J Inf Process Syst, Vol.17, No.2, pp.242~252, April 2021
https://doi.org/10.3745/JIPS.04.0211

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Industrial Process Monitoring and Fault Diagnosis Based on Temporal Attention Augmented Deep Network

Ke Mu*, Lin Luo*, Qiao Wang*, and Fushun Mao**

Abstract

Following the intuition that the local information in time instances is hardly incorporated into the posterior sequence in long short-term memory (LSTM), this paper proposes an attention augmented mechanism for fault diagnosis of the complex chemical process data. Unlike conventional fault diagnosis and classification methods, an attention mechanism layer architecture is introduced to detect and focus on local temporal information. The augmented deep network results preserve each local instance's importance and contribution and allow the interpretable feature representation and classification simultaneously. The comprehensive comparative analyses demonstrate that the developed model has a high-quality fault classification rate of 95.49%, on average. The results are comparable to those obtained using various other techniques for the Tennessee Eastman benchmark process.

Keywords

Deep Learning, Online Fault Classification, Recurrent Neural Networks, Temporal Attention Mechanism

1. Introduction

The industrial Internet of Things (IoT) development and measuring instruments make industrial records from numerous measurement variables available [1-3]. As a key component in the modern industrial system, data-driven process monitoring is commonly used to protect plant safety, reduce production costs, and improve the quality of products.

Multivariate techniques based on latent variable (LV) methods have been successfully used in the application of process monitoring [4-7], including principal component analysis (PCA), canonical correlation analysis (CCA) and slow feature analysis (SFA), among others. In general, a latent space in the LV model is explored to reveal the low-dimensional inherent structure of original measured variables, and its complementary residual space is to locate the noises and outliers. Once the model has been determined, MSPC control charts with these two spaces are required to detect faults, respectively, referred to as T^2 and SPE control charts. However, the original multivariate techniques share several drawbacks, including a mass of data are required to generalize well, and parameter selection is difficult for their nonlinear extensions.

As a branch of machine learning, deep networks have become powerful tools for effectively dealing

Manuscript received November 18, 2020; first revision December 28, 2020; accepted January 6, 2021.
Corresponding Author: Lin Luo (luolin@lnpu.edu.cn)
* Dept. of Information and Control Engineering, Liaoning Shihua University, Fushun, China (muku@lnpu.edu.cn, lin.l.csc@gmail.com, 63526336@qq.com)
** Synthetic Detergent Factory of Fushun Petrochemical Company, China National Petroleum Corporation, Fushun, China (12hgh@sina.com)

with large-scale data and deep representations [8-11], which greatly impact the final results. Recently, deep learning has been present in the application of process monitoring, such as deep belief network (DBN) [9], stacked sparse auto-encoder (SAE) [12], and recurrent neural network (RNN) [13]. The application of deep learning on monitoring process conditions is still developing, although it often provides more useful insights than the traditional shadow methods. For example, Luo et al. [10] studied an adaptive monitoring strategy with a tensor factorization layer merged into the deep neural network. They extracted fault-sensitive characteristics with the tensor representations, which enable efficient cross-layer knowledge. However, it is challenging to preserve the process's dynamic information, which is important for long-term real-time scenarios. Recently, a typical deep learning model applied to fault diagnosis is long short-term memory (LSTM) framework [14]. The feature extraction process suitably models the process dynamics with recurrent feedback. However, a major limitation of the existing LSTM for the chemical process is that the local information is hardly incorporated into the posterior model. To improve the generalization capability, the local temporal dependencies should be preserved across different time steps.

Motivated by the above observations, this paper proposes an attention augmented network for application to online fault detection and classification. In the proposed deep network, an extra layer is fused with an attention mechanism that makes the layer preserve a time-series dynamic nature and allows its application for an online scenario. Furthermore, the batch normalization procedure's design is utilized to reduce the internal covariate shift of LSTM. Contrary to the conventional shadow method-based fault diagnosis, where feature extraction and classification are generally independent of each other, the proposed method is trained in an end-to-end manner which simultaneously makes the interpretable model and the feature expression learnable. Experimental results on the Tennessee Eastman (TE) benchmark process show that the proposed network can highlight different temporal information's contribution, which helps further analysis on industrial fault features.

The layout of the paper is organized as follows. Section 2 briefly reviews the RNN-based process monitoring method. In Section 3, the proposed attention augmented network approach-based fault diagnosis model is put forward, with the design of the network structure and fault diagnosis procedure. In Section 4, comprehensive comparisons between the attention augmented network-based fault diagnosis method with the existing strategies are carried out with the TE benchmark process. Finally, concluding remarks are drawn in Section 5.

