Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 7387398, 15 pages
https://doi.org/10.1155/2019/7387398

Research Article

Application of Data Denoising and Classification Algorithm Based on RPCA and Multigroup Random Walk Random Forest in Engineering

Renchao Wang, Yanlei Wang, and Yuming Ma

State Key Laboratory of Hydraulic Engineering Simulation and Safety, Tianjin University, Tianjin 300350, China

Correspondence should be addressed to Yanlei Wang; tjuwyl@tju.edu.cn

Received 27 June 2019; Revised 30 October 2019; Accepted 30 November 2019; Published 24 December 2019

Academic Editor: John S. Sakellariou

Data classification algorithms are often used in the engineering field, but the data measured in the actual engineering often contains different types and degrees of noise, such as vibration noise caused by water flow when measuring the natural frequencies of aqueducts or other hydraulic structures, which will affect the accuracy of classification. In reality, these noises often appear disorganized and stochastic and some existing algorithms exhibit poor performance in the face of these non-Gaussian noise. Therefore, the classification algorithms with excellent performance are needed. To address this issue, a hybrid algorithm of robust principal component analysis (RPCA) combined multigroup random walk random forest (MRWRF) is proposed in this paper. On the one hand RPCA can effectively remove part of non-Gaussian noise, and on the other hand MRWRF can select a better number of decision trees (DTs), which can effectively improve random forest (RF) robustness and classification performance, and the combination of RPCA and MRWRF can effectively classify data with non-Gaussian distribution noise. Compared with other existing algorithms, this hybrid algorithm has strong robustness and preferable classification performance and can thus provide a new approach for data classification problems in engineering.

1. Introduction

Data classification is one of the data mining problems receiving enormous attention [1]; many scholars have carried out relevant research and made great progress in many fields [2–5], and it is also often used in engineering. Initially, it was mainly to solve engineering problems with a single classic classification algorithm. For example, the support vector machine (SVM) [6] is applied to identify structural damage [7], the relative change quantity of modal flexibility is input in the SVM classifier to identify the location and degree of structural damage, and analysis results indicate that this method is feasible to identify the location and degree of structure damage with low noise. A novel method of damage identification for beam using the artificial neural network (ANN) [8] based on statistical properties of structural dynamic responses is developed [9], and it takes the change in the structural response variance as the input and the damage state as the output of the ANN. Experimental results show that the ANN can correctly detect the damage location and identify the damage extent with high precision. The Bayesian network classifier is applied into transformer fault diagnosis it has been proven to evaluate the condition of transformers and predict their potential damage by using electrical test data such as dissolved gas analysis [10]. With the improvement of machine learning technology, many scholars have made various attempts to better solve engineering problems. The efforts are mainly divided into two types: (1) some scholars have proposed new data classification algorithms, such as the chaotic salp swarm algorithm [11] and data classification methods based on fuzzy logic [12] and (2) some scholars have improved an existing algorithm, and the improvement usually optimizes the parameters of the existing algorithm or combination two and more algorithms, such as combining the SVM with the KNN and applying the method to visual category recognition [13]. An example of damage

identification of a concrete frame structure shows that the combination algorithm with kernel principal component analysis (KPCA) and the SVM has a certain antinoise ability [14]. A multigroup particle swarm optimization algorithm based on real coding is proposed and applied to the classification of the damage data of large structures under environmental excitation, and the results show that the algorithm has better noise immunity [15].

When solving various engineering problems with various algorithms, noise effects are unavoidable in engineering data acquisition systems. Thus, advanced noise reduction algorithms should be applied to minimize the impact of data corruption in problem solving.

Many data denoising algorithms have emerged in the early time and these include wavelet transform, principal component analysis (PCA), independent component analysis (ICA), the adaptive filtering, neural networks (NNs), and empirical mode decomposition (EMD) [16]. These methods and the variations of these methods are widely used in signal data denoising and image data denoising [17, 18]. Besides, some new noise reduction algorithms are proposed and applied in new engineering fields in recent years, e.g., wavelet packet transform (WPT) is used to extract data from acoustic emission signals containing noisy data [19], noise reduction for desert seismic data using spectral kurtosis adaptive bandpass filter [20], the 1D undecimated discrete wavelet transform (UDWT) has been acquired to attenuate random noise and ground roll [21], a new denoising method was proposed for the simultaneous noise reduction and preservation of seismic signals based on variational mode decomposition (VMD) [22], PCA + linear discriminate analysis (LDA) is first used to extract and denoise the original data, and then nearest neighbor (NN) is used to classify the processed data [23]. These methods are usually for Gaussian distribution noise when performing robustness analysis. In practice, these seemingly chaotic noises do not necessarily obey Gaussian distribution.

Therefore, we firstly try to use robust principal component analysis (RPCA) [24] to denoise the signal data containing non-Gaussian noise in engineering field, which has excellent performance in the field of image noise reduction. It is worth mentioning that the data denoising algorithm is only to reduce the influence of noise as much as possible and cannot completely eliminate the noise. Therefore, we also need excellent classification algorithms to improve classification accuracy. To address this issue, multigroup random walk random forest (MRWRF) is proposed to classify RPCA-purified data. RPCA can effectively take out part of non-Gaussian noise from raw data, and MRWRF can select a better number of decision trees (DTs), which can effectively improve random forest (RF) robustness and classification performance. The hybrid of RPCA and MRWRF can effectively classify data with non-Gaussian distribution noise, and a detailed introduction will be given in the following sections.

2. Related Work

Many applications in engineering are characterized by large quantities of very high-dimensional data. Although these data often lie in very high-dimensional observation spaces, these many dimensions may express only a few intrinsic degrees of freedom [24]. It can be indicated in the following form:

$$M = A + E, \tag{1}$$

where M denotes the contaminated data matrix, A is low-rank, and E is a noise term.

Statistically, the above problem is equal to exploring the principal components of the data. When N follows Gaussian distribution with small variance, the PCA can handle the above problem well. However, the performance of PCA is limited due to lack of robustness to gross corruptions [25]. To overcome the disadvantages of PCA, some robust principal component analysis methods [24, 26] have emerged in recent years. In particular, Zhang et al. [27] established a RPCA method which is a powerful tool for various applications [27–29]. Its noise reduction process can be described by the following optimization problem:

$$\begin{cases} \min\limits_{A,E} \|A\|_* + \lambda\|E\|_1 \\ \text{s.t.} \quad M = A + E, \end{cases} \tag{2}$$

where $\|\cdot\|_*$ and $\|\cdot\|_1$ are the kernel norm and 1 norm of the matrix, respectively. If the singular vector of A is not related to the standard base, when $\lambda = n^{-1/2}$, the convex problem of (2) can be better solved [30]. Subsequently, λ was modified in [30] and it proves that RPCA is still robust when the noise is not so sparse. This paper uses the formula given in [31] as follows:

$$\lambda = C_1\left(4\sqrt{1-\rho} + \frac{9}{4}\right)^{-1}\sqrt{\frac{1-\rho}{\rho n}}, \tag{3}$$

where $C_1 > 0$ is a constant and $0 < \rho < 1$ is a sparsity parameter, which is the ratio of the number of nonzero entries of the noise matrix E to the total number in the original matrix.

The solution method of the RPCA algorithm mainly includes the following: the iterative thresholding approach (ITA), the accelerated proximal gradient approach (APGA), the dual approach (DA), and augmented Lagrangian multipliers (ALM) [32]. This paper uses the inexact augmented Lagrangian multiplier method (Inexact ALM) that was proposed in [32]. Its main formula is as follows.

First, construct augmented Lagrangian function for (2):

$$\Gamma(A, E, Y, \mu) = \|A\|_* + \lambda\|E\|_1 + \langle Y, M - L - S\rangle + \frac{\mu}{2}\|M - L - S\|_F^2, \tag{4}$$

where $Y \in R^{m\times n}$ denotes the Lagrange multiplier, $\mu > 0$ denotes a positive scalar, and $\langle\cdot\rangle$ denotes the standard inner product.

Then, iterate according to the following formula:

$$\begin{cases} A_{k+1} = \arg\min_{A} L\left(A, E_{k+1}, Y_k, \mu_k\right) = M_{1/\mu_k}\left(M - E_{k+1} + \dfrac{Y_k}{\mu_k}\right), \\ E_{k+1} = \arg\min_{E} L\left(A_{k+1}, E, Y_k, \mu_k\right) = M_{1/\mu_k}\left(M - A_{k+1} + \dfrac{Y_k}{\mu_k}\right). \end{cases} \tag{5}$$

Random forest (RF) is one of the classifier systems, and many theoretical and practical studies have proved that RF has high classification accuracy [33, 34]. Many scholars have improved it. For example, the fuzzy forest (FF) [35] algorithm combines the robustness of the classifier system and the flexibility of fuzzy logic theory to process data, and it has good classification accuracy in the absence of data. In the rotation forest (RoF) [36] algorithm, the feature space is divided into subspaces and the most important features are extracted from each subset using principal component analysis. The process is repeated to obtain the most distinguishable training data set and the basic classification of the different feature subspaces. Moreover, the cost-sensitive RoF algorithm has been proposed to reduce the classification cost of rotating deep forests [37] and KPCA combined with RoF and applied to linear indivisible data classification has achieved good results [38].

For the question of how to search the optimal number of DTs in RF. A small number of DTs optimal ensembles can be found exhaustively, but the burden of exponential complexity of such search limits its practical applicability for larger systems. Some methods have been proposed, such as heuristic forward search (FS), backward search (BS) [39], and genetic algorithm (GA) [40]. FS starts with a single classifier in each iteration, looking for a pair of classifiers to minimize majority voting errors. If majority voting error cannot be reduced for any pair of classifiers the algorithm stops with the combination built so far. BS represents a symmetrical to FS approach to classifier selection. GA as an effective evolutionary optimization algorithm bringing lots of applications in the machine learning domain [41, 42]. However, these algorithms have certain limitations, FS and BS are often reported to get caught in local maxima [39] and GA is more dependent on the choice of initial population, using a good initial population usually yields good results [43].

