This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2933009, IEEE Transactions on Industrial Informatics

A Novel Electricity Price Forecasting Approach Based on Dimension Reduction Strategy and Rough Artificial Neural Networks

Hamidreza Jahangir, Hanif Tayarani, Sina Baghali, Ali Ahmadian, Ali Elkamel,
Masoud Aliakbar Golkar, *Senior Member*, *IEEE*, and Miguel Castilla

Abstract**—** An accurate Electricity Price Forecasting (EPF) plays a vital role in the deregulated energy markets and has a specific effect on optimal management of the power system. Considering all the potent factors in determining the electricity prices - some of which have stochastic nature - makes this a cumbersome task. In this paper, first, Grey Correlation Analysis (GCA) is applied to select the effective parameters in EPF problem and eliminate redundant factors based on low correlation grades. Then, a deep neural network with Stacked Denoising Auto-Encoders (SDAEs) has been utilized to denoise data sets from different sources individually. After that, to detect the main features of the input data and putting aside the unnecessary features, Dimension Reduction (DR) process is implemented. Finally, the rough structure Artificial Neural Network (ANN) has been executed to forecast the day-ahead electricity price. The proposed method is implemented on the data of Ontario, Canada, and the forecasted results are compared with different structures of ANN, Support Vector Machine (SVM), Long Short-Term Memory (LSTM) - benchmarking methods in this field- and forecasting data reported by Independent Electricity System Operator (IESO) to evaluate the efficiency of the proposed approach. Furthermore, the results of this study indicate that the proposed method is efficient in terms of reducing error criterion and improves the forecasting error about 5 to 10 percent in comparison with IESO. This is a remarkable achievement in EPF field.

Index Terms**—Price forecasting; Dimension reduction; Deep learning; Rough neuron; Denoising.**

NOMENCLATURE

Parameters

n_0	Total number of input data components
m	Total number of input data in GCA
n	Total kind of input data parameters in GCA
N^s	Total number of neurons in layer *S*
N^l	Total number of hidden layers
ξ	Distinguishing factor in GCA

Variables

b	Bias vector for visible layer
b'	Bias vector for visible layer transposed
b_U^S	Upper bound bias vector for layer *S*
b_L^S	Lower bound bias vector for layer *S*
b_{LR}	Bias vector of Logistic Regression layer
CF	Compression Factor in dimension reduction
CL	Number of compression layers in dimension reduction
E	Total Sum square error
$E(k)$	Total Sum square error in iteration *k*
f	Activation function
f^S	Activation function for layer *S*
H	Hidden layer vector
H^S	Hidden layer vector for layer *S*
O^S	Output of layer *S*
O_U^S	Output of upper bound neuron for layer *S*
O_L^S	Output of lower bound neuron for layer *S*
W_U^S	Weight vector of upper bound neurons in layer *S*
W_L^S	Weight vector of lower bound neurons in layer *S*
W	Weight vector between visible and hidden layer
W'	Weight vector between visible and hidden layer transposed
W_{LR}	Weight vector of Logistic Regression layer
X	Input data vector
X_i	Input data vector for sample *i*
X'	Reconstructed data vector
$\hat{X}$	Contaminated data vector
Y_g	Output vector for sample *g*
$\hat{Y}_g$	Desired vector for sample *g*
$\hat{Y}_{g_{mean}}$	Mean value of desired vector in calculating sample
$\lambda_o{}^*(t)$	Normalized target data for at time *t* for GCA
$\lambda_z(t)$	Data for *z-th* data sample at time *t* for GCA
$\lambda_z{}^*(t)$	Normalized Data for *z-th* data sample at time *t* for GCA
α^S	Coefficient of upper bound neuron for layer *S*
β^S	Coefficient of lower bound neuron for layer *S*
γ	Momentum coefficient
ε_S	Weight variations for fine-tuning in layer *S*
η	Training coefficient
ψ	Random noise operator
$\pi_z(\lambda_o{}^*(t),\lambda_z{}^*(t))$	Grey coefficient between sequence $\lambda_o{}^*(t),\lambda_z{}^*(t)$
$\Gamma_z(\lambda_o{}^*(t),\lambda_z{}^*(t))$	Grey correlation grade between sequence $\lambda_o{}^*(t),\lambda_z{}^*(t)$

Indices

i	Index of visible layer sample
g	Index of output layer sample
k	Index of iteration number
S	Index of layer number
t	Index of input data time sample in GCA
z	Index of input data type in GCA

Hamidreza Jahangir, Hanif Tayarani, Sina Baghali, and Masoud Aliakbar Golkar are with the Faculty of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran (Emails: h.r.jahangir@email.kntu.ac.ir ; baghalisina@gmail.com; haniftayarani@email.kntu.ac.ir; Golkar@kntu.ac.ir).
Ali Ahmadian is with the Faculty of Engineering, University of Bonab, Bonab, Iran and the College of Engineering, University of Waterloo, Waterloo, ON, Canada, (Email: ahmadian@bonabu.ac.ir)

Ali Elkamel is with the College of Engineering, University of Waterloo, Waterloo, ON, Canada, and the College of Engineering, Khalifa University of Science and Technology, The Petroleum Institute, Abu Dhabi, United Arab Emirates (Email: aelkamel@uwaterloo.ca)

Miguel Castilla is with the Department of Electronic Engineering, Technical University of Catalonia, Spain (e-mail: miquel.castilla@upc.edu)

I. Introduction

A. Background and motivation

Deregulating energy markets in the late 1900s revolutionized the electricity pricing and introduced a sophisticated competitive market in which the Electricity Price (EP) changes momentarily. As a result of the EP intermittent behavior, the short-term EP forecasting with high precision is very much needed in the optimal management of the power system [1]. The economic benefits of an accurate Electricity Price Forecasting (EPF) cannot be neglected. As well, slightest improvement in the EPF accuracy would save millions of dollars for the industry. Determining the electricity price is contingent upon several parameters such as climate conditions (wind speed, temperature, precipitation, etc.) and consumption patterns (peak hours, weekdays, seasonal attributes, etc.) therefore, EPF should be seen as a "Big Data" problem [2].

To analyze this large volume of data, a strong data mining tool is necessary for extracting the main features of the input data, and data mining approaches are noteworthy these days in various applications [3].

B. Literature survey

Due to the critical importance of EPF, this subject has been under thorough studies since the deregulation of the energy market. These studies can be categorized into three general subsections: statistical methods, probabilistic methods and Computational Intelligence (CI) based approaches [4].

Generally, the statistical methods, are based on assumptions introduced by each algorithm which is a source of inaccuracy and adds deficiency to the model. Moreover, statistical methods perform poorly dealing with price spikes in the electricity markets [4]. Recently, researchers have been interested in hybrid approaches, combining statistical methods with other forecasting methods in order to limit the shortcomings of these procedures [5]–[7].

The probabilistic methods are other kinds of approaches in EPF studies. These methods usually employ probability distribution functions in the forecasting procedure. Due to the high uncertainty of EP series, the main advantage of these techniques is finding optimal prediction intervals (lower and higher bounds) in final results [4]. In probabilistic forecasting methods, the combination of different approaches such as Wavelet function [8], spatial interpolation [9], active learning [10] are considered, and this improves the accuracy of the forecasting results [11], [12].

CI gathers fundamental theories of learning, evolution, and fuzziness to create mechanisms that can cope with complex dynamic systems. Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), fuzzy systems, and evolutionary computation are the main modules of CI methods. CI methods are flexible and able to analyze problems with complex non-linear behaviors [13]. These features make CI a promising tool for solving short-term forecasting and presents an excellent performance in EPF [14]. ANN is one of the most popular CI methods that can forecast the expected outcomes solely by processing the known data. Besides, ANN does not require baseline hypotheses to generalize the validity of the models, and thereby additional inaccuracies are avoided [15]. Conventional ANNs consist of a few hidden layers that are defined as shallow networks [16], which got developed over the years from simple feed-forward to recurrent and radial base structures with multi-layers of neurons. Most of the studies in EPF applied conventional shallow structures of ANNs with low hidden layers [17]. In recent years, the concept of multilayer ANNs and Deep Learning (DL) with Stacked Auto-Encoders (SAEs) gave rise to novel studies with more sophisticated, better performing ANNs [8], [14]. The effects of a DL approach in EPF is elaborated in [18]. The main shortcoming of these approaches is that they are trained by the raw data and no deliberation is done on the input data; thus, ignoring the input noise results in defective forecasts. Stacked Denoising Auto-Encoders (SDAEs) became handy in this matter because they can efface the noise of the input data [14], [19].

