Dimensionality Reduction for Identification of Hepatic Tumor Samples Based on Terahertz Time-Domain Spectroscopy

Haishun Liu, Zhenwei Zhang, Xin Zhang, Yuping Yang, Zhuoyong Zhang, Xiangyi Liu, Fan Wang, Yiding Han, and Cunlin Zhang

***Abstract*—Terahertz time-domain spectroscopy (THz-TDS) combining with chemometrics methods was proposed for the identification of hepatic tumors. Two linear compression methods, principle component analysis and locality preserving projections (LPPs), and a nonlinear method, Isomap, were used to reduce the dimensionality of the measured dataset. Comparing two-dimensional (2-D) data reduced by these three dimensionality reduction techniques, only 2-D Isomap plot could separate the distances between two classes for the THz time-domain data and LPP had capacity of distinguishing two types of samples building on frequency-domain data. The best classification accuracies from 2-D time-domain data were 99.81 ± 0.30% and 99.69 ± 0.61% given by Isomap probabilistic neural network (PNN) and Isomap support vector machine (SVM), respectively, while the best classification results of 2-D frequency-domain data were 100.00 ± 0.00%, 99.75 ± 0.32% provided by LPP-PNN, LPP-SVM. The results showed that Isomap and LPP are appropriate techniques to reflect the nonlinear manifold of the THz data. The THz technology either in time-domain or frequency-domain coupled with Isomap-PNN or LPP-PNN could offer a potential procedure to identify hepatic tumors.**

***Index Terms*—Dimensionality reduction, Isomap, locality preserving projections (LPPs), principle component analysis (PCA), terahertz (THz).**

I. Introduction

LIVER cancer is a common cancer with a high morbidity and mortality worldwide [1], [2]. The most common type is hepatocellular carcinoma (HCC). Recently, several established techniques have been used for diagnosing HCC. Although CT and MRI possess of high sensitivity and specificity, they are impractical for frequent monitoring especially for the unnecessary radiation exposure to the patients [3]. As for α-fetoprotein or ultrasound, the low sensitivity and specificity maybe lead to misdiagnose [3]. A reliable method is urgently needed for screening the liver cancer.

On account of the attractive features of nonionizing, low energy, and good penetration of terahertz (THz) radiation [4], THz time-domain spectroscopy (THz-TDS) and THz pulse imaging (TPI) has been applied to many fields in this decade, such as composite materials nondestructive testing [5], drug quality testing [6], crop identification [7], cultural heritage detection [8], etc. Similarly, utilizing THz technologies to screen cancers have also made corresponding progress. Ashworth *et al.* [9] reported the different optical properties between normal breast tissue (including adipose and fibrous types) and cancerous tissue by means of THz spectroscopy technique. Ji *et al.* [10] identified early gastric cancer regions from fresh tissue samples by using TPI, largely because the THz reflection intensities of cancer regions were higher than those of normal tissues. Bowman *et al.* [11] differentiated the paraffin-embedded breast cancer tissues with different thickness by TPI and verified it with pathology images. For THz wave, the main reason of identifying these cancer regions lies in the normal and cancerous tissues could be regarded as two structural substances of different optical properties [9], [12]. Additionally, the tissue is not destroyed by the THz radiation. Generally, the absorption of cancer region is higher than that of the normal within THz spectrum [9], [13], [14].

By exploring the structures and relationships within the spectra data, chemometrics methods are capable of clarifying the differences between tissues to identify the tumors. So recently chemometrics combining THz-TDS has been widely applied to

Manuscript received October 12, 2017; revised January 26, 2018; accepted February 20, 2018. Date of publication April 9, 2018; date of current version May 1, 2018. This work was supported in part by the National Instrumentation Program under Grant 2012YQ14000508; in part by the National Key Basic Research and Development Program of China (973 Program) under Grant 2015CB755402; in part by the National Natural Science Foundation of China under Grant 11574408, Grant 21275101, and Grant 11204191; in part by the National Key Research and Development Program of China under Grant 2017YFB0405402); in part by Beijing Science and Technology Project under Grant D151100006015002 and Grant Z161100005016074; and in part by the Young-talent Plan of State Affairs Commission under Grant 2016-3-02. *(Corresponding authors: Zhenwei Zhang; Xiangyi Liu.)*

H. Liu, Z. Zhang, and C. Zhang are with the Beijing Key Laboratory for Terahertz Spectroscopy and Imaging, Key Laboratory of Terahertz Optoelectronics, Ministry of Education, and Beijing Advanced Innovation Center for Imaging Technology, Department of Physics, Capital Normal University, Beijing 100048, China (e-mail: liuhaishun2015@cnu.edu.cn; zhangzw_cnu@163.com; cunlin_zhang@cnu.edu.cn).

X. Zhang and Z. Zhang are with the Department of Chemistry, Capital Normal University, Beijing 100048, China (e-mail: xinkevinzhang@outlook.com; gusto2008@vip.sina.com).

