Monitoring Method of VOCs Based on PID in Soil-water-gas Environment

Dong Dong , Yongjun Ren , Mengmeng Zhang , Xiujuan Feng * , Kaiwei Liu *

Posted Date: 17 January 2024

doi: 10.20944/preprints202401.1137.v2

Keywords: Photoionization detector; VOCs; Principal Component Analysis; Genetic Algorithm

Article

Monitoring Method of VOCs Based on PID in Soil-Water-Gas Environment

Chengliang Dong [1,2,3], Yongjun Ren [4], Mengmeng Zhang [1,2,3] and Xiujuan Feng [1,2,3,*]

[1] School of Mines, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China
[2] Rare Earth Research Institute, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China
[3] Mechano Chemistry Research Institute, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China
[4] School of Computer science, Nanjing University of Information Science and Technology, Nanjing, Jiangsu 210044, China
* Correspondence: xjfeng@cumt.edu.cn

Abstract: In the moist environment of soil-water-air, there is a problem of low accuracy in monitoring Volatile Organic Compounds (VOCs) using a Photoionization Detector (PID). This paper analyzes the reasons for the low accuracy of the traditional Support Vector Machine (SVM) regression method. To address the issue, the PID signal is subjected to feature extraction and Principal Component Analysis (PCA) to reduce the data dimensionality. Moreover, the optimal SVM parameters are selected using a Genetic Algorithm (GA), and a combined approach of SVM regression with PCA and GA is utilized for PID signal regression analysis. And the effectiveness of the method is validated through extensive experiments and simulations. Furthermore, the influence of the sample quantity on the regression accuracy is analyzed, enabling accurate monitoring of VOCs concentration in a moist environment.

Keywords: photoionization detector; VOCs; principal component analysis; genetic algorithm

1. Introduction

Volatile organic compounds (VOCs) typically refer to a group of organic compounds with boiling points lower than 50°C to 260°C under normal atmospheric pressure [1]. VOCs themselves are toxic and are significant contributors to air pollution [2]. When volatile organic compounds are released into the atmosphere, they can cause widespread pollution, affecting air quality and posing health risks to individuals [1-2].

Benzene compounds, important components of VOCs, are commonly used organic solvents in the organic and coal chemical industries. If emitted into the environment in excess, they can cause symptoms of poisoning in humans [3]. Benzene compounds can also lead to dysfunction in the human nervous system [4]. Prolonged inhalation of benzene compounds can result in abnormal liver function, damage to the hematopoietic organs, and may even lead to symptoms of sepsis, causing abnormalities in human health and potentially triggering disorders such as aplastic anemia. In cases of large-scale vaporization of benzene, individuals may experience acute poisoning, which can lead to fatalities [5].

Efficient and precise VOCs monitoring methods are beneficial for accurately understanding the real-time dynamics of VOC concentrations. This helps enterprises, government agencies, and other departments to promptly implement corresponding measures, preventing harm to human health and the environment caused by VOCs. The Photoionization Detector (PID) is an important method for monitoring Volatile Organic Compounds (VOCs). Utilizing PID for VOC monitoring has characteristics such as high monitoring accuracy, non-destructive monitoring of VOCs, rapid response, long lifespan, and the ability to conduct VOC monitoring at atmospheric pressure [6]. PID typically can achieve monitoring of VOCs at the ppm level, and high-precision PIDs can even achieve monitoring at the ppb level (ppb is one thousandth of ppm). This allows for the detection of extremely

 doi:10.20944/preprints202401.1137.v2

low concentrations of volatile organic compounds. PID exhibits high sensitivity and enables non-destructive monitoring of VOCs. The VOCs being measured return to their original state after ionization in the ionization chamber. Due to its non-destructive monitoring characteristics, PID can also collaborate with other VOC monitoring technologies such as mass spectrometry, providing more information about the components and concentrations of the monitored VOCs. The application of PID is extensive, as it can monitor a wide range of volatile organic pollutants and is widely used in the monitoring of VOCs generated in industries such as coal chemical and petrochemical.

Currently, the calibration methods for Photoionization Detectors (PID) used in VOCs monitoring primarily involve determining the concentration of VOCs based on voltage values. This approach is evident in the research of Arnaud Termonia [7], Chung-hwan Je [8], Gianfranco Manes [9], Kentaro Oka [10], Qian [11], Wang J [12], Li [13], Wang Li [14], and others. However, when PID is applied to monitor VOCs in the soil-water-air environment, the signals generated by PID occurs simultaneously complex noise, impacting the numerical values of PID voltage and leading to misjudgments of VOC concentrations. Machine learning methods offer a solution by using multi-dimensional features of the signal instead of relying on a single voltage value for PID calibration. These multi-dimensional features help reduce the impact of noise on PID signals responding to VOC concentrations, thereby improving the robustness of the method [15].

Deep learning methods like artificial neural networks require the collection of a large number of VOCs samples, meaning the preparation of VOCs gas at specific concentrations [13-15]. It demands a substantial amount of human resources, materials, and financial investment [11-15]. And the Support Vector Machine (SVM) method, based on small-sample statistical learning theory, addresses this issue by constructing an optimal hyperplane that maximizes the distance between the hyperplane and different sets of samples in the sample or feature space [16]. The objective is to maximize the generalization ability. SVM demonstrates superior generalization ability compared to Deep learning methods such as artificial neural networks. In addition, the solution provided by SVM is the unique globally optimal solution. Therefore, this paper will investigate the method of using SVM regression (SVR) to monitor VOC concentrations in the soil-water-air environment.

2. Relate Work

2.1. Monitoring Method for VOCs based on PID

In 1997, Arnaud Termonia et al. [7] used gas chromatography-mass spectrometry in conjunction with PIDs to construct a VOCs monitoring system for monitoring VOCs in landfill sites. This method enables effective VOCs monitoring but it has issues such as high cost and maintenance difficulties. In 2007, Chung-hwan Je et al. [8] established an online VOCs monitoring system using a set of PID. They focused on the development and application of a multi-channel monitoring system based on PID for measuring, processing, and analyzing the concentration levels of VOCs emitted from a walk-in fume hood in hazardous waste management facilities. This system reduced the noise of PID signals by summing up the data over a time interval. In 2016, Gianfranco Manes et al. [9] addressed the issues of non-linear data, periodic calibration, and replacement distribution associated with PID in long-term monitoring in their VOCs online monitoring system based on PID. The first VOCs online monitoring system was installed in a petrochemical plant in Italy. Since its installation, the system has been continuously operational without human intervention. The successful operation of this system validates the feasibility of regional VOCs monitoring using PID.

