IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DEEP LEARNING ALGORITHMS FOR INTERNET OF MEDICAL THINGS

Received February 22, 2020, accepted February 27, 2020, date of publication March 2, 2020, date of current version March 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2977680

Deep Autoencoder for Mass Spectrometry Feature Learning and Cancer Detection

QINGGUO ZHOU[1], BINBIN YONG[1,2], QINGQUAN LV[1,3], JUN SHEN[4], (Senior Member, IEEE), AND XIN WANG[1]

[1]School of Information Science and Engineering, Lanzhou University, Lanzhou 730030, China
[2]School of Physical Science and Technology, Lanzhou University, Lanzhou 730030, China
[3]State Grid Gansu Electric Power Research Institute, Lanzhou 730030, China
[4]School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia

Corresponding author: Binbin Yong (yongbb@lzu.edu.cn)

This work was supported in part by State Grid Corporation Science and Technology Project under Grant No. SGGSKY00FJJS1800403, No. 522722160071 and No. 52272218002K, and Double first class Funding-International Cooperation and Exchange Program under Grant No. 227000-560001. Google Research Awards and Google Faculty Award. The work of Jun Shen was supported in part by the University of Wollongong's University Global Partnership Network Research Collaboration Fund University Internationalisation Committee Fund 2018-2019, in part by the National Science Foundation of China under Grant 61872079, and in part by the University of Wollongong's University Internationalisation Committee Fund International Exchange and Sabbatical Leave Program supporting his visit at Lanzhou University and Massachusetts Institute of Technology.

ABSTRACT Cancer is still one of the most life threatening disease and by far it is still difficult to prevent, prone to recurrence and metastasis and high in mortality. Lots of studies indicate that early cancer diagnosis can effectively increase the survival rate of patients. But early stage cancer is difficult to be detected because of its inconspicuous features. Hence, convenient and effective cancer detection methods are urgently needed. In this paper, we propose to utilize deep autoencoder to learn latent representation of high-dimensional mass spectrometry data. Meanwhile, as a contrast, traditional particle swarm optimization (PSO) optimization algorithm are also used to select optimized features from mass spectrometry data. The learned features are further evaluated on three cancer datasets. The experimental results demonstrate that the cancer detection accuracy by learned features is as high as 100%. As our main contribution, the deep autoencoder method used in this study is a feasible and powerful instrument for mass spectrometry feature learning and also cancer diagnosis.

INDEX TERMS Early cancer diagnosis, deep autoencoder, particle swarm optimization, mass spectrometry feature learning.

I. INTRODUCTION

Cancer is one of the most serious threats to human life and health in today's society, and it not only imposes death threats and heavy psychological burdens on patients, but also imposes a heavy financial burden on families and societies with very high treatment costs. According to the recent confirmed global cancer statistics report from American Cancer Society (ACS), nearly two million cancer patients were diagnosed in 2018 around the world [1]. An important reason for the low cure rate of cancer is that it is often diagnosed too late. For example, the five years' survival rate of breast cancer is more than 95% in the first stage, while it reduces to no more than 20% in the fourth stage. That is to say, early detection is crucial for higher survival rate for a cancer patient. Therefore, early diagnosis and treatment of cancer has become a key means of cancer prevention and treatment. Currently, the diagnosis of cancer depends mainly on imaging diagnosis [2]–[5] and pathological diagnosis [6]. The diagnosis of cancer is mainly conducted by doctors to examine for abnormal areas in the medical images, which costs a lot of energy and is prone to misdiagnosis. On the other hand, some cancers have no obvious symptoms in the medical images in the early stage. In this case, non-invasive and efficient early screening becomes an important research topic. Myomics and proteomics are the key research directions in the field of bioinformatics in recent years [7]. Some studies show that, when the body's health condition changes, some protein levels in the body will also change, which will result in differences in metabolic substances. These differences can be reflected in the measurement parameters of body fluids, such as blood and saliva, which can be detected by mass

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

spectrometry technology. Nowadays, biomass spectrometry technology has been widely used in proteomics and metabolomics. It can be used in the sequence analysis of amino acids in proteins, protein identification and quantitative analysis and disease diagnosis [8], [9]. Cancer biomarkers are substances produced directly by tumor cells or induced by tumor cells. The detection of cancer biomarkers can judge or predict the presence, pathogenesis and prognosis of the cancer. However, mass spectrometry technology is still not widely used in clinical diagnosis. One reason is the difficulty in collecting and processing relevant mass spectrometry data. Another reason is due to the high dimensional characteristics of mass spectrometry data [10], which makes the detection model prone to over-fitting . In order to detect cancer via the high-dimensional mass spectrometry data, in this paper, we utilize deep autoencoder and particle swarm optimization (PSO) algorithms to learn latent representation of high-dimensional mass spectrometry data, so as to reduce the dimension of mass spectrometry data for cancer detection. We evaluate the learned features by training and testing different machine learning models for cancer detection task.

A. CONTRIBUTIONS

In this paper, we make four main research contributions:

1) A new phemotocytoma and paraganglioma (PPGL) mass spectrometry dataset including 150 health samples and 131 cancer samples are collected for cancer detection experiments.
2) The characteristics of two ovarian cancer datasets and PPGL dataset are analyzed by principal component analysis (PCA) method.
3) Autoencoder and PSO based feature learning and selection methods are designed in this paper, and 1000 features are selected for cancer detection.
4) Cancer detection experiments are conducted based on learned features and traditional machine learning models, and the results are compared between autoencoder and PSO methods.

The above contributions can be represented by similar mass spectrometry based cancer detection studies, which are important for early screening and treatment of cancers.

B. ORGANIZATION

The rest of the paper is organised as follows. In section II, related work in mass spectrometry based cancer detection is presented. The datasets and preprocessing methods are introduced in section III. Then, the autoencoder and PSO based feature learning methods for cancer detection are presented in section IV. Section V depicts the experimental results and analysis. We address the conclusions and future work in section VI.

II. RELATED WORK

Mass spectrometry technology is being applied into cancer detection, and there have been many convincing studies.

Early in 2003, Stattin and Hakama combined a protein mass spectrometry based method and artificial intelligence (AI) algorithms to distinguish non-cancer groups, so as to determine better biomarkers [11]. In 2004, Yang *et al.* proposed the metabonomics technology based on high performance liquid chromatography (HPLC) [12]. They distinguished liver cancer patients and healthy volunteers by analyzing their urine samples. This method has better specificity and sensitivity. Then, Semmes *et al.* discussed the application of proteomic mass spectrometry technologies in prostate cancer diagnosis [13]. Later, Kojima *et al.* successfully utilized mass spectrometry data to detect stomach cancer [14]. De Petris *et al.* used SELDI-TOF-MS technology to detect the S100A6 in tumor cell lysate in 39 patients, which was very important for lung cancer cure test [15].

Recently, Liu *et al.* [16] proposed to detect tumor antigens from the ovarian cancer cell by proteomic-based methods. Their study indicated that the positive rate of serum HSP70 autoantibody in patients with ovarian cancer was 21.7%, which could be used to discriminate cancer patients. Jabbar *et al.* tried to detect pancreatic cancer by targeted mass spectrometry technology, and they achieved a best accuracy of 97% [17]. Liu *et al.* [18] developed a pipeline for biomarker development based on mass spectrometry technology, and their results showed a high predictive value with sensitivity of 95%.

