polymers

MDPI

*Article*

# Enhanced Soft Sensor with Qualified Augmented Samples for Quality Prediction of the Polyethylene Process

**Yun Dai [1], Angpeng Liu [1], Meng Chen [2], Yi Liu [1,*] and Yuan Yao [3,*]**

[1] Institute of Process Equipment and Control Engineering, Zhejiang University of Technology, Hangzhou 310023, China
[2] Guangdong Basic and Applied Basic Research Foundation, Guangzhou 510640, China
[3] Department of Chemical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan
* Correspondence: yliuzju@zjut.edu.cn (Y.L.); yyao@mx.nthu.edu.tw (Y.Y.); Tel.: +886-3-5713690 (Y.Y.)

**Abstract:** Data-driven soft sensors have increasingly been applied for the quality measurement of industrial polymerization processes in recent years. However, owing to the costly assay process, the limited labeled data available still pose significant obstacles to the construction of accurate models. In this study, a novel soft sensor named the selective Wasserstein generative adversarial network, with gradient penalty-based support vector regression (SWGAN-SVR), is proposed to enhance quality prediction with limited training samples. Specifically, the Wasserstein generative adversarial network with gradient penalty (WGAN-GP) is employed to capture the distribution of the available limited labeled data and to generate virtual candidates. Subsequently, an effective data-selection strategy is developed to alleviate the problem of varied-quality samples caused by the unstable training of the WGAN-GP. The selection strategy includes two parts: the centroid metric criterion and the statistical characteristic criterion. An SVR model is constructed based on the qualified augmented training data to evaluate the prediction performance. The superiority of SWGAN-SVR is demonstrated, using a numerical example and an industrial polyethylene process.

**Keywords:** soft sensor; polymerization process; data augmentation; data selection; generative adversarial network; support vector regression



**Citation:** Dai, Y.; Liu, A.; Chen, M.; Liu, Y.; Yao, Y. Enhanced Soft Sensor with Qualified Augmented Samples for Quality Prediction of the Polyethylene Process. *Polymers* **2022**, *14*, 4769. https://doi.org/10.3390/polym14214769

Academic Editor: Andrea Sorrentino

Received: 23 October 2022
Accepted: 4 November 2022
Published: 7 November 2022



## 1. Introduction

Data-driven soft sensor models [1–7] have been applied extensively to provide important real-time information for the quality prediction of industrial polymerization processes in modern industry [8,9]. Although without an in-depth understanding of the process mechanisms, data-driven soft sensors have been constructed for difficult-to-measure product quality using easy-to-measure process measurements. Various soft sensors have received increasing attention in recent years, including partial least squares regression [10,11], Gaussian process regression [12,13], support vector regression (SVR) [14–16], and neural networks [17–21]. Among them, SVR-based soft sensors have exhibited good performance in several nonlinear regression tasks.

The reliability of training data for the efficient development of data-driven soft sensors is a key aspect [22,23]. However, only limited labeled samples are obtained in many polyethylene processes, which is a phenomenon that has received less attention than it deserves. For example, in the case of frequent changes in operating conditions, manual operations result in large measurement intervals and long settling times. Consequently, the acquisition of sufficient training samples is intractable [24–26]. With limited available training data, it is difficult to capture the process characteristics and model the relationship between product quality and operating conditions. Hence, the development of a soft sensor model with insufficient data requires further investigation.

The virtual sample generation (VSG) technique is effective in handling the problem of the insufficient construction of soft sensors with limited training data [27–30]. Several

VSG methods have been developed for data augmentation, which can be divided into three types: sampling-based, information diffusion-based, and deep learning-based VSGs. For sampling-based VSG, a typical method named bootstrap [31] has been increasingly adopted for data augmentation, owing to its simple mechanism. Nevertheless, as copies of the original samples are generated, the virtual samples that are obtained via bootstrap may not carry new information and cannot fill the gaps between samples. Information diffusion-based VSG is based on the distribution function of the sample space. Typical methods include mega-trend diffusion [32] and tree-based trend diffusion [33]. These two methods employ the information diffusion principle to derive the diffusion function and generate new samples using the fuzzy set theory. However, an appropriate diffusion function and coefficient cannot easily be determined.

Deep learning-based VSG methods have gained increasing attention in fields such as imaging and natural language processing [34]. In recent years, deep learning-based VSG methods have also been adopted in the process industry [35]. The generative adversarial network (GAN) [36–39], as one promising generative model, has been well studied and is valued for its generative properties. By generating virtual data that resemble actual data, the GAN enlarges the sample capacity to enhance the prediction performance. Although various improvements have been made in GANs, including alternative loss functions and training strategies, the training process remains unstable [40–43]. Therefore, the quality of the generated samples remains uncertain. In practice, both suitable and unsuitable data exist simultaneously among the generated candidates. The prediction performance of the model will deteriorate if unsuitable virtual samples are included in the training set. Hence, data selection for the total number of virtual candidates is significant. Jiang and Ge [42] used the Mahalanobis and Euclidean distances to measure the similarity between different samples and, subsequently, selected qualified candidates for data augmentation. However, they focused on the original individual samples separately for data selection, resulting in the dilemma of local minima. Thus, a new data selection strategy that considers the general distribution is necessary.

This study aims to develop an enhanced data augmentation soft sensor framework to meet the challenge of limited labeled samples in the polyethylene process. The proposed soft sensor is named the selective Wasserstein GAN, with gradient penalty-based SVR (SWGAN-SVR). First, to expand the sample capacity and enrich the data information, virtual samples are generated using a Wasserstein GAN with a gradient penalty (WGAN-GP) network. Owing to the instability of the model training, both suitable and unsuitable samples are generated for the task simultaneously. Subsequently, a data selection strategy is adopted for sample filtering, which is composed of two parts: the centroid metric criterion and the statistical characteristic criterion. Moreover, the selected qualified samples serve as supplements for the original samples. Without the loss of generality, SVR is adopted as the base regression model. Consequently, using qualified augmented virtual samples, a more accurate and reliable prediction model can be constructed, compared to that using only the original data.

The remainder of this paper is organized as follows. Section 2 briefly introduces the preliminaries. In Section 3, the SWGAN-SVR soft sensor is presented, along with its algorithmic implementation. Section 4 demonstrates the effectiveness of SWGAN-SVR in the polyethylene process. Finally, our concluding remarks are presented in Section 5.

## 2. Preliminaries

### 2.1. Problem Statement

It has traditionally been assumed that the amount of available training data is sufficient for soft sensor modeling, and many studies have focused on the design and improvement of modeling methods. However, it is difficult to collect sufficient samples in many situations, such as for industrial processes with large measurement intervals or during the early stages of new working conditions [5]. GANs, which are unsupervised generative models, have been adopted to generate virtual samples for data augmentation. The connection and main

differences between traditional supervised soft sensor models and data-augmentation-based candidates are illustrated in Figure 1. Traditional methods use only the available limited labeled samples, which are denoted as $\{\mathbf{Z}_O\} = \{\mathbf{X}_O, \mathbf{Y}_O\}$ for the purposes of model construction, where $\{\mathbf{X}_O\} = \{\mathbf{x}_{Oi}\}_{i=1,\ldots,M}$ and $\{\mathbf{Y}_O\} = \{y_{Oi}\}_{i=1,\ldots,M}$ represent the input and output data with $M$ samples, respectively. In contrast, data-augmentation-based soft sensors are constructed based on the augmented dataset. To address the problem of insufficient models that are established with limited training data, virtual samples, which are denoted as $\{\mathbf{Z}_G\} = \{\mathbf{X}_G, \mathbf{Y}_G\}$, are generated using a GAN, where $\{\mathbf{X}_G\} = \{\mathbf{x}_{Gi}\}_{j=1,\ldots,N}$ and $\{\mathbf{Y}_G\} = \{y_{Gi}\}_{j=1,\ldots,N}$ are the input and output data, respectively, with $N$ generated samples.