In 2015, Qian Kun et al. [11] designed a low-power ZigBee sensor network and a data reception control framework between nodes based on a photoionization detector for monitoring VOCs in indoor environments. Their research focused on the design of a low-power ZigBee sensor network and a data reception control framework for real-time data acquisition and communication of VOCs air pollutant levels, enabling automated indoor VOCs monitoring. In 2019, Healy et al. [17] analyzed the principle and characteristics of VOCs detection using a photoionization detector. They also assessed the advantages and disadvantages of various VOCs detection methods and performed cost analysis. Additionally, they provided a detailed description of the circuit design and software system

 doi:10.20944/preprints202401.1137.v2

design for online monitoring systems. In 2020, Wang Jin et al. [12] discussed the issues of low accuracy and high cost associated with produced photoionization detectors in China, which make them difficult to widely deploy and utilize. Then they proposed a method to separate the sensor current detection module from the sensor radio frequency ultraviolet lamp driving module and designed a high-precision photoionization detector. In 2019, Li Hai et al. [13], based on the principles of photoionization technology, conducted theoretical analysis and simulations to determine various parameters of the ionization chamber in the photoionization detector (PID) according to practical conditions.

2.2. Quantitative analysis method for PID signal

Currently, the calibration method for PID used in VOCs monitoring primarily relies on voltage values to determine the current VOCs concentration, as demonstrated in the studies conducted by Arnaud Termonia [7], Chung-hwan Je [8], Gianfranco Manes [9], Kentaro Oka [10], Qian Kun [11], Wang Jin [12], Li Hai [13], Wang Lixin [14], and others. However, when PID is applied to monitor VOCs in real-world soil-water-air environments, the signals generated by PID can be affected by complex noise. This complex noise can impact the numerical value of PID voltage and result in inaccuracies in VOCs concentration estimation. Machine learning methods, on the other hand, can utilize multidimensional features of the PID signal instead of relying solely on voltage values for PID calibration. These multidimensional features help reduce the interference of noise when the PID signal responds to VOCs concentration, thus enhancing the accuracy of VOCs monitoring using PID.

In recent years, the artificial neural network method has been widely applied in various fields. However, artificial neural networks have drawbacks and limitations such as slow convergence speed, slow generalization, and a tendency to get trapped in local optima. Moreover, in practical VOCs engineering processes, artificial neural networks often require a large number of samples for training. In order to achieve precise VOCs concentration, it is necessary to prepare gas at a specific concentration. However, the preparation work is complex and requires a significant amount of manpower and financial resources [10-15].

The basic idea of SVM is to construct an optimal hyperplane that maximizes the distance between the hyperplane and the sample sets of different classes in the sample or feature space, aiming to achieve the goal of maximizing generalization ability [15, 16]. Unlike traditional artificial neural networks methods, SVM adopts a structural risk minimization criterion, minimizing the generalization error bound to achieve maximum generalization ability [18]. SVM has better generalization ability compared to artificial neural networks methods, and its solution is the unique global optimum. Therefore, in this paper, we apply SVR based on PID signal to monitor VOCs concentration.

3. PID Selection and Problem Statement

3.1. Calibration of PID for various VOCs

Different types of volatile organic compounds (VOCs) exhibit variations in the number of electrons generated and the extent of ionization after being ionized under high-energy ultraviolet light. As a result, the signal generated by the photoionization detector (PID) may reflect varying VOCs concentrations due to differences in their composition, despite the same concentration level. To establish a measurement standard, the PID employs a correction factor (*CF*) to compute the concentration of the monitored gas in relation to the standard gas [19]. In this study, benzene was used as the calibration gas for the PID, with a predefined calibration factor of 0.53. The *CF*, as defined by Equation (1), represents the correction factor for a specific component of VOCs gas, where C_b denotes the concentration of the standard benzene gas, R_b signifies the reading of the benzene gas used for calibration, C_m represents the concentration of a particular component of VOCs gas, and Rm denotes the reading of that specific component of VOCs gas.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2024 doi:10.20944/preprints202401.1137.v2

$$CF = \frac{R_b C_m}{C_b R_m} \quad (1)$$

It can be observed that different VOC gases have varying sensitivities in the PID. Certain VOC gases with high calibration coefficients exhibit low sensitivity in the PID, such as isobutanol and cyclohexane. On the other hand, some VOC gases with lower calibration coefficients demonstrate relatively higher sensitivity, such as styrene and chlorobenzene. This leads to varying detection accuracies of different gases by the PID. Hence, in practical VOC monitoring processes, the concentration of the actual monitored VOC gas needs to be multiplied by the corresponding response coefficient (*RF*) to obtain the respective VOC concentration, which can be determined by formula (2).

$$C_c = \rho * RF \quad (2)$$

In which C_c represents the actual concentration of the gas to be measured, and ρ represents the concentration displayed by the PID. Table 1 lists the response coefficients for some VOC gases [12].

Table 1. Response coefficients of common VOCs.

Chemicals	Response coefficient	Chemicals	Response coefficient	Chemicals	Response coefficient
benzene	1	acrolein	7.36	acetone	2.26
isobutanol	8.87	n-butyl	6.42	isobutene	1.887
cyclohexane	2.83	butyl acetate	4.53	butadiene	1.3
styrene	0.75	2-dimethylbenzene	1.02	propylene oxide	12.3
phenol	1.887	naphthalene	0.70	chlorobenzene	0.75

For example, if the PID response coefficient for isobutylene is 1, and the PID response coefficient for benzene is 0.5, it implies that when isobutylene generates a response value of 1V, benzene would produce a response value of 2 V.

