Adversarial Attacks for Neural Network-Based Industrial Soft Sensors: Mirror Output Attack and Translation Mirror Output Attack

Lei Chen, *Graduate Student Member, IEEE*, Qun-Xiong Zhu, and Yan-Lin He, *Member, IEEE*

***Abstract*—Soft sensing using the neural network technique has been increasingly applied to industrial processes. Recently, the security and robustness of neural network-based soft sensors have become primary concerns. In addition, current studies indicated that neural networks are vulnerable to adversarial attacks. In other words, small perturbations imposed on the input can lead to significant deviations in the output. If a soft sensor for key process variables is attacked, considerable damage may be brought to industrial processes. This article focuses on the attack methods for neural network-based industrial soft sensors. Considering the characteristics of industrial soft sensors, this article proposes two new adversarial attack methods. The first method, called the mirror output attack (MOA), is a subtle attack method that flips the output curve to change the direction of outputs. The second method, called the translation MOA (TMOA), is easy to make operators misoperate. TMOA translates the output curve while flipping the output curve to achieve the purpose of changing the output conditions. The effectiveness of MOA and TMOA is demonstrated in an industrial case study of the sulfur recovery unit process. Simulation results show that the neural network-based industrial soft sensors can be attacked by both the proposed adversarial attack methods. The study of adversarial attack methods can provide a basis for defending against attacks, thereby enhancing the security and robustness of soft sensors.**

***Index Terms*—Adversarial attack, mirror output attack (MOA), soft sensing, translation mirror output attack (TMOA).**

I. Introduction

MONITORING and controlling key quality variables is an effective way to ensure industrial process safety and improve product yield [1]. However, direct measurement of key quality variables can be challenging due to sensor limitations and the complexity of industrial sites. Fortunately, some auxiliary variables, such as temperature, pressure, and flow, can be easily measured. Soft sensors can predict key quality variables by using a mathematical model that maps related auxiliary variables to key quality variables [2]. The aim of soft sensors is to provide accurate measurements of key quality variables; in other words, to make the soft sensor outputs as close as possible to the actual values. With the aid of soft sensors, operators can have an overview of the current state of the industrial process. Currently, there are two main types of soft sensors: the knowledge-driven soft sensor and the data-driven soft sensor [3], [4]. The knowledge-based soft sensor uses physics equations based on the principles of industrial processes. Unfortunately, the large scale of modern industrial processes makes it challenging to fully understand the underlying principles and mechanisms. Consequently, developing and implementing knowledge-based soft sensors can be a time-consuming and costly process. [5]. On the other hand, the advent of the industrial Big Data era has provided the foundation for data-based soft sensors, making data-based soft sensors increasingly popular in industrial processes. In data-based soft sensors, the neural network (NN) technique has been widely used due to its strong ability in nonlinear regression [6], [7]. Several NN models such as recurrent NN (RNN) [8], long short-term memory [9], convolutional NN [10] and autoencoder [11] have been widely developed as soft sensors [12].

While NNs have shown great performance in industrial soft sensors, recent research has revealed that NNs have vulnerabilities in security. NNs are susceptible to adversarial samples [13], [14], [15]. The adversarial samples are carefully designed samples that can be used to cause significant changes in the output of the NN by adding imperceptible perturbations to the input. This type of attack is known as adversarial attack [16]. In recent years, a great deal of research has been done in the field of adversarial attacks, including the fast gradient sign method (FGSM) proposed by Goodfellow et al. [17], the projected gradient descent (PGD) method proposed by Madry et al. [18], the Carlini and Wagner attacks proposed by Carlini and Wagner [19], and the DeepFool attack proposed by Moosavi-Dezfooli et al. [20]. In addition, Xiao et al. [21] proposed an attack method called adversarial generative adversarial networks (AdvGAN) that uses the structure of the generative adversarial network (GAN).

Manuscript received 22 February 2023; revised 4 May 2023 and 11 June 2023; accepted 28 June 2023. Date of publication 3 July 2023; date of current version 19 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62073022 and Grant 61973024, and in part by the Fundamental Research Funds for the Central Universities under Grant JD2327. Paper no. TII-23-0606. *(Corresponding author: Yan-Lin He.)*

The authors are with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China, and also with the Engineering Research Center of Intelligent PSE, Ministry of Education of China, Beijing 100029, China (e-mail: ielabchenlei@163.com; zhuqx@mail.buct.edu.cn; heyanlin@ieee.org).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2023.3291717.

Digital Object Identifier 10.1109/TII.2023.3291717

TABLE I
SOME PROPERTIES OF POPULAR ATTACK METHODS

Method	Attack Type	Attack Frequency	Attack Task	Attack Target	Applicable To Regression
FGSM [17]	Gradient-Based	One-step	Classification	Both possible	Yes
PGD [18]	Gradient-Based	Iterative	Classification	Both possible	Yes
C&W [19]	Optimization-Based	Iterative	Classification	Targeted	No
DeepFool [20]	Optimization-Based	Iterative	Classification	Targeted	No
AdvGAN [21]	GAN-Based	Iterative	Classification	Both possible	Yes
DAO [26]	Gradient-Based	One-step	Soft sensor	Untargeted	Yes
IDAO [26]	Gradient-Based	Iterative	Soft sensor	Untargeted	Yes

Although adversarial attacks have aroused the interest of researchers, the main application has been focused on classification tasks. In the field of soft sensors, there has been little research. In the actual industrial process, NN-based soft sensors have been widely adopted. The safety of production relies on the accuracy and robustness of the NN-based soft sensors used. However, due to the sensitivity of NNs to adversarial samples, NN-based soft sensors have limitations. If the NN-based soft sensor is attacked, significant damage could be brought to industrial processes [22]. Furthermore, the intelligent upgrading of industry in recent years has broken down the "isolated island of information" of traditional industrial control systems (ICSs). Instead, ICSs connect production equipment through the industrial Internet, providing convenience for producers but also creating loopholes for attackers. Unfortunately, there are weak network security mechanisms and protective measures in ICSs. As a result, many factories and equipment still have significant security vulnerabilities and risks, making it possible for attackers to launch attacks against soft sensors [23], [24], [25]. Considering the unique characteristics of industrial processes, Kong and Ge [26] proposed directly attack output (DAO) and iterative DAO (IDAO) to attack NN-based soft sensors.