2. Related Works on Fault Diagnosis Using RNN

Assume that a multivariate time series with N samples and D dimensions can be defined as $\mathbf{X}_k \in \mathbb{R}^{D\times\Delta t}$, $k = 1, \cdots, N$, where contains a sequence of Δt sampling points. The input data defined as $\mathbf{X} = [\mathbf{x}_1, \cdots \mathbf{x}_T] \in \mathbb{R}^{D\times\Delta t}$ is fed into the input layer, where T is the time-steps for a sequence. In the hidden layers, RNN [13] maintains a sequence of hidden states $\mathbf{h}_{\Delta t}$ for each time step Δt,

$$\mathbf{h}_{\Delta t} = \tan(\mathbf{W}\mathbf{h}_{\Delta t-1} + \mathbf{U}\mathbf{x}_{\Delta t}) \tag{1}$$

where $\tan(\cdot)$ is the hyperbolic tangent function, $\mathbf{W} \in \mathbb{R}^{D_h\times D_h}$ is the recurrent weight matrix need to be estimated, and $\mathbf{U} \in \mathbb{R}^{D_h\times D}$ signifies the projection matrix. Note that D_h is the number of neurons in each

hidden layer whose values need to be pre-determined. A prediction $\mathbf{y}_{\Delta t}$ can be made using the softmax operation with a hidden state and a weight matrix,

$$\mathbf{y}_{\Delta t} = \text{softmax}(\mathbf{W}\mathbf{h}_{\Delta t-1}) \tag{2}$$

where $\mathbf{Y} = [\mathbf{y}_1, \cdots \mathbf{y}_T] \in \mathbb{R}^{D_h}$ is a tensor of the output.

One of the major issues in RNN, a vanishing gradient problem, has been found in many applications. A typical LSTM network to tackle this problem is to generate an associated sequence of outputs $\mathbf{y}_{\Delta t}$ by three gates and a memory cell. The computation at each time step is as follows,

$$\begin{aligned}
\mathbf{g}_{\Delta t}^{u} &= \sigma(\mathbf{W}^{u}\mathbf{h}_{\Delta t-1} + \mathbf{U}^{u}\mathbf{x}_{\Delta t}) \\
\mathbf{g}_{\Delta t}^{f} &= \sigma(\mathbf{W}^{f}\mathbf{h}_{\Delta t-1} + \mathbf{U}^{f}\mathbf{x}_{\Delta t}) \\
\mathbf{g}_{\Delta t}^{o} &= \sigma(\mathbf{W}^{o}\mathbf{h}_{\Delta t-1} + \mathbf{U}^{o}\mathbf{x}_{\Delta t}) \\
\mathbf{g}_{\Delta t}^{c} &= \tan(\mathbf{W}^{c}\mathbf{h}_{\Delta t-1} + \mathbf{U}^{c}\mathbf{x}_{\Delta t}) \\
\mathbf{m}_{\Delta t} &= \mathbf{g}_{\Delta t}^{f} \odot \mathbf{m}_{\Delta t-1} + \mathbf{g}_{\Delta t}^{u} \odot \mathbf{g}_{\Delta t}^{c} \\
\mathbf{m}_{\Delta t} &= \tan(\mathbf{g}_{\Delta t}^{o} \odot \mathbf{m}_{\Delta t})
\end{aligned} \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function, the symbol $\odot$ is the elementwise multiplication. $\mathbf{W}^{u}$, $\mathbf{W}^{f}$, $\mathbf{W}^{o}$are the weight matrices of input, forget and output gate, respectively. $\mathbf{W}^{c}$ is the weight matrix of memory cell.

3. Temporal Attention Augmented Network for Fault Diagnosis

3.1 Temporal Attention Augmented Layer

Although the layer learns independent temporal dependencies along with each mode, the difficulty with long-term dependencies still arises in the LSTM. It might make the signals about these dependencies tend to be hidden by the smallest fluctuations. This means that squashing local information of the entire sequence poses a potential bottleneck in the performance improvement of LSTM. Inspired by incorporating the position information into sequence-to-sequence learning [15], an attention augmented layer is proposed to overcome the short-term dependencies. Specifically, a vector generated from the sequence of the hidden states $\mathbf{c}_{\Delta t}$ is obtained by a weighted sum of these states $\mathbf{h}_k$, $k = 1, \cdots, T$, at position k,