Y. Yang is with the School of Science, Minzu University of China, Beijing 100081, China (e-mail: ypyang_cun@126.com).

X. Liu and F. Wang are with the Department of Laboratory Medicine, Beijing Tongren Hospital, Capital Medical University, Beijing 100730, China (e-mail: 13693328516@163.com; tryywf@126.com).

Y. Han is with the Department of Pathology, Beijing Tongren Hospital, Capital Medical University, Beijing 100730, China (e-mail: 771206965@qq.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TTHZ.2018.2813085

cancer diagnosis. Brun *et al.* [15] distinguished cancerous tissue and even tumor subclasses using principal component analysis (PCA). Qi *et al.* [16] used two fuzzy classifiers, fuzzy rule-building expert systems and fuzzy optimal associative memories, to classify the cervical cancer samples. Additionally, they also compared the performances of other two classifiers, support vector machine (SVM) and partial least squares-discriminant analysis, with different pretreatment methods when diagnosing cervical carcinoma [17].

During processing the spectra data by employing chemometrics tools, the curse of dimensionality is inevitable and often accompanied by the emergence of the undesired properties of high-dimensional spaces [18], computational cost, and bad performance of predictor [19].Thus, it is imperative to reduce the dimensionality. Many scholars usually choose PCA, a popular linear technique for dimensionality reduction [20], to reduce dimension. But PCA is difficult to maintain topological relationships for the nonlinear complex dataset, resulting in the false relative position between the faraway points and nearby points, which could affect the classification results. Isomap is a global manifold learning algorithm proposed by Tenenbaum *et al.* [21] in 2000 for dimensionality reduction of nonlinear data. It has been used in data visualization and classification [22], [23]. It is supposed that the data lie on or near an embedded lower-dimensional manifold. Isomap builds on multidimensional scaling (MDS) but tries to find a lower dimensional embedding through measuring the approximated geodesic distances that preserve the intrinsic geometry of the original data [21]. The difference between the geodesic distance and the Euclidean distance is that the former is the distance between the points on the manifold and the latter is that in Euclid space. Both PCA and Isomap try to preserve the overall structure of the data [21], [24], maybe ignoring the local structure of the data. However, locality preserving projections (LPPs), combining the advantage of linear and local nonlinear methods, is a novel linear dimensionality reduction algorithm proposed by He *et al.* [25]. LPP has proved the great effectiveness for finding local geometrical structure of the dataset, such as face recognition [24] and document representation [26]. LPP seeks an embedding that could preserve local information and acquire the key parts of the data manifold structure. After dimensionality reduction, the extracted features are used as the inputs of the classifiers. SVM and probabilistic neural network (PNN) are remarkable classifiers to validate the classification effects as they have made great achievements in pattern classification [27]–[30].

In this paper, we compared the separating capacities of two-dimensional (2-D) data building on PCA, LPP, and Isomap. SVM and PNN classified the THz data of hepatic tissue samples to further realize the classification effects of three dimensionality reduction techniques.

II. Experiment

A. Sample Preparation

Tumor and normal hepatic tissue samples placed on the silver mirrors were provided by Beijing Tongren Hospital, Capital Medical University. These samples were prepared at the same time in the pathology laboratory of the constant humidity and temperature. The detailed preparing procedure was described as follows: at first we soak the tumor and normal hepatic tissues in the 4% formalin solution to fix the cellular morphology. And next the tissues were dewatered with the ethanol solution. In order to make the paraffin permeate into the tissue to obtain a better effect, we put the tissues into xylene for hyalinization. At last, the tissues were soaked into the melting paraffin for embedding. After that, we obtained measured specimens in the following three steps. First, the paraffin-embedded normal and tumor hepatic slices with a thickness of about 45, ± 3 μm were sliced by slicer. Since the tissue sections were easily rolled together, they were put in a warm water pool and flatten under the tension of the water, and then the tissue sections were taken out and put on silver mirrors. Finally, the specimens were put on a drying plate and dewatered by a process of dehydration at a temperature of about 60 °C. The experiments using human hepatic tissue specimens were approved by the Ethics Committee in Beijing Tongren Hospital, Capital Medical University. Note that we have ever tried the sample thickness from 10 to 50 μm with the increment of 10 μm, the sample with the thickness of 40 μm was obtained the best identification effect. However, when slicing a sample with 50 μm, we found that this thickness was difficult to slice because the thicker slice was easier to roll together. And the thicker plicate slice was not entirely flatted with the tension of water. Additionally, the crisp characteristic of liver tissue also resulted in a broken form of the specimens. So we chose to the slices with a thickness of 45 μm.