In this article, we propose two attack methods based on AdvGAN, one is called mirror output attack (MOA), and the other is called translation MOA (TMOA) in this article. The contributions of this article are summarized as follows.

1) For the first time, the AdvGAN-based adversarial attack method is introduced into the field of industrial soft sensors. Due to its black box attack ability, AdvGAN can be effectively utilized to attack industrial soft sensors.
2) Two methods of mirroring output attacks are proposed in this article. The first attack method is for a single working condition called MOA. MOA changes the direction of the output so that the adversarial output is symmetrical with the actual output without changing the output interval. The second attack method, TMOA, is designed for multiple working conditions. Based on MOA, the output value is not only symmetrically flipped, but the interval of output is also changed in a controlled manner. Thus, the operator may mistakenly believe that the working state has changed.
3) An industrial case sulfur recovery unit (SRU) [27] is adopted to verify the attack effect of the proposed MOA and TMOA. By comparing with the FGSM- and PGD-based methods, MOA and TMOA are more aggressive in terms of white box attacks, and more reasonable and covert in terms of black box attacks. Thereby, MOA and TMOA have the potential to cause serious harm to industrial processes.

II. BRIEF SUMMARY OF CURRENT METHODS

In recent years, adversarial attacks have gained increasing attention in the field of the NN, particularly in the area of classification. Table I presents several popular adversarial attack methods, including FGSM [17], PGD [18], C&W [19], DeepFool [20], and AdvGAN [21] for classification, as well as DAO and IDAO [26] for the soft sensor. FGSM and PGD are two classic gradient-based adversarial attack methods that can set the output target or simply make the output different from the original output. FGSM generates adversarial examples in one step, while PGD generates perturbations by taking multiple small steps. Gradient-based attack methods require access to the gradient information, which means that the attacker needs to know the structure and parameters of the NN. However, in practice, attackers may only have access to the inputs and outputs of the NN, making gradient-based attack methods unable to attack black box models. C&W and DeepPool are two optimization-based targeted adversarial attack methods that rely on an optimization algorithm to generate adversarial perturbations. Optimization-based algorithms are very effective in classification problems, but their applicability is poor in regression problems. This is because regression problems require predicting continuous values, and these optimization algorithms often find it difficult to ensure that the generated perturbations are also continuous. AdvGAN performs exceptionally in generating high-quality adversarial samples, and it can work on black box models, allowing attackers to attack models without model access permissions. In regression problems, AdvGAN can use the regression loss such as mean square error (MSE) to measure the deviation of predicted values and use generator networks to generate adversarial perturbations with continuous value.

However, the adversarial attack method for classification cannot be simply transplanted to the soft sensor. This is because the adversarial attack against classification only considers whether the classification result is correct. But for soft sensors, it is not enough to only require the output of NN to differ greatly from the real value. Industrial processes are usually steady state,

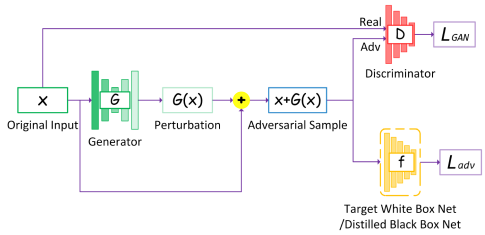

Fig. 1. Structure of AdvGAN.

and the output of adversarial examples also needs to be an orderly steady state that meets the requirements of the actual process. Otherwise, the operator will easily detect the problem, and the attack will fail. Aiming at this problem, DAO and IDAO methods were proposed to attack soft sensors. DAO is an attack method based on FGSM, while IDAO is an iterative version of DAO based on PGD. DAO and IDAO make the output maximum or minimum so as to smooth the output and achieve the purpose of the attack. However, DAO and IDAO still have some shortcomings. First, DAO and IDAO maximize or minimize the output to change the interval of output, but have no control over the size of the interval. The premise of the adversarial attack is that the perturbations applied to the input are imperceptible. If the input value changes modestly while the output value changes significantly, the operator may detect the problem, and then, the attack fails. At the same time, if the output is not constrained, it may cause nonconformance with the process, such as minimizing the output to a negative value or maximizing the output to exceed the upper limit of the actual process. Second, DAO and IDAO are methods based on FGSM and PGD, which are effective in white box attacks. However, when confronted with black box attacks, the effect of FGSM or PGD becomes unacceptable.

In this study, we adopt AdvGAN as the basic method to implement adversarial attacks against soft sensors. The structure of AdvGAN is shown in Fig. 1. In AdvGAN, a target NN f is added on the basis of a generator G and a discriminator D. The core idea of AdvGAN is to map the original samples $\boldsymbol{x}$ into adversarial perturbations $G(\boldsymbol{x})$ through the generator G. Next, perturbations are added to the corresponding original samples. The discriminator D is responsible for judging whether the samples are adversarial samples. The purpose of the discriminator is to encourage the generated adversarial samples $\boldsymbol{x}+G(\boldsymbol{x})$ to be indistinguishable from original samples $\boldsymbol{x}$. The target loss function L_{GAN} of discriminator D represents the difference between $\boldsymbol{x}$. and $\boldsymbol{x}+G(\boldsymbol{x})$. The target loss function L_{adv} of the attacked network f represents the difference between the output of f and the actual value.

$$L_{\text{GAN}} = E_x \log D(\boldsymbol{x}) + E_x \log\left(1 - D(\boldsymbol{x} + G(\boldsymbol{x}))\right) \quad (1)$$

$$L_{\text{adv}} = -E_x l_f(\boldsymbol{x} + G(\boldsymbol{x}), t) \quad (2)$$

where l_f is the objective function of the original network f, and t is the label of the target. In order to limit the size of perturbations, an additional loss L_{hinge} is added

$$L_{\text{hinge}} = E_x \max\left(0, \|G(\boldsymbol{x})\|_2 - c\right) \quad (3)$$

where c is a constant that controls the size of perturbations.

$$L = L_{\text{GAN}} + \alpha L_{\text{adv}} + \beta L_{\text{hinge}} \quad (4)$$

where α and β are constants, indicating the relative importance of each target.

