


## Article

# Statistical Feature Extraction Combined with Generalized Discriminant Component Analysis Driven SVM for Fault Diagnosis of HVDC GIS

Ruixu Zhou <sup>1,\*</sup> , Wensheng Gao <sup>1</sup>, Weidong Liu <sup>1</sup>, Dengwei Ding <sup>2</sup> and Bowen Zhang <sup>3</sup>

<sup>1</sup> State Key Laboratory of Power System and Generation Equipment, Department of Electrical Engineering, Tsinghua University, Beijing 100084, China; wsgao@tsinghua.edu.cn (W.G.); lwd-dea@mail.tsinghua.edu.cn (W.L.)

<sup>2</sup> Sichuan Energy Internet Research Institute, Tsinghua University, Chengdu 610213, China; sunnyall123@163.com

<sup>3</sup> China Electric Power Research Institute, Beijing 100192, China; zbw1986@126.com

\* Correspondence: Ruixuzhou\_tsinghua@yeah.net

**Abstract:** Accurately identifying the types of insulation defects inside a gas-insulated switchgear (GIS) is of great significance for guiding maintenance work as well as ensuring the safe and stable operation of GIS. By building a set of 220 kV high-voltage direct current (HVDC) GIS experiment platforms and manufacturing four different types of insulation defects (including multiple sizes and positions), 180,828 pulse current signals under multiple voltage levels are successfully measured. Then, the apparent discharge quantity and the discharge time, two inherent physical quantities unaffected by the experimental platform and measurement system, are obtained after the pulse current signal is denoised, according to which 70 statistical features are extracted. In this paper, a pattern recognition method based on generalized discriminant component analysis driven support vector machine (SVM) is detailed and the corresponding selection criterion of involved parameters is established. The results show that the newly proposed pattern recognition method greatly improves the recognition accuracy of fault diagnosis in comparison with 36 kinds of state-of-the-art dimensionality reduction algorithms and 44 kinds of state-of-the-art classifiers. This newly proposed method not only solves the difficulty that phase-resolved partial discharge (PRPD) cannot be applied under DC condition but also immensely facilitates the fault diagnosis of HVDC GIS.

**Keywords:** HVDC GIS; fault diagnosis; pulse current measurement; statistical feature extraction; generalized discriminant component analysis; SVM



**Citation:** Zhou, R.; Gao, W.; Liu, W.; Ding, D.; Zhang, B. Statistical Feature Extraction Combined with Generalized Discriminant Component Analysis Driven SVM for Fault Diagnosis of HVDC GIS. *Energies* **2021**, *14*, 7674. <https://doi.org/10.3390/en14227674>

Academic Editor:  
Saravanakumar Arumugam

Received: 1 November 2021

Accepted: 8 November 2021

Published: 16 November 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

At present, the power grid is developing towards the direction of high voltage, large capacity and intensification, and the power supply reliability requirement is gradually improved. In this context, the gas insulated switchgear (GIS), with fully enclosed structures, is increasingly being used by power grids at different levels [1]. Meanwhile, due to the increasing use of offshore wind energy and therefore increased demand for energy, transmission to onshore is required. For this energy collection, offshore platforms are needed where space is very expensive, and DC GIS offers a solution [2]. In order to improve the transmission capacity and power supply reliability, the long-term fault evolution and aging problem of HVDC GIS must be solved urgently, having important value in theoretical research and engineering applications. The damage degree of insulation defect to GIS is closely related to the type of insulation defect itself. Different types of insulation defects have different fault evolution laws and different influences on the aging of GIS insulating materials, which will also result in different maintenance and treatment measures. Therefore, accurately identifying the types of insulation defects inside GIS to perform fault

diagnosis will be of great significance for guiding the maintenance work as well as ensuring the safe and stable operation of the GIS.

Dimensionality reduction plays a vital role in the process of pattern recognition. On the one hand, effective dimensionality reduction technology can reduce computational complexity and save computational time of pattern recognition. On the other hand, too high dimensionality of training samples' recognition vectors, which consist of features used to discriminate different classes, may reduce the generalization ability of the classifier adopted in the process of pattern recognition [3]. Generally speaking, feature selection and subspace projection technology are two main methods that are widely adopted for dimensionality reduction in pattern recognition [4,5]. The frequently used feature selection methods involve the filter approach, the wrapper approach, the embedded approach [6], etc. The filtering methods utilize an independent measure to evaluate features without involving any learning algorithms [7], such as similarity-based methods (i.e., Fisher Score [4], ReliefF [8], etc.), statistics-based methods (i.e., *t*-test [5], etc.), correlation-based methods (i.e., CFS [8], etc.) and information theory-based methods (i.e., fast correlation-based filter (FCBF) [8], minimum-redundancy-maximum-relevancy (mRMR) [6], etc.), to name a few. The wrapper methods make use of learning algorithms to evaluate which features are optimal, for which metaheuristic-search-based algorithms can be used efficiently [9], such as genetic algorithms (GA), simulated annealing (SA), differential evolution (DE), ant colony optimization (ACO), particle swarm optimization (PSO), tabu search (TS), etc. The embedded methods combine both of the previous methods, in which feature selection and learning cannot be separated, such as random forests (RF) [5], Group LASSO (GLASSO) [10], SVM-RFE [5], sparse logistic regression with Bayesian regularization (BLogReg) [11], sparse multinomial logistic regression via Bayesian  $L_1$  regularization (SBMLR) [12], manifold regularized discriminative feature selection (MDFS) [13], etc. Obviously, filter methods ignore interactions with the learning algorithms. Meanwhile, the wrapper and embedded methods are specified classifiers, suffering risks of overfitting and being computationally intensive. Compared with feature selection, subspace projection technology uses all the information contained in the recognition vector, which can be roughly divided into two types: unsupervised subspace projection technology and supervised subspace projection technology [14]. Commonly used unsupervised subspace projection techniques mainly include principal component analysis (PCA), Kernelized PCA (KPCA) [15] and various variants of PCA, such as probabilistic PCA (PPCA) [16], multidimensional scaling (MDS) [17], t-SNE [18], local linear embedding (LLE) [19], Isomap [20], Laplacian eigenmaps (LE) [21], autoencoder (AE) [22], etc. Since the unsupervised subspace projection technology does not involve any class information, for example, although PCA satisfies the minimum mean square error criterion and the maximum entropy criterion when the recognition vector satisfies the joint Gaussian distribution [4], the effect of PCA used in pattern recognition is not good, and the supervised subspace projection technology is more conducive to pattern recognition [23]. Commonly used supervised subspace projection techniques mainly include NCA [24], supervised locality preserving projection (SLPP), locality sensitive discriminant analysis (LSDA), S-Isomap [25], Fisher Linear Discriminant Analysis (FDA) [26], Multi-dimensional FDA (MD-FDA) [23], successively orthogonal discriminant analysis (SODA) [27], principal-component discriminant component analysis (PC-DCA), regularized FDA (RFDA) [4] (RFDA is also referred to as BDCA for distinction), etc. In addition, there are many derived versions of FDA, such as local FDA (LFDA) [28], rotational invariant linear discriminant analysis (RILDA) [29], sparse uncorrelated linear discriminant analysis (SULDA) [30], robust linear discriminant analysis (RLDA) [31],  $L_1$ -norm-based global optimal locality preserving LDA (GLLDA- $L_1$ ) [32], etc. It can be concluded from the above investigation that the overwhelming majority of supervised subspace projection technologies are variants of FDA. However, the problem with SODA is that the within-class scatter matrix in each iteration may become ill-conditioned. Once the within-class scatter matrix in a certain iteration is a singular matrix, the direction of the subsequent projection vectors obtained by SODA after this iteration will become exactly the same, so that the

projection vectors obtained by SODA may be of redundancy as the projection vectors in the same direction cannot improve the recognition ability. PC-DCA still suffers the problem of numerical instabilities of MD-FDA caused by the ill-conditioned within-class scatter matrix. Furthermore, there exist some serious fundamental errors and unreasonable aspects regarding BDCA in [4]. Firstly, the division method of signal-subspace and noise-subspace by Professor S. Y. Kung is not universal. Secondly, the assumption that all the eigenvalues of BDCA's discriminant matrix corresponding to the noise-subspace approximate to 1 is incorrect. Thirdly, there exist problems of numerical instabilities in the kernelization form of BDCA given in [4]. Lastly, the theories of BDCA lack rigorously mathematical proofs. All the problems mentioned above will be resolved by generalized discriminant component analysis (GDCA) as well as its kernelization forms proposed in this paper.

In addition, alternating current (AC) partial discharge (PD) pattern recognition research has accounted for the main proportion and is very mature [18,33], while DC PD pattern recognition has gradually been involved but is still in its infancy and has not formed a unified standard, due to lack of phase information, which mainly comprises the Centor score method [34], chaotic analysis method [35], NoDi pattern method [36], compressed sensing theory [37], support vector machine (SVM) [38,39], etc. The main deficiencies in current DC PD pattern recognition are detailed as follows:

(a) There exist features extracted from the waveform of pulse current signal, which are influenced by the specific experimental platform and measurement system.

(b) Most of the related research papers are based on ideal defect models under some specific voltage level to perform PD tests, but the actual GIS operation site may have partial discharges from insulation defects of different voltage levels, types, locations and sizes. Even under the same defect type, the voltage level, defect size and defect location all have greater impacts on the DC PD pulse. The problem of DC PD pattern recognition for insulation defects with different sizes and locations under different voltage levels still needs to be effectively solved urgently.

(c) Most of the existing literature verifies the recognition accuracy of the corresponding PD pattern recognition method based on the assumption that sufficient experimental data can be obtained from the designed defect models in a laboratory environment. When the number of available discharges to be recognized is relatively small, whether the PD pattern recognition method can also be applied or not has not been verified.

(d) Except ensuring the recognition accuracy, how to reduce recognition time as much as possible so that reasonable measures will be taken as soon as possible to minimize the damage of insulation defects to GIS should be researched.

In order to solve the above problems, we built a set of 220 kV HVDC GIS experiment platform and manufactured four different types of insulation defects (including multiple sizes and locations). For each insulation defect, multiple voltage levels were set, ranging from the beginning of stable discharge to the final breakdown or the highest voltage that the experimental platform can provide, and stepwise-boosting voltages were applied with each voltage level lasting for 1 h. Finally, a total of 180,828 pulse current signals were successfully measured. Then, the apparent discharge quantity and the discharge time, two inherent physical quantities unaffected by the experimental platform and measurement system, were obtained after the pulse current signal was denoised, according to which 70 statistical features were extracted. In this paper, a pattern recognition method based on GDCA and its kernelization forms driven SVM is detailed and the corresponding selection criterion of involved parameters is established. Combining the Monte-Carlo experimental method with the cross-validation test strategy, a wealth of estimation indicators for classification results are calculated. The results show that the newly proposed pattern recognition method greatly improves the recognition accuracy in comparison with 36 kinds of state-of-the-art dimensionality reduction algorithms and 44 kinds of state-of-the-art classifiers. This newly proposed method not only solves the difficulty that phase-resolved partial discharge (PRPD) cannot be applied under DC conditions but also immensely facilitates the fault diagnosis of HVDC GIS.

The subsequent structure of this paper is arranged as follows: Section 2 introduces the GIS experimental platform and insulation defect settings; Section 3 describes the 70 statistical features extracted from the inherent physical quantities of pulse current signal; Section 4 proposes the theories and algorithms of GDCA and its kernelization forms; Section 5 gives the results and discussions of the newly proposed pattern recognition method based on GDCA and its kernelization forms driven SVM; and the paper is concluded in Section 6.

## 2. Experimental Platform and Insulation Defects

### 2.1. Experimental Platform

The schematic diagram of the experimental platform is shown in Figure 1, consisting of 220 V AC power supply (powered by WH38905 ultra-isolation transformer), ABB close switch AS, voltage regulator VR, step-up transformer BT (turns ratio is 1:1000), silicon stack  $D_1$  and  $D_2$  (rated rectifier current is 12 mA and rated inverse peak voltage value is 200 kV), protection resistor  $R_1$  (1.6 M $\Omega$ ), ZWF200-0.1 DC capacitor  $C_1$  (composed of two capacitors 0.1010  $\mu$ F and 0.1015  $\mu$ F in series, both of which have a rated voltage of 200 kV), resistor divider (RD, divider ratio is 8000:1), multimeter, protection resistor  $R_2$  (2.13 M $\Omega$ ), high voltage bushing, test sleeve (mainly consisted of HV electrode, insulator, low-voltage (LV) electrode, and insulation support),  $\text{SF}_6/\text{N}_2$  gas filling device, signal detection impedance  $Z_1$  and contrast detection impedance  $Z_2$  ( $Z_1$  and  $Z_2$  are both RLC type and identical), coupling capacitor  $C_k$  (197.8 pF), pulse current amplifier (PCAP), ultrasonic probe (UAP), ultrasonic amplifier (UAA), built-in UHF sensor (BUHFS), two DLM2054 oscilloscopes (the highest sampling rate is 2.5 GSa/s, and the bandwidth is 500 MHz) and one Agilent DSO-S 254 A oscilloscope (the highest sampling rate is 20 GSa/s, and the bandwidth is 2.6 GHz). Note that only pulse current signals are researched in this paper due to limited space; the other two kinds of signals, UHF signal and ultrasonic signal, will be researched in other papers.

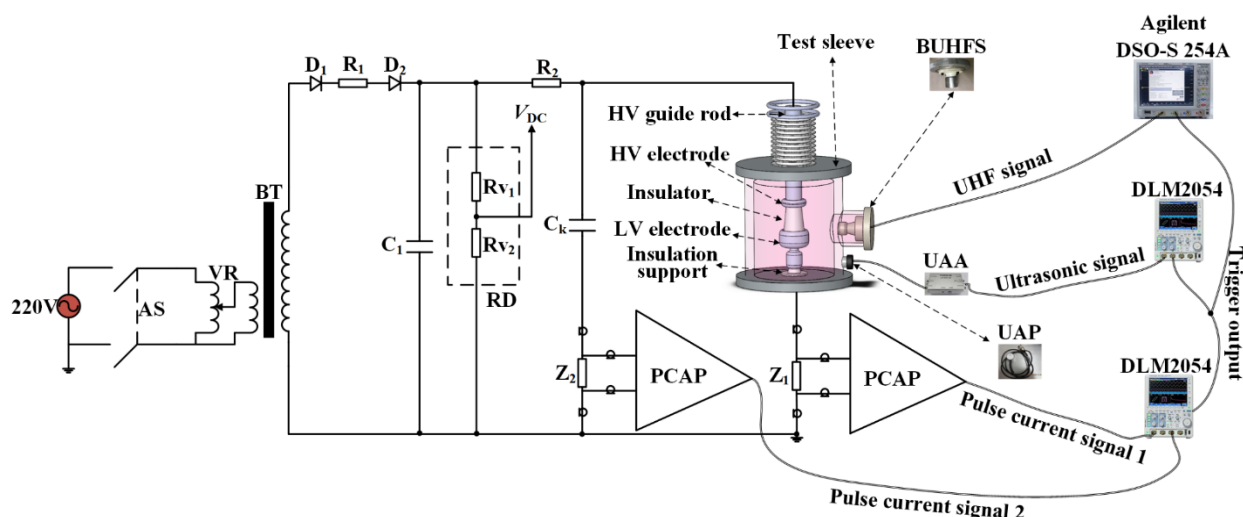


Figure 1. The schematic diagram of the experimental platform.

### 2.2. Insulation Defects

In this paper, four different types of insulation defects are manufactured, corresponding to solid insulation air gap discharge, surface discharge, floating discharge and point discharge, among which the solid insulation air gap defect adopts a self-made vacuum casting block using bisphenol-A epoxy resin shown in Figure 2, and the remaining three types of insulation defects are all set on the GIS post insulator shown in Figure 3. In order to take into account the influences of the defect's location and size on the pulse current signal, different defect locations or defect sizes are set for the same type of defect. The details are shown as Table 1.

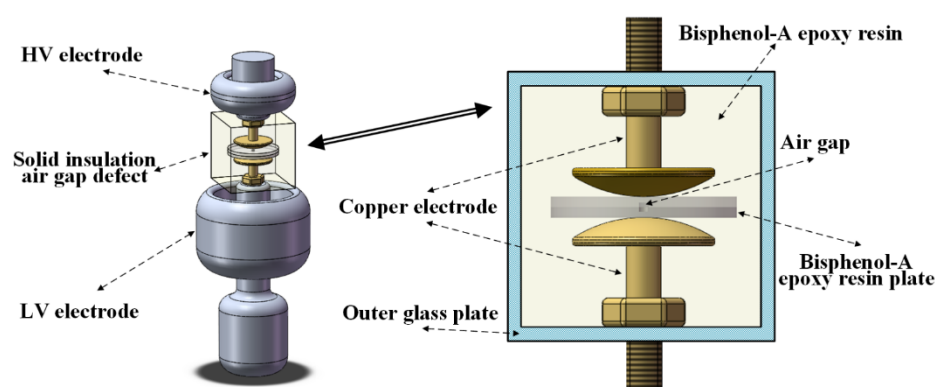


Figure 2. The schematic diagram of solid insulation air gap defect.