From these viewpoints, this study proposes an improved RF, the multigroup random walk random forest (MRWRF). The algorithm can select a better number of DTs in each training stage to achieve better classification results. Moreover, RPCA is introduced into the data classification to solve the case where there may be different degrees of data noise in the actual environment, which may not necessarily obey a Gaussian distribution. The RPCA is first used to recover the noise-contaminated data, and then the restored data is input into the MRWRF for data classification. Finally, a model example is used to prove the effectiveness of the method.

The rest of the paper is organized as follows. Section 3 presents the improved algorithm and the data classification steps of the RPCA-MRWRF algorithm. Then, Section 4 and Section 5 present the proposed method which was tested and verified using a concrete aqueduct's structural damage data. Finally, Section 6 concludes the paper.

3. The Improved Algorithm

3.1. Random Forest Based on the Thought of Random Walk (RWRF). RF is a classifier with multiple DTs, and its basic thought is to extract T samples from the original training set using Bootstrap sampling. Then, T decision tree models are established for T samples and T classification results are obtained. Finally, according to the T classification results, each record is voted to determine its final classification. Its core function is

$$T(X) = \arg\max_{Y} \sum_{i=1}^{n} I\left(t_i(X) = Y\right), \tag{6}$$

where $T(X)$ denotes the combined classifier model, $t(X)$ is single DT classification model, Y is the output variable, and $I(\cdot)$ is the voting decision function.

RF can obtain the upper bound of the generalization error according to the law of large numbers, as shown in the following:

$$\lim_{k\longrightarrow\infty} PE^{*} = P_{X,Y}\left\{P\left(t(X) = Y\right) - \max_{j\neq Y} P\left(t(X) = j\right) < 0\right\}, \tag{7}$$

where k denotes the number of DTs. As the number of DTs increases, the generalization error of RF will gradually tend to the upper bound of the above formula, which shows that RF has a good ability to prevent overfitting.

Random walk (RW) [44] is one of the most basic processes in dynamics. It has the ability to globally patrol and can be widely used, such as in image segmentation [45], biology, and electron transport. Its core thought is that the individual moves from the current state to the next state is random, that is, there are same probability of reaching the other locations in the next step. When moving to the next position, the individual will use the current position as the starting point and repeat the above process. In 2-dimensional space, random walk can be explained by Figure 1.

As can be seen from Figure 1, the red dot represents the position of the individual at the current moment and it has the same transition probability to reach any other position in the plane at the next moment. When the next position is reached, the individual repeats the above process with the following position as the starting point. This unpredictable random walk gives the individual the ability to find the best advantage within the set area.

In fact, it is a nonlinear optimization problem that chooses the number of DTs that enables the random forest algorithm to achieve the best classification effect.

The optimization problem is denoted as follows:

$$\begin{aligned} &\text{Find} && k \in R \\ &\text{Subject to} && \max(P) \in [0, 1], \end{aligned} \tag{8}$$

where P denotes the classification accuracy.

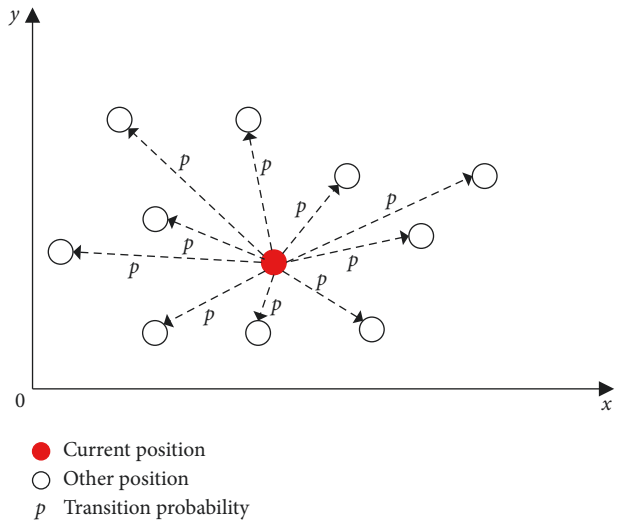


Figure 1: The progress of random walk.

In response to this problem, this paper proposes a random forest based on the thought of random walk. The specific algorithm steps are as follows:

Step 1: randomly select the initial number of DTs k_1. Set the initial walking step λ_1, the control precision ε, the iteration control times N, and the current number of iterations $i = 1$. Calculate the initial accuracy P_1, with equation (6).

Step 2: generate a random number u_i between −1 and 1. Calculate $k_{i+1} = k_i + [u_i\lambda_i]$ (where [] denotes the largest integer no more than $u\lambda$). Complete a step walk.

Step 3: calculate the value of P_{i+1}. If $P_{i+1} > P_i$, reset i to 1 and change P to P_0. Otherwise, return to step 2.

Step 4: if no better value can be found after N consecutive iterations, it is considered that the optimal solution is centered on the current optimal solution. At this point, if $\lambda_{i+1} < \varepsilon$, end the algorithm. Otherwise, set $\lambda = \alpha\lambda$, where $0 < \alpha < 1$ is the step reduction factor. Then, return to step 1 and start a new round of walks.

3.2. Multigroup Random Walk Random Forest (MRWRF). In practice, we find that although the random walk algorithm is simple to operate, its performance depends on the initial step size and the choice of the initial number of DTs. This phenomenon easily leads to the algorithm falling into the local optimum. Therefore, we propose the multigroup random walk random forest (MRWRF). First, establish three groups of the same level where each group is responsible for different areas, and then the individuals with different velocities randomly walk from different locations in each group. Update the best point of each area in each walk. If no better value is found over multiple successive iterations, then the best point of the area is considered to be the current point, and end the search. When the three groups find their respective best advantages, the three compare again and select the point with the highest accuracy as the optimal number of DTs in the random forest. Key steps of the MRWRF algorithm:

Step 1: initialize the MRWRF algorithm, create three groups, and create three individuals in each group. Set each individual's range of walk and randomly select the initial step size $\lambda^1_{mn}, m, n = 1, 2, 3$ and initial position of each individual $k^1_{mn}, m, n = 1, 2, 3$. λ^1_{mn} denotes the initial step size of the nth individual in the mth group, and k_{mn} denotes the initial position of the nth individual in the mth group. Forming the following matrix:

$$\lambda^1 = \begin{bmatrix} \lambda^1_{11} & \lambda^1_{12} & \lambda^1_{13} \\ \lambda^1_{21} & \lambda^1_{22} & \lambda^1_{23} \\ \lambda^1_{31} & \lambda^1_{32} & \lambda^1_{33} \end{bmatrix},$$
$$k^1 = \begin{bmatrix} k^1_{11} & k^1_{12} & k^1_{13} \\ k^1_{21} & k^1_{22} & k^1_{23} \\ k^1_{31} & k^1_{32} & k^1_{33} \end{bmatrix}. \quad (9)$$

Set the iteration control times N, the control precision ε, and the current number of iterations $i = 1$.

Step 2: calculate the classification accuracy of each individual's initial point $P^1_{mn}, m, n = 1, 2, 3$, with equation (6). Generate random number $u^1_{mn}, m, n = 1, 2, 3$ between −1 and 1. Calculate $k^{i+1}_{mn} = k^i_{mn} + [u^i_{mn}\lambda^i_{mn}], i = 1, 2, 3, \ldots, N$. Complete one step walk.

Step 3: calculate the accuracy of the latest location P^{i+1}_{mn} with equation (6). If $P^{i+1}_{mn} > P^i_{mn}$, change k_{mn} to k^{i+1}_{mn} and change P_{mn} to P^{i+1}_{mn}. k_{mn} is the best point of the nth individual in the mth group, P_{mn} denotes the optimal accuracy of the nth individual in the mth group, and reset i to 1. Otherwise, return to step 2.

Step 4: if no better value can be found for N consecutive iterations, it is considered that the optimal solution is centered on the current optimal solution. At this point, if $\lambda < \varepsilon$, end the algorithm. Otherwise, let $\lambda^{i+1}_{mn} = \alpha\lambda^i_{mn}$, where $0 < \alpha < 1$ is the step reduction factor; return to step 1 and start a new round of walks.

Step 5: after three individuals in a group find the best point of their respective regions, the three compare and select the best point of the group. When the best points of the three groups are found, the groups are compared and the overall optimal value is selected. Output the value of the accuracy P and the number of DTs k, which ends the algorithm.

The difficulty of the algorithm is that when the individual randomly walks, the same position may be swept many times, which will reduce the efficiency of the algorithm. We trim the algorithm as follows. The individual will mark the current position and classification accuracy rate during each walk and will automatically skip the marked position in subsequent walks, thereby improving the computational efficiency of the algorithm.

3.3. Data Denoising and Classification Process of RPCA-MRWRF. RPCA-MRWRF is a hybrid learning algorithm that differs from traditional random forest algorithms in two aspects: (1) it preprocesses raw data using RPCA and (2) a better number of DTs can be selected to effectively improve the classification accuracy. The specific classification process of the RPCA-MRWRF algorithm is as follows.

As shown in Figure 2, first, the raw data contaminated by random noise is input into the RPCA algorithm for preprocessing to remove noise. Then, the processed data is input into MRWRF for classification, and the specific classification process is given in Section 3.2. Finally, the classification results of different data and the corresponding accuracy rate are output.

4. Numerical Simulation Case

Generally, large-scale building damage identification is mainly based on the monitoring data that are obtained by monitoring sensors that are embedded in the interior of the building and the test data that are obtained by routine and special tests. These data usually contain varying degrees of noise, which will affect the accuracy of the classification. Therefore, in this paper, a finite element model is established for an aqueduct and the relevant parameters are obtained. Then, random noise of different degrees and intensity is added to simulate actual environmental noise and used as raw data to test the performance of the RPCA-MRWRF algorithm.