III. Proposed Adversarial Attack Methods: MOA and TMOA

The primary goal of an adversarial attack on soft sensors is to reduce the accuracy of soft sensors, resulting in a large error between the output of soft sensors and the actual value. Meanwhile, industrial processes are usually in a steady state, which means that the output does not fluctuate much. Therefore, the attacked output should also be in a steady state so as to deceive the operator. In order to achieve the abovementioned requirements, two adversarial attack methods are proposed and designed for NN-based soft sensors in this article. The first adversarial attack method is called MOA, which flips the output and changes the curve direction of the output. MOA can ensure that the output is still in a steady state while increasing the error between the output of soft sensors and the actual value. The second method is called as translation MOA (TMOA) that is adopted to change the interval of the output on the basis of changing the direction of the output curve. TMOA can be utilized to make the output large or small and control the interval of the output according to the requirements of the attacker. In industrial production processes, there are multiple working conditions. Different conditions are reflected in the different intervals of process data. The on-site operator needs to adjust the operation constantly to different working conditions. With a large or small output after an adversarial attack, the operator can be deceived into believing that the working conditions have changed, resulting in misoperations.

A. Mirror Output Attack

MOA is designed based on the structure of AdvGAN. Similar with AdvGAN, the objective function of MOA is also composed of three parts, L_{GAN}, L_{adv} and L_{hinge}. The aim of L_{GAN} is to make the discriminator not distinguish between the generated samples and the original samples. Thus, the perturbations generated by the generator cannot be detected. The purpose of L_{hinge} is to further limit the perturbations to make the perturbations undetectable. L_{GAN} and L_{hinge} are the same as (1) and (3) and L_{adv} needs to be adapted to the purpose of mirroring the output.

The optimization object for traditional soft sensors can be expressed as

$$\min_{\boldsymbol{\phi}} L(\boldsymbol{y}, f(\boldsymbol{x}; \boldsymbol{\phi})) \quad (5)$$

where L denotes the loss function that is generally the MSE loss, f is the target NN parameterized by $\boldsymbol{\phi}$, $\boldsymbol{x}$ is the input sample, $\boldsymbol{y}$ is the actual output, and $f(\boldsymbol{x}; \boldsymbol{\phi})$ is the soft sensor output. The optimization objective of (5) can be understood as finding the best parameters $\boldsymbol{\phi}$ of NN to minimize the error between the actual value and the soft sensor output. In order to achieve MOA, the optimization objective needs to be modified.

When one develops NN-based soft sensors, the output can be standardized to the range of $[-1, 1]$. In this way, the purpose of mirror output can be achieved by changing the expected output of NN-based soft sensors to the inverse number of the actual output. As a result, a new optimization objective is used in the proposed MOA

$$\min_{\boldsymbol{\eta}} L\left(-\boldsymbol{y}, f\left(\boldsymbol{x}+\boldsymbol{\eta}; \boldsymbol{\phi}\right)\right)$$
$$\text{s.t. } \|\boldsymbol{\eta}\|_{\infty} \leq \varepsilon \tag{6}$$

where $\boldsymbol{\eta}$ is the added perturbations, $f(\boldsymbol{x}+\boldsymbol{\eta}; \boldsymbol{\phi})$ is the adversarial output of the soft sensor, and ε is a constant. This optimization objective can be understood as seeking the best perturbations $\boldsymbol{\eta}$ with threshold ε so that the adversarial output is close to the inverse number of the actual output. In this way, (2) can be rewritten to meet the optimization objective

$$L_{\text{adv}} = E_x l_f\left(\boldsymbol{x}+G\left(\boldsymbol{x}\right), -\boldsymbol{y}\right). \tag{7}$$

So, the loss function of MOA can be expressed as

$$\begin{aligned} L_{\text{MOA}} = {} & E_x \log D\left(\boldsymbol{x}\right) + E_x \log\left(1 - D\left(\boldsymbol{x}+G\left(\boldsymbol{x}\right)\right)\right) \\ & + \alpha E_x l_f\left(\boldsymbol{x}+G\left(\boldsymbol{x}\right), -\boldsymbol{y}\right) \\ & + \beta E_x \max\left(0, \|G\left(\boldsymbol{x}\right)\|_2 - \varepsilon\right). \end{aligned} \tag{8}$$

B. Translation Mirror Output Attack

The designed TMOA is also based on the structure of Adv-GAN. Compared with MOA, (1) and (3) remain the same and only (2) needs to be modified. TMOA can be adopted to achieve the translation of the output interval on basis of MOA, which can be regarded as a change in working conditions in actual industrial processes.

After standardization, most of the data are concentrated in the interval of $[-1, 1]$. In TMOA, if the adversarial output goes upward, the corresponding value should turn larger by adding a constant. Otherwise, if the adversarial output goes downward, the value should turn smaller by subtracting a constant. The size of the constant value can be set according to the requirements of the attack. A new optimization objective is adopted in the designed TMOA

$$\min_{\boldsymbol{\eta}} L\left(n-\boldsymbol{y}, f\left(\boldsymbol{x}+\boldsymbol{\eta}; \boldsymbol{\phi}\right)\right)$$
$$\text{s.t. } \|\boldsymbol{\eta}\|_{\infty} \leq \varepsilon \tag{9}$$

where n represents the added constant and n can be positive or negative. This optimization objective can be understood as seeking the best perturbations $\boldsymbol{\eta}$ with threshold ε so that the adversarial output is close to the mirror output adding or subtracting the constant. In this way, (2) can be rewritten to meet the optimization objective

$$L_{\text{adv}} = E_x l_f\left(\boldsymbol{x}+G\left(\boldsymbol{x}\right), n-\boldsymbol{y}\right). \tag{10}$$

So, the loss function of TMOA can be expressed as

$$\begin{aligned} L_{\text{MOA}} = {} & E_x \log D\left(\boldsymbol{x}\right) + E_x \log\left(1 - D\left(\boldsymbol{x}+G\left(\boldsymbol{x}\right)\right)\right) \\ & + \alpha E_x l_f\left(\boldsymbol{x}+G\left(\boldsymbol{x}\right), n-\boldsymbol{y}\right) \\ & + \beta E_x \max\left(0, \|G\left(\boldsymbol{x}\right)\|_2 - \varepsilon\right). \end{aligned} \tag{11}$$

Algorithm: MOA and TMOA.