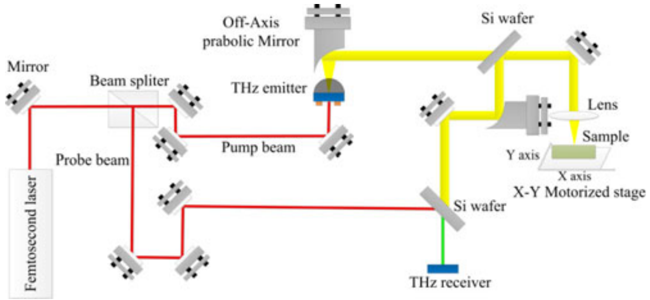


Fig. 1. Schematic of THz reflection setup.

B. Instrument Setup

A schematic of homemade THz-TDS instrument in reflection geometry was given in Fig. 1. The setup is based on the ultrafast fiber laser (Toptica Photonics). The femtosecond laser was divided into two beams by beam splitter. One was pump beam that was collimated on the photoconductive antenna to generate THz wave that was reflected vertically to the sample. The sample was scanned by moving the 2-D motorized stage. The other was probe beam merging with the THz wave carrying the sample information. They entered the detecting crystal ZnTe. After the processing of A/D converter and the lock-in amplifier, the time-domain signal was shown on the computer. In order to avoid system echo and vapor absorption in the atmosphere [31], the impulse functions were extracted by deconvolving each waveform of sample with that of reference. The more description about this process can be seen in the literature [32]. Here, the reference signal was acquired by using silver mirror.

III. Principle of Data Analysis Methods

A. PCA

The basic idea of PCA [33] is to find a linear mapping to project high-dimensional data to a lower dimensional subspace, while maximizing the variance of the data and minimizing the mean squared reconstruction error [34]. This is achieved by computing the covariance matrix of original data so that a set of eigenvectors is obtained in terms of corresponding eigenvalues. The eigenvectors also called principal components represent new composite variables and the eigenvalues express the projection variance of data. Generally, these orthogonal principal components are selected by cumulative variance contribution rate. The more detail of PCA can be seen elsewhere [35]. In this experiment, the total variance rates of top two principal components were 94.9% and 90.6% for extracting from time-domain and frequency-domain data, respectively.

B. Isomap

Isomap is conducted by measuring the geodesic distances on the neighborhood graph and applying MDS to obtain the low-dimensional representations of dataset. Isomap works as the following three steps [21].

Step 1 Construct neighborhood graph G. The Euclidean distances between points in the input space are calculated. Each point x_i is connected to all points x_j within a certain radius ε or x_i and x_j are *K*-nearest neighbors. So the neighborhood graph G with the edge $d(i,j)$ can be acquired.

Step 2 Compute the shortest path of G. In graph G, we initialize $d_{\mathrm{G}}(i,j) = d(i,j)$ if an edge links two points x_i and x_j, otherwise $d_{\mathrm{G}}(i,j) = +\infty$. Next, for each value of $m = 1, 2, \ldots, n$, all entries $d_{\mathrm{G}}(i,j)$ are replaced by $\min\{d_{\mathrm{G}}(i,j), d_{\mathrm{G}}(i,m) + d_{\mathrm{G}}(m,j)\},$. The final matrix $D_{\mathrm{G}} = \{d_{\mathrm{G}}(i,j)\}$ will be comprise the shortest paths distances between each pair of points in G.

Step 3 Construct *d*-dimensional embedding. Classical MDS is applied to the matrix D_G and *d*-dimensional embedding coordinate vectors Y_i of Euclidean space Y are constructed. The vectors Y_i in Y are used to minimize the cost function

$$E = \|\tau(D_{\mathrm{G}}) - \tau(D_Y)\|_{L^2} \tag{1}$$

where D_Y refers to the matrix of Euclidean distances $\{d_Y(i,j) = \|y_i - y_j\|\}$ and $\|A\|_{L^2}$ is the matrix norm$\sqrt{\sum_{i,j} A_{ij}^2}$. The operator $\tau(D) = -\mathrm{HSH}/2$, where S is the matrix of squared distances $\{S_{ij} = D_{ij}^2\}$, and H is the centering matrix $\{H_{ij} = \delta_{ij} - 1/n\}$. Let λ_p, ν_p be the pth eigenvalue and eigenvector of the matrix $\tau(D_{\mathrm{G}})$. The global minimum of (1) is ultimately achieved by setting the coordinates y_i to the top *d* eigenvectors of the matrix $\tau(D_{\mathrm{G}})$, that is $Y = [y_1, \ldots, y_n] = [\sqrt{\lambda_1}\nu_1, \ldots, \sqrt{\lambda_d}\nu_d]^T$.

