden representation H through a deterministic mapping as in Eq. (1). Then, decoding layer maps H onto a reconstruction $\tilde{X}_b$ as in Eq. (2).

$$H = f(WX_b + b_1) \tag{1}$$

$$\tilde{X}_b = \rho(\widetilde{W}H + b_2) \tag{2}$$

where W, $\widetilde{W}$, f, ρ, b_1, and b_2 denote the weight matrices, the activation functions, and the biases of the encoding layer and the decoding layer, respectively.

Traditional AE considers the training error as objective function, which is prone to overfitting. In contrast, the LiSSA [49] does not only focus on the training error but also the sensitivity with respect to unseen samples with small differences (perturbations) from training samples, to learn more informative features and enhance the generalization capability of the AE model. The detailed description of LiSSA can be found in Appendix A. In addition, deep architecture ANN could better capture the characteristics of the load. Thus, we use the trained encoders of several LiSSAs via layer-by-layer stacking to initialize D-LiSSA. Let LiSSA$_l$ be the l^{th} $(l = 1,2,\cdots, L)$ LiSSA, where L and H_l denote the total number of hidden layers and the hidden representations, respectively. The LiSSA$_l$ is trained independently using H_{l-1} as inputs and outputs the corresponding reconstructed data $\widetilde{H}_{l-1}$, where H_0=X_b. Eventually, a nonlinear fully connected feedforward neural network as regression layer is appended on top of H_L. This layer takes the hidden representation H_L of the

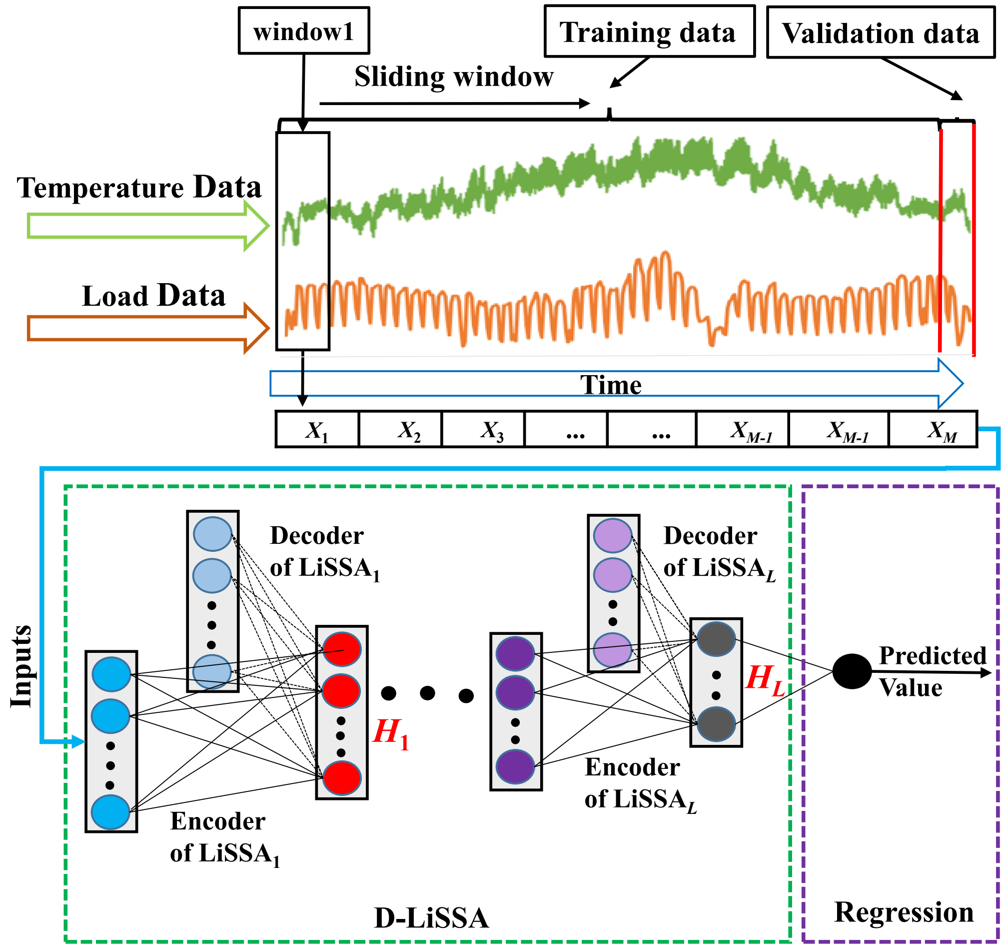


Fig. 2. Overview of D-LiSSA for load forecasting.

last LiSSA$_L$ as inputs and outputs a single value representing the predicted load in a future time stamp. Fig. 2 shows the structure of D-LiSSA for STLF.

2.3 Training of D-LiSSA

The training of D-LiSSA consists of the initialization and fine-tuning phases. The D-LiSSA follows the standard layer-by-layer training rule to train each LiSSA via minimization of the PE between the previous hidden layer outputs H_{l-1} as inputs and the corresponding reconstruction $\widetilde{H}_{l-1}$. For the l^{th} individual LiSSA$_l$, connection weights of both input to hidden layers and hidden to output layers are optimized by using error backpropagation algorithm. Detailed derivations of the connection weights of each individual LiSSA can be found in [49].

In the fine-tuning phase, the initialized D-LiSSA is further trained. Firstly, a nonlinear fully connected feedforward neural network as a regression layer is appended after the L^{th} stacking H_L. Parameters of this layer are randomly initialized from [-1,1] as prior knowledge about their values is unavailable. Then, the whole D-LiSSA is trained with historical load data. The backpropagation algorithm is applied to fine tune weights and biases of all layers in D-LiSSA. The objective function is as follows:

$$\begin{aligned} &\arg\min_{W} \frac{1}{2M}\sum_{b=1}^{M}\left((\beta(X_b)-y_b)^2+E\left[\left(\beta(X_b+\Delta X)-\beta(X_b)\right)^2\right]\right) \\ &=\arg\min_{W} \frac{1}{2M}\sum_{b=1}^{M}((\beta(X_b)-y_b)^2+\frac{1}{C}\sum_{c=1}^{C}\left(\beta(X_b+\Delta X_c)-\beta(X_b)\right)^2\Bigg) \end{aligned} \tag{3}$$

where y_b and $\beta(\cdot)$ denote the target output value of the b^{th} sample and the output of D-LiSSA, respectively. C denotes the number of generated uniformly distributed random points $\Delta X_c \in \mathbb{R}^n$ $(c = 1, \dots, C)$ with each coordinate range from$[-Q, Q]$, where Q is a given value like 0.01. The detailed introduction of Q can be found in literature [49]. Let $\{H_1, H_2, \dots, H_l\}$ and $\{W_1, W_2, \dots, W_l\}$ $(l = 1,2,\cdots, L)$ be outputs and connection weights of each layer, respectively. Let f_l be the activation function of the l^{th} layer. Thus, the output of the l^{th} layer is as follows:

$$H_l = f_l(W_{l-1}H_{l-1}) \tag{4}$$

The objective function adopts a second-order normal form for mean square error and sensitivity terms, as follows:

$$L(W) = \frac{1}{M}(\beta(X_b) - y_b)^T(\beta(X_b) - y_b) \tag{5}$$

and

$$R(W) = \frac{1}{2M}\frac{1}{C}\big(\beta(X_b + \Delta X_c) - \beta(X_b)\big)^T\big(\beta(X_b + \Delta X_c) - \beta(X_b)\big) \tag{6}$$

Then, the objective function can be written as

$$J(W) = L(W) + R(W) \tag{7}$$

To learn the optimal parameters of the neural network model, the gradient descent method is used to obtain the partial derivatives of the weights W_l of the l^{th} layer.

$$\begin{aligned}\frac{\partial J(W)}{\partial W_l} &= \frac{\partial L(W)}{\partial W_l} + \frac{\partial R(W)}{\partial W_l} \\ &= \frac{1}{2M}\sum_{b=1}^{M}\left(\frac{\partial L(W)}{\partial \beta(X_b)}\frac{\partial \beta(X_b)}{\partial H_l}\frac{\partial H_l}{\partial W_l}\right) \\ &\quad + \frac{1}{C}\sum_{c=1}^{C}\left(\frac{\partial R(W)}{\partial \beta(X_b + \Delta X_c)}\frac{\partial \beta(X_b + \Delta X_c)}{\partial H_l}\frac{\partial H_l}{\partial W_l}\right)\end{aligned} \tag{8}$$

Then, terms in Eq. (8) can be further expanded as follows:

$$\frac{\partial L(W)}{\partial \beta(X_b)} = 2(\beta(X_b) - y_b) - \frac{2}{C}\sum_{c=1}^{C}\big(\beta(X_b + \Delta X_c) - \beta(X_b)\big) \tag{9}$$

$$\frac{\partial \beta(X_b)}{\partial H_l} = \frac{\partial \beta(X_b)}{\partial H_L}\frac{\partial H_L}{\partial H_{L-1}}\cdots\frac{\partial H_{l+1}}{\partial H_l} \tag{10}$$

$$\frac{\partial H_{l+1}}{\partial H_l} = W_l f_l{'}(W_l H_l) \tag{11}$$

$$\frac{\partial H_l}{\partial W_l} = H_{l-1} f_l{'}(W_l H_{l-1}) \tag{12}$$

$$\frac{\partial R(W)}{\partial \beta(X_b + \Delta X_c)} = 2\big(\beta(X_b + \Delta X_c) - \beta(X_b)\big) \tag{13}$$

$$\frac{\partial \beta(X_b + \Delta X_c)}{\partial H_l} = \frac{\partial \beta(X_b + \Delta X_c)}{\partial H_L}\frac{\partial H_L}{\partial H_{L-1}}\cdots\frac{\partial H_{l+1}}{\partial H_l} \tag{14}$$

where $f_l{'}$ denotes the partial derivative of f_l.

Algorithm 1 Feature learning and predicting of D-LiSSA

Input: $M \times n$ input data D, where w and M denote the window size and the number of samples, respectively. m_l denotes the number of hidden neurons on the encoding layer of LiSSA_l, $l = 1, 2, \cdots, L$.
Output: Target output value $\beta(X_b)$.
Feature learning:
1: Scale each input feature to the range of [0, 1].
2: Set $H_0 = D$.
3: For $l = 1$ to L do
  3.1: Train LiSSA_l with m_l neurons on the encoding layer using H_{l-1} as the input.
  3.2: Compute the outputs of the encoding layer for all training samples to form a $M \times m_l$ matrix H_l.
4: End for
Predicting:
1: Initializing the weights of D-LiSSA through L stacking H_l and adding regression layer.
2: Using BP algorithm to optimize the weights of each layer of D-LiSSA by Eq. (15).
3: Output the predicted value $\beta(X_b)$, where $b = 1,2,\cdots,M$.