An aqueduct in an empty tank overhaul state, the main body of which is a single-slot ribbed belt tie rod structure. The section size is 6.0 m ∗ 5.4 m, the length of a single span is 30 m, the thickness of a sidewall is 0.60 m, and a 2.0 m wide sidewalk board is at the top. Side ribs and bottom ribs are added to the aqueduct body with widths of 0.5 m and heights of 0.7 m and 0.9 m, respectively. A tie rod is placed on the top of the transverse wall. Its section size is 0.3 m ∗ 0.4 m. The rib spacing is 2.5 m. The density of the concrete material was set at 2550 kg/m^3, the elasticity modulus was set at 34.5 GPa, and the Poisson's ratio was set at 0.167. The analysis modeling was built using SOLID95 mechanics elements in ANSYS. The three-dimensional map and free meshing map of the aqueduct is shown in Figure 3.

Next, we simulate damage in the middle of the bottom plate. The crack width is 4 mm. The degree of aqueduct damage is expressed by η, which is the ratio of the depth of the crack to the thickness of the bottom plate. In this paper, η was 0, 5%, 10%, 20%, 30%, 40%, and 50%. In addition, four types of damage are defined: no damage (ND, $\eta = 0$), general damage (GD, $\eta = 5\%, 10\%$), heavier damage (HD, $\eta = 20\%, 30\%$), and serious damage (SD, $\eta = 40\%, 50\%$).

The input parameters consist of the first ten natural frequencies $w_i, i = 1, 2, \ldots, 10$, of the aqueduct before and after the damage is calculated by ANSYS, as shown in Table 1.

In actual measurement, it is necessary to measure the natural frequencies of the structure multiple times to reduce the influence of environmental noise. Therefore, we performed the following processing on the data set in this simulation. $\eta = 0$, which means that no damage was recorded 184 times. $\eta = 5\%$, 10%, 20%, 30%, 40%, and 50% of the data were recorded 92 times, respectively, forming a $736 * 10$-dimensional data set. Non-Gaussian random noise $[-\sigma, \sigma]$ of different degrees and intensity is added to it. $\Delta = \sigma/w_i, i = 1, 2, \ldots, 10$, is used to represent the strength of the random noise, and w_i is the input parameter when $\eta = 0$. ρ in equation (3) is used to indicate the degree of random noise. In this paper, Δ is 0, 0.1%, 1%, 2%, 5%, 10%, 20%, and 30% and ρ is 0.5 and 0.6, respectively.

5. Algorithm Performance Verification and Result Analysis

In this section, we mainly test the classification performance of the RPCA-MRWRF algorithm and classify noise-contaminated aqueduct modal data. The main contents include the following: (1) discussing the influence of the number of DTs in RF on its classification performance and comparing the ability of multigroup random walk (MRW), FS, BS, and GA algorithm to find the better number of DTs in RF, (2) comparing the classification performance of the RPCA-MRWRF algorithm with RF algorithm under different degrees of random noise, and (3) comparing the classification performance of the RPCA-MRWRF algorithm with other existing classification algorithms under random noise with the same degree and intensity.

5.1. Discussing the Influence of the Number of DTs in RF and MRW Performance Verification. In this paper, 586 groups were randomly selected from the 736 sets of data as the training set and the remaining 150 sets of data were used as the test sets. The effect of the number of DTs on the RF's performance is tested when Δ is 30% and ρ is 0.5. As shown in Figure 4, the RF's performance is best when the number of DTs is 98, but the RF's performance is the worst when the number is 140. When the number of DTs increases to a certain extent, the classification accuracy of the RF is basically unchanged. Therefore, it can be stated that the number of DTs has an impact on the RF's performance.

The better number of DTs found by MRW, FS, BS, and GA and the corresponding accuracy rates are shown in the Tables 2 and 3. It is worth mentioning that these are performed under the conditions of the RPCA processing the raw noise data.

When ρ is 0.5, the number of DTs sought by MRW corresponds to the highest accuracy rate except in the case of $\Delta = 2\%$, and when ρ is 0.6, the number of DTs sought by MRW corresponds to the highest accuracy rate except in the case of $\Delta = 10\%$ and $\Delta = 30\%$. The better number of DTs searched by FS is less than MRW, and the better number of DTs searched by BS is more than MRW. There is a similar situation when ρ is 0.6. A likely reason is that the two methods are relatively easy to fall into local optimum [39]. GA has better performance than FS and BS, and it is not easy to fall into local optimum, but it has poor performance compared with MRW. The reason for this phenomenon may be that GA is more dependent on the choice of initial values,

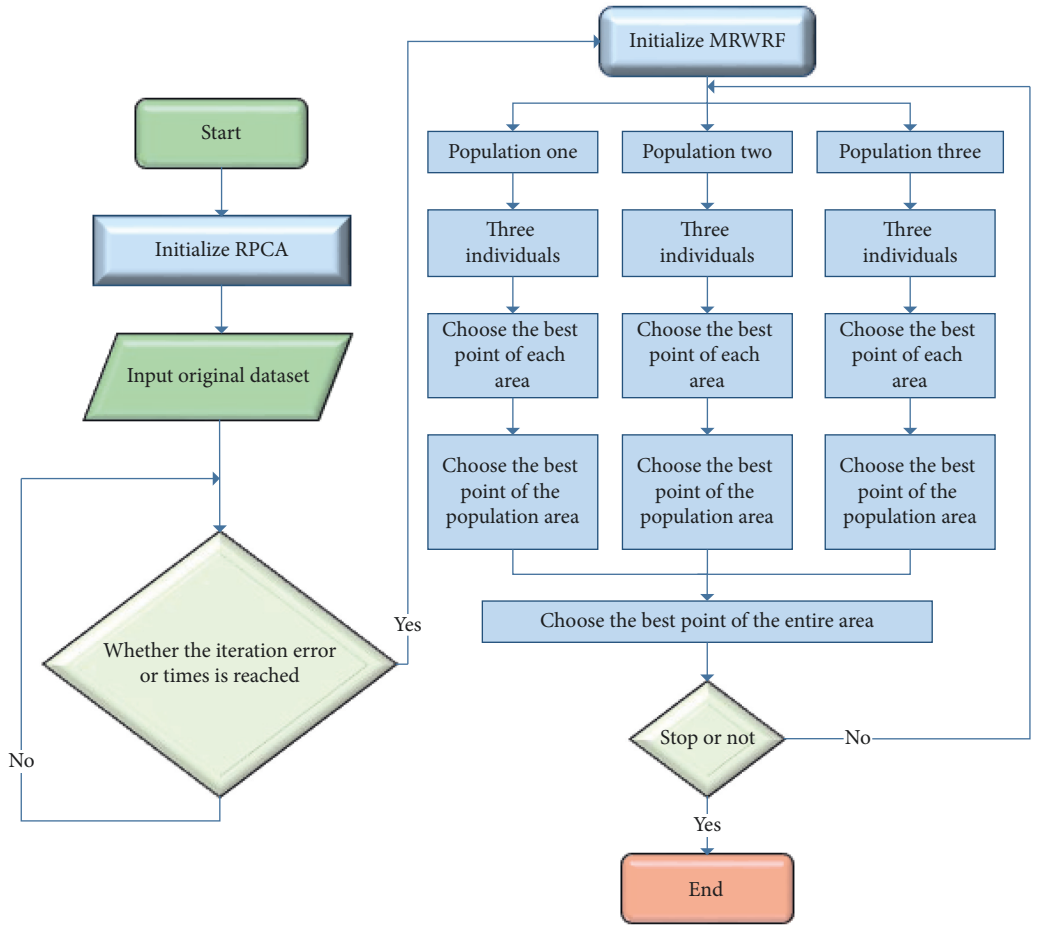
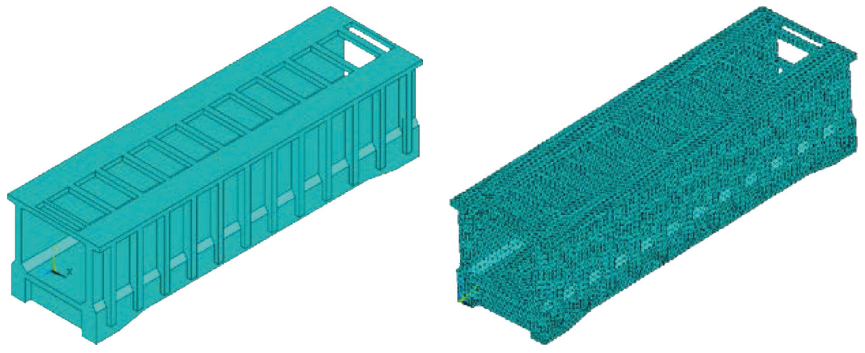


Figure 2: The data noise reduction and classification process of the RPCA-MRWRF.



(a) (b)

Figure 3: (a) The three-dimensional map of the aqueduct and (b) the free meshing map of the aqueduct.

and it is difficult to find the optimal or near-optimal number of DTs in a limited iteration [46].

The above analysis can show that MRW has better performance than FS, BS, and GA. Since the test set and the training set are randomly selected during each classification process, the better number of preferred DTs will be different each time. However, it is always able to find a better number of DTs to improve the classification accuracy of the data in each run.

5.2. Performance Comparison between RPCA-MRWRF and RF. This paper compares the classification performance of the RPCA-MRWRF algorithm and the RF algorithm when ρ is 0.5 and 0.6, respectively. The following indicators are used to evaluate the performance of the algorithm:(1) Classification Precision, (2) Confusion Matrix, and (3) Cohen's Kappa Coefficient.

As can be seen from Figure 5, RPCA can effectively reduce the impact of noise on the data but cannot completely

Table 1: Input parameters.

