Novel SVD integrated with GBDT based Virtual Sample Generation and Its Application in Soft Sensor

Qun-Xiong Zhu[1,2], **Xiao-Lu Song**[1,2], **Ning Zhang**[1,2], **Ye Tian**[1,2], **Yuan Xu**[1,2], **Yan-Lin He**[1,2,*]

[1]College of Information Science and Technology, Beijing University of Chemical Technology (BUCT), Beijing, 100029, China

[2]Engineering Research Center of Intelligent PSE, Ministry of Education of China, Beijing, 100029, China

(e-mail: heyl@mail.buct.edu.cn)

Abstract: With the coming of the big data era, data-driven based modeling approaches have become the hot research topic in recent years. Unfortunately, due to the limitation of the actual process, the data is basically in a steady state and it is difficult to obtain enough high-quality data, which is defined as the small sample (SS) problem. Recently, to deal with the SS problem, a virtual sample generation (VSG) approach based on the distribution of the original data has been taken into account. In this paper, a VSG method based on singular value decomposition (SVD) feature decomposition and gradient boosting decision tree (GBDT) prediction model (SVD-GBDT) is proposed. In the proposed SVD-GBDT method, firstly, the distribution characteristics of the original data are used to extract the main features and expand the number of samples by using the SVD algorithm. Then the GBDT algorithm is used to find the virtual output of the virtual samples by the SVD method. Finally, SVD and GBDT are combined to complete the sample expansion (SVD-GBDT-VSG). In this paper, we choose the purified terephthalic acid (PTA) industrial process to verify the effectiveness of the proposed methodology. Simulation results show that compared with related methods, the proposed SVD-GBDT-VSG algorithm in this paper can achieve sample expansion well and at the same time can effectively improve the accuracy performance of soft measurement.

Keywords: Soft Sensor, Singular Value Decomposition, Gradient Boosting Decision Tree, Virtual Sample Generation

1. INTRODUCTION

In recent years, as the process industry is moving toward digitalization and intelligence, it has become a fashionable research topic to mine valid information from process data and build data-driven models(He et al., 2020). In the data-driven modeling popularity, the number and distribution of samples directly affect the prediction performance of data-driven models. Only when the samples are sufficient and the distribution of samples is reasonable, the accuracy of the model can be guaranteed. However, the complexity of the actual chemical process scale and the increased difficulty of collecting data on core variables make data-driven modeling lack sufficient data information. Therefore, a series of methods have been proposed to improve the accuracy of data modeling with SS.

To address the problem of SS size, scholars use gray predictive modeling, Gaussian mixture modeling method to process the sample. Besides, scholars have proposed machine learning models (MLM) based on statistical theory to solve the SS problem, and typical methods include Bayesian Network (BN) (Sánchez-Franco et al., 2019) and Support Vector Machine (SVM) (Arora et al., 2019). Furthermore, the VSG method is another useful way to solve the SS problem. The main idea of VSG is to use the prior knowledge of the sample and the known distribution to generate new samples based on the original SS. The main advantage of this method is that it can fill the data gaps and rationalize the distribution of the data. The generated virtual samples are combined with the original SS to get the purpose of sample expansion and improve the accuracy and generalization of the model.

In recent years, the virtual sample generation technique has become a popular method to solve small data problems, and a series of VSG methods have been generated. Among them, the most representative methods are: Mega-Trend-Diffusion (MTD) (Kang et al., 2019), Tree Trend Diffusion (TTD) (Li et al., 2012) Monte Carlo (Arndt, 2009), and some midpoint interpolation (MI) (Campo, 2020) Kriging interpolation method (KIM) (Zhu et al., 2020), SMOTE (Maldonado et al., 2019) and bootstrap (Da Silva et al., 2015). The advantages of Bootstrap and SMOTE of sampling-based methods are computational simplicity and low cost. Bootstrap is mainly based on the principle of repeated sampling of raw sample, and the sampled sample is used as a virtual sample to complete the sample data expansion. SMOTE is to generate new samples by local nearest neighbors. Both methods rely on the overall distribution of samples, however it is difficult to obtain the real distribution of the data for high-dimensional data cannot fundamentally solve the SS problem. Unlike the above methods, MTD and TTD based on information diffusion technology generate virtual samples by expanding the sample

This research is supported by the National Natural Science Foundation of China under Grant Nos.61973024,61973022 and 62073022.

attribute domain based on diffusion function and using fuzzy theory. These methods get rid of the distribution problem of samples, but the selection of diffusion function and diffusion coefficient has encountered difficulties.