3.2. PID Selection in the Work

The ultraviolet lamp in the PID is a crucial component that directly determines the performance of the photoionization detector. It has significant effects on important functional indicators of the PID, including the detection limits and accuracy of VOC monitoring. Furthermore, the ultraviolet lamp directly influences key performance indicators of the PID, such as power consumption, lifespan, and size [20, 21]. To ensure proper transmission, the ultraviolet lamp requires a material with specific lattice constants as the window material to facilitate the transmission of vacuum ultraviolet photons necessary for photoionization detection. Inert gas is filled inside the ultraviolet lamp to extend its lifespan and accelerate the ignition speed, thereby enhancing the output light intensity.

Currently, PIDs utilize ultraviolet lamps with three primary energy levels: 10.6 eV, 9.8 eV, and 8.4 eV [22]. The 10.6 eV ultraviolet lamp demonstrates stable performance during operation. In this research, a 10.6 eV AC-powered ultraviolet lamp was selected for the PID. It emits ultraviolet light with an energy level of 10.6 eV. The window material of this ultraviolet lamp is magnesium fluoride, enabling the detection of VOCs with ionization potentials lower than 10.6 eV, including benzene derivatives, alkenes, esters, and aldehydes.

The PID used in this study has a response time of less than 3 seconds and exhibits varying response values for different VOCs. The detection limit can be obtained by calculating the response coefficient. According to the PID manual, the maximum voltage that the photoionization detector can generate is 2.9V. If the detection limit for isobutylene by the PID is 1ppm-2000ppm, the theoretical detection limit for benzene should be 1ppm-1000ppm. VOCs with varying compositions have different detection limits, Others can be calculated using the response coefficient. The operational temperature range of the PID is -20 to 60 °C.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2024 doi:10.20944/preprints202401.1137.v2

3.3. Existing Problem in VOCs monitoring with PID

When monitoring VOCs in soil-water-gas systems using PID, it is necessary to volatilize the VOCs from the soil-water environment for measurement. To achieve this, monitoring devices are often equipped with heating devices. However, since VOCs are present in the soil-water medium, volatilization of VOCs is accompanied by a significant amount of water. The presence of this water content increases the humidity in the PID monitoring environment. In an environment with increased humidity, the PID's photocathode surface may exhibit barriers and uneven distribution, resulting in abundant low-frequency noise in the PID signal. This noise can affect the accuracy of the PID voltage readings, thereby leading to misinterpretation of VOCs concentrations.

SVM exhibits superior generalization capability compared to traditional machine learning methods such as Artificial Neural Networks. Moreover, SVM provides a unique global optimum solution. Hence, the research investigates the utilization of SVR for monitoring VOC concentration in soil-water-air environments.

4. Analysis of VOC concentration based on traditional SVR

4.1. SVR

Assuming a given training sample set $\{(x_1, y_1),\ldots\ldots,(x_n, y_n)\}x_i, y_i \in R$, considering the use of a linear regression function [16, 23].

$$F(x) = wx + b \tag{3}$$

To ensure the flatness of function $F(x)$, it is crucial to find the smallest value of w. Therefore, the generalization of the Euclidean space is minimized. Assuming that all training data points (x_i, y_i) can be approximated by a linear function within an accuracy of ε, the problem of finding the minimum value of w can be formulated as a convex optimization problem.

$$\min \frac{1}{2} \| w \|^2 \tag{4}$$

The constraint condition is as follows.

$$\begin{cases} y_i - w \cdot x_i - b \le \varepsilon \\ w \cdot x_i + b - y_i \le \varepsilon \end{cases} \tag{5}$$

In consideration of allowing fitting errors, a relaxation factor is introduced: $\xi_i \ge 0$ and $\xi_i^* \ge 0$. Similar to maximizing the classification margin in the optimal classification hyperplane, the problem of regression estimation is transformed into the following two equations.

$$\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} (x_i + x_i^*) \tag{6}$$

The constraint condition is as follows

$$\begin{cases} y_i - w \cdot x_i - b \le \varepsilon + \xi_i \\ w \cdot x_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i \ge 0 \\ \xi_i^* \ge 0 \end{cases} \quad i = 1,\ldots n \tag{7}$$

The constant C > 0 is used to balance the flatness of the regression function F and the number of sample points with a bias greater than ε. Equations (6) and (7) are derived from the ε-insensitive loss function represented by the following equation (8). The function $|\xi|_\varepsilon$ is expressed as follows:

$$|\xi|_\varepsilon := \begin{cases} 0 & \text{if } |\xi| \le \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases} \tag{8}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2024 doi:10.20944/preprints202401.1137.v2

When dealing with a limited number of samples, the solution to SVM is commonly approached using the duality theory, which transforms it into a quadratic programming problem. To accomplish this, the Lagrange equation is established.

$$l(w, \xi, \xi^*) = \frac{1}{2}(w \cdot w) + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i(\varepsilon + \xi_i + y_i - w, x_i - b) - \ldots \\ \ldots\sum_{i=1}^{n}\alpha_i(\varepsilon + \xi_i^* + y_i - w, x_i - b) - (\eta_i\xi_i + \eta_i^*\xi_i^*) \tag{9}$$

The partial derivatives of the parameter w, b, ξ_i, ξ_i^* should all be equal to zero. Substituting this condition into equation (9) results in the dual optimization problem.

$$\min\frac{1}{2}\sum_{i,\ j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i,x_j + \sum_{i=1}^{n}\alpha_i(\varepsilon - y_i) + \sum_{i=1}^{n}\alpha_i^*(\varepsilon + y_i) \tag{10}$$

$$\text{s.t}\begin{cases}\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\ \alpha_i,\ \alpha_i^* \in [0, C]\end{cases} \tag{11}$$

For nonlinear regression problems, assuming that the samples X are mapped to a high-dimensional space using a nonlinear function [24], the regression problem is then transformed into minimizing the function under the constraint equations (12).

$$\frac{1}{2}\sum_{i,\ j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) < \phi(x_i),\phi(x_j) > + \sum_{i=1}^{n}\alpha_i(\varepsilon - y_i) + \sum_{i=1}^{n}\alpha_i^*(\varepsilon + y_i) \tag{12}$$

As a result, the following equation is derived: $w = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\phi(x_i)$.