According to the large dimension of input data, feature extraction methods have been considered in CI studies. The Principal Component Analysis (PCA) is the most common method for feature extraction task which omits the redundant features with mapping data in linear space [20]. Kernel PCA is a newer feature extracting method which works in non-linear space and has better accuracy [21]. As stated by Hinton [22], the feature extraction methods based on deep auto-encoders outperform the PCA based approaches.

To sum up this discussion, the CI-based approaches and data-driven techniques are the most effective methods in EPF problems with large dimension input data [18].

As far as the authors know, some developments are missing from the studies in the field of ANN-based EPF which can be categorized as follows:

- Data mining methods which have acceptable performance in feature extraction of input data have been rarely applied.
- ANNs with the rough structure which have high ability in handling uncertainties is not employed.

C. Paper contribution

As mentioned, EPF studies involve high dimensional inputs, and the SAEs are primarily used for data de-noising. On the other hand, SAEs can be employed for reducing the volume of data to extract the main features of the input data, as well as to reduce the deficiencies caused by the noise. In the feature extraction task, the input data has been mapped with a nonlinear operator to a new space with a lower dimension which consists of the main features. In this study, firstly, a large number of different input data variables are considered, to select the effective parameters, the Grey Correlation Analysis (GCA) is employed. Then, SDAEs are applied to reduce input data noise, and after that, a Dimension Reduction (DR) strategy with optimal compression rate is applied to extract crucial features of the input data. Indeed, the GCA and DR methods are employed for feature selection and extraction, respectively. The main scope of this study is an hourly day-ahead EPF which is required in electricity market studies, but this methodology can be extended for EPF studies with wider time spans as well. The main objective of data reduction in this study is to provide a forecasting method with better accuracy. According to high fluctuations in EP profile, and to verify the spike points in a better way, a Sinusoidal Rough Artificial Neural Network (SR-ANN) is implemented which is more responsive to rapid variations [23]. Rough neurons, by employing interval weights handle the EP uncertainty with more precision than conventional neurons.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2933009, IEEE Transactions on Industrial Informatics

In this study, implementing GCA for selecting important input data parameters and optimal DR strategy for feature extraction approach with rough neurons to address the uncertainty are utilized altogether in EPF problem for the first time. In fact, in this paper, the potential of the feature selection, feature extraction and forecasting procedure with rough neuron-based ANN are employed together to develop the robust forecasting method in EPF task. Indeed, in this study, we introduce hybrid method with high precision in electricity price forecasting based on various techniques, which are effective in forecasting task for a specific goal. The following is the list of applied techniques and their application in this study.

- GCA: preprocessing feature selection.
- Rough neurons: handling uncertainty.
- DR: robust feature extraction task with optimal extraction rate.

The main contributions of this study can be enumerated as follows:

- The DR method based on deep auto-encoders is implemented to improve the forecasting results effectively by extracting the superfluous features of input data.
- The optimal structure of the DR method is determined according to the data from the case study.
- The EPF is considered as a "big data" problem with different kinds of parameters as input data, and effective parameters are selected based on the GCA method.
- A deep rough structure ANN is employed in order to reach more accurate modeling of the EP with its uncertainties.

D. Organization of the paper

The main body of this paper is defined as follows: Section II gives a brief review of the implemented methods in this paper. In Section III, a comprehensive description of the proposed method is provided. In Section IV, numerical results and case study details are presented. Finally, Section V summarizes the findings of this study.

II. A Brief Review of the Basic Methodologies

A. Deep Learning Method

DL is basically an ANN with multi hidden layers in which the input data is assessed more thoroughly -compared to the shallow structures of ANNs- in order to perform better in feature extraction of the data. In rudimentary ANNs, all the features are expected to be dug out in a single layer of artificial neurons, while in a deep structure, each layer detects different features of the input data. Indeed, DL allows the ANNs, which are constructed with multi hidden layers, to learn the representation of the features with multifarious level of extractions. DL has proved its superiority in large dimension studies such as image and speech recognition, diverse forecasting studies, etc. In this study, DL method is employed in EPF which is a large dimension study and has a high dynamic and intermittent behavior. Therefore, a strong feature extraction tool is needed to find the accurate forecasting result. More details about implementing the deep network training procedure are expressed by pseudo-code in Appendix section, part A.

B. Dimension Reduction

Nowadays, the availability of data in large scales provides the opportunity of accessing a myriad of useful information for researchers. This seemingly beneficial advantage makes every database study overwhelming and is the main impediment for time sufficient simulations. DR is presented as a solution to curtail the size of data into a smaller set. Reducing data using deep ANNs was introduced by Hinton [22] utilizing SAEs. SAEs have a high potential in feature extraction task and present an acceptable performance in various forecasting tasks. In DR method, the main goal is to eliminate the redundant features of input data, which have low efficiency in forecasting the target data. In this procedure, first, all AEs are trained separately, then, stacked together, and finally, the fine-tuning procedure is applied to the holistic network. The core objective of this method is to produce a low dimension data, containing the main features of the data set. As a matter of fact, reducing data is shown to have better accuracy in the results to a certain extent. This feature will be discussed thoroughly in the upcoming sections.

C. Stacked Denoising Auto-encoders

AEs are known to be the simple models of ANNs that get input data, rebuild them by encoding and decoding, and convert them to the output with the minimum discrepancy [22]. In this approach, SAE is engendered by fixing AEs and deep ANN. Discriminating significant features of the input data is the main duty of SAEs. Concerning the robustness of SAEs, the Stacked Denoising Auto-encoder (SDAE) adds noise to the data in the midst of the procedure. Input data should be regained by SDAEs from the contaminated data which enhances the ability of feature extraction in this technique. In this study, various parameters from different sources are considered as input data; denoising task is necessary work in these situations. In fact, the denoising process helps us to make a robust approach in confronting different kinds of input data which may have noise or bad data.

D. ANN-based on the Rough Structure

A simple definition of a rough neuron is a neuron created by pairing a couple of neurons called the upper bound and lower bound. Prior works confirm that the rough ANNs (R-ANNs) which are used to consider the uncertainties of input data, perform better than any other structures. G. Ahmadi et al. in [23] indicate that in many fields such as traffic volume prediction, reducing image noise and medical diagnostic support system, R-ANN is highly effective. In this research, an ANN with a rough structure has been applied accordingly to consider the uncertainty of input data. The rough neurons, with their interval weights configuration, handle the uncertainty in an acceptable way and affect the results significantly. This effect will be illustrated in the numerical results.

III. Proposed Methodology

The proposed method consists of four parts, which are described in the following sections.

A. *Selecting Input Data with GCA*

The GCA is a tool for determining the correlation grade between different input and target data which has an acceptable performance in EPF problem [21]. In this study, the input data structure for EPF is determined based on GCA results, and the parameters with low grey correlation grades are neglected.

Matrix D is defined based on the input data as follows:

$$D = \begin{bmatrix} \lambda_1(1) & \cdots \lambda_z(1) & \cdots & \lambda_n(1) \\ \vdots & \vdots & & \vdots \\ \lambda_1(t) & \lambda_z(t) & & \lambda_n(t) \\ \vdots & \vdots & & \vdots \\ \lambda_1(m) & \cdots \lambda_z(m) & \cdots & \lambda_n(m) \end{bmatrix} \quad (1)$$

The rows define the time sample and columns define the input data index [21].
To apply GCA, input data can be normalized as follows:

$$\lambda_z{}^*(t) = \frac{\lambda_z(t) - min\,\lambda_z(t)}{max\,\lambda_z(t) - min\,\lambda_z(t)} \quad (2)$$

then, the grey coefficient is determined as:

$$\pi_z\left(\lambda_o{}^*(t), \lambda_z{}^*(t)\right) = \frac{\Delta_{min} + \xi\,\Delta_{max}}{\Delta_{oz}(t) + \xi\,\Delta_{max}} \quad \xi \in (0,1) \quad (3)$$

$$\Delta_{oz}(t) = |\lambda_o{}^*(t) - \lambda_z{}^*(t)| \quad (4)$$

$$\Delta_{max} = max\,|\lambda_o{}^*(t) - \lambda_z{}^*(t)| \quad z = 1,..,n \quad (5)$$

$$\Delta_{min} = min\,|\lambda_o{}^*(t) - \lambda_z{}^*(t)| \quad z = 1,..,n \quad (6)$$

where ξ is a distinguishing factor and has the value of 0.5 to avoid the sharp selection in GCA task and keeps the useful features as much as possible [21], [24]. The final grey correlation values for different input data is calculated by:

$$\Gamma_z\left(\lambda_o{}^*(t), \lambda_z{}^*(t)\right) = \frac{\sum_{t=1}^{m} \pi_z\left(\lambda_o{}^*(t), \lambda_z{}^*(t)\right)}{m} \quad (7)$$

B. *Denoising*

SDAE is used in this paper as a pre-training methodology. SDAE which is a type of AE collects its input $X \in [0\,,1]$ and based on the properties of the hidden layers produces the outcome $H \in [0\,,1]$. The coding equivalent of this procedure can be obtained using the following equation:

$$H = f(WX + b) \quad (8)$$

In the encoding section, the output of the previous section is treated as the input and the remodeled form of the input $X' \in [0\,,1]$ is created by maintaining the same dimension as follows:

$$X' = f(W'H + b') \quad (9)$$

Afterward, the difference in the values of the real input and the remodeled one should be minimized. An unsupervised mode is employed in order to achieve this goal. The loss function (L) that needs to be minimized is defined as follow:

$$L(X, X') = \|X - X'\|^2 \quad (10)$$

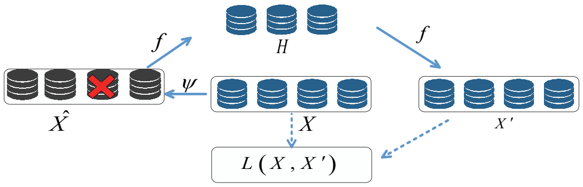


Fig. 1. The procedure of a typical SDAE.