Based on the above problems, a VSG method based on singular value decomposition (SVD) (Li et al., 2019) feature decomposition and gradient augmented decision tree (GBDT) (Zhu et al., 2021) called SVD-GBDT-VSG is proposed in this paper. In SVD-GBDT-VSG, the data is first processed by SVD to obtain feature values and feature vectors, then the most representative features are selected to expand the data, and finally GBDT is used to find the output of the expanded virtual samples. Meanwhile, the purified terephthalic acid (PTA) chemical data is chosen in this paper to verify the effectiveness of the proposed SVD-GBDT-VSG method. The simulation results show that SVD-GBDT-VSG can significantly enhance the accuracy of prediction model generalization capacity compared with other related methods. It is proved that SVD-GBDT-VSG can effectively solve the small data problem in our case.

The remainder of this paper is listed as follows: in section 2, the SVD and GBDT methods are briefly introduced. Section 3 details the modeling and simulation process of SVD-GBDT-VSG algorithm. Section 4 verifies the proposed method through PTA simulation experiments and analyzes the simulation results, and conclusions are given in Section 5.

2. RELATED METHODS

In this section, we briefly introduce the related methods used for VSG and modeling, including singular value decomposition (SVD) and gradient augmented decision tree (GBDT).

2.1 SVD algorithm

The commonly used eigenvalue decomposition (EVD) has high requirements for matrices, and the decomposed matrix must be a real symmetric square matrix. In contrast, SVD is a matrix decomposition method applicable to arbitrary matrices. This method can extract the eigenvalues and eigenvectors of the data. Based on the extracted eigenvectors, we can get representative eigenvectors and use them to expand the data.

The specific process of SVD is as follows. Suppose a matrix $\boldsymbol{A} \in \mathbb{R}^{a \times b}$, the SVD of $\boldsymbol{A}$ is defined as

$$\boldsymbol{A} = \boldsymbol{U\Sigma V}^T \tag{1}$$

where $\boldsymbol{U}$ is $a \times a$ matrix named left singular matrix, $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}_i)\ \ (i = 1, 2, \cdots, p)$, $\boldsymbol{\sigma}_i$ is the singular value, $\boldsymbol{V}$ is $b \times b$ matrix. By doing the eigenvalue decomposition of $\boldsymbol{AA}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$, we can obtain $\boldsymbol{U}$ and $\boldsymbol{V}$, and thus the corresponding eigenvectors $\boldsymbol{\sigma}$. Based on the obtained eigenvectors, the high-quality part is selected to expand the original sample size.

2.2 GBDT algorithm

GBDT is a novel regression prediction model, the main idea is to train a decision tree model using the learning strategy of Gradient Boosting. GBDT is a multi-round iterative model, in each round iteration produces a weak classifier, where the weak classifier is generally chosen as Classification and Regression Tree (CART). GBDT uses an additive model, which means that a new CART tree is created in the direction of the gradient of residual reduction in each iteration of GBDT, and after several iterations, the residual converges to 0 to obtain the optimal fitting performance by the training data. Finally, the results of all decision trees are accumulated to obtain the final prediction results.

3. THE PROPOSED SVD-GBDT METHODS

To improve the model accuracy, we need to perform sample expansion for SS. In this section, we propose a novel SVD-GBDT virtual sample generation method (SVD-GBDT-VSG) based on SVD-GBDT using a specific small sample dataset to expand the data based on the original SS. The SVD-GBDT-VSG method starts from the sample distribution and can generate new samples that match the original sample distribution. The ability of the model to generalize to the original sample test set is effectively enhanced by adding the newly generated samples to the original sample set. The SVD-GBDT-VSG method consists of the following four steps:

Step 1 Training the GBDT model based on original sample set. The choice of a good regression model plays a crucial role in the prediction of the sample. The basic learner of GBDT is a decision tree, which is constructed at each step of the iteration to compensate for the shortcomings of the existing model by reducing the loss in the direction of the steepest gradient. This idea gives it the natural advantage of being able to discover multiple distinguishing features and combinations of features, and this advantage gives GBDT a good generalization capability. Suppose the small samples data is $\boldsymbol{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $\boldsymbol{y}_i \in \mathbb{R}^1$, $i = [1, N]$. We divide the dataset $\boldsymbol{D}$ into a training set $\boldsymbol{D}_{train}$ and a test set $\boldsymbol{D}_{test}$, determine the hyperparameters, and then train the GBDT model. The specific procedure of the algorithm is shown in the following table.