Finally, the weight update formula for the l^{th} layer is as follows:

$$\begin{aligned} W_l &= W_l - \alpha \frac{\partial J(W)}{\partial W_l} \\ &= W_l \\ &- \alpha \left(\frac{1}{2M} \sum_{b=1}^{M} \left(\frac{\partial L(W)}{\partial \beta(X_b)} \frac{\partial \beta(X_b)}{\partial H_L} \frac{\partial H_L}{\partial H_{L-1}} \cdots \frac{\partial H_{l+1}}{\partial H_l} \frac{\partial H_l}{\partial W_l} \right. \right. \\ &+ \left. \left. \frac{1}{C} \sum_{c=1}^{C} \left(\frac{\partial R(W)}{\partial \beta(X_b + \Delta X_c)} \frac{\partial \beta(X_b + \Delta X_c)}{\partial H_L} \frac{\partial H_L}{\partial H_{L-1}} \cdots \frac{\partial H_{l+1}}{\partial H_l} \frac{\partial H_l}{\partial W_l} \right) \right) \right) \end{aligned} \tag{15}$$

where α denotes the learning rate. Algorithm 1 shows the detailed steps of the features learning and the prediction by D-LiSSA.

2.4 Computational and Space Complexity of the D-LiSSA

In case of a sigmoid nonlinearity, computing the SS (or its gradient) has about the same cost as computing the reconstruction error (or its gradient). Computation of the SS can be found in Appendix A. Suppose the input dimension is n for the input layer and m_l hidden neurons for the $LiSSA_l$. Using H Halton points to compute the SS for the $LiSSA_l$, the computational complexity is $O\left(H(m_l^2 m_{l-1})\right)$. Thus, the overall computational complexity of the D-LiSSA is $O\left(H(\sum_{l=1}^{L-1} m_l^2 m_{l+1} + m_l)\right)$, where $l = 1,2,\cdots,L$. For an individual $LiSSA_l$ with m_l hidden neuro nodes, there are $n \times m_l + m_l$ weight parameters. Let $m_0 = n$, the overall time space complexity of the D-LiSSA with L hidden layers is $(n \times m_1 + m_1) + (m_1 \times m_2 + m_2) + \cdots + (m_{L-1} \times m_L + m_L) + m_L \times 1 = \sum_{l=0}^{L-1} (m_l + 1) m_{l+1} + m_L$. In addition, it is found that the run time of D-LiSSA is acceptable for the current computers, not to mention future computers for future practical application from Subsection 3.5.

2.5 Performance Evaluation Criteria

The accuracy of different models is assessed by comparing the predicted load with the actual load data. The root mean square error (RMSE): RMSE $= \sqrt{\frac{1}{M}\sum_{b=1}^{M}(y_b - \hat{y}_b)^2}$ MW, mean absolute error (MAE): MAE $= \frac{1}{M}\sum_{b=1}^{M}|y_b - \hat{y}_b|$ MW, mean absolute percentage error (MAPE): MAPE $= \frac{1}{M}\sum_{b=1}^{M}\frac{|y_b-\hat{y}_b|}{y_b} \times 100\%$, R^2 score: $R^2 = 1 - \frac{\sum_{b=1}^{M}(y_b-\hat{y}_b)^2}{\sum_{b=1}^{M}(y_b-\bar{y})^2}$, and the explained variance (EV) score: EV $= 1 - \frac{Var(y-\hat{y})}{Var(y)}$ are employed as five evaluation criteria for the precision of prediction, where $\hat{y}_b$ is the predicted value of the b^{th} sample, y_b is the corresponding true value, and $\bar{y} = \frac{1}{M}\sum_{b=1}^{M} y_b$. $\hat{y}$, y, and Var denote the predicted target output, the corresponding (correct) target output, and the variance, respectively.

With ARIMA model [11] as the benchmark, the improvement of the models over the baseline is defined as follows:

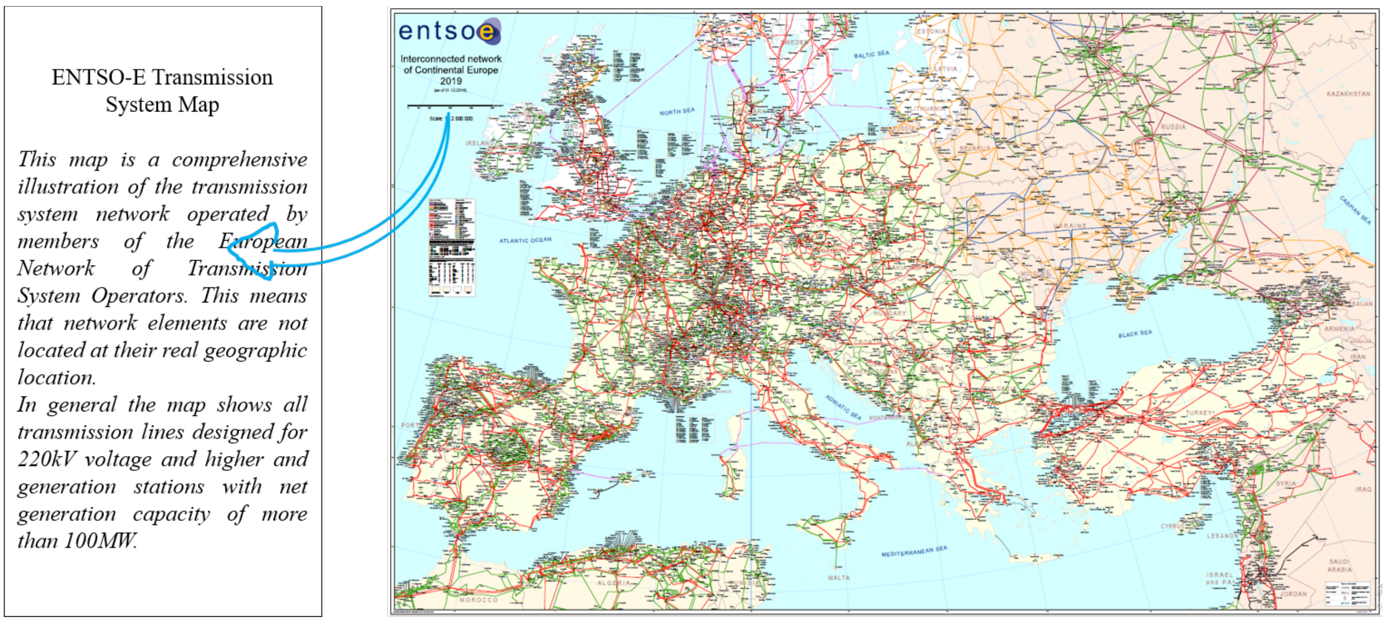
$$\mathrm{Imp} = \left(\left|1 - \frac{error}{error_b}\right|\right) \times 100 \quad \% \tag{16}$$

where $error$ is one of the metrics RMSE, MAE, and MAPE, and $error_b$ is the corresponding $error$ of ARIMA.

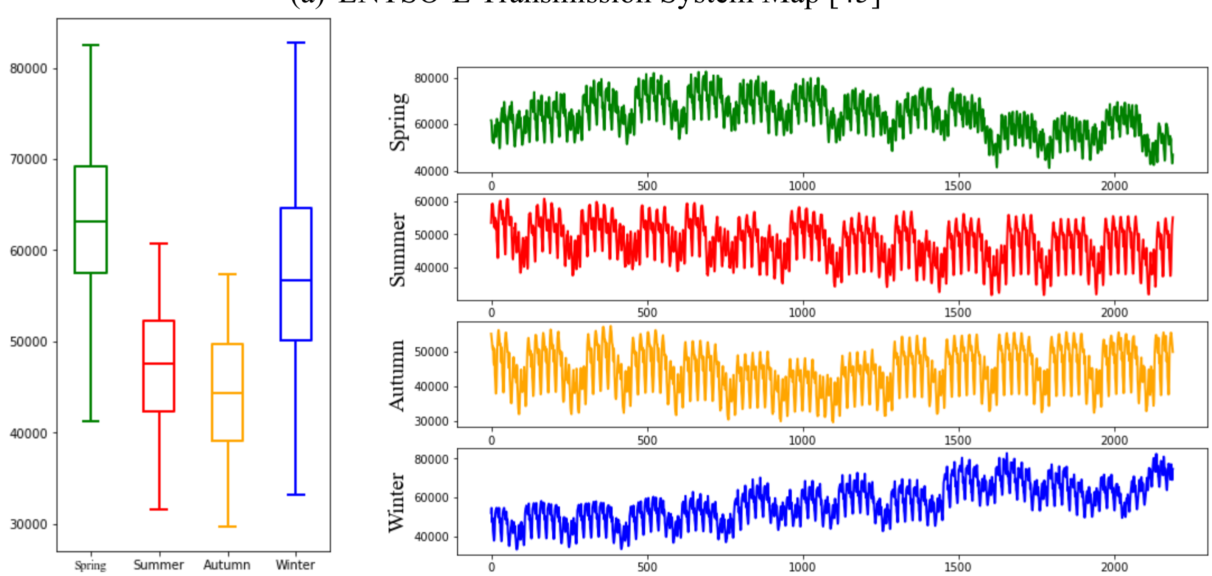
3. Numerical Results And Discussion

3.1. Description of Datasets

In our experiments, we show the performance of D-LiSSA with two hidden layers to forecast 1 hour-ahead load consumption data using 24 hours window size historical load data. In the first part, we test the model on the real-world electricity markets of France (FR), Germany (GR), Romania (RO), and Spain (ES) in Europe. All the load consumption data are obtained from the electricity market, i.e., the ENTSO-E dataset [51]. Moreover, the corresponding temperature of the countries obtained from the Renewables dataset [53] is also included in our experiment. Nationally aggregated temperature output data with hourly intervals were acquired from system operators in the four countries. These data series were converted to Greenwich Mean Time and aggregated hourly resolution for compatibility with MERRA-2 [53]. Load and temperature data with one-hour interval of the two characteristics are used in the study to cover the period from 1st January to 31st December for 2014 and 2015. For simplicity and without loss of generality, the load forecasting strategy is applied to the 2015 period that recorded the latest load in the given ENTSO-E dataset.