C. Locality Preserving Projection

It attempts to find a transformation matrix with a rule of maintaining the local manifold structure of data, which is used to transform input data to a lower dimensional subspace. The detailed steps of algorithm are listed as follows [25]. Given a dataset of l dimensional and n data points $X = \{x_1, x_2, \ldots, x_n\} \in \boldsymbol{R}^l$, we need to seek a proper transformation matrix *A* that maps X to $Y = \{y_1, y_2, \ldots, y_n\} \in \boldsymbol{R}^d$, so we can acquire $y_i = A^T x_i$, which minimizes the object function as follows:

$$\sum_{ij} (y_i - y_j)^2 S_{ij} \tag{2}$$

where S is a similarity matrix measuring the local relationships of each data point. S is defined as follows. If x_i and x_j are neighbors, then

$$S_{ij} = e^{-\frac{\|x_i - x_j{}^2\|}{t}}, t \in \boldsymbol{R} \tag{3}$$

otherwise $S_{ij} = 0$. Next, following some simple algebraic steps:

$$\begin{aligned}
&\frac{1}{2}\sum_{ij} (y_i - y_j)^2 S_{ij} \\
&= \frac{1}{2}\sum_{ij} \left(A^T x_i - A^T x_j\right)^2 S_{ij} \\
&= \frac{1}{2}\sum_{ij} \left(A^T x_i - A^T x_j\right)\left(A^T x_i - A^T x_j\right)^T S_{ij} \\
&= A^T X (D - S) X^T A \\
&= A^T X L X^T A
\end{aligned} \tag{4}$$

where *D* is a diagonal matrix, $D_{ii} = \sum_j S_{ji}$ and $L = D - S$ represents the Laplacian matrix. So minimizing the objective function turns to deal with a generalized eigenvalue problem

$$XLX^T A = \lambda X D X^T A. \tag{5}$$

In terms of the first *d* minimum eigenvalues, the vectors $a_i (i = 1, 2, \ldots, d)$ being the solutions to (5) constitute the transformation matrix *A*. More discussion of LPP can be consulted in [25].

In this paper, as the number of features from THz time-domain was larger than the number of data points, to avoid singular of the matrix XDX^T we projected the dataset to a lower dimensional subspace by utilizing PCA at first. We selected the first ten principal components of retaining 99.7% original information.

D. Support Vector Machine

The main idea of SVM is to seek an optimal hyperplane that separate the two classes as much as possible by mapping the data to a high-dimensional space through nonlinear transformation [36], [37].The nonlinear transformation is determined by the kernel function. Here we chose the radial basis function. The kernel parameter γ and penalty parameter *C* were obtained by using a mesh grid search combined with a five-fold cross validation.

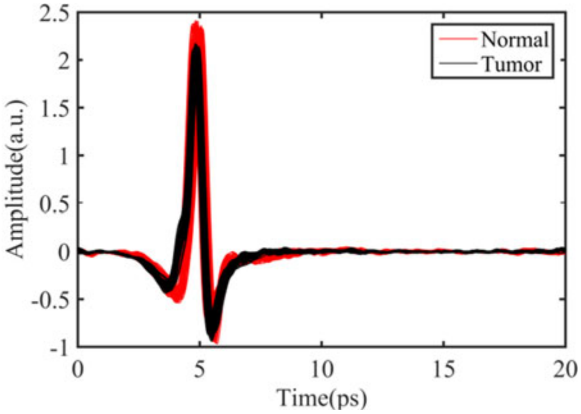


Fig. 2. THz time-domain signals of normal hepatic and tumor samples.

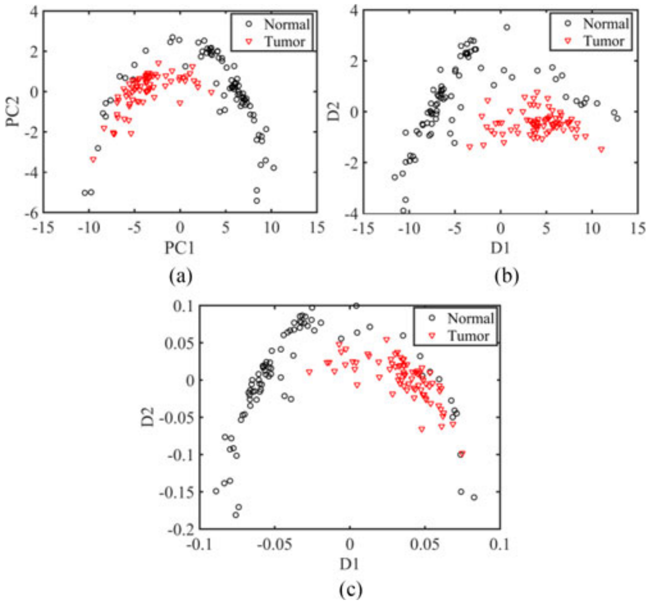


Fig. 3. Two-dimensional scatter plots of time-domain data reduced by (a) PCA. (b) Isomap. (c) PCA+LPP.