$$\boldsymbol{c}_{\Delta t} = \sum_{k=1}^{T} \alpha_{\Delta t,k} \boldsymbol{h}_k \tag{4}$$

where $\alpha_{\Delta t,k}$ is the weight of each hidden state, which can be given as,

$$\alpha_{\Delta t,k} = \frac{\exp(e_{\Delta t,k})}{\sum_{j=1}^{T} e_{\Delta t,j}} \tag{5}$$

where the alignment model $e_{i,j}$ is learned by the following equation,

$$e_{\Delta t,k} = \mathbf{v}_{\alpha}^{T} \tan(\mathbf{W}^{\alpha}\mathbf{h}_{\Delta t-1} + \mathbf{U}^{\alpha}\mathbf{h}_k) \tag{6}$$

where $\mathbf{v}_{\alpha}^{T}$ is learnable row vector, $\mathbf{W}^{\alpha}$ and $\mathbf{U}^{\alpha}$ are learnable weights. The parameter vector $\mathbf{v}_{\alpha}^{T}$ and matrix

$\mathbf{W}^{\alpha}$, $\mathbf{U}^{\alpha}$ can be learned from a two-layer multi-layer perceptron without bias.

Using the hidden state $\mathbf{h}_k$ and $\mathbf{h}_{\Delta t-1}$ from the recurrent unit in the decoder module at time $\Delta t-1$, the alignment model matches the inputs around position Δt and the output at position k. The softmax function in Eq. (5) makes the model produce the generated vectors $\boldsymbol{c}_{\Delta t}$ that concerns a specific component of the input sequence. To represent the overall information of the sequence, multiple hops of attention need to be performed so that multiple of the generated vectors $\boldsymbol{c}_{\Delta t}$ focuses on different parts of the sequence. The graphical illustration of the classical LSTM and the proposed model are shown in Fig. 1.

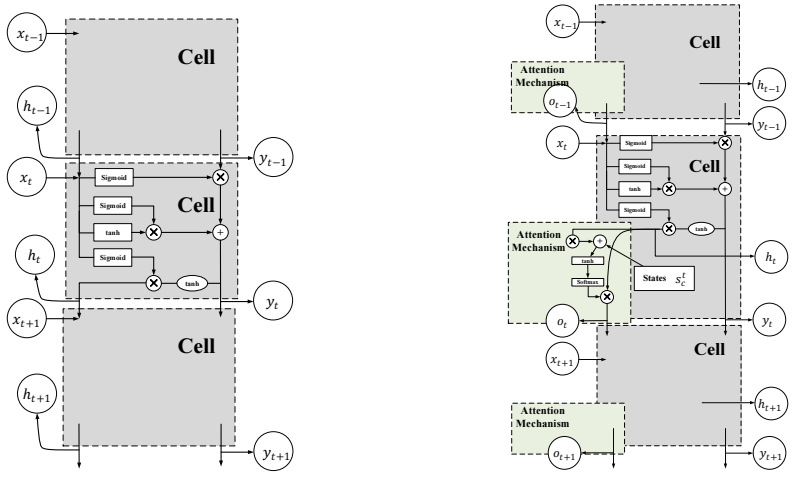


Fig. 1. Illustration of deep network architecture with (a) LSTM and (b) the proposed temporal attention augmented layer, respectively.

3.2 Fault Diagnosis with Temporal Attention Augmented Network

From the hidden state $\mathbf{h}_{\Delta t}$ concerned to the previous state $\mathbf{h}_{\Delta t-1}$, the output $\mathbf{y}_{\Delta t-1}$ and the generated states $\boldsymbol{c}_{\Delta t}$, the output of the last hidden layer has the following form,

$$\mathbf{y}_{\Delta t} = \text{softmax}(\mathbf{W}_{\text{out}}\mathbf{h}_{\Delta t} + \mathbf{b}_{\text{out}}) \tag{7}$$

and the softmax layer calculates a conditional probability of each output neuron for the industrial system health conditions.