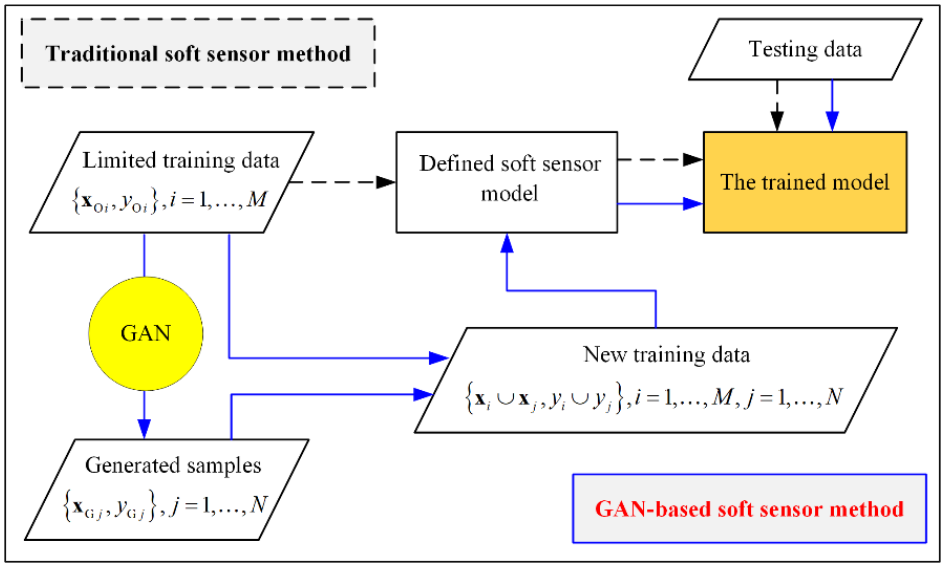



**Figure 1.** Flowchart of traditional and GAN-based soft sensor frameworks.

Unfortunately, owing to the unstable training of GANs [40–43], unsuitable samples are inevitably generated. It is worth noting that the distributions of unsuitable samples do not properly match the distribution of the real data. If these unsuitable samples are merged with the original data for establishing the model, the prediction of the soft sensor may be degraded. In this study, a data selection strategy is proposed to improve the quality of the generated samples. It is expected that a more reliable soft sensor model can be obtained by introducing these newly qualified virtual samples into the training data.

### *2.2. WGAN-GP Data Augmentation Approach*

GANs have recently attracted significant attention owing to their good distribution-learning capabilities. The vanilla GAN uses the Jensen–Shannon (JS) divergence to measure the distance between the generated and original data. However, this often causes problems, such as mode collapse and vanishing gradients [44]. To address these problems, Arjovsky et al. proposed the WGAN [44], which uses the Earth-Mover distance rather than JS divergence as a distance measurement. In the WGAN, to enforce the Lipschitz constraint, the weights of the discriminator are clipped to lie within a compact space $[-r, r]$, where $r$ is a constant. The discriminator attempts to distinguish between the real and generated samples and concentrates its parameter distribution on the two extremes of the maximum and minimum; that is, $r$ and $-r$. Consequently, a WGAN often becomes stuck in a poor regime and fails to learn.

To solve the problem caused by the weight clipping of the WGAN, Gulrajani et al. proposed a WGAN with a gradient penalty (WGAN-GP) [45]. Specifically, a penalty constraint is imposed on the gradient norm of the discriminator. The weight of the discriminator is reduced to an extremely small range using the gradient penalty strategy, which accelerates

the model convergence and solves the gradient explosion problem. The objective function of the WGAN-GP is as follows:

$$\min_G \max_D V(D,G) = \underbrace{\mathbb{E}_{\mathbf{z}_O \sim p_{\text{data}}(\mathbf{z}_O)}[D(\mathbf{z}_O)] - \mathbb{E}_{\mathbf{z}_G \sim p_G(\mathbf{z}_G)}[D(\mathbf{z}_G)]}_{\text{original WGAN}} \underbrace{-\lambda \mathbb{E}_{\hat{\mathbf{z}} \sim p_{\hat{\mathbf{z}}}}\left[\left(\| \nabla_{\hat{\mathbf{z}}} D\left(\hat{\mathbf{z}}\right) \|_2 - 1\right)^2\right]}_{\text{the penalty term}}, \quad (1)$$

where $\lambda$ is the penalty coefficient, $\hat{\mathbf{z}}$ is sampled through random interpolation on the connecting line of the original data $\mathbf{z}_O$ and generated data $\mathbf{z}_G$; that is, $\hat{\mathbf{z}} = \theta \mathbf{z}_O + (1-\theta)\mathbf{z}_G$, and $\theta$ is a random number in [0, 1].

## 3. The SWGAN-Based Soft Sensor Framework

### *3.1. Virtual Sample Selection Strategy*

Owing to the unstable training of the WGAN-GP, the quality of the generated virtual samples varies significantly. A data selection strategy is proposed for sample filtering, to eliminate the negative effects of unqualified virtual samples that are generated by the WGAN-GP for model construction. The selection strategy includes a centroid metric criterion, which is denoted as S1, and a statistical characteristic criterion, which is denoted as S2. The distribution scatters of the original and rough virtual samples are plotted in Figure 2. The distribution of most virtual samples conforms to the real data distribution. However, the WGAN-GP also generates samples that are located in regions A and B, which are far from the distribution of the original data. If the generated samples in regions A and B, which are regarded as unqualified, are added to the original set, the prediction performance of the model may deteriorate. A detailed description of the proposed selection strategy is provided below.

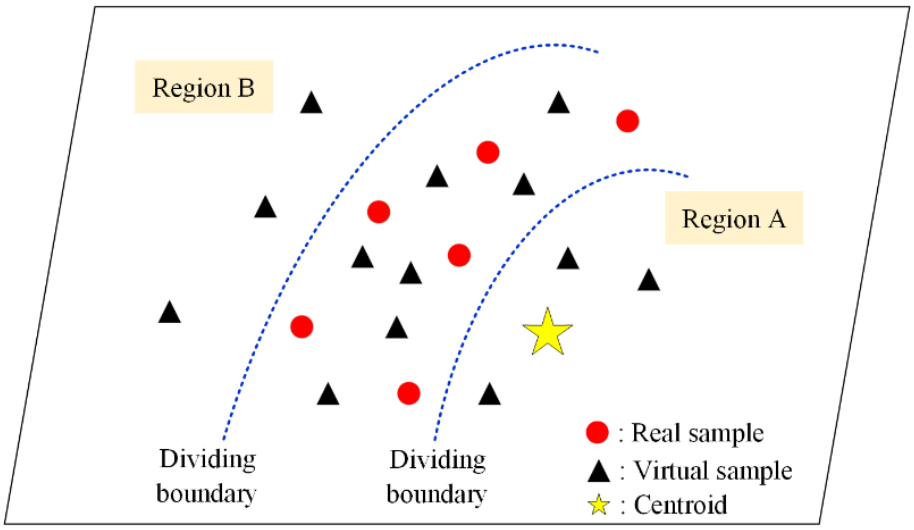



**Figure 2.** Distribution scattering of the real and virtual samples.

First, the S1 criterion is developed to filter the virtual samples in region A. The samples in region A are too close to the centroid $\mathbf{z}_C$ of the original samples. Furthermore, the distribution of these samples is not uniform compared to that of the original sample. Thus, the virtual samples around the centroid are considered to be information-poor and

unqualified. The centroid $\mathbf{z}_C$ is defined as the closest point in space to the original data, as follows:

$$\mathbf{z}_C = (\mu_{OX}, \mu_{OY}) = \frac{1}{M}(\sum_{i=1}^{M} \mathbf{x}_{Oi}, \sum_{i=1}^{M} y_{Oi}), \tag{2}$$

where $\mu_{OX}$ and $\mu_{OY}$ are the process and target variables of $\mathbf{z}_C$, respectively.