Parameter	$\eta = 0$	$\eta = 5\%$	$\eta = 10\%$	$\eta = 20\%$	$\eta = 30\%$	$\eta = 40\%$	$\eta = 50\%$
w_1	11.055	11.05	11.052	11.051	11.05	11.05	11.05
w_2	14.75	14.741	14.746	14.744	14.742	14.745	14.742
w_3	18.548	18.538	18.542	18.542	18.539	18.542	18.539
w_4	18.85	18.854	18.857	18.855	18.85	18.853	18.855
w_5	20.395	20.39	20.392	20.391	20.387	20.389	20.391
w_6	27.24	27.231	27.233	27.229	27.23	27.227	27.225
w_7	27.581	27.566	27.569	27.566	27.563	27.564	27.566
w_8	28.722	28.722	28.725	28.724	28.721	28.721	28.727
w_9	40.424	40.412	40.417	40.412	40.412	40.415	40.417
w_{10}	42.097	42.074	42.072	42.07	42.055	42.045	42.027

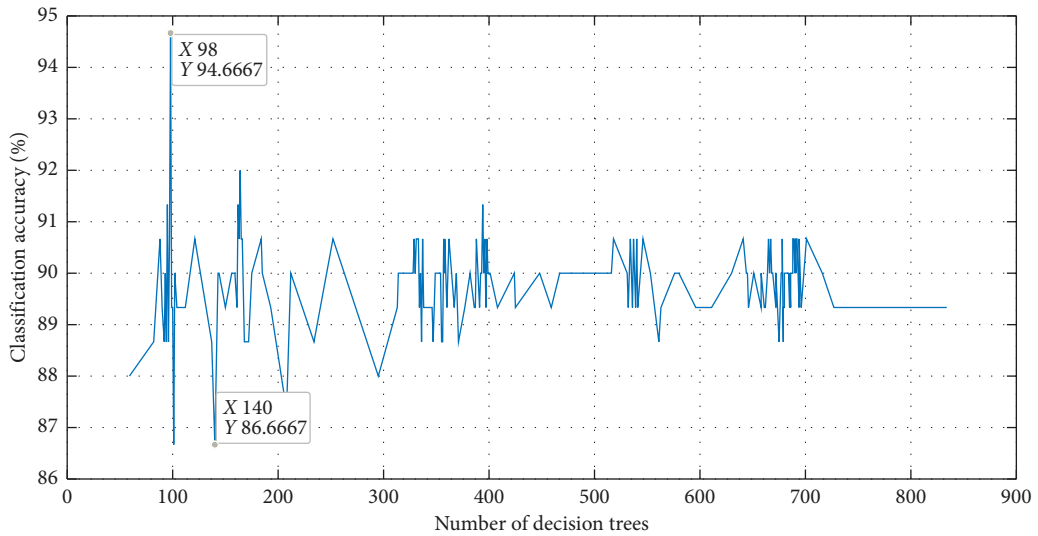


Figure 4: The influence of the number of DTs on the RF's performance.

Table 2: The better number of DTs found by MRW, FS, BS, and GA and the corresponding accuracy rates when ρ is 0.5.

Δ	MRW		FS		BS		GA	
	DTs	Accuracy (%)	DTs	Accuracy (%)	DTs	Accuracy (%)	DTs	Accuracy (%)
0.1%	73	**98**	73	**98**	799	90.67	73	**98**
1%	93	**93.33**	69	88.67	535	85	151	91.33
2%	524	94	135	91.33	481	95.33	163	**96**
5%	172	**92.67**	91	86.67	267	90	97	90.67
10%	204	**94.67**	109	90.67	247	91	173	93.33
20%	206	**94.67**	55	82	385	89.33	537	88.67
30%	98	**94.67**	91	90.67	163	88	133	90.67

Table 3: The better number of DTs found by MRW, FS, BS, and GA and the corresponding accuracy rates when ρ is 0.6.

Δ	MRW		FS		BS		GA	
	DTs	Accuracy (%)	DTs	Accuracy (%)	DTs	Accuracy (%)	DTs	Accuracy (%)
0.1%	93	**88**	67	86	269	81.3	93	**88**
1%	195	**86.67**	95	83.33	395	79.33	263	84.67
2%	243	**84**	79	81.33	243	84	89	80.67
5%	166	**84.67**	153	84	197	82.67	684	77
10%	293	83.33	53	79.33	325	81.33	159	**84**
20%	97	**85.33**	97	**85.33**	651	75.33	463	81.33
30%	779	84	35	78.67	779	84	211	**85.33**

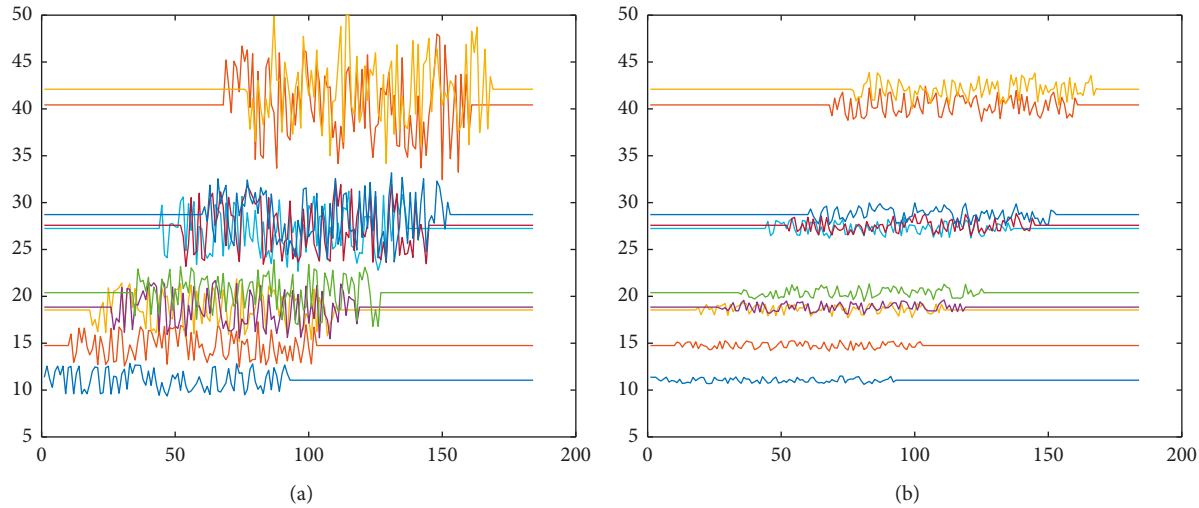


Figure 5: The denoising effect of RPCA of when ρ is 0.5 and Δ is 20%. (a) The data before RPCA processing and (b) the data after RPCA processing.

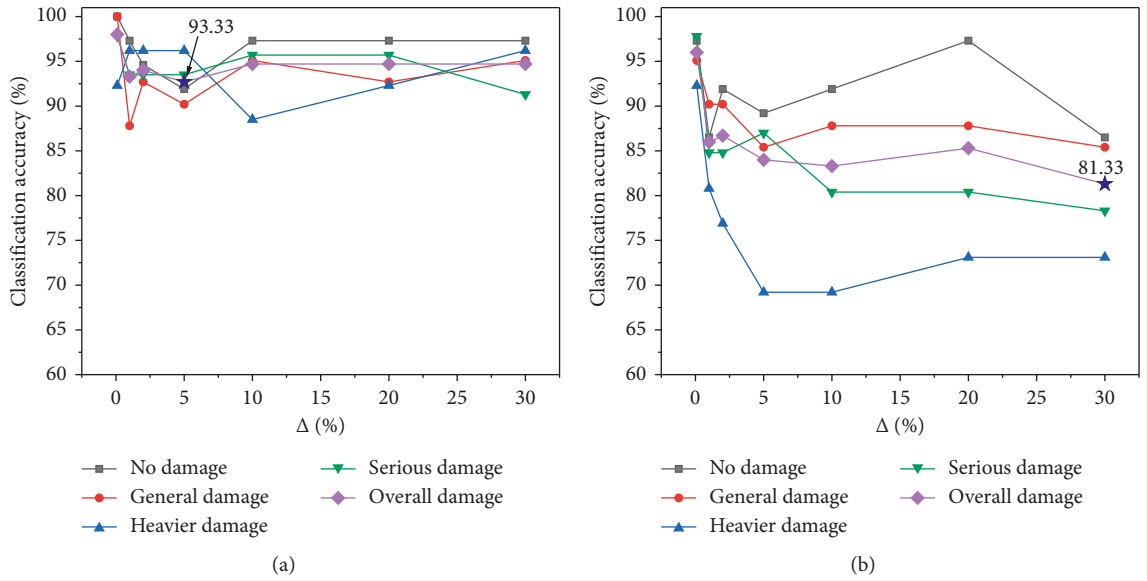


Figure 6: The classification accuracy rates of (a) RPCA-MRWRF and (b) RF when ρ is 0.5.

eliminate it. Classification precision results are shown in Figures 6 and 7. When ρ is 0.5, the overall accuracy of the RPCA-MTWRF is above 93%, the lowest is 93.3%, and the accuracy of various types of damage is also above 85%. However, the classification accuracy of the RF is slightly different. When Δ is small, its classification accuracy is higher, but as Δ increases, the accuracy decreases rapidly. The overall recognition rate is basically below 85% except in the case of lower noise intensity ($\Delta = 0.1\%$ or 1%), and the lowest is 81.33%. When ρ is 0.6, the overall recognition rate of RPCA-MRWRF is reduced, which is basically approximately 85%, and the lowest is 84%. However, the overall recognition rate of the RF is below 78% and the lowest is 68%.

The confusion matrix of RPCA-MRWRF and RF are shown in Figures 8–11. Most of the classification results of the two algorithms are distributed in the diagonal area of the confusion matrix, indicating that they all have certain classification ability. In the ideal state ($\Delta = 0\%$), both RPCA-MRWRF and RF have 100% accuracy. As ρ and Δ increase, the number of RF classification errors increases significantly, but RPCA-MRWRF does not change much.

According to the confusion matrix of RPCA-MRWRF and RF, the Kappa coefficient under different conditions are calculated, as shown in Figure 12. The results show that the Kappa value of RPCA-MRWRF is always larger than RF under different levels of noise.