Kernel function

In SVR, the kernel function is employed to simplify nonlinear approximation [25]. If the kernel function $k(x, x')$ satisfying $k(x, x') = \phi(x), \phi(x')$, the following equation is formulated.

$$\min\frac{1}{2}\sum_{i,\ j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x, x') + \sum_{i=1}^{n}\alpha_i(\varepsilon - y_i) + \sum_{i=1}^{n}\alpha_i^*(\varepsilon + y_i) \tag{13}$$

A kernel function $k(x, x')$ is a symmetric positive definite function that must satisfy the Mercer condition:

$$\iint k(x, x')g(x)g(x')dxdx' > 0,\ g \in L_2 \tag{14}$$

The selection and construction of kernel functions were discussed in Ref. [16]. According to it, in the work, the construction of the SVR was performed using the following Gaussian radial basis function (RBF) kernel.

$$k(x, x') = exp(-\frac{||x - x'||^2}{2\delta^2}) \tag{15}$$

4.2. Analysis of VOC concentration based on traditional SVR

Based on the characteristics of SVR in constructing regression models for small-sample data, this work utilizes SVM to build a regression model for VOC concentration data based on PID response.

Initially, a total of 84 sets of VOC concentration data were randomly arranged, with 67 sets used as the training set and 17 sets as the testing set. Each set comprised 6,001 values, where the first 6,000 values represented the PID response to VOC concentration, and the 6,001st value represented the true VOC concentration.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2024 doi:10.20944/preprints202401.1137.v2

To eliminate the adverse effects caused by anomalous samples, both the training and testing sets were normalized. It is essential to normalize the data for SVM as it ensures feature scaling, preventing elliptical feature distributions that could hinder model training and lead to convergence issues or poor prediction accuracy.

The parameters of SVR were set, and the Gaussian radial basis kernel function was chosen as the core function for SVR. Different values of kernel function parameters and penalty factor C were experimented to observe their impact on the accuracy of SVR. Since no prior knowledge was available, the penalty factor C was initially set to 0.1 and the kernel function parameter was temporarily set to 1. Under these settings, the R^2 value of the SVR was 0.21. When C was set to 0.2, the R^2 value decreased to 0.13, and further reduced to 0.015 with C set to 0.3. The model accuracy was not satisfactory. To improve it, C was subsequently set to 1, resulting in an R^2 value of 0.30. Interestingly, when C was increased to 2, 3, and 5, the R^2 value remained at 0.30, but it slightly increased to 0.38 with C set to 4. However, no clear mathematical relationship between C and model accuracy was identified. Thus, for SVR, a penalty factor of 4 was provisionally in the work.

The traditional SVR was ultimately employed to construct a model using a penalty factor of C = 4 and kernel function parameter of $\beta = 0.8$ for 84 different concentrations of VOCs. The test set results of this model are shown in Figure 1. The mean squared error (MSE) of this model was 69458.6, and the coefficient of determination (R^2) was 0.38. However, there was a significant discrepancy between the predicted VOCs concentrations and the actual VOCs concentrations, indicating a large prediction error.

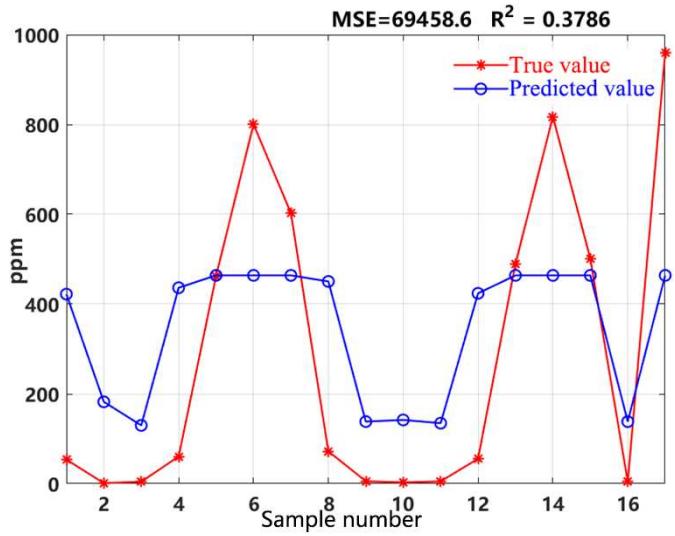

Figure 1. Quantitative analysis effect of VOCs concentration based on traditional SVR.

After repeated comparisons and analysis, the main reasons for the high errors of the model are as follows. Firstly, the uncertainty in determining optimal values for the penalty factor C and kernel function parameters makes it difficult for determination of their optimal values. Secondly, the original training data used in this model contains redundancies, where relevant information is not well-identified while irrelevant information impacts the accuracy of the model construction. Hence, this work will optimize it from the two aspects.

5. Discussion

In the section, the genetic algorithm (GA) is employed to automatically select the values of the penalty factor C and the parameters of the kernel function in SVR. This approach utilizes the expansive search space and global search capability of the genetic algorithm. As a result, an optimized SVR model is built, and the concentration of VOCs is determined through signal generation using PID.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2024 doi:10.20944/preprints202401.1137.v2

5.1. SVR Based on PCA of PID Signal Features

5.1.1. Subsubsection

Principal Component Analysis (PCA) is an effective and widely applied dimensionality reduction algorithm. It decomposes the principal components into mutually orthogonal directions, thereby effectively eliminating redundant and overlapping information among the original data. Generally, a few significant principal components can cover the majority of information regarding the signal produced by VOCs in PID response. The computational procedure of the PCA algorithm is as follows [16].

(1) Perform zero-mean normalization on the sample set of dimensionality d and $O_i = (O_1, O_2, \ldots\ldots, O_n)$

$$O_i = O_i - \frac{1}{n}\sum_{i=1}^{n} O_i \tag{16}$$

(2) Compute the covariance matrix $\sum$ of vector O;
(3) Use the method of singular value decomposition to obtain the eigenvalues and eigenvectors of the covariance matrix $\sum$.
(4) Take the eigenvectors corresponding to the top v eigenvalues to form a new matrix, where v should be smaller than n.
(5) Obtain a new low-dimensional sample set and calculate the contribution rate of each principal component and the cumulative contribution rate.

