IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 7, 2019, accepted October 13, 2019, date of publication October 21, 2019, date of current version October 31, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2948367

Time Series Prediction With Incomplete Dataset Based on Deep Bidirectional Echo State Network

QIANG WANG, LINQING WANG, (Member, IEEE), YING LIU, (Member, IEEE), JUN ZHAO, (Member, IEEE), AND WEI WANG, (Senior Member, IEEE)
School of Control Sciences and Engineering, Dalian University of Technology, Dalian 116023, China

Corresponding author: Jun Zhao (zhaoj@dlut.edu.cn)

This work was supported in part by the National Key R&D Program of China under Grant 2017YFA0700300, in part by the National Natural Sciences Foundation of China under Grant 61833003, Grant 61533005, and Grant 61873048, in part by the Fundamental Research Funds for the Central Universities under Grant DUT18TD07 and Grant DUT19JC40, and in part by the Outstanding Youth Sci-Tech Talent Program of Dalian under Grant 2018RJ01.

ABSTRACT In the complex industrial environment, data missing situation is often occurred in the process of data acquisition and transition. The major contribution of the paper is the proposal of a deep bidirectional echo state network (DBESN) framework for time series prediction with such incomplete dataset. Instead of data imputation methodology, a bidirectional fusion reservoir is here designed to extract the deep bidirectional feature along with forward and backward time scales, based on which a deep autoencoder echo state network (DAESN) and a deep bidirectional state echo state network (DBSESN) are constructed for the incomplete output and input samples, respectively. As for such two networks, a bidirectional echo state network (BESN) is proposed for connecting them to constitute the DBESN framework for prediction. To verify the effectiveness of the proposed method, one synthetic time series as well as two real-world industrial datasets are employed to conduct the comparative experiments. The experimental results demonstrate that the proposed method outperforms other comparative ones at various missing rates.

INDEX TERMS Deep learning, echo state network, incomplete dataset, prediction, time series.

I. INTRODUCTION

With the development of modern industrial information technology, the amount of the industrial data accumulated in process of manufacturing is increasing at an unprecedented rate [1]. To monitor and analyze the state of energy utilization in industrial process, it is necessary to establish a prediction model for some crucial variables based on these process data [2].

Since, in the process of industrial data acquisition and transition, due to the collector faults, transmission errors, memory failures and human errors, the data absence phenomenon may occur, which often brings a huge challenge for data analysis and processing [3]. Furthermore, the existence of missing data may largely increase the difficulties to describe a system by using data-driven approaches [4]. Therefore, it becomes essential to construct a prediction model for the incomplete data to provide a significative guidance for energy scheduling to avoid energy waste in the industrial process.

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Moinul Hossain.

In literature, a series of researches exist on the time series prediction with incomplete dataset in recent years, but most of them only considered the issue of data imputation [5]. For example, a multiple imputation using Markov Chain Monte Carlo (MCMC) algorithm were adopted in [6] to impute the missing data by using partial least squares regression model. Besides, periodicity imputation method, mean imputation method and cubic spline imputation method were employed in [7] for incomplete time series data imputation before modeling, in which the differences between them were analyzed. Furthermore, a polyfit line-fitting algorithm [8], a Gaussian process [9] and a nearest neighbor method [10] were also applied to data imputation. However, one of the major drawbacks of data imputation lies in that the original feature of the data may be retorted by imputation, which could exhibit an adverse effect on the prediction accuracy.

Deep learning, featuring in more hidden layers and massive training data, transforms the samples from the original space into a new feature space to reveal more useful feature and ultimately improve the accuracy of prediction [11]–[14]. For example, a deep architecture was constructed in [15] for time series prediction, in which a deep belief network (DBN)

was employed at the bottom for feature learning in an unsupervised fashion and a multitask regression layer was built at the top for supervised prediction. Also, considering the inherent spatial and temporal correlations, a deep learning-based prediction model was proposed in [16], in which a greedy layer wise training-based stacked autoencoder (SAE) model was used to learn the feature of the samples. Besides, a deep learning-based prediction method was reported in [17], which adopted a multi-task convolutional neural network model to automatically extract features from the time series before prediction. Although the abovementioned studies were applied to time series prediction, they belonged to a class of static memoryless network, which shown limited ability in modeling time series.

With respect to the memory-based deep learning model, a cascade connecting long short-term memory (LSTM) network with multi-layers based on memory units was constructed in [18], in which an origin destination correlation matrix was integrated in the LSTM network via full connected layers and vector generators to capture the feature of time series. Moreover, an improved deep recurrent neural network model was reported in [19], in which a computationally-efficient optimization framework was designed by using Bayesian optimization and Gaussian processes to reduce the computational cost. However, these studies were completed only for complete datasets without any missing points.

Aiming at time series prediction with incomplete dataset, a deep bidirectional echo state network (DBESN) framework is proposed in this study. In order to dispose the missing points in the output and input samples of the proposed framework instead of data imputation, a deep autoencoder echo state network (DAESN) and a deep bidirectional state echo state network (DBSESN) are constructed based on a designed bidirectional fusion reservoir for extracting the deep bidirectional feature of the samples in both past and future time. Then, a bidirectional echo state network (BESN) is proposed for prediction of the deep features extracted by such two constructed networks to compose the proposed framework. To verify the effectiveness of the proposed method, one synthetic time series and two industrial time series datasets coming from real world practice are adopted here for prediction task. The experimental results with three missing rates are further analyzed, and the effectiveness of the proposed method is demonstrated.

The rest of this paper is organized as follows. The related works and preliminaries are reviewed in Section 2. The main contributions including the establishment of the time series prediction model with incomplete dataset based on the DBESN framework is given in Section 3. And, the experimental analyses are conducted in Section 4. Finally, the section 5 draws the conclusions.

II. RELATED WORKS AND PRELIMINARIES

A. RELATED WORKS

Most of the existing prediction studies adopted complete data-based methods via imputation processing when facing with incomplete dataset. In data imputation field, an expectation-maximization (EM) based classification algorithm was reported in [20], which imputed the data by similar target object. However, such an algorithm required to classify the dataset, which made the imputation performance depend on the classification accuracy. In addition, three kinds of correlation analysis-based imputation methods were proposed in [21] to interpolate the missing data. But, these three methods constructed the correlation equations with linear interpolation, which might result in low imputation accuracy. Furthermore, a data imputation algorithm was proposed in [22], where the nearest neighbor data objects of the missing data were found by the classification method, and its average attribute value was used as the estimated value of the missing data. However, the value of k was difficult to determine in practice, which limited the practicability of the method.

In the field of data imputation-based practical application, a quartic polynomial was used to impute the missing data of wind tower in [23] by fitting the wind speed data at different heights of the same tower, but the fitting parameters would dynamically vary with the height, thus the fitting process was very complex. Besides, a sector linear regression method for wind direction data imputation was proposed in [24], where the wind direction data was divided into several categories by wind sector, and the missing data was imputed by linear regression method. However, a large amount of high-quality historical data was required for such a method.

With respect to the field of prediction, a backpropagation neural network with generalized type-2 fuzzy weights was proposed in [25] for time series prediction, where the fuzzy inference was used to optimize the weights of the neural network. Besides, a single neuron model based on dendritic mechanisms and a phase space reconstruction (PSR) method were proposed in [26] to analyze the time series, where the PSR observed the properties of the samples in a reconstructed phase space. Furthermore, an oblique random forest model was proposed in [27], where a least square classifier was employed to replace the single optimal feature-based orthogonal classification algorithm used by standard random forest in each node of the decision trees for forecasting. However, the abovementioned methods were all based on shallow network, in which their modeling ability were largely limitative while handling the complicated data with high noise, and the satisfactory result could not be obtained.