(a) ENTSO-E Transmission System Map [45]



(b) France

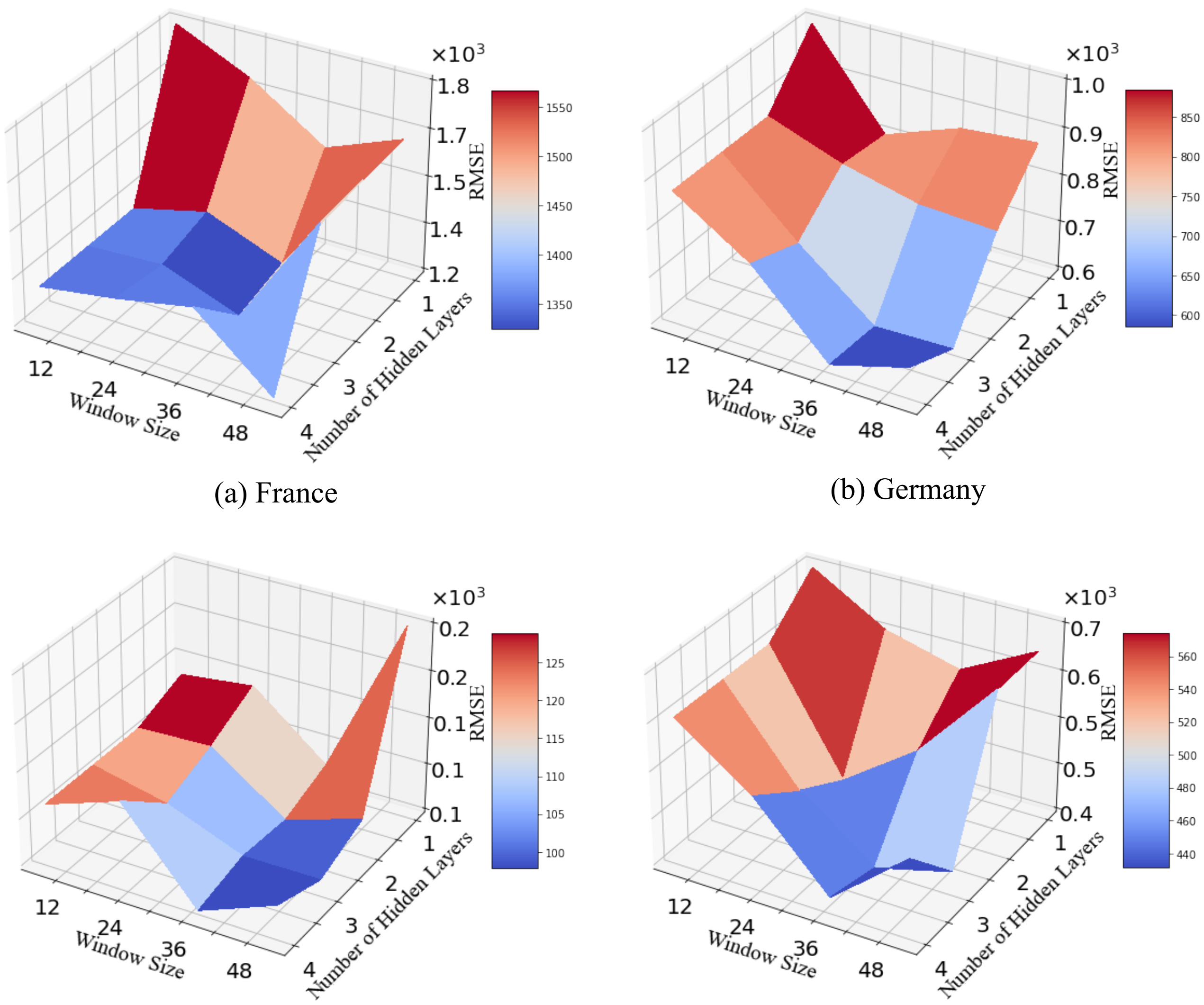
Fig. 3. (a) ENTSO-E Transmission System Map [45]; (b) data description for France.

In our case experiments, 90% of the historical data are used for model training, 10% are reserved for the model validation in 2014, and the hourly load of the year 2015 in the ENSTO-E data is used for testing data. Fig. 3 shows the ENTSO-E and the decomposition electric load for France market in four seasons (2015).

3.2. D-LiSSA Architecture Selection

To identify the optimal architecture of D-LiSSA, a grid search is performed. Fig. 4 shows the RMSE of different D-LiSSA in four countries. The window size and number of hidden layers to establish D-LiSSA is chosen from {12h, 24h, 36h, 48h} and {1,2,3,4}, respectively. Experimental results reveal that as predictive horizon increases D-LiSSA with larger window size obtains lower RMSE in four countries. Moreover, augmented window size needs to be combined with more hidden layers to capture the characteristic of the load data.

However, both larger window and more layers will lead to more intricate structures that are prone to over-



(a) France (b) Germany

(c) Romania (d) Spain

Fig. 4. RMSE of D-LiSSA with different window size and hidden layers in four countries.

fitting. Besides, the vanishing gradient problem triggered by deep architecture also hampers the fine-tuning process of D-LiSSA. Thus, for short horizon forecasting, the simplification of training could relieve the shortcoming of decreasing input size and depth and promote the predictive accuracy. Consequently, optimizing the architecture of D-LiSSA with different window and hidden layers is significant and indispensable. The optimal structure yields the least validation error compared to various structures as shown in Fig. 4. In this paper, D-LiSSA with window sizes and hidden layers of (48h,4), (36h, 3), (36h, 4), and (36h,3) are used for prediction in France, Germany, Romania, and Spain, respectively.

3.3. Case Study 1: Comparison with Other Algorithms in the Four European Countries

1) *Comparison with other Methods*: In this section, we compare D-LiSSA with other methods in different European countries. As a generic prediction model, D-LiSSA has good generalization ability and stability as its main advantages. The model can be directly applied to other countries or regions as well. The temperature affects the load consumption of the residents' activities and different countries have various changing patterns of temperature. In addition, the past nonlinear behavior of electricity price has a significant impact on the future electric load forecasting. Thus, researchers can consider a variety of effective factors on electric load. A load forecasting model should recognize the important features of this signal's behavior in the past and consider them as inputs to further improve accuracy of electric load forecasting. For several reasons, there are some relevant differences in electric loads between the four countries we selected, as follows:

• The countries are in different latitudes and longitudes. Obviously, for countries that are far from the Atlantic Ocean and in the interior of Europe, the weather can be very hot and cold in summer and winter, respectively. Romania's electricity demand is quite large by the usage of electric fans, air conditioners, and

Table 1. Comparisons of 1 hour-ahead forecasting performance based on different metrics in four countries.

Nations	Methods	EV	Imp	MAE	Imp	MAPE	Imp	R^2	Imp	RMSE	Imp
FR	ARIMA	0.969	-	1540	-	2.91	-	0.969	-	2065	-
	LSTM	0.971	0.2	1469	5	2.78	4.6	0.972	0.2	2007	3
	ResNet	0.975	0.7	1331	14	2.40	17.6	0.975	0.7	1834	11
	DBN	0.990	2.3	820	47	1.56	46.4	0.990	2.3	1140	45
	BDT	0.985	1.7	854	45	1.52	47.9	0.985	1.7	1418	31
	D-LiSSA	**0.995**	**2.9**	**587**	**62**	**1.07**	**63.2**	**0.995**	**2.7**	**900**	**56**
GR	ARIMA	0.978	-	1078	-	1.90	-	0.978	-	1511	-
	LSTM	0.982	0.4	1018	6	1.82	4.2	0.982	0.4	1385	8
	ResNet	0.986	0.8	864	20	1.51	20.2	0.986	0.8	1202	20
	DBN	0.991	1.3	660	39	1.20	36.5	0.991	1.3	974	36
	BDT	0.993	1.5	606	44	1.08	42.9	0.993	1.5	854	43
	D-LiSSA	**0.998**	**2.0**	**523**	**51**	**0.94**	**50.3**	**0.997**	**2.0**	**688**	**54**
RO	ARIMA	0.954	-	139	-	2.33	-	0.954	-	189	-
	LSTM	0.961	0.79	127	9	2.10	9.6	0.961	0.8	173	8
	ResNet	0.967	1.39	130	6	2.17	6.5	0.962	0.9	170	10
	DBN	0.979	2.77	92	34	1.56	32.9	0.979	2.7	127	33
	BDT	0.976	2.39	96	31	1.63	29.8	0.976	2.4	135	28
	D-LiSSA	**0.989**	**3.75**	**80**	**42**	**1.34**	**42.2**	**0.990**	**3.8**	**109**	**42**
ES	ARIMA	0.960	-	696	-	2.49	-	0.960	-	932	-
	LSTM	0.984	2.53	631	9	2.24	10.2	0.965	0.5	874	6
	ResNet	0.984	2.53	439	37	1.54	37.9	0.984	2.5	596	36
	DBN	0.988	2.94	351	50	1.29	48.3	0.988	3.0	508	45
	BDT	0.982	2.32	421	39	1.47	40.9	0.982	2.3	624	33
	D-LiSSA	**0.995**	**3.71**	**304**	**56.33**	**1.07**	**57.20**	**0.990**	**3.60**	**435**	**53**

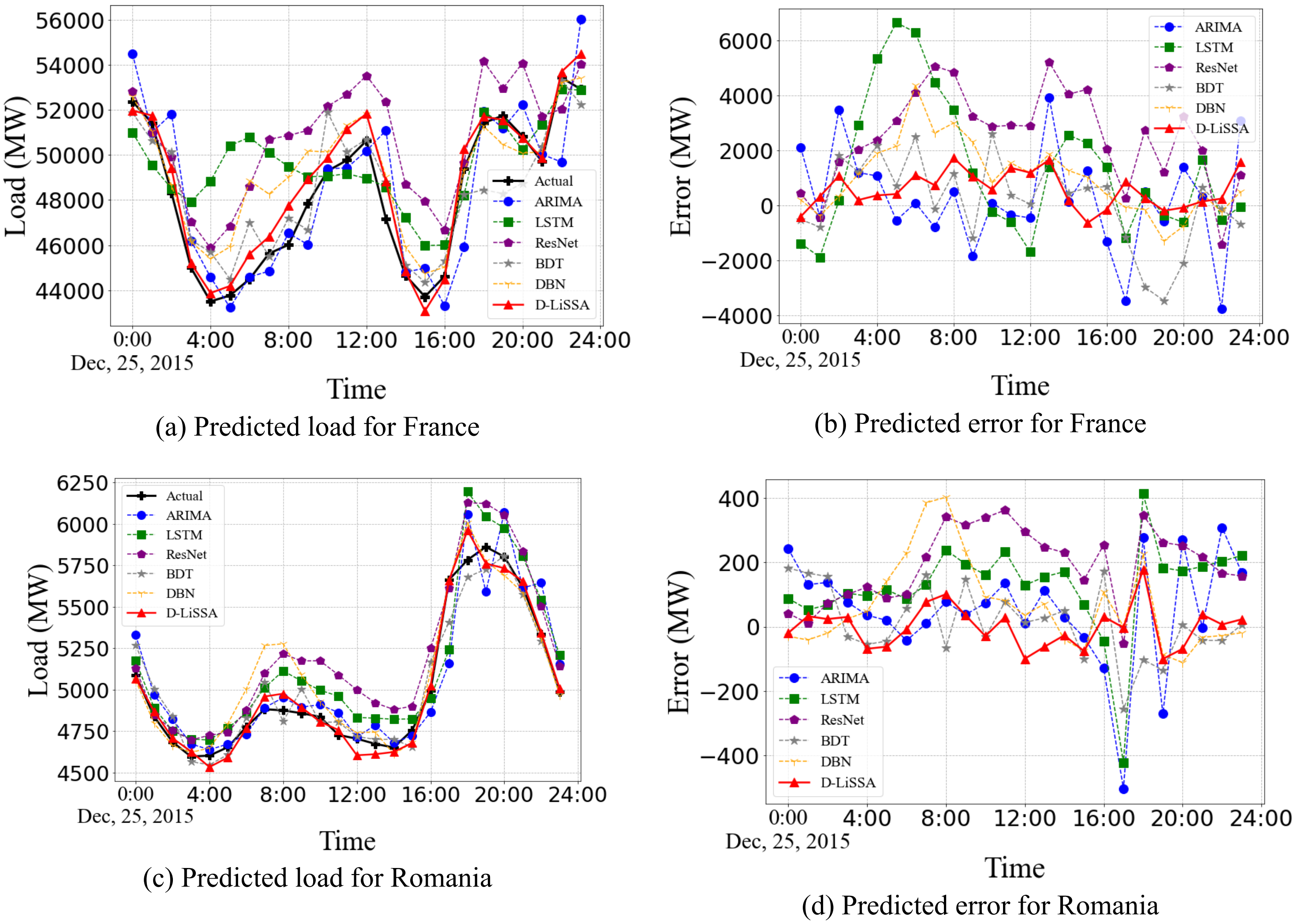
refrigerators. Especially, there are several days that are extremely hot in August. To ensure the daily electricity demand for residents is met, the government asks for closing most commercial activities, offices, and heavy or light industries during these periods [38]. France has a similar weather pattern as well.