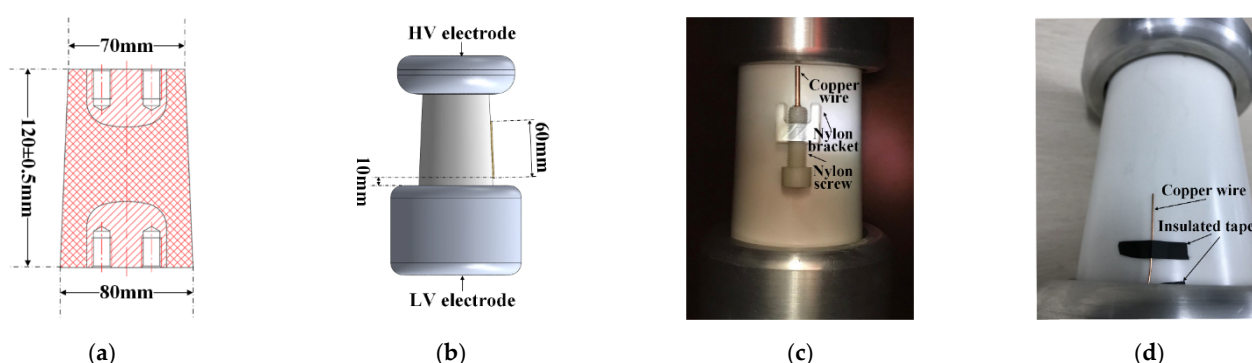


Figure 3. The dimensioning of insulator and the schematic diagrams of No. 4 surface defect, No. 2 floating defect and No. 3 point defect. (a) Dimensioning of insulator; (b) No. 4 surface defect; (c) No. 2 floating defect; (d) No. 3 point defect.

Table 1. Locations and sizes of four different insulation defects.

Solid Insulation Air Gap Defect				Post Insulator Defects	Surface Defect				Floating Defect			Point Defect		
label	1	2	3	label	1	2	3	4	1	2	3	1	2	3
Diameter/mm	2	1	0.5	Diameter/mm	1	0.5	0.5	0.5	0.6	2.5	0.6	0.6	2.5	0.6
Height/mm	2	1	0.5	Length/mm	60	60	30	60	20	20	20	30	30	30
				Distance to HV electrode/mm	10	10	10		1	1		0	0	
				Distance to LV electrode/mm				10			1			0

### 3. Statistical Features Extraction from the Inherent Physical Quantities

As stated in Sections 1 and 2, for each insulation defect, multiple voltage levels were set, ranging from the beginning of stable discharge to the final breakdown or the highest voltage that the experimental platform can provide, and stepwise-boosting voltages were applied with each voltage level lasting for 1 h. Finally, a total of 180,828 pulse current signals were successfully measured, consisting of 540 sample points (one sample point comprises the whole discharge data recorded during the 1 h experiment of the specific insulation defect under the corresponding voltage level, containing at least 50 continuous discharge signals). For each single pulse after being denoised, the corresponding apparent discharge quantity and discharge time, two inherent physical quantities unaffected by the experimental platform and measurement system as well as reflecting the inherent properties of the discharge sources, can be obtained accurately. All the statistical features extracted in this section are based on the above-mentioned two inherent physical quantities.



In general, the extraction of statistical feature quantities is commonly based on the distribution function (continuous case) or probability distribution (discrete case) of one-dimensional or multi-dimensional random variables, which we extend to PD data modes in this paper. PD data modes refer to the statistical relationship diagrams involved with the discharge time, the apparent discharge quantity or their corresponding differences, not necessarily representing probability distributions. The following discussion is focused on a certain discharge sample point.

Assume that the apparent discharge quantity sequence of the current discharge sample point is denoted as  $Q = \{q_r \mid r = 1, 2, \dots, SN\}$ , where  $q_r$  denotes the apparent discharge quantity of the  $r$ th single pulse belonging to the sample point and  $SN$  denotes the number of pulses in the current discharge sample point; the discharge time sequence is denoted as  $PDT = \{PDT_r \mid r = 1, 2, \dots, SN\}$ , where  $PDT_r$  denotes the discharge time of the  $r$ th discharge. With regard to the  $r$ th discharge, the forward discharge time interval is  $\Delta t_{pre} = PDT_r - PDT_{r-1}$  and the backward discharge time interval is  $\Delta t_{suc} = PDT_{r+1} - PDT_r$ ; the first-order difference of apparent discharge quantity is  $\Delta q = q_r - q_{r-1}$  and the first-order difference of discharge time interval is  $\Delta(\Delta t) = PDT_r - 2PDT_{r-1} + PDT_{r-2}$ . In addition,  $T$  denotes the duration of the sample point;  $n(q_r)$  and  $f_{PD}(q_r)$  denote the discharge number and discharge repetition rate corresponding to the pluses with apparent discharge quantity equal to  $q_r$ ;  $U$  denotes the DC voltage applied across the insulation defect when obtaining the sample point and  $U_s$  denotes the corresponding initial voltage of partial discharge;  $WP_r$  ( $r = 1, 2, \dots, SN$ ) denotes the energy of the  $r$ th discharge and  $CP$  denotes the partial discharge cumulative product [34].

As stated above, a PD data mode does not always represent a kind of probability distribution. For a two-dimensional PD data mode, uniformly expressed as  $y_i = f(x_i)$ , it needs to be first analogized to be the probability distribution of a discrete random variable  $X$ . Let all possible values of  $X$  be denoted as  $x_1, x_2, \dots, x_n$  ( $x_1 < x_2 < \dots < x_n$ ), the corresponding probabilities are  $p_1, p_2, \dots, p_n$ . The transformation formula is shown as Equation (1).

$$p_i = \frac{y_i}{\sum_{i=1}^n y_i} \quad (1)$$

By Equation (1), we can calculate the statistical features of any two-dimensional PD data mode (when the PD data mode is a histogram of a certain random variable) or analogical statistical features (when the PD data mode is not a histogram of a certain random variable). The involved features of two-dimensional PD data modes comprise expectation (denoted as  $m_1$ ), standard deviation (denoted as  $m_2$ ), skewness (denoted as *Skewness*), kurtosis (denoted as *Kurtosis*) and the number of peaks (denoted as *Peaks*).

When the two-dimensional PD data mode represents a variable histogram, namely the probability distribution (it should be called the frequency distribution to be more precise, an estimate of the actual probability distribution using experimental data), we can use Weibull distribution (when the random variable is non-negative) or one-dimensional kernel density estimation [23] to fit the corresponding probability distribution. Using the maximum likelihood estimation method to fit the Weibull distribution, the corresponding scale parameter  $\alpha$  and shape parameter  $\beta$  can be obtained. The kernel density estimation is a non-parametric method of estimating the probability density function. Assuming that the unary probability density function to be estimated is denoted as  $g$ , its kernel density estimation function is denoted as  $\hat{g}$  in Equation (2), where  $K$  is a non-negative kernel function and  $h$  is a smoothing parameter or referred to as bandwidth. We adopt the adaptive kernel density estimator based on the linear diffusion process proposed by [40] to estimate the optimal smoothing parameter  $h_{best}$ .

$$\left\{ \begin{array}{l} \hat{g}(x|h) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) \\ h_{best} = \underset{h}{\operatorname{argmin}} \left\{ E_f \left[ \int (\hat{g}(x|h) - g(x))^2 dx \right] \right\} \end{array} \right\} \quad (2)$$

Similarly, when the three-dimensional PD data mode represents a binary histogram, namely a two-dimensional probability distribution, two-dimensional kernel density estimation can be used to fit the corresponding probability distribution, and finally, two optimal smoothing parameters can be calculated [40], denoted as  $H_x$  and  $H_y$ . The energy entropy of the binary histogram can also be calculated by Equation (3), denoted as *Entropy*, where  $F_i$  denotes the probability (frequency to be more precise) corresponding to the  $(i, j)$ th binary grid.

$$Entropy = -\sum_i \sum_j F_i \ln F_i \quad (3)$$

We first introduce each PD data mode used in this paper labelled from A to L, and then detail the extracted statistical features based on the corresponding PD data mode.

A.  $H_n(q)$

$H_n(q)$  is the frequency density histogram of  $q$ . The extracted statistical features consist of  $m_1$ ,  $m_2$ , *Skewness*, *Kurtosis*, *Peaks*, Weibull distribution fitting parameters  $\alpha$  and  $\beta$ , as well as the optimal smoothing parameter  $h_{\text{best}}$ .

B.  $H_n(WP)$

$H_n(WP)$  is the frequency density histogram of  $WP$ . The extracted statistical features consist of  $m_1$  and  $m_2$ .

C.  $H_n(\Delta q)$

$H_n(\Delta q)$  is the frequency density histogram of  $\Delta q$ . The extracted statistical features consist of  $m_1$ ,  $m_2$ , *Skewness*, *Kurtosis*, *Peaks*, and the optimal smoothing parameter  $h_{\text{best}}$ .

D.  $H_n(\ln(\Delta t))$

$H_n(\ln(\Delta t))$  is the frequency density histogram of  $\ln(\Delta t)$ . The extracted statistical features consist of  $m_1$ ,  $m_2$ , *Skewness*, *Kurtosis*, *Peaks*, and the optimal smoothing parameter  $h_{\text{best}}$ . In addition, the Weibull distribution can be used to fit the probability distribution function of  $\Delta t$ , which must always be non-negative, to obtain the fitting parameters  $\alpha$  and  $\beta$ .

E.  $Hq(CP)$

$Hq(CP)$  is the two-dimensional relationship diagram of PD cumulative product  $CP$  [34] calculated by Equation (4) and apparent discharge quantity  $q$ . The extracted statistical features consist of  $m_1$ ,  $m_2$ , *Skewness*, and *Kurtosis*.

$$\begin{cases} CP(q) = q \cdot \sum_{q_r \geq q} f_{PD}(q_r) = \frac{q \cdot \sum_{q_r \geq q} n(q_r)}{T} \\ q \in [\min(\mathbf{Q}), \max(\mathbf{Q})] \end{cases} \quad (4)$$

Since the discharge time interval  $\Delta t$  generally does not keep the same, it is necessary to sort all  $\Delta t$  first, and then set an appropriate interval range to divide  $\ln(\Delta t)$  at equal intervals in order to make a PD data mode of  $q$  and  $\Delta t$ . Let the total number of intervals be  $NI$ ,  $q_n$  ( $n = 1, 2, \dots, NI$ ) denote the average of all the apparent discharge quantities in the  $n$ th interval and  $q_{\max}$  denote the corresponding maximum of all the apparent discharge quantities in the  $n$ th interval. Then, we can derive the following four PD data modes, of which the extracted statistical features consist of  $m_1$ ,  $m_2$ , *Skewness*, *Kurtosis* and *Peaks*.

F.  $Hq_n(\ln(\Delta t_{\text{suc}}))$

$Hq_n(\ln(\Delta t_{\text{suc}}))$  is the two-dimensional relationship diagram between  $q_n$  and  $\ln(\Delta t_{\text{suc}})$ .

G.  $Hq_{\max}(\Delta t_{\text{suc}})$

$Hq_{\max}(\Delta t_{\text{suc}})$  is the two-dimensional relationship diagram between  $q_{\max}$  and  $\ln(\Delta t_{\text{suc}})$ .

H.  $Hq_n(\ln(\Delta t_{\text{pre}}))$

$H_{qn}(\Delta t_{pre})$  is the two-dimensional relationship diagram between  $q_n$  and  $\ln(\Delta t_{pre})$ .

#### I. $H_{q_{max}}(\ln(\Delta t_{pre}))$

$H_{q_{max}}(\Delta t_{pre})$  is the two-dimensional relationship diagram between  $q_{max}$  and  $\ln(\Delta t_{pre})$ .

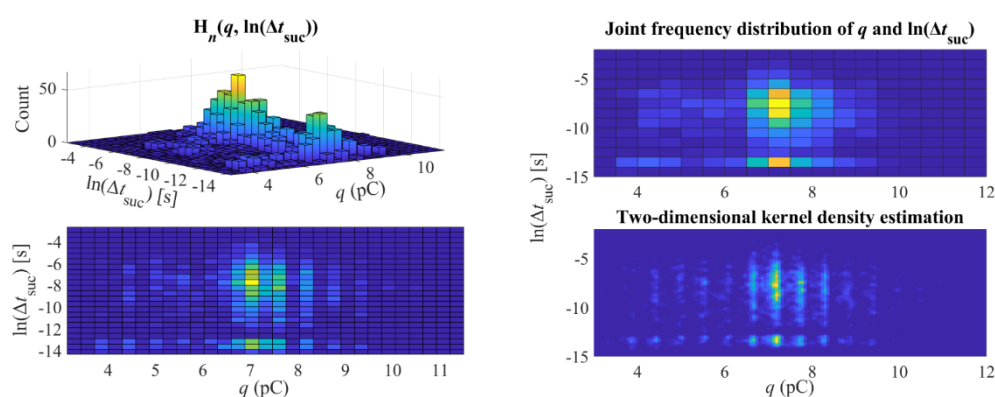
Analogous to the AC PD phase distribution pattern, describing the difference in the distribution shapes of the two-dimensional relationship diagrams corresponding to the positive and negative half cycle of power frequency period, the above four PD data modes can be used to construct two combination diagrams, one of which combines  $H_{qn}(\ln(\Delta t_{suc}))$  with  $H_{qn}(\ln(\Delta t_{pre}))$  and the other of which combines  $H_{q_{max}}(\ln(\Delta t_{suc}))$  with  $H_{q_{max}}(\ln(\Delta t_{pre}))$ . From the combination diagrams, the extracted statistical features consist of cross-correlation factor and degree of asymmetry, denoted as *CC* and *Asymmetry*, respectively. Let  $y_{1i}$  ( $i = 1, 2, \dots, n$ ) represent the ordinate values of  $H_{qn}(\Delta t_{suc})$  or  $H_{q_{max}}(\Delta t_{suc})$  and  $y_{2i}$  ( $i = 1, 2, \dots, n$ ) represent the ordinate values of  $H_{qn}(\Delta t_{pre})$  or  $H_{q_{max}}(\Delta t_{pre})$ ; *CC* and *Asymmetry* can be calculated as Equation (5).

$$CC = \frac{\sum_{i=1}^n y_{1i}y_{2i} - \frac{1}{n} \sum_{i=1}^n y_{1i} \cdot \sum_{i=1}^n y_{2i}}{\sqrt{\left[ \sum_{i=1}^n y_{1i}^2 - \frac{1}{n} \left( \sum_{i=1}^n y_{1i} \right)^2 \right] \cdot \left[ \sum_{i=1}^n y_{2i}^2 - \frac{1}{n} \left( \sum_{i=1}^n y_{2i} \right)^2 \right]}}, Asymmetry = \frac{\sum_{i=1}^n y_{2i}}{\sum_{i=1}^n y_{1i}} \quad (5)$$

In this paper, binary joint distributions are also used as three-dimensional PD data modes labelled from J to L, which take the apparent discharge quantity  $q$ , the discharge time interval  $\Delta t$  or their corresponding differences as the joint variables, of which the extracted statistical features consist of  $H_x$  and  $H_y$ , two optimal smoothing parameters of two-dimensional kernel density estimation, as well as energy entropy *Entropy*.

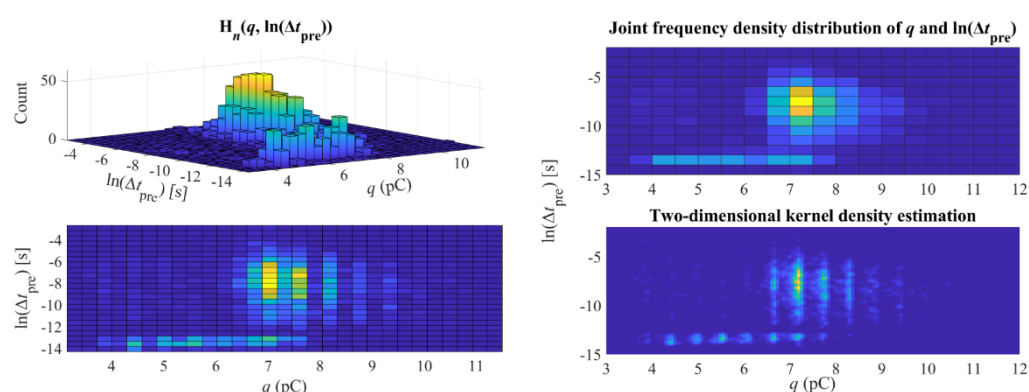
#### J. $H_n(q, \ln(\Delta t))$

$H_n(q, \ln(\Delta t))$  is a binary histogram of the apparent discharge quantity  $q$  and the natural logarithm of discharge time interval  $\ln(\Delta t)$ , with two cases  $H_n(q, \ln(\Delta t_{suc}))$  and  $H_n(q, \ln(\Delta t_{pre}))$ , illustrated as Figures 4 and 5, respectively, in which the fitting results of two-dimensional kernel density estimation are also given.



**Figure 4.** The illustration of  $H_n(q, \ln(\Delta t_{suc}))$  and the corresponding result of two-dimensional kernel density estimation.