The Euclidean distance is commonly used to measure the distance between two samples. A large distance indicates that the samples are far from one another. The square of the distance between $\mathbf{z}_C$ and a finite number of original samples is formulated as follows:

$$d_C = \sum_{i=1}^{M} \|\mathbf{z}_C - \mathbf{z}_{Oi}\|_2^2 = \min(\sum_{i=1}^{M} \|\mathbf{z}_r - \mathbf{z}_{Oi}\|_2^2), \tag{3}$$

where $\mathbf{z}_{Oi}$ is the $i^{\text{th}}$ original sample and $\mathbf{z}_{Oi} = (\mathbf{x}_{Oi}, y_{Oi})$, and $\mathbf{z}_r$ is any point in space.

Similarly, the square of the distance between the $j^{\text{th}}$ generated sample and the original samples is calculated as follows:

$$d_j = \sum_{i=1}^{M} \left\|\mathbf{z}_{Gj} - \mathbf{z}_{Oi}\right\|_2^2, \tag{4}$$

where $\mathbf{z}_{Gj} = (\mathbf{x}_{Gj}, y_{Gj})$ is the $j^{\text{th}}$ generated sample.

According to the definitions of $d_j$ and $d_C$, $d_j \geq d_C$. A smaller $d_j$ means that $\mathbf{z}_{Gj}$ is closer to $\mathbf{z}_C$, indicating a more dissimilar distribution of $\mathbf{z}_{Gj}$ to the original samples. A sample in region A satisfies $d_j < \rho d_C$, where $\rho \geq 1$ is a parameter. Therefore, the qualified samples, based on the S1 criterion, are defined as:

$$d_j \geq \rho d_C, \rho \geq 1. \tag{5}$$

Subsequently, the S2 criterion is adopted to filter the unsuitable virtual samples in area B. The samples in region B are far away from the distribution of the original data and tend to be outliers. The samples can be screened according to the statistical characteristics of the original samples. Based on the probability density function $p(\mathbf{x})$ for each normal operating data point of the initial samples, the $100\beta\%$ confidence bound can be defined as the likelihood threshold $h$ that satisfies the following formula:

$$\int_{\mathbf{x}:p(\mathbf{x})>h} p(\mathbf{x})d\mathbf{x} = \beta, \tag{6}$$

where $p(\mathbf{x})$ is a multivariate Gaussian distribution and the above confidence bounds can be found in a previous paper [46]. In particular, when the generated sample $\mathbf{x}_{Gj}$ satisfies the following formula, it is considered as an outlier, as follows:

$$D^2 = (\mathbf{x}_{Gj} - \mu_{OX})^{\mathrm{T}} \mathrm{C}_{OX}^{-1} (\mathbf{x}_{Gj} - \mu_{OX}) > \chi_q^2(\beta), \tag{7}$$

where $\mathrm{C}_{OX}^{-1}$ is the covariance of the input data of the original samples and $\chi_q^2(\beta)$ is the $\beta$-fractile of the Chi-square distribution, with a degree of freedom, $q$.

In summary, according to the aforementioned two-stage data selection strategy, $k$-qualified samples are selected from the rough generated data and are denoted as $\left\{\mathbf{x}_{Sj}, y_{Sj}\right\}_{j=1,\ldots,k}$. This data selection strategy makes the selected virtual samples more homogeneous, in agreement with the original data distribution.

### 3.2. SWGAN-SVR Soft Sensor Model

In this case, SVR is adopted as the base soft-sensor model for nonlinear processes. SVR is a statistical learning method that uses the structural risk-minimization criterion

instead of the empirical risk-minimization criterion for model construction [15]. The target function of the SVR is as follows [15]:

$$\begin{aligned} &\min J(\mathbf{w}, b, \xi_i, \xi_i^*) = \tfrac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ &s.t. \begin{cases} y_i - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) - b \le \varepsilon + \xi_i \\ \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1, \ldots, n \end{cases} \end{aligned}, \tag{8}$$

where $b$ is the bias, $\mathbf{w}$ is the weight vector, $\xi_i$ and $\xi_i^*$ are slack variables, $\gamma$ is a regularization parameter that controls the penalty for samples exceeding the fitting error, $\phi$ is a nonlinear kernel function, $\varepsilon$ is an insensitivity coefficient, and $n$ is the number of samples for the SVR model.

The constrained optimization can be solved using the Lagrange function by introducing Lagrange multipliers. Subsequently, Equation (8) is converted into a dual problem, as'follows:

$$\begin{aligned} &\min_{\alpha,\alpha^*} \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(\mathbf{x}_i, \mathbf{x}_j) + \varepsilon \sum_{i=1}^{n} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) \\ &s.t. \begin{cases} \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le \gamma, i = 1, 2, \ldots, n \end{cases} \end{aligned}, \tag{9}$$

where $\alpha_i$ and $\alpha_i^*$ are the Lagrange multipliers and $K(\cdot,\ \cdot)$ represents a kernel function. In this study, the radial basis function (RBF) is adopted:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\psi \|\mathbf{x}_i - \mathbf{x}_j\|^2), \tag{10}$$

where $\psi > 0$ is a controlling parameter for the RBF kernel width.

Therefore, the SVR model can be described as

$$f(\mathbf{x}) = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) K(\mathbf{x}_i, \mathbf{x}) + b, \tag{11}$$

A flowchart of the SWGAN-SVR model is presented in Figure 3. It is difficult to develop a reliable SVR soft sensor for the initial limited training data, $\{\mathbf{x}_{Oi}, y_{Oi}\}_{i=1,\ldots,M}$. In such a situation, the WGAN-GP is adopted for data augmentation and $N$ virtual samples are generated, which is denoted as $\{\mathbf{x}_{Gj}, y_{Gj}\}_{j=1,\ldots,N}$. Furthermore, considering the unstable training process of the WGAN-GP, unsuitable virtual samples are generated, which need to be screened out from the group of rough virtual samples. Consequently, after employing the proposed two-stage data selection strategy; that is, the centroid metric criterion S1 and statistical characteristic criterion S2, $k$-qualified samples $\{\mathbf{x}_{Sj}, y_{Sj}\}_{j=1,\ldots,k}$ are obtained. By combining qualified virtual samples with the initial limited training samples, a new augmented training sample set is obtained, which can be denoted as $\left\{\mathbf{x}_{Oi} \cup \mathbf{x}_{Sj}, \mathrm{y}_{Oi} \cup \mathrm{y}_{Sj}\right\}_{i=1,\ldots,M, j=1,\ldots,k}$. Subsequently, an SVR soft sensor is constructed for quality prediction. Note that other supervised soft-sensor modeling methods, such as partial least squares regression and Gaussian process regression, can also replace SVR in this framework.