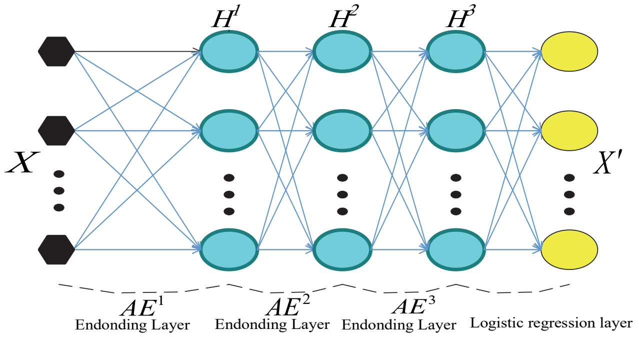


Fig. 2. The structure of an SDAE with three AEs.

The summary of the SDAE process is illustrated in Fig. 1. As shown in this figure, the function ψ stochastically maps the input and adds error to the data. Then, another function transforms the contaminated data into a hidden layer. At last, the reformed version of input is created based on these data. This process casts out bad data and supports SDAES's robustness.
After each independent training of DAEs, all of them are put together and trained by the input data in the supervised mode as shown in Fig. 2. The purpose of this section is just to eliminate input data noise; thus, the input and output data dimension in this section are the same. The loss function for fine-tuning is considered as the following cross-entropy function:

$$L(X, X') = -\sum_{i=1}^{n_0}[X_i \log(\tau_i) + (1 - X_i)\log(1 - \tau_i)] \quad (11)$$

where,

$$\tau_i = \frac{1}{1 + exp(-W_{LR}H + b_{LR})} \quad (12)$$

More information on cross-entropy function is given in [19].

C. *Dimension Reduction*

DR process is illustrated in Fig. 3. This procedure consists of multiple AEs stacked together. Excluding the unnecessary data and extracting the main features are the ultimate goals of the DR which is recommended for all dataset studies, mainly with large dimension input data. In this paper, to prevent overfitting and improving SAE performance, the initial weight of each AE is determined independently by a Restricted Boltzmann Machines (RBMs), based on their proven capacities stated in [25]. Reducing the data is a delicate matter, in order to prevent excessive data loss, the optimal number of compression layers (CLs) and the compression factor (CF) needs to be determined. In this paper, as illustrated in Fig. 3(a), the structure of SAE is achieved by examining the different arrangement of layers with various CFs. The next step is training each AE based on the two consecutive layers. This training is done autonomously for each AE, and in the final fine-tuning, these AEs come together to be trained by the input data in the mini-batch mode as shown in Fig. 3(b). The error that defines the optimal structure of the AEs in DR task is calculated in this state.

D. *Forecasting*

Going through the denoising and DR process, a data set containing the main features of the study is available to estimate future prices. A rough-ANN is implemented in this study to forecast the EP. Fig. 4 illustrates a rough neuron with its main parameters at layer S. Equations defining the relations of these parameters are as follows:

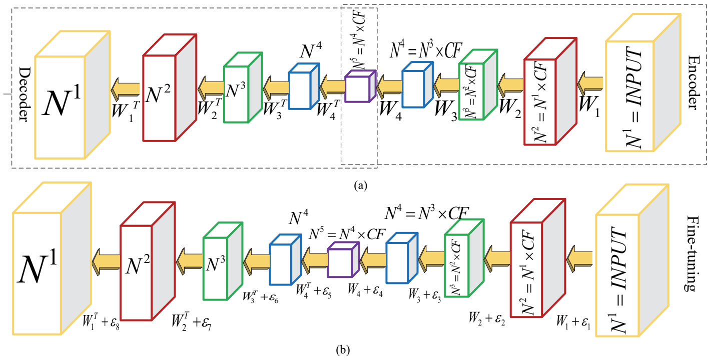


Fig. 3. The overall configuration of the DR strategy:
(a) The training process for auto-encoders. (b) Auto-encoders fine-tuning process.

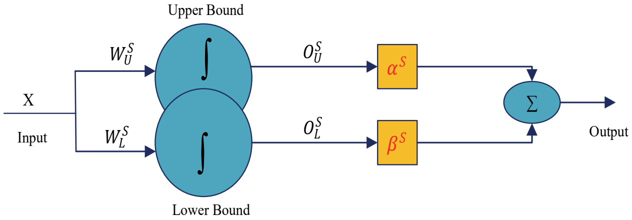


Fig. 4. Rough neuron structure.

$$O_U^S = Max(f^S(W_U^S X + b_U^S), f^S(W_L^S X + b_L^S)) \tag{13}$$

$$O_L^S = Min(f^S(W_U^S X + b_U^S), f^S\,(W_L^S X + b_L^S)) \tag{14}$$

$$O^S = \alpha^S O_U^S + \beta^S O_L^S \tag{15}$$

Normally, sigmoid activation functions are used in ANNs. However, this function can easily get saturated especially dealing with large input values.

Consequently, in this study, a sinusoidal activation function is used to circumvent this saturation. Additionally, this function has better performance in following abrupt changes in the phenomena under study [23]. Back Propagation equations are described as follows:

$$W^S(k) = W^S(k-1) - \eta \frac{\partial E(k-1)}{\partial W^S(k-1)} + \gamma \Delta W^S(k-1) \tag{16}$$

$$\alpha^S(k) = \alpha^S(k-1) - \eta \frac{\partial E(k-1)}{\partial \alpha^S(k-1)} + \gamma \Delta \alpha^S(k-1) \tag{17}$$

$$\beta^S(k) = \beta^S(k-1) - \eta \frac{\partial E(k-1)}{\partial \beta^S(k-1)} + \gamma \Delta \beta^S(k-1) \tag{18}$$

Momentum coefficient γ is defined between 0 and 1.

According to (15), if $f^{S+1}(W_U^{S+1} O^1 + b_U^{S+1}) \geq f^{S+1}(W_L^{S+1} O^1 + b_L^{S+1}))$, the training process of different parameters of the ANN with two rough layers, which is illustrated in Fig. 5, can be seen in the Appendix section, part B. The step by step procedure of the overall proposed methodology is illustrated in Fig. 6. Each step is implemented autonomously, and the results of each section are relegated to the next step. For the final training procedure, all different parts are stacked together. Deep networks are remarkable tools for forecasting problems with high intermittent behavior, but with a large number of training parameters, they have some shortages such as instability and overfitting. To solve this problem, in this study, three commonly used techniques in DL concept such as Mini-Batch Gradient Descent (MBGD), dropout, and L2 regularization algorithms are employed [26], [27].

IV. Numerical Study

A. Data Description

The data used in this paper is based on the EP records of the province of Ontario, Canada [29]. This province gets its required energy from multiple sources that contain sustainable energies, such as wind and solar and etc. As stated in [21], the EPF should be considered as a “Big Data” problem, and many parameters affect this issue.