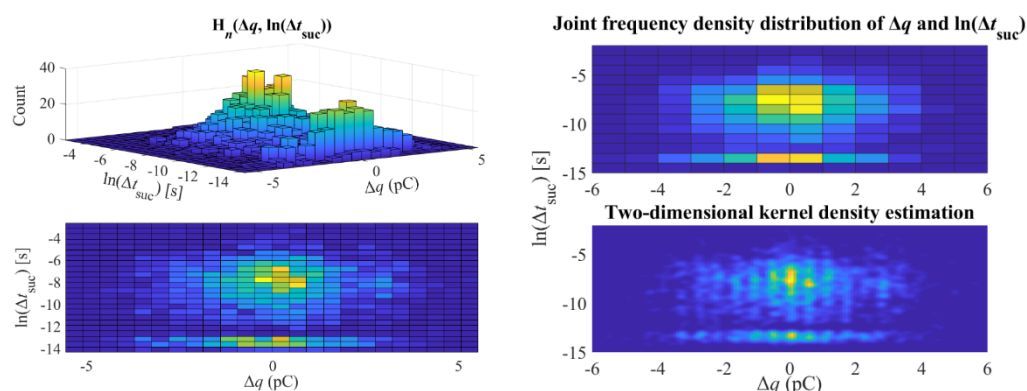




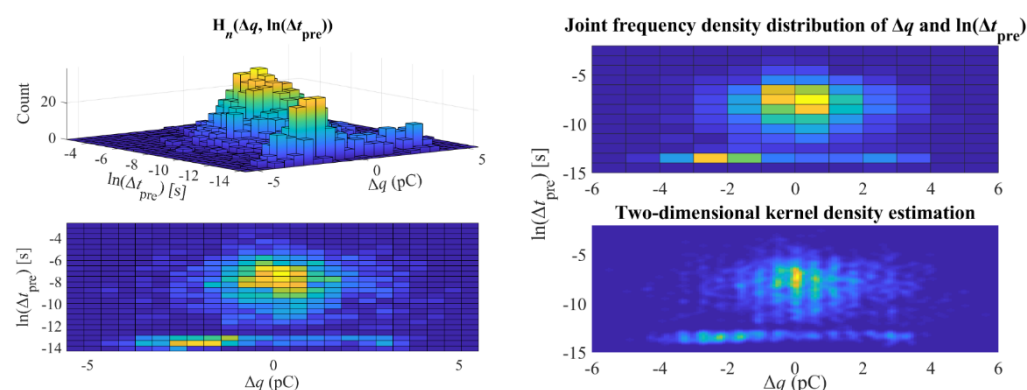
**Figure 5.** The illustration of  $H_n(q, \ln(\Delta t_{pre}))$  and the corresponding result of two-dimensional kernel density estimation.

#### K. $H_n(\Delta q, \ln(\Delta t))$

$H_n(\Delta q, \ln(\Delta t))$  is a binary histogram of the first-order difference of apparent discharge quantity  $\Delta q$  and the natural logarithm of discharge time interval  $\ln(\Delta t)$ , with two cases  $H_n(\Delta q, \ln(\Delta t_{suc}))$  and  $H_n(\Delta q, \ln(\Delta t_{pre}))$ , illustrated as Figures 6 and 7, respectively, in which the fitting results of two-dimensional kernel density estimation are also given.



**Figure 6.** The illustration of  $H_n(\Delta q, \ln(\Delta t_{suc}))$  and the corresponding result of two-dimensional kernel density estimation.

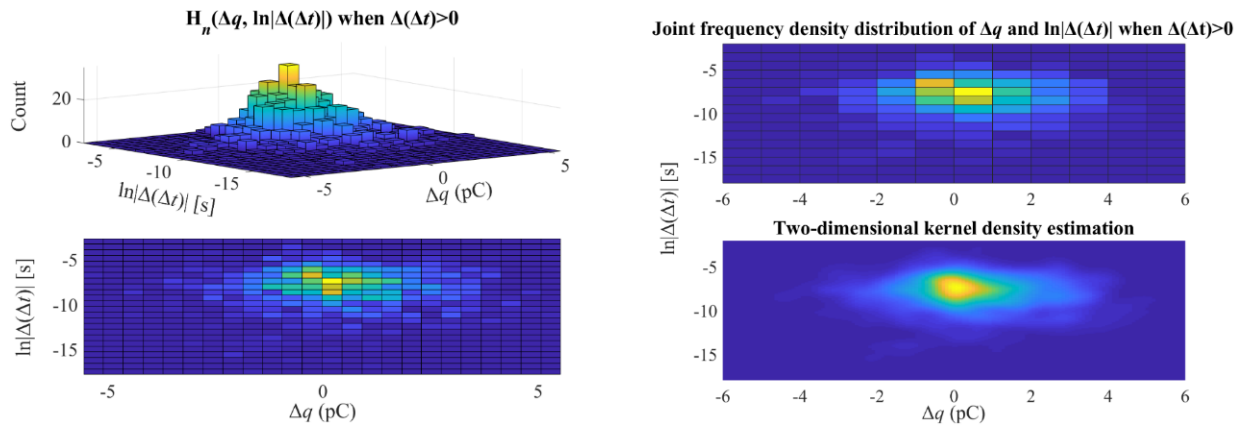


**Figure 7.** The illustration of  $H_n(\Delta q, \ln(\Delta t_{pre}))$  and the corresponding result of two-dimensional kernel density estimation.

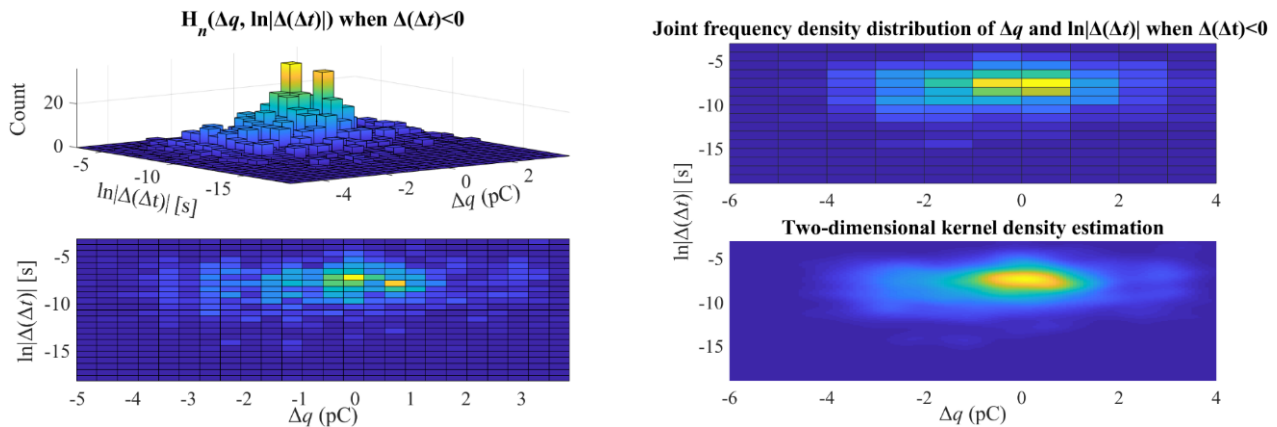
#### L. $H_n(\Delta q, \ln|\Delta(\Delta t)|)$

$H_n(\Delta q, \ln|\Delta(\Delta t)|)$  is a binary histogram of the first-order difference of apparent discharge quantity  $\Delta q$  and the natural logarithm of the absolute value of the first-order

difference of discharge time interval  $\ln |\Delta(\Delta t)|$ . Considering that  $\Delta(\Delta t)$  is not always a positive number, the two cases of  $\Delta(\Delta t) > 0$  and  $\Delta(\Delta t) < 0$  are taken into account separately, illustrated as Figures 8 and 9, respectively, in which the fitting results of two-dimensional kernel density estimation are also given.



**Figure 8.** The illustration of  $H_n(\Delta q, \ln |\Delta(\Delta t)|)$  and the corresponding result of two-dimensional kernel density estimation when  $\Delta(\Delta t) > 0$ .



**Figure 9.** The illustration of  $H_n(\Delta q, \ln |\Delta(\Delta t)|)$  and the corresponding result of two-dimensional kernel density estimation when  $\Delta(\Delta t) < 0$ .

#### 4. GDCA and Its Kernelization Forms

This section promotes the supervised subspace projection technology from BDCA to GDCA and its kernelization forms. Some indispensable terminologies and symbols should first be introduced [4,23]:

**Recognition Vector:** denoted as  $\mathbf{x}_i \in \mathbb{R}^{M \times 1}$  ( $i = 1, 2, \dots, N$ ), consists of all statistical features of the  $i$ th discharge sample point. The corresponding ones in the intrinsic vector space and the empirical vector space are denoted as  $\vec{\phi}(\mathbf{x}) \in \mathbb{R}^{J \times 1}$  and  $\vec{k}(\mathbf{x}) \in \mathbb{R}^{N \times 1}$ , respectively, where  $M$  denotes the number of features,  $J$  denotes the dimensionality of the reproducing kernel Hilbert space (i.e., intrinsic vector space) and  $N$  denotes the number of samples. According to Section 3,  $M = 70$  and  $N = 540$ .

**Feature Sample Matrix:** denoted as  $\mathbf{X} \in \mathbb{R}^{M \times N}$ , consists of all available recognition vectors.

**Class Number:** denoted as  $CN$ , the total number of all classes. According to Section 2,  $CN = 4$ .

**Within-class Scatter Matrix:** denoted as  $\mathbf{S}_W \in \mathbb{R}^{M \times M}$ .

**Between-class Scatter Matrix:** denoted as  $\mathbf{S}_B \in \mathbb{R}^{M \times M}$ .

**Center-adjusted Scatter Matrix:** denoted as  $\mathbf{S}_C \in \mathbb{R}^{M \times M}$ .

Different from the derivation of BDCA [4], the newly proposed GDCA in this paper starts directly from PC-DCA shown in Equation (6) and improves the corresponding constraint condition, which can be more robust and flexible than BDCA. Then, BDCA can be regarded as a special case of the proposed GDCA under a specific parameter value.

$$\mathbf{W}_P = \operatorname{argmax}_{\mathbf{W} \in \mathbb{R}^{M \times m}} \left\{ \operatorname{trace} \left[ \mathbf{W}^T (\mathbf{S}_C + \rho \mathbf{I}) \mathbf{W} \right] \mid \mathbf{W}^T \mathbf{S}_W \mathbf{W} = \mathbf{I} \right\} \quad (6)$$

At first, the projection matrix of PC-DCA in Equation (6) is divided into two parts: signal-subspace projection matrix  $\mathbf{W}_{PS}$  and noise-subspace projection matrix  $\mathbf{W}_{PN}$ . Without loss of generality, suppose that the projected dimensionality  $m$  is larger than  $\operatorname{rank}(\mathbf{S}_B)$ , and then Equation (6) can be transformed into Equation (7) according to the signal-subspace and the noise-subspace.

$$\begin{cases} \mathbf{W}_{PS} = \operatorname{argmax}_{\mathbf{W} \in \mathbb{R}^{M \times \operatorname{rank}(\mathbf{S}_B)}} \left\{ \operatorname{trace} \left[ \mathbf{W}^T (\mathbf{S}_C + \rho \mathbf{I}) \mathbf{W} \right] \mid \mathbf{W}^T \mathbf{S}_W \mathbf{W} = \mathbf{I} \right\} \\ \mathbf{W}_{PN} = \operatorname{argmax}_{\mathbf{W} \in \mathbb{R}^{M \times [m - \operatorname{rank}(\mathbf{S}_B)]}} \left\{ \operatorname{trace} \left[ \mathbf{W}^T (\mathbf{S}_C + \rho \mathbf{I}) \mathbf{W} \right] \mid \mathbf{W}^T \mathbf{S}_W \mathbf{W} = \mathbf{I}, \mathbf{W}^T \mathbf{S}_B \mathbf{W} = 0 \right\} \\ = \operatorname{argmax}_{\mathbf{W} \in \mathbb{R}^{M \times [m - \operatorname{rank}(\mathbf{S}_B)]}} \left\{ \operatorname{trace} (\mathbf{W}^T \mathbf{W}) \mid \mathbf{W}^T \mathbf{S}_W \mathbf{W} = \mathbf{I}, \mathbf{W}^T \mathbf{S}_B \mathbf{W} = 0 \right\} \end{cases} \quad (7)$$

PC-DCA can be promoted to GDCA by improving the constraint  $\mathbf{W}^T \mathbf{S}_W \mathbf{W} = \mathbf{I}$  of Equation (7) to  $\mathbf{W}^T (\mathbf{S}_W + \delta \mathbf{I}) \mathbf{W} = \mathbf{I}$  ( $\delta > \rho$ ,  $\delta \rightarrow 0^+$  and  $\rho \rightarrow 0$ ). Note that signal-subspace projection matrix is also denoted as  $\mathbf{W}_{PS}$  and noise-subspace projection matrix is also denoted as  $\mathbf{W}_{PN}$  in GDCA, shown as Equation (8):

$$\begin{cases} \mathbf{W}_{PS} = \operatorname{argmax}_{\mathbf{W} \in \mathbb{R}^{M \times \operatorname{rank}(\mathbf{S}_B)}} \left\{ \operatorname{trace} \left[ \mathbf{W}^T (\mathbf{S}_C + \rho \mathbf{I}) \mathbf{W} \right] \mid \mathbf{W}^T (\mathbf{S}_W + \delta \mathbf{I}) \mathbf{W} = \mathbf{I} \right\} \\ \mathbf{W}_{PN} = \operatorname{argmax}_{\mathbf{W} \in \mathbb{R}^{M \times [m - \operatorname{rank}(\mathbf{S}_B)]}} \left\{ \operatorname{trace} (\mathbf{W}^T \mathbf{W}) \mid \mathbf{W}^T (\mathbf{S}_W + \delta \mathbf{I}) \mathbf{W} = \mathbf{I}, \mathbf{W}^T \mathbf{S}_B \mathbf{W} = 0 \right\} \\ = \operatorname{argmax}_{\mathbf{W} \in \mathbb{R}^{M \times [m - \operatorname{rank}(\mathbf{S}_B)]}} \left\{ \operatorname{trace} \left( \frac{\mathbf{I} - \mathbf{W}^T \mathbf{S}_C \mathbf{W}}{\delta} \right) \mid \mathbf{W}^T (\mathbf{S}_W + \delta \mathbf{I}) \mathbf{W} = \mathbf{I}, \mathbf{W}^T \mathbf{S}_B \mathbf{W} = 0 \right\} \\ = \operatorname{argmin}_{\mathbf{W} \in \mathbb{R}^{M \times [m - \operatorname{rank}(\mathbf{S}_B)]}} \left\{ \operatorname{trace} \left[ \mathbf{W}^T (\mathbf{S}_C + \rho \mathbf{I}) \mathbf{W} \right] \mid \mathbf{W}^T (\mathbf{S}_W + \delta \mathbf{I}) \mathbf{W} = \mathbf{I}, \mathbf{W}^T \mathbf{S}_B \mathbf{W} = 0 \right\} \end{cases} \quad (8)$$

It can be seen from Equation (8) that GDCA degenerates into BDCA when  $\rho = 0$ . If we temporarily ignore the constraint that  $\mathbf{W}^T \mathbf{S}_B \mathbf{W} = 0$ , it can be derived that the discriminant matrix of GDCA is  $(\mathbf{S}_W + \delta \mathbf{I})^{-1} (\mathbf{S}_C + \rho \mathbf{I})$  and the signal-subspace projection matrix  $\mathbf{W}_{PS}$  consists of the eigenvectors of  $(\mathbf{S}_W + \delta \mathbf{I})^{-1} (\mathbf{S}_C + \rho \mathbf{I})$  corresponding to the  $\operatorname{rank}(\mathbf{S}_B)$  larger eigenvalues while the noise-subspace projection matrix  $\mathbf{W}_{PN}$  consists of the eigenvectors of  $(\mathbf{S}_W + \delta \mathbf{I})^{-1} (\mathbf{S}_C + \rho \mathbf{I})$  corresponding to the  $m - \operatorname{rank}(\mathbf{S}_B)$  smaller eigenvalues. It is only necessary to further prove that  $\mathbf{W}_{PN}$  has automatically approximately satisfied the constraint that  $\mathbf{W}_{PN}^T \mathbf{S}_B \mathbf{W}_{PN} = 0$  as follows:

Let the  $m - \operatorname{rank}(\mathbf{S}_B)$  smaller eigenvalues of  $(\mathbf{S}_W + \delta \mathbf{I})^{-1} (\mathbf{S}_C + \rho \mathbf{I})$  be arranged in ascending order to form a diagonal matrix  $\mathbf{\Sigma}_{PN}$ , so Equation (9) can be deduced.

$$\begin{aligned} (\mathbf{S}_W + \delta \mathbf{I})^{-1} (\mathbf{S}_C + \rho \mathbf{I}) \mathbf{W}_{PN} &= \mathbf{W}_{PN} \mathbf{\Sigma}_{PN} \\ \Leftrightarrow (\mathbf{S}_W + \delta \mathbf{I})^{-1} [\mathbf{S}_B + (\rho - \delta) \mathbf{I}] \mathbf{W}_{PN} &= \mathbf{W}_{PN} (\mathbf{\Sigma}_{PN} - \mathbf{I}) \\ \Leftrightarrow \mathbf{W}_{PN}^T [\mathbf{S}_B + (\rho - \delta) \mathbf{I}] \mathbf{W}_{PN} &= \mathbf{W}_{PN}^T (\mathbf{S}_W + \delta \mathbf{I}) \mathbf{W}_{PN} (\mathbf{\Sigma}_{PN} - \mathbf{I}) \end{aligned} \quad (9)$$