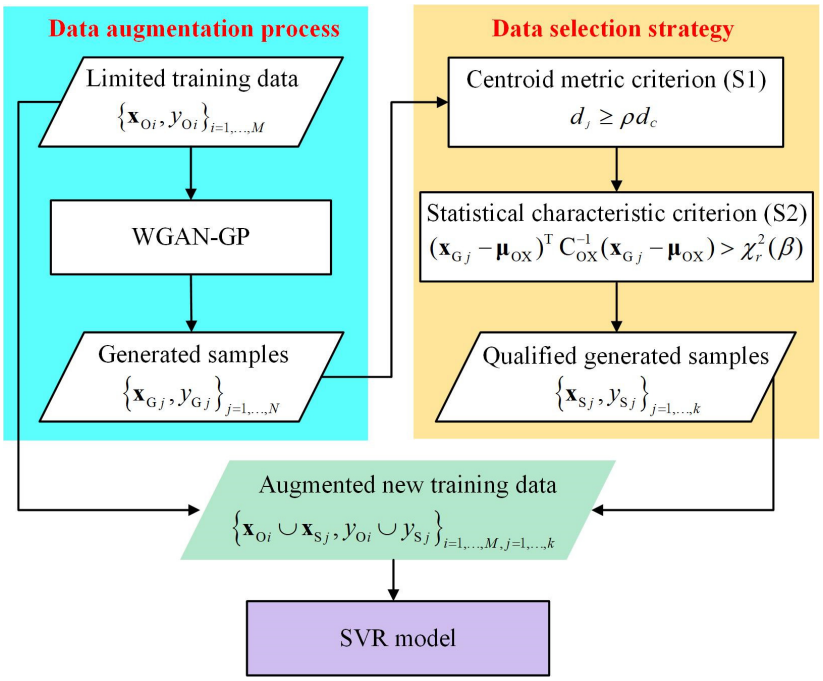



**Figure 3.** Flowchart of the proposed SWGAN-SVR method.

## 4. Results and Discussion

A numerical example and an industrial polyethylene process were adopted to validate the effectiveness of the proposed SWGAN-SVR modeling method. The commonly used root-mean-square error (RMSE), coefficient of determination ($R^2$), and mean absolute error (MAE) indices were used for the performance evaluation and are expressed as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{t=1}^{m}(y_t - \hat{y}_t)^2}, \tag{12}$$

$$\mathrm{R}^2 = 1 - \sum_{t=1}^{m}(y_t - \hat{y}_t)^2 \Big/ \sum_{t=1}^{m}(y_t - \overline{y}_t)^2, \tag{13}$$

$$\mathrm{MAE} = \frac{1}{m}\sum_{t=1}^{m}|y_t - \hat{y}_t|, \tag{14}$$

where $y_t$ and $\hat{y}_t$ are the quality measurement and prediction values of the $t^{\mathrm{th}}$ observation, respectively, and $m$ is the sample size.

### *4.1. Numerical Example*

A numerical example with a two-dimensional input and one-dimensional output was constructed to simulate the process of insufficient initial training samples:

$$\begin{aligned} x_1 &= 3u^2 + 4u,\ u = -10, -9.8, -9.6, \ldots, 10 \\ x_2 &= 8u + 2\cos(\pi u/3),\ u = -10, -9.8, -9.6, \ldots, 10 \\ y &= x_1 + x_2 + e \end{aligned}, \tag{15}$$

where $x_1$ and $x_2$ are two state variables that are constructed using the variable $u$, $y$ is an output variable, and $e$ is Gaussian noise with a zero mean and a variance of 0.01.

In this study, 100 samples were collected. To build the soft-sensor model, 50 samples were randomly selected as the training data and 50 samples were used for testing. In such a situation, using only limited training samples to train an SVR soft sensor may be insufficient. Therefore, it is essential to generate virtual samples to increase the data capacity and enrich the data diversity.

First, we investigated the number of generated samples that were sufficient for this example, using a 10-fold cross-validation algorithm. Specifically, a new training set containing both the original samples and generated virtual samples was divided into 10 non-overlapping subsets. Subsequently, based on the $i^{th}$ subset, which was regarded as a temporary test set, and extra subsets other than the $i^{th}$ subset, which was regarded as a temporary training set, an SVR model was constructed. Each subset was used as a temporary test set, in turn. Consequently, the total prediction result for a certain number of generated samples was obtained across 10 trials. The RMSE results for different numbers of generated samples are depicted in Figure 4. As the number of virtual samples increased, the RMSE value first decreased and then increased. This is mainly because the generated virtual samples filled the information gap in the initial training stage, which improved the model prediction accuracy. When the size of the virtual samples was sufficiently large, the influence of the initial samples was weakened, and more significant differences occurred between the initial and virtual samples. Therefore, as illustrated in Figure 4, the appropriate number of virtual samples for this example was 450.

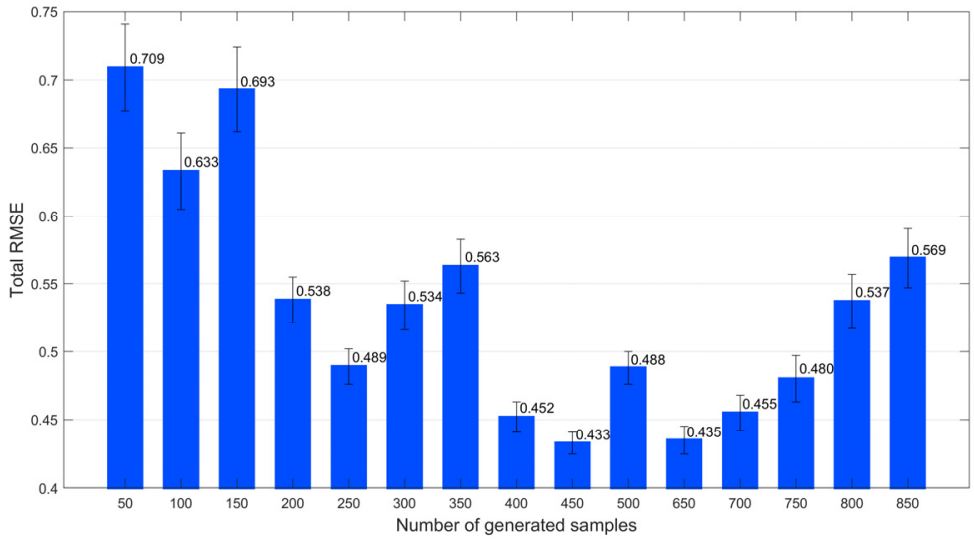
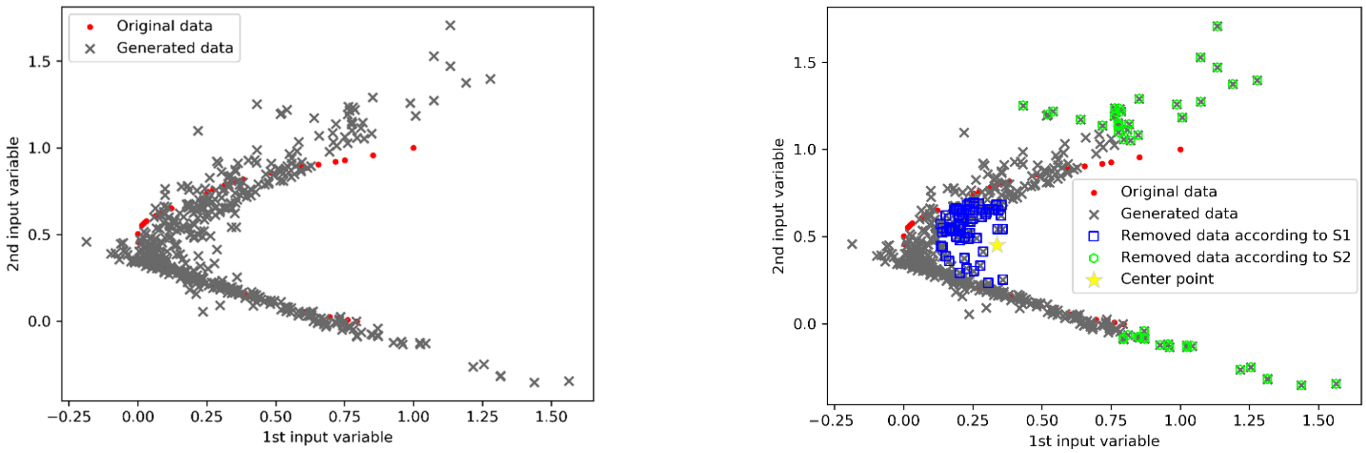