In this regard, the input data (January 1, 2016 to December 30, 2017 with one-hour time intervals) of different parameters such as: generation power data -nuclear, gas, hydro, wind, solar, biofuel, total generation output power (all with the unit of MW)-, forecasted EP data -Hour 1 pre-dispatch (H1P), Hour 2 pre-dispatch (H2P) and Hour 3 pre-dispatch (H3P) (all are forecasted price for one, two and three hours ahead, respectively with $/MWh unit)- , weather condition data –temperature (℃), dew point temperature (℃) real humidity (%), and load data- market demand and Ontario demand (both of them based on MW) -, are considered in EPF problem. The GCA result is depicted in Fig. 7. Based on the GCA results, the parameters with low grey correlation grades (less than 0.6) including nuclear, hydro, temperature, dew point temperature, real humidity are considered as irrelevant data and have been eliminated in this study. The resultant correlation factors are justifiable from the power system and electricity market point of view as well. Since the hydro and nuclear power units are providing the base load demand, they do not have the same effect on EP as the other power units such as gas units which participate in the peak load demands.

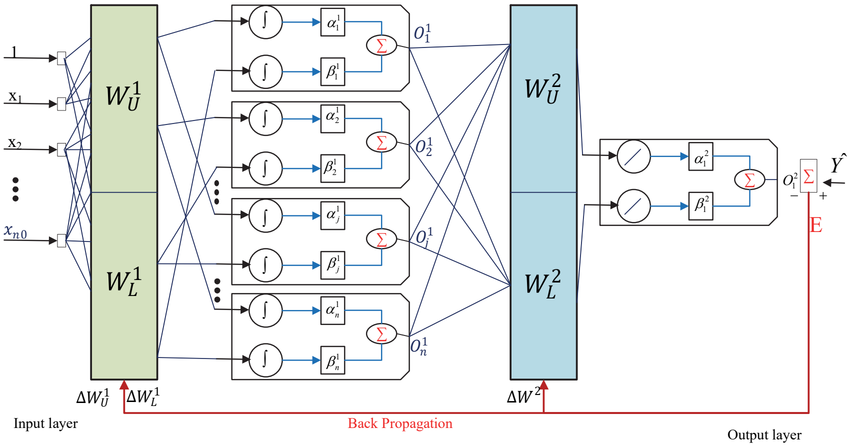


Fig. 5. Proposed rough NN.

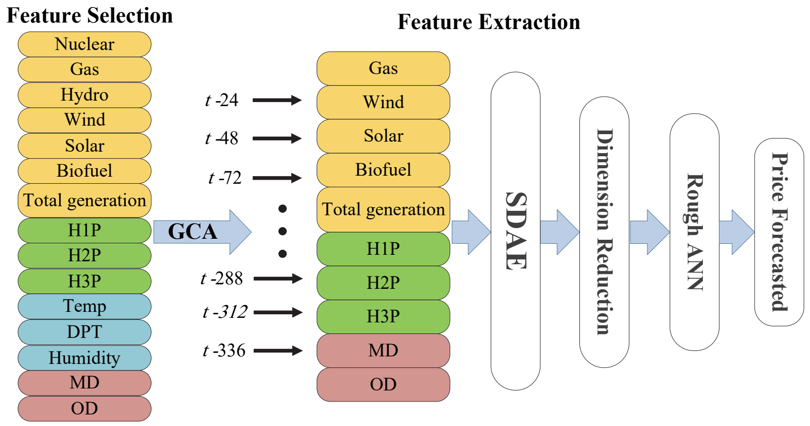


Fig. 6. The overall structure of the proposed methodology.

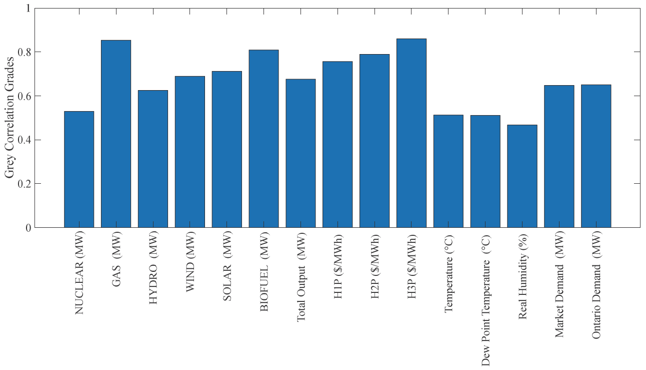


Fig. 7. GCA result.

The parameters such as temperature, dew point temperature, and humidity depend on the regional conditions. Ontario has a cold climate which causes the temperature to be somewhat constant. Therefore, these parameters also play less significant role in the EP. On the other hand, if we had studied data from other regions with a greater variety of climate conditions, the correlation factors of the mentioned parameters might have higher values. In this regard, the EP for each hour of a day in 2018 is forecasted based on the different parameters data - gas, wind, biofuel, total output, H1P, H2P, H3P, market demand, and Ontario demand - of the similar hour in the last 14 days according to the high grey correlation grades with the target data. According to Fig. 6, 140 data is employed for forecasting the EP of each hour.

B. Error Calculation Strategy

In this study well-known error criteria including the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE) [14] are considered as follows:

$$MAE = \frac{1}{n_0}\sum_{g=1}^{n_0}(|\widehat{Y_g} - Y_g|) \tag{19}$$

$$MAPE = \frac{1}{n_0}\sum_{g=1}^{n_0}\left(\left|\frac{\widehat{Y_g}-Y_g}{\widehat{Y_g}_{mean}}\right|\right) \tag{20}$$

$$RMSE = \sqrt{\frac{1}{n_0}\sum_{g=1}^{n_0}(|\widehat{Y_g} - Y_g|)^2} \tag{21}$$

In addition, R-squared, which is the square of the correlation between the target and the forecasted values, is also considered. The R-squared range is [0,1] denoting that a higher value implies a better forecasting result [30].

$$R^2 = 1 - \frac{\sum_{g=1}^{n_0}(Y_g-\hat{Y}_g)^2}{\sum_{g=1}^{n_0}(Y_g-\bar{Y}_g)^2} \tag{22}$$

C. Optimum Dimension Reduction Structure

As mentioned before, one of the main challenges in reducing data is determining the number of compression layers (CL) and the compression factor (CF). In order to decide which structure is the optimum one, four different number of layers were examined with different compression factors. For this purpose, the day-ahead EPF in 2016 and 2017 are employed. The mean value of different error criteria for each structure is presented in Table I. According to this table, the best choice is CF=0.85 for three compression layer (CL=3). As shown in Fig. 8, by increasing the compression (decreasing the CF) the error criteria decrease until they reach the saddle point which is the optimum point and increases afterward. This enhancement of accuracy is the result of omitting unnecessary features and the increments of error criteria after the saddle point means losing the vital features of data. Based on these results, by increasing the number of CLs from 1 layer to 4 layers, the optimum CF has been increased from 0.75 to 0.95. It should be noted that in implementing the DR method, determining the optimal number of CL is very important. In this case, the application of DR with four CLs is not adequate as it significantly increases the error value rate.

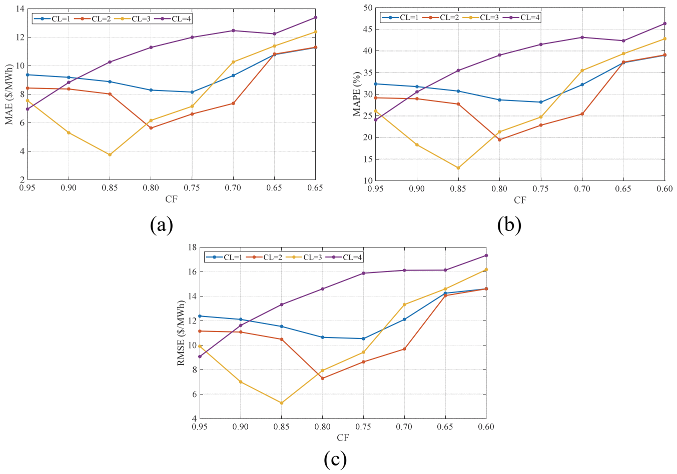
This means that the DR structure must be designed based on the available data set and that increasing the number of compression layers is not always useful. The optimum configurations (for CL=1, 2 and 4) are shown in Fig. 9. Among these four different CLs, according to Table I and Fig.8, the deep ANN had the best performance with three compression layers and CF=0.85. Consequently, in the following, this optimal structure is employed to forecast a day-head EP in 2018. As stated above, data from January 1, 2016, to December 30, 2017, with one-hour time intervals consisting 17520 data is applied for training the proposed method. Regression figures of training (80% of input data with 14016 data set), validation and test (both contain 10% of input data with 1752 data set) processes of this structure are shown in Fig. 10.