Input:
$\boldsymbol{x}$: original input samples
$\boldsymbol{y}$: original output samples
f : the attacked soft sensor with parameters $\boldsymbol{\phi}$
ε : the value of threshold
Output:
$\boldsymbol{x}+\boldsymbol{\eta}$: adversarial samples
Process:
Standardize $\boldsymbol{x}$ and $\boldsymbol{y}$
for number of iterations **do**
  for k steps **do**
    Sample a batch of m examples $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}\}$;
    Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_D} \frac{1}{m} \sum_{i=1}^{m} \begin{bmatrix} \log D(\boldsymbol{x}^{(i)}) \\ +\log(1 - D(G(\boldsymbol{x}^{(i)}) + \boldsymbol{x}^{(i)})) \end{bmatrix}$$

  end for
  Sample a batch of m examples $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}\}$;
  Update the generator by descending its stochastic gradient:
  For MOA:

$$\nabla_{\theta_G} \frac{1}{m} \sum_{i=1}^{m} \begin{bmatrix} \log(1 - D(G(\boldsymbol{x}^{(i)}) + \boldsymbol{x}^{(i)})) \\ +\alpha l_f(\boldsymbol{x}^{(i)} + G(\boldsymbol{x}^{(i)}), -\boldsymbol{y}^{(i)}) \\ +\beta \max(0, \|G(\boldsymbol{x})\|_2 - \varepsilon) \end{bmatrix}$$

  For TMOA:

$$\nabla_{\theta_G} \frac{1}{m} \sum_{i=1}^{m} \begin{bmatrix} \log(1 - D(G(\boldsymbol{x}^{(i)}) + \boldsymbol{x}^{(i)})) \\ +\alpha l_f(\boldsymbol{x}^{(i)} + G(\boldsymbol{x}^{(i)}), n - \boldsymbol{y}^{(i)}) \\ +\beta \max(0, \|G(\boldsymbol{x})\|_2 - \varepsilon) \end{bmatrix}$$

end for
$\boldsymbol{\eta} = G(\boldsymbol{x})$;
Clip the adversarial perturbations $\boldsymbol{\eta}$ so that $\|\boldsymbol{\eta}\|_{\infty} \leq \varepsilon$.

The pseudocodes of MOA and TMOA are shown in Algorithm.

C. Practical Deployment Process of Adversarial Attack

According to the background knowledge of the attacked NN-based soft sensor model, the adversarial attack can be divided into the white box attack and the black box attack. For white box attacks, the structure and parameters of the NN-based soft sensor are known. Thus, perturbations can be generated directly from the NN-based soft sensor. Therefore, the white box attack is easy to achieve. In contrast, black box attacks have no knowledge of the NN-based soft sensor. Thus, a distilled model needs to be built to achieve attacks. The distilled model is trained by querying the input and output of the NN-based soft sensor. The purpose of the distilled model is to copy or approximate the NN-based soft sensor. The attack strategy of the white box attack is then carried out on the well-built distilled model. Hence, compared with white box attacks, the deployment of black box attacks is more complex. In addition, some necessary statements and assumptions are given to adversarial attacks on NN-based soft sensors. It is important to note that research on adversarial

attacks is based on these statements and assumptions, otherwise the research is not valid.

Statement 1: Both MOA and TMOA attack methods target NN-based soft sensors. Currently, all adversarial attack methods are aimed at NN-based models.

Statement 2: The queried dataset refers to the auxiliary data used for soft sensors during the online period of industrial processes. The queried dataset can be understood as the test samples in the online test stage. The generated adversarial perturbations need to be imposed on the queried dataset.

Assumption 1: The attacker has the right to get and modify the queried dataset.

Assumption 2: The black box attack can query the historical input and output of the NN-based soft sensor to establish the distilled model.

Based on the abovementioned statements and assumptions, the specific steps to deploy MOA and TMOA in white box attacks are as follows.

Step 1: Obtain the queried dataset and the knowledge of the NN-based soft sensor from ICS.

Step 2: Generate adversarial perturbations based on the queried dataset and the NN-based soft sensor. Impose the adversarial perturbations to the queried dataset to obtain the adversarial dataset.

Step 3: Input the adversarial dataset into the NN-based soft sensor to get the adversarial output.

The specific steps to deploy MOA and TMOA in black box attacks are as follows.

Step 1: Obtain the queried dataset and the historical input and output of the NN-based soft sensor from ICS.

Step 2: Train the distilled model by the historical input and output of the NN-based soft sensor.

Step 3: Generate adversarial perturbations based on the queried dataset and the distilled model. Impose the adversarial perturbations to the queried dataset to obtain the adversarial dataset.

Step 4: Input the adversarial dataset into the NN-based soft sensor to get the adversarial output.

According to the abovementioned discussion of the white box attack and the black box attack, Fig. 2 shows the deployment diagram of the white and black box attacks.

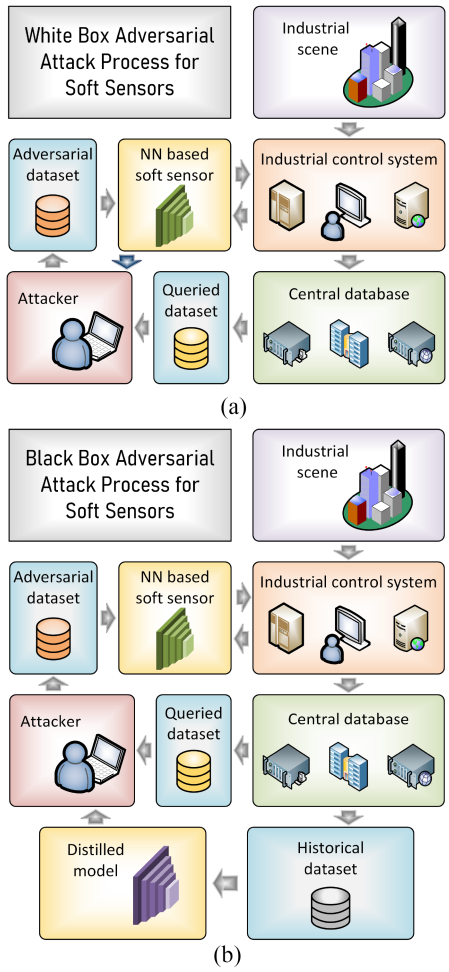

Fig. 2. Deployment process of adversarial attack for soft sensors. (a) White box attack. (b) Black box attack.