Generally, when the body's organs become of lesions or cancerous, the metabolites in the body change with the body's lesions. After years of development, mass spectrometry technology has been more mature. Now the mass spectrometry technology is regarded as the most promising technology for cancer diagnosis. In the diagnosis of tumor, to find the features, which can best represent the characteristics of the tumor, is the focus of metabolic histology research. On the other hand, machine learning models can represent complex nonlinear mapping relationships and are widely used to find the inherent rules of data. In this paper, the deep autoencoder (DAE) and PSO algorithms are applied into the feature learning of high-dimensional cancer mass spectrometry data. Then, machine learning models are used to classify the mass spectrometry data based on selected features. Experiments show that the autoencoder and PSO algorithms can learn the features that contribute to classification well. Also, it indicates that mass spectrometry technology combined with machine learning method is feasible and has certain advantages for cancer detection.

III. DATASET

The datasets used in this paper include two ovarian cancer datasets, denoted as OC_4302 and OC_8702, and one PPGL dataset. We select one health sample and one cancer sample from these three datasets separately, and their heatmaps containing health samples and cancer samples are plotted in Fig. 1. We can see that ovarian cancer datasets have more obvious intensity contrast than PPGL datasets. However, it is difficult to distinguish health and cancer

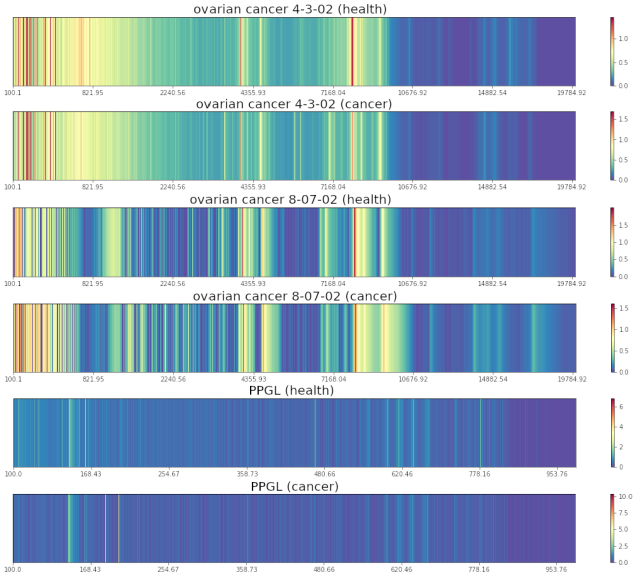

FIGURE 1. The heatmaps of mass spectrometry datasets.

samples by the intensity contrast of figures. Next, we will introduce the preprocessing and the initial feature screening methods.

A. OVARIAN CANCER DATASETS

The ovarian cancer datasets were acquired from the America Clinical Proteomics Program Database [19], which aim to develop and apply proteomics technology. In this paper, we will test on the ovarian cancer 4-3-02 dataset (OC_4302) and the ovarian cancer 8-7-02 dataset (OC_8702). The OC_4302 is a dataset with low-resolution collected using a WCX2 chip. It contains 100 ovarian cancer samples and 100 health samples, and every sample has 15,154 features. Baseline correction was manually carried out on these samples. The OC_8702 is also a dataset with low-resolution, which contains 162 cancer samples and 100 normal samples with 15,154 features. This dataset is curated with machine, and it is quite different from the OC_4302 dataset.

B. PPGL DATASET

PPGL is a glandular nerve chain tumor that can be classified as phemotocytoma and paraganglioma, depending on where it occurs. The PPGL dataset used in this paper is collected from a hospital in Shanghai city of China.

We firstly collected blood samples from PPGL patients and healthy people who were not diagnosed with PPGL. Then, the serum samples were processed and analyzed by matrix-assisted laser analysis of ionizing flight time mass spectrometry (MALDI-TOF-MS), and the mass spectrometry data of these samples was consequently obtained. The PPGL dataset includes 150 healthy samples and 131 PPGL samples, each data has 104,960 original features. The mass spectrometry data is composed of the mass-to-charge ratio (m/z) in horizontal coordinate and the intensity in the ordinated

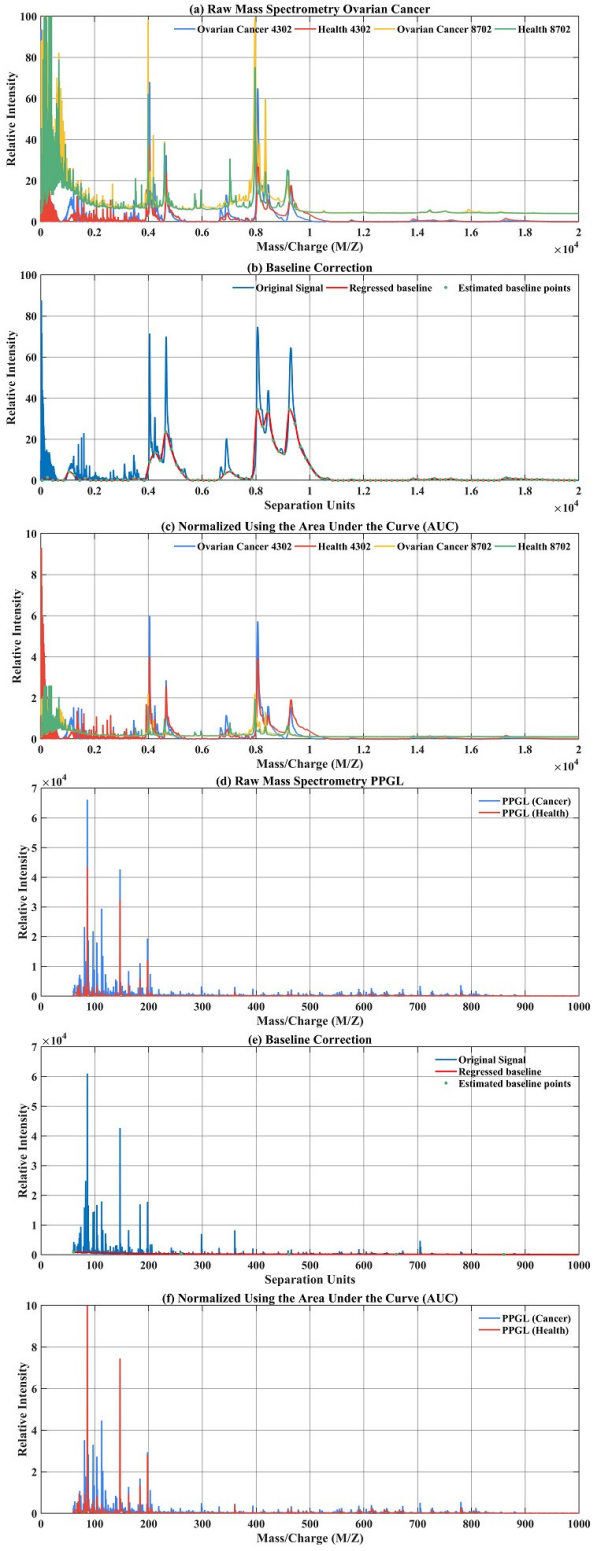

FIGURE 2. The original samples, baseline calibration samples and normalization samples of mass spectrometry data.

coordinate. By analyzing mass spectrometry data, we can find features containing classification information, so as to detect and identify cancer.