• France has a particularly heavy load consumption in winter from the cold winds of the western Atlantic, where the electricity is used for heating as an alternative to conventional fuels such as natural gas.

• Unlike other countries, Germany was one of the countries with rapid development of heavy industry in the world in the 20^{th} century. The industrial load is the main factor of heavy load consumption and largely unaffected by seasonal patterns. Thus, the annual fluctuation of electric load in Germany tends to be stable.

• The annual gross domestic product growth of the four countries from recent years (from 2012 to 2015) are continuing to grow, and the corresponding total consumption electric load for four countries from 2012 to 2015 is growing at a corresponding rate. These real data are obtained from the World Bank [54].

The proposed D-LiSSA model and other comparative models, including the one benchmark method such as auto-regressive integrated moving average (ARIMA) [11], long-short term memory (LSTM) [25], residual neural network (ResNet) [27], deep belief network (DBN) [28], and binary decision tree (BDT) [52] are applied to the study for the four countries. Our model is trained using computationally efficient Adam optimizer [55] with 300 epochs. The Adam has not only little memory requirements but also is well suited for problems that are large in terms of parameters in deep learning. As for computation of the SS can be found in Appendix A, Q and H are set to 0.01 and 50, respectively. For reference, the deep structures of LSTM, ResNet, and DBN adopt the same optimizer and with default parameters as given in [25, 27, 28] for training models. In the ARIMA model, the parameters p, d, and q are set to 1, 1, and 2, respectively. In this way, representative results for the whole year of 2015 can be obtained. Similar procedures have been used in previous works in this area. The various evaluation criteria computed by these models are reported in Table 1. Since ARIMA as the benchmark model, the improvement (i.e., Imp) of the model in Table 1 is neglected in our experiments. As seen from this table, there are some differences in the results for different countries as measured by evaluation criteria. Despite such differences, it is worth noting that the performance of D-LiSSA in different European countries (the average EV, MAPE, R^2 in the year 2015) are similar. Furthermore, as the temperature data are considered, the average criteria of the obtained results in these criteria are also similar to those reported in Table 3 and Table 4 by other models for different countries. But D-LiSSA yields the best results in all evaluated criteria. This demonstrates the effectiveness of D-LiSSA. Other artificial neural networks including ResNet and DBN obtained the second-best results. ResNet model first uses basic deep residual networks structure with several fully connected layers to produce preliminary forecast of the two hours. The deep residual network integrates domain knowledge and builds different neural network blocks. This ensemble of strategy could enhance the prediction capability of the deep residual network by combining multiple individual networks. Although using the individual CNN and LSTM for building different blocks make the deep neural networks to have high flexibility and effectiveness, as the number of layers increases, the depth model becomes more and more difficult to train. Thus, the number of hidden layers is often considered small so that it reduces the effectiveness of ResNet. In [28], raw data sources are normalized using Box-Cox transformation. The deep belief network (DBN) adopts less hidden neurons to overcome the limitations of overfitting of the traditional neural networks. It can learn to the feature pattern of the input data owing to the use of multiple layers of nonlinear transformation. Thus, DBN produces exceptional results. Binary decision tree (BDT) obtains the similar results because it is a supervised machine learning method. However, LSTM is sensitive to data scale and has more hyper-parameters that needs to be tuned effectively, which limits the performance of the model. In addition, parameter tuning can be affected on the load data with extremely high volatility. But LSTM is also superior to ARIMA which demonstrates the artificial neural network models significantly outperform the traditional statistical models. Because the ANN models can better learn the nonlinear forecasting relationship of load signal.



(a) Predicted load for France
(b) Predicted error for France
(c) Predicted load for Romania
(d) Predicted error for Romania

Fig. 5. Christmas one-day load forecasting result for France (upper row), and Romania (lower row). Christmas one-day predicted load (left column); Christmas one-day predicted error (right column) by ARIMA, LSTM, ResNet, BDT, DBN, and D-LiSSA, respectively.

2) *Application to Holiday Prediction*: The load during special holidays remains to be predicted with the greatest difficulties as discussed by other authors in the literature review. Thus, actual load and forecast error in the winter test day, i.e., Christmas one-day test data is taken from the retained data to demonstrate the performance of D-LiSSA and the other selected load forecasting models. We provide the results for France (upper row) and Romania (lower row). Figs. 5 (a) and (c) left column provide the load forecasting curve results and Figs. 5 (b) and (d) right column demonstrate the prediction error for the same period. When considering the challenging testing period, D-LiSSA produces results that match the real data better in both figures. Only small deviations are seen in Fig. 5 (a). In Fig. 5 (b), the upper and lower bounds of the error are within the minimum interval and smaller than other comparative methods.

3) *Week-ahead Peak Load Forecasting*: In both 1 hour-ahead and holiday forecasting cases, the proposed D-LiSSA proves to be more effective than the other models. To further validate the capacity of D-LiSSA, the forecasting results in the peak loads are examined. On the other hand, the peak load is considered as an important factor for the grid reliability in week-ahead load forecasting practices [28]. Even with high forecasting accuracy, an underestimation of the peak load may result in a power outage. In certain instances, the forecasting of the weekly peak load is the objective of short-term forecasting as the peak load is the most important one in a certain time interval [28]. Based on the domain knowledge, the majority of the peak loads

Table 2. Performance evaluation of week-ahead load forecasting on ENTSO-E data in 2015.

Seasons		**Winter**		**Spring**		**Summer**		**Fall**	
Nations	**Methods**	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**	**RMSE**
FR	ARIMA	8.8	7900	6.9	5092	5.7	4224	8.8	6099
	LSTM	7.5	7210	5.7	4468	4.4	2926	7.5	5403
	ResNet	8.2	8105	6.9	5181	**3.7**	**2440**	8.2	6113
	DBN	7.5	7013	6.1	4489	4.3	2627	7.5	4991
	BDT	9.0	8218	8.9	6545	5.9	4578	9.0	5772
	D-LiSSA	**6.7**	**6049**	**5.5**	**4140**	4.3	2640	**6.7**	**4920**
GR	ARIMA	7.6	7389	9.3	8163	9.3	8121	7.6	7454
	LSTM	**5.0**	**4666**	5.3	5442	3.2	3064	2.5	2177
	ResNet	9.2	9055	5.4	5439	**2.0**	**1737**	3.4	2730
	DBN	6.1	6970	4.7	4723	2.6	2273	2.3	2031
	BDT	6.8	4938	6.0	6148	2.6	2476	3.2	4092
	D-LiSSA	5.2	5622	**4.4**	**4527**	2.4	2035	**2.3**	**2028**
RO	ARIMA	5.9	590	6.6	611	6.4	553	6.2	5328
	LSTM	4.9	508	4.3	425	3.9	325	3.2	272
	ResNet	5.3	542	5.5	484	4.2	336	5.0	478
	DBN	5.3	500	4.1	406	3.3	284	3.1	260
	BDT	5.5	539	5.3	570	5.9	508	6.2	563
	D-LiSSA	**4.8**	**470**	**3.4**	**371**	**3.3**	**281**	**2.5**	**217**
ES	ARIMA	6.6	2945	6.9	2955	8.4	3809	6.7	2774
	LSTM	**5.3**	**2542**	4.6	2115	**5.2**	**2225**	**3.5**	**1413**
	ResNet	6.7	2946	4.9	2044	8.0	3480	4.0	1777
	DBN	6.2	2728	4.5	1954	5.5	2334	3.6	1465
	BDT	7.0	3272	6.0	2701	7.5	3687	5.1	2232
	D-LiSSA	5.9	2623	**4.3**	**1950**	5.5	2331	3.6	1446

occur at different time periods in spring, summer, fall, and winter. Hence, it is necessary to evaluate the week-ahead peak load forecasting using D-LiSSA and other comparative models in this subsection. The computational results for week-ahead peak load forecasting cases are presented in Table 2. It is observed that D-LiSSA outperforms the other five models in the most cases. The values of MAPE and RMSE indicate significant differences between D-LiSSA and the other models. The comparative analysis confirms the effectiveness of the proposed D-LiSSA in improving the peak load forecasting performances.

3.4 Case Study 2: Electric Load for Different Seasons

Validating the efficiency and effectiveness of the proposed forecasting model requires comparisons with comparative models in solving the same problem. In this work, four countries corresponding to four seasonal electric load and weather data of year 2015 are considered in the season test case. Statistical method and several based-ANN forecasting models including ARIMA [11], LSTM [25], ResNet [27], DBN [28], and tree-based BDT [52] have used these cases because the electric loads have a significant seasonal variation in the whole year and especially have a maximum peak value in summer. Thus, they are considered here so that the proposed D-LiSSA model can be compared with the state-of-the-art methods. For the fair comparisons, D-LiSSA also considers the same evaluation criteria for the comparative models. To analyze the results from different perspectives of the same prediction model, Tables 3 and Table 4 provide the statistical error metrics results of the four seasons. Meanwhile, Fig. 6 only shows the short-term load forecasting performance of the

Table 3. EV results for 1 hour-ahead forecasting of the four countries on ENTSO-E data in 2015.