IV. Case Study

In this section, a set of real industrial dataset SRU [27] is used to verify the effectiveness of the proposed MOA and TMOA in the white box attack and the black box attack, respectively. Two adversarial attack algorithms, FGSM and PGD, are used for comparison. That is, MOA and TMOA based on FGSM and PGD will be used as comparison methods.

A. SRU Process

Acid gas is extremely dangerous. SRU absorbs the acid gas before it is discharged, and recovers the sulfur in the acid gas. The flowchart of SRU is shown in Fig. 3. SRU takes MEA containing H_2S and SWS containing H_2S and NH_3 as inputs, and collects sulfur to generate SO_2. The tail gas of SRU contains residual H_2S

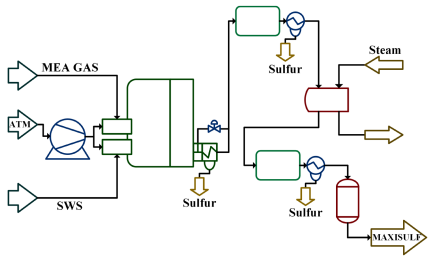

Fig. 3. Flowchart of SRU [27].

and SO_2. It is required to measure H_2S and SO_2 in time to control SRU. However, H_2S and SO_2 are easy to damage the sensors, so it is very necessary to carry out soft sensors for H_2S and SO_2 during sensor maintenance. In this article, SO_2 is selected as the output. According to the relevant process knowledge, five input auxiliary variables are selected, which are x_1: gas flow MEA_GAS, x_2: air flow AIR_MEA, x_3: secondary air flow AIR_MEA_2, x_4: gas flow in the SWS zone, and x_5: air flow in the SWS zone. Considering the process dynamics, the data are temporalized. In other words, in order to obtain the sampling of the output at time t, the previous sampling data with hysteresis

are added to the input variable. The formula is shown in (12). There are 4000 sets of data, 70% of which are used for training and the rest for adversarial attack.

$$y(t) = f\begin{pmatrix} x_1(t), x_1(t-5), x_1(t-10), x_1(t-15) \\ x_2(t), x_2(t-5), x_2(t-10), x_2(t-15) \\ \vdots \\ x_5(t), x_5(t-5), x_5(t-10), x_5(t-15) \end{pmatrix} \tag{12}$$

where $x_i(t-n)$ represents the sampling of the ith input at $t-n$ time.

Before simulation, the data are standardized according to (13). Three indicators are used to evaluate the output results: the mean absolute error (MAE), root mean square error (RMSE), and decision coefficient R^2

$$\boldsymbol{x} = \frac{\boldsymbol{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{13}$$

where $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\sigma}$ is the variance.

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i' - y_i| \tag{14}$$

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i' - y_i)^2} \tag{15}$$

$$R^2 = \frac{\sum_{i=1}^{m}(y_i' - \bar{y})^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{m}(y_i - y_i')^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{16}$$

where m is the total number of samples, y' is the output value of the model, and $\bar{y}$ is the average value of the output actual value.

B. White Box Attack Case

The details of the NN-based soft sensor in white box attacks are known. We assume that the NN-based soft sensor is a deep NN (DNN) model. The structure of DNN is 20-30-50-30-20-1, and the activation function is rectified linear unit (LeakyReLU). In MOA, the structure of the discriminator is 20-30-50-30-20-10-1 and the activation function is Sigmoid. The structure of the generator is 20-30-40-50-30-20, and the activation function is tanh. The perturbation size is set to $[-0.3, 0.3]$. Both α and β are set to 1. TMOA has the same structure as MOA. n in TMOA is set to 1.5, which means translating the output curve upward. At the same time, using AdvGAN as a comparison, the structure of AdvGAN is the same as MOA. The loss function of AdvGAN is changed from the cross-entropy loss used for classification to the MSE loss used for regression, and the rest is the same as the original AdvGAN. The number of iterations for PGD is set to 1000 for MOA. Like MOA and TMOA, the perturbation threshold of FGSM and PGD is set to 0.3.

Fig. 4 shows the distribution of the output of DNN and the actual output. It can be seen that the DNN is able to track the actual values well (in fact, the DNN is just an NN-based soft sensor with acceptable accuracy). The simulation results can be seen in Table II. Although it can be seen from Table II that AdvGAN has the largest attack errors, it is not suitable for adversarial attacks.

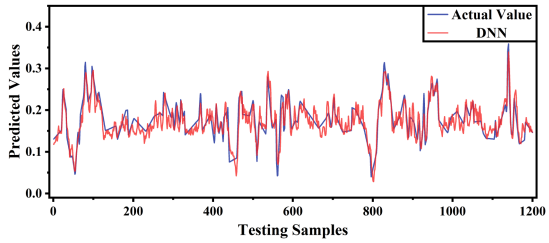

Fig. 4. Distributions of DNN and the actual value.

TABLE II
SIMULATION RESULT

Method	The concentration of SO_2		
	RMSE	MAE	R^2
DNN	0.0158	0.0199	0.7886
AdvGAN	0.1504	0.1367	-10.971
FGSM-MOA	0.0439	0.0340	-0.0232
PGD-MOA	0.0683	0.0525	-1.4715
MOA	**0.0504**	**0.0713**	**-1.6951**
FGSM-TMOA	0.0903	0.0869	-3.3149
PGD-TMOA	0.1057	0.0922	-4.9156
TMOA	**0.0900**	**0.1094**	**-5.3365**

The bold entities represent the proposed two methods in this paper.

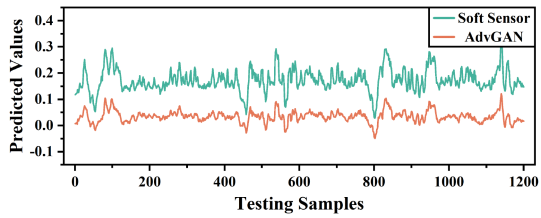

Fig. 5. Distributions of AdvGAN and the soft sensor.