A fault detection problem is a classification task to indicate which condition the system belongs to. Two different data sets can be trained by the attention augmented network, one is the operation data from the normal operation, and the other is from the abnormal condition. Along with the accuracy, the other two criteria commonly assessed for the model performance, fault detection rate (FDR) and false alarm rate (FAR), should be defined as follow:

$$\text{FDR} = \frac{\text{Total of faulty samples with fault label}}{\text{Total of faulty samples}} \tag{8}$$

$$\text{FAR} = \frac{\text{Total of normal samples with fault label}}{\text{Total of normal samples}} \tag{9}$$

4. Experiment on Tennessee Eastman Process

In this section, the proposed method's performance evaluation is carried on the TE process, which is a benchmark process for the process modeling and monitoring. A brief description of the TE process is provided firstly, and feature engineering is utilized to improve the later deep network's performance. To evaluate the traditional LSTM network's performance, LSTM with batch normalization (BNLSTM) and the attention augmented network (AAN), a series of experiments are then conducted in the fault detection and classification of the multivariate TE sequential data.

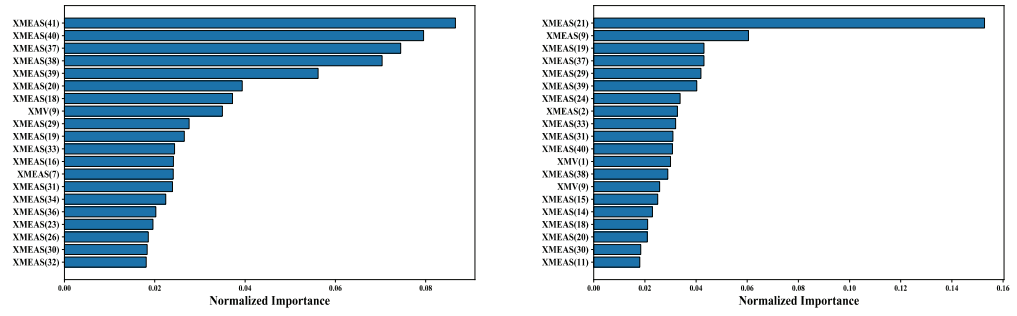
4.1 Process Description

The TE process has been extensively explored in-process monitoring and control communities as a source of the available dataset for comparing various process control and monitoring techniques. The process contains five major process units: a reboiled striper, a cooling condenser, a flash separator, an exothermic two-phase reactor, and a recycling compressor. There are a total of 52 measurements available, in which 41 and 11 measurements are for process variables and manipulated variables, respectively, and a set of 20 programmed fault modes are defined in [16].

For the normal operation, each data set contains a simulation run of 25 hours with a sampling interval of 3 minutes, and it consists of 500 samples. For the faulty operation, each test data set for one fault mode (introduced at 160th sample) consists of 960 samples. All the samples were normalized to zero mean and unit variance.

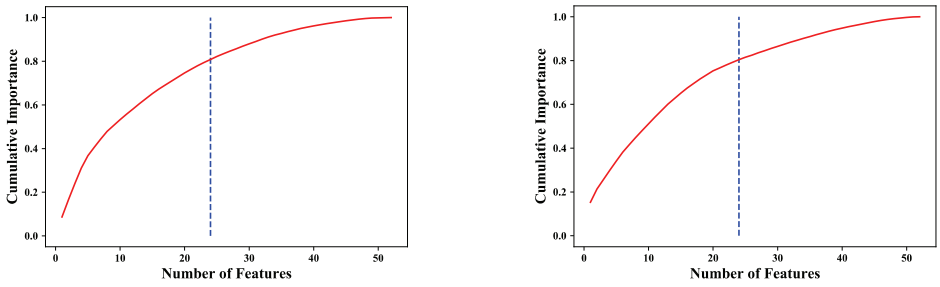
4.2 Feature Selection

Feature selection is one of the core concepts in fault detection and classification, which impacts the model's performance. To identify nonlinear feature interactions and reliably extract relevant features, the importance of features from a model can be automatically estimated by a gradient boosting machine (GBM) implemented by the LightGBM. The importance score is calculated for the individual decision tree by the number of split points that improve the area under the curve (AUC). The feature importance is then averaged over all of the decision trees within the model. The training procedure is repeated 10 times to reduce the variance in the resulting score.



(a) (b)

Fig. 2. The sorting features according to the cumulative importance in (a) IDV15 and (b) IDV17, respectively.



(a) (b)

Fig. 3. The cumulative feature importance versus the number of features in (a) IDV15 and (b) IDV17, respectively.

Fig. 2 shows the 20 most important features in IDV15 and IDV17 on a normalized scale where the features sum to 1, respectively. Meanwhile, it also allows cumulative feature importance to find the number of features. A threshold is used to identify the number of features required to reach a specified cumulative feature importance. We set the threshold to 0.8 in the experiments, which means the number of features accounts for 80% of the total importance. For example, Fig. 3 shows that there are 23 and 28 features that contributed to the specified cumulative importance in IDV15 and IDV17, respectively.