**Figure 4.** Total RMSE values of the different numbers of generated samples for the numerical example.

The scatter distributions of the original samples and the 450 generated samples are presented in Figure 5a. Several unsuitable samples did not conform to the initial data distribution. According to the proposed S1 and S2 data selection criteria, the scatter distribution of the qualified virtual samples, rough virtual samples, and initial limited samples are shown in Figure 5b. Unsuitable samples that were too close to the centroid and distant outliers were filtered. Consequently, the qualified virtual samples matched the distribution of the original samples. When combined with the original training data, the qualified virtual samples served as complements to the initial samples. The SWGAN-SVR model was built, based on the qualified augmented training samples; the prediction results for the test set are listed in Table 1. For comparison, the prediction results of SVR, WGAN-SVR, and WGAN-SVR using the S1 criterion (denoted as WGAN-SVR(S1)), and WGAN-SVR using the S2 criterion (denoted as WGAN-SVR(S2)), are also listed in Table 1. WGAN-SVR, WGAN-SVR(S1), WGAN-SVR(S2), and SWGAN-SVR outperformed the SVR method, with smaller RMSE and MAE values and larger $R^2$ values. This is mainly because the generated samples increased the diversity of the training samples. The prediction

performances of WGAN-SVR(S1) and WGAN-SVR(S2) were further enhanced, compared to the results of WGAN-SVR. By adopting only one data selection criterion, unsuitable virtual samples around the centroid or far-away outliers were screened out, which improved the quality of the augmented samples. This also demonstrates that unsuitable virtual samples result in the insufficient construction of reliable soft sensors. Furthermore, after simultaneously adopting the S1 and S2 criteria, SWGAN-SVR achieved the best prediction performance among the five methods. This indicates that a two-stage data selection strategy is beneficial for selecting qualified augmented samples and improving the performance of the base SVR soft sensor.



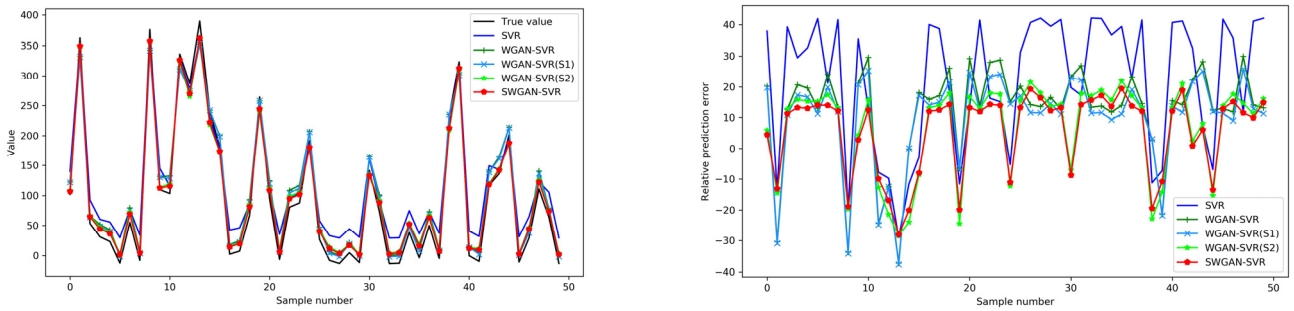

(**a**) (**b**)

**Figure 5.** Scatter distributions for the numerical example: (**a**) original data and rough generated data, and (**b**) qualified virtual samples.

**Table 1.** Performance comparison of SWGAN-SVR and other methods for the numerical example.

| | RMSE | $R^2$ | MAE |
|---|---|---|---|
| SVR | 30.399 | 0.933 | 27.248 |
| WGAN-SVR | 20.083 | 0.971 | 18.548 |
| WGAN-SVR(S1) | 18.222 | 0.976 | 16.653 |
| WGAN-SVR(S2) | 16.249 | 0.981 | 15.478 |
| SWGAN-SVR | **14.144** | **0.986** | **13.407** |

For a better illustration, the detailed prediction results and relative prediction errors of the five soft sensors on the test set are presented in Figure 6a,b, respectively. As shown in Figure 6a, SWGAN-SVR tracked the real trajectory better than the other four soft sensors, and the prediction curve of SWGAN-SVR was the one that was most consistent with the real curve. As illustrated in Figure 6b, the prediction errors of the proposed SWGAN-SVR were much smaller for the entire test set, and the errors were mostly around zero. A boxplot of the absolute prediction error values for the five methods is shown in Figure 7. SWGAN-SVR had a narrower error range, which was closer to zero, than the other four methods. Furthermore, as demonstrated through a comparison of the red lines in the boxes, the median value of the absolute error was smaller than that of the other four methods, indicating a better prediction performance for SWGAN-SVR.




(**a**) (**b**)

**Figure 6.** Prediction performance of five methods for the numerical example: (**a**) the assay and predicted values, and (**b**) the relative prediction errors.

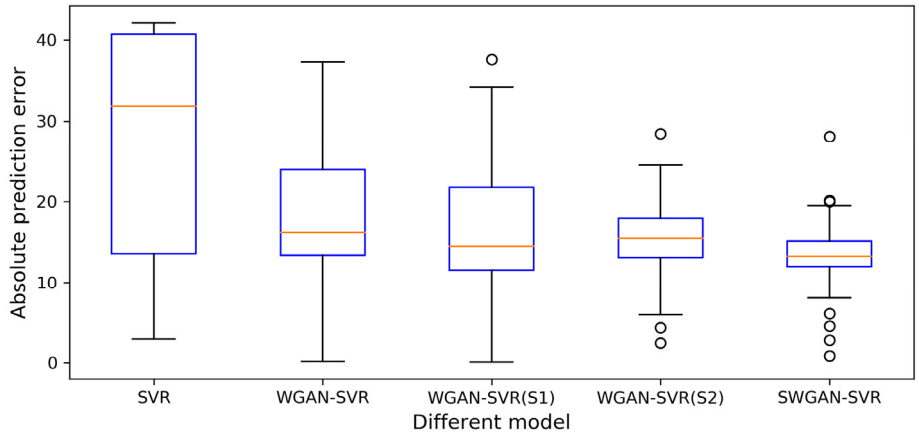



**Figure 7.** Absolute values of the prediction errors for the numerical example.

### *4.2. Industrial Polyethylene Process*

An industrial polyethylene process [4] was utilized to verify the necessity and superiority of the proposed method for practical applications. The product of the polyethylene manufacturing process was sampled once daily from the laboratory. Hence, in the initial stage of a new product grade, the collected quality variables (that is, the melt index (MI)) are insufficient for the development of a reliable soft sensor. After using a simple 3-sigma criterion to remove outliers, 60 samples were investigated. The dataset was partitioned into two parts: 30 randomly selected samples were used as the training data and the remaining 30 samples were used for testing.