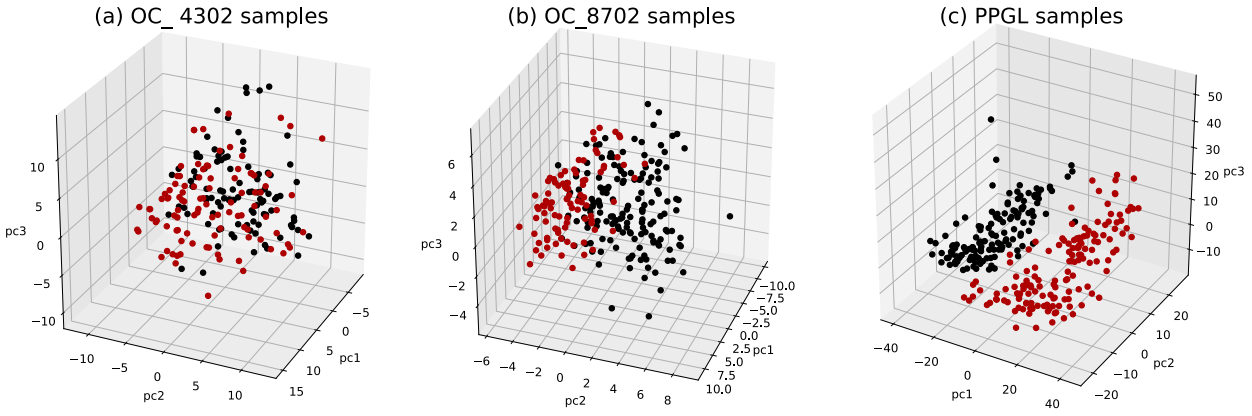

FIGURE 3. The mass spectrometry sampels by using first three principle components of three datasets.

C. MASS SPECTROMETRY PREPROCESSING

In this paper, the OC_4302 dataset is preprocessed based on baseline calibration and normalization method, which is actually using the area under the curve (AUC) method. In order to conveniently preprocess the data, we set the intensity values to range of 0 and 10.

The original mass spectrometry samples of OC_4302 and OC_8702 are shown in Fig. 2(a), which includes cancer samples and normal samples. The baseline calibrated mass spectrometry signals are shown in Fig. 2(b). Fig. 2(c) shows the signals after normalization ([0,10]). PPGL mass spectrometry data is exported by compassXport software, and the data is organized with mzXML format. Unlike the ovarian cancer dataset, PPGL data is metabolic mass spectrometry data. The mass-to-charge ratio ranges between 60 and 1,000, and the number of features is 104,960. The intensity values of mass-to-charge ratio between 60 and 100 are useless for cancer diagnose. Hence, we actually eliminate data between this range. And the number of features reduces to 95,022. Similarly, the preprocessed results of PPGL are shown in Fig. 2(d), Fig. 2(e) and Fig. 2(f). We can see that the baseline wandering problem is well relieved after baseline calibration and normalization for both ovarian cancer and PPGL data.

After preprocessing, we need to carry out initial feature screening to reduce the number of features.

D. INITIAL FEATURE SCREENING

$$\chi^2 = \sum \frac{(A-T)^2}{T} \tag{1}$$

Generally, mass spectrometry has the characteristics of high dimension but very small number of samples. Therefore, we need to screen the mass spectrometry data. In this paper, the single variable chi-square test (χ^2test) is used to screen the mass spectrometry data. The initial feature number of ovarian cancer samples is 15,154, and the initial number of features of PPGL samples is 104,960. We firstly calculate the χ^2 values of these features by Eq. (1). In which A denotes the observation value and T denotes the expected value. Then, these features are sorted by these values. Finally, we select

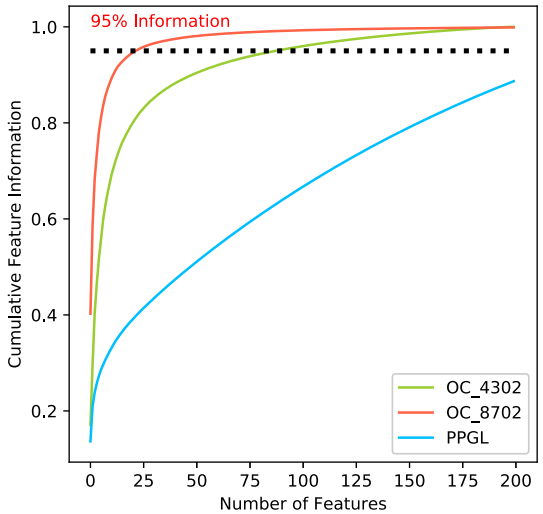

FIGURE 4. The cumulative information of different number of features.

9000 features for ovarian cancer samples, and 100,000 features for PPGL samples after the initial screening in the first stage.

E. FEATURES ANALYSIS

Generally, mass spectrometry data has significantly larger features than samples, which makes it relatively difficult to collect large enough samples through experiments. On the contrast, numerous features are prone to train a model with the risk of over-fitting. Therefore, we need to reduce the number of features of mass spectrometry data before using it for cancer detection [20]. Based on PCA method, we plot the first three principle components extracted from samples of three datasets, and these samples with only three principle components are visualized in Fig. 3. We can see that all three datasets tend to be divided into two categories, in which black dots represent cancer samples and red dots represent healthy samples. However, OC_8702 dataset and PPGL dataset can be obviously separated, whereas OC_4302 dataset has unclear classification boundary.

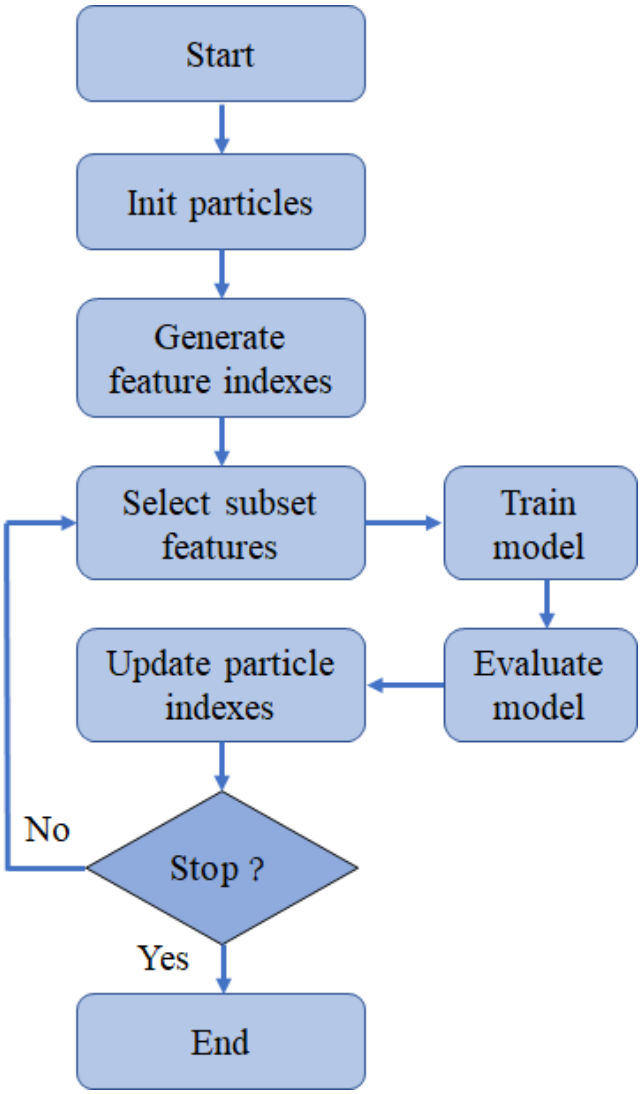

FIGURE 5. The flowchart of PSO based feature learning method.