The output curves for AdvGAN and the soft sensor are shown in Fig. 5. It can be seen that the output of AdvGAN is negative at some sampling points, which is not reasonable. It is easy for the operator to detect the failure of the soft sensor, leading to the failure of the attack. Thus, it is not feasible to simply increase the gap between the adversarial output and the output of the soft sensor. Further restrictions on the output are required. From Table II, it can be seen that MOA and TMOA have larger output errors compared to the FGSM- and PGD-based methods. To show the advantages of MOA and TMOA intuitively, Figs. 6 and 7 are drawn. In Fig. 6, the black line is the mean of the output of the soft sensor. In Fig. 7, the black line is the mean of all outputs of TMOA and the soft sensor. As can be seen from Fig. 6(a), FGSM-MOA only achieves mirroring attacks at a small number of sampling points. In Fig. 6(b), compared with FGSM-MOA, PGD-MOA implements mirroring attacks at most sampling points, but the output curve is relatively smooth. In Fig. 6(c), MOA achieves mirroring of the output at all sampling points. The output fluctuations of MOA are similar to the outputs of the soft sensor, which can confuse the operator. As shown in Fig. 7, TMOA implements the purpose of translation and mirroring the output. Meanwhile, the output fluctuations of TMOA are similar to the outputs of the soft

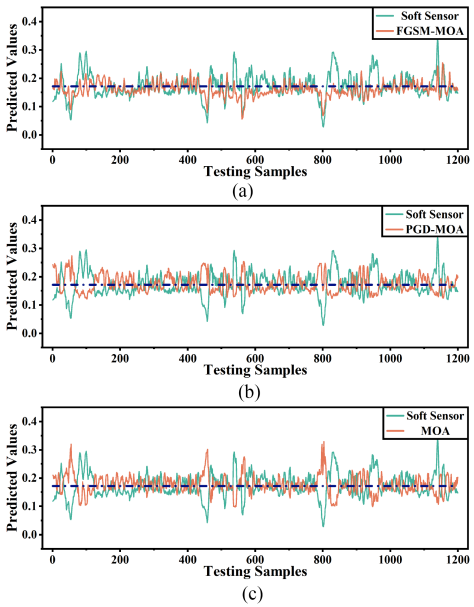

Fig. 6. Distributions of the original output and the adversarial output of (a) FGSM-MOA, (b) PGD-MOA and (c) MOA.

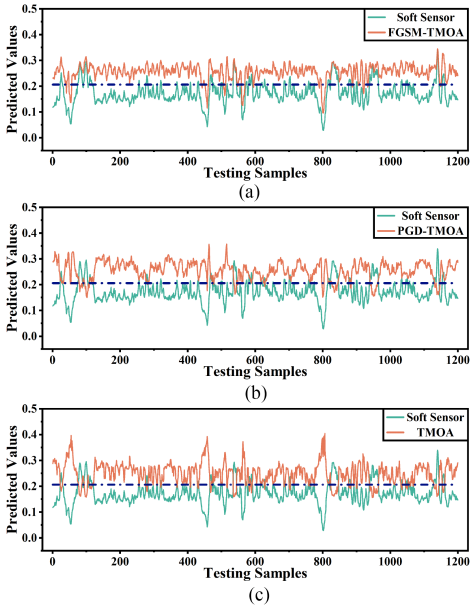

Fig. 7. Distributions of the original output and the adversarial output of (a) FGSM-TMOA, (b) PGD-TMOA and (c) TMOA.

sensor. It can be seen that the one-time attack strength of FGSM-based methods is insufficient. The iterative attack strength of PGD-based methods has increased, but there is still room for improvement. The two gradient-based attack methods will show greater drawbacks in the next section on black box attacks.

Fig. 8 shows the changes in the output of the NN-based soft sensor after being attacked. The green line in the first half refers to the DNN output before the attack, and the orange line in the second half refers to the output after the attack. The black dotted line refers to the mean value of the output curve. As can be seen from Fig. 8(a), the mean value of MOA is almost the same as the original mean value. Moreover, the output of the soft

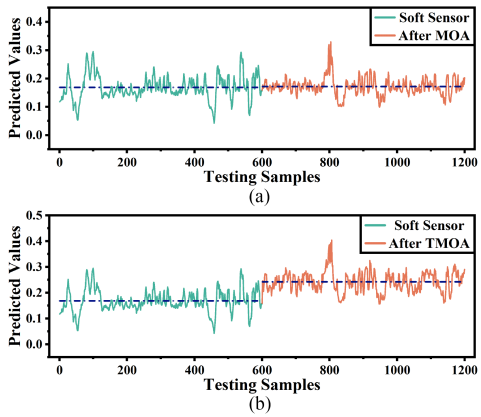

Fig. 8. Output curve after being attacked. (a) MOA. (b) TMOA.

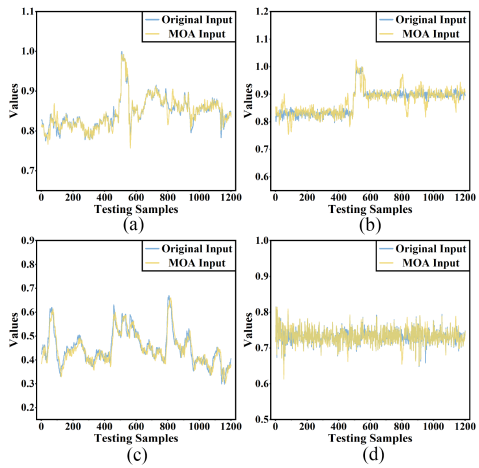

Fig. 9. Adversarial samples of MOA. (a) Input #1. (b) Input #2. (c) Input #3. (d) Input #4.

sensor remains steady state even after the attack. This means that the soft sensor has been attacked without the operators being aware. It can be seen from Fig. 8(b) that the mean value of the output changes after TMOA, which indicates that the output switches from the original steady state to a new steady state. TMOA implements the switching of working conditions. It is easy for operators to misoperate, thus causing problems in industrial processes.