D. Forecasting Results

In this study, different types of ANNs (Rough SDAE (R-SDAE), SDAE, Conventional NN (C-NN)), SVM and Long

Short Term Memory (LSTM) [14], [21], as benchmarking data-driven approaches in EPF field, and the hourly forecasted EP by Independent Electricity System Operator (IESO) from Ontario power system [29] are employed in order to evaluate the proposed methodology i.e. the Dimension Reduction Rough SDAE (DR-R-SDAE) with three CLs and optimal CF (0.85). The numerical study is done using MATLAB 9.3 software on a PC with an Intel Core i7, 3.4 GHz CPU and 16 GB of RAM. More details about the tested networks are given as follows:

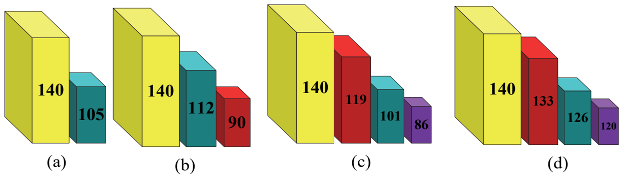
As stated before, according to the pre-training process by the RBMs, for training SDAEs, the smaller learning rate is required, so that the information in the trained weights is not entirely lost. In this paper, RBM and SDAEs training rate are considered as 0.005 and 0.001, respectively. L2 regularization rate, dropout rate (probability of retaining a hidden unit in the network) are defined as 0.001 and 0.6, respectively. Maximum epoch and validation frequency are considered as 1000 and 10, respectively. These training parameters are applied in all of the ANN-based approaches (DR-R-SDAE, R-SDAE, SDAE, and C-NN). In this study, the benchmarking SVM method is considered as a least-squares SVM with sigmoid kernel function which is a powerful tool in EPF studies. Basic super parameters of SVM which have a high effect on SVM performance are adjusted with the cross-validation method. Furthermore, the benchmarking LSTM method as one of the powerful recurrent networks, which has an internal self-looped cell has been considered by 20 hidden layers. For more details about SVM and LSTM training procedure see [21] and [28] respectively. The training procedure of the ANN-based methods is done by stochastic gradient descent method which is much faster than gradient descent method according to its high-frequency updating rate of the training parameters. Stochastic gradient descent attempts to find minimums or maximums by iteration from some randomly picked training examples instead of all the training data set. In this way, the error is typically noisier than gradient descent. Given this, the stochastic gradient descent can escape shallow local minimum points more efficiently in comparison with the gradient descent method and results in a more accurate forecast.



(a) (b)

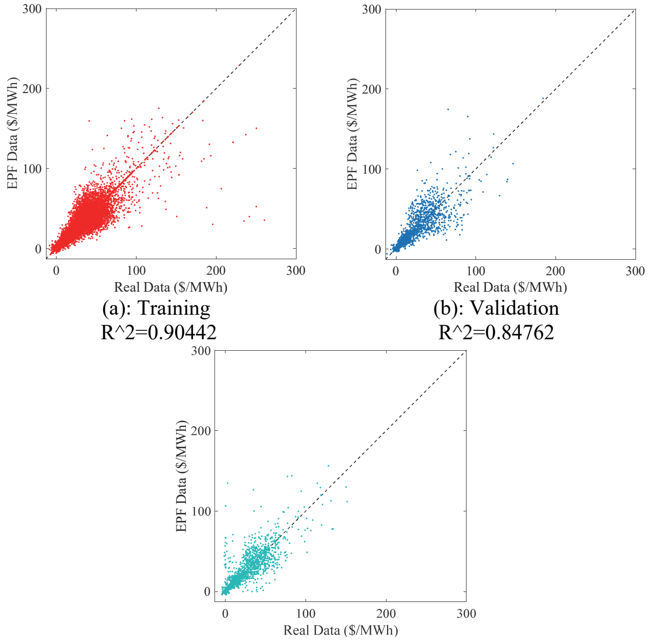
(c)

Fig. 8. Different error criteria variation for various CL and CFs: (a) MAE ($/MWh) (b) MAPE (%) (c) RMSE ($/MWh).



(a) (b) (c) (d)

Fig. 9. Optimal configuration of data reduction for different CLs: (a) CL=1, CF=0.75 (b) CL=2, CF=0.8 (c) CL=3, CF= 0.85 (d) CL=4, CF= 0.95.



(a): Training
R^2=0.90442

(b): Validation
R^2=0.84762

(c): Test
R^2=0.85259

Fig. 10. Training procedure of the proposed method.

Table I. The mean value of different error criterions for different CF.

CF	One compression layer			Two compression layers			Three compression layers			Four compression layers		
	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)
0.95	32.38	9.36	12.37	29.15	8.43	11.15	26.11	7.552	9.91	24.07	6.96	9.06
0.9	31.75	9.18	12.11	28.94	8.37	11.07	18.31	5.29	6.99	30.54	8.83	11.62
0.85	30.70	8.88	11.53	27.73	8.01	10.48	12.97	3.75	5.28	35.50	10.26	13.32
0.8	28.67	8.29	10.64	19.46	5.63	7.29	21.31	6.16	7.94	39.05	11.29	14.59
0.75	28.18	8.15	10.53	22.85	6.61	8.63	24.74	7.15	9.43	41.49	12.00	15.88
0.7	32.20	9.31	12.11	25.44	7.35	9.68	35.50	10.26	13.32	43.12	12.47	16.12
0.65	37.28	10.78	14.24	37.42	10.82	14.04	39.38	11.38	14.60	42.34	12.24	16.13
0.6	39.00	11.28	14.60	39.08	11.30	14.60	42.82	12.38	16.17	46.32	13.39	17.33

The computation time of different methods are given in Table II. As shown in Table II, in this study, the large number of input data (17520×140) is employed for training, validating and testing tasks which caused high computational time in comparison to the other DL based EPF papers [14]. Fig. 11 and Fig.12 show the hourly EPF results for August 30, 2018, and November 30, 2018, which are randomly selected dates with different fluctuation rate in EP data, respectively. In the same manner, for a month time span, the EP for the August 2018 and November 2018 with low and high EP spike are also forecasted, and the results are shown in Fig. 13 and Fig. 14, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2933009, IEEE Transactions on Industrial Informatics

The forecasted results from different methods are shown in Table III, and it is evident that DR-R-SDAE, in general, outperforms the other methods in a day or a month time span. This precision stems from utilizing DR method alongside the rough structure, and LSTM is the second best. LSTM is a strong tool in time series forecasting which is capable of memorizing the long (static) and short (recurrent) terms of data, and choosing the useful past terms. In the proposed method, the important parameters are selected by GCA, and main features are extracted by DR technique. In fact, this method accomplishes the LSTM duty in another way. As illustrated in the GCA results, different input parameters such as generating power data, weather condition data and so on, have a high correlation with EP profile, and in this study, the input data sequence has a large dimension (140 data in each sequence). In this situation, as illustrated by the forecasting results in different time spans, feature extraction-based method has a better performance than LSTM. Because every component of the input data sequence has a different effect and in LSTM, all of them are selected or forgotten together, however in the proposed method main features are selected and extracted with high precision. Furthermore, LSTM has a gradient vanishing problem in large dimension input data and feature extraction methods are built for solving this situation [cite power system]. After LSTM, R-SDAE has a sufficient performance; as an example, in November 2018, the SDAE without DR and rough structure, the MAPE, MAE, and RMSE augmented from 12.698%, 3.097($/MWh), and 4.343($/MWh) to 16.266% 4.418($/MWh), and 6.062($/MWh), respectively. Moreover, as shown in Fig. 11, 12, 13, and 14, the proposed method outperforms the LSTM networks in sharp spike points. Indeed, high ability of the proposed method in feature extraction task and handling the uncertainty, which is achieved by deep auto-encoders and rough neurons, is more illustrated in high spike points, and this is so imperative in power system operation problems. Comparing the error criterions of R-SDAE with SDAE points out the rough structure's effect on forecasting results that can improve accuracy by examining data in lower and higher bound intervals. Considering these intervals handles the uncertainty more effectively. The error criterions of C-NN in Table III reveal the influence of reducing the noise of the data via SDAE. Input data in this study is derived from diverse sources through different origins of noise. Therefore, employing the SDAE has an appreciable effect on reducing the error of the forecasting results. As shown in Figs. 11 and 12, day-ahead EPF is more accurate on August 30, 2018, than November 30, 2018. This is because of the smoother variation of EP on August 30 than November 30, 2018. On November 30 the best forecasting result employed by DR-R-SDAE has MAPE error equal to 12.69%, whereas, on August 30, DR-R-SDAE presented the MAPE error equal to 5.62% which is less than half of MAPE value on November 30. This is also evident in IESO forecasted results as well. It should be mentioned that by the proposed method, on November 30 –a day with more fluctuation in EP data– the accuracy of EPF result is improved compared to IESO forecasted data as its MAPE criterion is reduced about 9.55%. As shown in Fig. 14, the extremely large price spikes occurred at November 9, 2018. In this day, EP raised from 14.35 $/MWh to 365.64 $/MWh and drops to 33.74 $/MWh in the proceeding hour, which had high influence in forecasting results in this month. This might have happened based on market players' behavior. These changes were compared to the relevant days of previous years, and further investigation of these points revealed that other related parameters did not share the drastic changes such as price. In the same situation, the forecasting accuracy by IESO, SVM and LSTM based on MAPE were 38.30%, 39.97%, and 33.91%, while our proposed method has improved it to 30.95%. Indeed, the proposed method decreased forecasting error in comparison with other strong benchmark techniques. In order to clarify more on this topic, it is better to analyze the EPF results for August 2018 which are depicted in Fig. 13. As shown in this Figure, the proposed method was able to forecast the EP more correctly because there were fewer unprecedented peaks in this month. As stated in Table III, the values of the error criteria in August 2018 are also lower than those of November 2018, which supports this assertion.