Next, we calculate the cumulative information of different number of features, and the results are shown in Fig. 4. We can see that OC_8702 dataset needs minimum number of features to express 95% information, and PPGL dataset needs maximum number of features to express 95% information. Although mass spectrometry data has a high dimension, we can obtain more than 100% cumulative information with 1000 features for all three datasets. Therefore, in the next step, we will try to learn 1000 features to represent these samples for cancer detection.

IV. FEATURE LEARNING METHODS

In this section, we will introduce the PSO method and two types of deep autoencoder models for mass spectrometry feature learning.

A. PSO BASED FEATURE LEARNING METHOD

Firstly, we design a PSO algorithm to select features from initial features, and the flowchart is shown in Fig. 5. PSO is a global optimization algorithm, which is developed based on concept of swarm intelligence [21]. It has been widely used in many fields, such as neural network optimization, time series analysis and feature selection [22].

According to the analysis earlier, we can find that 1000 features can express 100% information. Hence, we select 1000 features to further test cancer detection models. We first select different features from the initial screening features randomly to generate 30 particle populations, and each population contains 1000 feature indexes of subset features. Then we use 70% samples extracted based on these subset features in these populations to train linear classifiers, and the classification accuracy is set as the objective function, which are used to select the global optimal features in this round. Next, these populations are updated, which means the feature indexes are updated. The classifiers are trained again to update the global optimal features in the next round. We repeat this optimization process 20 times to complete the feature selection. At last, these selected features can be used to detect cancer, and we test on the remaining 30% samples to test the classification results by these selected features.

B. DEEP AUTOENCODER BASED FEATURE LEARNING METHOD

The deep autoencoder model is a commonly used deep learning method for feature learning. The autoencoder model is shown in Fig. 6. In this paper, five layers are designed to learn features from original mass spectrum features based on deep autoencoder, which are denoted as L1, L2, L3, L4 and L5. The input vector and the output vector of the autoencoder are both the mass spectrum data after preprocessing. Therefore, the input dimension of OC_4302 dataset is 9,000, and the input dimension of OC_8702 dataset is 100,000. Herein, two types of autoencoder are designed for feature learning. First type is the deep autoencoder model with fully-connected (FC) neuron nodes, denoted as DAE, and another type is deep convolutional autoencoder (DCAE). The parameter setting of these two types models are shown in Table 1. For both OC_4302 and OC_8702 datasets, the DAE models are both designed with $2000 \times 1000 \times 1000$ structure. For ovarian cancer dataset, DCAE model consists of 3 convolutional layers, one max-pooling layer and one upsampling layer. The number of convolutional kernels is set as 16, 1 and 1, and the kernel sizes are set as 5, 3 and 1. The stride of convolution operation is set as 1, and the pooling size is set as 9. On the whole, ReLU function is selected as the activation function. For PPGL dataset, DCAE model also consists of 3 convolutional layers, one max-pooling layer and one upsampling layer. However, the stride of first convolution layer is set as 10, and the pooling size is set as 10, due to the high dimension of PPGL dataset. Furthermore, according to these model structures, the middle L3 layer has just 1000 nodes, corresponding to 1000 learned features, which can be easily used for cancer classification.

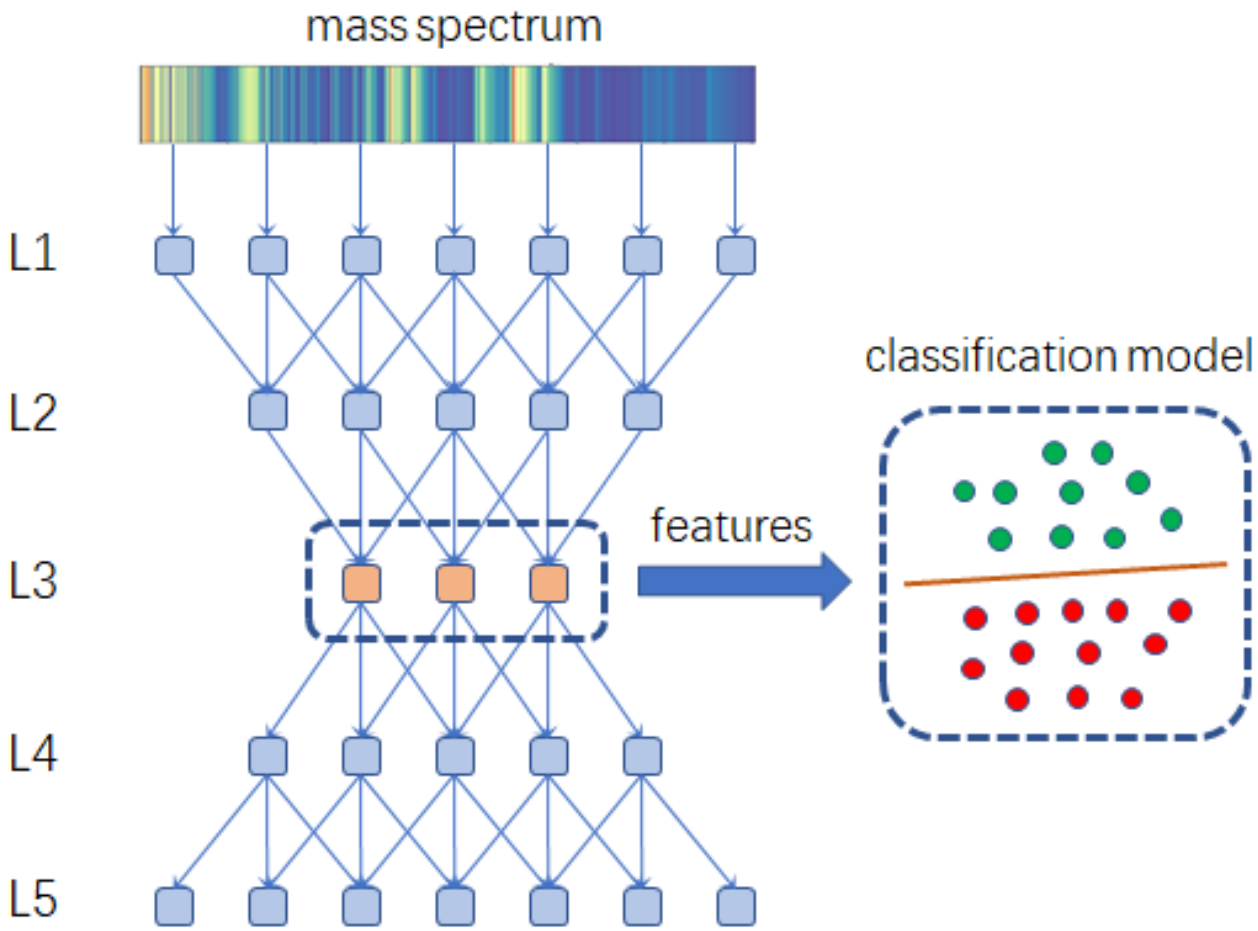

FIGURE 6. The feature learning method based on deep autoencoder.