E. Probabilistic Neural Network

PNN, a feed forward neural network building on Bayesian decision theory, was first proposed in [38].The network structure of PNN is consisted of input layer, pattern layer, summation layer, and output layer. The input layer distributes the input variables to the neurons of the pattern layer that is comprised of some pattern units. And then a nonlinear operation, using a Gaussian kernel [38], on the value of dot product between the input pattern vector and weight vector is performed in each unit. The summation layer obtains the maximum likelihood of the specific pattern classified into corresponding category by summarizing and averaging the neurons of the same class. Based on the output from summation layer, the output layer classifies each pattern in terms of Bayesian decision theory. The smoothing factor, referring to the width of the Gaussian kernel, could affect the classification accuracy of PNN. We optimized this factor by trial and error. To be specific, we calculated the accuracy of the dataset by setting the factor from 0.1 to 1 with the interval of 0.1. So we found that the best result was acquired when the value of smoothing factor was set to 0.1.

IV. Classification the THz Time-Domain Dataset

For the normal and tumor tissue slices, a total number of eighty sample spots of each type were selected randomly from the collected THz reflection waveforms. Every waveform contained 2000 points within a range of 20 ps. As shown in Fig. 2, the THz signals of normal and tumor samples overlapped and we probably could not distinguish them entirely depending on the small differences of the THz pulses. Thus, it is necessary to discriminate them rely on other tools. Dimensional reduction methods of PCA, Isomap, and PCA+LPP were executed to compare the effect of separation between two classes. For k nearest neighbors of Isomap, we found that the optimal result was achieved when $k = 46$. And the number of k nearest neighbors in PCA+LPP was set to two.

A. Comparing in 2-D Data Visualization

It can be seen from the score scatter plot of PCA: the structure of dataset was severely nonlinear. This indicated that PCA was not able to discover the inner structure of the nonlinear complex data because of the overlapping between the two classes. Confronting with the global nonlinear dataset, similar to PCA, PCA+LPP was not able to reveal the true structure of this dataset as well. So PCA and PCA+LPP were not able to truly reflect the differences of time-domain data from normal hepatic and tumor tissue samples. Nevertheless, Isomap performed well to preserve the geodesic distances between data points (i.e., maintaining the internal structure of the whole data). So in the comparison of Fig. 3, only Isomap could separate the distances between two classes entirely. Therefore, a conclusion is achieved that there likely exists a global nonlinear manifold relationship within this type of THz time-domain data.

B. Results of Classification

In order to further realize the classification effects of three kinds of reduction methods, the effect of SVM and PNN based on PCA, PCA+LPP, and Isomap was compared.

The 2-D data from three dimensionality reduction methods were used as input variables, SVM and PNN were applied to classify data. A total of 160 data points were divided into training set (including 60 data points of each class) and test set (including 20 data points of each class). The classification process was repeated ten times, the averaged accuracy of ten times was regarded as a classification criterion. The results were listed in Table I.

It can be seen that the classifiers based on Isomap was obviously better than the other two. The total accuracies based on Isomap were best and Isomap-PNN performed better than Isomap-SVM. In the remaining two classifiers based on PCA+LPP and PCA, PNN also performed better than SVM, and the effect of classifiers building on PCA were better than PCA+LPP. Besides, the classification effect built on Isomap

TABLE I
CLASSIFICATION ACCURACIES WITH DIFFERENT MODELS BASED ON THz TIME DOMAIN DATA

Model	Accuracy in normal (%)		Accuracy in tumor (%)		Total accuracy (%)
	Train. (%)	Test. (%)	Train. (%)	Test. (%)	
PCA-PNN	95.00 ± 1.57	92.00 ± 6.32	100.0 ± 0.00	97.50 ± 3.50	96.75 ± 0.65
PCA-SVM	95.50 ± 1.58	92.00 ± 7.53	99.50 ± 0.81	95.00 ± 5.27	96.50 ± 1.15
Isomap-PNN	100.00 ± 0.00	99.50 ± 1.58	100.00 ± 0.00	99.00 ± 2.11	99.81 ± 0.30
Isomap-SVM	99.83 ± 0.53	98.00 ± 3.50	100.00 ± 0.00	100.00 ± 0.00	99.69 ± 0.61
PCA+LPP-PNN	92.83 ± 2.49	90.00 ± 7.45	100.00 ± 0.00	97.50 ± 2.64	95.75 ± 0.92
PCA+LPP-SVM	94.33 ± 3.70	92.00 ± 5.37	98.67 ± 1.72	94.50 ± 5.50	95.69 ± 1.73

(Train. represents the training set of the samples and test. represents the testing set of the samples.)

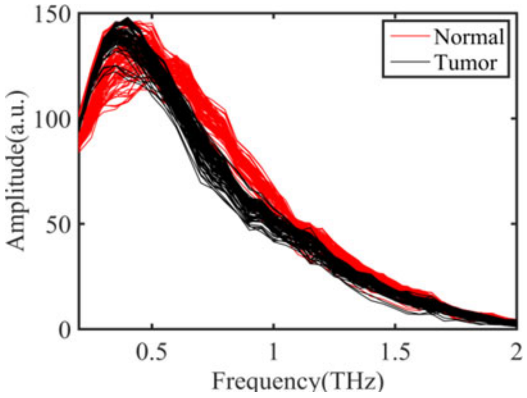


Fig. 4. THz frequency-domain spectra of normal hepatic and tumor specimens.

was more stable, which can be seen from the standard deviation. So the results indicated that Isomap-PNN performed the best for distinguishing hepatic tumor and normal samples. The classification results also manifested that Isomap could preserve the geometrical structure of the THz time-domain data well. The global structures of THz time-domain data were more crucial than those of the local properties.