Combining the constraint condition  $\mathbf{W}_{PN}^T (\mathbf{S}_W + \delta \mathbf{I}) \mathbf{W}_{PN} = \mathbf{I}$  in Equations (8) and (9) can be equivalently converted to Equation (10), from which Equation (11) can be further obtained.

$$\mathbf{W}_{PN}^T \mathbf{S}_B \mathbf{W}_{PN} = \mathbf{\Sigma}_{PN} - \mathbf{I}_{[m - \operatorname{rank}(\mathbf{S}_B)] \times [m - \operatorname{rank}(\mathbf{S}_B)]} + (\delta - \rho) \mathbf{W}_{PN}^T \mathbf{W}_{PN} \quad (10)$$

$$\begin{cases} \frac{\mathbf{w}_i^T \mathbf{S}_B \mathbf{w}_i}{\|\mathbf{w}_i\|^2} = \frac{\lambda_i - 1}{\|\mathbf{w}_i\|^2} + \delta - \rho \\ \frac{\mathbf{w}_i^T \mathbf{S}_B \mathbf{w}_j}{\|\mathbf{w}_i\| \cdot \|\mathbf{w}_j\|} \leq \frac{\mathbf{w}_i^T \mathbf{S}_B \mathbf{w}_j}{\langle \mathbf{w}_i, \mathbf{w}_j \rangle} = \delta - \rho \\ j \neq i \text{ and } i, j = \text{rank}(\mathbf{S}_B) + 1, \text{rank}(\mathbf{S}_B) + 2, \dots, m \end{cases} \quad (11)$$

Combining Equation (11) with Equation (A11) in the Appendix A and the conclusion that  $\lambda_i < 1$  ( $i = \text{rank}(\mathbf{S}_B) + 1, \text{rank}(\mathbf{S}_B) + 2, \dots, m$ ) in the Appendix, Equation (12) can be further derived.

$$\begin{cases} \frac{\mathbf{w}_i^T \mathbf{S}_B \mathbf{w}_i}{\|\mathbf{w}_i\|^2} = \frac{\lambda_i - 1}{\sum_{j=1}^M \frac{1}{\mu_j} u_{ji}^2} + \delta - \rho \leq \delta \lambda_i - \rho < \delta - \rho \\ i = \text{rank}(\mathbf{S}_B) + 1, \text{rank}(\mathbf{S}_B) + 2, \dots, m \end{cases} \quad (12)$$

Based on the fact that  $\delta > \rho$ ,  $\delta \rightarrow 0^+$  and  $\rho \rightarrow 0$ , it can be concluded from Equations (11) and (12) that  $\mathbf{W}_{PN}$  has indeed automatically approximately satisfied the constraint that  $\mathbf{W}_{PN}^T \mathbf{S}_B \mathbf{W}_{PN} = 0$ .

The GDCA algorithm can be given as follows:

#### GDCA algorithm

- (0) Prepare Essential Parameters
  - (0.1) Choose the projected dimensionality  $m$ ;
  - (0.2) Choose regularization parameters  $\delta$  and  $\rho$ , which must satisfy the condition that  $\delta > \rho$ ,  $\delta \rightarrow 0^+$  and  $\rho \rightarrow 0$  (for simplicity, let  $\rho = \alpha\delta$ ,  $\delta \rightarrow 0^+$  and  $\alpha < 1$ ).
- (1) Calculate the between-class scatter matrix  $\mathbf{S}_B$ , within-class scatter matrix  $\mathbf{S}_W$ , and center-adjusted scatter matrix  $\mathbf{S}_C$ 
  - (1.1) Use data preprocessing methods, such as standard normal density (SND) or min-max normalization (MMN), to preprocess the original recognition vectors [41];
  - (1.2) Denote recognition vectors after preprocessing as  $\mathbf{x}_i \in \mathbb{R}^{M \times 1}$  ( $i = 1, 2, \dots, N$ ), then calculate  $\mathbf{S}_B$ ,  $\mathbf{S}_W$  and  $\mathbf{S}_C$ .
- (2) Calculate the projection matrix  $\mathbf{W}_{GDCA}$ 
  - (2.1) If  $m$  is not more than  $\text{rank}(\mathbf{S}_B)$ ,  $\mathbf{W}_{GDCA}$  is consisted of the eigenvectors of  $(\mathbf{S}_W + \delta \mathbf{I})^{-1}(\mathbf{S}_C + \rho \mathbf{I})$  corresponding to the  $\text{rank}(\mathbf{S}_B)$  larger eigenvalues arranged in descending order.
  - (2.2) If  $m$  is larger than  $\text{rank}(\mathbf{S}_B)$ , the signal-subspace projection matrix  $\mathbf{W}_{PS}$  is consisted of the eigenvectors of  $(\mathbf{S}_W + \delta \mathbf{I})^{-1}(\mathbf{S}_C + \rho \mathbf{I})$  corresponding to the  $\text{rank}(\mathbf{S}_B)$  larger eigenvalues arranged in descending order while the noise-subspace projection matrix  $\mathbf{W}_{PN}$  is consisted of the eigenvectors of  $(\mathbf{S}_W + \delta \mathbf{I})^{-1}(\mathbf{S}_C + \rho \mathbf{I})$  corresponding to the  $m - \text{rank}(\mathbf{S}_B)$  smaller eigenvalues arranged in ascending order. Finally,  $\mathbf{W}_{GDCA} = [\mathbf{W}_{PS}, \mathbf{W}_{PN}]$ .
- (3) Normalize projection vectors
 

Let each column vector of  $\mathbf{W}_{GDCA}$  divide its own 2-norm. For any column vector of  $\mathbf{W}_{GDCA}$ , multiply itself by  $-1$  if the element with the largest absolute value is negative.
- (4) Calculate the feature sample matrix after projection  $\mathbf{Y}_{GDCA} = \mathbf{W}_{GDCA}^T \mathbf{X}$
- (5) Whether to change the values of  $\delta$  and  $\rho$ ? Return to 0.2 if yes and go to next step if no.
- (6) Whether to change  $m$ ? Return to 0.1 if yes and output  $\mathbf{W}_{GDCA}$  and  $\mathbf{Y}_{GDCA}$  if no.

We have proved that GDCA algorithm does meet the SNR criterion in the signal-subspace and the noise-power criterion in the noise-subspace, which means SNRs of projected components in the signal-subspace are arranged in descending order while the noise powers of projected components in the noise-subspace are arranged in ascending order; the details of the proof are shown in the Appendix A. Then, we can extend GDCA to the nonlinear case, KGDCA-Intrinsic-Space and KGDCA-Empirical-Space, by means of Gaussian radial basis kernel function (RBF)  $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$ ,  $\gamma > 0$ . Recognition vectors in the original vector space  $\mathbf{x}_i \in \mathbb{R}^{M \times 1}$  are first mapped to the intrinsic or

empirical vector space, and then GDCA is used with regard to intrinsic vectors  $\vec{\phi}(\mathbf{x}) \in \mathbb{R}^{I \times 1}$  or empirical vectors  $\vec{k}(\mathbf{x}) \in \mathbb{R}^{N \times 1}$ , respectively.

## 5. Results and Discussions

In order to demonstrate the advantages of the newly proposed pattern recognition method based on GDCA and its kernelization forms driven SVM, the test strategy of recognition effect based on the combination of Monte-Carlo experimental method and cross-validation is put forward firstly in this section, by which a wealth of estimation indicators for classification results can be calculated. Then, the criterion aimed at finding the optimal  $(\alpha, \delta)$  value-pair for GDCA and the optimal  $(\gamma, \alpha, \delta)$  value-pair for GDCA's kernelization forms is given, through which it is possible to optimally select the parameters involved in GDCA and its kernelization forms in advance without using the estimation indicators for classification results, greatly shortening the time of pattern recognition and ensuring the optimal recognition effect. Finally, results and discussions are detailed.

### 5.1. Test Strategy

First, random sampling is performed on the uniform distribution, so that the recognition vectors of all the discharge sample points are equally divided into five disjoint folds, and then 5-fold cross-validation is performed. Furthermore, the estimation indicators of each fold are calculated separately and the results of 5 folds are averaged. The above process can be regarded as one Monte-Carlo experiment. In order to reduce the impact of the randomness of the data set division on the estimation indicators of the final classification result, the above process is repeated 10 times, which means 10 Monte-Carlo experiments are performed. Finally, the results of the 10 Monte-Carlo experiments are averaged. The whole test strategy is shown in Figure 10. The Binary-SVM presented in Figure 10 adopts two-class support vector classification machine based on soft constraints, also referred to as Binary C-SVC [4]. Let  $\mathbf{x}_i$  ( $i = 1, 2, \dots, N$ ) denote the PD recognition vector corresponding to the  $i$ th sample point input to Binary-SVM;  $y_i$  ( $i = 1, 2, \dots, N$ ), only equal to 1 or  $-1$ , denotes the class label of the  $i$ th sample point. It is worth noting that  $\mathbf{x}_i$  ( $i = 1, 2, \dots, N$ ) can be recognition vectors in the original vector space or after being projected by means of GDCA or its kernelization forms. After solving the corresponding quadratic programming problem, we can obtain the decision function  $f(\mathbf{x})$  with regard to any PD recognition vector  $\mathbf{x}$ .

Since the kernel matrix is a dense matrix in general and may be too large to store, Professor Chih-Jen Lin et al. [42] developed the LIBSVM toolbox, which is widely used for solving classification and regression problems due to its convenience of adjusting parameters, adopting an SMO-type decomposition method [43,44] to solve the quadratic programming problem.

The above-mentioned Binary-SVM is only specified for a two-class situation, and there mainly exist two kinds of methods to extend Binary SVM to Multiclass-SVM, namely one-versus-one scheme and one-versus-all scheme. The one-versus-one scheme needs to train one Binary-SVM for each possible pair of  $CN$  classes, which results in  $CN(CN - 1)/2$  Binary-SVMs. The one-versus-all scheme consisted of  $CN$  Binary-SVMs, each of which is trained for one class and all the other classes. This paper adopts the one-versus-one scheme. However, basic Binary-SVM can only obtain the decision value of the test sample. In order to obtain posterior class probabilities, we firstly adopt Equation (13) to convert the decision values output by Binary-SVM into the estimated pairwise class probabilities  $r_{ij}$  ( $i \neq j$  and  $i, j = 1, 2, \dots, CN$ ), where parameters  $A$  and  $B$  can be obtained by solving the regularized maximum likelihood problem of maximizing the log-likelihood function in Equation (14), a kind of relative entropy or Kullback–Leibler divergence [45]. In Equation (14),  $t_i$  denotes the maximum a posteriori (MAP) estimation for the target probability shown as Equation (15), consisted of two values, namely  $t_+$  and  $t_-$ , corresponding to positive and negative samples,



respectively. Compared with  $t_+ = 1$  and  $t_- = 0$ , Equation (15) can effectively avoid the overfitting of Equation (13).

$$r_{ij} = \frac{1}{1 + e^{Af+B}}, i \neq j \text{ and } i, j = 1, 2, \dots, CN \quad (13)$$

$$\max_{A,B} \left\{ \sum_i [t_i \ln r_{ij} + (1 - t_i) \ln(1 - r_{ij})] \right\} \quad (14)$$

$$t_i = \begin{cases} t_+ = \frac{N_+ + 1}{N_+ + 2}, y_i = 1 \\ t_- = \frac{1}{N_- + 2}, y_i = -1 \end{cases} \quad (15)$$

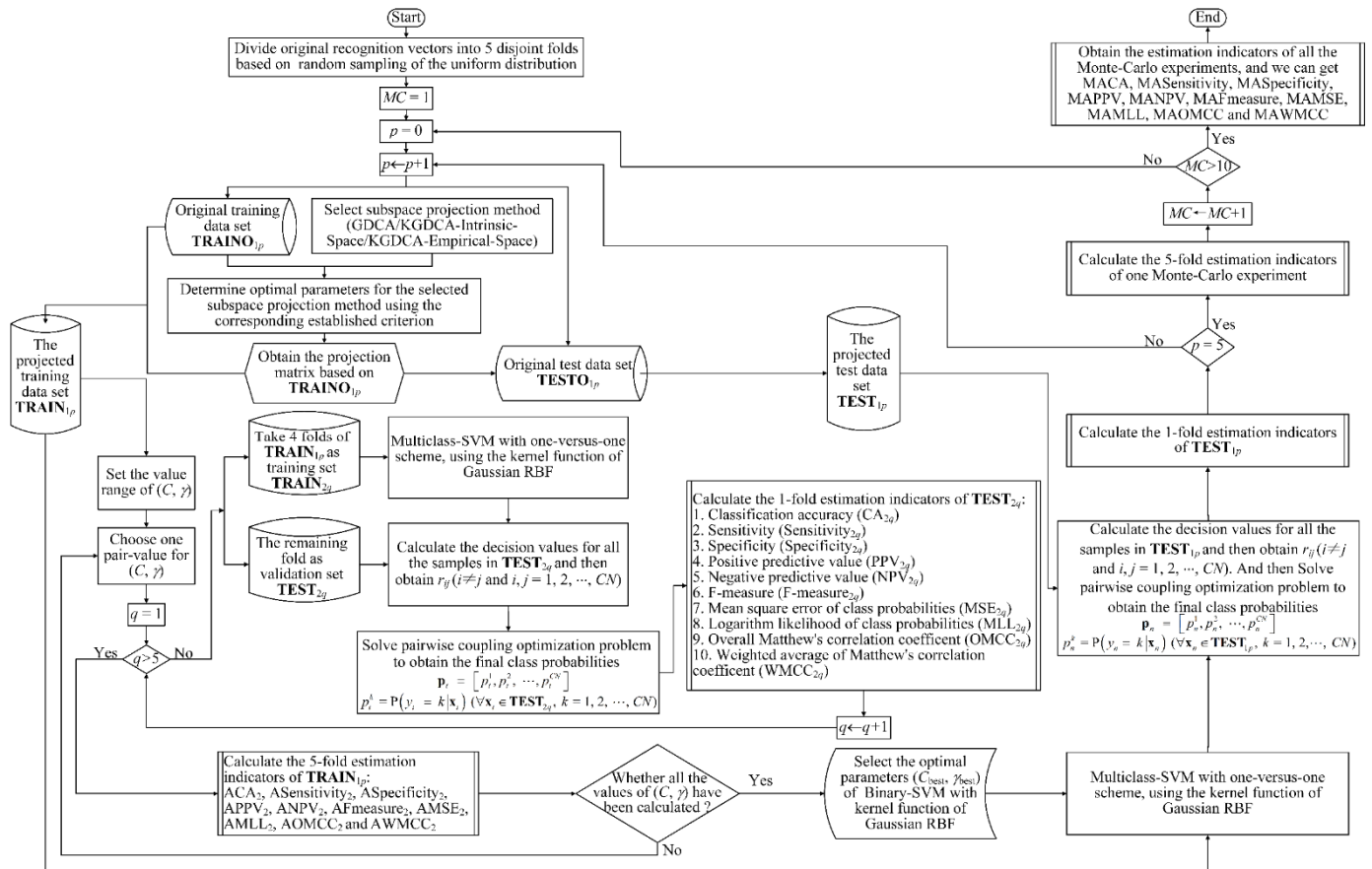


Figure 10. Test strategy of the newly proposed pattern recognition by combining Monte-Carlo experiment with cross-validation.

In addition, in order to ensure the unbiasedness of the decision values used to estimate the parameters  $A$  and  $B$ , all the decision values in Equation (13) are obtained through 5-fold cross-validation of the training data set, which means the decision function is firstly obtained with 4-fold samples, then the decision values of the remaining 1-fold samples are calculated, and we repeat the above process until each training sample has a decision value. For the sample  $x_t$  to be classified, we should firstly obtain its estimated pairwise class probabilities  $r_{ij}$  ( $i \neq j$  and  $i, j = 1, 2, \dots, CN$ ), then the optimization problem based on the pairwise coupling method in Equation (16) [46] is solved to obtain the final class probabilities  $p_t^k = P(y_t = k | x_t)$  ( $k = 1, 2, \dots, CN$ ) of Multiclass-SVM. Finally, according to the maximum posterior probability criterion, the class that maximizes  $p_t^k$  is used as the

predicted class of the test sample. The above pattern recognition can achieve the Bayes optimal decision under the condition of equal cost.

$$\begin{aligned} \min_{\mathbf{P}_t} & \sum_{i=1}^{CN} \sum_{j=1, j \neq i}^{CN} (r_{ji} p_t^i - r_{ij} p_t^j)^2 \\ \text{subject to the constraints:} & \\ & \sum_{k=1}^{CN} p_t^k = 1 \text{ and } \forall k, p_t^k \geq 0 \end{aligned} \quad (16)$$

It can be seen from Figure 10 that there exist ten estimation indicators to evaluate the recognition results [41,47]. By extending two-class estimation indicators of pattern recognition to the multi-class situation using class ratio as weight, we can obtain each 1-fold estimation indicator. Then, each 5-fold estimation indicator can be obtained by averaging the corresponding results of all folds.