TABLE 1. The architectural of deep autoencoders.

	OC_4302			OC_8702			PPGL		
model	DAE	DCAE		DAE	DCAE		DAE	DCAE	
	fc	type	size	fc	type	size	fc	type	size
L1	9000	conv	16×5	9000	conv	16×5	100000	conv	16×5, s=10
L2	2000	conv	1×3	2000	conv	1×3	2000	pool	10
L3	1000	pool	9	1000	pool	9	1000	conv	1×3
L4	2000	upsample	9	2000	upsample	9	2000	upsample	10
L5	9000	conv	1×2	9000	conv	1×2	100000	conv	1×2

V. EXPERIMENTS AND ANALYSIS

In this section, we first show and analyze the learned features of these samples based on PSO and autoencoder, followed with the setting of the parameters of machine learning classification models, including extreme learning machine (ELM), BPNN, support vector machine (SVM), k nearest neighbors (KNN) and random forest (RF). At last, the cancer detection results by learned features are illustrated and compared.

A. FEATURE LEARNING RESULTS

We conducted experiments to train feature learning models for three datasets. For each dataset, we select one healthy sample and one cancer sample to show the features as Fig. 7, Fig. 8 and Fig. 9. The original features, 1000 features learned by PSO, DAE, DCAE are shown in sub-figures, the blue curves represent the features of healthy samples, and the red curves represent the features of cancer samples. For ovarian cancer datasets, we can see that the original features and learned features are distributed in the whole data for both healthy samples and cancer samples. For PPGL dataset, the original features and features learned by DCAE are mainly distributed in the front. The differences in these features may reflect the differences in the expression of some proteins between healthy people and cancer patients, and

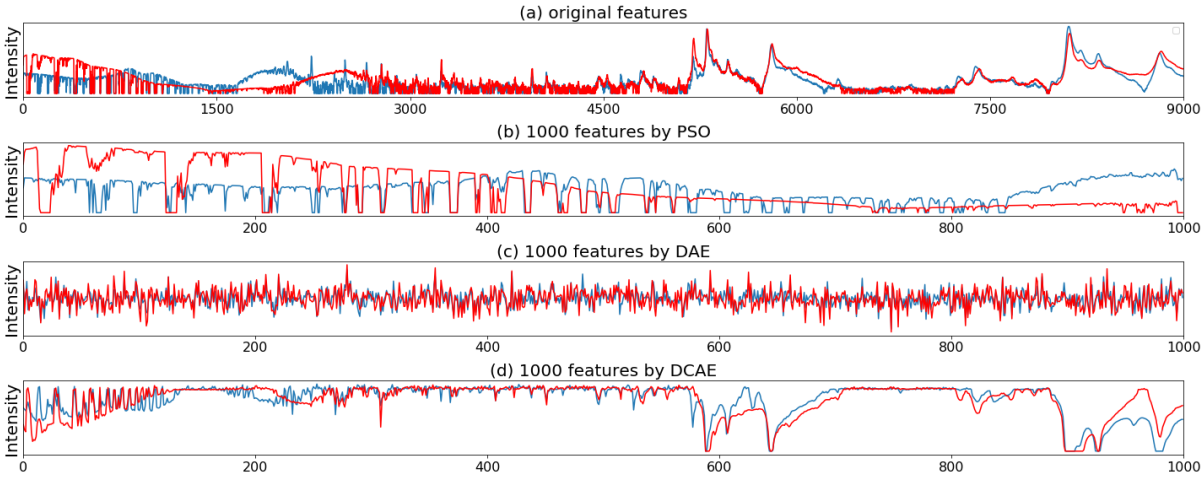

FIGURE 7. Learned features of one health sample (blue) and one cancer sample (red) of OC_4302 dataset.

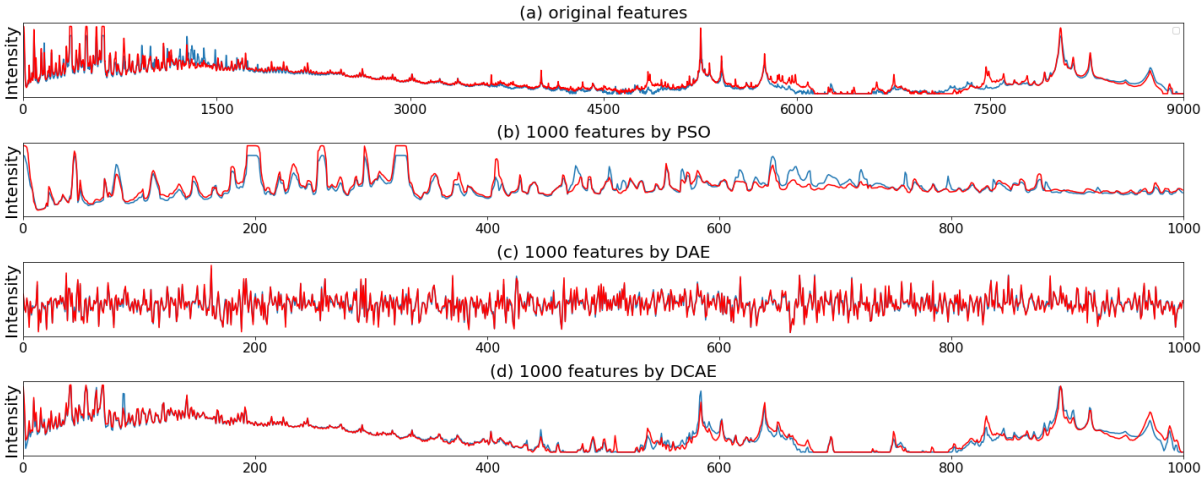

FIGURE 8. Learned features of one health sample (blue) and one cancer sample (red) of OC_8702 dataset.

further exploration of these features may provide some basis for the diagnosis of cancer.

Further, we can observe that these features learned by PSO and DAE are different from the original features, and we cannot observe obvious characteristics of these features. We can also find that features learned by DCAE have similar trends for OC_8702 and PPGL datasets. Meanwhile, these features has opposite trend characteristics for OC_4302 dataset. It indicates that DCAE can better learn the features of the original data. We will further conduct cancer classification experiments based on these features.

B. MEASURE METRICS AND MODEL PARAMETERS

In our experiments, we use three metrics, including Accuracy, Sensitivity and Specificity, to evaluate the performances of cancer detection. These three metrics are frequently used in classification problems, which are based on four basic metrics including TP (true positive), TN (true negative), FP (false positive) and FN (false negative). For classification problems, the Accuracy metric is often used to evaluate classifiers, which represents the percentage of correct identification. As a quite intuitive metric, Accuracy does not fully reveal the ability of the classification models, especially when the samples are not balanced. Sensitivity and Specificity are also two important evaluation metrics in disease diagnosis studies, especially in cancer detection. Sensitivity (Recall) can be used to describe the proportion of detected cancer samples to the whole cancer samples. Therefore, a high Sensitivity denotes a low probability of missed diagnosis. Specificity can be used to describe the proportion of healthy samples to the whole health samples. It indicates the ability for detecting healthy people. Generally, a high Specificity means a low probability of false diagnosis.