The time-domain and frequency-domain features of the PID response signals to VOCs were utilized as the original dataset for PCA to extract the principal components. The contributions of the mean, mean frequency, centroid frequency, root mean square frequency, frequency standard deviation, standard deviation, skewness, kurtosis, and maximum value are 0.4495，0.2439，0.1082，0.0787，0.0557，0.0300，0.0152，0.0080，0.0065, respectively. The cumulative contribution rate of these nine features is 0.9958. Therefore, these nine features are used to replace the original 16 features.

5.1.2. SVR after PCA of PID signal features

Perform PCA algorithm on the 12 time-domain features and 4 frequency-domain features of 84 different sets of VOCs concentration signals. Use the feature data obtained from principal component analysis of 67 signal sets as the training set, and the feature data obtained from principal component analysis of 17 signal sets as the testing set. Each data set consists of 10 values, where the first 9 values are the PCA-based feature parameters of the PID response signal to VOCs concentration, and the 10th value represents the true concentration of VOCs. Then, the training and testing sets are normalized, and the SVM parameters are set with the Gaussian radial basis kernel function chosen as the kernel function for SVR. Set the parameters of the radial basis kernel function to 0.8 and the penalty factor C in SVR as 4. Finally, the results of the testing set are shown in Figure 2. The model has a mean squared error of 372 and an R^2 value of 0.996. From the Figure 2, it can be seen that the regression performance of the model is improved, but it still has some errors. Based on the analysis, these errors are attributed to suboptimal choices of the penalty factor C and kernel function parameters. To address this issue, the work employs GA algorithm to optimize them.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2024 doi:10.20944/preprints202401.1137.v2

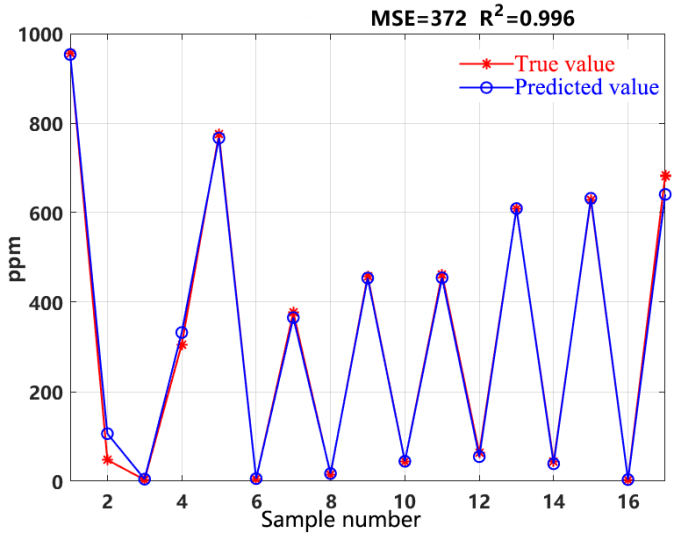

Figure 2. VOCs concentration regression by SVR after PCA of PID signal features.

5.2. Proposed method based on PCA-GA-SVR

5.2.1. SVR after PCA of PID signal features

The genetic algorithm simulates the problem-solving process as a biological evolution, generating the next generation of solutions through operations such as reproduction, crossover, and mutation. It gradually eliminates solutions with low fitness values and increases solutions with high fitness values. After evolving for N generations, it is highly likely to obtain individuals with high fitness values, which represent the optimal results of the objective function [26]. The steps for selecting the optimal kernel function parameters and penalty factor using genetic algorithm are as follows.

(1) The dataset consisting of 84 group different concentrations of PID response to VOCs was split into an 80% training set and a 20% testing set.
(2) Normalize the input of the training and testing sets.
(3) Set the parameters of the genetic algorithm, such as the population size, iteration count, crossover probability, mutation probability, etc. Here, the chromosome dimension is set to 2, where the two numbers in the chromosome represent δ and C.
(4) Initialize the population by initializing each chromosome and calculating its objective function value.
(5) Begin iterative loop.
(6) Selection operator.
(7) Crossover and mutation operators (simulated binary crossover and polynomial mutation).
(8) Recalculate the objective function value for the updated chromosomes, where the objective function is the minimum mean squared error.
(9) Update the optimal objective of the global best chromosome.
(10) Proceed to the next iteration until the maximum iteration count is reached.
(11) Export the global best chromosome and C values, and plot the iteration curve.

5.2.2. SVR after PCA of PID signal features

When applying genetic algorithms to problem-solving, there are two encoding methods for chromosomes: binary encoding and floating-point encoding. The floating-point chromosome encoding is suitable for solving problems with a large value range, while the binary encoding is suitable for problems with a smaller value range. Since this paper applies a genetic algorithm to optimize SVM regression and involves processing small-sample data, the binary encoding method is adopted in this paper. Through the study of parameter settings for support vector machine regression in domestic and foreign research, the minimum values for penalty factor and kernel function

 doi:10.20944/preprints202401.1137.v2

parameters in this study were determined as 0.001. The maximum value for the penalty factor was set to 100, and the maximum value for the kernel function parameter was set to 10. Genetic algorithm operations are performed based on a population. During iterations, an initial population is provided to the genetic algorithm, and subsequent iterations are performed using this population. The population size was set to 20. Through experiments, it was determined that the convergence of the iteration curve occurs within 50 iterations, hence the iteration limit was set to 50. The population information was defined as a structure, and the population loop was initiated. Chromosomes are formed by encoding, and for each gene of the chromosome, a random number is generated between its maximum and minimum values. This represents the chromosome. The penalty factor is represented by the first number of the chromosome, and the kernel function parameters are represented by the second number of the chromosome.

GA distinguishes individuals based on the evaluation of the fitness function value for each chromosome. In GA, the larger the fitness value of a chromosome, the better the individual it represents. After initialization, the optimal objective and its corresponding chromosome are identified, followed by iterative processes, selection operators, computation of current objective fitness, calculation of the fitness proportion for each chromosome, and generation of nonzero random numbers within the population.