4.3 Effects of Temporal Instances

In the encoder-decoder architecture, such as LSTM, where the entire sequence's information is squashed to a single vector, the local information in time instances is hardly incorporated into the posterior sequence. The problem may degrade the efficiency of the encoder-decoder architecture. Attention augmented mechanism solves this problem by introducing additional weights which contain information surrounding a particular time instance in the past sequence.

In the following experiments, the proposed AAN, the classical LSTM, and BN-LSTM were implemented using Python and the TensorFlow backend. The input layer in all the networks used a sigmoid activation function. The networks were initialized by the Xavier initialization [17] to ensure the signals do not vanish away, and the Adam was selected as the optimizer during the training step. The candidate structures and parameters for these methods are listed in Table 1, where the entire network structure is the number of neurons in input, hidden, and output layers. Regarding regularization techniques, dropout was applied with a percentage of 0.5 to all hidden layers' output.

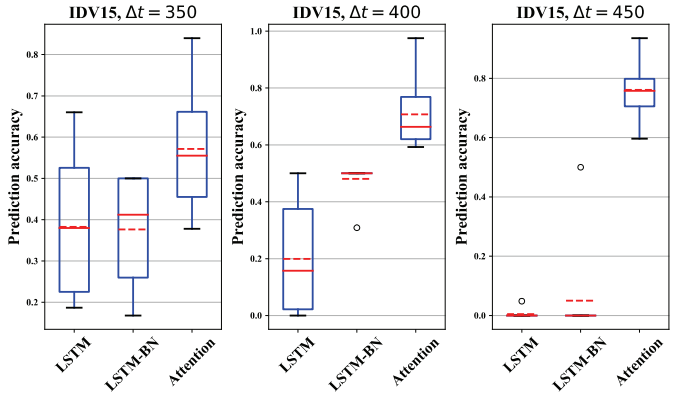
Table 1. The candidate structures and parameters for LSTM, BN-LSTM, and AAN

	LSTM	**BN-LSTM**	**ANN**
Architecture	{20, 64, 128, 64, 2}	{20, 64, 128, 64, 2}	{20, 64, 128, 64, 2}
Optimizer	Adam	Adam	Adam
Learning rate	0.0005	0.0005	0.0005
Decay rates	0.01	0.01	0.01
Hyper-parameter β_1	0.9	0.9	0.9
Hyper-parameter β_2	0.999	0.999	0.999
Activation[a]	{sigmoid, tanh}	{sigmoid, tanh}	{sigmoid, tanh}

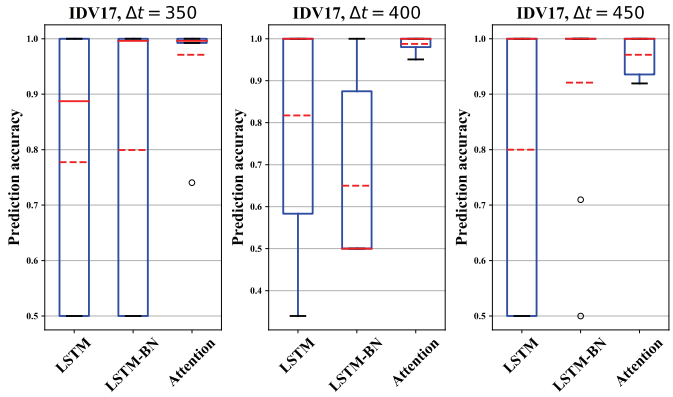
[a] Input activation is sigmoid, hidden activation is tanh.

The dimension of inputs was associated with the number of features required to the 80% cumulative importance. In total, all configurations were trained for a maximum of 40 epochs with a mini-batch size of 32 samples. To evaluate the encoder-decoder structure and attention-based model in the local temporal representation, we constructed three different baseline configurations with $\Delta t = \{350,400,450\}$. Meanwhile, each configuration was repeated 20 times.

Fig. 4 reports the experiment results in IDV15 and IDV17 with three baseline configurations, respectively. As shown in Fig. 4(a), it is clear that the attention augmented mechanism outperforms other competing models as gradually increasing of temporal instances, e.g., $\Delta t = 450$. Inspecting the box-plot in Fig. 4(b) finds that more instances we used in training, the attention augmented mechanism holds the more stable and accurate representation on the sequence. Although being with batch normalization, the BN-LSTM model is inferior to the proposed one. It is because the information in the attention layer is capable of seizing the temporal features across time steps.