### 5.2. Criterion for Selecting Optimal Parameters of GDCA and Its Kernelization Forms

In the actual application of GDCA or its kernelization forms, the optimal  $(\alpha, \delta)$  or  $(\gamma, \alpha, \delta)$  value pair should be determined according to a certain criterion in order to establish the final pattern recognition algorithm, which also shown in Figure 10. In this section, we establish a criterion  $criterionf_m$  as shown in Equation (17), which integrates the three technical indicators, namely  $SNR_i$  ( $i = 1, 2, \dots, \text{rank}(\mathbf{S}_B)$ ),  $OSNR_m$  and  $WOSNR_m$ . In Equation (17),  $h_i$  ( $i = 1, 2, 3$ ) denote the weights, and  $N_a, N_b$  and  $N_c$  denote the number of calculated values of  $\gamma, \alpha$  and  $\delta$ , respectively.

$$\left\{ \begin{aligned} (\gamma_{\text{opt}}, \alpha_{\text{opt}}, \delta_{\text{opt}}) &= \arg \max_{(\gamma, \alpha, \delta)} \left\{ criterionf_m(\gamma, \alpha, \delta) = \sum_{i=1}^3 h_i \bullet TIN_i \right\} \\ TIN_1 &= \frac{N_a N_b N_c \sum_{i=1}^{\text{rank}(\mathbf{S}_B)} SNR_i(\gamma, \alpha, \delta) - \sum_a \sum_b \sum_c \sum_{i=1}^{\text{rank}(\mathbf{S}_B)} SNR_i(\gamma, \alpha, \delta)}{\sqrt{N_a N_b N_c \sum_a \sum_b \sum_c \left[ \sum_{i=1}^{\text{rank}(\mathbf{S}_B)} SNR_i(\gamma, \alpha, \delta) \right]^2 - \left[ \sum_a \sum_b \sum_c \sum_{i=1}^{\text{rank}(\mathbf{S}_B)} SNR_i(\gamma, \alpha, \delta) \right]^2}} \\ TIN_2 &= \frac{N_a N_b N_c \cdot OSNR_m(\gamma, \alpha, \delta) - \sum_a \sum_b \sum_c OSNR_m(\gamma, \alpha, \delta)}{\sqrt{N_a N_b N_c \sum_a \sum_b \sum_c [OSNR_m(\gamma, \alpha, \delta)]^2 - \left[ \sum_a \sum_b \sum_c OSNR_m(\gamma, \alpha, \delta) \right]^2}} \\ TIN_3 &= \frac{N_a N_b N_c \cdot WOSNR_m(\gamma, \alpha, \delta) - \sum_a \sum_b \sum_c WOSNR_m(\gamma, \alpha, \delta)}{\sqrt{N_a N_b N_c \sum_a \sum_b \sum_c [WOSNR_m(\gamma, \alpha, \delta)]^2 - \left[ \sum_a \sum_b \sum_c WOSNR_m(\gamma, \alpha, \delta) \right]^2}} \end{aligned} \right. \quad (17)$$

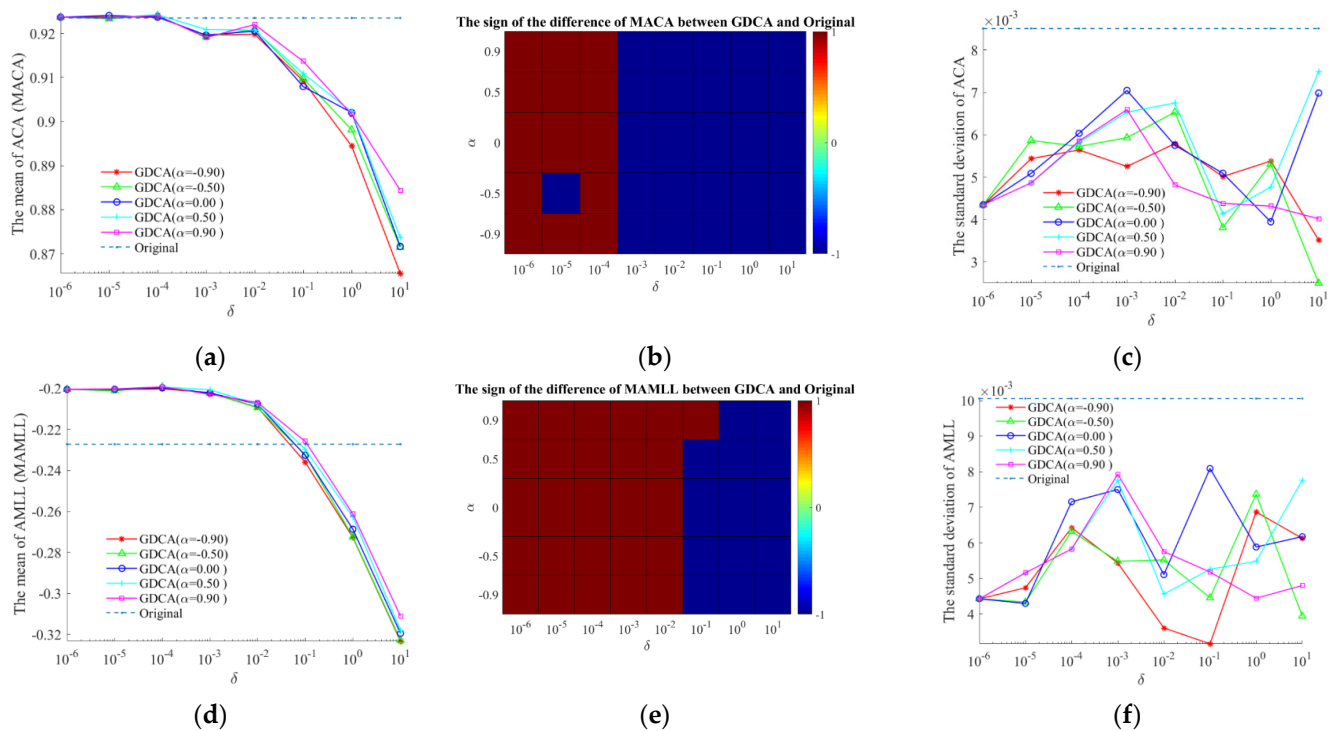
### 5.3. Recognition Effects of GDCA and Its Kernelization Forms Driven SVM

This section gives the recognition effects of GDCA and its kernelization forms driven SVM. Many comparisons are performed, together with the effects of different values of regularization coefficient  $\alpha$  on GDCA and its kernelization forms driven SVM researched.

#### 5.3.1. Recognition Effect of GDCA Driven SVM

According to the flowchart shown in Figure 10, the estimation indicators of all the Monte-Carlo experiments of GDCA driven SVM and original SVM are calculated, and then we obtain the mean of each estimation indicator by averaging 10 Monte-Carlo experiments, which are denoted as MACA, MASensitivity, MASpecificity, MAPPV, MANPV, MAFmeasure, MAMSE, MAMLL, MAOMCC and MAWMCC. The mean and standard deviation of each estimation indicator of all the Monte-Carlo experiments as well as the sign of the difference between GDCA driven SVM (with MMN preprocessing) and original SVM for the mean of each estimation indicator are shown in Figure 11 (only the results of MACA and MAMLL are displayed due to limited space). Results show that there exist values of  $\delta$ , with which GDCA driven SVM outperforms original SVM with regard to all the estimation indicators except MASpecificity under all the values of  $\alpha$ . All the standard deviations of

10 Monte-Carlo experiments for each estimation indicator by GDCA driven SVM are less than those by original SVM, which means GDCA driven SVM is more robust than original SVM. In addition, by maximizing  $criteria_{f_{10}}$  established in Section 5.2, it is derived that  $\alpha_{opt} = 0.9$ ,  $\delta_{opt} = 10^{-4}$ , under which MACA can arrive at the maximum of 92.41%. The consuming time of executing all the Monte-Carlo experiments by GDCA driven SVM has the mean of 403.9399 s and the standard deviation of 53.3158 s, while the time consumed by original SVM is 1697.9388 s. The above time is counted using MATLAB on a personal computer with 2.50 GHz CPU and 16.0 GB RAM.

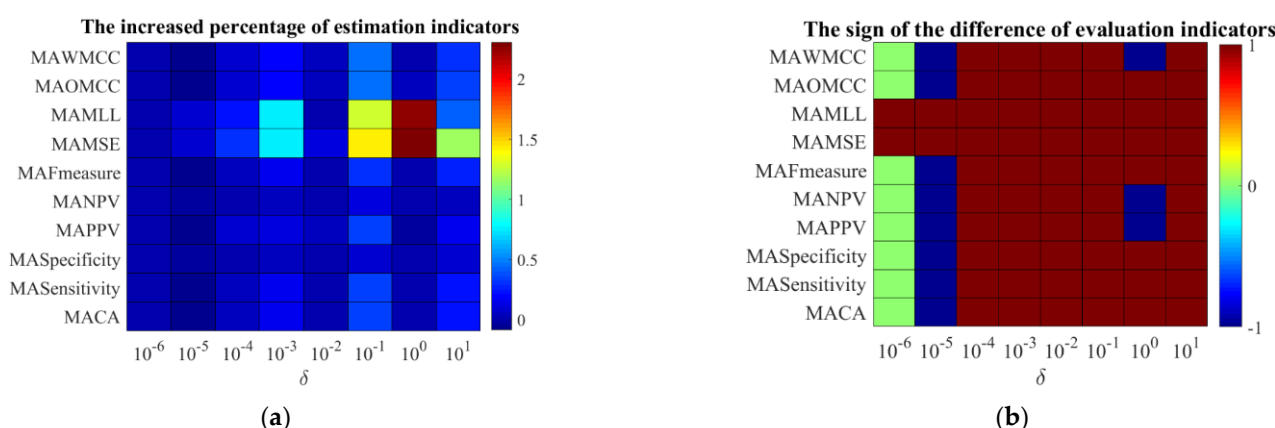


**Figure 11.** Comparisons between GDCA driven SVM and original SVM regarding the mean and standard deviation of all the Monte–Carlo experiments for ACA and AMLL. (a) The mean of ACA (MACA). (b) The sign of the difference of MACA. (c) The standard deviation of ACA. (d) The mean of AMLL (MAMLL). (e) The sign of the difference of MAMLL. (f) The standard deviation of AMLL.

The increased percentage and the sign of estimation indicators comparing GDCA driven SVM with BDCA driven SVM under different values of  $\alpha$  are shown in Figure 12 (only display the case of  $\alpha = 0.5$  due to limited space). Results show that the range of  $\delta$ , in which GDCA outperforms BDCA with regard to all the estimation indicators, generally expands as  $\alpha$  expands. Especially, GDCA ( $0 < \alpha < 1$ ) is superior to BDCA in the majority span of  $\delta$ .

### 5.3.2. Recognition Effect of GDCA’s Kernelization Forms Driven SVM

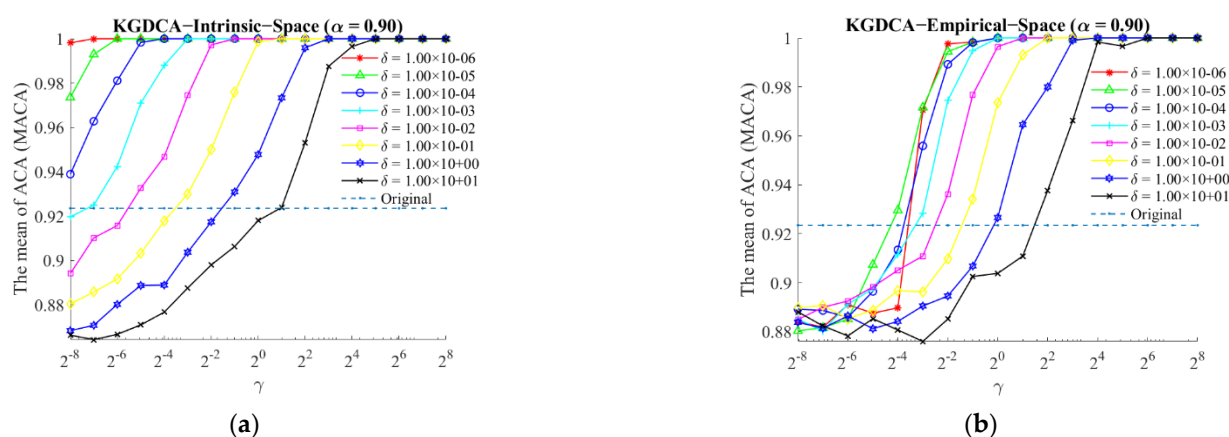
According to the flowchart shown in Figure 10, the estimation indicators of all the Monte-Carlo experiments by KGDCA-Intrinsic-Space- and KGDCA-Empirical-Space driven SVM under different values of  $\gamma$ ,  $\alpha$  and  $\delta$  are calculated. The consuming time of executing all the Monte-Carlo experiments by KGDCA-Intrinsic-Space/KGDCA-Empirical-Space driven SVM has means of 515.5661 s/468.8253 s and standard deviations of 241.1514 s/276.1790 s, respectively.



**Figure 12.** Comparisons between GDCA driven SVM ( $\alpha = 0.5$ ) and BDCA driven SVM regarding all the estimation indicators. (a) The increased percentage. (b) The sign of the difference.

- Comparisons between KGDCA-Intrinsic-Space/KGDCA-Empirical-Space driven SVM and original SVM

Comparisons of the mean of each estimation indicator by averaging all the 10 Monte-Carlo experiments between KGDCA-Intrinsic-Space/KGDCA-Empirical-Space driven SVM and original SVM under different values of  $\alpha$  are shown in Figure 13 (only the cases under  $\alpha = 0.9$  are displayed due to limited space). The results show that there exist large numbers of combinations of  $(\gamma, \alpha, \delta)$ , by which KGDCA-Intrinsic-Space/KGDCA-Empirical-Space driven SVM is superior to original SVM. In addition, the range of  $\gamma$ , in which KGDCA-Intrinsic-Space/KGDCA-Empirical-Space outperforms original SVM, generally expands as  $\delta$  decreases. In addition, by maximizing  $criterion_{f_{10}}$  established in Section 5.2, it is derived that  $\gamma_{opt} = 2^{-3}$ ,  $\alpha_{opt} = 0.9$ ,  $\delta_{opt} = 10^{-5}$  for KGDCA-Intrinsic-Space driven SVM, under which MACA can arrive at the maximum of 100%, meaning the results that all the test samples of all the Monte-Carlo experiments have been classified successfully and  $\gamma_{opt} = 2^{-1}$ ,  $\alpha_{opt} = 0.9$ ,  $\delta_{opt} = 10^{-6}$  for KGDCA-Empirical-Space driven SVM, under which MACA can arrive at 99.83%.

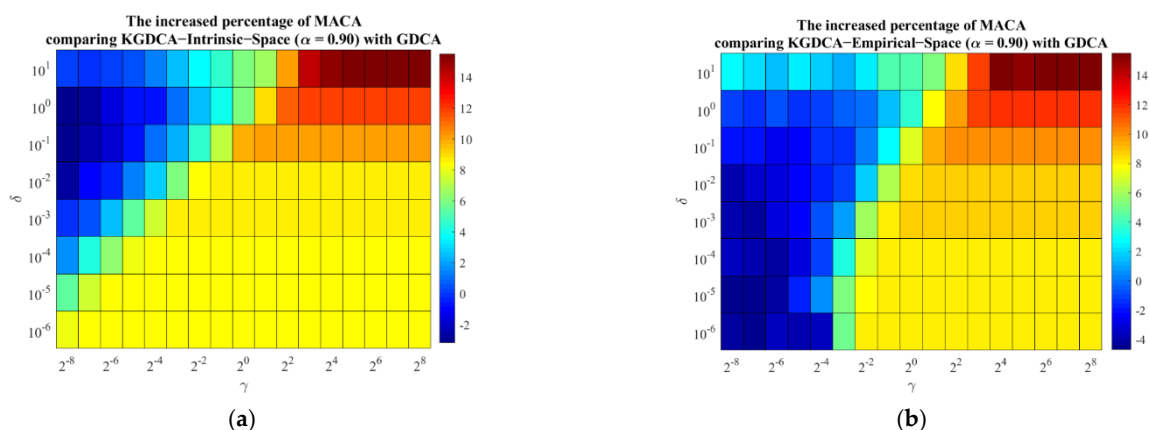


**Figure 13.** Comparisons of MACA between GDCA's kernelization forms driven SVM and original SVM when  $\alpha = 0.9$ . (a) KGDCA–Intrinsic–Space driven SVM; (b) KGDCA–Empirical–Space driven SVM.