Table 1*. The GBDT algorithm training process.

```
Algorithm: GBDT
Input:
(1) Initializing CART
        $f_0(\boldsymbol{x}) = arg\min_{c} \sum_{i=1}^{m} L(\boldsymbol{y}_i, c)$
    where $c$ is the mean of $y_i$, $L(y, f(x)) = (y - f(x))^2$ is the loss function of the algorithm.
(2) for t = 1,2,······,T
        for i = 1,2, ······,m
            $r_{ti} \approx -\left[\frac{\partial L(\boldsymbol{y}_i, f(\boldsymbol{x}_i))}{\partial f(\boldsymbol{x}_i)}\right]_{f(\boldsymbol{x}) = f_{t-1}(\boldsymbol{x})}$
        end
        Fitting a CART tree using $(\boldsymbol{x}_i, r_{ti})$, The t-th regression tree is obtained, and the corresponding leaf node region is $R_{ij}$, J is the number of leaf nodes of the t-th regression tree.
```

```
for j = 1,2, ······,J, calculate the best-fit value $c_{tj}$ .
    $c_{tj} = arg\min_{c} \sum_{x_{i\in R_{tj}}} L(y_i, f_{t-1}(x_i) + c)$
end
$f_t(x) = f_{t-1}(\boldsymbol{x}) = \sum_{j=1}^{J} c_{tj}, I(x \in R_{tj})$
(3) The final regression model was obtained as
$f(\boldsymbol{x}) = f_0(\boldsymbol{x}) + \sum_{t=1}^{T} \sum_{j=1}^{J} c_{tj}, I\,(x \in R_{tj})$
```

Step 2 Using SVD to generate inputs of virtual samples.
The SVD can decompose any matrix, and the singular values in the matrix of singular values after decomposition are arranged from largest to smallest, and the reduction rate is particularly fast. In most cases, the sum of the first 10% of the singular values accounts for more than 99% of the sum of all the singular values, so the first 10% of the singular values of the SVD are chosen to construct a new matrix to generate new samples.

Suppose the training dataset has *a* samples, each of which is *b* dimension. We can represent this data set as a matrix $\boldsymbol{A} \in \mathbb{R}^{a\times b}$. After SVD decomposition, the $\boldsymbol{A}$ is represented by three matrices $\boldsymbol{U} \in \mathbb{R}^{a\times a}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{a\times b}$ and $\boldsymbol{V}^T \in \mathbb{R}^{b\times b}$. Using the properties of SVD, $\boldsymbol{A}$ is approximated by the singular values of the top *m (m<b, m<a)* largest and the corresponding left singular vectors $\boldsymbol{U}' \in \mathbb{R}^{a\times m}$ and right singular vectors $\boldsymbol{V}^{T\prime} \in \mathbb{R}^{m\times b}$:

$$\boldsymbol{A}' = U'_{a\times m} \begin{bmatrix} \partial_1 & 0 & 0 & 0 \\ 0 & \partial_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \partial_m \end{bmatrix}_{m\times m} V'^{T}_{m\times b} \qquad (2)$$

The $\boldsymbol{A}'$ obtained is the new distribution that approximates the $\boldsymbol{A}$ matrix, i.e., the features of the new virtual samples are generated. In other words, we get the inputs of the virtual samples.

It is worth noting that this obtains input x_{vi} of *a* new samples . That is, double the original samples. But in practical applications we often need more samples. Our solution is that for ∂_i, if it satisfies $\partial_1 + \partial_2 + \cdots + \partial_i \geqslant k$ (k is the sum of the singular values, which is generally required to be greater than or equal to 80% of the sum of all singular values), then we can take the first i ∂'s and compute $\boldsymbol{A}'$ once .Thus if there are *n* ∂'s that satisfy the requirement, we expand *n* times $\boldsymbol{A}$.

Step 3 Using GBDT to obtain the output virtual samples
The input $\boldsymbol{x}_{vir}$ of the virtual samples obtained by SVD is used as the input of the trained GBDT model. The output $\boldsymbol{y}_{vir}$ of the corresponding input is predicted by the GBDT model. All the virtual samples $\boldsymbol{D}_{vi} = \{(\boldsymbol{x}_{vi}, \boldsymbol{y}_{vi})\}$ generated by SVD-GBDT are obtained, where $\boldsymbol{x}_{vi} \in \mathbb{R}^d, \boldsymbol{y}_{vi} \in \mathbb{R}^1$ and $i = [1, N]$.