As for deep model-based prediction, a graph deep learning model was proposed in [28], where the temporal features of samples were extracted by using a LSTM network, and then a scalable graph convolutional deep learning architecture was constructed for time series prediction. Moreover, a deep learning-based method was proposed in [29], where a convolutional neural network combining with correlated variable analysis was constructed to predict dynamic time delay sequences. In addition, a deep belief echo-state network was proposed in [30], where a DBN was used for feature extracting and an echo-state learning mechanism was employed for time series prediction. However, the abovementioned

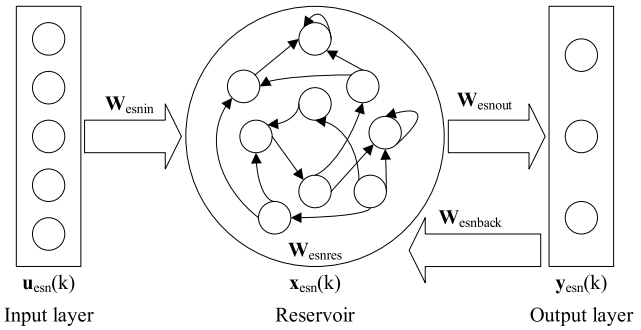


FIGURE 1. The network structure of ESN.

deep learning studies were not concerned with missing data issue.

B. PRELIMINARIES

To make the study self-contained, a brief review of the echo state network (ESN) is given as preliminaries. Such a network, as a kind of recurrent neural network, has shown good performance in time series prediction [31]. Its structure is shown in Fig. 1.

ESN consists of three parts: input layer, state reservoir, output layer [32]. The input signal $\mathbf{u}_{esn}(k)$ is connected to the reservoir by input connection weight matrix $\mathbf{W}_{esnin}$, and the output $\mathbf{y}_{esn}(k)$ is fed back to the reservoir by feedback connection weight matrix $\mathbf{W}_{esnback}$. Besides, the internal state vector is composed of the reservoir output $\mathbf{x}_{esn}(k)$, which is connected to the output layer by output connection weight matrix $\mathbf{W}_{esnout}$. Furthermore, the internal neurons in the reservoir are connected with each other by internal connection weight matrix $\mathbf{W}_{esnres}$ [33], [34]. The relationship among the input, state and output of ESN are

$$\mathbf{x}_{esn}(k+1)=f(\mathbf{W}_{esnin}\mathbf{u}_{esn}(k+1)+\mathbf{W}_{esnres}\mathbf{x}_{esn}(k) + \mathbf{W}_{esnback}\mathbf{y}_{esn}(k)) \quad (1)$$

and

$$\mathbf{y}_{esn}(k+1)=g(\mathbf{W}_{esnout}[\mathbf{u}_{esn}(k+1), \mathbf{x}_{esn}(k+1), \mathbf{y}_{esn}(k)]) \quad (2)$$

where $f(\cdot)$ and $g(\cdot)$ denote the activation function of reservoir neurons and output layer neurons, respectively.

III. DEEP BIDIRECTIONAL ECHO STATE NETWORK FRAMEWORK -BASED TIME SERIES PREDICTION

Due to the irregularities of industrial processes, industrial time series usually contains missing data. In this paper, a deep learning-based prediction model is proposed for time series data with incomplete dataset.

A. OVERALL STRUCTURE OF THE DEEP BIDIRECTIONAL ECHO STATE NETWORK FRAMEWORK

A deep network framework called DBESN is proposed here for time series prediction with incomplete dataset, whose overall structure is shown in Fig. 2.

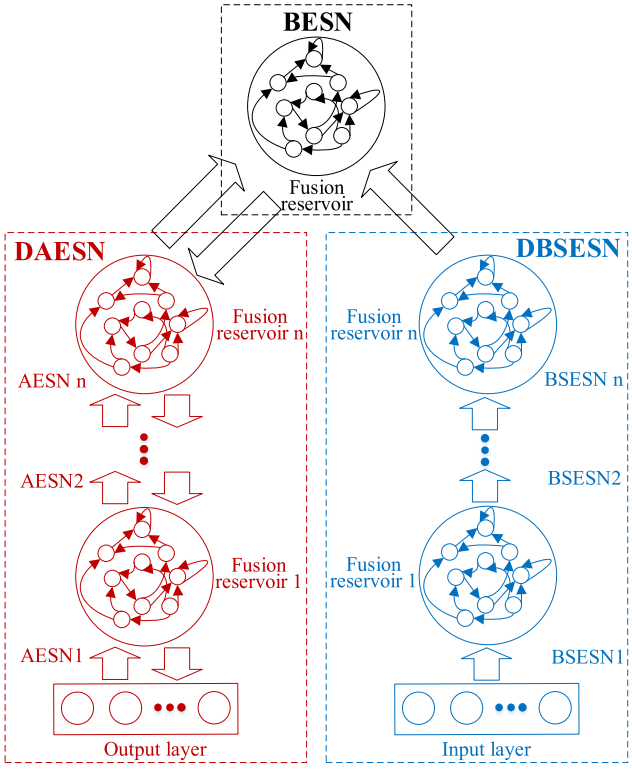


FIGURE 2. The overall structure of the DBESN framework.

The DBESN framework consists of three parts, i.e., the processing for output samples and input samples with incomplete dataset and the feature-based prediction. In the first two parts, such framework adopts multilayer feature extraction instead of data imputation to deal with the missing data, in which a DAESN and a DBSESN are constructed for unsupervised deep feature extraction of the output and input samples with missing points respectively. Then, the third part establishes a BESN to connect the DAESN and the DBSESN for supervised prediction with the deep features extracted by them.

B. DEEP AUTOENCODER ECHO STATE NETWORK

In order to extract the deep feature of the output sample with missing points of the DBESN framework, a DAESN model is proposed in this study, which stacks multilayer autoencoder echo state network (AESN) to form deep structure for feature extraction.

1) AUTOENCODER ECHO STATE NETWORK

Considering the time series data with missing points, both the past information and the future information play an important role in feature extraction of current time. However, the reservoir state of an ESN in current time is only related to the input data of current time, the reservoir state of past time and the output data of past time, that is to say, ESN collects the state information of past time and neglects the future time.

To overcome this issue, a bidirectional fusion state network called AESN is proposed in this study, which collects the state

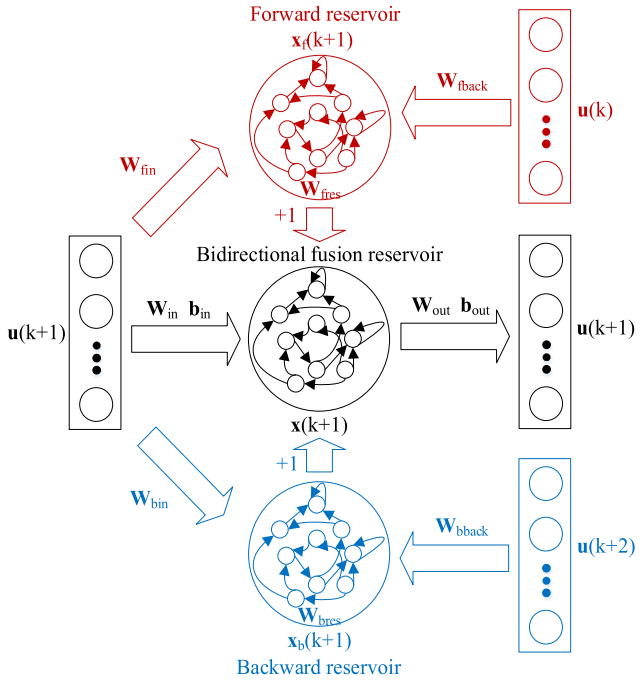


FIGURE 3. The structure of AESN.

information along with forward and backward time scales. The structure of AESN is shown in Fig. 3.

The AESN consists of L input units $\mathbf{u}(k+1) = [u_1(k+1), u_2(k+1), \ldots, u_L(k+1)]^T$, M output units $\mathbf{y}(k+1) = [y_1(k+1), y_2(k+1), \ldots, y_M(k+1)]^T$ and internal state units $\mathbf{x}(k+1)$ at time $k+1$. To extract the feature of $\mathbf{u}(k+1)$, the output of the AESN is set to be equal to the input, i.e., $\mathbf{y}(k+1) = \mathbf{u}(k+1)$. Therefore, $\mathbf{x}(k+1)$ is the feature of $\mathbf{u}(k+1)$.

a: BIDIRECTIONAL FUSION RESERVOIR

In order to collect the state information of both past time and future time at time $k+1$, this study proposes a bidirectional fusion reservoir, which fuses a forward reservoir and a backward reservoir.