(a)



(b)

Fig. 4. Box-plot of prediction accuracy under three different baseline configurations with $\Delta t = \{350,400,450\}$, repeating for 20 times, in (a) IDV15 and (b) IDV17, respectively. The solid and dashed lines are the median and mean of the resulting accuracy, respectively. The circles denote outlier points.

Besides, the attention mechanism in fault detection and classification gives opportunities for interpreting and visualizing the contribution of the temporal instances being attended to. An additional

layer with the same number of output parameters as the input layer is applied to observe how each of the $\Delta t = \{20,50\}$ events in the input instances contributes to the decision function, see Fig. 5.

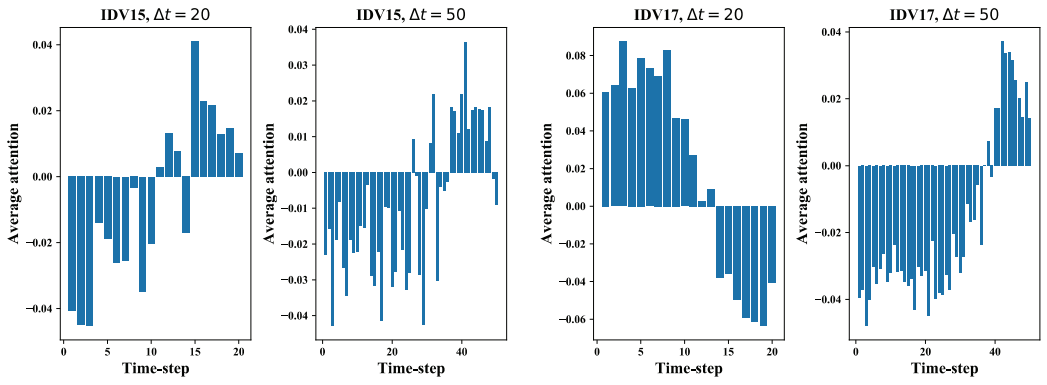


Fig. 5. Contribution of the temporal instances being attended to (a) IDV15 and (b) IDV17 with $\Delta t = \{20,50\}$, respectively.

Visualizing the average attention is considered to each temporal instance during the training process. This would mean that decoder pays much attention to the next state if the attention value is significant.

4.4 Fault Detection Results

When validation data was available in the offline modeling phase, 20 datasets were merged, each containing 480 normal samples and extra 800 samples collected under one fault mode as the validation data set. The proposed AAN, BN-LSTM, DBN with Gaussian activation function [8], deep artificial neural networks (DANN) [18] were constructed for comparative analysis on the fault detection performance. The results from different methods are summarized in Table 2. The ANN model shows the best overall fault detection rate than the other three methods. It can be seen that AAN provides a lower misclassification rate than the other three methods for faults IDV2, IDV8, IDV9, IDV11, IDV13, IDV15, IDV18, and IDV20.

Moreover, they show similar performances for other fault IDs. The improved accuracy to fault classification in ANN lies in the fact that the attention weights retain the long-term dependencies at each time step. The temporal attention can determine the local hidden state, referring to the previous states across all time steps. However, in the case of IDV5, IDV16, and IDV19, better fault detection rates can be found in the DANN method, while DBN performed better in the case of IDV17. Our deep network has not been completely optimized in terms of time length selection, counting that there still exists the possibility for improvements on the different types of faults mentioned above.

Furthermore, the results on the classification rates of all the faults are provided using the proposed method and other deep networks, e.g., hierarchical neural network (HNN) [19], stacked SAE [12] and DANN. The results are illustrated in Fig. 6. It can be seen that almost all the samples can be classified correctly by ANN. Inspecting the simulation results can be concluded that AAN has superior performance in fault detection and classification. This is due to the introduction of the temporal attention mechanism in the ANN model.