Nations	Methods	Winter	Spring	Summer	Fall	Mean
FR	ARIMA	0.936	0.952	0.926	0.939	0.938
	LSTM	0.927	0.956	0.937	0.952	0.943
	ResNet	0.927	0.966	0.971	0.971	0.959
	DBN	0.977	0.985	0.978	0.985	0.981
	BDT	0.945	0.984	0.985	0.983	0.974
	D-LiSSA	**0.981**	**0.095**	**0.994**	**0.995**	**0.991**
GR	ARIMA	0.978	0.979	0.980	0.972	0.977
	LSTM	0.979	0.979	0.985	0.982	0.981
	ResNet	0.977	0.988	0.94	0.991	0.988
	DBN	0.990	0.990	0.993	0.990	0.991
	BDT	0.990	0.990	0.985	0.993	0.993
	D-LiSSA	**0.997**	**0.997**	**0.998**	**0.998**	**0.998**
RO	ARIMA	0.951	0.950	0.947	0.937	0.946
	LSTM	0.961	0.957	0.947	0.954	0.955
	ResNet	0.961	0.962	0.970	0.959	0.963
	DBN	0.980	0.977	0.973	0.977	0.977
	BDT	0.974	0.975	0.969	0.972	0.972
	D-LiSSA	**0.989**	**0.979**	**0.976**	**0.983**	**0.982**
ES	ARIMA	0.956	0.949	0.970	0.950	0.956
	LSTM	0.956	0.965	0.978	0.950	0.962
	ResNet	0.980	0.981	0.986	0.986	0.983
	DBN	0.984	0.986	0.992	0.987	0.987
	BDT	0.980	0.987	0.974	0.985	0.981
	D-LiSSA	**0.990**	**0.993**	**0.988**	**0.994**	**0.991**

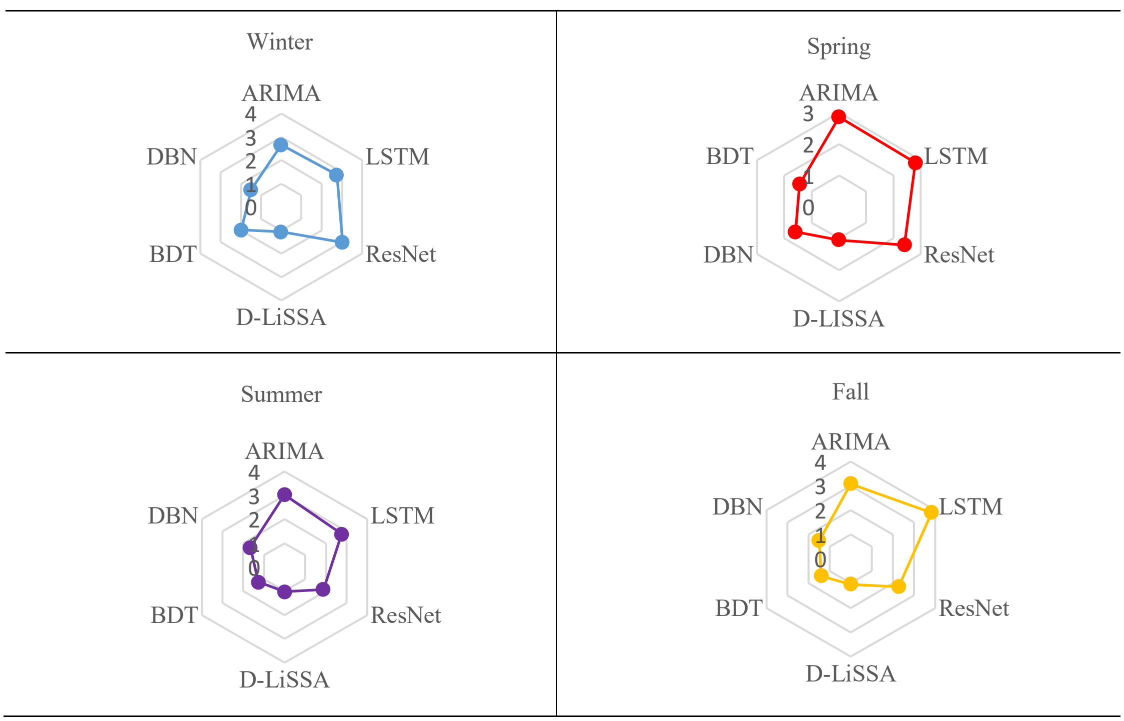


Fig. 6. Short-term load forecasting performance of the proposed method and the other comparative methods based on France.

proposed method and the other comparative methods for four seasons in France, based on the MAPE error measures. Because D-LiSSA uses historical data as well as the unseen samples surrounding in the *Q*-neigh-

Table 4. MAPE results for 1 hour-ahead forecasting of the four countries on ENTSO-E data in 2015.

Nations	Methods	Winter	Spring	Summer	Fall	Mean
FR	ARIMA	2.66	2.86	3.03	3.08	2.91
	LSTM	2.75	2.81	2.76	2.78	2.78
	ResNet	3.03	2.42	1.88	2.27	2.408
	DBN	1.49	1.59	1.65	1.50	1.56
	BDT	1.99	1.44	1.25	1.40	1.52
	D-LiSSA	**1.07**	**1.04**	**1.03**	**1.03**	**1.04**
GR	ARIMA	1.83	1.87	1.81	2.07	1.90
	LSTM	1.96	1.98	1.65	1.68	1.82
	ResNet	2.29	1.41	1.08	1.29	1.52
	DBN	1.29	1.25	1.08	1.20	1.20
	BDT	1.26	1.09	0.94	1.05	1.08
	D-LiSSA	**0.78**	**0.77**	**0.66**	**0.65**	**0.71**
RO	ARIMA	2.28	2.33	2.226	2.47	2.33
	LSTM	2.06	2.11	2.19	2.04	2.10
	ResNet	2.09	2.44	1.94	2.23	2.14
	DBN	1.38	1.61	1.74	1.51	1.56
	BDT	1.56	1.63	1.70	1.64	1.63
	D-LiSSA	**1.13**	**1.48**	**1.47**	**1.28**	**1.34**
ES	ARIMA	2.80	2.52	2.09	2.56	2.49
	LSTM	2.70	2.06	1.76	2.44	2.24
	ResNet	1.81	1.50	1.51	1.37	1.55
	DBN	1.55	1.28	1.13	1.19	1.29
	BDT	1.75	1.27	1.59	1.27	1.47
	D-LiSSA	**1.04**	**0.95**	**1.14**	**0.89**	**1.01**

borhood. With the more powerful forecast engine of D-LiSSA, D-LiSSA achieves higher EV as compared to five models. The EV with best possible value 1.0, measure how well the model could learn and represent the variance of load data, indicating the goodness of fit and the ability of the model to forecast future samples. The results of Tables 3 illustrate the effectiveness for different models, including seasonal variations. According to Table 4 results, the mean MAPE and each MAPE value of the proposed model are better than in the other forecasting models for each country. For different seasons, the average MAPE of the proposed model changes from 1.04% for France to 1.01% for Spain. Concurrently, the average MAPE of ARIMA, LSTM, ResNet, DBN, and BDT change from 2.91% to 2.49%, 3.83% to 2.84%, 2.40% to 1.55%, 1.56% to 1.29%, 1.52% to 1.47%, respectively. It is obvious that the proposed D-LiSSA has the highest forecast accuracy because the MAPE of the proposed method changes least. That implies that the seasonal changes do impact he performance of electric load forecasting, however, the proposed model is more stable, and the prediction ability shows a minor fluctuation.

3.5.Computational Efficiency

A good model must have a high efficiency with accuracy and stability. Since the model efficiency is related to the computational time and a short run time corresponds to a high efficiency, we apply the run time to represent the computational efficiency. Furthermore, we carry out the all experiments on the deep learning platform PyCharm with a GeForce RTX 2080Ti graphics processing unit, Window10 operator system with 8GB inner memory and Intel Core i5-9500 3.00GHz central processing unit. Considering the run time, the individual models spend less time for their oversimplified construction. For example, the run time of ARIMA is about 1s, LSTM is about158s, ResNet is about 241s, BDT is about 2s, and DBN is about 80s in France. Table 5 indicates that single models like ARIMA and BDT obtain shorter run time and the model efficiency reduces with increase in model complexity. Similarly, with the increasing model complexity, the artificial neural network models spend more time. The proposed model attains high computational efficiency in artificial neural network models because of using H Halton points; to compute the stochastic sensitivity (SS) for the l^{th} individual LiSSA$_l$ significantly increasing the run time. Although the proposed D-LiSSA has the longest run time, it improves the forecast accuracy and achieving the significant advancement of the reliability and validity. In fact, the run time of D-LiSSA is acceptable for the current computers with good enough graphics processing unit and central processing unit performance. The detail computation of SS can be found in Appendix A.

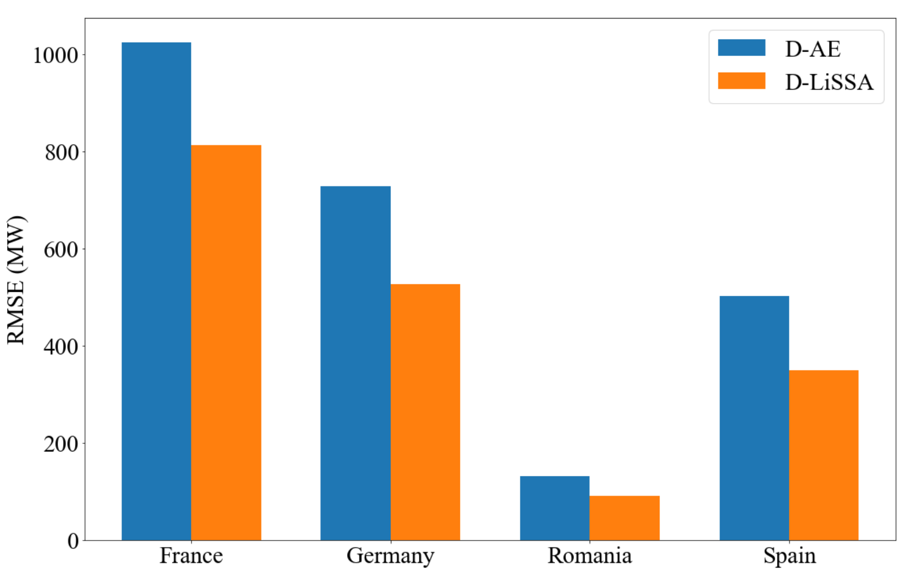
Table 5. The run time of different methods for four countries.

Run Time (s)	FR	GR	RO	ES	Average
ARIMA	1	1	1	1	1
LSTM	158	198	149	146	163
ResNet	241	225	242	222	232
BDT	2	1	1	1	1
DBN	80	80	78	76	78
D-LiSSA	715	819	749	723	751

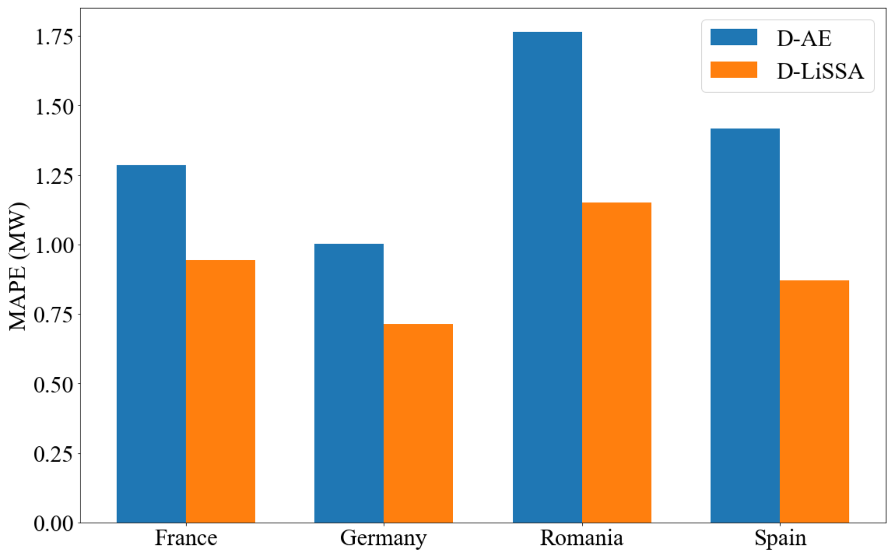
Moreover, Fig. 7 shows the forecasting load with D-LiSSA, LSTM, ResNet, DBN, BDT, and ARIMA as the benchmark model. The figures present the results for summer (from a1 to f1) and winter (from a2 to f2) in Germany. The solid straight black line indicates that the actual and forecasting load demands match exactly, i.e., the actual load equals the predicted load. For instance, the blue dots are the predicted load data, and they are close to the solid black line, and indicate accurate load forecasting results. The high intensity of blue/green/purple/red dot marks concentrate around the solid black line shows small differences between the actual and forecasted load values. Simple observation shows that D-LiSSA is remarkably close to the perfect-match line in different seasons. Besides, the R^2 values are R^2 =0.9963 in summer and R^2=0.9945 in winter, which indicates a good prediction ability of the D-LiSSA for forecasting short-term load data points.