The forward reservoir and the backward reservoir collect the past state information and the future state information, respectively, and then such two separated reservoirs are fused together to extract the feature of input sequence. Since the two opposite direction reservoirs are not connected with each other, each reservoir network can be regarded as an ESN structure, which guarantees that its unfolded chart is non-cyclic. The internal state updating equation of forward reservoir is shown in (1), and the backward one is shown as follows,

$$\mathbf{x}_b(k+1) = f(\mathbf{W}_{bres}\mathbf{x}_b(k+2) + \mathbf{W}_{bin}\mathbf{u}(k+1) + \mathbf{W}_{bback}\mathbf{u}(k+2)) \quad (3)$$

where, $\mathbf{W}_{bres}$ denotes the connection weight matrix of internal neurons in backward reservoir, $\mathbf{W}_{bin}$ denotes the input weight matrix of backward reservoir, $\mathbf{W}_{bback}$ denotes the feedback connection weight matrix of backward reservoir, and $\mathbf{x}_b(k+1)$ denotes the state of backward reservoir.

Since the bidirectional fusion reservoir is composed of forward and backward reservoirs, its internal state updating equation is shown as

$$\begin{aligned}\mathbf{x}(k+1) &= f(\mathbf{x}_f(k+1) + \mathbf{x}_b(k+1)) \\ &= f(f(\mathbf{W}_{fres}\mathbf{x}_f(k) + \mathbf{W}_{fin}\mathbf{u}(k+1) + \mathbf{W}_{fback}\mathbf{u}(k)) \\ &\quad + f(\mathbf{W}_{bres}\mathbf{x}_b(k+2) + \mathbf{W}_{bin}\mathbf{u}(k+1) \\ &\quad + \mathbf{W}_{bback}\mathbf{u}(k+2))) \end{aligned} \quad (4)$$

where, $\mathbf{x}(k+1)$ denotes the state of the bidirectional fusion reservoir, $\mathbf{x}_f(k+1)$ denotes the state of the forward reservoir, $\mathbf{W}_{fres}$ denotes the connection weight matrix of internal neurons in forward reservoir, $\mathbf{W}_{fin}$ denotes the input weight matrix of forward reservoir, $\mathbf{W}_{fback}$ denotes the feedback connection weight matrix of forward reservoir.

However, due to the missing points in the sample matrix $\mathbf{U}$, (4) is incalculable. To solve this problem, this paper proposes principle 1 to deal with the missing points.

Principle 1: The missing points do not participate in matrix multiplication and the remaining points participate in matrix multiplication in their original positions.

This principle ensures that the feature extraction is carried out without changing the original feature of the samples with missing points.

b: INPUT WEIGHT MATRIX AND OUTPUT WEIGHT MATRIX

In order to form deep structure by stacking multiple AESNs and apply fine-tuning strategy to optimize the weights of the DAESN, this study defines an input weight matrix $\mathbf{W}_{in}$, an input bias $\mathbf{b}_{in}$, an output weight matrix $\mathbf{W}_{out}$, and an output bias $\mathbf{b}_{out}$.

As for $\mathbf{W}_{in}$, it multiplied by the input sample equals the feature of the input sample extracted by the proposed bidirectional fusion reservoir. The process can be shown as

$$h(\mathbf{W}_{in}\mathbf{u}(k+1) + \mathbf{b}_{in}) = \mathbf{x}(k+1) \quad (5)$$

where $h(\cdot)$ denote the activation function.

Since the output of the AESN equals to its input, the relationship between the feature of the input sample and the output can be summarized as follows,

$$h(\mathbf{W}_{out}\mathbf{x}(k+1) + \mathbf{b}_{out}) = \mathbf{u}(k+1) \quad (6)$$

where the matrix multiplication with missing points is disposed by the principle 1.

To get the optimal values of $\mathbf{W}_{in}$, $\mathbf{b}_{in}$, $\mathbf{W}_{out}$ and $\mathbf{b}_{out}$, one can minimize the following cost function with respect to them by the gradient descent method, i.e.,

$$J(\mathbf{W}_{in}, \mathbf{b}_{in}) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\|h(\mathbf{W}_{in}\mathbf{u}_i(k+1) + \mathbf{b}_{in}) - \mathbf{x}_i(k+1)\|^2\right) + \frac{\lambda}{2}\sum_{j=1}^{s_1}\sum_{k=1}^{s_2}(\mathbf{W}_{in}^{jk})^2 \quad (7)$$

and

$$J(\mathbf{W}_{out}, \mathbf{b}_{out}) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\|h(\mathbf{W}_{out}\mathbf{x}_i(k+1)+\mathbf{b}_{out})-\mathbf{u}_i(k+1)\|^2\right) + \frac{\lambda}{2}\sum_{j=1}^{s_1}\sum_{k=1}^{s_2}(\mathbf{W}_{out}^{jk})^2 \quad (8)$$

where, m, s_1, and s_2 denote the number of input samples, the number of rows of the weight matrix, and the number of columns of the weight matrix, respectively. Furthermore, the first term in the right-hand side denotes an average sum-of-squares error term, and the second term with its coefficient λ denotes a regularization term decreasing the magnitude of the weights as well as preventing overfitting.

However, due to the missing points in $\mathbf{U}$, the calculation process of the gradient descent method cannot be continued. Therefore, this study proposes the following method to solve this problem as principle 2.

Principle 2: When the gradient descent method is used for optimizing the parameters, we calculate the gradient for each sample at a time, in which the case of matrix multiplication with missing points is solved by principle 1. Then, all the gradients are added together and divided by the number of samples to obtain the average of these gradients for updating the parameters.

2) DEEP NETWORK FORMATION AND UNSUPERVISED FINE-TUNING

By utilizing the AESN, the state of the bidirectional fusion reservoir is the feature of the input samples. However, such feature is so shallow that it cannot reveal the essential information of the input samples. In order to further acquire deep features, a DAESN is proposed in this study, in which the reservoir state of an AESN is taken as the input sample of another AESN in order that multiple AESNs are stacked layer by layer to form deep structure.

As for the DAESN, an improved unsupervised wake-sleep algorithm including a wake phase and a sleep phase is also proposed in this paper to fine-tune the deep network, in which all the weights can be optimized to improve the performance of the whole network.

In the wake phase, the feature of current layer $\mathbf{X}^{(i)'}$ is generated by the product of the feature of lower layer $\mathbf{X}^{(i-1)'}$ and the upward input weight $\mathbf{W}_{in}^{(i-1)}$, and a reconstructed feature of lower layer $\mathbf{X}^{(i-1)''''}$ is generated by the product of the downward output weight $\mathbf{W}_{out}^{(i-1)}$ and $\mathbf{X}^{(i)'}$. By calculating the residual error between $\mathbf{X}^{(i-1)'}$ and $\mathbf{X}^{(i-1)''''}$, the gradient descent algorithm can be used to modify $\mathbf{W}_{out}^{(i-1)}$. The updating formula of output weight $\Delta\mathbf{W}_{out}^{(i-1)}$ is shown as

$$\begin{aligned}\Delta\mathbf{W}_{out}^{(i-1)} &= \varepsilon\mathbf{X}^{(i)'}(\mathbf{X}^{(i-1)'}-\mathbf{X}^{(i-1)''''}) \\ &= \varepsilon\mathbf{W}_{in}^{(i-1)}\mathbf{X}^{(i-1)'}(\mathbf{X}^{(i-1)'}-\mathbf{W}_{out}^{(i-1)}\mathbf{W}_{in}^{(i-1)}\mathbf{X}^{(i-1)'})\end{aligned} \quad (9)$$

where, ε denotes the learning rate.