The cumulative fitness proportion is computed by iterating through the population, and when it exceeds the random number, the last accumulated individual is chosen as the selected individual. Select offspring of the same size as the original population, calculate the optimization variable dimension and population number, and then determine the crossover probability. Compare a random number with the crossover probability to decide whether to perform crossover. If crossover is performed, randomly select two different chromosomes. We export the two selected chromosomes and iterate through each dimension of the chromosome. The crossover operator simulates binary crossover and then limits the boundaries of the crossed individuals, replacing values exceeding the maximum with the maximum value and values below the minimum with the minimum value. Copy the resulting individuals back into the population, and then loop through the population. Generate a random number and compare it with the crossover mutation probability. If a mutation occurs, randomly select an individual for mutation preparation. Loop through each gene of the chromosome, perform polynomial mutation on the selected chromosome. Then copy the mutated individuals back into the population. Recalculate the objectives of the offspring from the crossover-mutated individuals to obtain the best and worst objectives. Then compare the current best objective with the global best objective. Replacing the worst with the historical global best increases the probability of the population iterating towards better individuals. Record the average objective and the best objective of the current generation. After the iteration, export C and δ.

The GA optimizes the iterative curve of SVR parameters, as shown in Figure 3. The iteration curve has converged at 16 iterations, which means that the minimum value of MSE can be obtained after the 16th iteration.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2024 doi:10.20944/preprints202401.1137.v2

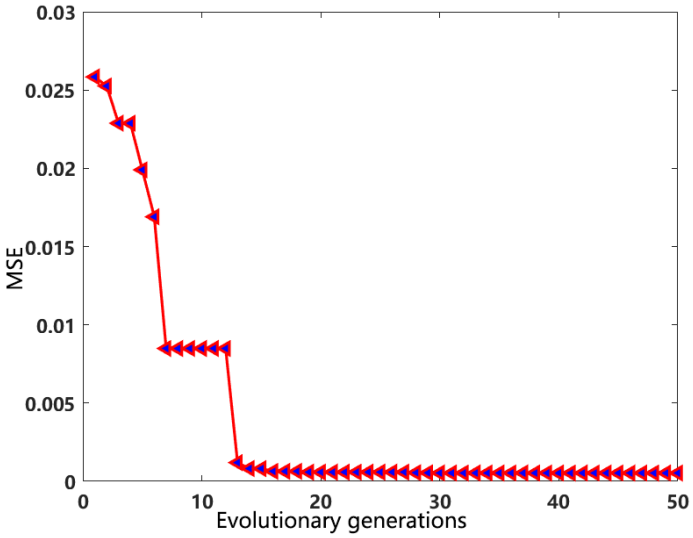

Figure 3. trend of prediction changes with the evolution generations.

5.2.3. Results

In the genetic algorithm optimized SVM regression method based on PID signal time-domain and frequency-domain with principal component analysis, 84 sets of different PID generate 9 principal component features that reflect the VOCs concentration signal. These 9 features are considered as the characteristics of the signal. Among these, 67 sets of features obtained from the principal component analysis are used as the training set, while 17 sets of features obtained from the principal component analysis are used as the test set. Each set of data consists of 10 values, where the first 9 values represent the principal component analysis parameters of the PID response signal to VOCs concentration, and the 10th value represents the true concentration of VOCs.

The training set and test set are normalized. Through genetic algorithm, the optimal parameters for the radial basis kernel function are found to be 0.0101, and the optimal penalty factor C is 7.8783. Based on the chosen parameters, the support vector machine regression function is derived as F(x) = w*x + b. The results of the test set are shown in Figure 4. The mean squared error of this model is 0.000059, and the R^2 value is 0.9999. Compared with the R^2 values obtained by Wang Jin, which is 99.8%, and Li Hai, which is 98.2%, 98.5%, 96.9%, etc., the proposed research method demonstrated superiority.

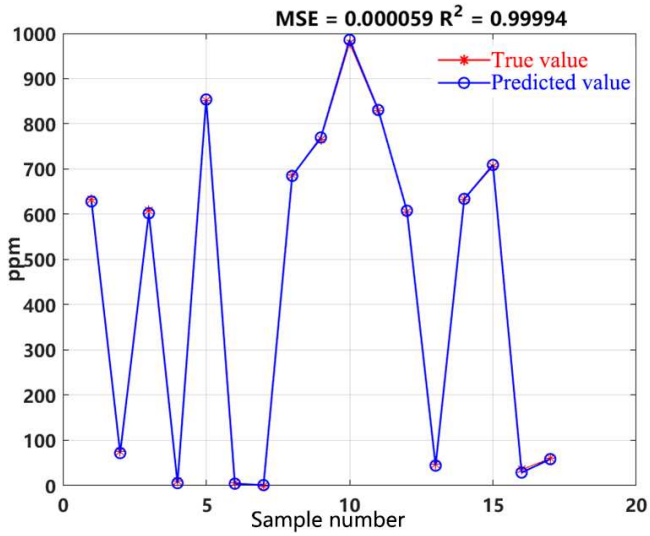

Figure 4. Analysis results for VOCs concentration based on PCA-GA-SVR.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2024 doi:10.20944/preprints202401.1137.v2

5.2.4. Results

In order to verify the effectiveness of the high-accuracy models for the 84 samples in this study and to determine the minimum sample size required for constructing the PCA-GA-SVM model, the study conducted experiments based on our training data and 17 sets of testing data. By sequentially reducing 4 sets of training data and 1 set of testing data, the study aimed to find the minimum sample size for model establishment. Among them, A represents the original data; B represents the data with a reduction of 4 sets of training data and 1 set of testing data; C represents the data with a reduction of 8 sets of training data and 2 sets of testing data; D represents the data with a reduction of 12 sets of training data and 3 sets of testing data; E represents the data with a reduction of 16 sets of training data and 4 sets of testing data; F represents the data with a reduction of 20 sets of training data and 5 sets of testing data. The accuracy results of the PCA-GA-SVR are shown in Figure 5.