Table II. The computational time for different methods.

Algorithms	DR-R-SDAE	LSTM	R-SDAE	SDAE	C-NN	SVM
Time (S)	48.325	32.559	30.727	28.682	23.641	31.365

Table III. Error criteria for different methods on selected days - 30 August, 30 November. – and selected months – August and November. 2018.

Method	August 30, 2018			November 30, 2018			August 2018			November 2018		
	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)
DR-R-SDAE	5.62	1.59	1.95	12.69	3.09	4.34	12.96	3.75	5.28	30.95	7.31	16.32
R-SDAE	8.16	2.30	2.77	13.18	3.77	5.27	14.70	4.25	5.45	34.67	7.75	17.29
SDAE	10.26	2.90	3.88	16.26	4.41	6.06	17.80	5.14	6.71	34.79	8.19	17.84
C-NN	12.90	3.65	4.28	19.27	4.88	6.68	19.55	5.65	7.40	37.97	8.97	18.23
LSTM	8.02	2.09	2.41	13.03	3.22	5.01	14.10	4.08	5.39	33.91	7.63	16.87
SVM	11.92	3.37	4.70	22.81	6.31	8.34	24.58	7.10	9.16	39.97	9.44	20.36
IESO	11.60	3.28	3.93	22.24	5.98	7.90	23.31	6.74	10.79	38.30	9.05	18.56

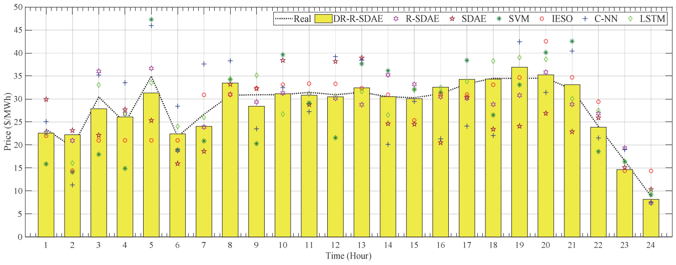


Fig. 11. Forecasting results for different methods on August 30, 2018.

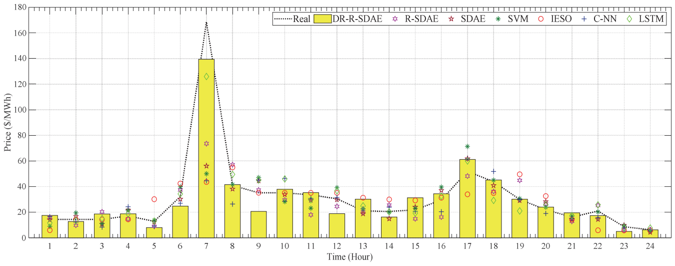
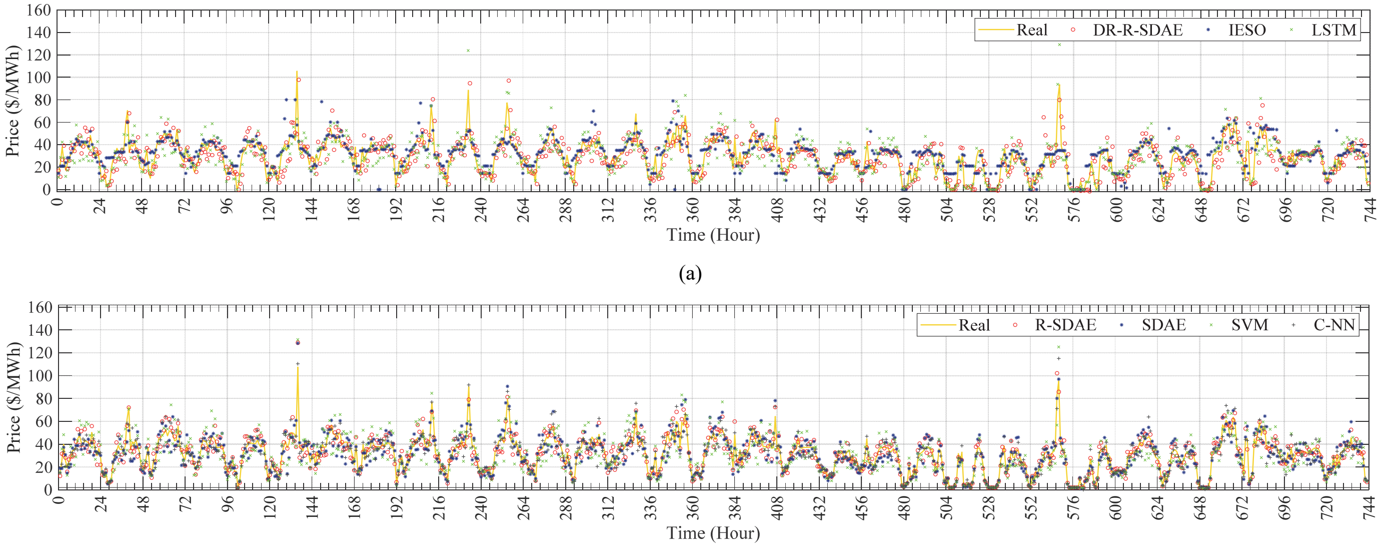


Fig. 12. Forecasting results for different methods on November 30, 2018.

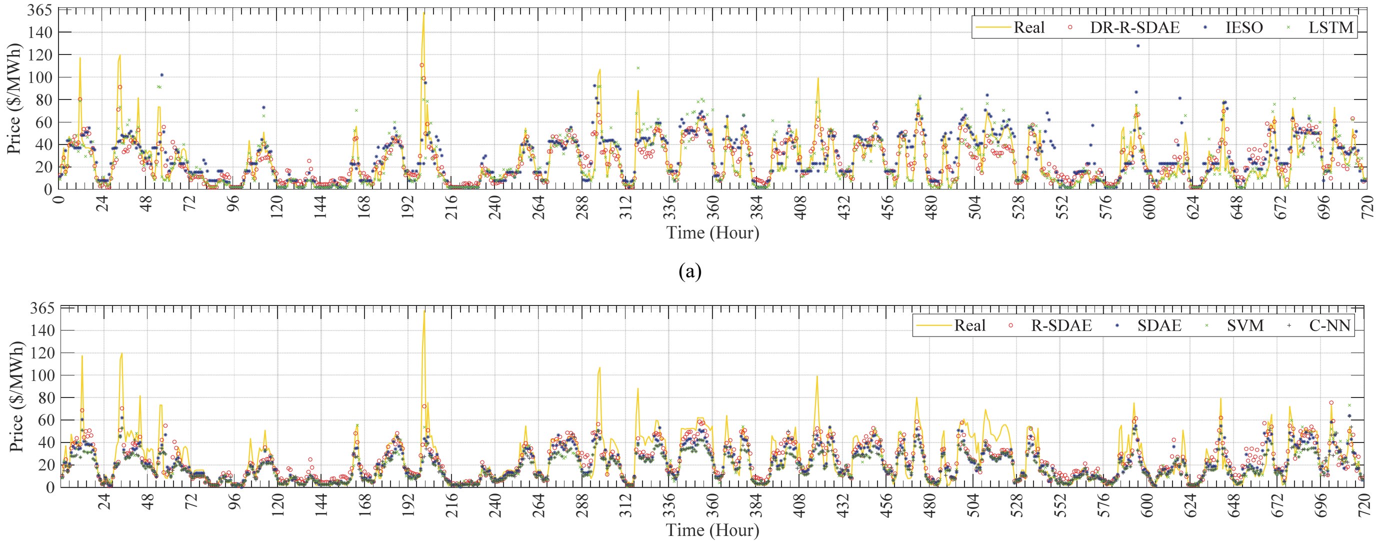
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2933009, IEEE Transactions on Industrial Informatics



(a)

(b)

Fig. 13. Forecasting results for different methods for August 2018; (a): DR-R-SDAE, IESO, LSTM; (b): R-SDAE, SDAE, SVM, C-NN.