- Comparisons between KGDCA-Intrinsic-Space/KGDCA-Empirical-Space driven SVM and GDCA driven SVM

The increased percentages of all the estimation indicators comparing KGDCA-Intrinsic-Space/KGDCA-Empirical-Space driven SVM with GDCA driven SVM under different values of  $\gamma$ ,  $\alpha$  and  $\delta$  are shown in Figure 14 (only the case under  $\alpha = 0.9$  for MACA is

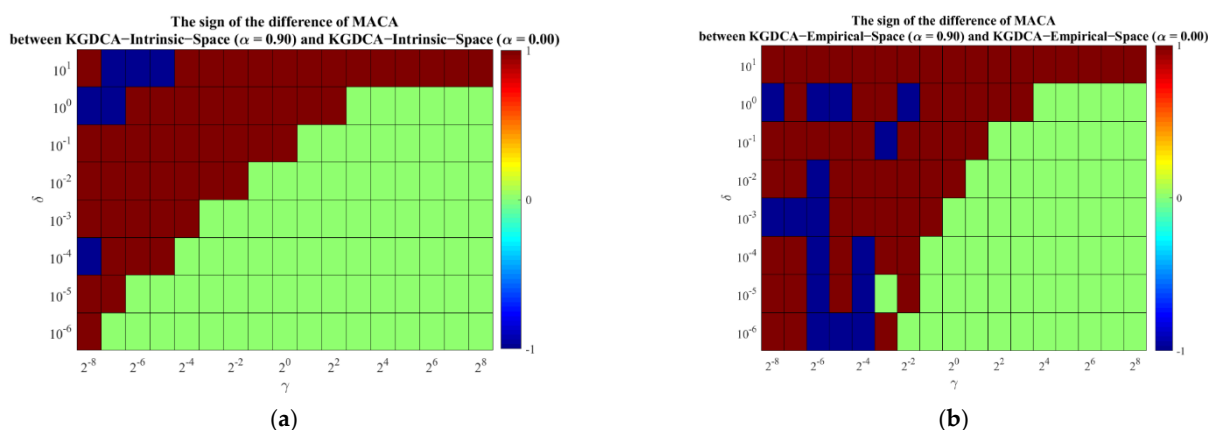
displayed due to limited space). The results show that in the overwhelmingly major combinations of  $\gamma$  and  $\delta$  under all the values of  $\alpha$ , KGDCA-Intrinsic-Space/KGDCA-Empirical-Space driven SVM outperforms GDCA driven SVM. The maximum increased ratios with regard to MACA, MASensitivity, MASpecificity, MAPPV, MANPV, MAFmeasure, MAMSE, MAMLL, MAOMCC and MAWMCC for KGDCA-Intrinsic-Space driven SVM are 15.53%, 15.53%, 5.48%, 15.19%, 4.86%, 15.82%, 99.79%, 94.95%, 21.84% and 22.40%, respectively, while the corresponding ones for KGDCA-Empirical-Space driven SVM are 15.53%, 15.53%, 5.48%, 15.19%, 4.86%, 15.82%, 99.78%, 95.24%, 21.84% and 22.40%, respectively.



**Figure 14.** The increased percentages of MACA comparing GDCA's kernelization forms driven SVM with GDCA driven SVM when  $\alpha = 0.9$ . (a) KGDCA–Intrinsic–Space driven SVM; (b) KGDCA–Empirical–Space driven SVM.

- Effect of  $\alpha$  on KGDCA-Intrinsic-Space/KGDCA-Empirical-Space driven SVM

The relationships describing the sign of the difference of all the estimation indicators between KGDCA-Intrinsic-Space/KGDCA-Empirical-Space ( $0 < \alpha < 1$ ) and KGDCA-Intrinsic-Space/KGDCA-Empirical-Space ( $\alpha = 0$ ) driven SVM varying with the values of  $\gamma$  and  $\delta$  are shown as Figure 15 (only the case under  $\alpha = 0.9$  for MACA is displayed due to limited space). The results show that KGDCA-Intrinsic-Space/KGDCA-Empirical-Space ( $0 < \alpha < 1$ ) driven SVM outperforms KGDCA-Intrinsic-Space/KGDCA-Empirical-Space ( $\alpha = 0$ ) driven SVM in the overwhelmingly major combinations of  $\gamma$  and  $\delta$  except the range in which they are tied.

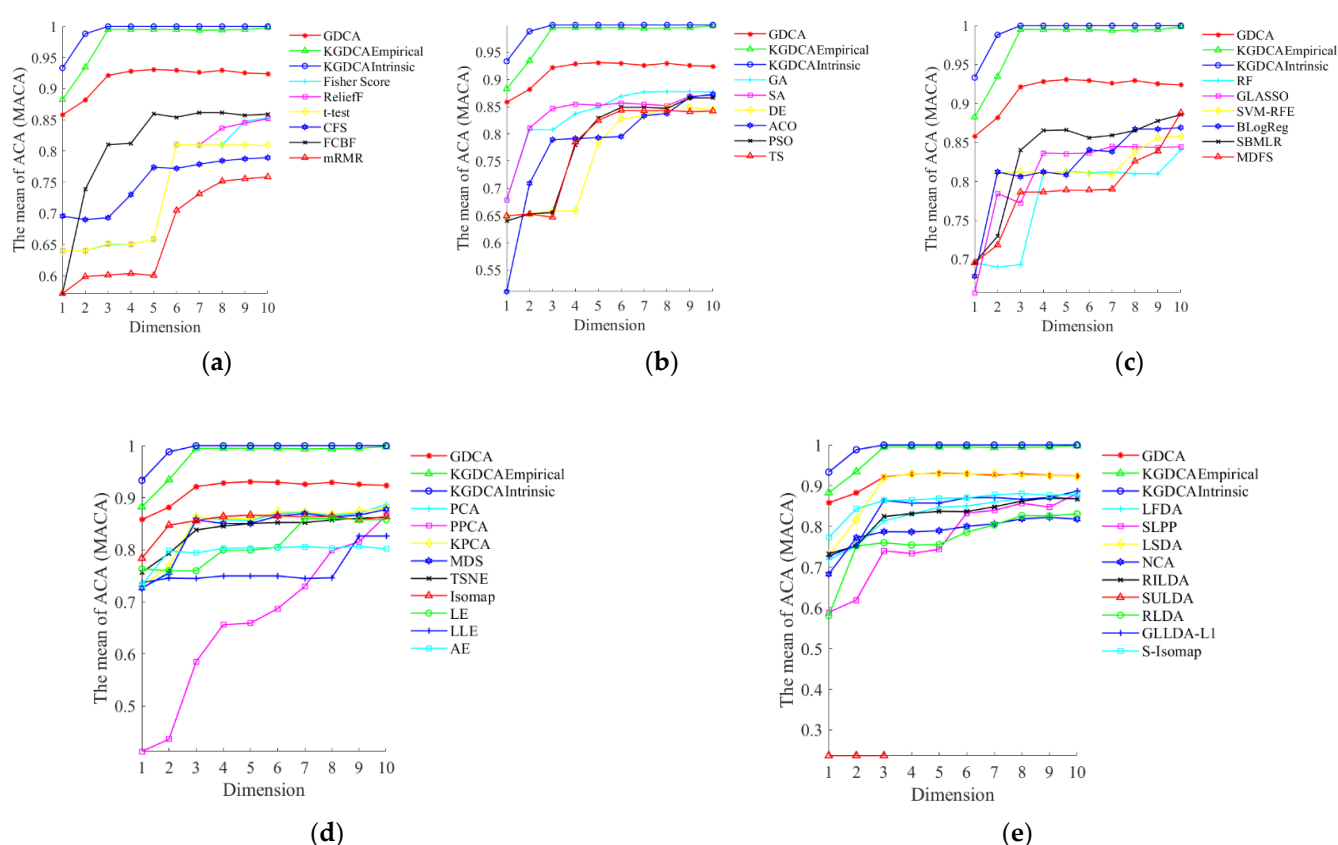


**Figure 15.** The sign of the difference of MACA between KGDCA–Intrinsic–Space/KGDCA–Empirical–Space ( $\alpha = 0.9$ ) driven SVM and KGDCA–Intrinsic–Space/KGDCA–Empirical–Space ( $\alpha = 0$ ) driven SVM under different values of  $\gamma$  and  $\delta$ . (a) KGDCA–Intrinsic–Space driven SVM; (b) KGDCA–Empirical–Space driven SVM.



#### 5.4. Comparisons with Other Dimensionality Reduction Algorithms

In this section, we use the test strategy in Section 5.1 to compare the newly proposed method with 36 kinds of state-of-the-art dimensionality reduction algorithms. Although we calculate 10 estimation indicators to evaluate the recognition results, only results of MACA are shown in Figure 16 due to limited space. It can be seen from the results that the proposed method outperforms all the compared feature selection ones, comprising filter type, wrapper type and embedded type, since the proposed method uses all the information contained in the recognition vectors but feature selection methods inevitably discard some useful information. In the meantime, the wrapper type and embedded type are always classifier-dependent and not only computationally intensive but also at risk of overfitting. In addition, the proposed method also outperforms all the compared unsupervised subspace projection ones, which can be attributed to the fact that the unsupervised subspace projection technology does not involve any class information. Even compared with supervised subspace projection technologies, our proposed method still demonstrates competitive performances with significant advantages.



**Figure 16.** Comparisons with 36 kinds of state-of-the-art dimensionality reduction algorithms for MACA. (a) Feature selection algorithms of filter type. (b) Feature selection algorithms of wrapper type. (c) Feature selection algorithms of embedded type. (d) Unsupervised subspace projection algorithms. (e) Supervised subspace projection algorithms.

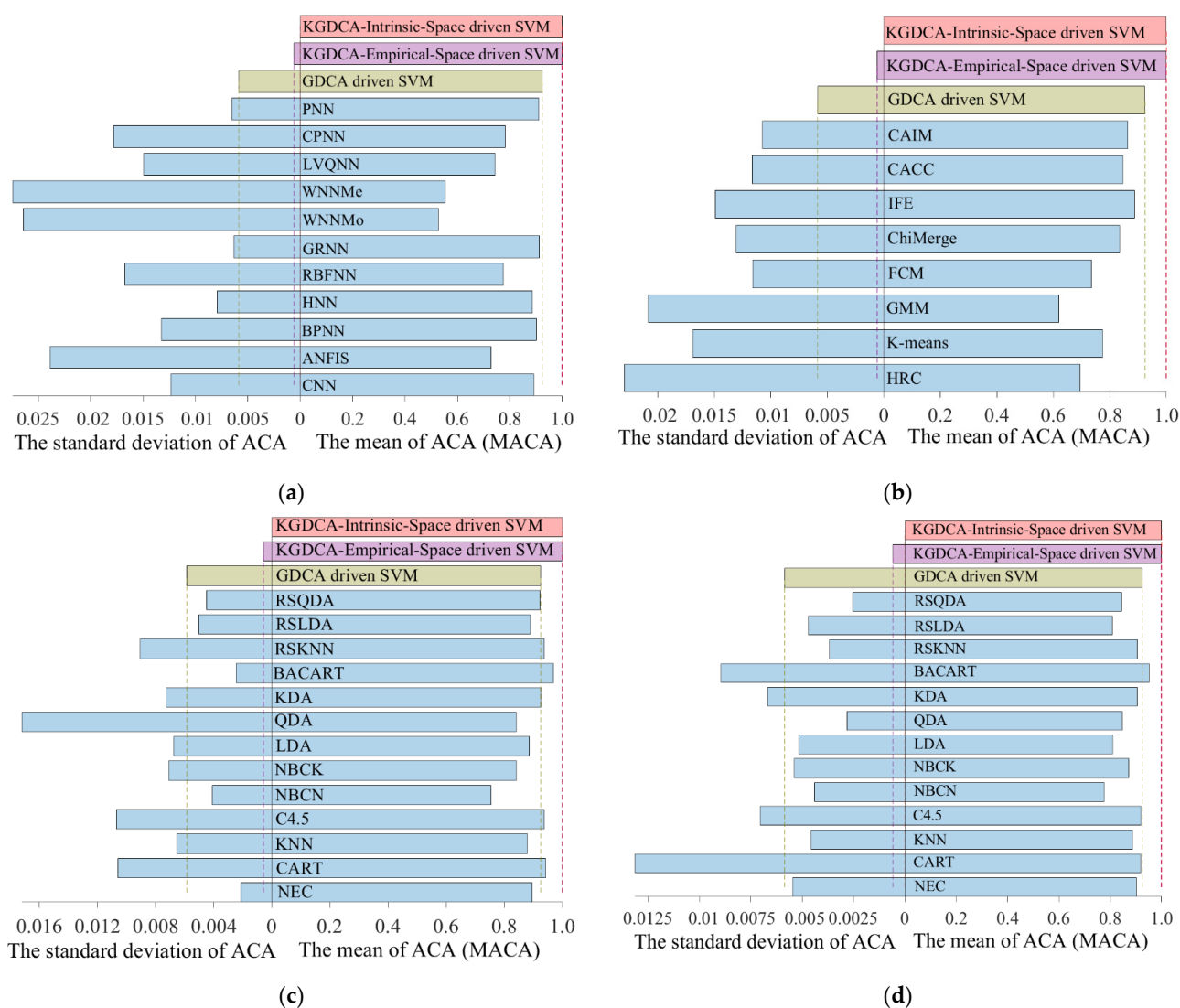
#### 5.5. Comparisons with Other Classifiers

In this section, we use the test strategy in Section 5.1 to compare the newly proposed method with other state-of-the-art classifiers adopted in mainstream pattern recognition methods, composed of ten kinds of neural networks [48], classical rough set (CRS) [49], neighborhood classifier (NEC) [50], K nearest neighbor classifier (KNN), classification and regression tree (CART), C4.5, Naive Bayes classifier (NBC), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), kernelized discriminant analysis (KDA) [4] and ensemble algorithms (including bootstrap aggregating and random subspace ensemble).

bles) [51], as well as their combinations with feature selection (also referred to as attribute reduction) using neighborhood rough set (NRS) [50]. All the involved parameters in the above methods are chosen according to 5-fold cross validation.

### 5.5.1. Comparisons with Ten Kinds of Neural Networks

Ten representative kinds of neural networks, comprising convolutional neural network (CNN), adaptive neuro-fuzzy inference system (ANFIS), back-propagation neural network (BPNN), Hopfield neural network (HNN), radial basis function neural network (RBFNN), generalized regression neural network (GRNN), wavelet neural network (including two cases of Morlet wavelet and Mexican hat wavelet, referred to as WNNMo and WNNMe, respectively), learning vector quantization neural network (LVQNN), counter propagation neural network (CPNN) and probabilistic neural network (PNN), were chosen to make comparisons with the newly proposed pattern method. The results of the mean and standard deviation of all the Monte-Carlo experiments for ACA are shown in Figure 17a, from which it can be seen that both GDCA driven SVM and GDCA's kernelization forms driven SVM are superior to the chosen neural networks.



**Figure 17.** The results of the mean and standard deviation of all the Monte-Carlo experiments for ACA. (a) Comparisons with ten kinds of neural networks. (b) Comparisons with classical rough set. (c) Comparisons with the remaining recognition methods without attribute reduction of NRS. (d) Comparisons with the remaining recognition methods with attribute reduction of NRS.

### 5.5.2. Comparisons with CRS

CRS is specified for discrete features, so it is indispensable to carry out proper discretization of continuous features before using CRS. Eight discretization methods were chosen, comprising four unsupervised algorithms based on hierarchical clustering (HRC), K-means, Gaussian mixing model (GMM) [48], fuzzy C-means clustering (FCM) [52], together with four supervised algorithms based on ChiMerge, information entropy (IFE), class-attribute contingency coefficient (CACC) and class-attribute interdependence maximization (CAIM) [53], the corresponding results of which are shown in Figure 17b, from which it can be seen that both GDCA driven SVM and GDCA's kernelization forms driven SVM are superior to CRS with all the eight discretization methods.

### 5.5.3. Comparisons with the Remaining Classifiers

- Without attribute reduction of NRS

The results of the remaining recognition methods, namely NEC, CART, KNN, C4.5, NBC (considering two cases of normal probability density estimation and kernel density estimation, referred to as NBCN and NBCK, respectively), LDA, QDA, KDA, bootstrap aggregating of CART (BACART) and random subspace ensembles of KNN, LDA and QDA (referred to as RSKNN, RSLDA and RSQDA, respectively), without attribute reduction of NRS, are shown in Figure 17c.

- with attribute reduction of NRS

The results of the remaining recognition methods, namely NEC, CART, KNN, C4.5, NBCN, NBCK, LDA, QDA, KDA, BACART, RSKNN, RSLDA and RSQDA, with attribute reduction of NRS, are shown in Figure 17d.