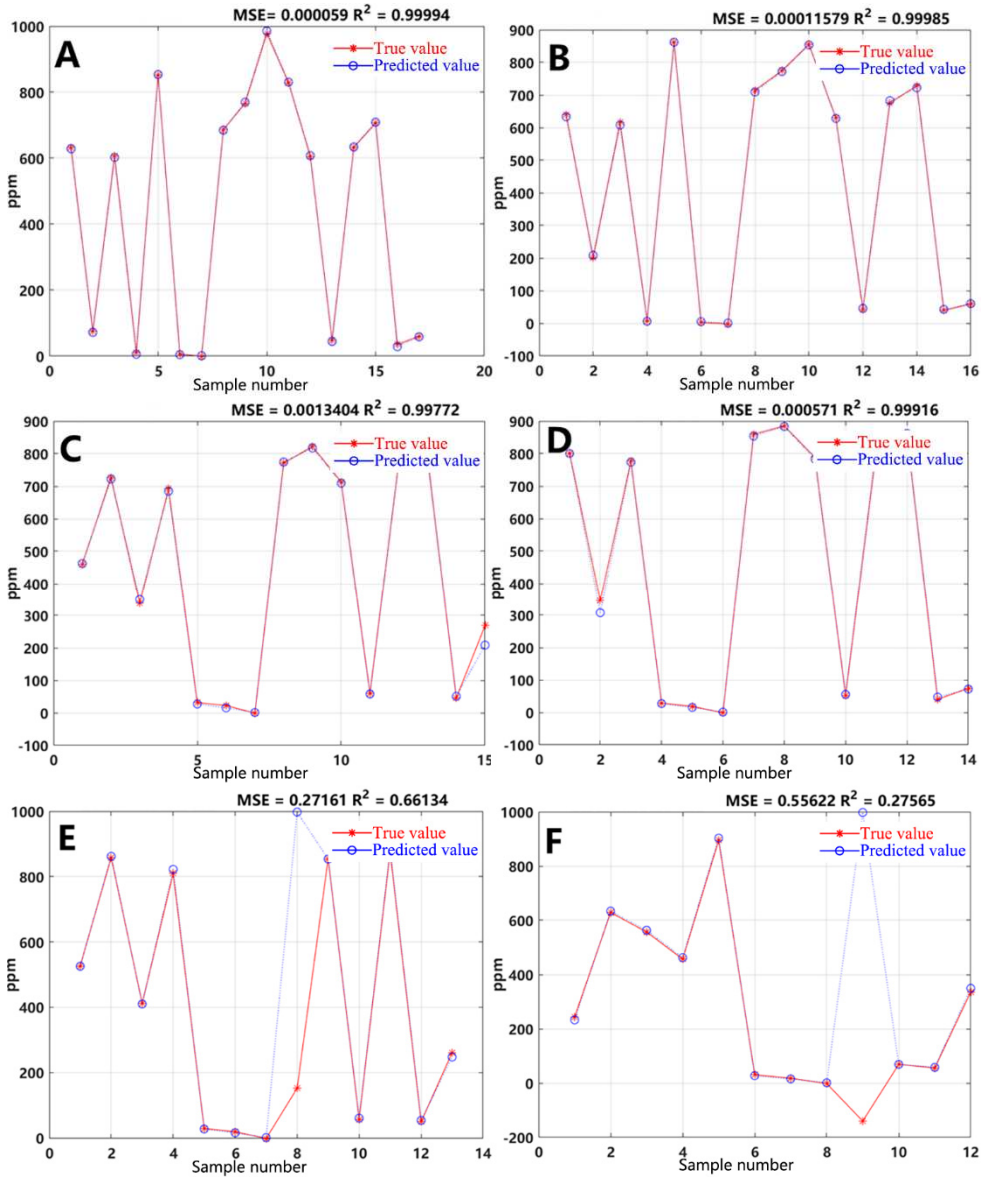

Figure 5. Effect of sample size on regression accuracy.

By reducing the number of training sets to 12 and testing sets to 3, the R^2 of the model can still be maintained above 0.99. However, if the number of training and testing sets is further reduced, the accuracy of the PCA-GA-SVM model will drastically decrease. Therefore, in the calculation of VOCs concentration using the PCA-GA-SVR based on signals generated by PID, it is necessary to ensure

that the sample size is greater than 69. In this study, the PID response to VOCs signals consists of 84 groups, thus meeting the requirements for model establishment.

6. Conclusion

This paper addresses the issue of long computation time and low accuracy in SVR caused by redundant data information. It conducts PCA on time-frequency features to reduce the data dimensionality. Furthermore, it solves the difficulty in determining the optimal values for the penalty factor C and kernel function parameters in traditional SVR by utilizing genetic algorithms. This approach effectively improves the generalizability and robustness of the SVR. The mean squared error of signal time-frequency feature extraction - PCA-GA-SVR is 0.000059, with an R^2 of 0.9999.

Moreover, this paper analyses the impact of the number of experimental samples on the regression accuracy of the signal time-frequency feature extraction-PCA-GA-SVR. The model maintains a high level of accuracy when the number of samples exceeds 69 groups. It also confirms that the 84 sets of data in the study meet the sample requirements for the regression. This demonstrates the effectiveness of the PCA-GA-SVR method for VOCs monitoring in a humid environment and validates the effectiveness and robustness of the proposed method in the paper.