The above analysis shows that the RPCA-MRWRF algorithm has better noise immunity and better classification performance than RF.

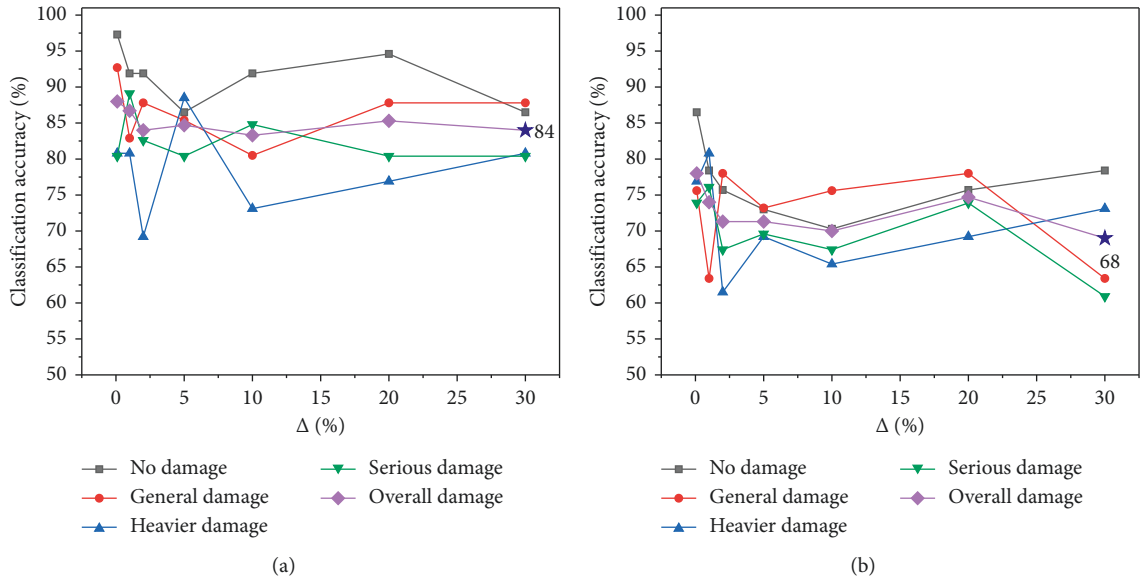


Figure 7: The classification accuracy rates of (a) RPCA-MRWRF and (b) RF when ρ is 0.6.

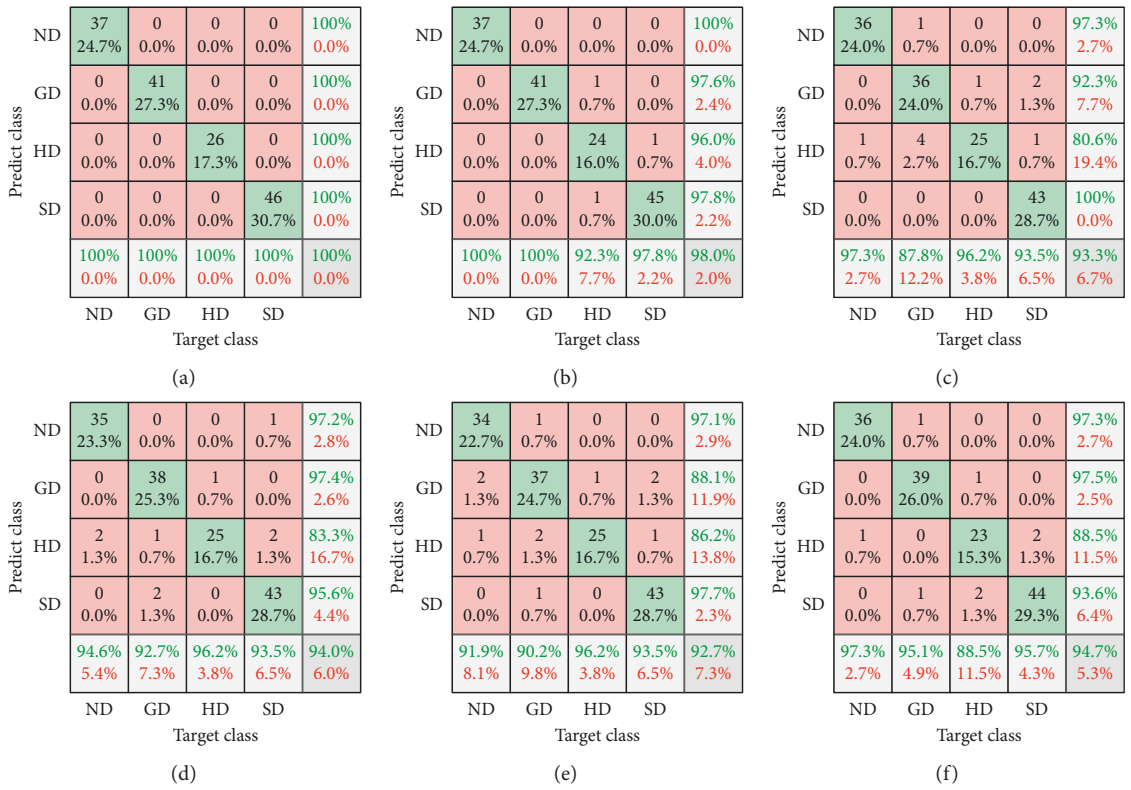


Figure 8: Continued.

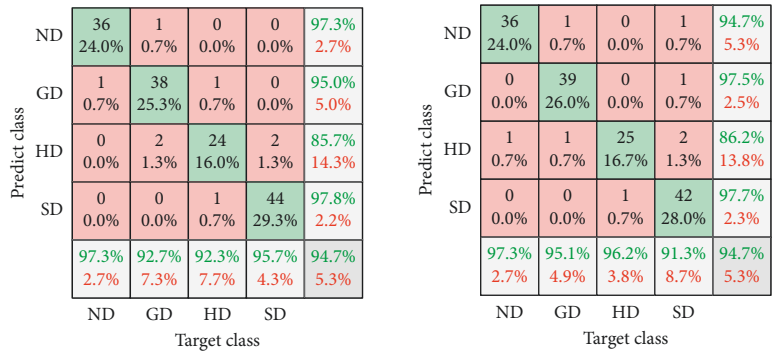


Figure 8: When ρ is 0.5, (a), (b), (c), (d), (e), (f), (g), and (h) are the confusion matrices of RPCA-MRWRF for Δ is 0, 0.1%, 1%, 2%, 5%, 10%, 20%, and 30%, respectively, where ND denotes no damage, GD denotes general damage, HD denotes heavier damage, and SD denotes serious damage.

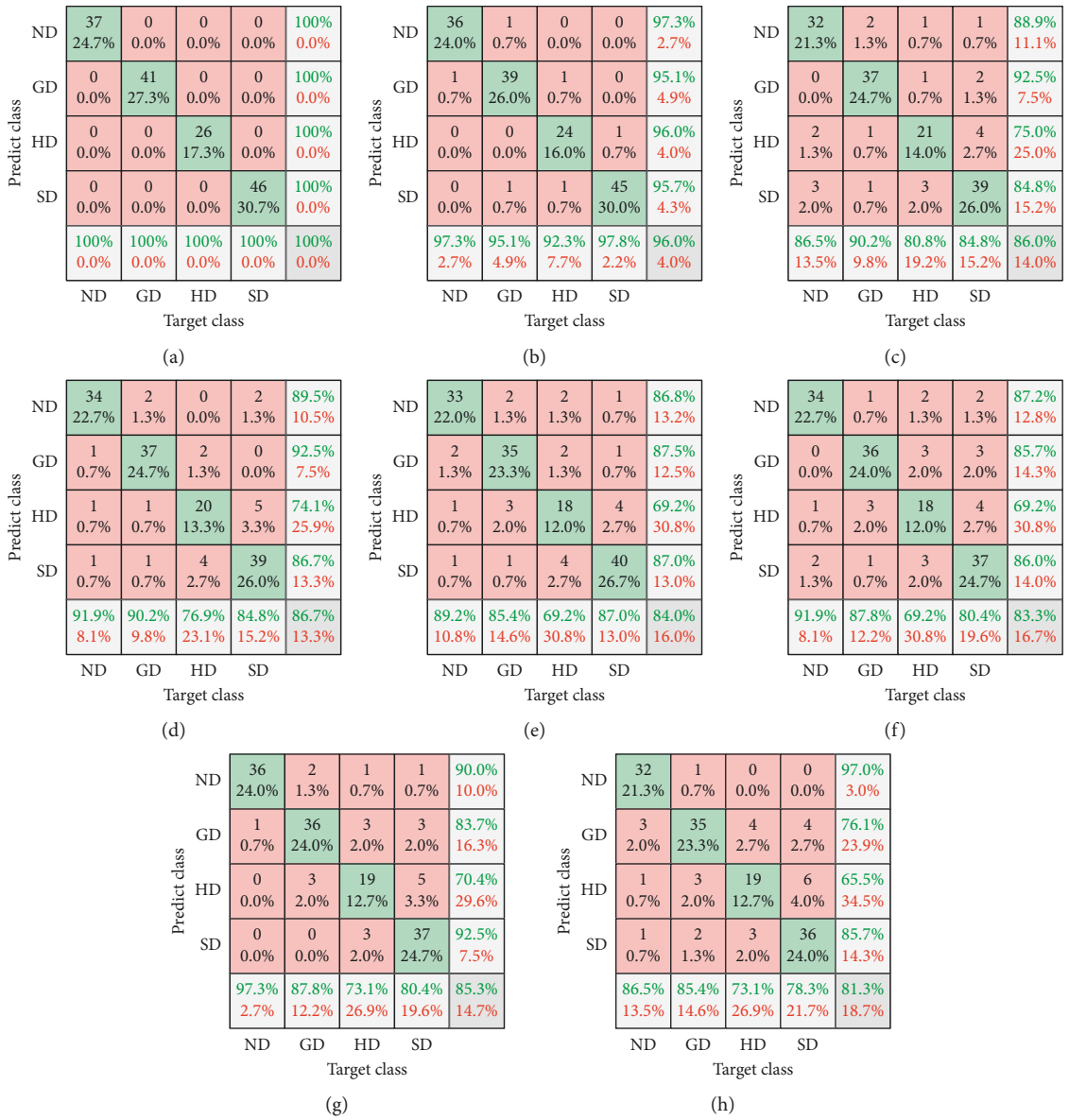


Figure 9: When ρ is 0.5, (a), (b), (c), (d), (e), (f), (g), and (h) are the confusion matrices of RF for Δ is 0, 0.1%, 1%, 2%, 5%, 10%, 20%, and 30%, respectively, where ND denotes no damage, GD denotes general damage, HD denotes heavier damage, and SD denotes serious damage.