In the sleep phase, the feature of lower layer $\mathbf{X}^{(i-1)''}$ is generated by the product of the feature of current layer $\mathbf{X}^{(i)''}$ and $\mathbf{W}_{out}^{(i-1)}$, and a reconstructed feature of current layer $\mathbf{X}^{(i)'''}$ is generated by the product of $\mathbf{W}_{in}^{(i-1)}$ and $\mathbf{X}^{(i-1)''}$. By calculating the residual error between $\mathbf{X}^{(i)''}$ and $\mathbf{X}^{(i)'''}$, the gradient descent algorithm can be used to modify $\mathbf{W}_{in}^{(i-1)}$. The updating formula of input weight $\Delta\mathbf{W}_{in}^{(i-1)}$ is shown as

$$\begin{aligned}\Delta\mathbf{W}_{in}^{(i-1)} &= \varepsilon\mathbf{X}^{(i-1)''}(\mathbf{X}^{(i)''}-\mathbf{X}^{(i)'''}) \\ &= \varepsilon\mathbf{W}_{out}^{(i-1)}\mathbf{X}^{(i)''}(\mathbf{X}^{(i)''}-\mathbf{W}_{in}^{(i-1)}\mathbf{W}_{out}^{(i-1)}\mathbf{X}^{(i)''})\end{aligned} \quad (10)$$

By multi-layer feature extraction of the DAESN optimized by the improved unsupervised wake-sleep algorithm, the feature of the topmost layer is the deep feature Ldf of the output samples, which is expressed as

$$Ldf = \mathbf{W}_{in}^{(1)}\mathbf{W}_{in}^{(2)}\mathbf{W}_{in}^{(3)}\cdots\mathbf{W}_{in}^{(n-1)}\mathbf{U} \quad (11)$$

C. DEEP BIDIRECTIONAL STATE ECHO STATE NETWORK

With respect to the input samples with missing points of the DBESN framework, a DBSESN model stacked by multilayer bidirectional state echo state network (BSESN) is proposed in this study to extract its deep feature instead of data imputation, in which the reservoir state of an BSESN is taken as the input sample of another BSESN to form deep structure. The network structure of the BSESN is shown in Fig. 4.

The BSESN also adopts the proposed bidirectional fusion reservoir to collect the state information $\mathbf{x}_1(k+1)$ fused by the past state information $\mathbf{x}_{1f}(k+1)$ and the future state information $\mathbf{x}_{1b}(k+1)$ for the incomplete input samples $\mathbf{u}_1(k+1)$, in which the missing points are treated by principle1 and 2. In this way, $\mathbf{x}_1(k+1)$ is the feature of $\mathbf{u}_1(k+1)$, which can be obtained by (4).

By collecting the fusion reservoir state layer by layer, the deep feature Rdf of the input samples with missing points of the DBESN framework can be extracted on the topmost layer of the DBSESN model.

D. BIDIRECTIONAL ECHO STATE NETWORK

By utilizing the DBSESN and DAESN models, the deep features of the input samples and output samples with missing points of the DBESN framework are extracted. Based on these deep features, this paper proposes a BESN model for prediction, whose structure is shown in Fig. 5.

In order to meet the prediction demand, the BESN adds an output matrix $\mathbf{W}_{out}^{BESN}$ to its reservoir, meanwhile, it considers $Rdf(k+1)$ as the input samples and $Ldf(k+1)$ as the output samples to calculate $\mathbf{W}_{out}^{BESN}$ for prediction. Furthermore, such network also fuses a forward and a backward reservoir to collect the past state information $\mathbf{x}_{2f}(k+1)$ and the future state information $\mathbf{x}_{2b}(k+1)$ of $Rdf(k+1)$. Its updating equation of the bidirectional fusion reservoir state $\mathbf{x}_2(k+1)$ is the same as (4), and the output equation is the same as (2).

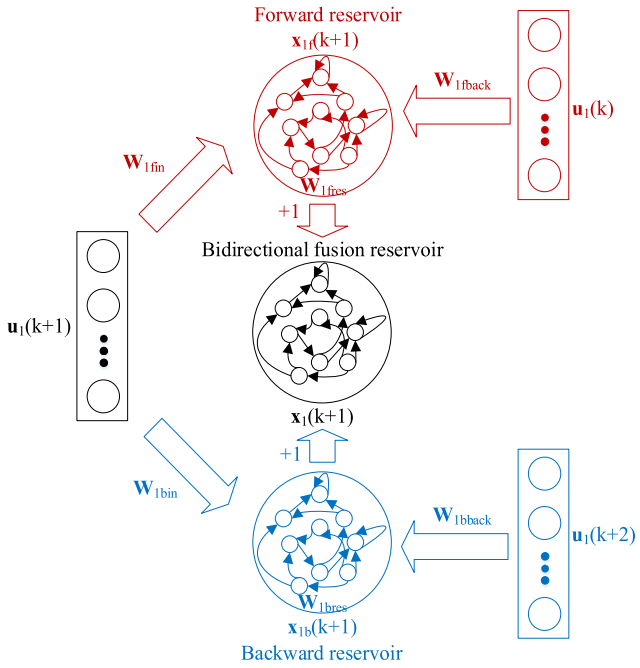


FIGURE 4. The network structure of BSESN.

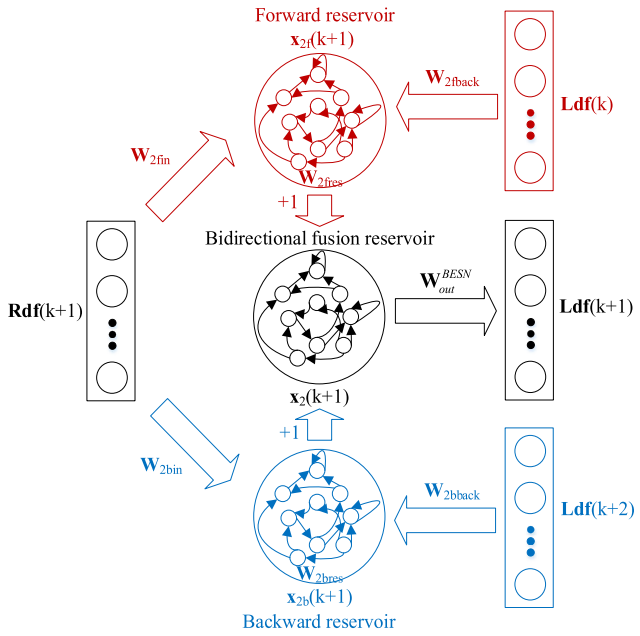


FIGURE 5. The network structure of BESN.

E. SUPERVISED FINE-TUNING

In order to further improve the prediction performance of the DBESN framework, such network is divided into a left part and a right part by the boundary of the reservoir of BESN for supervised fine-tuning. The left part includes the output state of the BESN reservoir and the DAESN, which is used for output prediction. The right part includes DBSESN and the reservoir input of the BESN, which is used for extracting deep feature of the input sample.

In this way, this study uses the error back-propagation algorithm [35] to fine-tune the left part, in which the output state of the reservoir of BESN is considered as the input of the left part and the input of the bottom DAESN is considered as the output of the left part. Thus, its cost function can be expressed as

$$\begin{aligned} J(\mathbf{W}_{out}^{(l)}, \mathbf{b}_{out}^{(l)}) &= \frac{1}{m}\sum_{i=1}^{m}\Big(\frac{1}{2}\Big\|h(\mathbf{W}_{out}^{(l)}(h(\mathbf{W}_{out}^{(2)}(h(\mathbf{W}_{out}^{(1)}\mathbf{x}_i^{BESN}(k+1) \\ &\quad + \mathbf{b}_{out}^{(1)})) + \mathbf{b}_{out}^{(2)})\cdots) + \mathbf{b}_{out}^{(l)}) - \mathbf{u}_i^{DABESN}(k+1)\Big\|^2\Big) \\ &\quad + \frac{\lambda}{2}\sum_{l=1}^{n}\sum_{j=1}^{s_1}\sum_{k=1}^{s_2}(\mathbf{W}_{out}^{(l)jk})^2 \end{aligned} \tag{12}$$

where, $\mathbf{W}_{out}^{(l)}$ denotes the output weight of the l layer, $\mathbf{b}_{out}^{(l)}$ denotes the output bias of the l layer, m denotes the number of samples, $h(\cdot)$ denote the activation function, $\mathbf{x}_i^{BESN}(k+1)$ denotes the output state of the reservoir of BESN, $\mathbf{u}_i^{DABESN}(k+1)$ denotes the input of the bottom DAESN, n denotes the number of layers, s_1 denotes the number of rows of the output weight matrix, and s_2 denotes the number of columns of the output weight matrix. Moreover, the first term in the right-hand side denotes an average sum-of-squares error term, and the second term with its coefficient λ denotes a regularization term decreasing the magnitude of the weights as well as preventing overfitting.

Therefore, the optimal values of the decision variables $\mathbf{W}_{out}^{(l)}$ and $\mathbf{b}_{out}^{(l)}$ can be obtained by using the gradient descent algorithm to minimize the cost function (12) with respect to them, in which the missing points are treated by principle 1 and 2.