An important premise for a successful adversarial attack is that the perturbations imposed are small and hard to be detected. Figs. 9 and 10 show the distribution of original samples and adversarial samples generated by MOA and TMOA. It can be seen that the adversarial samples almost coincide with the original samples, which ensures that the adversarial samples are imperceptible. Due to the length limitation of the article, only the first four input variables are shown here. In fact, the remaining input variables are the same as the variables shown, which almost coincide with the original samples.

C. Black Box Attack Case

Unlike white box attacks, the details of the NN-based soft sensor in black box attacks are not known. First, attackers establish a distilled model, DNN, with a structure of 20-30-20-15-10-1.

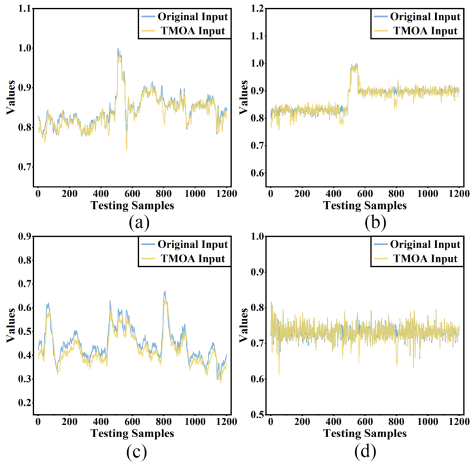

Fig. 10. Adversarial samples of TMOA. (a) Input #1. (b) Input #2. (c) Input #3. (d) Input #4.

The structure of MOA and TMOA is the same as that of the white box attack. The parameters for FGSM and PGD are also set in the same way as for the white box attack.

The evaluation values of each method are given in Table SI of the supplementary material. It can be seen that MOA and TMOA are less effective in the black box attack than the white box attack. However, the NN-based soft sensor can still be attacked with the help of the distilled model. The output after the MOA and TMOA attacks is shown in Figs. S2 and S3 of the supplementary material. In Fig. S2(a) and (b), the output of FGSM-MOA and PGD-MOA is very disordered. It is easy for the operator to realize that there is an attack in the soft sensor, which makes the attack fail. In Fig. S2(c), MOA achieves mirroring of the output and the output of MOA is steady state. Then, as can be seen in Fig. S3(a) and (b), neither FGSM-TMOA nor PGD-TMOA meet the objective of translating upwards and mirroring the output. Also, some outputs are negative, making the output unreasonable and the attack of FGSM-TMOA and PGD-TMOA fail. TMOA implements the purpose of translation and mirroring the output in Fig. S3(c), and the output is steady state. Therefore, in terms of black box attack, MOA and TMOA still complete the attack on the NN-based soft sensor.

D. Summary and Measures for Preventing Attacks

This article proposes two novel adversarial attack strategies, MOA and TMOA. Simulation results indicate that these methods can successfully attack the NN-based soft sensor. This is a concern because such attacks can lead to incorrect outputs, which can have serious consequences in critical applications such as ICS. The NN-based soft sensor is particularly vulnerable to adversarial attacks. Therefore, it is crucial to identify potential vulnerabilities and develop effective defenses.

Some specific ways to prevent attacks on the NN-based soft sensor are as follows.

1) *Identifying vulnerabilities:* Subjecting the NN-based soft sensor to various types of adversarial attacks and analyzing its structure and parameters to pinpoint potential weaknesses that can be exploited by attackers.
2) *Developing defenses:* Using adversarial samples to train the NN-based soft sensor to be more robust to attacks, including techniques such as data augmentation, adversarial training, and regularization. In addition, methods such as input sanitization, anomaly detection, and model ensembling can be used to detect and filter out adversarial inputs.
3) *Improving accuracy and robustness:* Testing the NN-based soft sensor against adversarial attacks can help to identify weaknesses and refine the network to make it more accurate and robust, including adjusting the structure or hyperparameters of the NN-based soft sensor or using transfer learning to improve its generalization ability.
4) *Enhancing security standards:* Developing new security standards and guidelines for NN-based soft sensors, including performing security audits, implementing secure coding practices, and developing protocols for secure data transmission and storage.

In addition, some concrete countermeasures have been provided for operator training as follows.

1) Educating operators about the nature of adversarial attacks and their potential impact on soft sensor outputs, as well as providing them with strategies to identify and mitigate such threats.
2) Training operators to recognize suspicious changes in output curves, which may indicate an ongoing attack. This could include the use of visual aids, such as highlighting abnormal trends or deviations in the output data.
3) Providing operators with guidelines on how to respond to potential threats, such as verifying the output of the soft sensor with alternative measurement tools or even temporarily shutting down the process for further investigation.
4) Exploring alternative output representations, such as quantitative values or different graphical representations to help operators better detect subtle changes in the output that may indicate an attack.

V. Conclusion

This article investigates the adversarial attack of NN-based soft sensors. According to the actual characteristics of industrial processes, two attack methods MOA and TMOA are proposed. The MOA implements the flip of the output curve of the soft sensor while the output is ordered and is steady state. On the basis of mirroring the output, TMOA implements the translation of the output curve to change the working condition, which makes it easier for operators to misoperate. In addition, this article uses an industrial case to validate the proposed approach. The simulation results show that MOA and TMOA can produce reasonable attacks. In terms of white box attacks, the proposed methods are powerful. In terms of black box attacks, only the proposed MOA and TMOA are able to achieve the set attack goals. The compared methods only produce disordered and irrational attack outputs.

This research aims to expose the vulnerability of soft sensors. The ultimate goal of making soft sensors more accurate, safer and more robust will be achieved through subsequent research. Methods such as adversarial training, noise reduction, and anomaly detection on input may be able to achieve defense against such attacks. Work on adversarial defense is left in future research.

REFERENCES

[1] C. Liu, K. Wang, Y. Wang, and X. Yuan, "Learning deep multimanifold structure feature representation for quality prediction with an industrial application," *IEEE Trans. Ind. Inform.*, vol. 18, no. 9, pp. 5849–5858, Sep. 2022.

[2] Y. He, L. Chen, and Q. Zhu, "Quality regularization based semisupervised adversarial transfer model with unlabeled data for industrial soft sensing," *IEEE Trans. Ind. Inform.*, early access, May 3, 2023, doi: 10.1109/TII.2023.3272690.