We obtained 1,000 features for classification by training feature learning models based on PSO and autoencoder. Then, samples based on these 1,000 features from three

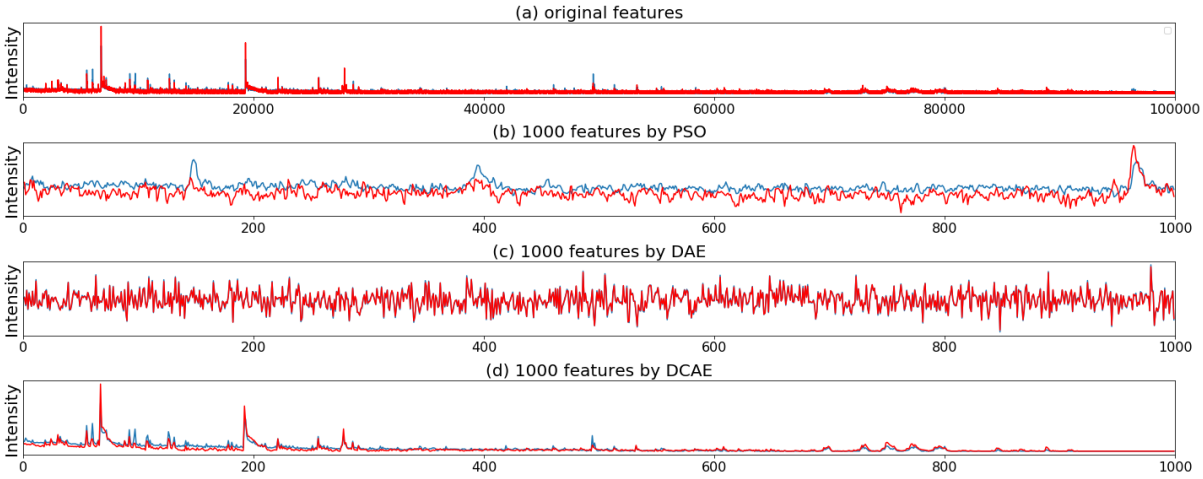

FIGURE 9. Learned features of one health sample (blue) and one cancer sample (red) of PPGL dataset.

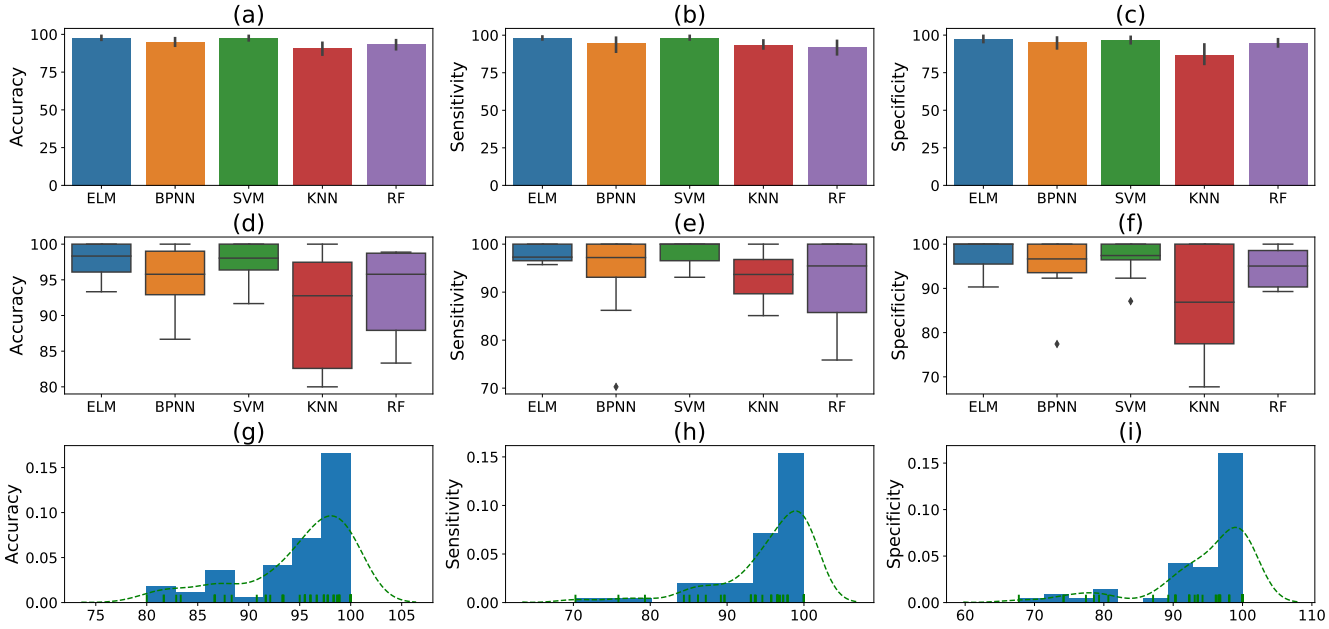

FIGURE 10. Experimental results: (a-c) average classification results of ELM, BPNN, SVM, KNN and RF; (d-f) the classification result distribution of ELM, BPNN, SVM, KNN and RF; (g-i) the distribution of three metrics.

datasets are generated to train classifiers, including ELM, BPNN, SVM, KNN and RF. In our experiments, these classifiers are tested to get better model parameters. For these models, the input dimension is equal to the number of original mass spectrometry or 1,000. For ELM model, it is based on FC neural network with one hidden layer. We use grid method to search appropriate number of hidden nodes, which is set as 5,000 finally. And the activation function is set as ReLu function, which is equal to function $max(0, x)$ and proved to be effective for our experiments. ELM is a random model, and its forecast results are fluctuating. Therefore, we conduct five experiments and take the average result as the final results of ELM. For BPNN, it is designed with one hidden layers with 20 hidden nodes after many experiments. For the SVM model, the kernel type is selected as the radial basis function (RBF). For KNN, we set the parameter k (number of neighbors) to 15, due to the small number of samples. For RF, there are no important parameter to set.

C. CANCER DETECTION RESULTS

In this subsection, we will compare the classification results between original, PSO and deep autoencoder features. After training these classifiers with 70% samples, 30% samples are used for test, and the detection results are shown in Fig. 10, Fig. 11 and Table 2.