F. PROCEDURE OF THE PROPOSED PREDICTION MODEL

To make the structure of the proposed prediction method much easier to be understood, the whole prediction procedure is summarized as follows.

Step 1: Extract the deep feature *Rdf* of the training input samples by the DBSESN.

Step 2: Collect the fusion reservoir state of the DAESN with the training output samples layer by layer. Then calculate $\mathbf{W}_{in}^{(l)}$, $\mathbf{b}_{in}^{(l)}$, $\mathbf{W}_{out}^{(l)}$ and $\mathbf{b}_{out}^{(l)}$ defined in (5) and (6).

Step 3: Fine-tune the DAESN by the improved unsupervised wake-sleep algorithm to optimize $\mathbf{W}_{in}^{(l)}$, $\mathbf{b}_{in}^{(l)}$, $\mathbf{W}_{out}^{(l)}$ and $\mathbf{b}_{out}^{(l)}$. Then make a forward propagation to obtain the deep feature *Ldf* of the training output samples.

Step 4: Train the BESN by taking *Rdf* as the input samples and taking *Ldf* as the output samples to obtain $\mathbf{W}_{out}^{BESN}$.

Step 5: Fine-tune part of the DBESN framework including the DAESN and the output part of the BESN by the error back-propagation algorithm to optimize $\mathbf{W}_{out}^{BESN}$, $\mathbf{W}_{out}^{(l)}$ and $\mathbf{b}_{out}^{(l)}$.

Step 6: Make prediction. Extract the deep feature *Pdf* of the testing samples by the DBSESN and the input part of

TABLE 1. Parameters of the five methods for the incomplete noisy mackey-glass time-series.

Methods	Parameters					
EM-SAE	Embedded dimension	70	Number of hidden layers	2	Number of hidden layer neurons	60
EM-ESN	Embedded dimension	80	Number of neurons in reservoir	200	Spectral radius	0.75
R-SAE	Embedded dimension	70	Number of hidden layers	2	Number of hidden layer neurons	60
R-ESN	Embedded dimension	80	Number of neurons in reservoir	200	Spectral radius	0.75
Proposed Method	Embedded dimension	80	Number of hidden layers	2	Number of neurons in reservoir	200

the BESN. Then multiply Pdf by $\mathbf{W}_{out}^{BESN}$, $\mathbf{W}_{out}^{(l)}$ and add $\mathbf{b}_{out}^{(l)}$ layer by layer to obtain the final prediction results.

IV. EXPERIMENTS AND ANALYSIS

To verify the effectiveness of the proposed method, the Mackey-Glass benchmark dataset and two industrial datasets i.e., blast furnace gas generation flow and coke oven gas generation flow in the gas pipe network system of a steel plant in China are employed for validation. These datasets are chosen due to their own specific features: The Mackey-Glass time series is a standard dataset commonly used for verifying the prediction method. And the two industrial datasets are employed here for demonstrating the practical applicability of the proposed method.

It is well known that due to the equipment failure and the complex industrial environment, the data missing often occurs in supervisory control and data acquisition system. In this study, we employ a period of complete data and randomly create some missing points with 10%, 20% and 30% missing rates to conduct the experiments. For practical demand, we designate the prediction horizon as 200 points. (the sampling interval is 1 min.)

To demonstrate the performance of the proposed method, a series of comparative experiments by some imputation and prediction methods are also presented here. These methods include SAE [36] prediction along with EM [37] imputation (EM-SAE), ESN prediction along with EM imputation (EM-ESN), SAE prediction along with regression imputation [38] (R-SAE) and ESN prediction along with regression imputation (R-ESN). The reasons for choosing these four combination experiments are explained as follows. ESN belongs to shallow dynamic memorial network with a strong ability of modeling time series, and SAE is a deep static memoryless network with a powerful capacity of feature extraction. Therefore, EM-SAE along with EM-ESN, R-SAE along with R-ESN are employed to compare the prediction performance between deep static memoryless network and shallow dynamic memorial network with the same imputation method. EM-SAE along with R-SAE, EM-ESN along with R-ESN are employed to show the influence of different imputation methods on the same prediction method.

To further quantify the performance of the proposed method, the root mean square error (RMSE) and the mean absolute percentage error (MAPE) are employed as the evaluation indices for judging the prediction accuracy of different methods, in which RMSE measures the deviation between the predicted value and real value and MAPE reflects the ratio between the error and the real value. Their expressions are given as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y(i) - y_d(i))^2} \tag{13}$$

and

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y(i) - y_d(i)|}{y_d(i)} \tag{14}$$

where n denotes the predicted length, and $y(i)$ and $y_d(i)$ refer to the predicted value and real value, respectively.

A. MACKEY-GLASS TIME-SERIES DATA

Mackey-Glass system is a time-delay differential system, which is described in the following form

$$\frac{dx(t)}{dt} = \frac{ax(t-\tau)}{1 + x(t-\tau)^{10}} - bx(t) \tag{15}$$

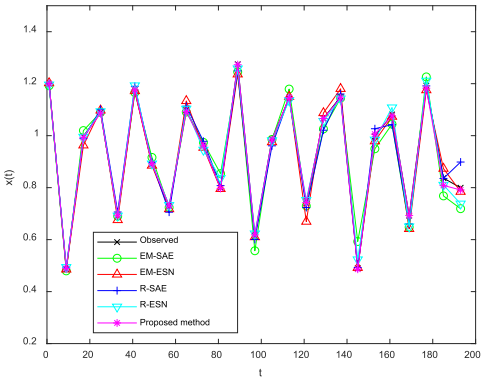
In this experiment, the parameters are set as $a = 0.2$, $b = -0.1$ and $\tau = 17$. In addition, the forth-order Runge-Kutta method is employed to sample a standard Mackey-Glass time-series data, where the sampling period and the initial condition $x(0)$ are set to be 2-s and 1.2, respectively. Furthermore, a noisy Mackey-Glass time-series data generated by the standard one with additive white Gaussian noise with the variance 0.01 is employed to verify the accuracy and robustness of the proposed model.

40,000 data points of the noisy Mackey-Glass time-series data serve as the experimental data. The principles of parameter setting are given as follows. As for the SAE network, the number of neurons in hidden layer is less than that of input layer, and the number of hidden layers is more than one. With respect to the ESN, the number of neurons in reservoir is much more than that of the input layer. The number of hidden layers and the number of neurons in reservoir of the proposed method are set similarly as SAE and ESN. The detailed parameters determined offline by trial and error are listed in Table 1.

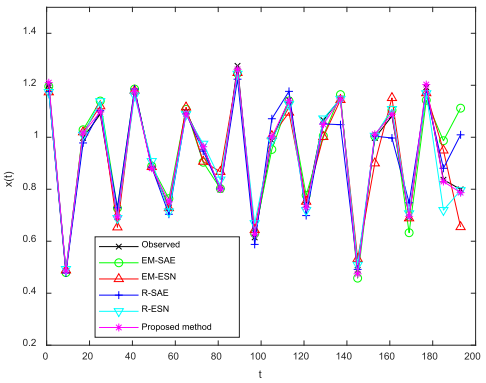
The comparison curves of the real values and the predicted ones by the five methods are plotted in Fig. 6. Besides, to present the comparison curves more clearly, every 8 prediction points by these methods are exhibited here. At 10% missing rate, the prediction curves of the five methods can all track the real values curve with small difference.

TABLE 2. Prediction errors of the five methods for the incomplete noisy mackey-glass time-series.

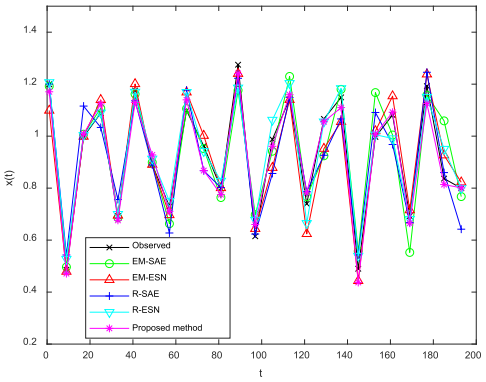
Methods	10%		20%		30%	
	RMSE	MAPE (%)	RMSE	MAPE (%)	RMSE	MAPE (%)
EM-SAE	0.036	3.250	0.070	4.776	0.093	7.921
EM-ESN	0.024	2.052	0.064	4.585	0.082	6.92
R-SAE	0.031	2.701	0.067	4.617	0.088	7.591
R-ESN	0.022	1.918	0.059	4.552	0.077	6.494
Proposed Method	0.007	0.652	0.018	1.434	0.036	3.224



(a)



(b)



(c)

FIGURE 6. Prediction curves of different methods for the incomplete noisy mackey-glass time-series: (a) 10% missing (b) 20% missing (c) 30% missing.