3.6. Generalization Capability of D-LiSSA

To further verify the generalization performance of D-LiSSA, D-AE is built by replacing LiSSA in D-LiSSA with basic autoencoders (with the same architecture) but trained via a minimization of mean squared error. The RMSE and MAPE yielded by D-LiSSA and D-AE are shown in Fig. 8. D-LiSSA yields the lowest



(a) RMSE



(b) MAPE

Fig. 8. RMSE and MAPE of D-LiSSA and D-AE.

Table 6. Comparison of RMSE, MAE on solar and wind datasets.

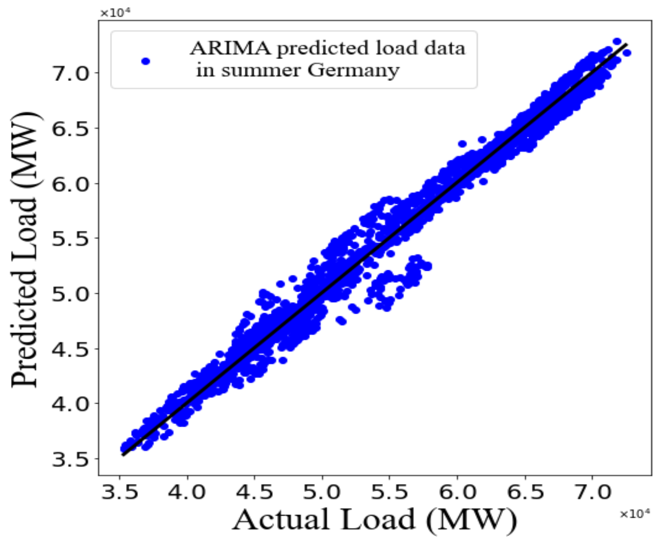
Dataset	Model	RMSE	MAE
Solar	D-AE	26.7 MW	13.3 MW
	D-LiSSA	**26.0 MW**	**12.8 MW**
Wind	D-AE	0.218 m/s	0.359 m/s
	D-LiSSA	**0.191 m/s**	**0.346 m/s**

values in comparison with D-AE on four countries. This experiment verifies D-LiSSA learns more informative hidden representations for SLFT. Moreover, we utilize D-LiSSA in other application situations to demonstrate the generalization performance of the D-LiSSA, including wind speed and solar irradiance forecasting. We conduct 30 min-ahead solar irradiance forecasting on the National Solar Radiation Data Base (NSRDB) [56] and 10 min-ahead wind speed forecasting on wind farm Golden in Colorado, USA [57]. Table 6 gives the RMSE and MAE for two models and shows that D-LiSSA yields the least testing error. These results demonstrate that D-LiSSA has a good generalization ability for the short-term forecasting tasks.

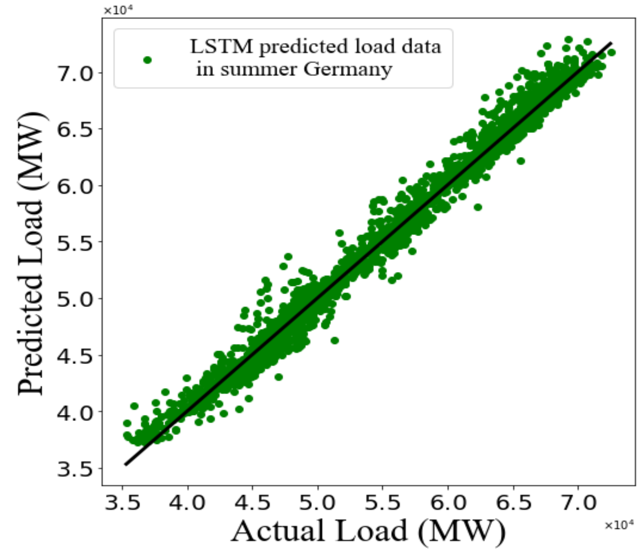
4. Conclusions

This work presents a novel model based on deep autoencoder with localized stochastic sensitivity (D-LiSSA) for short-term load forecasting. D-LiSSA can learn informative hidden representations by adding the perturbations strategy in the Q- neighborhood surrounding in the training samples. Thus, the proposed model is sensitive to similar unseen samples and is effective for features extraction from the historical load data. A nonlinear feedforward neural network as a regression model utilizes the last hidden layer representations by D-LiSSA for load forecasting was developed. To verify the performance of the proposed model, four real-world public electricity datasets from ENTSO-E are used. The results demonstrate that D-LiSSA outperforms other methods due to the reduction of sensitivity and the enhancement of generalization ability.

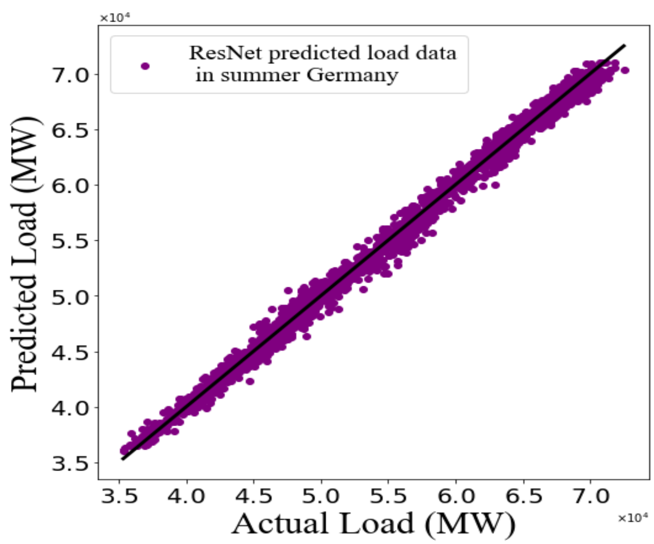
In smart grids, accurate load forecasting with prediction interval, and short/long-term load forecasting always play an important role. Thus, we aim to further explore the load prediction intervals, the load pricing prediction, and long-term forecasting approach based on D-LiSSA. Moreover, an unreliable short-term load forecasting gives challenges to the full utilization of renewable energies in the increasingly complex power market pricing strategies in smart grids. Considering that lower electricity costs require the design of efficient energy management systems and dispatch strategies, we will also apply our method to the household level load forecasting in future work. Since D-LiSSA has a good generalization ability, it is an interesting future work to research on how D-LiSSA can be generalized to provide multi-step ahead predictions.