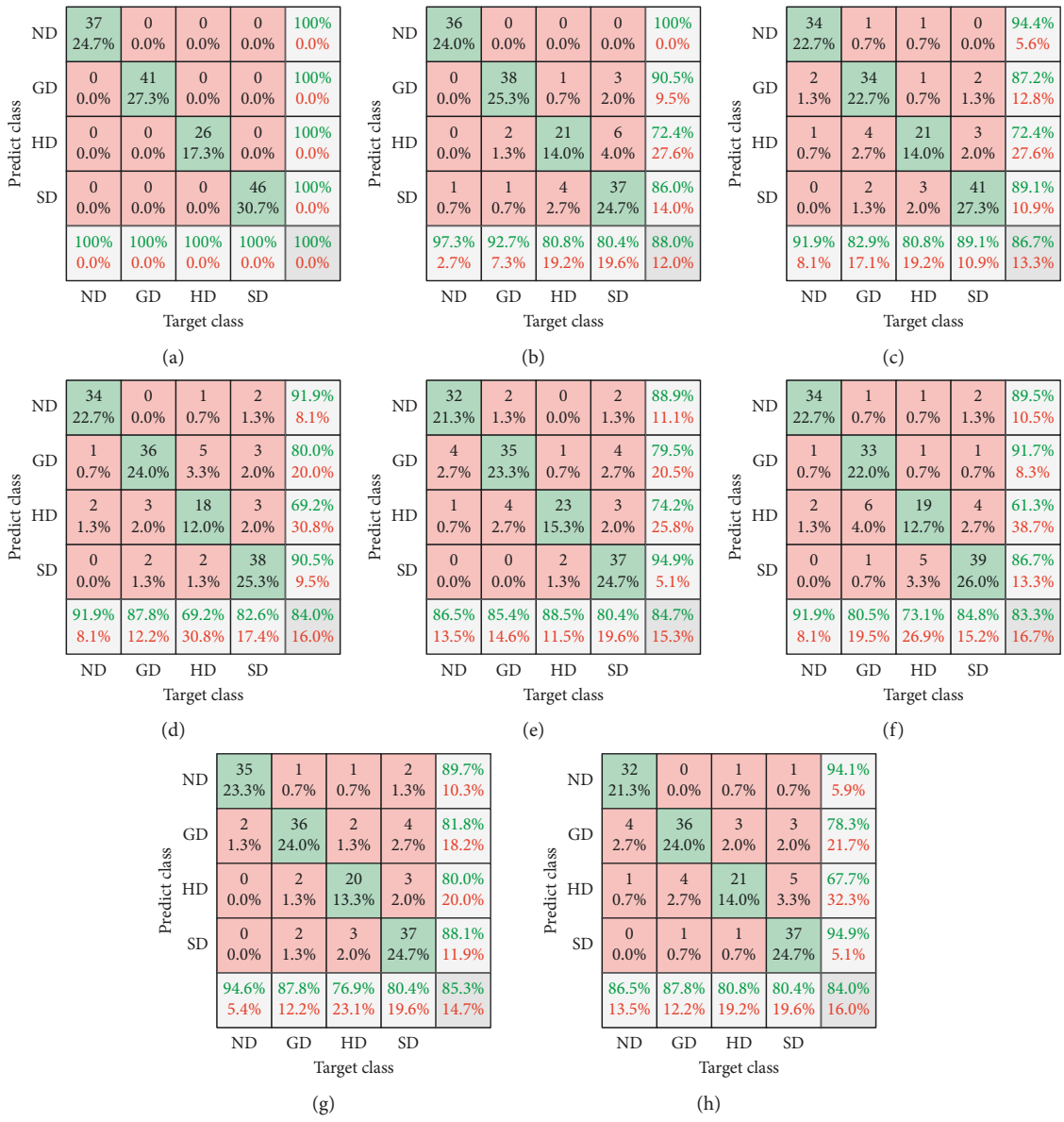


Figure 10: When ρ is 0.6, (a), (b), (c), (d), (e), (f), (g), and (h) are the confusion matrices of RPCA-MRWRF for Δ is 0, 0.1%, 1%, 2%, 5%, 10%, 20%, and 30%, respectively, where ND denotes no damage, GD denotes general damage, HD denotes heavier damage, and SD denotes serious damage.

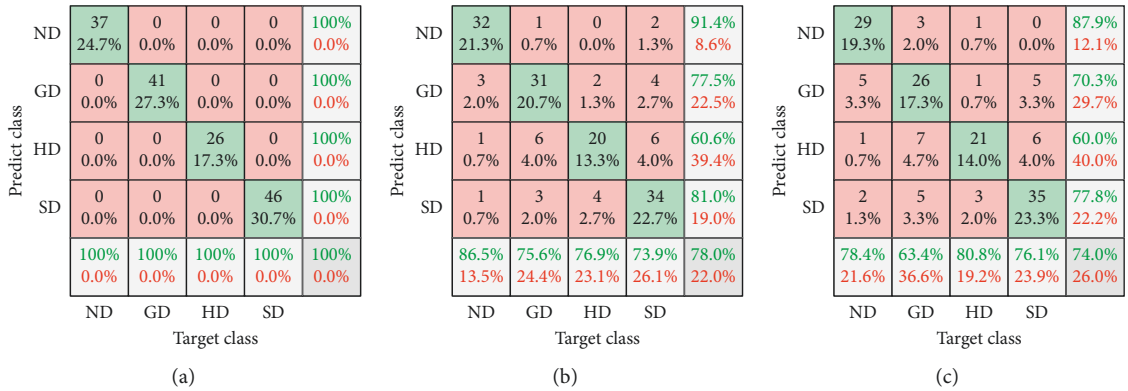


Figure 11: Continued.

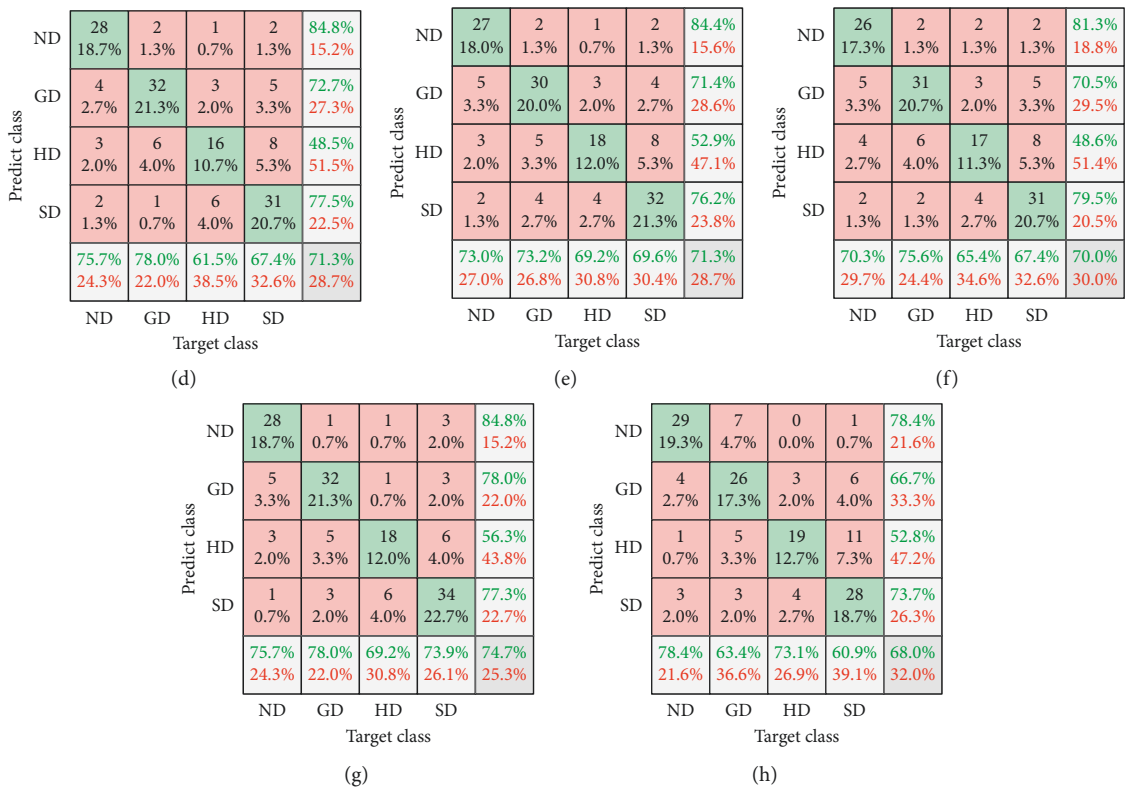


Figure 11: When ρ is 0.6, (a), (b), (c), (d), (e), (f), (g), and (h) are the confusion matrices of RPCA-MRWRF for Δ is 0, 0.1%, 1%, 2%, 5%, 10%, 20%, and 30%, respectively, where ND denotes no damage, GD denotes general damage, HD denotes heavier damage, and SD denotes serious damage.