At 20% missing rate, such four comparison methods begin to deviate from the real values, especially at the individual points of the last period. At 30% missing rate, these

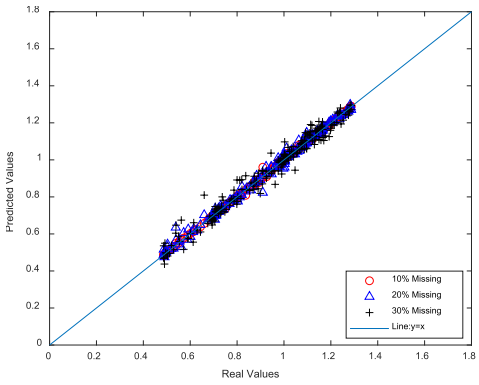


FIGURE 7. Comparison of real and predicted values for the incomplete noisy mackey-glass time-series.

comparison methods show a great fluctuation in about last 50 points. On the contrary, the prediction values of the proposed method have the smallest deviation from the real values.

Furthermore, to come up with a complete comparison of the prediction results by the proposed method, the real and predicted values with three missing rates are compared in Fig. 7. It is clearly seen that all the points are basically distributed around the diagonal at the three missing rates, which indicates a higher prediction accuracy.

To further quantify the prediction accuracy, the RMSE and MAPE of the prediction results of the five methods are listed in Table 2. Due to a certain regularity of the noisy Mackey-Glass time-series data, the prediction accuracy of the five methods at the three missing rates are high. As for the imputation accuracy, EM algorithm is lower than regression algorithm with the same prediction method. Furthermore, the prediction accuracy of ESN is higher than that of SAE with the same imputation method at the three missing rates owing to the memory ability of the reservoir of ESN and the low noise of the data. While, with the capable of memory and the avoidance of data imputation, the prediction accuracy of the proposed method is higher than that of the other four methods.

B. GENERATION FLOW DATA OF THE BLAST FURNACE GAS SYSTEM

In the production process of iron and steel enterprise, the blast furnace gas (BFG) is produced by four blast furnaces. BFG is a very important secondary energy, but it is also an air pollutant. Therefore, it is of great significance to

TABLE 3. Parameters of the five methods for the incomplete #3 BFG generation flow data.

Methods	Parameters					
EM-SAE	Embedded dimension	90	Number of hidden layers	2	Number of hidden layer neurons	80
EM-ESN	Embedded dimension	80	Number of neurons in reservoir	250	Spectral radius	0.75
R-SAE	Embedded dimension	90	Number of hidden layers	2	Number of hidden layer neurons	80
R-ESN	Embedded dimension	80	Number of neurons in reservoir	250	Spectral radius	0.75
Proposed Method	Embedded dimension	90	Number of hidden layers	2	Number of neurons in reservoir	250

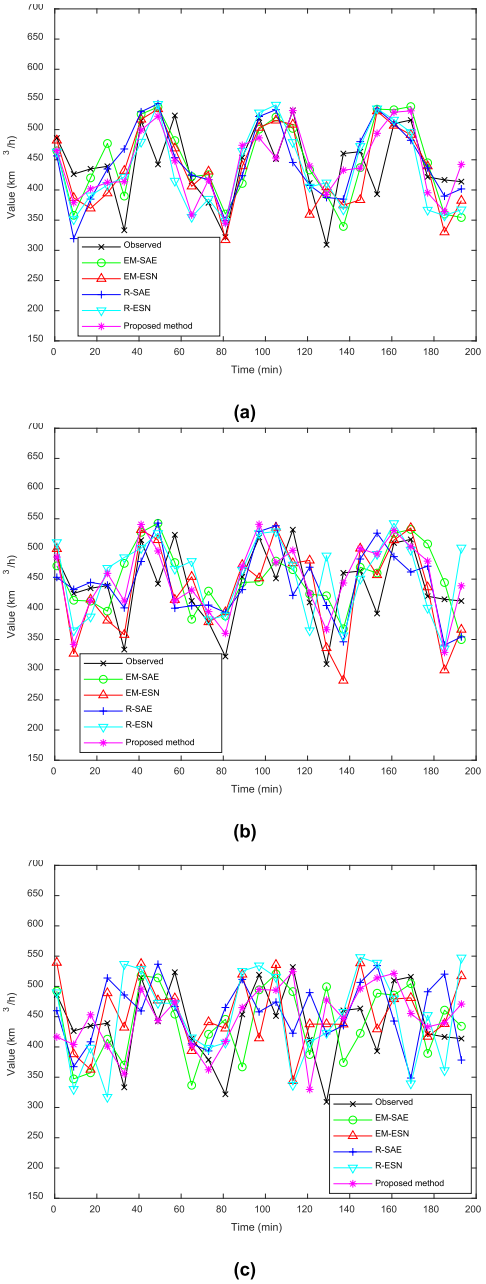


FIGURE 8. Prediction curves of different methods for the incomplete #3 BFG generation flow data: (a) 10% missing (b) 20% missing (c) 30% missing.

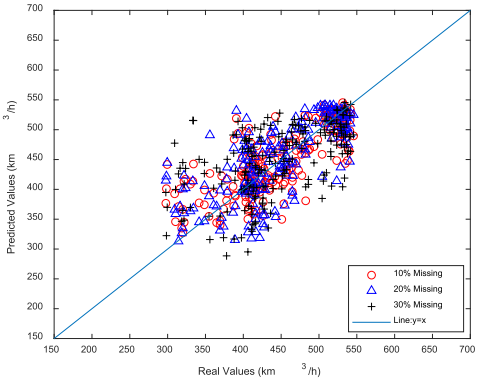


FIGURE 9. Comparison of real and predicted values for the incomplete #3 BFG generation flow data.

accurately predict the quantity of BFG generation flow for energy scheduling. In this section, the generation flow data of the third blast furnace (#3 BFG) in May 2016 with 40,000 points are employed to verify the effectiveness of the proposed method. The dataset with such period can reflect the typical flow state of the BFG, therefore, accurate prediction for it has representative significance at regardless of the time. The parameters of the five methods are offline determined by trial and error in Table 3.

The comparative results of different prediction methods with three missing rates are shown in Fig. 8. As can be seen from the comparison curves, the difference between the predicted values and the real values of the four comparison methods is acceptable at 10% missing rates. But it exhibits unsatisfied performance as the missing rates go on, especially for the regression imputation method at 30% missing rate. However, the proposed method shows an obvious advantage in the aspect of prediction tendency compared with the other four methods along with the increase of the missing rate. In addition, the scatter plot of the proposed method at 10%, 20% and 30% missing rates in Fig. 9 also can further verify the accuracy in a statistical fashion.