Using a 10-fold cross-validation method, a suitable number of virtual samples was first determined for this example. The complete RMSE indices for different numbers of virtual samples are presented in Figure 8. The RMSE value was smallest when the number of generated samples was 150. Hence, 150 virtual samples were generated as an appropriate supplement to the initial limited samples. The proposed data selection strategy was adopted to improve the quality of the generated virtual samples. Subsequently, the proposed SWGAN-SVR model was built, based on the qualified augmented samples. Furthermore, SVR, WGAN-SVR, WGAN-SVR(S1), and WGAN-SVR(S2) were built to predict the MI value. The details of the prediction performance of the five methods on the test set are listed in Table 2. According to the prediction results, the SVR method achieved the largest RMSE value and smallest $R^2$ value, indicating the worst prediction accuracy among the five methods. This occurred because the initial training data were insufficient for the construction of reliable soft sensors. With this data augmentation strategy, the WGAN-SVR, WGAN-SVR(S1), WGAN-SVR(S2), and SWGAN-SVR methods can improve the prediction accuracy, compared to the SVR approach. The generated virtual samples fill the information gap in the initial data and increase the sample capacity. Moreover, by

adopting the two-stage data selection criteria, the SWGAN-SVR method achieved the best prediction performance among the five methods. The SWGAN-SVR method attempts to select the qualified virtual samples and, subsequently, to improve the quantity and quality of the initial training data. Note that owing to the strong nonlinearity of this example, the $R^2$ index was relatively smaller than that of the numerical example described in Section 4.1.

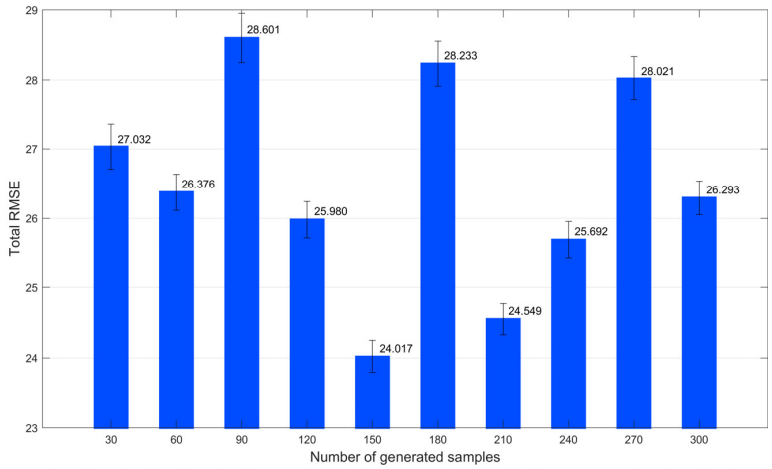



**Figure 8.** Total RMSE values of the different numbers of generated samples for the polyethylene process.

**Table 2.** Performance comparison of SWGAN-SVR and other methods for industrial polyethylene.

| | RMSE | $R^2$ | MAE |
|---|---|---|---|
| SVR | 50.925 | 0.015 | 33.061 |
| WGAN-SVR | 36.227 | 0.502 | 22.550 |
| WGAN-SVR(S1) | 32.923 | 0.588 | 22.835 |
| WGAN-SVR(S2) | 34.597 | 0.546 | 21.828 |
| SWGAN-SVR | **28.854** | **0.684** | **19.379** |

The scatter distribution of the rough generated samples and selected unsuitable samples are presented in Figure 9. Virtual samples close to the centroid and the distant outliers were filtered. The remaining qualified samples matched well with the distribution of the original samples. Moreover, the diversity of the original samples increased with the incorporation of the qualified samples.

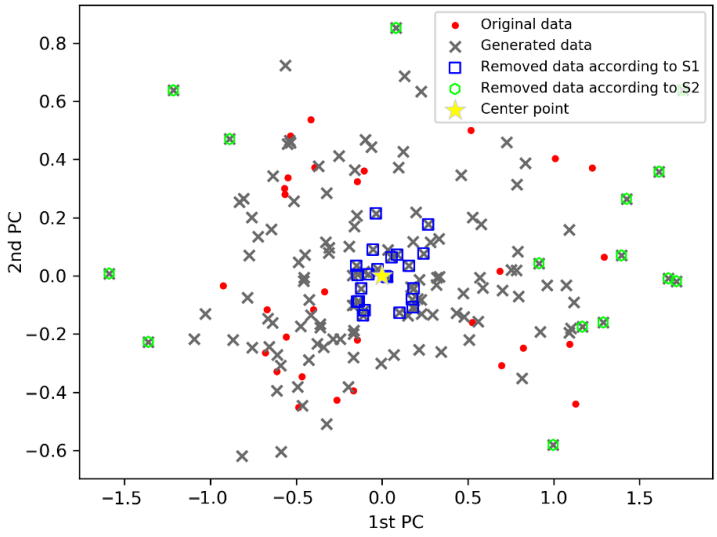



**Figure 9.** Scatter distribution comparison of the qualified virtual samples for the polyethylene process.

The detailed prediction results of the five soft sensors on the test set are depicted in Figure 10. The proposed SWGAN-SVR method was superior to the other four methods in terms of tracking the real trend of the output variable. The prediction of SWAGN-SVR was in good agreement with the actual trajectory of the MI value, and, thus, exhibited a much smaller deviation. The relative prediction errors of the five methods are shown in Figure 11. The SWGAN-SVR method achieved the best prediction performance and yielded the smallest prediction error at most sampling points. Consequently, the obtained results indicate that the proposed SWGAN-SVR soft sensor can enhance prediction performance when dealing with insufficient training samples.