We can see that ELM and SVM achieved better average classification results than BPNN, KNN and RF from Fig. 10

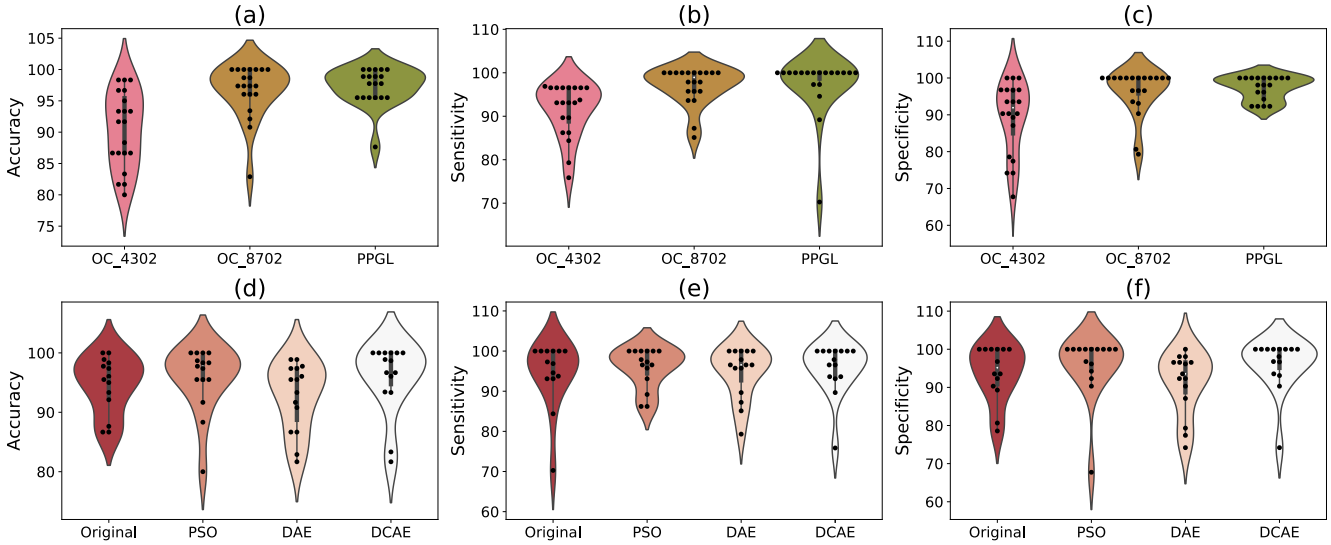

FIGURE 11. Experimental results: (a-c): average classification results based on three datasets; (d-f): average classification results based on different feature learning methods.

TABLE 2. Experimental results: classification results by learned features.

	OC_4302 (%)			OC_8702 (%)			PPGL (%)		
model	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe
				Original					
ELM	**98.33**	**96.88**	**100.0**	**100.0**	**100.0**	**100.0**	*	*	*
BPNN	93.33	93.10	93.55	97.37	100.0	93.55	87.64	70.27	100.0
SVM	95.00	93.10	96.77	**100.0**	**100.0**	**100.0**	95.51	100.0	92.31
KNN	86.67	93.75	78.57	92.11	100.0	80.65	97.75	94.59	100.0
RF	86.67	84.38	89.29	96.05	100.0	90.32	**98.88**	**97.30**	**100.0**
				PSO					
ELM	**98.33**	**96.55**	**100.0**	**100.0**	**100.0**	**100.0**	95.51	97.30	94.23
BPNN	91.67	86.21	96.77	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
SVM	98.33	96.55	100.0	**100.0**	**100.0**	**100.0**	97.75	100.0	96.15
KNN	80.00	93.10	67.74	97.37	95.74	100.0	95.51	89.19	100.0
RF	88.33	86.21	90.32	98.68	97.87	100.0	95.51	100.0	92.31
				DAE					
ELM	**93.33**	**96.55**	**90.32**	**97.37**	**95.74**	**100.0**	95.51	100.0	92.31
BPNN	86.67	96.55	77.42	96.05	95.74	96.55	95.51	100.0	92.31
SVM	91.67	96.55	87.10	97.37	97.87	96.55	**98.88**	**100.0**	**98.08**
KNN	81.67	89.66	74.19	82.89	85.11	79.31	97.75	100.0	96.15
RF	86.67	79.31	93.55	90.79	87.23	96.55	**98.88**	**100.0**	**98.08**
				DCAE					
ELM	**96.67**	**96.55**	**96.77**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
BPNN	93.33	93.10	93.55	98.68	97.87	100.0	**100.0**	**100.0**	**100.0**
SVM	96.67	96.55	96.67	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
KNN	81.67	89.66	74.19	93.42	93.62	93.10	**100.0**	**100.0**	**100.0**
RF	83.33	75.86	90.32	96.05	93.62	100.0	98.88	100.0	98.08

The best results are **bolded**.

(a-c), and SVM has abnormal Accuracy and Specificity in 10 (d-f). We can also observe from 10 (g-i) that the overall Accuracy, Sensitivity and Specificity are close to 100%. Meanwhile, we can see from Table 2 that ELM outperforms other classifiers for ovarian cancer datasets, and SVM also achieves best classification results for OC_8702 dataset based

on features learned by DCAE. For PPGL dataset, ELM cannot be trained based on original data, due to its high dimension, which is one reason for feature learning. RF, BPNN, SVM (RF), and ELM (BPNN, SVM, KNN) achieved best results for original features, PSO features, DAE features and DCAE features. In fact, on the whole, the differences between Sensitivity and Specificity of ELM and SVM are smaller than other classifiers. It indicates that ELM and SVM can predict with both low probability of missed diagnosis and low probability of error diagnosis. Hence, ELM and SVM are more suitable for mass spectrometry based cancer detection.

From Fig. 11 (a-c), we can see that PPGL dataset can be classified with highest Accuracy and Specificity, while OC_8702 dataset has a little advantage in Sensitivity. OC_4302 dataset performs worst for three metrics in three datasets. From Table 2 we can see that we achieved better classification results on OC_8702 dataset and PPGL dataset than that on OC_4302 dataset, which is consistent with our previous analysis for these dataset in Section III.

From Fig. 11 (d-f) we can see that classification results based on DAE features are similar to results based on original features. Results on DAE features are even a little worse than results of original features from Table 2, because of the fewer features than before. However, results based on PSO features and DCAE features are better than results of original features. It indicates that PSO and DCAE can satisfactorily learn the feature representation and remove the effects of noise.

From Table 2, we can also see that PSO features improve the classification performance on OC_8702 and PPGL dataset. While the classification results based on DCAE features are improved on both three datasets. Therefore, DCAE model can be well applied in mass spectrometry feature learning for cancer detection, and the highest Sensitivity for three datasets reach 96.55%, 100% and 100% based on learned features.

VI. CONCLUSION

Medical data collected from sensor network is often used for disease diagnosis [23], [24]. Mass spectrometry analysis are important research directions in the field of bioinformatics in recent years, especial for medical data. In this paper, we designed machine learning models to diagnose cancer based on the features learned from mass spectrometry data. PSO and deep autoencoder are utilized to learn the representation features of mass spectrometry data for cancer detection. Meanwhile, five machine learning models including ELM, BPNN, SVM, KNN and RF are utilized to detect cancer based on learned mass spectrometry features. Results show that DCAE model is more suitable for high-dimension feature learning than traditional PSO method, and ELM and SVM model can be well utilized to detect cancers. In future, we will focus on the model optimization, and more mass spectrometry data should be collected to improve and verify the classification models.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Sep. 2018.

[2] W. Wei, X.-L. Yang, B. Zhou, J. Feng, and P.-Y. Shen, "Combined energy minimization for image reconstruction from few views," *Math. Problems Eng.*, vol. 2012, pp. 1–15, Oct. 2012.

[3] Q. Ke, J. Zhang, W. Wei, R. Damasevicius, and M. Wozniak, "Adaptive independent subspace analysis of brain magnetic resonance imaging data," *IEEE Access*, vol. 7, pp. 12252–12261, 2019.

[4] W. Wei, B. Zhou, D. Połap, and M. Woźniak, "A regional adaptive variational PDE model for computed tomography image reconstruction," *Pattern Recognit.*, vol. 92, pp. 64–81, Aug. 2019.