In order to further demonstrate the advantage of the proposed method, we present the prediction errors in Table 4. Due to the high noise and poor regularity of the #3 BFG generation flow data, the RMSE and MAPE of the four comparison methods are large. In addition, the regression imputation method performs worse than the EM algorithm with the same prediction method at the three missing rates. Moreover, the performance of ESN is inferior to SAE with the same imputation method, that means deep network can better deal with the high noise and poor regularity data than the shallow network due to the powerful ability of feature extraction. Furthermore, owing to the deep framework, the proposed method

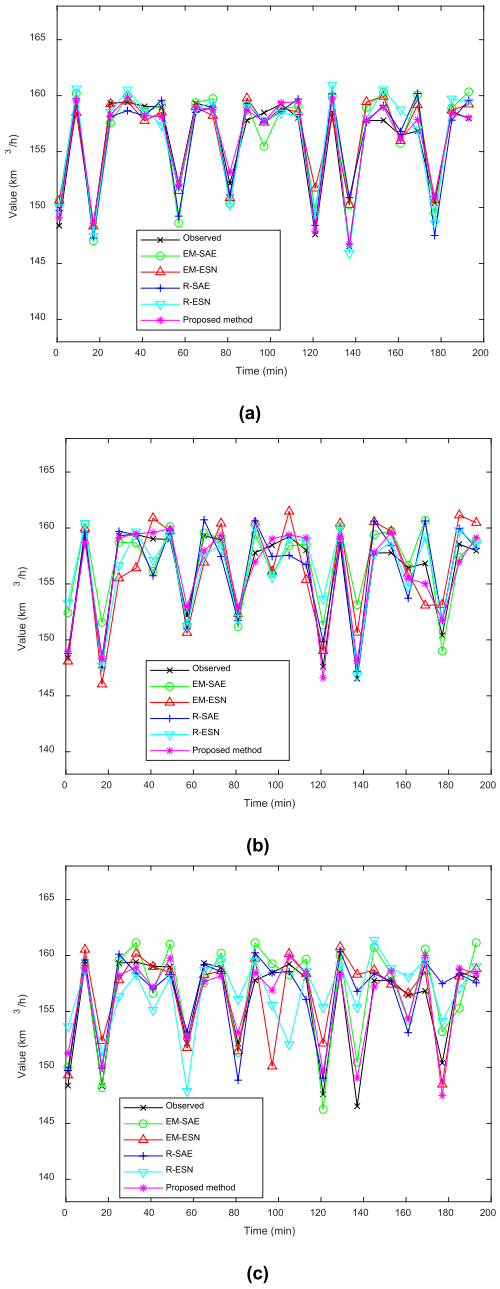


FIGURE 10. Prediction curves of different methods for the incomplete #1 COG generation flow data: (a) 10% missing (b) 20% missing (c) 30% missing.

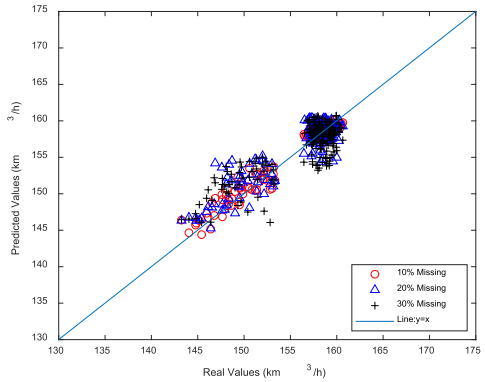


FIGURE 11. Comparison of real and predicted values for the incomplete #1 COG generation flow data.

uses multilayer feature extraction to replace data imputation to get a more excellent accuracy than other methods.

C. GENERATION FLOW DATA OF THE COKE OVEN GAS SYSTEM

The coke oven gas (COG) is an essential part of the energy system in the production of steel enterprise. Accurate prediction of generation flow of COG has an important impact on the rational scheduling in energy system. The COG is produced via six coke ovens in the energy system, and its generation flow data with 40,000 points of the first coke oven (#1 COG) in June 2016 are taken here as the samples, which may reflect the typical flow state of the COG. Hence, accurate prediction for such dataset has representative significance at regardless of the time. Table 5 gives the detailed parameters of the five methods determined offline by trial and error.

Fig. 10 gives the performances of the five prediction methods with different missing rates, and Fig. 11 presents the comparison of real and predicted values of the proposed method. It can intuitively reflect that the prediction curve of the proposed method is closer to the real values curve than the other four comparison methods with no larger deviation points at 10% missing rate. With the increase of missing rate, the prediction curves of the four comparison methods gradually show larger fluctuation, and some of the prediction values have been significantly deviated from the real values. However, the proposed method still performs better than other methods.

According to the error statistics of the five methods presented in Table 6, the prediction errors of the #1 COG generation flow data are between the noisy Mackey-Glass time-series and the #3 BFG generation flow data. At 10% missing rate, the regression imputation method and the ESN prediction method outperform EM imputation method and the SAE prediction method with the same prediction method and the same imputation method, respectively. Nevertheless, at 20% and 30% missing rates, their performances gradually become similar and opposite.

The fluctuation of the prediction accuracy shows that different imputation methods affect the final prediction accuracy seriously at different missing rate. Furthermore, with some low noise and strong regularity data, shallow dynamic memorial network performs better than deep static memoryless network at low missing rate due to its memory ability. However, with the increase of missing rate, the capability of feature extraction is becoming more important than memory ability in modeling incomplete time series data.

As for the proposed method, it can give full play to the powerful modeling ability of dynamic memorial network for time series data and the strong feature extraction ability of the deep network for avoiding the shortcomings of data

TABLE 4. Prediction errors of the five methods for the incomplete #3 BFG generation flow data.

Methods	10%		20%		30%	
	RMSE	MAPE (%)	RMSE	MAPE (%)	RMSE	MAPE (%)
EM-SAE	51.060	9.510	59.843	11.041	72.469	13.568
EM-ESN	54.561	9.952	61.425	11.681	76.001	14.006
R-SAE	55.930	10.403	64.879	12.161	77.243	14.456
R-ESN	57.772	10.881	65.425	12.381	79.567	14.949
Proposed Method	42.623	7.862	48.256	8.856	59.834	11.101

TABLE 5. Parameters of the five methods for the incomplete #1 COG generation flow data.

Methods	Parameters					
EM-SAE	Embedded dimension	80	Number of hidden layers	2	Number of hidden layer neurons	80
EM-ESN	Embedded dimension	80	Number of neurons in reservoir	210	Spectral radius	0.75
R-SAE	Embedded dimension	80	Number of hidden layers	2	Number of hidden layer neurons	80
R-ESN	Embedded dimension	80	Number of neurons in reservoir	210	Spectral radius	0.75
Proposed Method	Embedded dimension	80	Number of hidden layers	2	Number of neurons in reservoir	210

TABLE 6. Prediction errors of the five methods for the incomplete #1 COG generation flow data.

Methods	10%		20%		30%	
	RMSE	MAPE (%)	RMSE	MAPE (%)	RMSE	MAPE (%)
EM-SAE	1.902	0.992	2.573	1.260	3.161	1.542
EM-ESN	1.798	0.862	2.606	1.281	3.412	1.693
R-SAE	1.852	0.949	2.382	1.151	3.225	1.628
R-ESN	1.714	0.851	2.407	1.182	3.534	1.795
Proposed Method	1.061	0.558	1.873	0.931	2.349	1.159

imputation, so both the RMSE and MAPE of the proposed method are significantly lower than those of the other four comparison methods at three missing rates.

V. CONCLUSION

In this study, a DBESN framework is proposed for time series prediction with incomplete dataset. As for the output and input samples with missing points of the proposed framework, a DAESN and a DBSESN are constructed to deal with them respectively, in which a bidirectional fusion reservoir is designed to extract the deep bidirectional feature of the incomplete samples along with forward and backward time scales instead of data imputation. With respect to the deep bidirectional features extracted by these two networks, a BESN model is proposed for prediction based on them to constitute the DBESN framework. To verify the effectiveness of the proposed method, one benchmark dataset and two practical industrial datasets are employed for prediction. The experimental results show that the proposed method outperforms other comparative methods at different missing rates.

REFERENCES

[1] J. Hu, M. Wu, X. Chen, S. Du, P. Zhang, W. Cao, and J. She, "A multilevel prediction model of carbon efficiency based on the differential evolution algorithm for the iron ore sintering process," *IEEE Trans. Ind. Electron.*, vol. 65, no. 11, pp. 8778–8787, Nov. 2018.

[2] Z. Geng, L. Qin, Y. Han, and Q. Zhu, "Energy saving and prediction modeling of petrochemical industries: A novel ELM based on FAHP," *Energy*, vol. 122, pp. 350–362, Mar. 2017.

[3] J. Zhu, Z. Ge, Z. Song, and F. Gao, "Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data," *Annu. Rev. Control*, vol. 46, pp. 107–133, Oct. 2018.

[4] X. Yuan, Z. Ge, B. Huang, and Z. Song, "A probabilistic just-in-time learning framework for soft sensor development with missing data," *IEEE Trans. Control Syst. Technol.*, vol. 25, no. 3, pp. 1124–1132, May 2017.

[5] H. C. Valdiviezo and S. Van Aelst, "Tree-based prediction on incomplete data using imputation or surrogate decisions," *Inf. Sci.*, vol. 311, pp. 163–181, Aug. 2015.