V. CLASSIFICATION THE THz FREQUENCY-DOMAIN DATASET

Applying FFT to the above THz signals, we acquired the THz frequency-domain spectra of each class in the range of 0.2–2.0 THz, with the resolution of spectra being 50 GHz. Similar to the THz time-domain signal, the spectra of two classes could not be differentiated completely because some normal tissue spectra overlapped the tumor spectra, as shown in Fig. 4. But we may distinguish them with the aid of above techniques.

Here PCA, Isomap, and LPP were employed as a dimensional reduction tool to extract the features of the spectra. For k nearest neighbors of LPP, we found that the optimal result was achieved when $k = 2$. And the number of k nearest neighbors in Isomap was set to ten. So we acquired corresponding 2-D scatter plots and compared with each other.

A. Comparing in 2-D Data Visualization

By observing from above plots, we found that only scatter plot based on LPP could separate the distances between two classes best while some data points of two classes mingled with each other for other two plots. PCA and Isomap are all dimensional reduction techniques that retain the global structure of the dataset, whereas the former is a linear method for dimensionality reduction and the latter is a nonlinear one building on manifold learning. The overall separation effect of Isomap was better than that of PCA, as displayed in Fig. 5(a) and (b), because the between-cluster distance in Isomap was larger. However, some local points between two classes could not be separated either using PCA or Isomap. This indicated that perhaps parts of nonlinear manifold relationship hid in the structure of dataset. Hence some neighbor local structure of dataset should be considered. LPP preserved local structure well, outperforming PCA and Isomap, as verified in Fig. 5. Two types of tissue samples were separated well by employing LPP. Although LPP is a linear method, different from PCA, for discovering the local nonlinear structure of the data, it was more capable of recovering some key parts of the nonlinear manifold structure by comparing with Isomap. We, therefore, infer that parts of the THz frequency-domain data in this type may occupy a nonlinear submanifold of the frequency spectra space.

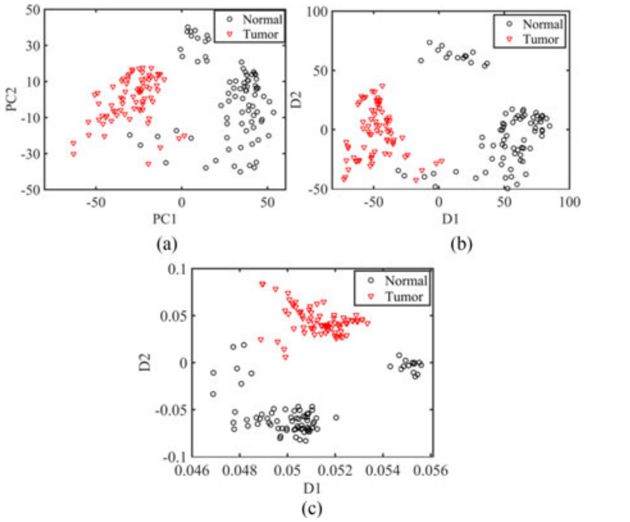


Fig. 5. Two-dimensional scatter plots of frequency-domain data reduced by (a) PCA (b) Isomap (c) LPP.

B. Results of Classification

We used two classifiers, SVM and PNN, based on PCA, Isomap, and LPP to more systematic and accurately compare

TABLE II
CLASSIFICATION ACCURACIES WITH DIFFERENT MODELS BASED ON FREQUENCY DOMAIN DATA

Model	Accuracy in normal (%)		Accuracy in tumor (%)		Total accuracy (%)
	Train. (%)	Test. (%)	Train. (%)	Test. (%)	
PCA-PNN	97.17 ± 1.37	92.00 ± 4.83	99.17 ± 1.18	94.50 ± 4.38	96.94 ± 0.55
PCA-SVM	97.50 ± 2.52	94.00 ± 5.16	97.50 ± 1.96	95.50 ± 4.97	96.81 ± 0.69
Isomap-PNN	97.83 ± 1.58	93.50 ± 5.30	99.50 ± 1.12	96.50 ± 3.37	97.75 ± 0.67
Isomap-SVM	97.67 ± 2.38	94.00 ± 6.58	98.83 ± 1.77	97.00 ± 2.58	97.56 ± 0.95
LPP-PNN	100.00±0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
LPP-SVM	100.00 ± 0.00	98.00 ± 2.58	100.00 ± 0.00	100.00 ± 0.00	99.75 ± 0.32

the classification effects. The classification process and criterion were the same to the previous section. The results were shown in Table II.