## 6. Conclusions

By building a set of 220 kV HVDC GIS experiment platform and manufacturing four different types of insulation defects (including multiple sizes and positions), we successfully measured 180,828 pulse current signals under multiple voltage levels. After being denoised, the apparent discharge quantity and the discharge time, two inherent physical quantities unaffected by the experimental platform and measurement system, were obtained, according to which 70 statistical features were extracted. We detailed a pattern recognition method based on generalized discriminant component analysis and its kernelized forms driven SVM, and established the corresponding selection criterion of involved parameters. Combining the Monte-Carlo experimental method with the cross-validation test strategy, 10 evaluation indicators for classification results were calculated. Then, recognition effects of GDCA and its kernelization forms driven SVM including comparisons between each other were analyzed in detail. Finally, comparisons between the newly proposed pattern recognition method and 36 kinds of state-of-the-art dimensionality reduction algorithms together with 44 kinds of state-of-the-art classifiers were performed. The following conclusions can be drawn:

- (1) All the problems of BDCA mentioned in Section 1 can be resolved by GDCA as well as its kernelization forms proposed in this paper. The range of  $\delta$ , in which GDCA outperforms BDCA with regard to all the estimation indicators, generally expands as  $\alpha$  expands. Especially, GDCA ( $0 < \alpha < 1$ ) is superior to BDCA in the majority span of  $\delta$ . In the overwhelmingly major combinations of  $\gamma$  and  $\delta$  under all the values of  $\alpha$ , KGDCA-Intrinsic-Space/KGDCA-Empirical-Space outperformed GDCA.
- (2) By establishing an effective criterion to optimally select the parameters involved in GDCA and its kernelization forms in advance without using the evaluation indicators of classification results, the time of pattern recognition can be shortened considerably to ensure the optimal recognition effect simultaneously.
- (3) The newly proposed pattern recognition method greatly improved the recognition accuracy in comparison with 36 kinds of state-of-the-art dimensionality reduction algorithms and 44 kinds of state-of-the-art classifiers.

Due to the fact that only the apparent discharge quantity and the discharge time, two inherent physical quantities unaffected by the experimental platform and measurement system, are needed, this newly proposed method not only solves the difficulty that phase-resolved partial discharge (PRPD) cannot be applied under DC conditions, but also immensely facilitates the fault diagnosis of HVDC GIS.

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

**Funding:** This research was funded by National Basic Research Program (973 Program), grant number 2014CB239506-2.

**Acknowledgments:** The authors would like to thank the fund and supports derived from 973 Program of “The Failure Evolution Process and Aging Law of HVDC Transmission Pipeline” (No. 2014CB239506-2).

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

## Appendix A

The following gives the rigorous mathematical proof of the proposition that GDCA algorithm does meet the SNR criterion in the signal-subspace and the noise-power criterion in the noise-subspace. The whole proof is consisted of Proposition 1, Proposition 2 and Proposition 3 as follows.

**Proposition A1:** *The rank( $\mathbf{S}_B$ ) larger eigenvalues of GDCA’s discriminant matrix  $(\mathbf{S}_W + \delta\mathbf{I})^{-1}(\mathbf{S}_C + \rho\mathbf{I})$  are larger than 1, the other  $M - \text{rank}(\mathbf{S}_B)$  eigenvalues are smaller than 1, and all the eigenvalues are at least equal to  $\rho/\delta$ .*

### ■ Proof of Proposition A1

Firstly, we provide the evidence that the rank( $\mathbf{S}_B$ ) larger eigenvalues of GDCA’s discriminant matrix  $(\mathbf{S}_W + \delta\mathbf{I})^{-1}(\mathbf{S}_C + \rho\mathbf{I})$  are larger than 1, and the other  $M - \text{rank}(\mathbf{S}_B)$  eigenvalues are smaller than 1.

Since  $\mathbf{S}_W$  is a real-symmetric positive semi-definite matrix and  $\delta \rightarrow 0^+$ ,  $\mathbf{S}_W + \delta\mathbf{I}$  must be a real-symmetric positive definite matrix. According to the Cholesky decomposition theorem,  $\mathbf{S}_W + \delta\mathbf{I}$  can be expressed as the product of a lower triangular matrix  $\mathbf{L}_{WI}$  whose diagonal elements are all positive and its transpose, shown as Equation (A1):

$$\mathbf{S}_W + \delta\mathbf{I} = \mathbf{L}_{WI} \bullet \mathbf{L}_{WI}^T \quad (\text{A1})$$

Let  $\Sigma_{GDCA}$  be a diagonal matrix whose diagonal elements are consisted of the rank( $\mathbf{S}_B$ ) larger eigenvalues of  $(\mathbf{S}_W + \delta\mathbf{I})^{-1}(\mathbf{S}_C + \rho\mathbf{I})$  arranged in descending order and the  $M - \text{rank}(\mathbf{S}_B)$  smaller eigenvalues arranged in ascending order. Then, Equation (A2) can be derived ( $m = M$  here).

$$\begin{aligned} (\mathbf{S}_W + \delta\mathbf{I})^{-1}(\mathbf{S}_C + \rho\mathbf{I}) \cdot \mathbf{W}_{GDCA} &= \mathbf{W}_{GDCA} \cdot \Sigma_{GDCA} \\ \Leftrightarrow (\mathbf{S}_W + \delta\mathbf{I})^{-1}[\mathbf{S}_B + (\rho - \delta)\mathbf{I}] \cdot \mathbf{W}_{GDCA} &= \mathbf{W}_{GDCA} \cdot (\Sigma_{GDCA} - \mathbf{I}) \\ \Leftrightarrow (\mathbf{L}_{WI} \bullet \mathbf{L}_{WI}^T)^{-1}[\mathbf{S}_B + (\rho - \delta)\mathbf{I}] \cdot \mathbf{W}_{GDCA} &= \mathbf{W}_{GDCA} \cdot (\Sigma_{GDCA} - \mathbf{I}) \\ \Leftrightarrow \mathbf{L}_{WI}^{-1}[\mathbf{S}_B + (\rho - \delta)\mathbf{I}] \cdot \mathbf{W}_{GDCA} &= \mathbf{L}_{WI}^T \mathbf{W}_{GDCA} \cdot (\Sigma_{GDCA} - \mathbf{I}) \end{aligned} \quad (\text{A2})$$

We further define a matrix  $\mathbf{H}_I$  as Equation (A3):

$$\mathbf{H}_I = \mathbf{L}_{WI}^T \mathbf{W}_{GDCA} \quad (\text{A3})$$

Substituting Equation (A3) into Equation (A2) can obtain Equation (A4):

$$\begin{aligned} \mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}] \cdot \mathbf{W}_{\text{GDCA}} &= \mathbf{L}_{\mathbf{W}\mathbf{I}}^{\text{T}} \mathbf{W}_{\text{GDCA}} \cdot (\Sigma_{\text{GDCA}} - \mathbf{I}) \\ \Leftrightarrow \mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}] \left( \mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1} \right)^{\text{T}} \cdot \mathbf{H}_{\mathbf{I}} &= \mathbf{H}_{\mathbf{I}} \cdot (\Sigma_{\text{GDCA}} - \mathbf{I}) \end{aligned} \quad (\text{A4})$$

It can be easily seen from Equations (A2) and (A4) that the eigenvalues of  $(\mathbf{S}_{\mathbf{W}} + \delta\mathbf{I})^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}]$  and  $\mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}] \left( \mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1} \right)^{\text{T}}$  are totally identical. Due to the fact that  $\text{rank}(\mathbf{S}_{\mathbf{B}}) \leq CN - 1$  and in general  $CN - 1 < M$ ,  $\mathbf{S}_{\mathbf{B}}$  is rank-deficient. Because  $\mathbf{S}_{\mathbf{B}}$  is a real-symmetric positive semi-definite matrix,  $\mathbf{S}_{\mathbf{B}}$  has  $\text{rank}(\mathbf{S}_{\mathbf{B}})$  eigenvalues greater than 0 and  $M - \text{rank}(\mathbf{S}_{\mathbf{B}})$  repeated eigenvalues 0. Furthermore, based on the condition that  $\delta > \rho$ ,  $\delta \rightarrow 0^+$  and  $\rho \rightarrow 0$ , it can be deduced that  $\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}$  has  $\text{rank}(\mathbf{S}_{\mathbf{B}})$  eigenvalues larger than 0 and  $M - \text{rank}(\mathbf{S}_{\mathbf{B}})$  repeated negative eigenvalues  $\rho - \delta$  (only requiring that  $\delta - \rho$  should be less than the smallest positive eigenvalue of  $\mathbf{S}_{\mathbf{B}}$ ). Additionally,  $\mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}] \left( \mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1} \right)^{\text{T}}$  and  $\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}$  are congruent, which means  $\mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}] \left( \mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1} \right)^{\text{T}}$  and  $\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}$  have the same positive and negative inertia indices. Thus,  $\mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}] \left( \mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1} \right)^{\text{T}}$  also has  $\text{rank}(\mathbf{S}_{\mathbf{B}})$  eigenvalues larger than 0 and  $M - \text{rank}(\mathbf{S}_{\mathbf{B}})$  eigenvalues smaller than 0. It can be seen from Equation (A4) that the diagonal elements of  $\Sigma_{\text{GDCA}} - \mathbf{I}_{M \times M}$  are just the eigenvalues of  $\mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}] \left( \mathbf{L}_{\mathbf{W}\mathbf{I}}^{-1} \right)^{\text{T}}$ , so the  $\text{rank}(\mathbf{S}_{\mathbf{B}})$  larger eigenvalues of GDCA's discriminant matrix  $(\mathbf{S}_{\mathbf{W}} + \delta\mathbf{I})^{-1}(\mathbf{S}_{\mathbf{C}} + \rho\mathbf{I})$  are larger than 1, and the other  $M - \text{rank}(\mathbf{S}_{\mathbf{B}})$  eigenvalues are smaller than 1.

Secondly, we prove all the eigenvalues of  $(\mathbf{S}_{\mathbf{W}} + \delta\mathbf{I})^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}]$  are not less than  $\rho/\delta$ . The proof can be performed by contradiction, and the details are as follows:

Assume that there exists a real  $\lambda$  smaller than  $\rho/\delta$  and  $\lambda$  is an eigenvalue of  $(\mathbf{S}_{\mathbf{W}} + \delta\mathbf{I})^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}]$ , which corresponds to the eigenvector  $\mathbf{v}$ . Then, Equation (A5) can be derived. It can be deduced from  $\rho < \delta$  and  $\lambda < \rho/\delta$  that  $1 - \lambda > 0$  and  $\rho - \lambda\delta > 0$ . Because of the fact that both  $\mathbf{S}_{\mathbf{B}}$  and  $\mathbf{S}_{\mathbf{W}}$  are real-symmetric positive semi-definite matrices as well as the fact that  $(\rho - \lambda\delta)\mathbf{I}$  is a real-symmetric positive definite matrix,  $\mathbf{S}_{\mathbf{B}} + (1 - \lambda)\mathbf{S}_{\mathbf{W}} + (\rho - \lambda\delta)\mathbf{I}$  is actually a real-symmetric positive definite matrix, which is contradictory to Equation (A5). Therefore, all the eigenvalues of  $(\mathbf{S}_{\mathbf{W}} + \delta\mathbf{I})^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}]$  are not less than  $\rho/\delta$ . Particularly, when  $\mathbf{S}_{\mathbf{B}} + (1 - \rho/\delta)\mathbf{S}_{\mathbf{W}}$  is a singular matrix, the smallest eigenvalue of  $(\mathbf{S}_{\mathbf{W}} + \delta\mathbf{I})^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}]$  is exactly equal to  $\rho/\delta$ .  $\square$

$$\begin{aligned} (\mathbf{S}_{\mathbf{W}} + \delta\mathbf{I})^{-1}(\mathbf{S}_{\mathbf{C}} + \rho\mathbf{I})\mathbf{v} &= \lambda\mathbf{v} \\ \Rightarrow |\mathbf{S}_{\mathbf{C}} + \rho\mathbf{I} - \lambda(\mathbf{S}_{\mathbf{W}} + \delta\mathbf{I})| &= 0 \\ \Rightarrow |\mathbf{S}_{\mathbf{B}} + (1 - \lambda)\mathbf{S}_{\mathbf{W}} + (\rho - \lambda\delta)\mathbf{I}| &= 0 \end{aligned} \quad (\text{A5})$$

**Proposition A2:** SNRs of projected components obtained by the signal-subspace projection matrix  $\mathbf{W}_{\text{PS}}$  are arranged in descending order.

### ■ Proof of Proposition A2

Let  $\Sigma_{\text{PS}}$  be a diagonal matrix whose diagonal elements are consisted of the  $\text{rank}(\mathbf{S}_{\mathbf{B}})$  larger eigenvalues  $\lambda_i$  ( $i = 1, 2, \dots, \text{rank}(\mathbf{S}_{\mathbf{B}})$ ) of  $(\mathbf{S}_{\mathbf{W}} + \delta\mathbf{I})^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}]$  arranged in the descending order. Denote the  $i$ th projection vector of  $\mathbf{W}_{\text{PS}}$  as  $\mathbf{w}_i$  ( $i = 1, 2, \dots, \text{rank}(\mathbf{S}_{\mathbf{B}})$ ). It can be deduced from Proposition 1 that  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{\text{rank}(\mathbf{S}_{\mathbf{B}})} > 1$ . Firstly, Equation (A6) can be derived.

$$\begin{aligned} (\mathbf{S}_{\mathbf{W}} + \delta\mathbf{I})^{-1}(\mathbf{S}_{\mathbf{C}} + \rho\mathbf{I})\mathbf{W}_{\text{PS}} &= \mathbf{W}_{\text{PS}}\Sigma_{\text{PS}} \\ \Leftrightarrow (\mathbf{S}_{\mathbf{W}} + \delta\mathbf{I})^{-1}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}]\mathbf{W}_{\text{PS}} &= \mathbf{W}_{\text{PS}}(\Sigma_{\text{PS}} - \mathbf{I}) \\ \Leftrightarrow \mathbf{W}_{\text{PS}}^{\text{T}}[\mathbf{S}_{\mathbf{B}} + (\rho - \delta)\mathbf{I}]\mathbf{W}_{\text{PS}} &= \mathbf{W}_{\text{PS}}^{\text{T}}(\mathbf{S}_{\mathbf{W}} + \delta\mathbf{I})\mathbf{W}_{\text{PS}}(\Sigma_{\text{PS}} - \mathbf{I}) \end{aligned} \quad (\text{A6})$$



Moreover, Equation (A7) can be obtained.

$$\begin{cases} \mathbf{w}_i^T [\mathbf{S}_B + (\rho - \delta) \mathbf{I}] \mathbf{w}_i = (\lambda_i - 1) \mathbf{w}_i^T (\mathbf{S}_W + \delta \mathbf{I}) \mathbf{w}_i \\ \text{for } i = 1, 2, \dots, \text{rank}(\mathbf{S}_B) \end{cases} \quad (\text{A7})$$

It can be seen from Equation (8) that  $\mathbf{W}_{\text{GDCA}} = [\mathbf{W}_{\text{PS}}, \mathbf{W}_{\text{PN}}]$  satisfies the constraint shown as Equation (A8), so  $\mathbf{W}_{\text{GDCA}}$  can be decomposed as Equation (A9), where  $\mathbf{V}_{\text{GDCA}} \mathbf{\Lambda}_{\text{GDCA}} \mathbf{V}_{\text{GDCA}}^T$  is the spectral decomposition of  $\mathbf{S}_W + \delta \mathbf{I}$ .

$$\mathbf{W}_{\text{GDCA}}^T (\mathbf{S}_W + \delta \mathbf{I}) \mathbf{W}_{\text{GDCA}} = \mathbf{I} \quad (\text{A8})$$

$$\begin{cases} \mathbf{W}_{\text{GDCA}} = \mathbf{V}_{\text{GDCA}} \mathbf{\Lambda}_{\text{GDCA}}^{-\frac{1}{2}} \mathbf{U}_{\text{GDCA}} \\ \mathbf{U}_{\text{GDCA}}^T \mathbf{U}_{\text{GDCA}} = \mathbf{I}_{m \times m} \\ \mathbf{U}_{\text{GDCA}} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m] \end{cases} \quad (\text{A9})$$

Moreover,

$$\mathbf{W}_{\text{GDCA}}^T \mathbf{W}_{\text{GDCA}} = \mathbf{U}_{\text{GDCA}}^T \mathbf{\Lambda}_{\text{GDCA}}^{-1} \mathbf{U}_{\text{GDCA}} \quad (\text{A10})$$

Let  $\mathbf{\Lambda}_{\text{GDCA}}$  be a diagonal matrix whose diagonal elements are consisted of all the eigenvalues  $\mu_i$  ( $i = 1, 2, \dots, M$ ) of  $\mathbf{S}_W + \delta \mathbf{I}$  arranged in the descending order. Due to the fact that  $\mathbf{S}_W$  is a real-symmetric positive semi-definite matrix and in general must have at least one positive eigenvalue,  $\mu_1 \geq \mu_2 \geq \dots \geq \mu_M \geq \delta > 0$  and  $\exists i \in [1, M]$  so as to make  $\mu_i$  larger than  $\delta$ . Therefore, Equation (A11) can be derived from Equation (A10).

$$\begin{cases} \frac{1}{\mu_1} \leq \|\mathbf{w}_i\|^2 = \sum_{j=1}^M \frac{1}{\mu_j} u_{ji}^2 < \frac{1}{\delta} \\ i = 1, 2, \dots, m \end{cases} \quad (\text{A11})$$

Combining Equations (A7) and (A8), Equation (A12) can be obtained.

$$\begin{cases} \text{SNR}_i(\delta, \rho) = \frac{\mathbf{w}_i^T \mathbf{S}_B \mathbf{w}_i}{\mathbf{w}_i^T \mathbf{S}_W \mathbf{w}_i} \\ = \frac{(\lambda_i - 1) \mathbf{w}_i^T (\mathbf{S}_W + \delta \mathbf{I}) \mathbf{w}_i - (\rho - \delta) \|\mathbf{w}_i\|^2}{\mathbf{w}_i^T \mathbf{S}_W \mathbf{w}_i} \\ = \lambda_i - 1 + \frac{(\delta \lambda_i - \rho) \|\mathbf{w}_i\|^2}{1 - \delta \|\mathbf{w}_i\|^2} \\ = \lambda_i \left( 1 + \frac{\delta - \rho / \lambda_i}{\|\mathbf{w}_i\|^{-2} - \delta} \right) - 1, \quad i = 1, 2, \dots, \text{rank}(\mathbf{S}_B) \end{cases} \quad (\text{A12})$$

Based on the fact that  $\delta > \rho$ ,  $\delta \rightarrow 0^+$  and  $\rho \rightarrow 0$ , Equation (A13) can be deduced from Equations (A11) and (A12).

$$\begin{cases} \text{SNR}_i \approx \lim_{\delta \rightarrow 0^+, \rho \rightarrow 0} \text{SNR}_i(\delta, \rho) \\ = \lim_{\delta \rightarrow 0^+, \rho \rightarrow 0} \left[ \lambda_i \left( 1 + \frac{\delta - \rho / \lambda_i}{\|\mathbf{w}_i\|^{-2} - \delta} \right) - 1 \right] \\ = \lambda_i - 1, \quad i = 1, 2, \dots, \text{rank}(\mathbf{S}_B) \end{cases} \quad (\text{A13})$$

Because  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{\text{rank}(\mathbf{S}_B)} > 1$ , it can be seen from Equations (A12) and (A13) that SNRs of projected components obtained by the signal-subspace projection matrix  $\mathbf{W}_{\text{PS}}$  are all larger than zero and arranged in descending order.  $\square$

**Proposition A3:** The noise powers of projected components obtained by the noise-subspace projection matrix  $\mathbf{W}_{\text{PN}}$  are arranged in ascending order.