[6] S. Steiner, Y. Zeng, T. M. Young, D. J. Edwards, F. M. Guess, and C.-H. Chen, "A study of missing data imputation in predictive modeling of a wood-composite manufacturing process," *J. Qual. Technol.*, vol. 48, no. 3, pp. 284–296, 2016.

[7] J. B. Li, X. Zhang, and J. X. Zhang, "Prediction and imputation for missing data at random in time series," *Chin. J. Health Statist.*, vol. 29, no. 6, pp. 790–793, 2012.

[8] J. Hu and X. B. Cao, "Time series modeling for missing data at random," *Math. Pract. Theory*, vol. 44, no. 20, pp. 248–252, 2014.

[9] F. Rodrigues, K. Henrickson, and F. C. Pereira, "Multi-output Gaussian processes for crowdsourced traffic data imputation," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 2, pp. 594–603, Feb. 2019.

[10] G. Tutz and S. Ramzan, "Improved methods for the imputation of missing data by nearest neighbor methods," *Comput. Statist. Data Anal.*, vol. 90, pp. 84–99, Oct. 2015.

[11] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[12] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[13] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J. C. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa, "Deep learning for understanding faces: Machines may be just as good, or better, than humans," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 66–83, Jan. 2018.
[14] B. S. Freeman, G. Taylor, B. Gharabaghi, and J. Thé, "Forecasting air quality time series using deep learning," *J. Air Waste Manage. Assoc.*, vol. 68, no. 8, pp. 866–886, 2018.
[15] W. Huang, G. Song, H. Hong, and K. Xie, "Deep architecture for traffic flow prediction: Deep belief networks with multitask learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2191–2201, Oct. 2014.
[16] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.
[17] M. Qiu, P. Zhao, K. Zhang, J. Huang, X. Shi, X. Wang, and W. Chu, "A short-term rainfall prediction model using multi-task convolutional neural networks," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 395–404.
[18] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: A deep learning approach for short-term traffic forecast," *IET Intell. Transp. Syst.*, vol. 11, no. 2, pp. 68–75, 2017.
[19] A. Rahman and A. D. Smith, "Predicting heating demand and sizing a stratified thermal storage tank using deep learning algorithms," *Appl. Energy*, vol. 228, pp. 108–121, Oct. 2018.
[20] T. Schneider, "Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values," *J. Climate*, vol. 14, no. 5, pp. 853–871, 2001.
[21] Y. Wang and H. Zhong, "Analysis of the interpolation method for wind farm's lost wind data," *Renew. Energy Resour.*, vol. 30, no. 3, pp. 14–17, 2012.
[22] C.-C. Liu, D.-Q. Dai, and H. Yan, "The theoretic framework of local weighted approximation for microarray missing value estimation," *Pattern Recognit.*, vol. 43, no. 8, pp. 2993–3002, 2010.
[23] J. H. Xie, P. P. Wang, and H. Y. Zhang, "Interpolation and correction of anemometric data on wind power plant," *Energy Eng.*, vol. 6, pp. 35–37, Dec. 2010.
[24] X. T. Zhang, Z. H. Chen, Y. Xu, and P. J. Sun, "Comparison and analysis of interpolation and correction methods of lost wind measurement data under the circumstance of complex mountains terrain," *Wind Energy*, vol. 1, pp. 82–86, Jan. 2015.
[25] F. Gaxiola, P. Melin, F. Valdez, and O. Castillo, "Generalized type-2 fuzzy weight adjustment for backpropagation neural networks in time series prediction," *Inf. Sci.*, vol. 325, pp. 159–174, Dec. 2015.
[26] T. Zhou, S. Gao, J. Wang, C. Chu, Y. Todo, and Z. Tang, "Financial time series prediction using a dendritic neuron model," *Knowl.-Based Syst.*, vol. 105, pp. 214–224, Aug. 2016.
[27] X. Qiu, L. Zhang, P. N. Suganthan, and G. A. J. Amaratunga, "Oblique random forest ensemble via Least Square Estimation for time series forecasting," *Inf. Sci.*, vol. 420, pp. 249–262, Dec. 2017.
[28] M. Khodayar and J. Wang, "Spatio-temporal graph deep neural network for short-term wind speed forecasting," *IEEE Trans. Sustain. Energy*, vol. 10, no. 2, pp. 670–681, Apr. 2019.
[29] B. Yang and H. Li, "A novel convolutional neural network based approach to predictions of process dynamic time delay sequences," *Chemometrics Intell. Lab. Syst.*, vol. 174, pp. 56–61, Mar. 2018.
[30] X. Sun, T. Li, Q. Li, Y. Huang, and Y. Li, "Deep belief echo-state network and its application to time series prediction," *Knowl.-Based Syst.*, vol. 130, pp. 17–29, Aug. 2017.
[31] L. Wang, H. Hu, X.-Y. Ai, and H. Liu, "Effective electricity energy consumption forecasting using echo state network improved by differential evolution algorithm," *Energy*, vol. 153, pp. 801–815, Jun. 2018.
[32] M. Chen, W. Saad, C. Yin, and M. Debbah, "Echo state networks for proactive caching in cloud-based radio access networks with mobile users," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3520–3535, Jun. 2017.
[33] S. E. Lacy, S. L. Smith, and M. A. Lones, "Using echo state networks for classification: A case study in Parkinson's disease diagnosis," *Artif. Intell. Med.*, vol. 86, pp. 53–59, Mar. 2018.
[34] F. M. Bianchi, "Prediction of telephone calls load using echo state Network with exogenous variables," *Neural Netw.*, vol. 71, pp. 204–213, Nov. 2015.
[35] L. Wang, Y. Zeng, and T. Chen, "Back propagation neural network with adaptive differential evolution algorithm for time series forecasting," *Expert Syst. Appl.*, vol. 42, no. 2, pp. 855–863, 2015.
[36] 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, Jul. 2017.
[37] W. Zhang, Y. P. Jiang, G. D. Yin, and L. A. Yu, "Handling missing values in software effort data based on naive Bayes and EM algorithm," *Syst. Eng.-Theory Pract.*, vol. 37, no. 11, pp. 2965–2974, 2017.
[38] H. Shahbazi, S. Karimi, V. Hosseini, D. Yazgi, and S. Torbatian, "A novel regression imputation framework for Tehran air pollution monitoring network using outputs from WRF and CAMx models," *Atmos. Environ.*, vol. 187, pp. 24–33, Aug. 2018.

QIANG WANG received the B.S. and M.S. degrees from Dalian Jiaotong University, Dalian, China, in 2008 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Dalian University of Technology, Dalian.

His research interests include machine learning and deep learning.

LINQING WANG received the B.S., M.S., and Ph.D. degrees from Northeastern University, Shenyang, China, in 2007, 2010, and 2014, respectively.

She is currently a Lecturer with the School of Control Science and Engineering, Dalian University of Technology. Her current research interests include industrial scheduling and optimization, data mining, and machine learning.

YING LIU received the B.S. and Ph.D. degrees from the Dalian University of Technology, China, in 2004 and 2010, respectively.

She is currently an Assistant Professor with the School of Control Sciences and Engineering, Dalian University of Technology. Her research interests include integrated production planning and scheduling, system simulation and modeling, intelligent optimization and application, and artificial neural networks.

JUN ZHAO received the B.S. degree in control theory from Dalian Jiaotong University, Dalian, China, in 2003, and the Ph.D. degree in engineering from the Dalian University of Technology, Dalian, in 2008.

He is currently a Professor with the School of Control Science and Engineering, Dalian University of Technology. His research interests include industrial production scheduling, computer integrated manufacturing, intelligent optimization, and machine learning.

WEI WANG received the B.S., M.S., and Ph.D. degrees from Northeastern University, Shenyang, China, in 1982, 1986, and 1988, respectively, all in industrial automation.

He was a Postdoctoral Fellow with the Division of Engineering Cybernetics, Norwegian Science and Technology University, Trondheim, Norway, from 1990 to 1992, a Professor and the Vice Director of the National Engineering Research Center of Metallurgical Automation of China, Beijing, China, from 1995 to 1999, and a Research Fellow with the Department of Engineering Science, University of Oxford, Oxford, U.K., from 1998 to 1999. He is currently a Professor with the School of Control Sciences and Engineering, Dalian University of Technology, Dalian, China. His current research interests include adaptive controls, computer integrated manufacturing, and computer controls of industrial processes.