(a)

(b)

Fig. 14. Forecasting results for different methods for November 2018; (a): DR-R-SDAE, IESO, LSTM; (b): R-SDAE, SDAE, SVM, C-NN.

Table IV. Error criteria values for different methods on selected days - 30 March, 30 July. – and selected months – March and July. 2018.

Method	March 30, 2018			July 30, 2018			March 2018			July 2018		
	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)
DR-R-SDAE	11.68	2.08	2.99	5.64	0.83	1.17	17.10	3.18	4.76	7.66	1.13	1.58
R-SDAE	13.32	2.39	3.68	6.56	0.97	1.38	20.74	3.72	5.57	10.33	1.53	2.06
SDAE	14.46	2.59	4.09	7.95	1.18	1.67	21.38	3.84	5.80	11.98	1.77	2.43
C-NN	18.31	3.27	5.15	9.03	1.33	1.83	24.86	4.46	6.64	16.12	2.39	3.25
LSTM	12.91	2.31	3.58	6.11	0.90	1.28	19.65	3.53	5.33	8.31	1.23	1.71
SVM	17.68	3.22	4.89	10.05	1.49	2.03	26.10	4.69	7.10	16.47	2.44	3.28
IESO	17.17	3.08	4.63	10.41	1.54	2.16	25.77	4.63	6.97	15.95	2.36	3.25

To prove the performance of the proposed method, another forecasting test is implemented in July and March, 2018, and the forecasting results are presented in Table IV. As the forecasting results show, the proposed method has high accuracy and outperforms the other approaches.

To investigate the robustness of the proposed method with low input data parameters, another forecasting task is also considered. In this study, each hour of a day in 2018 is forecasted based on the different input parameters data - gas, wind, biofuel, total output, H1P, H2P, H3P, market demand, and Ontario demand - of the similar hour in the last 14 days. If we don't have this information about our case study, we can employ the similar hour of the last 30 days for low input data parameters (just electricity price and load demand of the similar hour of the last 30 days).

To verify the robustness of the proposed method with low input data parameters, the forecasting result on November 9, 2018 (a day with sharp spikes), with low input parameters in comparison with LSTM (the best benchmark approach) network result is shown in Fig. 15. Different error criteria of low and full input parameters are presented in Table V. Based on the numerical results; the proposed method shows more robustness with low input parameters in comparison with LSTM method. Indeed, based on the MAPE criterion, the proposed method just losses 4.23% in the accuracy of the forecasting results but the LSTM losses 8.4%, which implies the acceptable performance of the proposed method with low input parameters. This happens because of the high ability of the proposed method in feature extraction of the large dimension data sets and handling the uncertainty. Accordingly, to improve the robustness of the proposed method in case studies with low input parameters, we can employ more previous data of the available parameters based on the high ability of the proposed method in feature extraction task with large dimension data sets. The superiority of the proposed method to LSTM network is more illustrated on November 9, 2018, so that DR-R-SDAE method can follow the sharp spike points more precisely in comparison with LSTM.

Finally, another comparison based on R-square as a performance metric is calculated for August 30, 2018 – random day-ahead forecasting – and the results are shown in Fig. 16. In Fig. 16, a higher R-squared specifies the better-forecasted EP which moves more relatively in line with the real data. The R-squared value defines the relationship between forecasted EP and real EP. As the results show, the proposed method, significantly increased the R-squared, and it is greater than other approaches which validates the efficiency of the proposed method. The findings illustrate the effects of the DR task with optimal DR rate and rough neuron-based networks on decreasing the forecasting error in daily and monthly time spans. DR and rough neuron-based networks handle the feature extraction and data uncertainty with acceptable performance, respectively.

Table V. Error criteria values for different input parameters on November 9, 2018.

Case Study	Full input data parameters including gas, wind, biofuel, total output, H1P, H2P, H3P, market demand, and Ontario demand			Low input data parameters including H1P, H2P, H3P, market demand, and Ontario demand		
Methods	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)	MAPE (%)	MAE ($/MWh)	RMSE ($/MWh)
DR-R-SDAE	32.38	9.36	12.37	29.15	8.43	11.15
LSTM	31.75	9.18	12.11	28.94	8.37	11.07

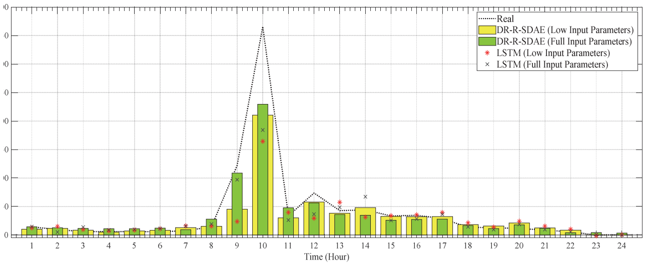


Fig. 15. Forecasting results for different input parameters on November 9, 2018.

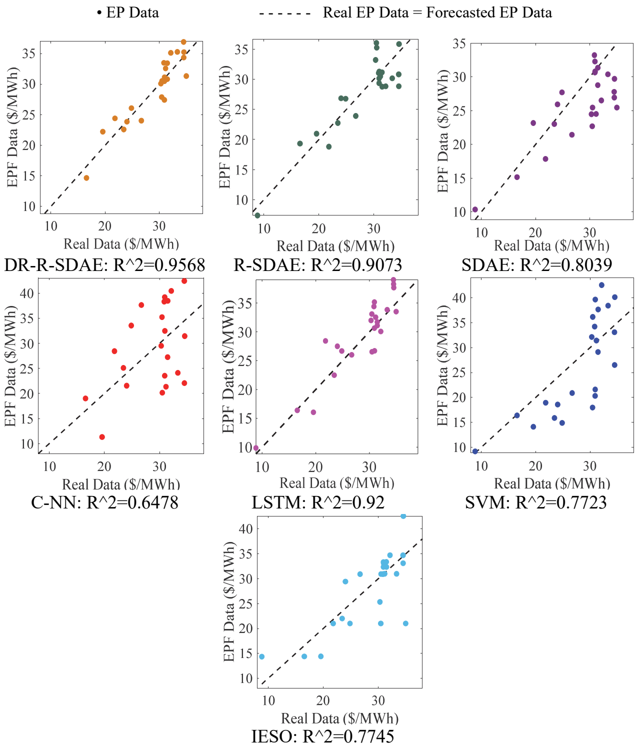


Fig. 16. R-squared values for different approaches on August 30, 2018.

V. Conclusion

A short-term Electricity Price Forecasting (EPF) problem has been studied in this paper. Previous records of electricity price, load demands, different power sources data are considered as input data. In this regard, to select effective parameters in EPF, Grey correlation analysis is applied which finds the vital parameters in EPF task and improves the forecasting accuracy. Then, data from different sources are denoised via stacked denoising auto-encoders in order to eliminate the noise deficiencies. To enhance the accuracy, a feature extraction method based on Dimension Reduction (DR) strategy is implemented to obtain the crucial features of the phenomena under study. As the numerical results show, increasing compression factor is not always adequate, and for each case study, the optimal configuration of the DR method should be determined. It should be mentioned that over DR will have a negative impact on forecasting task. To validate the proposed method, in addition to the different ANN-based methods, EPF results are compared with independent electricity system operator (IESO) forecasting data, and benchmarking methods SVM and LSTM. As illustrated in numerical results, the proposed method outperforms these approaches based on different performance metric criterions which shows the robustness of this technique. The numerical results point out the impact of feature extraction tool and rough neurons which improve the forecasting results about 6-12 % and 3-8%, respectively, in comparison to IESO forecasting results. This improvement implies the superiority of the proposed approach. This approach is recommended to other studies with a large volume of input data. For future work, it is suggested to employ the proposed method in forecasting the renewable energy resources power and load demand, and find the accurate forecasting effects in power system optimal dispatch problem.