Table 2. Fault detection rates of different data-driven methods

Fault ID	Fault detection rates (%)			
	DBN	DANN	BN-LSTM	AAN
IDV1	98	100	90	100
IDV2	95	99.51	100	100
IDV3	100	-	89.26	92.27
IDV4	100	100	90	100
IDV5	79	100	95.02	99.62
IDV6	100	100	100	100
IDV7	100	100	95.06	100
IDV8	89	98.06	90	100
IDV9	66	-	20.13	73.46
IDV10	98	93.96	97.36	98.85
IDV11	91	97.20	91.36	97.71
IDV12	72	98.69	100	99.62
IDV13	91	95.78	85.50	96.56
IDV14	91	99.97	89.31	100
IDV15	0	-	36.25	78.23
IDV16	0	95.41	88.16	91.60
IDV17	100	95.93	95.50	97.53
IDV18	78	94.15	97.47	100
IDV19	98	99.18	90.23	98.47
IDV20	93	93.62	95	98.09
Overall	81.95	97.73*	86.94	95.49

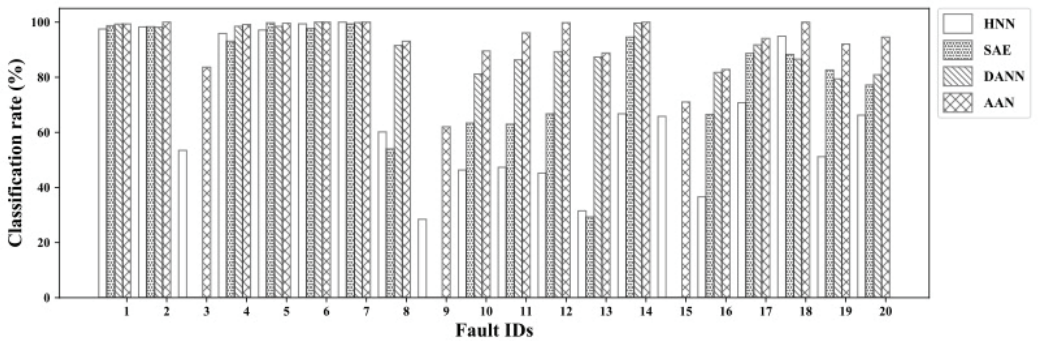


Fig. 6. Fault classification rates of different deep network models.

5. Conclusion

In this paper, we proposed fault detection and diagnosis scheme based on a deep network, where the temporal attention mechanism is designed on the network layer. The proposed scheme has the following notable features due to the local mechanism: the ANN training procedure integrates into an end-to-end manner. It is possible to realize parameter update of the feature extraction and fault classification synchronously. Moreover, the feature extraction relying on the handcrafted operation is significantly reduced. AAN explicitly considers the importance and contribution of each temporal instance and allows further analysis of the time instances of interest. AAN adaptively analyzes the dynamic information of

the industrial process with the usage of LSTM. Case studies on the TE process demonstrated that the AAN-based approach shows superior performance over the conventional classification methods and enhances the interpretability of the hidden state's feature. A promising direction is to address batch process monitoring problems with the attention augmented network in future work. Moreover, the information contained in the temporal and spatial domains should be shared across layers to enable efficient and general knowledge. Hence, the design on the shared layers should be in a further direction.

Acknowledgement

This paper is supported by National Natural Science Foundation of China (No. 61703191), the Foundation of Liaoning Educational Committee (No. L2017LQN028), the Scientific Research Foundation of Liaoning Shihua University (No. 2017XJJ-012).

References

[1] F. A. P. Peres and F. S. Fogliatto, "Variable selection methods in multivariate statistical process control: a systematic literature review," *Computers & Industrial Engineering*, vol. 115, pp. 603-619, 2018.

[2] H. Lahdhiri, M. Said, K. B. Abdellafou, O. Taouali, and M. F. Harkat, "Supervised process monitoring and fault diagnosis based on machine learning methods," *The International Journal of Advanced Manufacturing Technology*, vol. 102, no. 5, pp. 2321-2337, 2019.

[3] Y. Wang, Z. Pan, X. Yuan, C. Yang, and W. Gui, "A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network," *ISA Transactions*, vol. 96, pp. 457-467, 2020.

[4] S. J. Qin and L. H. Chiang, "Advances and opportunities in machine learning for process data analytics," *Computers & Chemical Engineering*, vol. 126, pp. 465-473, 2019.

[5] Q. Jiang, X. Yan, and B. Huang, "Review and perspectives of data-driven distributed monitoring for industrial plant-wide processes," *Industrial & Engineering Chemistry Research*, vol. 58, no. 29, pp. 12899-12912, 2019.

[6] L. Luo, L. Xie, U. Kruger, K. Alzebdeh, and H. Su, "A novel Bayesian robust model and its application for fault detection and automatic supervision of nonlinear process," *Industrial & Engineering Chemistry Research*, vol. 54, no. 18, pp. 5048-5061, 2015.