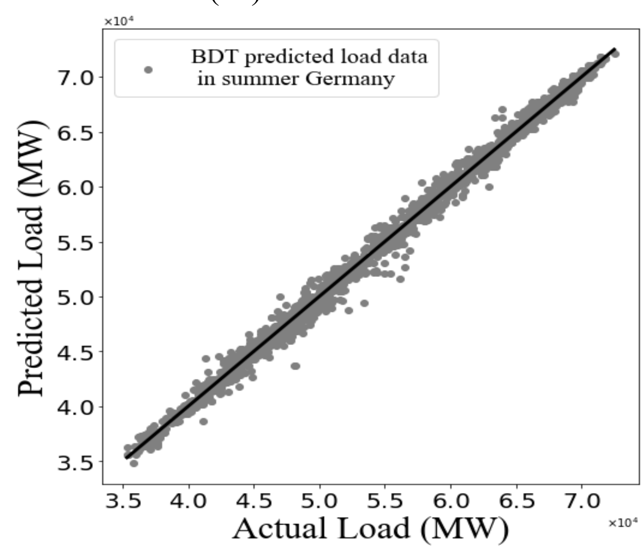
(a1) ARIMA



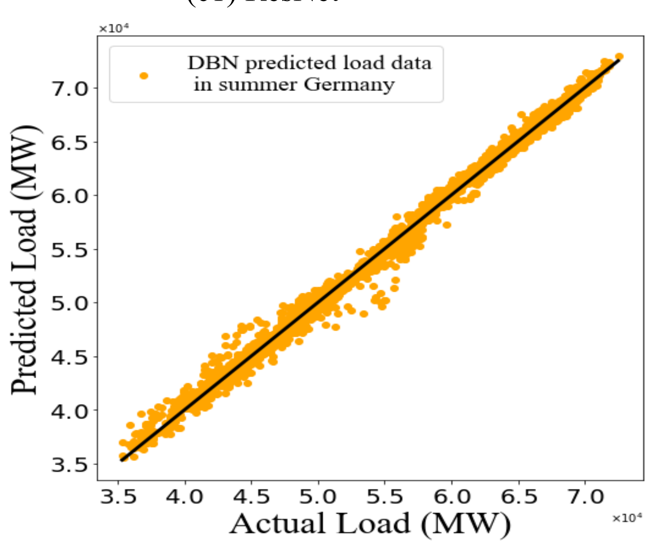
(b1) LSTM



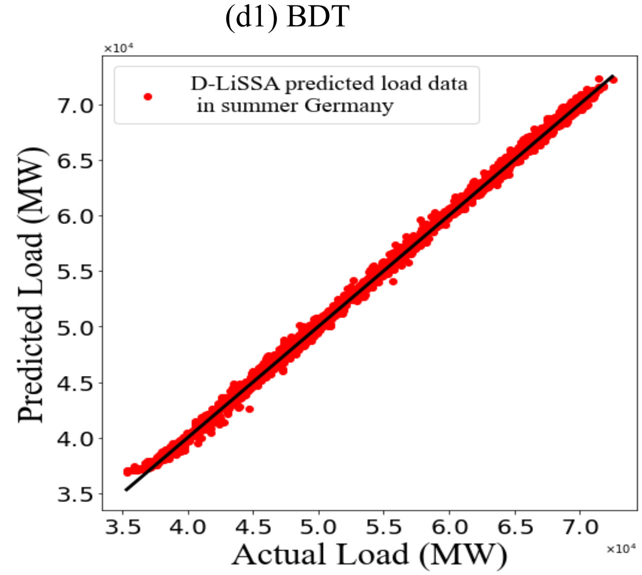
(c1) ResNet



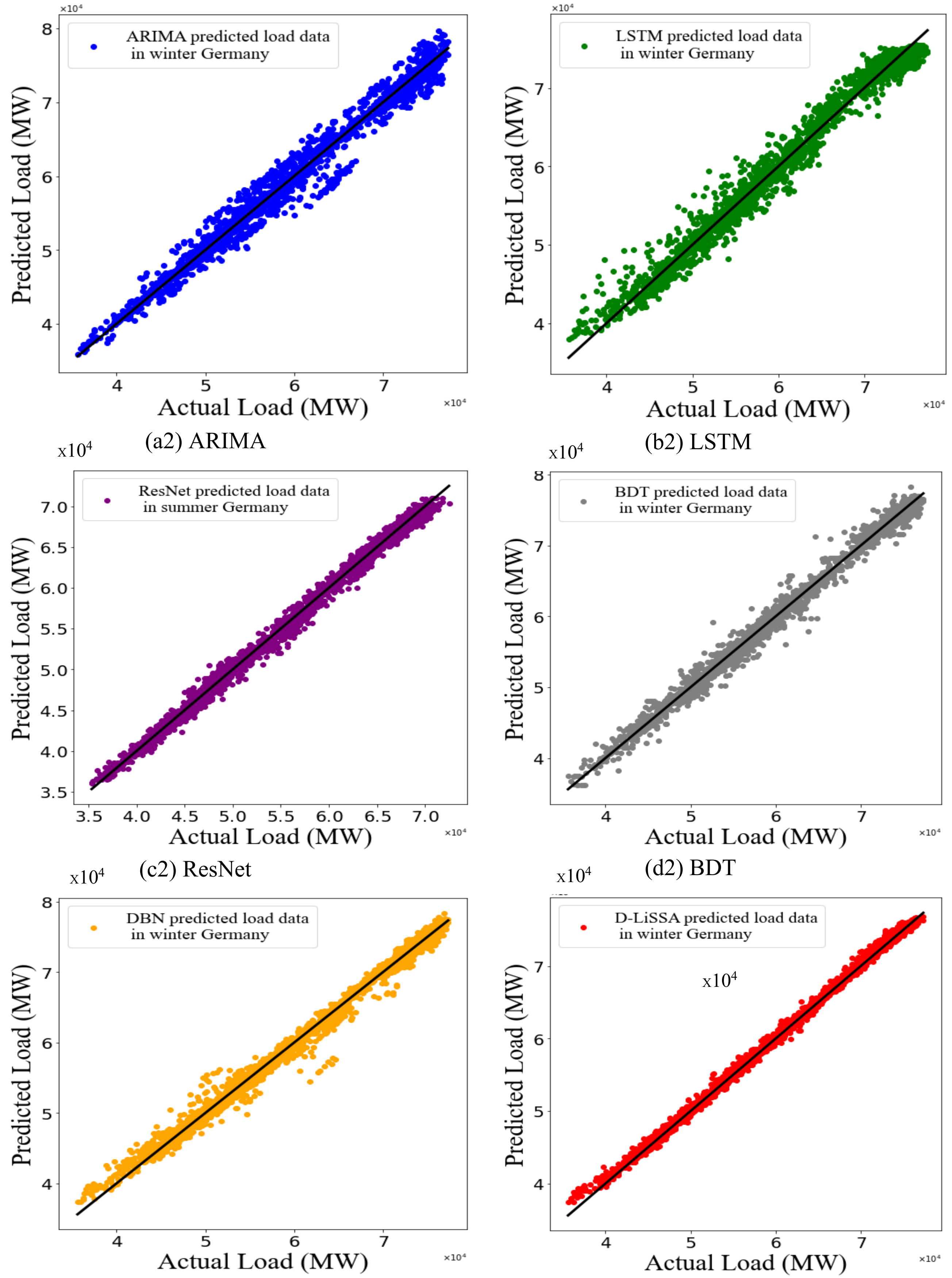
(d1) BDT



(e1) DBN



(f1) D-LiSSA



(a2) ARIMA (b2) LSTM

(c2) ResNet (d2) BDT

(e2) DBN (f2) D-LiSSA

Fig. 7. Actual and predicted load data by D-LiSSA and other comparison methods for Germany in Summer (from a1 to f1), Winter (from a2 to f2): (a1 and a2) ARIMA; (b1 and b2) LSTM; (c1 and c2) ResNet; (d1 and d2) BDT; (e1 and e2) DBN; (f1 and f2) D-LiSSA, respectively.

ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China under Grant 61876066; Guangdong Province Science and Technology Plan Project (Collaborative Innovation and Platform Environment Construction) 2019A050510006; Brunel University London, UK BRIEF Funding; Department of Finance and Education of Guangdong Province 2016 [202]: Key Discipline Construction Program, China; and the Education Department of Guangdong Province: New and Integrated Energy System Theory and Technology Research Group [Project Number 2016KCXTD022].

REFERENCES

[1] Z. Zhao, R. Cheng, B. Yan, J. Zhang, Z. Zhang, M. Zhang and L. L. Lai, "A dynamic particles MPPT method for photovoltaic systems under partial shading conditions," *Energy Conversion and Management*, vol. 220, no. 11, pp. 113070, 2020.

[2] X. Wu, C. S. Lai, C. Bai, L. L. Lai, Q. Zhang and B. Liu, "Optimal Kernel ELM and Variational Mode Decomposition for Probabilistic PV Power Prediction," *Energies*, vol. 13, no.14, pp. 3592, 2020.

[3] A. Vaccaro, I. Pisica, L. L. Lai and A. F. Zobaa, "A review of enabling methodologies for information processing in smart grids," *International Journal of Electrical Power and Energy Systems, Elsevier*, vol. 107, pp. 516–522, 2019.

[4] C. S. Lai, L. L. Lai and Q.H. Lai, "Smart City," in *Smart Grids and Big Data Analytics for Smart Cities. Springer, Cham*, 2021, [Online]. Available: https://doi.org/10.1007/978-3-030-52155-4_1

[5] A. C. Croonenbroeck and G. Stadtmann, "Renewable generation forecast studies – Review and good practice guidance," *Renewable and Sustainable Energy Reviews*, vol. 108, pp. 312-322, 2019.

[6] J. Luo, T. Hong and S. Fang, "Robust regression models for load forecasting," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5397-5404, 2019.

[7] X. Liye, S. Wei, W. Chen, Z. Kequan and L. Haiyan, "Research and application of a hybrid model based on multi-objective optimization for electrical load forecasting," Applied Energy, vol. 180, pp. 213-233, 2016.

[8] Y. Lin, U. Kruger, J. Zhang, Q. Wang, L. Lamont and L. E. Chaar, "Seasonal analysis and prediction of wind energy using random forests and ARX model structures," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 5, pp. 1994-2002, 2015.

[9] E. Erdem and J. Shi, "ARMA based approaches for forecasting the tuple of wind speed and direction," *Applied Energy*, vol. 88, pp. 1405-1414, 2011.

[10]H. Liu, H. Tian and Y. Li, "An EMD-recursive ARIMA method to predict wind speed for railway strong wind warning system," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 141, pp. 27-38, 2015.

[11]Ü. Ç. Büyükşahin and Ş. Ertekin, "Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition," *Neurocomputing*, vol. 361, pp. 151-163, 2019.

[12]M. Zhang, H. Bao, L. Yan, J. Cao and J. Du, "Research on processing of short-term historical data of daily load based on Kalman filter, " *Power System Technol*, pp.10-9, 2003.

[13]H. Sheng, S. Member, J. Xiao, Y. Cheng, S. Member and Q. Ni, "Short-term solar power forecasting based on weighted Gaussian process regression," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 1, pp. 300–308, 2017.

[14]L. Ruiz, M. Cuéllar, M. Calvo-Flores and M. Jiménez, "An application of non-linear autoregressive neural networks to predict energy consumption in public buildings," *Energies*, vol. 9, no. 9, 2016.

[15]A. Tanveer and C. Huanxin, "Nonlinear autoregressive and random forest approaches to forecasting electricity load for utility energy management systems," *Sustainable Cities and Society*, vol. 45, pp. 460-473, 2019.

[16]H. Nie, G. Liu, X. Liu and Y. Wang, "Hybrid of ARIMA and SVMs for short-term load forecasting," *Energy Procedia*, vol. 16, pp. 1455–60, 2012.

[17] Y. Tao, F. Zhao, H. Yuan, C. S. Lai, Z. Xu, W. Ng, R. Li, X. Li and L. L. Lai, "Revisit Neural Network based Load Forecasting," *Proceedings of 20th International Conference on Intelligent System Application to Power Systems*, pp. 10-14, 2019.

[18]B. Huang, D. Wu, C. S. Lai, X. Cun, H. Yuan, F. Xu, L. L. Lai and K. F. Tsang, "Load forecasting based on deep long short-term memory with consideration of costing correlated factor," *Proceedings of 2018 IEEE 16th International Conference on Industrial Informatics*, 2018.

[19]M. Cai, M. Pipattanasomporn and S. Rahman, "Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques," *Applied Energy*, vol. 236, pp. 1078-1088, 2019.

[20]E. Ceperic, V. Ceperic and A. Baric, "A strategy for short-term load forecasting by support vector regression machines," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4356–4364, Nov. 2013.

[21]Y. Bodyanskiy, S. Popov and T. Rybalchenko, "Multilayer neuro-fuzzy network for short-term electric load forecasting," *Proceedings of the 3rd International Conference on Computer Science*, vol. 5010, pp. 339–48, 2008.

[22]C. Cecati, J. Kolbusz, P. Różycki, P. Siano and B. M. Wilamowski, "A novel RBF training algorithm for short-term electric load forecasting and comparative studies," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 10, pp. 6519-6529, 2015.

[23]X. Kong, C. Li, F. Zheng and C.Wang, "Improved deep belief network for short-term load forecasting considering demand-side management," *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1531-1538, 2019.

[24]L. M. Konila Sriram, M. Gilanifar, Y. Zhou, E. Erman Ozguven and R. Arghandeh, "Causal Markov Elman network for load forecasting in multinetwork systems," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 2, pp. 1434-1442, 2019.

[25]W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841-851, 2019.

[26]S. Ryu, J. Noh and H. Kim, "Deep neural network-based demand side short term load forecasting," *Energies*, vol. 10, no. 1, p. 3, 2016.

[27]K. Chen, K. Chen, Q. Wang, Z. He, J. Hu and J. He, "Short-term load forecasting with deep residual networks," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3943-3952, 2019.

[28]T. Ouyang, Y. He, H. Li, Z. Sun and S. Baek, "Modeling and forecasting short-term power load with copula model and deep belief network," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 2, pp. 127-136, 2019.

[29]Y. Zhang and H. Chiang, "Enhanced ELITE-load: A novel CMPSOATT methodology constructing short-term load forecasting model for industrial applications," *IEEE Transactions on Industrial Informatics*, *doi: 10.1109/TII.2019.2930064*

[30]S. Wen, R. Hu, Y. Yang, T. Huang, Z. Zeng and Y. Song, "Memristor-based echo state network with online least mean square," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 9, pp. 1787-1796, 2019.