### ■ Proof of Proposition A3

Let  $\mathbf{\Sigma}_{\text{PN}}$  be a diagonal matrix whose diagonal elements are consisted of the  $m - \text{rank}(\mathbf{S}_B)$  smaller eigenvalues  $\lambda_i$  ( $i = M, M - 1, \dots, M - m + \text{rank}(\mathbf{S}_B) + 1$ ) of  $(\mathbf{S}_W + \delta \mathbf{I})^{-1} [\mathbf{S}_B + (\rho - \delta) \mathbf{I}]$

arranged in the ascending order. Correspondingly,  $\mathbf{W}_{\text{PN}} = [\mathbf{w}_{\text{rank}(\mathbf{S}_B)+1}, \mathbf{w}_{\text{rank}(\mathbf{S}_B)+2}, \dots, \mathbf{w}_m]$ . It can be deduced from Proposition 1 that  $\lambda_M \leq \lambda_{M-1} \leq \dots \leq \lambda_{M-m+\text{rank}(\mathbf{S}_B)+1} < 1$ .

It can be seen from Equation (8) that  $\mathbf{W}_{\text{PN}}$  satisfies the constraint shown as Equation (A14). Furthermore, Equation (A15) can be obtained.

$$\mathbf{W}_{\text{PN}}^T \mathbf{S}_B \mathbf{W}_{\text{PN}} = 0 \quad (\text{A14})$$

$$\begin{aligned} \mathbf{W}_{\text{PN}}^T \mathbf{S}_W \mathbf{W}_{\text{PN}} &= \mathbf{W}_{\text{PN}}^T (\mathbf{S}_W + \mathbf{S}_B) \mathbf{W}_{\text{PN}} \\ &= \mathbf{W}_{\text{PN}}^T (\mathbf{S}_C + \rho \mathbf{I} - \rho \mathbf{I}) \mathbf{W}_{\text{PN}} \\ &= \mathbf{W}_{\text{PN}}^T (\mathbf{S}_C + \rho \mathbf{I}) \mathbf{W}_{\text{PN}} - \rho \mathbf{W}_{\text{PN}}^T \mathbf{W}_{\text{PN}} \end{aligned} \quad (\text{A15})$$

Combined with Equation (9), Equation (A15) can be further transformed into Equation (A16).

$$\mathbf{W}_{\text{PN}}^T \mathbf{S}_W \mathbf{W}_{\text{PN}} = \mathbf{W}_{\text{PN}}^T (\mathbf{S}_W + \delta \mathbf{I}) \mathbf{W}_{\text{PN}} \Sigma_{\text{PN}} - \rho \mathbf{W}_{\text{PN}}^T \mathbf{W}_{\text{PN}} \quad (\text{A16})$$

Moreover, Equation (A16) can be transformed into Equation (A17).

$$\mathbf{W}_{\text{PN}}^T \mathbf{S}_W \mathbf{W}_{\text{PN}} \cdot (\mathbf{I} - \Sigma_{\text{PN}}) = \mathbf{W}_{\text{PN}}^T \mathbf{W}_{\text{PN}} (\delta \Sigma_{\text{PN}} - \rho \mathbf{I}) \quad (\text{A17})$$

Thus, Equation (A18) can be easily derived from Equation (A17).

$$\begin{cases} (1 - \lambda_{M-i+1}) \mathbf{w}_{\text{rank}(\mathbf{S}_B)+i}^T \mathbf{S}_W \mathbf{w}_{\text{rank}(\mathbf{S}_B)+i} = (\delta \lambda_{M-i+1} - \rho) \|\mathbf{w}_{\text{rank}(\mathbf{S}_B)+i}\|^2 \\ \text{for } i = 1, 2, \dots, m - \text{rank}(\mathbf{S}_B) \end{cases} \quad (\text{A18})$$

Finally, Equation (A19) can be derived.

$$\begin{cases} \text{NoisePower}_i = \frac{\mathbf{w}_i^T \mathbf{S}_W \mathbf{w}_i}{\|\mathbf{w}_i\|^2} \\ = \frac{\delta \lambda_{M+\text{rank}(\mathbf{S}_B)+1-i} - \rho}{1 - \lambda_{M+\text{rank}(\mathbf{S}_B)+1-i}} \\ = -\delta + \frac{\delta - \rho}{1 - \lambda_{M+\text{rank}(\mathbf{S}_B)+1-i}} \\ \text{for } i = \text{rank}(\mathbf{S}_B) + 1, \text{rank}(\mathbf{S}_B) + 2, \dots, m \end{cases} \quad (\text{A19})$$

It is noted from Equation (A19) that projection vectors should be normalized by their own 2-norm in order to avoid the influence of projection vectors' norm on the noise power. On account of the fact that  $\delta > \rho$  and  $\rho/\delta \leq \lambda_M \leq \lambda_{M-1} \leq \dots \leq \lambda_{M-m+\text{rank}(\mathbf{S}_B)+1} < 1$ , it can be seen from Equation (A19) that the noise powers of projected components obtained by the noise-subspace projection matrix  $\mathbf{W}_{\text{PN}}$  are arranged in ascending order and non-negative.  $\square$

## References

- Wenger, P.; Beltle, M.; Tenbohlen, S.; Riechert, U.; Behrmann, G. Combined characterization of free-moving particles in HVDC-GIS using UHF PD, high-speed imaging, and pulse-sequence analysis. *IEEE Trans. Power Deliv.* **2019**, *34*, 1540–1548. [\[CrossRef\]](#)
- Magier, T.; Tenzer, M.; Koch, H. Direct current gas-insulated transmission lines. *IEEE Trans. Power Deliv.* **2018**, *33*, 440–446. [\[CrossRef\]](#)
- Ridder, D.D.; Tax, D.M.J.; Lei, B.; Xu, G.; Feng, M.; Zou, Y.; van der Heijden, F. *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2017.
- Kung, S.Y. *Kernel Methods and Machine Learning*; Cambridge University Press: Cambridge, UK, 2014.
- Hira, Z.M.; Gillies, D.F. A review of feature selection and feature extraction methods applied on microarray data. *Adv. Bioinform.* **2015**, *2015*, 198363. [\[CrossRef\]](#)
- Kis, K.B.; Fodor, Á.; Büki, M.I. Adaptive, Hybrid Feature Selection (AHFS). *Pattern Recognit.* **2021**, *116*, 107932.
- Yun, L.; Tao, L.; Liu, H. Recent advances in feature selection and its applications. *Knowl. Inf. Syst.* **2017**, *53*, 551–577.
- Liu, J.; Lin, Y.; Lin, M.; Wu, S.; Zhang, J. Feature selection based on quality of information. *Neurocomputing* **2017**, *225*, 11–22. [\[CrossRef\]](#)
- Meenachi, L.; Ramakrishnan, S. Metaheuristic Search Based Feature Selection Methods for Classification of Cancer. *Pattern Recognit.* **2021**, *119*, 108079. [\[CrossRef\]](#)

10. Zini, L.; Noceti, N.; Fusco, G.; Odone, F. Structured multi-class feature selection with an application to face recognition. *Pattern Recognit. Lett.* **2015**, *55*, 35–41. [\[CrossRef\]](#)
11. Cawley, G.C.; Talbot, N.L.C. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics* **2006**, *22*, 2348–2355. [\[CrossRef\]](#)
12. Bernhard, S.; John, P.; Thomas, H. Sparse Multinomial Logistic Regression via Bayesian L1 Regularization. *Adv. Neural Inf. Process. Syst.* **2017**, *19*, 209–216.
13. Zhang, J.; Luo, Z.; Li, C.; Zhou, C.; Li, S. Manifold regularized discriminative feature selection for multi-label learning. *Pattern Recognit.* **2019**, *95*, 136–150. [\[CrossRef\]](#)
14. Su, B.; Ding, X.; Wang, H.; Wu, Y. Discriminative dimensionality reduction for multi-dimensional sequences. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 77–91. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Yang, J.; Frangi, A.F.; Yang, J.Y.; Zhang, D.; Jin, Z. KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 230–244. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Tipping, M.; Bishop, C. Probabilistic principal component analysis. *J. R. Stat. Soc.* **1999**, *61*, 611–622. [\[CrossRef\]](#)
17. Kruskal, J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **1964**, *29*, 230–244. [\[CrossRef\]](#)
18. Wu, M.; Cao, H.; Cao, J.; Nguyen, H.L.; Gomes, J.B.; Krishnaswamy, S.P. An overview of state-of-the-art partial discharge analysis techniques for condition monitoring. *IEEE Electr. Insul. Mag.* **2015**, *31*, 22–35. [\[CrossRef\]](#)
19. Roweis, S.T.; Saul, L.K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2000**, *290*, 2323–2326. [\[CrossRef\]](#)
20. Tenenbaum, J.B.; de Silva, V.; Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319–2323. [\[CrossRef\]](#)
21. Sundaresan, A.; Chellappa, R. Model Driven Segmentation of Articulating Humans in Laplacian Eigenspace. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1771–1785. [\[CrossRef\]](#)
22. Li, Y.; Chai, Y.; Zhou, H.; Yin, H. A novel dimension reduction and dictionary learning framework for high-dimensional data classification. *Pattern Recognit.* **2021**, *112*, 107793. [\[CrossRef\]](#)
23. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2001.
24. Goldberger, J.; Hinton, G.; Roweis, S.; Salakhutdinov, R. Neighborhood Components Analysis. *Adv. Neural Inf. Process. Syst.* **2005**, *17*, 513–520.
25. Masoudimansour, W.; Bouguila, N. Supervised dimensionality reduction of proportional data using mixture estimation. *Pattern Recognit.* **2020**, *105*, 107379. [\[CrossRef\]](#)
26. Bian, W.; Tao, D. Asymptotic Generalization Bound of Fisher’s Linear Discriminant Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2325–2337. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Yu, Y.; McKelvey, T.; Kung, S.Y. A classification scheme for ‘high-dimensional-small-sample-size’ data using soda and ridge-SVM with microwave measurement applications. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 36–31 May 2013; pp. 3542–3546.
28. Rahulamathavan, Y.; Phan, R.C.; Chambers, J.A.; Parish, D.J. Facial Expression Recognition in the Encrypted Domain Based on Local Fisher Discriminant Analysis. *IEEE Trans. Affect. Comput.* **2013**, *4*, 83–92. [\[CrossRef\]](#)
29. Lai, Z.; Xu, Y.; Yang, J.; Shen, L.; Zhang, D. Rotational invariant dimensionality reduction algorithms. *IEEE Trans. Cybern.* **2016**, *47*, 3733–3746. [\[CrossRef\]](#)
30. Zhang, X.; Chu, D.; Tan, R.C. Sparse uncorrelated linear discriminant analysis for undersampled problems. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 1469–1485. [\[CrossRef\]](#)
31. Zhao, H.; Wang, Z.; Nie, F. A new formulation of linear discriminant analysis for robust dimensionality reduction. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 629–640. [\[CrossRef\]](#)
32. Zhang, D.; Li, X.; He, J.; Du, M. A new linear discriminant analysis algorithm based on L1-norm maximization and locality preserving projection. *Pattern Anal. Appl.* **2018**, *21*, 685–701. [\[CrossRef\]](#)
33. Peng, X.; Yang, F.; Wang, G.; Wu, Y.; Li, L.; Li, Z.; Bhatti, A.A.; Zhou, C.; Hepburn, D.M.; Reid, A.J.; et al. A convolutional neural network based deep learning methodology for recognition of partial discharge patterns from high voltage cables. *IEEE Trans. Power Deliv.* **2019**, *34*, 1460–1469. [\[CrossRef\]](#)
34. Morshuis, P.H.F.; Smit, J.J. Partial discharges at DC voltage: Their mechanism, detection and analysis. *IEEE Trans. Dielectr. Electr. Insul.* **2005**, *12*, 328–340. [\[CrossRef\]](#)
35. Seo, I.J.; Khan, U.A.; Hwang, J.S.; Lee, J.G.; Koo, J.Y. Identification of insulation defects based on chaotic analysis of partial discharge in HVDC superconducting cable. *IEEE Trans. Appl. Supercond.* **2015**, *25*, 1–5. [\[CrossRef\]](#)
36. Pirker, A.; Schichler, U. Partial discharges at DC voltage - measurement and pattern recognition. In Proceedings of the IEEE International Conference on Condition Monitoring and Diagnosis, Xi’an, China, 25–28 September 2016.
37. Yang, F.; Sheng, G.; Xu, Y.; Hou, H.; Qian, Y.; Jiang, X. Partial discharge pattern recognition of XLPE cables at DC voltage based on the compressed sensing theory. *IEEE Trans. Dielectr. Electr. Insul.* **2017**, *24*, 2977–2985. [\[CrossRef\]](#)
38. Wenrong, S.; Junhao, L.; Peng, Y.; Yanming, L. Digital detection, grouping and classification of partial discharge signals at DC voltage. *IEEE Trans. Dielectr. Electr. Insul.* **2008**, *15*, 1663–1674. [\[CrossRef\]](#)
39. Varol, Y.; Oztup, H.F.; Avci, E. Estimation of thermal and flow fields due to natural convection using support vector machines (SVM) in a porous cavity with discrete heat sources. *Int. Commun. Heat Mass Transf.* **2008**, *35*, 928–936. [\[CrossRef\]](#)

40. Botev, Z.I.; Grotowski, J.F.; Kroese, D.P. Kernel density estimation via diffusion. *Ann. Stat.* **2010**, *38*, 2916–2957. [[CrossRef](#)]
41. Umbaugh, S.E. *Digital Image Processing and Analysis: Applications with MATLAB and CVIPtools*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2018.
42. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [[CrossRef](#)]
43. Fan, R.E.; Chen, P.H.; Lin, C.J.; Joachims, T. Working set selection using second order information for training support vector machines. *J. Mach. Learn. Res.* **2005**, *6*, 1889–1918.
44. Chen, P.H.; Fan, R.E.; Lin, C.J. A study on SMO-type decomposition methods for support vector machines. *IEEE Trans. Neural Netw.* **2006**, *17*, 893–908. [[CrossRef](#)] [[PubMed](#)]
45. Platt, J.C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **2000**, *10*, 61–74.
46. Wu, T.F.; Lin, C.J.; Weng, R.C. Probability estimates for multi-class classification by pairwise coupling. *J. Mach. Learn. Res.* **2004**, *5*, 975–1005.
47. Mak, M.W.; Guo, J.; Kung, S.Y. PairProSVM: Protein subcellular localization based on local pairwise profile alignment and SVM. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2008**, *5*, 416–422. [[PubMed](#)]
48. Brunton, S.L.; Kutz, J.N. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*; Cambridge University Press: Cambridge, UK, 2019.
49. Pawlak, Z. Rough sets. *Int. J. Parallel Program.* **1982**, *11*, 341–356. [[CrossRef](#)]
50. Hu, Q.H.; Yu, D.R.; Xie, Z.X. Neighborhood classifiers. *Expert Syst. Appl.* **2008**, *34*, 866–876. [[CrossRef](#)]
51. Zhou, Z.H. *Ensemble Methods Foundations and Algorithms*; CRC Press: Boca Raton, FL, USA, 2012.
52. Li, Y.; Gou, J.; Fan, Z.W. Educational data mining for students’ performance based on fuzzy C-means clustering. *IET Gener. Transm. Distrib.* **2019**, *2019*, 8245–8250. [[CrossRef](#)]
53. Garcia, S.; Luengo, J.; Sáez, J.A.; Lopez, V.; Herrera, F. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 734–750. [[CrossRef](#)]