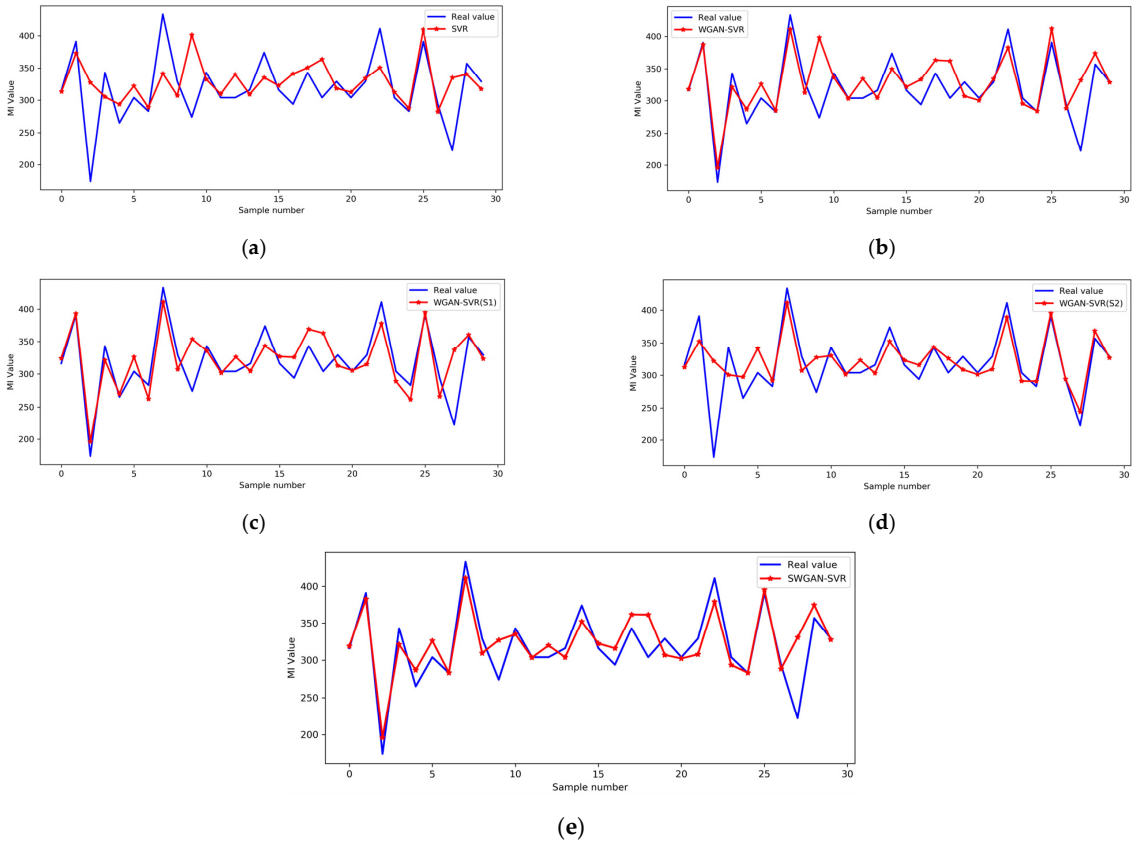



**Figure 10.** Assay and predicted values for the polyethylene process: (**a**) SVR, (**b**) WGAN-SVR, (**c**) WGAN-SVR(S1), (**d**) WGAN-SVR(S2), (**e**) SWGAN-SVR.

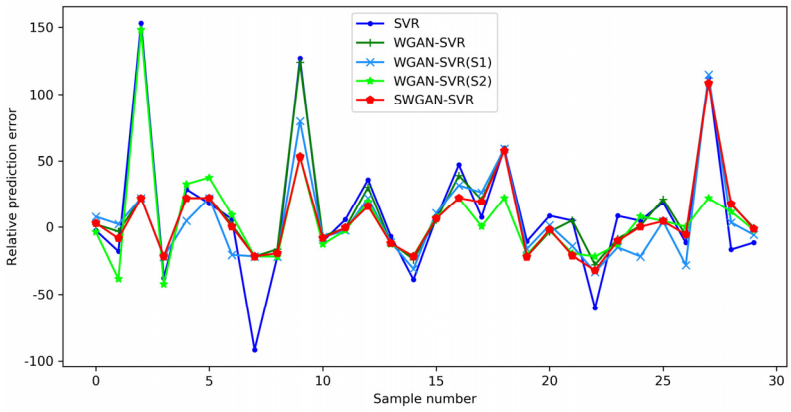



**Figure 11.** Relative prediction errors for the polyethylene process.

## 5. Conclusions

In this study, a reliable soft sensor framework is developed to enhance prediction performance by introducing augmented data. Because having limited training data will be insufficient for establishing a reliable soft sensor, rough virtual samples are generated using the WGAN-GP method to enrich the sample information. Subsequently, based on a two-stage data selection strategy, qualified augmented samples are gradually selected to eliminate the negative effects of unsuitable samples on the prediction performance. Based on the qualified augmented training samples, the SWGAN-SVR method is designed to capture the process characteristics, which is beneficial for regression. The prediction results for the two examples demonstrate the advantages of the proposed approach. Further investigations will aim to enhance the quality of the generated samples, using GANs. Additionally, the combination of the process characteristics to generate more informative samples for practical applications is an interesting topic.

**Author Contributions:** Data curation, Y.D. and Y.L.; funding acquisition, Y.L. and Y.Y.; investigation, Y.D., A.L., M.C. and Y.L.; methodology, Y.D., A.L., Y.L. and Y.Y.; project administration, Y.L.; writing—original draft, Y.D., A.L., M.C. and Y.L.; writing—review and editing, Y.L. and Y.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the National Natural Science Foundation of China (Grant Nos. 62022073 and 61873241) and the National Key Research Program of China (Grant No. 2019YFB1705904). Yuan Yao was supported in part by the Ministry of Science and Technology, ROC under Grant No. MOST 111-2221-E-007-005.

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Fortuna, L.; Graziani, S.; Rizzo, A.; Xibilia, M. *Soft Sensors for Monitoring and Control of Industrial Processes*; Springer: New York, NY, USA, 2007.
2. Kadlec, P.; Grbic, R.; Gabrys, B. Review of adaptation mechanisms for data-driven soft sensors. *Comput. Chem. Eng.* **2011**, *35*, 1–24. [CrossRef]
3. Sun, Q.; Ge, Z. A survey on deep learning for data-driven soft sensors. *IEEE Trans. Ind. Inform.* **2021**, *17*, 5853–5866. [CrossRef]
4. Liu, Y.; Yang, C.; Zhang, M.; Dai, Y.; Yao, Y. Development of adversarial transfer learning soft sensor for multigrade processes. *Ind. Eng. Chem. Res.* **2020**, *59*, 16330–16345. [CrossRef]
5. Liu, Y.; Xie, M. Rebooting data-driven soft-sensors in process industries: A review of kernel method. *J. Process Control* **2020**, *89*, 58–73. [CrossRef]
6. Fuentes-Cortes, L.; Flores-Tlacuahuac, A.; Nigam, K. Machine learning algorithms used in PSE environments: A didactic approach and critical perspective. *Ind. Eng. Chem. Res.* **2022**, *61*, 8932–8962. [CrossRef]
7. Jiang, Y.; Yin, S.; Dong, J.; Kaynak, O. A review on soft sensors for monitoring, control, and optimization of industrial processes. *IEEE Sens. J.* **2021**, *21*, 12868–12881. [CrossRef]
8. Verma, A.; Kumar, R.; Parashar, A. Enhanced thermal transport across a bi-crystalline graphene–polymer interface: An atomistic approach. *Phys. Chem. Chem. Phys.* **2019**, *21*, 6229–6237. [CrossRef]
9. Verma, A.; Parashar, A.; Packirisamy, M. Effect of grain boundaries on the interfacial behaviour of graphene-polyethylene nanocomposite. *Appl. Surf. Sci.* **2019**, *470*, 1085–1092. [CrossRef]
10. Sharmin, R.; Sundararaj, U.; Shah, S.; Griend, L.V.; Sun, Y.-J. Inferential sensors for estimation of polymer quality parameters: Industrial application of a PLS-based soft sensor for a LDPE plant. *Chem. Eng. Sci.* **2006**, *61*, 6372–6384. [CrossRef]
11. Qin, S.; Dong, Y.; Zhu, Q.; Wang, J.; Liu, Q. Bridging systems theory and data science: A unifying review of dynamic latent variable analytics and process monitoring. *Annu. Rev. Control* **2020**, *50*, 29–48. [CrossRef]
12. Rasmussen, C.; Williams, C. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
13. Liu, Y.; Chen, T.; Chen, J. Auto-switch Gaussian process regression-based probabilistic soft sensors for industrial multigrade processes with transitions. *Ind. Eng. Chem. Res.* **2015**, *54*, 5037–5047. [CrossRef]
14. Kaneko, H.; Funatsu, K. Development of soft sensor models based on time difference of process variables with accounting for nonlinear relationship. *Ind. Eng. Chem. Res.* **2011**, *50*, 10643–10651. [CrossRef]