[31]A. Shahbazia, J. Aghaeia, S. Pirouzib, M. Shafie-khahc and J. P.S.Catalão, "Hybrid stochastic/robust optimization model for resilient architecture of distribution networks against extreme weather conditions," *International Journal of Electrical Power & Energy Systems*, vol. 126, pp. 106576, 2021.

[32]T. Li, B. Wang, M. Zhou and J. Watada, "Short-term load forecasting using optimized LSTM networks based on EMD," [Online]. Available: *https://arxiv.org/abs/1809.10108*

[33]N. Sinha, L. L. Lai, P. K. Ghosh and Y. Ma, "Wavelet-GA-ANN based hybrid model for accurate prediction of short-term load forecast," *Proceedings of the International Conference on Intelligent Systems Applications to Power Systems*, 2007.

[34]Y. Xie, P. Hu, N. Zhu, F. Lei and Q. Sun, "A hybrid short-term load forecasting model and its application in ground source heat pump with cooling storage system," *Renewable Energy*, vol. 161. Pp. 1244-1259, 2020.

[35]Z. J. liang, W. Y. Ming, L. D. Zhi, T. Z. Fu and Z. J. Hua, "Short term electricity load forecasting using a hybrid model," *Energy*, vol. 158, pp. 774-481, 2018.

[36]G. F. Fan, L. L. Peng and W. C. Hong, "Short term load forecasting based on phase space reconstruction algorithm and bi-square kernel regression model," *Applied Energy*, vol. 224, pp 13-33, 2018.

[37]H. Hua, Y. Qin, C. Hao and J. Cao, "Stochastic optimal control for energy internet: A bottom-up energy management approach," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1788-1797, 2019.

[38]M. Tucci, E. Crisostomi, G. Giunta and M. Raugi, "A multi-objective method for short-term load forecasting in European countries," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3537-3547, 2016.

[39]T. K. Chau, S. S. Yu, T. Fernando, H. H. Iu and M. Small, "A load-forecasting-based adaptive parameter optimization strategy of STATCOM using ANNs for enhancement of LFOD in power systems," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 6, pp. 2463-2472, June 2018.

[40]F. Y. Xu, X. Cun, M. Yan, H. Yuan, Y. Wang and L. L. Lai, "Power market load forecasting on neural network with beneficial correlated regularization," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 5050-5059, 2018.

[41]M. Q. Raza, M. Nadarajah, J. Li and K. Y. Lee, "Multivariate ensemble forecast framework for demand prediction of anomalous days," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 1, pp. 27-36, 2020.

[42]C. S. Lai, Y. Yang, K. Pan, J. Zhang, H. Yuan, Wing 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 Transactions on Power Systems (Early Access)*, 03 Dec 2020, *doi: 10.1109/TPWRS.2020.3042389*

[43]J. Cordova, L. M. Konila Sriram, A. Kocatepe, Y. Zhou, E. E. Ozguven and R. Arghandeh, "Combined electricity and traffic short-term load forecasting using bundled causality engine," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 9, pp. 3448-3458, 2019.

[44]X. Tang, Y. Dai, Q. Liu, X. Dang and J. Xu, "Application of bidirectional recurrent neural network combined with deep belief network in short-term load forecasting," *IEEE Access*, vol. 7, pp. 160660-160670, 2019.

[45]M. R. Kazemzadeh, A. Amjadian and T. Amraee, "A hybrid data mining driven algorithm for long term electric peak load and energy demand forecasting," *Energy*, vol. 204, pp. 117948, 2020.

[46]G. Hafeez, K. S. Alimgeer, I. Khan, "Electric load forecasting based on deep learning and optimized by heuristic algorithm in smart grid," *Applied Energy*, vol. 269, pp. 114915, 2020.

[47]Z. Cao, C. Wan, Z. Zhang, F. Li and Y. Song, "Hybrid ensemble deep learning for deterministic and probabilistic low-voltage load forecasting," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 1881-1897, May 2020, doi: 10.1109/TPWRS.2019.2946701.

[48]P. Singh and P. Dwivedi, "A novel hybrid model based on neural network and multi-objective optimization for effective load forecast," *Energy*, vol. 182, pp. 606-662, 2019.

[49]T. Wang, W. W. Y. Ng, M. Pelillo and S. Kwong, "LiSSA: Localized stochastic sensitive autoencoders," *IEEE Transactions on Cybernetics*, doi: 10.1109/TCYB.2019.2923756.

[50]C. S. Lai, Z. Mo, T. Wang, H. Yuan, W.W. Ng and L. L. Lai, "Load forecasting based on deep neural network and historical data augmentation," *IET Generation, Transmission & Distribution, 2020*, doi: 10.1049/iet-gtd.2020.0842.

[51]ENTSO-E. Data portal. [Online]. Available: https://www.entsoe.eu/data/data-portal/consumption/. Accessed on March 5, 2020.

[52]T. Ahmad, et al., "Supervised based machine learning models for short, medium and long-term energy prediction in distinct building environment," *Energy*, vol. 158, pp. 17-32, 2018.

[53]S. Iain and P. Stefan, "Using bias-corrected reanalysis to simulate current and future wind power output," *Energy*, vol. 114, pp. 1224-1239, 2016.

[54] The World Bank. [Online]. Available: https://data.worldbank.org/. Accessed on Feb 26, 2020.
[55] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of International Conference for Learning Representations*, 2015.
[56] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin and J. Shelby, "The National Solar Radiation Data Base (NSRDB)," *Renewable and Sustainable Energy Reviews*, vol. 89, pp: 51-60, 2018.
[57] D. Jager and A. Andreas, "NREL National Wind Technology Center (NWTC): M2 Tower; Boulder, Colorado (Data)," *National Renewable Energy Laboratory*, 2014. [Online]: https://dx.doi.org/10.7799/1052222.

Appendix A: Description of localized stochastic sensitivity autoencoder (LiSSA).

For training an autoencoder (AE), a training dataset D with M samples $\{X_b \in R^{1\times n}\}$, is given from the problem domain where X_b denotes a n-dimensional input vector of the b^{th} training sample. The aim of AE is to learn a better hidden representation (learned feature) to minimize reconstruction error between input X_b and output $\tilde{X}_b \in R^{1\times n}$, as shown in Fig. 1. Usually, an AE is trained by minimizing the mean square error (MSE) between inputs and outputs of the AE as follows:

$$\begin{aligned} R_{emp} &= \frac{1}{M}\sum_{b=1}^{M}\left\|X_b - \tilde{X}_b\right\|^2 \\ &= \frac{1}{Mn}\sum_{b=1}^{M}\sum_{j=1}^{n}\left(\alpha_j(X_b) - X_{bj}\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n} R_{emp}(j) \end{aligned} \tag{A.1}$$

where $\alpha_j(\cdot)$ and X_{bj} denote the output of the j^{th} hidden neuron and the j^{th} input feature of the b^{th} training sample, respectively.

The training dataset is expected to be representative, therefore in general unseen samples should not deviate too much from training samples. And LiSSA is an AE that optimizes weight matrices and bias vectors via minimization of localized perturbation error (LPE) for a given training dataset. Thus, the LPE measures the sensitivity of the model to unseen samples that are similar to training samples. The difference of between an unseen sample X and its corresponding training sample X_b is defined as $\Delta X = (\Delta x_1, \Delta x_2, \cdots, \Delta x_n)^T = X - X_b$, where n denotes the number of input features, Δx_i denotes the perturbations to the i^{th} input feature, the output vector of an AE consists of n outputs x_{bk}, and $|\Delta x_i| \leq Q, i = 1, 2, \cdots, n$. When Q value is given, we define the Q neighborhood of X_b is $S_Q(X_b) = \{X | X = X_b + \Delta X\}$. Then, the LPE of the k^{th} $S_Q(X_b)$ is defined as follows:

$$R(x_{bk}, Q) = \int_{S_Q(X_b)}\left(x_{bk} - \alpha_k(X_b)\right)^2 P(X)dX \tag{A.2}$$

where $\alpha_k(\cdot)$ and $P(X)$ denote the k^{th} output unit of LiSSA and the unknown probability density function of X in $S_Q(X_b)$, respectively. The LPE of all outputs of the AE for a given training dataset of M samples is as follows:

$$R_{LPE}(X, Q) = \frac{1}{Mn}\sum_{b=1}^{M}\sum_{k=1}^{n} R(x_{bk}, Q) \tag{A.3}$$

The Hoeffdings inequality is applied to Eq. (A.3), with a probability of $1-\eta$, the LPE is given as:

$$R_{LPE}(X, Q) \leq \left(R_{emp}(k) + E((\Delta\alpha_k)^2)\right)^2 + \varepsilon \tag{A.4}$$

where $R_{emp}(k)$, $E((\Delta\alpha_k)^2)$, and $\varepsilon = B\left(\sqrt{\ln\eta / -2M}\right)$ denote the empirical error, the stochastic sensitivity (SS), and the confidence of the upper bound, respectively. Therefore, the LPE includes three parts namely, 1) the training error; 2) the SS; and 3) constants defined by the training dataset.

The training error is the reconstruction error between inputs and outputs of the LiSSA, namely training mean square error (MSE)

$$R_{emp} = \frac{1}{Mn}\sum_{b=1}^{M}\sum_{k=1}^{n}(\alpha_k(X_b) - x_{bk})^2 \tag{A.5}$$

where x_{bk} denotes the k^{th} output feature of the b^{th} training sample.

The expectation of squared differences between the outputs of the training samples and their corresponding perturbed samples (i.e., $X_b + \Delta X$) is defined as the SS, the SS is given as follows:

$$E((\Delta\alpha_k)^2) = \frac{1}{Mn}\sum_{b=1}^{M}\sum_{k=1}^{n} E\left[\left(\alpha_k(X_b + \Delta X) - \alpha_k(X_b)\right)^2\right] \tag{A.6}$$

By adopting the concept of the Monte Carlo method to compute the SS. The Algorithm A.1 shows the computation process of the SS by using a uniform random sampling.

If we neglect constant term in Eq. (A.4), the major components affecting $R_{LPE}(X,Q)$ are $R_{emp}(k)$ and $E((\Delta\alpha_k)^2)$. Therefore, the final objective function of LiSSA is:

$$R_{LPE} = R_{emp} + E((\Delta\alpha_k)^2) \tag{A.7}$$

Algorithm A.1 Computation of SS

Input: Q, H and $\alpha(\cdot)$

Output: The SS value

1: Generate H uniformly distributed random points $\Delta x_h \in \mathbb{R}^n, h = 1, \dots, H$ with each coordinate range from$[-Q, Q]$;

2: For each training sample X_b, compute each outputs SS:

$$\delta(X_i) = \frac{1}{Hn}\sum_{h=1}^{H}\sum_{k=1}^{n}\left(\alpha_k(X_b + \Delta X_h) - \alpha_k(X_b)\right)^2$$

3: Compute the SS of the whole LiSSA:

$$E((\Delta\alpha_k)^2) = \frac{1}{M}\sum_{b=1}^{M}\delta(X_b)$$