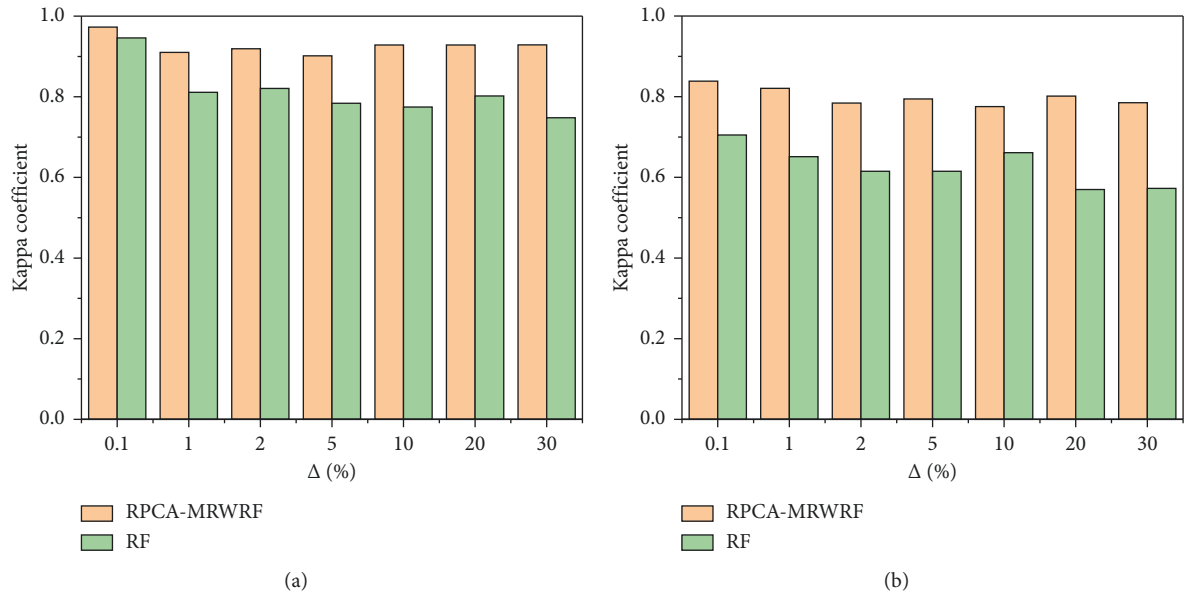


Figure 12: (a) The Kappa coefficient when ρ is 0.5 and (b) the Kappa coefficient when ρ is 0.6.

5.3. The Performance of Other Existing Algorithms. In addition, we tested the performance of other existing algorithms when ρ is 0.5, including the following: multiwavelet and mutation particle swarm optimization algorithm (MW-MPSO) [5], the KPCA-SVM [14], the wavelet packet transform (WPT) analysis and fully connected deep neural

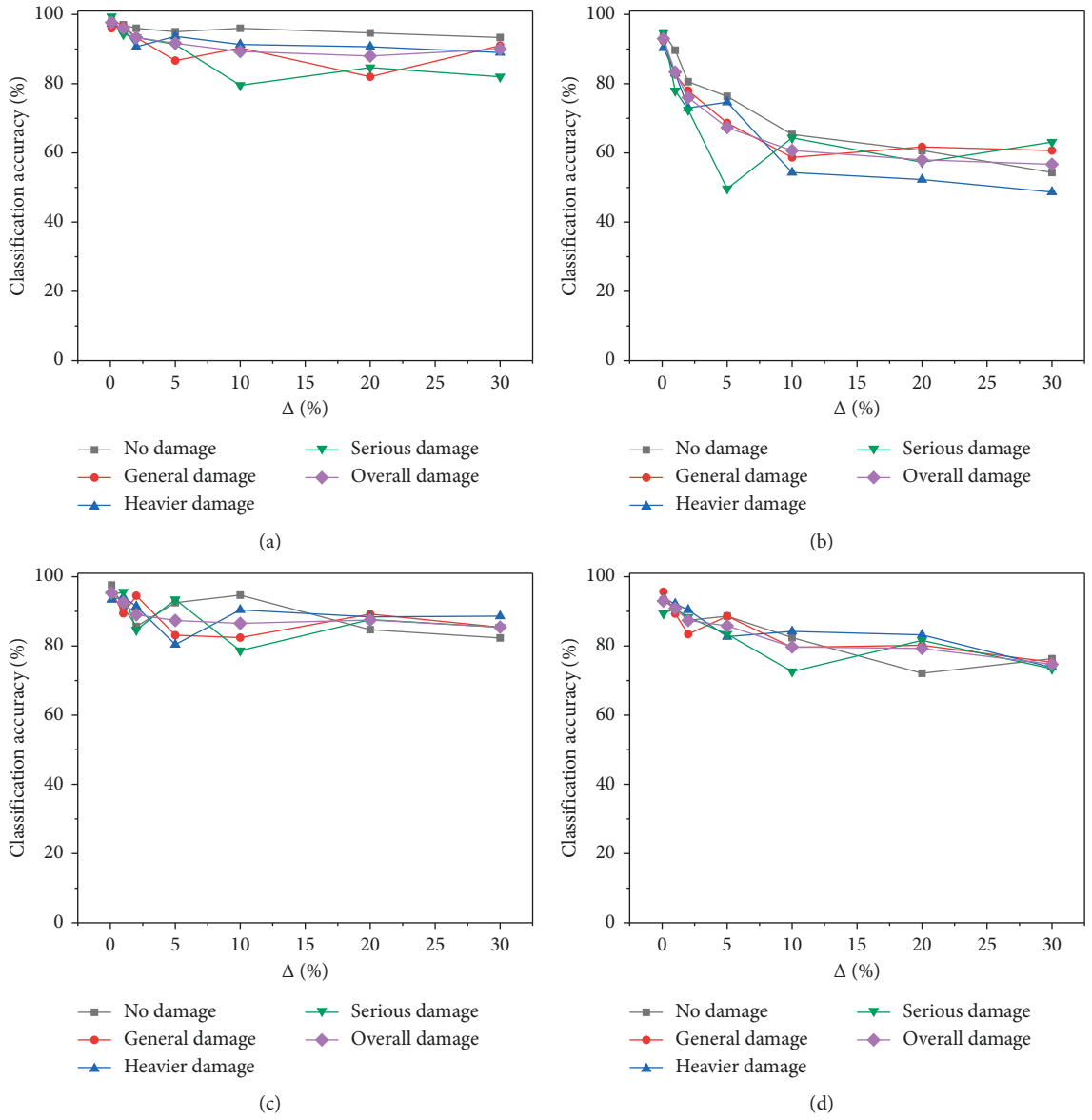


Figure 13: (a), (b), (c), and (d) are the classification accuracy rates of MW-MPSO, KPCA-SVM, WPT-FCDNN, and PCA-LDA-NN when ρ is 0.5, respectively.

network (WPT-FCDNN) [19], and the PCA-LDA-NN [23]. The classification accuracy of each algorithm is shown in Figure 13 and Table 4.

In the ideal state ($\Delta = 0\%$), several algorithms all have 100% accuracy. As the noise level increases, although the accuracy of several algorithms is decreasing, the accuracy of several methods is maintained above 80% except KPCA-SVM and PCA-LDA-NN. When Δ is small, KPCA-SVM has good performance, but when Δ is larger than 5%, the accuracy of KPCA-SVM is less than 70%. The accuracy of PCA-LDA-NN is lower than 80%, and it is maintained at around 75% when $\Delta u\lambda$ is greater than 20%. This may be due to the limited noise reduction capability of the PCA-LDA and the lack of robustness of the corresponding classifier. The lowest accuracy of RF, MW-MPSO, and WPT-FCDNN is more than 80% under different degrees of non-Gaussian noise pollution, which indicates that they all have certain antinoise ability. RF benefits from the integration of multiple decision trees to reduce the risk of misjudgment. The performance of MW-MPSO and WPT-FCDNN may be attributed to the favourable denoising capability and the preponderance of robustness of the corresponding classifier. Meanwhile, it can be distinctly observed that the accuracy of the developed algorithm is more than 93%, which is the most excellent in existing algorithms.

It is feasible to apply the RF, MW-MPSO, and WPT-FCDNN algorithm for the case when there is not a strict requirement for the classification accuracy and no serious environmental noise. However, the RPCA-MRWRF, the proposed algorithm in this paper, is the best choice for the case when there is a more stringent requirement for classification accuracy and more serious environmental noise. In

Table 4: Comparison of the overall recognition rates of several classification algorithms when ρ is 0.5.

Classification algorithm	$\Delta = 0$ Rate (%)	$\Delta = 0.1\%$ Rate (%)	$\Delta = 1\%$ Rate (%)	$\Delta = 2\%$ Rate (%)	$\Delta = 5\%$ Rate (%)	$\Delta = 10\%$ Rate (%)	$\Delta = 20\%$ Rate (%)	$\Delta = 30\%$ Rate (%)
RPCA-MRWRF	100	98	93.33	94	93.33	94.67	94.67	94.67
RF	100	96	86	86.67	84	83.33	85.33	81.33
MW-MPSO	100	97.67	96	93.33	91.67	89.3	88	90
KPCA-SVM	100	93	83.33	76	67.33	60.67	58	56.7
WPT-FCDNN	100	95.33	92.7	89.33	87	86.7	87.33	84.67
PCA-LDA-NN	100	93	90.7	87.33	85.33	78.7	78	74.67

general, the RPCA-MRWRF algorithm has excellent anti-noise ability and can provide a new method for the identification of aqueduct structural damage and other engineering classification problems.

6. Conclusion

In this paper, a hybrid algorithm of RPCA-MRWRF is proposed. RPCA can reduce the impact of noise on the original data as much as possible, and MRWRF can choose the better DTS number to improve the classification accuracy. Experimental results show that RPCA-MRWRF has better classification performance than other existing classification algorithms and can thus provide a new approach for data classification problems in engineering.