Appendix

A. Deep learning method Pseudo-code

Training procedure of Deep artificial neural network is presented by pseudo-code as follows:

Algorithm

Deep learning by back-propagation algorithm

```
Begin operation Back-Propagation
  while validation criterion is not satisfied do
    for $epoch$ = 1 to $epoch_{max}$ do
      for each input sequence X do
        divide data set in mini-batches
        dropout neurons
        forward propagation
        calculate loss function for each sequence with momentum and L2 regulation coefficients
        for $S$ = 1 to $N^l$ do //with $N^l$ number of hidden layers
          back-propagate error on the hidden layer $S$
        end for
        calculate the updated weights of the output layer
        for $i$ = 1 to $N^l$ do
          update weights of the hidden layer $i$
        end for
      end for
        for each input sequence of validation part do
          calculate loss function
        end for
    end for
  end while
end operation
```

B. Training formulation of rough network

The training procedure of the rough neuron-based network with two hidden layers is defined as follows:

$$\frac{\partial E}{\partial w_U^2}=\frac{\partial E}{\partial o^2}\frac{\partial o^2}{\partial o_U^2}\frac{\partial o_U^2}{\partial w_U^2} \tag{1}$$

$$\frac{\partial E}{\partial w_L^2}=\frac{\partial E}{\partial o^2}\frac{\partial o^2}{\partial o_L^2}\frac{\partial o_L^2}{\partial w_L^2} \tag{2}$$

$$\frac{\partial E}{\partial \alpha^2}=\frac{\partial E}{\partial o^2}\frac{\partial o^2}{\partial \alpha^2} \tag{3}$$

$$\frac{\partial E}{\partial \beta^2}=\frac{\partial E}{\partial o^2}\frac{\partial o^2}{\partial \beta^2} \tag{4}$$

$$\frac{\partial E}{\partial w_U^1}=\frac{\partial E}{\partial o^2}\frac{\partial o^2}{\partial o_U^2}\frac{\partial o_U^2}{\partial o^1}\frac{\partial o^1}{\partial o_U^1}\frac{\partial o_U^1}{\partial w_U^1}+\frac{\partial E}{\partial o^2}\frac{\partial o^2}{\partial o_L^2}\frac{\partial o_L^2}{\partial o^1}\frac{\partial o^1}{\partial o_U^1}\frac{\partial o_U^1}{\partial w_U^1} \tag{5}$$

$$\frac{\partial E}{\partial w_L^1}=\frac{\partial E}{\partial o^2}\frac{\partial o^2}{\partial o_U^2}\frac{\partial o_U^2}{\partial o^1}\frac{\partial o^1}{\partial o_L^1}\frac{\partial o_L^1}{\partial w_L^1}+\frac{\partial E}{\partial o^2}\frac{\partial o^2}{\partial o_L^2}\frac{\partial o_L^2}{\partial o^1}\frac{\partial o^1}{\partial o_L^1}\frac{\partial o_L^1}{\partial w_L^1} \tag{6}$$

$$\frac{\partial E}{\partial \alpha^1}=\frac{\partial E}{\partial o^2}\frac{\partial o^2}{\partial o_U^2}\frac{\partial o_U^2}{\partial o^1}\frac{\partial o^1}{\partial \alpha^1}+\frac{\partial E}{\partial o^2}\frac{\partial o^2}{\partial o_L^2}\frac{\partial o_L^2}{\partial o^1}\frac{\partial o^1}{\partial \alpha^1} \tag{7}$$

$$\frac{\partial E}{\partial \beta^1}=\frac{\partial E}{\partial o^2}\frac{\partial o^2}{\partial o_U^2}\frac{\partial o_U^2}{\partial o^1}\frac{\partial o^1}{\partial \beta^1}+\frac{\partial E}{\partial o^2}\frac{\partial o^2}{\partial o_L^2}\frac{\partial o_L^2}{\partial o^1}\frac{\partial o^1}{\partial \beta^1} \tag{8}$$

References

[1] 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 Trans. Ind. Informatics*, 2018.

[2] X. Li and X. Li, "Big Data and Its Key Technology in the Future," *Comput. Sci. Eng.*, vol. 20, no. 4, 2018.

[3] A. I. Sarwat, M. Amini, A. Domijan, A. Damnjanovic, and F. Kaleem, "Weather-based interruption prediction in the smart grid utilizing chronological data," *J. Mod. Power Syst. Clean Energy*, vol. 4, no. 2, pp. 308–315, 2016.

[4] J. Nowotarski and R. Weron, "Recent advances in electricity price forecasting: A review of probabilistic forecasting," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 1548–1568, Jan. 2018.

[5] G. Osório, M. Lotfi, M. Shafie-khah, V. Campos, and J. Catalão, "Hybrid Forecasting Model for Short-Term Electricity Market Prices with Renewable Integration," *Sustainability*, vol. 11, no. 1, p. 57, 2019.

[6] A. Bello, D. W. Bunn, J. Reneses, and A. Munoz, "Medium-Term Probabilistic Forecasting of Electricity Prices: A Hybrid Approach," *IEEE Trans. Power Syst.*, vol. 32, no. 1, pp. 334–343, 2017.

[7] M. H. Amini, A. Kargarian, and O. Karabasoglu, "ARIMA-based decoupled time series forecasting of electric vehicle charging demand for stochastic power system operation," *Electr. Power Syst. Res.*, vol. 140, pp. 378–390, 2016.

[8] M. Rafiei, T. Niknam, and M. H. Khooban, "Probabilistic Forecasting of Hourly Electricity Price by Generalization of ELM for Usage in Improved Wavelet Neural Network," *IEEE Trans. Ind. Informatics*, vol. 13, no. 1, pp. 71–79, 2017.

[9] A. Bello, J. Reneses, A. Muñoz, and A. Delgadillo, "Probabilistic forecasting of hourly electricity prices in the medium-term using spatial interpolation techniques," *Int. J. Forecast.*, 2016.

[10] P. Kou, D. Liang, L. Gao, and J. Lou, "Probabilistic electricity price forecasting with variational heteroscedastic Gaussian process and active learning," *Energy Convers. Manag.*, 2015.

[11] C. Wan, Z. Xu, Y. Wang, Z. Y. Dong, and K. P. Wong, "A hybrid approach for probabilistic forecasting of electricity price," *IEEE Trans. Smart Grid*, 2014.

[12] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond," *International Journal of Forecasting*, 2016.

[13] H. Jahangir *et al.*, "Charging Demand of Plug-in Electric Vehicles: Forecasting Travel Behavior Based on a Novel Rough Artificial Neural Network Approach," *J. Clean. Prod.*, vol. 229, pp. 1029–1044, 2019.

[14] L. Wang, Z. Zhang, and J. Chen, "Short-Term Electricity Price Forecasting with Stacked Denoising Autoencoders," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2673–2681, 2017.

[15] D. Singhal and K. S. Swarup, "Electricity price forecasting using artificial neural networks," *Int. J. Electr. Power Energy Syst.*, vol. 33, no. 3, pp. 550–555, 2011.

[16] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2014.

[17] G. Marcjasz, B. Uniejewski, and R. Weron, "On the importance of the long-term seasonal component in day-ahead electricity price forecasting with NARX neural networks," *Int. J. Forecast.*, Jan. 2018.

[18] J. Lago, F. De Ridder, and B. De Schutter, "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms," *Appl. Energy*, 2018.

[19] C. Tong, J. Li, C. Lang, F. Kong, J. Niu, and J. J. P. C. Rodrigues, "An efficient deep model for day-ahead electricity load forecasting with stacked denoising auto-encoders," *Journal of Parallel and Distributed Computing*, 2017.

[20] W. Kristjanpoller and M. C. Minutolo, "A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis," *Expert Syst. Appl.*, 2018.

[21] K. Wang, C. Xu, Y. Zhang, S. Guo, and A. Zomaya, "Robust Big Data Analytics for Electricity Price Forecasting in the Smart Grid," *IEEE Trans. Big Data*, pp. 1–1, 2017.

[22] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science (80-.).*, vol. 313, no. 5786, pp. 504–507, 2006.

[23] G. Ahmadi and M. Teshnehlab, "Designing and implementation of stable sinusoidal rough-neural identifier," *IEEE Trans. neural networks Learn. Syst.*, vol. 28, no. 8, pp. 1774–1786, 2017.

[24] S. Liu and Y. Lin, "Introduction to grey systems theory," *Underst. Complex Syst.*, 2010.

[25] R. Salakhutdinov and G. Hinton, "Deep Boltzmann Machines," in *AISTATS*, 2009, vol. 1, no. 3, pp. 448–455.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, 2014.

[27] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*, 2013, pp. 1139–1147.

[28] Y. Zhou, F.-J. Chang, L.-C. Chang, I.-F. Kao, and Y.-S. Wang, "Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts," *J. Clean. Prod.*, vol. 209, pp. 134–145, Feb. 2019.

[29] "The Independent Electricity System Operator (IESO)," 2018. [Online]. Available: http://ieso.ca/. [Accessed: 05-Dec-2018].

[30] S. M. Hashemi and M. Sanaye-Pasand, "A New Predictive Approach to Wide-Area Out-of-Step Protection," *IEEE Trans. Ind. Informatics*, pp. 1–1, 2018.