It can be seen that the classifiers based on LPP was obviously better than other two, especially, LPP-PNN obtained the accuracy of 100%. Also, the standard deviation of the classifiers based on LPP was smaller than others. It meant that the classification effect built on LPP was more stable. Besides that, for total accuracy in comparison of LPP-PNN and LPP-SVM, PNN performed a little better than SVM. And for standard deviation of total accuracy, PNN was smaller than that of SVM. Overall, PNN outperformed than SVM either in accuracy or stability. Furthermore, the classification results testified that LPP was possessed of more discriminating power than PCA and Isomap. Thus, the discriminating capability of LPP-PNN is the best for frequency-domain data of this experiment.

VI. CONCLUSION

We compared PCA, Isomap, and LPP for dimensionality reduction and classification normal hepatic and tumor tissue specimens coupled with THz time-domain and frequency-domain data. Based on THz time-domain signals, either 2-D data visualization or classification results building on classifiers manifested that the constructed feature space of Isomap was better than that of PCA and PCA+LPP. However, the same operation conducted in the frequency-domain dataset manifested that LPP performed better than PCA and Isomap. The comparative results showed that there was evident global nonlinear structure in this measured THz time-domain data while local nonlinear structure existed within the frequency-domain dataset of this experiment. Isomap could guarantee the true structure of the severe nonlinear data. LPP is a perfect tool to preserve the local structure of THz spectra data and captures more useful local information. Thus, the datasets based on THz time-domain signals and frequency-domain spectra embodied dissimilar physicochemical properties. It can be seen from the experiment, Isomap and LPP are feasible methods to reflect the nonlinear manifold of the THz data. Building on THz-TDS technology, the proposed scheme is a promising method to identify the hepatic tumors.

REFERENCES

[1] F. X. Bosch, J. Ribes, R. Cléries, and M. Díaz, "Epidemiology of hepatocellular carcinoma," *Clin. Liver Dis.*, vol. 7, no. 1, pp. 5–15, 2005.

[2] P. Stefaniuk, J. Cianciara, and A. Wiercinska-Drapalo, "Present and future possibilities for early diagnosis of hepatocellular carcinoma," *World J. Gastroentero.*, vol. 16, no. 4, pp. 418–424, 2010.

[3] M. Patel *et al.*, "Hepatocellular carcinoma: diagnostics and screening," *J. Eval. Clin. Pract.*, vol. 18, no. 2, pp. 335–42, 2012.

[4] J. El Haddad, B. Bousquet, L. Canioni, and P. Mounaix, "Review in terahertz spectral analysis," *TrAC Trends Anal. Chem.*, vol. 44, pp. 98–105, 2013.

[5] I. Amenabar, F. Lopez, and A. Mendikute, "In introductory review to THz non-destructive testing of composite mater," *J. Infrared Millim. THz. Waves*, vol. 34, no. 2, pp. 152–169, 2013.

[6] Y.-C. Shen and P. F. Taday, "Development and application of terahertz pulsed imaging for nondestructive inspection of pharmaceutical tablet," *IEEE J. Sel. Topics Quantum Electron.*, vol. 14, no. 2, pp. 407–415, 2008.

[7] W. Xu *et al.*, "Discrimination of transgenic rice containing the Cry1Ab protein using terahertz spectroscopy and chemometrics," *Sci. Rep.*, vol. 5, 2015. Art. no. 11115.

[8] J. B. Jackson *et al.*, "A survey of terahertz applications in cultural heritage conservation science," *IEEE Trans. THz. Sci. Technol.*, vol. 1, no. 1, pp. 220–231, Sep. 2011.

[9] P. C. Ashworth *et al.*, "Terahertz pulsed spectroscopy of freshly excised human breast cancer," *Opt. Express*, vol. 17, no. 15, pp. 12444–12454, 2009.

[10] Y. B. Ji *et al.*, "Feasibility of terahertz reflectometry for discrimination of human early gastric cancers," *Biomed.Opt. Express*, vol. 6, no. 4, pp. 1398–1406, 2015.

[11] T. C. Bowman, M. El-Shenawee, and L. K. Campbell, "Terahertz imaging of excised breast tumor tissue on paraffin sections," *IEEE Trans. Antennas Propag.*, vol. 63, no. 5, pp. 2088–2097, May 2015.

[12] E. Berry *et al.*, "Multispectral classification techniques for terahertz pulsed imaging: An example in histopathology," *Med. Eng. Phys.*, vol. 26, no. 5, pp. 423–430, 2004.

[13] V. P. Wallace *et al.*, "Terahertz pulsed spectroscopy of human Basal cell carcinoma," *Appl. Spectrosc.*, vol. 60, no. 10, pp. 1127–1133, 2006.

[14] S. Yamaguchi, Y. Fukushi, O. Kubota, T. Itsuji, T. Ouchi, and S. Yamamoto, "Brain tumor imaging of rat fresh tissue using terahertz spectroscopy," *Sci Rep.*, vol. 6, 2016, Art. no. 30124.