Step 4 Retraining the GBDT model
We integrate the original small sample $\boldsymbol{D}$ with the generated virtual sample $\boldsymbol{D}_{vir}$ to form a new sample set. Then, we use the new training to reconstruct the GBDT model. The original small sample and the new sample are used to test the newly trained GBDT model. The SVD-GBDT-VSG is summarized in the following flow chart.

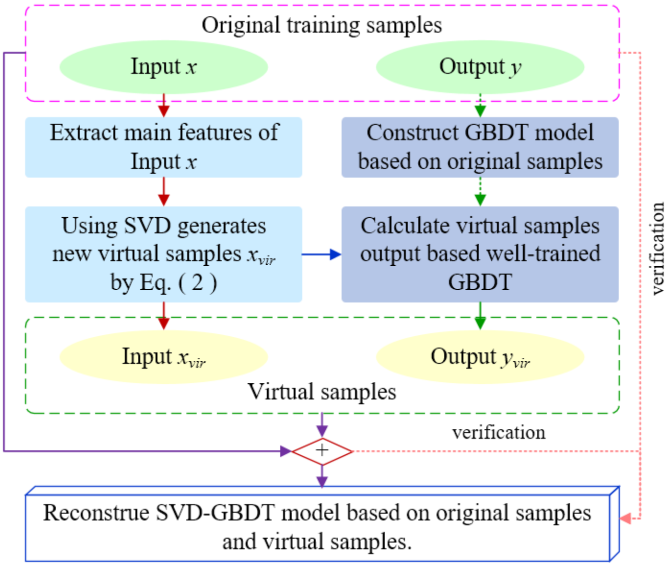
Original training samples
Input x
Output y
Extract main features of Input x
Construct GBDT model based on original samples
Using SVD generates new virtual samples x_{vir} by Eq. (2)
Calculate virtual samples output based well-trained GBDT
Input x_{vir}
Output y_{vir}
Virtual samples
verification
+
verification
Reconstrue SVD-GBDT model based on original samples and virtual samples.

Figure 1. Virtual samples generation flow chart.

4. CASE STUDY

This section, the proposed SVD-GBDT-VSG is simulated in a real industrial dataset to verify its performance. To ensure the reproducibility of the experiments, this experiment was conducted in Matlab 2020b and Python 3.7. During the simulation, the PTA industrial dataset from the petrochemical industry is selected as the experimental data. To validate the experimental effect, mean square error (MSE), mean absolute error (MAE) and coefficient of determination (R^2) are selected as the metrics of the prediction model in the simulation, and the specific equations are as follows:

$$MSE \doteq \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \qquad (3)$$

$$MAE \doteq \frac{1}{m} \sum_{i=1}^{m} |(y_i - \hat{y}_i)| \qquad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{m} (y_i - \overline{y}_i)^2} \qquad (5)$$

The chemical industry PTA is utilized in the production of polyester fibers. The PTA process data includes seventeen variables that affect the output, one output is the solvent dehydration tower conductivity and the measure of system efficiency is acetic acid. The industrial flow diagram for the PTA process is shown in figure 2.

There are 39 PTA samples (30 training data, 9 testing data) are selected for the simulation experiments. In the simulation experiments, GBDT prediction model is built from the original PTA training data. Then, SVD is performed on the training data as well, the top 90% of singular values are selected and the data is reconstructed to generate the input of the samples. The generated virtual samples are put into the trained GBDT model to generate the output of the virtual samples.