[7] J. C. Kabugo, S. L. Jamsa-Jounela, R. Schiemann, and C. Binder, "Industry 4.0 based process data analytics platform: a waste-to-energy plant case study," *International Journal of Electrical Power & Energy Systems*, vol. 115, article no. 105508, 2020. https://doi.org/10.1016/j.ijepes.2019.105508

[8] Q. Jiang and X. Yan, "Learning deep correlated representations for nonlinear process monitoring," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 12, pp. 6200-6209, 2018.

[9] Z. Zhang and J. Zhao, "A deep belief network based fault diagnosis model for complex chemical processes," *Computers & Chemical Engineering*, viol. 107, pp. 395-407, 2017.

[10] L. Luo, L. Xie, and H. Su, "Deep learning with tensor factorization layers for sequential fault diagnosis and industrial process monitoring," *IEEE Access*, vol. 8, pp. 105494-105506, 2020.

[11] M. Aamir, Y. F. Pu, Z. Rahman, W. A. Abro, H. Naeem, F. Ullah, and A. M. Badr, "A hybrid proposed framework for object detection and classification," *Journal of Information Processing Systems*, vol. 14, no. 5, pp. 1176-1194, 2018.

[12] F. Lv, C. Wen, Z. Bao, and M. Liu, "Fault diagnosis based on deep learning," in *Proceedings of 2016 American Control Conference (ACC)*, Boston, MA, 2016, pp. 6851-6856.

[13] H. Zhao, S. Sun, and B. Jin, "Sequential fault diagnosis based on LSTM neural network," *IEEE Access*, vol. 6, pp. 12929-12939, 2018.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998-6008, 2017.
[16] A. Bathelt, N. L. Ricker, and M. Jelali, "Revision of the Tennessee Eastman process model," *IFAC-PapersOnLine*, vol. 48, no. 8, pp. 309-314, 2015.
[17] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, 2010, pp. 249-256.
[18] S. Heo and J. H. Lee, "Fault detection and classification using artificial neural networks," *IFAC-PapersOnLine*, vol. 51, no. 18, 470-475, 2018.
[19] R. Eslamloueyan, "Designing a hierarchical neural network based on fuzzy clustering for fault diagnosis of the Tennessee–Eastman process," *Applied Soft Computing*, vol. 11, no. 1, pp. 1407-1415, 2011.

Ke Mu **https://orcid.org/0000-0001-6028-2247**

He received the B.S. degree in industrial automation from Liaoning Shihua University, Liaoning, China, in 1990, and the M.S. degrees from Northeastern University, Shenyang, China, in 2008. He was a lecturer with the Department of Auto, Liaoning Shihua University, from 1995 to 1997, where he was an Associate Professor with the Institute of Electrics and Electronics, from 1998 to 2000, and is currently a Professor with Electrical Engineering. His current research interests include advanced process control theory and applications, state monitoring of power apparatus.

Lin Luo **https://orcid.org/0000-0002-1226-0745**

He received the B.Eng. and M.Eng. degrees from Liaoning Shihua University, Fushun, China, in 2007 and 2010, respectively, and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2015. From May to October 2014, he was a research assistant with the Sultan Qaboos University. In 2016, he became a lecturer with the Department of Electrical and Control Engineering, Liaoning Technical University. Since 2017, he has been with the Department of Information and Control Engineering, Liaoning Shihua University. His research interests include monitoring, optimization and control of industrial process, and soft sensor.

Qiao Wang **https://orcid.org/0000-0003-4389-9135**

He received the Ph.D. degree in control theory and control application from Zhejiang University, Hangzhou, China, in 2015. In 2015–2017, he holds a postdoctor position at the college of electrical engineering, Zhejiang University. Since 2017, he has been an Associate Professor with the College of Information and Control Engineering, Liaoning Shihua University. His research interests include control and monitoring of electrical power system.

Fushuo Mao **https://orcid.org/0000-0003-0636-5943**

He received the B.Eng. and M.Eng. degrees from Liaoning Shihua University, Fushun, China, in 2007 and 2010, respectively. Since 2011, he has been a Quality Manager with the Fushun Petrochemical Synthetic Detergent Factory. He is currently working in Fushun Petrochemical Synthetic Detergent Factory, China National Petroleum Corporation. His working experience from the workshop staff of Ethoxylation workshop and BOPP workshop to the director of Mechanical Engineering Department engaged in information and equipment management with the responsibility for ERP project, visualization project, equipment renovation project, and etc.