[3] X. Zhang, C. Song, J. Zhan, and Z. Xu, "Deep Gaussian mixture adaptive network for robust soft sensor modeling with a closed-loop calibration mechanism," *Eng. Appl. Artif. Intell.*, vol. 122, 2023, Art. no. 106124.

[4] X. Song, Y. He, X. Li, Q. Zhu, and Y. Xu, "Novel virtual sample generation method based on data augmentation and weighted interpolation for soft sensing with small data," *Expert Syst. Appl.*, vol. 225, 2023, Art. no. 120085.

[5] J. P. Jordanou, E. A. Antonelo, and E. Camponogara, "Echo state networks for practical nonlinear model predictive control of unknown dynamic systems," *IEEE Trans. Neural. Netw. Learn. Syst.*, vol. 33, no. 6, pp. 2615–2629, Jun. 2022.

[6] Y. He, L. Chen, Y. Gao, J. Ma, Y. Xu, and Q. Zhu, "Novel double-layer bidirectional LSTM network with improved attention mechanism for predicting energy consumption," *ISA Trans.*, vol. 127, pp. 350–360, 2022.

[7] L. Ranzan, L. F. Trierweiler, B. Hitzmann, and J. O. Trierweiler, "Avoiding misleading predictions in fluorescence-based soft sensors using autoencoders," *Chemometrics Intell. Lab. Syst.*, Vol. 223, 2022, Art. no. 104527.

[8] L. Feng, C. Zhao, and Y. Sun, "Dual attention-based encoder—decoder: A customized sequence-to-sequence learning for soft sensor development," *IEEE Trans. Neural. Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3306–3317, Aug. 2021.

[9] J. Zhou, X. Wang, C. Yang, and W. Xiong, "A novel soft sensor modeling approach based on difference-LSTM for complex industrial process," *IEEE Trans. Ind. Inform.*, vol. 18, no. 5, pp. 2955–2964, May 2022.

[10] J. M. M. de Lima and F. M. U. de Araujo, "Ensemble deep relevant learning framework for semi-supervised soft sensor modeling of industrial processes," *Neurocomputing*, vol. 462, pp. 154–168, 2021.

[11] Y. He, X. Li, J. Ma, Q. Zhu, and S. Lu, "Attribute-relevant distributed variational autoencoder integrated with LSTM for dynamic industrial soft sensing," *Eng. Appl. Artif. Intell.*, vol. 119, 2023, Art. no. 105737.

[12] Y. Jiang, S. Yin, J. Dong, and O. Kaynak, "A review on soft sensors for monitoring, control, and optimization of industrial processes," *IEEE Sens. J.*, vol. 21, no. 11, pp. 12868–12881, Jun. 2021.

[13] R. V. Mendonça et al., "Intrusion detection system based on fast hierarchical deep convolutional neural network," *IEEE Access*, vol. 9, pp. 61024–61034, 2021.

[14] A. Kurniawan, Y. Ohsita, and M. Murata, "Experiments on adversarial examples for deep learning model using multimodal sensors," *Sensors*, vol. 22, 2022, Art. no. 8642.

[15] O. D. Okey et al., "BoostedEnML: Efficient technique for detecting cyberattacks in IoT systems using boosted ensemble machine learning," *Sensors*, vol. 22, 2022, Art. no. 7409.

[16] F. Croce and M. Hein, "Minimally distorted adversarial examples with a fast adaptive boundary attack," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 2196–2205.

[17] I. Goodfellow, J. Shlens, and C. Szeged, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015, pp. 1–11.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 5849–5858.

[19] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.

[20] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2574–2582.

[21] C. Xiao, B. Li, J. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," 2018, *arXiv:1801.02610*.

[22] X. Jiang and Z. Ge, "Attacks on data-driven process monitoring systems: Subspace transfer networks," *IEEE Trans. Artif. Intell.*, vol. 3, no. 3, pp. 470–484, Jun. 2022.

[23] M. Kravchik and A. Shabtai, "Efficient cyber attacks detection in industrial control systems using lightweight neural networks and PCA," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 4, pp. 2179–2197, Jul./Aug. 2022.

[24] X. Zhou, Z. Xu, L. Wang, K. Chen, C. Chen, and W. Zhang, "APT attack analysis in SCADA systems," in *Proc. MATEC Web Conf., EDP Sci.*, vol. 173, 2018, Art. no. 01010.

[25] Y. Zhuo, Z. Yin, and Z. Ge, "Attack and defense: Adversarial security of data-driven FDC systems," *IEEE Trans. Ind. Inform.*, vol. 19, no. 1, pp. 5–19, Jan. 2023.

[26] X. Kong and Z. Ge, "Adversarial attacks on neural-network-based soft sensors: Directly attack output," *IEEE Trans. Ind. Inform.*, vol. 18, no. 4, pp. 2443–2451, Apr. 2022.

[27] L. Chen, Y. Xu, Q. Zhu, and Y. He, "Adaptive multi-head self-attention based supervised VAE for industrial soft sensing with missing data," *IEEE Trans. Automat. Sci. Eng.*, early access, Jun. 5, 2023, doi: 10.1109/TASE.2023.3281336.

Lei Chen (Graduate Student Member, IEEE) received the B.Sc. degree in automation from the College of Information Engineering, Beijing Institute of Petrochemical Technology, Beijing, China, in 2020. He is currently working toward the Ph.D. degree in control science and engineering with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing.

His research interests include soft sensor, process modeling, neural networks, and machine learning.

Qun-Xiong Zhu received the Ph.D. degree in control science and engineering from the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China, in 1996.

He is currently a Professor with the College of Information Science and Technology, Beijing University of Chemical Technology. His research interests include fault diagnosis, process modeling, soft sensor, machine learning, and computational intelligence.

Yan-Lin He (Member, IEEE) received the B.Sc. degree in automation and the Ph.D. degree in control science and engineering from the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China, in 2011 and 2016, respectively.

He is currently a Professor with the College of Information Science and Technology, Beijing University of Chemical Technology. His research interests include fault diagnosis, process modeling, soft sensor, machine learning, and computational intelligence.