However, there are still many shortcomings in this paper. For example, the time consumption of the RPCA-MRWRF algorithm is slightly longer than that of other types of algorithms, and it will be discussed as key points in subsequent research.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This article was funded by The National Key R&D Program of China (Grant number: 2018YFC0406902).

References

[1] S. M. Kamruzzaman and A. M. J. Sarkar, "A new data mining scheme using artificial neural networks," *Sensors*, vol. 11, no. 5, pp. 4622–4647, 2011.

[2] M. Hongjin and N. Yufeng, "Mixed noise removal algorithm combining adaptive directional weighted mean filter and improved adaptive anisotropic diffusion model," *Mathematical Problems in Engineering*, vol. 2018, Article ID 6492696, 19 pages, 2018.

[3] J. Chai, J. Du, K. K. Lai, and Y. P. Lee, "A hybrid least square support vector machine model with parameters optimization for stock forecasting," *Mathematical Problems in Engineering*, vol. 2015, Article ID 231394, 7 pages, 2015.

[4] B. Sornkitja, C. Chunxiang, C. Wei, X. Ni, M. Xu, and B. Acharya, "The classification of noise-afflicted remotely sensed data using three machine-learning techniques: effect of different levels and types of noise on accuracy," *ISPRS International Journal of Geo-Information*, vol. 7, no. 7, p. 274, 2018.

[5] H. Sun, K. Li, H. Wang, P. Chen, and Y. Cao, "Intelligent mechanical fault diagnosis based on multiwavelet adaptive threshold denoising and MPSO," *Mathematical Problems in Engineering*, vol. 2014, Article ID 142795, 15 pages, 2014.

[6] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[7] Y. Yang and T. Y. Liu, "Application of support vector machine in structure damage identification," in *Proceedings of the 2009 International Conference on Information Engineering and Computer Science*, Wuhan, China, December 2009.

[8] I. Pisa, I. Santín, J. Vicario, A. Morell, and R. Vilanova, "ANN-based soft sensor to predict effluent violations in wastewater treatment plants," *Sensors*, vol. 19, no. 6, p. 1280, 2019.

[9] Z. X. Li and X. M. Yang, "Damage identification for beams using ANN based on statistical property of structural responses," *Computers & Structures*, vol. 86, no. 1-2, pp. 64–71, 2008.

[10] J. Quan, L. Ruan, Z. Xie, X. Li, and X. Lin, "The application of Bayesian network theory in transformer condition assessment," in *Proceedings of the 2013 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, Kowloon, Hong Kong, December 2013.

[11] A. E. Hegazy, M. A. Makhlouf, and G. S. El-Tawel, "Feature selection using chaotic salp swarm algorithm for data classification," *Arabian Journal for Science and Engineering*, vol. 44, no. 4, pp. 3801–3816, 2019.

[12] S. Taneja, B. Suri, S. Gupta, H. Narwal, A. Jain, and A. Kathuria, "A fuzzy logic based approach for data classification," in *Advances in Intelligent Systems and Computing*, pp. 605–616, Springer, Singapore, 2018.

[13] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: discriminative nearest neighbor classification for visual category recognition," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Volume 2 (CVPR'06)*, pp. 2126–2136, New York, NY, USA, June 2006.

[14] S. Yanli, Y. Na, Z. Zhengtao et al., "Structural damage identification study based on kernel component analysis and support vector machine," *Journal of Basic Science and Engineering*, vol. 26, pp. 888–900, 2018.

[15] L. He, J. Lian, and B. Ma, "Intelligent damage identification method for large structures based on strain modal parameters," *Journal of Vibration and Control*, vol. 20, no. 12, pp. 1783–1795, 2014.

[16] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2009.

[17] H.-P. Müller, G. Nolte, and S. N. Erné, "Using independent component analysis for noise reduction of magnetocardiographic data in case of exercise with an ergometer," *Journal of Medical Engineering & Technology*, vol. 30, no. 3, pp. 158–165, 2006.

[18] A. Stateczny and M. Wlodarczyk-Sielicka, "Self-organizing artificial neural networks into hydrographic big data reduction process," *Rough Sets and Intelligent Systems Paradigms*, Lecture Notes in Computer Science, vol. 8537, no. 6, pp. 335–342, Springer, Berlin, Germany, 2014.

[19] M. Sohaib and J.-M. Kim, "Data driven leakage detection and classification of a boiler tube," *Applied Sciences*, vol. 9, no. 12, p. 2450, 2019.

[20] H. Ma, Z. Qian, Y. Li, H. Lin, D. Shao, and B. Yang, "Noise reduction for desert seismic data using spectral kurtosis adaptive bandpass filter," *Acta Geophysica*, vol. 67, no. 1, pp. 123–131, 2018.

[21] R. Ostori, A. Goudarzi, and B. Oskoi, "GPR random noise reduction using BPD and EMD," *Journal of Geophysics and Engineering*, vol. 15, no. 2, pp. 347–353, 2018.

[22] W. Liu, S. Cao, and Z. Wang, "Application of variational mode decomposition to seismic random noise reduction," *Journal of Geophysics and Engineering*, vol. 14, no. 4, pp. 888–899, 2017.

[23] S.-I. Choi, T. Eom, and G.-M. Jeong, "Gas classification using combined features based on a discriminant analysis for an electronic nose," *Journal of Sensors*, vol. 2016, Article ID 9634387, 9 pages, 2016.

[24] E. Candés, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?: recovering low-rank matrices from sparse errors," in *Proceedings of the 2010 IEEE Sensor Array and Multichannel Signal Processing Workshop*, Jerusalem, Israel, October 2010.

[25] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," 2009, https://arxiv.org/abs/0912.3599.

[26] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3047–3064, 2012.

[27] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma, "TILT: transform invariant low-rank textures," *International Journal of Computer Vision*, vol. 99, no. 1, pp. 1–24, 2012.

[28] K. Min, Z. Zhang, J. Wright, and Y. Ma, "Decomposing background topics from keywords by principal component pursuit," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management—CIKM'10*, Toronto, Canada, October 2010.

[29] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proceedings of the 18th International Conference on Multimedia 2010*, Firenze, Italy, October 2010.

[30] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, pp. 1–37, 2011.

[31] A. Ganesh, J. Wright, X. Li, E. J. Candes, and Y. Ma, "Dense error correction for low-rank matrices via principal component pursuit," in *Proceedings of the 2010 IEEE International Symposium on Information Theory*, Austin, TX, USA, June 2010.

[32] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," vol. 1, 2010, https://arxiv.org/abs/1009.5055.

[33] X. Xue, Y. Dexin, and Z. Wei, "Data calibration based on multisensor using classification analysis: a random forests approach," *Mathematical Problems in Engineering*, vol. 2015, Article ID 708467, 8 pages, 2015.

[34] T. Han, D. Jiang, Q. Zhao, L. Wang, and K. Yin, "Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery," *Transactions of the Institute of Measurement and Control*, vol. 40, no. 8, pp. 2681–2693, 2018.

[35] P. Bonissone, J. M. Cadenas, M. Carmen Garrido, and R. Andrés Díaz-Valladares, "A fuzzy random forest," *International Journal of Approximate Reasoning*, vol. 51, no. 7, pp. 729–747, 2010.

[36] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: a new classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619–1630, 2006.

[37] H. Lu, L. Yang, K. Yan, Y. Xue, and Z. Gao, "A cost-sensitive rotation forest algorithm for gene expression data classification," *Neurocomputing*, vol. 228, pp. 270–276, 2017.

[38] H. Lu, Y. Meng, K. Yan, and Z. Gao, "Kernel principal component analysis combining rotation forest method for linearly inseparable data," *Cognitive Systems Research*, vol. 53, pp. 111–122, 2019.

[39] D. Ruta and B. Gabrys, "Classifier selection for majority voting," *Information Fusion*, vol. 6, no. 1, pp. 63–81, 2005.

[40] J. Wang, J. Kang, and G. Hou, "Real—time fault repair scheme based on improved genetic algorithm," *IEEE Access*, vol. 7, pp. 35805–35815, 2019.

[41] S.-B. Cho, "Pattern recognition with neural networks combined by genetic algorithms," *Fuzzy Seta and Systems*, vol. 103, no. 2, pp. 339–347, 1999.

[42] L. I. Kuneheva and L. C. Jain, "Designing classifier fusion systems by genetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 4, pp. 327–336, 2000.

[43] M. A. Rahman and M. Z. Islam, "A hybrid clustering technique combining a novel genetic algorithm with K-means," *Knowledge-Based Systems*, vol. 71, pp. 345–365, 2014.

[44] R. Metzler and J. Klafter, "The random walk's guide to anomalous diffusion: a fractional dynamics approach," *Physics Reports*, vol. 339, no. 1, pp. 1–77, 2000.

[45] C. Qin, G. Zhang, Y. Zhou, W. Tao, and Z. Cao, "Integration of the saliency-based seed extraction and random walks for image segmentation," *Neurocomputing*, vol. 129, pp. 378–391, 2014.

[46] D. Jiang and Z. Fan, "The algorithm for algorithms: an evolutionary algorithm based on automatic designing of genetic operators," *Mathematical Problems in Engineering*, vol. 2015, Article ID 474805, 15 pages, 2015.


Advances in
Operations Research
Advances in
Decision Sciences
Journal of
Applied Mathematics
The Scientific
World Journal
Journal of
Probability and Statistics
International
Journal of
Mathematics and
Mathematical
Sciences
Hindawi
Submit your manuscripts at
www.hindawi.com
Journal of
Optimization
International Journal of
Engineering
Mathematics
International Journal of
Analysis
Journal of
Complex Analysis
Advances in
Numerical Analysis
Mathematical Problems
in Engineering
International Journal of
Differential Equations
Discrete Dynamics in
Nature and Society
International Journal of
Stochastic Analysis
Journal of
Mathematics
Journal of
Function Spaces
Abstract and
Applied Analysis
Advances in
Mathematical Physics