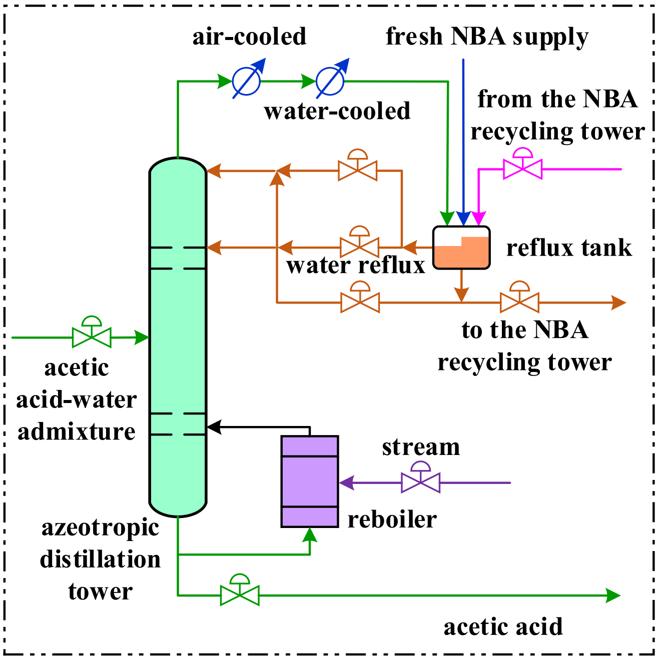
air-cooled
fresh NBA supply
water-cooled
from the NBA recycling tower
reflux tank
water reflux
to the NBA recycling tower
acetic acid-water admixture
stream
reboiler
azeotropic distillation tower
acetic acid

Figure 2. PTA flow chart.

It is worth noting that the error results in Table 1, 2, 3 and 4 are based on the test data of the original dataset without including the test data of the virtual samples, so the validity of our method can be well verified. Table 1 shows the error of the prediction model with different virtual samples numbers added using SVD-GBDT-VSG method. In addition, bootstrap, MTD and TTD are chosen to compare with SVD-GBDT-VSG to verify the performance of VSG. As can be seen from Table 2 3 and Table 4, the virtual samples generated by the proposed SVD-GBDT-VSG make the model have more error reduction and higher prediction accuracy compared to the related methods. However, it is not the case that more samples are better. As small data samples have certain features, too many virtual samples may hide their main features and result in a decrease in the accuracy of the model.

Table 1. The errors of adding different virtual samples.

Size of training sample	MAE (%)	MSE (%)	R^2
30	36.08	19.05	0.418
30+30	30.78	14.87	0.546
30+60	**24.32**	**10.29**	**0.686**
30+90	26.58	11.06	0.662
30+120	26.40	11.06	0.662
30+150	29.88	13.72	0.580
30+180	31.31	14.88	0.545

The output values of several VSG methods with and without virtual samples are depicted in figure 3. As can be seen from figure 3, the SVD-GBDT-VSG curve is closer to the real value than the other methods, which indicates that SVD-GBDT-VSG can generate high quality virtual samples and enhance the model performance.

In summary, the SVD-GBDT-VSG method has great advantages in improving the modeling accuracy. The simulation results show that SVD-GBDT-VSG is suitable for the generation of virtual samples for small data sets.

Table 2. The MSE (%) of different methods

Size of virtual sample	Methods			
	SVD-GBDT	Bootstrap	MTD	TTD
0	19.08	19.08	19.08	19.08
30	**14.87**	19.53	14.96	21.28
60	**10.29**	15.33	16.45	14.53
90	**11.06**	13.84	23.00	18.17
120	**11.06**	18.03	17.34	18.90
150	**13.72**	14.17	13.25	15.30
180	**14.88**	17.91	25.68	17.16

Table 3. The MAE (%) of different methods

Size of virtual sample	Method			
	SVD-GBDT	Bootstrap	MTD	TTD
0	36.08	36.08	36.08	36.08
30	**30.78**	31.25	33.02	38.74
60	**24.32**	25.45	34.55	32.26
90	**26.58**	28.34	43.15	37.62
120	**26.40**	30.03	34.05	36.60
150	**29.88**	27.87	31.82	36.03
180	**31.31**	33.06	36.09	34.47

Table 4. The R^2 of different methods

Size of virtual sample	Method			
	SVD-GBDT	Bootstrap	MTD	TTD
0	0.418	0.418	0.418	0.418
30	**0.546**	0.403	0.543	0.349
60	**0.686**	0.531	0.497	0.556
90	**0.662**	0.577	0.297	0.444
120	**0.662**	0.449	0.470	0.422
150	**0.580**	0.567	0.595	0.532
180	**0.545**	0.453	0.215	0.475

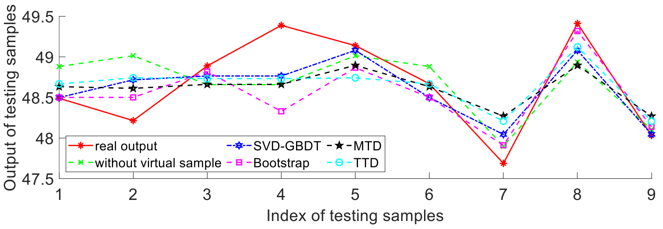

Figure 3. The output of testing set in different methods.