15. Zhang, M.; Liu, X. A real-time model based on optimized least squares support vector machine for industrial polypropylene melt index prediction. *J. Chemometr.* **2016**, *30*, 324–331. [CrossRef]
16. Chitralekha, S.; Shah, S. Application of support vector regression for developing soft sensors for nonlinear processes. *Can. J. Chem. Eng.* **2010**, *88*, 696–709. [CrossRef]
17. Liu, K.; Ma, Z.; Liu, Y.; Yang, J.; Yao, Y. Enhanced defect detection in carbon fiber reinforced polymer composites via generative kernel principal component thermography. *Polymers* **2021**, *13*, 825. [CrossRef]
18. Zhang, J.; Zhao, C. Condition-driven probabilistic adversarial autoencoder with nonlinear Gaussian feature learning for nonstationary process monitoring. *J. Process Control* **2022**, *117*, 140–156. [CrossRef]
19. Liu, Q.; Jia, M.; Gao, Z.; Xu, L.; Liu, Y. Correntropy long short term memory soft sensor for quality prediction in industrial polyethylene process. *Chemometrics Intell. Lab. Syst.* **2022**, *231*, 104678. [CrossRef]
20. Wang, C.; Peng, X.; Shang, C.; Fan, C.; Zhao, L.; Zhong, W. A deep learning-based robust optimization approach for refinery planning under uncertainty. *Comput. Chem. Eng.* **2021**, *155*, 107495. [CrossRef]
21. Roman, A.; Qin, S.; Rodriguez, J.; Gonzalez, L.; Zavala, V.; Osswald, T. Natural rubber blend optimization via data-driven modeling: The implementation for reverse engineering. *Polymers* **2022**, *14*, 2262. [CrossRef]
22. Yuan, X.; Ou, C.; Wang, Y. Development of NVW-SAEs with nonlinear correlation metrics for quality-relevant feature learning in process data modeling. *Meas. Sci. Technol.* **2020**, *32*, 015006. [CrossRef]
23. Wang, X.; Liu, H. Data supplement for a soft sensor using a new generative model based on a variational autoencoder and Wasserstein GAN. *J. Process Control* **2019**, *85*, 91–99. [CrossRef]
24. Wu, H.; Lo, Y.; Zhou, L.; Yao, Y. Process modeling by integrating quantitative and qualitative information using a deep embedding network and its application to an extrusion process. *J. Process Control* **2022**, *115*, 48–57. [CrossRef]
25. Jin, H.; Li, Z.; Chen, X.; Qian, B.; Yang, B.; Yang, J. Evolutionary optimization based pseudo labeling for semi-supervised soft sensor development of industrial processes. *Chem. Eng. Sci.* **2021**, *237*, 116560. [CrossRef]
26. Zhang, X.; Xu, Y.; He, Y.; Zhu, Q. Novel manifold learning based virtual sample generation for optimizing soft sensor with small data. *ISA Trans.* **2021**, *109*, 229–241. [CrossRef] [PubMed]
27. Poggio, T.; Vetter, T. Recognition and structure from one 2D model view: Observations on prototypes object classes and symmetries. *Mass. Inst. Technol.* **1992**, *1347*, 1–25.
28. Cho, S.; Jang, M.; Chang, S. Virtual sample generation using a population of networks. *Neural Process. Lett.* **1997**, *5*, 21–27. [CrossRef]
29. He, Y.; Hua, Q.; Zhu, Q.; Lu, S. Enhanced virtual sample generation based on manifold features: Applications to developing soft sensor using small data. *ISA Trans.* **2022**, *126*, 398–406. [CrossRef]
30. Wei, S.; Chen, Z.; Arumugasamy, S.; Chew, I. Data augmentation and machine learning techniques for control strategy development in bio-polymerization process. *Env. Sci. Ecotechnol.* **2022**, *11*, 100172. [CrossRef]
31. Chao, G.; Tsai, T.; Lu, T.; Hsu, H.; Bao, B.; Wu, W.; Lin, M.; Lu, T. A new approach to prediction of radiotherapy of bladder cancer cells in small dataset analysis. *Expert Syst. Appl.* **2011**, *38*, 7963–7969. [CrossRef]
32. Li, D.; Wu, C.; Tsai, T.; Lina, Y. Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. *Comput. Oper. Res.* **2007**, *34*, 966–982. [CrossRef]
33. Li, D.; Chen, C.; Chang, C.; Lin, W. A tree-based-trend-diffusion prediction procedure for small sample sets in the early stages of manufacturing systems. *Expert Syst. Appl.* **2012**, *39*, 1575–1581. [CrossRef]
34. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
35. Lyu, Y.; Chen, J.; Song, Z. Synthesizing labeled data to enhance soft sensor performance in data-scarce regions. *Control Eng. Practice* **2021**, *115*, 104903. [CrossRef]
36. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. In Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014.
37. Gao, S.; Dai, Y.; Li, Y.; Jiang, Y.; Liu, Y. Augmented flame image soft sensor for combustion oxygen content prediction. *Meas. Sci. Technol.* **2023**, *34*, 015401. [CrossRef]
38. Gao, S.; Dai, Y.; Li, Y.; Liu, K.; Chen, K.; Liu, Y. Multiview Wasserstein generative adversarial network for imbalanced pearl classification. *Meas. Sci. Technol.* **2022**, *33*, 085406. [CrossRef]
39. Liu, K.; Zheng, M.; Liu, Y.; Yang, J.; Yao, Y. Deep autoencoder thermography for defect detection of carbon fiber composites. *IEEE Trans. Ind. Inform.* **2022**. [CrossRef]
40. Xu, X.; Lin, K.; Yang, Y.; Hanjalic, A.; Sheng, H. Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3030–3047. [CrossRef]
41. Zhang, T.; Chen, J.; Li, F.; Pan, T.; He, S. A small sample focused intelligent fault diagnosis scheme of machines via multimodules learning with gradient penalized generative adversarial networks. *IEEE Trans. Ind. Electron.* **2021**, *68*, 10130–10141. [CrossRef]
42. Jiang, X.; Ge, Z. Data augmentation classifier for imbalanced fault classification. *IEEE Trans. Autom. Sci. Eng.* **2021**, *18*, 1206–1217. [CrossRef]
43. Li, Z.; Xia, P.; Tao, R.; Niu, H.; Li, B. A new perspective on stabilizing gans training: Direct adversarial training. *IEEE Trans. Emerg. Top. Comput. Intell.* **2022**, 1–12. [CrossRef]

44. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein Generative Adversarial Networks. In Proceedings of the Machine Learning Research, Sydney, Australia, 6–11 August 2017.
45. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. A Improved training of Wasserstein GANs. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
46. Berger, J. *Statistical Decision Theory and Bayesian Analysis*; Springer-Verlag: New York, NY, USA, 1985.