[5] Q. Ke, J. Zhang, W. Wei, D. Połap, M. Woźniak, L. Kośmider, and R. Damaševičius, "A neuro-heuristic approach for recognition of lung diseases from X-ray images," *Expert Syst. Appl.*, vol. 126, pp. 218–232, Jul. 2019.

[6] C.-J. Tseng, C.-J. Lu, C.-C. Chang, and G.-D. Chen, "Application of machine learning to predict the recurrence-proneness for cervical cancer," *Neural Comput. Appl.*, vol. 24, no. 6, pp. 1311–1316, 2014.

[7] F. Wen, L. Chu, P. Liu, and R. C. Qiu, "A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning," *IEEE Access*, vol. 6, pp. 69883–69906, 2018.

[8] D. J. Ryan, D. Nei, B. M. Prentice, K. L. Rose, R. M. Caprioli, and J. M. Spraggins, "Protein identification in imaging mass spectrometry through spatially targeted liquid micro-extractions," *Rapid Commun. Mass Spectrometry*, vol. 32, no. 5, pp. 442–450, 2017.

[9] A. Arendowski, J. Nizioł, and T. Ruman, "Silver-109-based laser desorption/ionization mass spectrometry method for detection and quantification of amino acids," *J. Mass Spectrometry*, vol. 53, no. 4, pp. 369–378, Feb. 2018.

[10] C. R. Goodwin, S. D. Sherrod, C. C. Marasco, B. O. Bachmann, N. Schramm-Sapyta, J. P. Wikswo, and J. A. McLean, "Phenotypic mapping of metabolic profiles using self-organizing maps of high-dimensional mass spectrometry data," *Anal. Chem.*, vol. 86, no. 13, pp. 6563–6571, Jun. 2014.

[11] P. Stattin and M. Hakama, "Correspondence re: B-L. Adam et al., serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res., 62: 3609-3614, 2002.," *Cancer Res.*, vol. 63, no. 10, p. 2701, 2003.

[12] J. Yang, G. Xu, Y. Zheng, H. Kong, T. Pang, S. Lv, and Q. Yang, "Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases," *J. Chromatography B*, vol. 813, nos. 1–2, pp. 59–65, Dec. 2004.

[13] O. J. Semmes, G. Malik, and M. Ward, "Application of mass spectrometry to the discovery of biomarkers for detection of prostate cancer," *J. Cellular Biochemistry*, vol. 98, no. 3, pp. 496–503, 2006.

[14] T. Kojima, K. Yoshikawa, S. Saga, T. Yamada, S. Kure, T. Matsui, T. Uemura, Y. Fujimitsu, M. Sakakibara, Y. Kodera, and H. Kojima, "Detection of elevated proteins in peritoneal dissemination of gastric cancer by analyzing mass spectra data of serum proteins," *J. Surgical Res.*, vol. 155, no. 1, pp. 13–17, Jul. 2009.

[15] L. D. Petris, L. M. Orre, L. Kanter, M. Pernemalm, H. Koyi, R. Lewensohn, and J. Lehtiö, "Tumor expression of S100A6 correlates with survival of patients with stage i non-small-cell lung cancer," *Lung Cancer*, vol. 63, no. 3, pp. 410–417, Mar. 2009.

[16] X.-X. Liu, H. Ye, P. Wang, L.-X. Li, Y. Zhang, and J.-Y. Zhang, "Proteomic-based identification of HSP70 as a tumor-associated antigen in ovarian cancer," *Oncol. Rep.*, vol. 37, no. 5, pp. 2771–2778, Mar. 2017.

[17] K. S. Jabbar, L. Arike, C. S. Verbeke, R. Sadik, and G. C. Hansson, "Highly accurate identification of cystic precursor lesions of pancreatic cancer through targeted mass spectrometry: A phase IIc diagnostic study," *J. Clin. Oncol.*, vol. 36, no. 4, pp. 367–375, 2017.

[18] X. Liu, W. Zheng, W. Wang, H. Shen, L. Liu, W. Lou, X. Wang, and P. Yang, "A new panel of pancreatic cancer biomarkers discovered using a mass spectrometry-based pipeline," *Brit. J. Cancer*, vol. 117, no. 12, pp. 1846–1854, Nov. 2017.

[19] C for Cancer Research. (Dec. 2003). *Clinical Proteomics Program*. [Online]. Available: https://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

[20] Q. Ke, J. Zhang, H. Song, and Y. Wan, "Big data analytics enabled by feature extraction based on partial independence," *Neurocomputing*, vol. 288, pp. 3–10, May 2018.
[21] C. Yue, B. Qu, and J. Liang, "A multiobjective particle swarm optimizer using ring topology for solving multimodal multiobjective problems," *IEEE Trans. Evol. Comput.*, vol. 22, no. 5, pp. 805–817, Oct. 2018.
[22] T. Xiong, Y. Bao, Z. Hu, and R. Chiong, "Forecasting interval time series using a fully complex-valued RBF neural network with DPSO and PSO algorithms," *Inf. Sci.*, vol. 305, no. C, pp. 77–92, Jun. 2015.
[23] W. Wei, H. Song, W. Li, P. Shen, and A. Vasilakos, "Gradient-driven parking navigation using a continuous information potential field based on wireless sensor network," *Inf. Sci.*, vol. 408, no. 2, pp. 100–114, Oct. 2017.
[24] W. Wei, Q. Xu, L. Wang, X. H. Hei, P. Shen, W. Shi, and L. Shan, "GI/Geom/1 queue based on communication model for mesh networks," *Int. J. Commun. Syst.*, vol. 27, no. 11, pp. 3013–3029, 2013.

QINGGUO ZHOU received the B.S. and M.S. degrees in physics and the Ph.D. degree in theoretical physics from Lanzhou University, in 1996, 2001, and 2005, respectively. He is currently a Professor with Lanzhou University, where he is also working with the School of Information Science and Engineering. His research interests include safety-critical systems, embedded systems, and real-time systems. He is also a Fellow of IET. He was a recipient of the IBM Real-Time Innovation Award, in 2007, the Google Faculty Award, in 2011, and the Google Faculty Research Award, in 2012.

BINBIN YONG received the Ph.D. degree in computer science and technology from Lanzhou University, in 2017. He is currently a Postdoctoral Researcher with Lanzhou University. His research interests include parallel computing of GPU, machine learning, and artificial neural networks.

QINGQUAN LV received the M.S. degree in computer Science and Technology from Lanzhou University, in 2012, where he is currently pursuing the Ph.D. degree. He is also a Senior Engineer with the State Grid Gansu Electric Power Research Institute. His research interests include new energy, machine learning, and artificial neural networks.

JUN SHEN (Senior Member, IEEE) received the Ph.D. degree from Southeast University, China, in 2001. He is currently an Associate Professor with the School of Computing and Information Technology, University of Wollongong, Wollongong. He has published more than 200 articles in journals and conferences in CS/IT areas. His expertise include computational intelligence, web services, and cloud computing. He is also a Senior Member of two institutions: ACM and ACS. He has been an Editor, the PC Chair, a Guest Editor, and a PC Member for numerous journals and conferences published by IEEE, ACM, Elsevier, and Springer.

XIN WANG received the bachelor's degree in information security from Sichuan University, in 2014. She is currently pursuing the master's degree with the Distributed and Embedded Laboratory, Lanzhou University. Her research interests include artificial neural networks, machine learning, deep learning, and big data.