[15] M. A. Brun, F. Formanek, A. Yasuda, M. Sekine, N. Ando, and Y. Eishii, "Terahertz imaging applied to cancer diagnosis," *Phys. Med. Biol.*, vol. 55, no. 16, pp. 4615–4623, 2010.

[16] N. Qi, Z. Zhang, Y. Xiang, Y. Yang, and P. B. Harrington, "Terahertz time-domain spectroscopy combined with fuzzy rule-building expert system and fuzzy optimal associative memory applied to diagnosis of cervical carcinoma," *Med. Oncol.*, vol. 32, no. 1, p. 383, 2015.

[17] N. Qi *et al.*, "Terahertz time-domain spectroscopy combined with support vector machines and partial least squares-discriminant analysis applied to diagnosis of cervical carcinoma," *Anal. Methods*, vol. 7, no. 6, pp. 2333–2338, 2015.

[18] L. O. Jimenez and D. A. Landgrebe, "Supervised classification in high-dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data," *IEEE Trans. Syst. Man Cybern. C*, vol. 28, no. 1, pp. 39–54, 1998.

[19] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[20] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: A comparative," *J. Mach. Learn. Res.*, vol. 10, pp. 66–71, 2009.

[21] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[22] I. S. Lim, P. D. H. Ciechomski, S. Sarni, and D. Thalmann, "Planar arrangement of high-dimensional biomedical data sets by isomap coordinates," in *Proc. IEEE Conf. Comput.-Based Med. Syst.*, 2003, pp. 50–55.

[23] Y. Bu, F. Chen, and J. Pan, "Stellar spectral subclasses classification based on Isomap and SVM," *New Astron.*, vol. 28, no. 28, pp. 35–43, 2014.

[24] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[25] X. He and P. Niyogi, "Locality preserving projections," *Adv. Neural Inf. Process. Syst.*, vol. 16, no. 1, pp. 186–197, 2003.

[26] X. He, D. Cai, H. Liu, and W. Y. Ma, "Locality preserving indexing for document representation," in *Proc. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, 2004, pp. 96–103.

[27] T. S. Furey *et al.*, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[28] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, no. 1, pp. 45–66, 2001.

[29] K. Z. Mao, K. C. Tan, and W. Ser, "Probabilistic neural-network structure determination for pattern classification," *IEEE Trans. Neural Netw.*, vol. 11, no. 4, pp. 1009–1016, 2000.

[30] M. Hajmeer and I. Basheer, "A probabilistic neural network approach for modeling and classification of bacterial growth/no-growth data," *J. Microbiol. Methods*, vol. 51, no. 2, pp. 217–226, 2002.

[31] C. B. Reid *et al.*, "Terahertz pulsed imaging of freshly excised human colonic tissues," *Phys. Med. Biol.*, vol. 56, no. 14, pp. 4333–53, 2011.

[32] R. M. Woodward *et al.*, "Terahertz pulse imaging in reflection geometry of human skin cancer and skin tissue," *Phys. Med. Biol.*, vol. 47, no. 21, pp. 3853–3863, 2002.

[33] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Edu. Psychol.*, vol. 24, no. 6, pp. 417–520, 1933.

[34] S. De Backer, Unsupervised Pattern Recognition: dimensionality reduction and classification, Universiteit Antwerpen¸ Antwerp, Belgium, 2002.

[35] R. Bro and A. K. Smilde, "Principal component analysis," *Anal. Methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

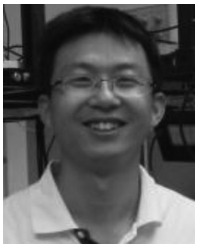
[36] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 2000, pp. 988–999.

[37] V. Kecman, *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models*. Cambridge, MA, USA: MIT Press, 2001.

[38] D. F. Specht, "Probabilistic neural networks," *Neural Netw.*, vol. 3, no. 1, pp. 109–118, 1990.

Haishun Liu was born in 1991. He received the B.S. degree in physics from Yanbian University, Yanji, China, in 2013. He is working toward the master degree in physics at Capital Normal University, Beijing, China.

His research interest focuses on the application of THz spectroscopy.



Zhenwei Zhang was born in 1977. He received the M.S. degree in physics from Capital Normal University, Beijing, China, in 2006.

He is currently an Experimentalist with the Key Laboratory of Terahertz Optoelectronics, Ministry of Education, Capital Normal University. His research interests include nondestructive testing of THz wave and THz spectroscopy.

Xin Zhang, photograph and biography not available at the time of publication.

Yuping Yang, photograph and biography not available at the time of publication.

Zhuoyong Zhang, photograph and biography not available at the time of publication.

Xiangyi Liu, photograph and biography not available at the time of publication.

Fan Wang, photograph and biography not available at the time of publication.

Yiding Han, photograph and biography not available at the time of publication.

Cunlin Zhang, photograph and biography not available at the time of publication.