5. CONCLUSIONS

This paper proposes an SVD-GBDT based VSG method which consists of two parts. In SVD-GBDT-VSG, SVD is used to propose the main eigenvalues and eigenvectors associated with the model and expand the small samples according to the main eigenvectors; GBDT is used to synthesize the output of the virtual samples. The MSE and MAE are chosen as the detection indexes of virtual sample quality. Finally, the performance of the proposed SVD-GBDT-VSG is verified on the PTA industrial process. The simulation results show that the SVD-GBDT-VSG proposed in this paper can have a more satisfactory performance compared with other related methods. Meanwhile, in our future research, the SVD method will be further improved and other advanced techniques will be further investigated.

REFERENCES

He, Y. L., Tian, Y., Xu, Y. and Zhu, Q. X. (2020). Novel soft sensor development using echo state network integrated with singular value decomposition: Application to complex chemical processes. *Chemometrics and Intelligent Laboratory Systems*, 200, 103981.

Sánchez-Franco, M. J., Navarro-García, A., & Rondán-Cataluña, F. J. (2019). A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services. *Journal of Business Research*, 101, 499-506.

Arora, S., Du, S. S., Li, Z., Salakhutdinov, R., Wang, R., & Yu, D. (2019). Harnessing the power of infinitely wide deep nets on small-data tasks. *arXiv preprint arXiv*:1910.01663.

Kang, G., Wu, L., Guan, Y. and Peng, Z. (2019) A Virtual Sample Generation Method Based on Differential Evolution Algorithm for Overall Trend of Small Sample Data: Used for Lithium-ion Battery Capacity Degradation Data, *IEEE Access*, 7, 123255-123267.

Li, D. C., Chen, C. C., Chang, C. J., and Lin, W. K. (2012). A tree-based-trend-diffusion prediction procedure for small sample sets in the early stages of manufacturing systems. *Expert Systems with Applications*, 39(1), 1575-1581.

Arndt, M. F. (2009). A method for the generation of uniform point densities in samples that have an axis of symmetry and the monte carlo integration of functions whose domains have an axis of symmetry. Nuclear Instruments & Methods in Physics Research, 603(3), 532-540.

Campo, A. (2020). Simplistic solution of the graetz problem in the upstream sub-region x → 0 using nonlinear interpolation of the three-term, mean bulk temperature series for the downstream sub-region x →∞: a historical perspective. International Journal of Heat and Mass Transfer, 162.

Zhu, Q. X., Chen, Z. S., Zhang, X. H., Rajabifard, A., Xu, Y., & Chen, Y. Q. (2020). Dealing with small sample size problems in process industry using virtual sample generation: a Kriging-based approach. *Soft Computing*, 24(9), 6889-6902.

Maldonado, S., López, J., & Vairetti, C. (2019). An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing*, 76, 380-389.

Da Silva, A. N. C., dos Santos Amaral, R., dos Santos Junior, J. A., Vieira, J. W., & Menezes, R. S. C. (2015). Statistical analysis of discrepant radioecological data using Monte Carlo Bootstrap Method. *Journal of Radioanalytical and Nuclear Chemistry*, 306(3), 571-577.

Li, H., Liu, T., Wu, X., & Chen, Q. (2019). Research on bearing fault feature extraction based on singular value decomposition and optimized frequency band entropy. *Mechanical Systems and Signal Processing*, 118, 477-502.

Zhu, Q. X., Liu, D. P., Xu, Y., & He, Y. L. (2021). Novel space projection interpolation based virtual sample generation for solving the small data problem in developing soft sensor. *Chemometrics and Intelligent Laboratory Systems*, 217, 104425.

Yuan, X., Ou, C., Wang, Y., Yang, C., and Gui, W. (2019). A layer-wise data augmentation strategy for deep learning networks and its soft sensor application in an industrial hydrocracking process. *IEEE transactions on neural networks and learning systems*.

Tian, Y., Xu, Y., Zhu, Q. X., and He, Y. L. (2021). Novel Virtual Sample Generation Using Target-Relevant Autoencoder for Small Data-Based Soft Sensor. *IEEE Transactions on Instrumentation and Measurement*, 70, 1-10.