Author Contributions: Conceptualization, X.F.; methodology,X.F,L.K; software,C.D. and Y.R.; validation, Y.R., M.Z. and X.F.; formal analysis, X.F,L.K; investigation, C.D., Y.R., M.Z. ,L.Kand X.F.; resources, X.F.; data curation, Y.R. and M.Z.; writing—original draft preparation, Y.R.; writing—review and editing, X.F.; visualization, Y.R ,,L.K and M.Z.; supervision, C.D., R.Y., M.Z. and X.F.; project administration, X.F.; funding acquisition, X.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shandong Province Key R&D Program (Typical heavy industry soil pollution monitoring, early warning and remediation technology integration and equipment research and development, grant number 2021CXGC011206".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Zhong, Y.; Wang, Y. M.; Ji, Y. F.; Zhang, X. T.; Wang, X. M. Biomass carbon-based composites for adsorption/photocatalysis degradation of VOCs: A comprehensive review. *Colloid and Interface Science Communications* **2023,** 57, 100749.
2. Wang, S. S.; Zhang, J.; Zhang, Y.; Wang, L. W.; Sun, Z. X.; Wang, H. L. Review on Source Profiles of Volatile Organic Compounds (VOCs) in Typical Industries in China. *Atmosphere* **2023,** 14, (5), 878.
3. Li, X. Q.; Zhang, L.; Yang, Z. Q.; He, Z. Q.; Wang, P.; Yan, Y. F.; Ran, J. Y. Hydrophobic modified activated carbon using PDMS for the adsorption of VOCs in humid condition. *Sep. Purif. Technol.* **2020,** 239, 116517.
4. Kuranchie, F. A.; Angnunavuri, P. N.; Attiogbe, F.; Nerquaye-Tetteh, E. N. Occupational exposure of benzene, toluene, ethylbenzene and xylene (BTEX) to pump attendants in Ghana: Implications for policy guidance. *Cogent Environmental Science* **2019,** 5, (1), 1603418.
5. Wang, Y.; Zhou, B.; Yang, M. R.; Xiao, G.; Xiao, H.; Dai, X. R. Bibliometrics and Knowledge Map Analysis of Research Progress on Biological Treatments for Volatile Organic Compounds. *Sustainability* **2023,** 15, (12), 9274.
6. Liu, H.; Meng, G.; Deng, Z.; Li, M.; Chang, J.; Dai, T.; Fang, X. Progress in Research on VOC Molecule Recognition by Semiconductor Sensors. *Acta Physico-Chimica Sinica* **2022,** 38, (5).
7. Termonia, A.; Termonia, M. Characterisation and on-site monitoring of odorous organic compounds in the environment of a landfill site. *International Journal of Environmental Analytical Chemistry* **1999,** 73, (1), 43-57.
8. Je, C.-h.; Stone, R.; Oberg, S. G. Development and application of a multi-channel monitoring system for near real-time VOC measurement in a hazardous waste management facility. *Sci. Total Environ.* **2007,** 382, (2-3), 364-374.

9. Manes, G.; Collodi, G.; Gelpi, L.; Fusco, R.; Ricci, G.; Manes, A.; Passafiume, M. Realtime Gas Emission Monitoring at Hazardous Sites Using a Distributed Point-Source Sensing Infrastructure. *Sensors* **2016,** 16, (1), 121.
10. Oka, K.; Iizuka, A.; Inoue, Y.; Mizukoshi, A.; Noguchi, M.; Yamasaki, A.; Yanagisawa, Y. Development of a Combined Real Time Monitoring and Integration Analysis System for Volatile Organic Compounds (VOCs). *International Journal of Environmental Research and Public Health* **2010,** 7, (12), 4100-4110.
11. Peng, C.; Qian, K.; Wang, C. Design and Application of a VOC-Monitoring System Based on a ZigBee Wireless Sensor Network. *IEEE Sens. J.* **2015,** 15, (4), 2255-2268.
12. Wang, J.; Hao, X. W.; Dong, J. G.; Xiong, J. J.; Hong, Y. P. Design of high precision photoionization detector. *Infrared and Laser Engineering* **2020,** 49, (8), 248-255.
13. Li, H. VOCs Detection Based on Photoionization Technology. Master, Chongqing University of Posts and Telecommunications, Chongqing, 2020.
14. Wang, L.; Cheng, Y.; Gopalan, S.; Luo, F.; Amreen, K.; Singh, R. K.; Goel, S.; Lin, Z.; Naidu, R. Review and Perspective: Gas Separation and Discrimination Technologies for Current Gas Sensors in Environmental Applications. *Acs Sensors* **2023,** 8, (4), 1373-1390.
15. Liu, Z.; Feng, X.; Dong, C.; Jiao, M. Study on Denoising Method of Photoionization Detector Based on Wavelet Packet Transform. *Chemosensors* **2023,** 11, (2), 146.
16. Roy, A.; Chakraborty, S. Support vector machine in structural reliability analysis: A review. *Reliability Engineering & System Safety* **2023,** 233.
17. Healy, R. M.; Wang, J. M.; Karellas, N. S.; Todd, A.; Sofowote, U.; Su, Y.; Munoz, A. Assessment of a passive sampling method and two on-line gas chromatographs for the measurement of benzene, toluene, ethylbenzene and xylenes in ambient air at a highway site. *Atmospheric Pollution Research* **2019,** 10, (4), 1123-1127.
18. Shi, J.; Teh, J. Load forecasting for regional integrated energy system based on complementary ensemble empirical mode decomposition and multi-model fusion. *Applied Energy* **2024,** 353, 122146.
19. Zhu, H.; Nidetz, R.; Zhou, M.; Lee, J.; Buggaveeti, S.; Kurabayashi, K.; Fan, X. Flow-through microfluidic photoionization detectors for rapid and highly sensitive vapor detection. *Lab on a Chip* **2015,** 15, (14), 3021-3029.
20. Bilek, J.; Marsolek, P.; Bilek, O.; Bucek, P. Field Test of Mini Photoionization Detector-Based Sensors-Monitoring of Volatile Organic Pollutants in Ambient Air. *Environments* **2022,** 9, (4), 49.
21. Liu, R. Y.; Hu, H. Design of Photoionization Sensor for VOC Gas Detection. *Instrument Technique and Sensor* **2020,** 7, 1-5.
22. Zhou, Q.; Zhang, S.; Zhang, X.; Ma, X.; Zhou, W. Development of a Novel Micro Photoionization Detector for Rapid Volatile Organic Compounds Measurement. *Applied Bionics and Biomechanics* **2018,** 2018, 5651315.
23. Das, S.; Khanwelkar, D. R.; Maiti, J. A semi-automated coding scheme for occupational injury data: An approach using Bayesian decision support system. *Expert Systems with Applications* **2024,** 237, 121610.
24. Rymarczyk, T.; Klosowski, G.; Hola, A.; Sikora, J.; Tchorzewski, P.; Skowron, L. Optimising the use of Machine learning algorithms in electrical tomography of building Walls: Pixel oriented ensemble approach. *Measurement* **2022,** 188, 110581.
25. Ding, S.; Zhao, X.; Zhang, J.; Zhang, X.; Xue, Y. A review on multi-class TWSVM. *Artificial Intelligence Review* **2019,** 52, (2), 775-801.
26. Wang, Y.; Xue, W. Sustainable development early warning and financing risk management of resource-based industrial clusters using optimization algorithms. *Journal of Enterprise Information Management* **2022,** 35, (4/5), 1374-1391.