Predictive data clustering of laser-induced breakdown spectroscopy for brain tumor analysis

GEER TENG,1,2 QIANQIAN WANG,1,2,* XUTAI CUI,1,2 GUOYAN CHEN,1,2 KAI WEI,1,2 XIANGJUN XU,1,2 BUSHRA SANA IDREES,1,2 AND M. NOUMAN KHAN1,2

1 School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China
2 Key Laboratory of Photonic Information Technology, Ministry of Industry and Information Technology, Beijing Institute of Technology, Beijing 100081, China
*qqwang@bit.edu.cn

Abstract: Limited by the lack of training spectral data in different kinds of tissues, the diagnostic accuracy of laser-induced breakdown spectroscopy (LIBS) is hard to reach the desired level with normal supervised learning identification methods. In this paper, we proposed to apply the predictive data clustering methods with supervised learning methods together to identify tissue information accurately. The meanshift clustering method is introduced to compare with three other clustering methods which have been used in LIBS field. We proposed the cluster precision (CP) score as a new criterion to work with Calinski-Harabasz (CH) score together for the evaluation of the clustering effect. The influences of principal component analysis (PCA) on all four kinds of clustering methods are also analyzed. PCA-meanshift shows the best clustering effect based on the comprehensive evaluation combined CH and CP scores. Based on the spatial location and feature similarity information provided by the predictive clustering, the PCA-Meanshift can improve diagnosis accuracy from less than 95% to 100% for all classifiers including support vector machine (SVM), k nearest neighbor (k-NN), soft independent modeling of class analogy (Simca) and random forests (RF) models.

1. Introduction

As a novel atomic spectroscopy technique, laser-induced breakdown spectroscopy (LIBS) has been applied in many identification fields [1,2]. The pulsed laser is focused on the sample surface and ablates the trace materials. The excited sample forms a plasma in local thermodynamic equilibrium (LTE). By detecting the optical radiation generated by the energy level transition during the plasma cooling process, the characteristic spectrum of each substance can be obtained. Based on the difference of the corresponding spectral signals, the classification and identification of substances can be carried out. Especially for the biomedical field, with advantages like micro damage, real-time, no need to inject molecular markers, LIBS may show greater potential.

Spectral lines of specific wavelengths reflect the corresponding element information. Through the mapping of the sample surface, the elemental distribution can be analyzed [3,4]. Meanwhile, LIBS can also identify the half-life cycles of elements in biological tissues [5]. Combined with chemometrics, it can also reveal the dietary intake of different livestock [6]. Besides, LIBS has been used in the diagnosis of several diseases. For instance, the diagnoses of Alzheimer’s disease [7], melanoma [8], blood cancers [9] and cervical cancer [10] based on LIBS technique all achieve high accuracies than traditional medical diagnostic methods. Especially for the tumor diagnostic field, the diagnostic accuracy exceeds 90% for most tumors [8–10].

Since the spectral difference of the analyte is usually very subtle, a chemometric model or machine learning model is usually established for identification [11]. However, building a
classification or identification model usually need a large number of label-known data. In some cases, it is difficult to collect enough label-known data and sample categories not involved in modeling will affect the identification accuracy. Even though some methods to generate the simulated spectra can solve the problem of insufficient data \cite{12}, the accuracy of the supervised learning models trained with simulation data still needs to be improved. Especially in the biomedical field, biological tissue types vary widely and some types are unknown. This makes it difficult to improve the accuracy of the model by simulating existing spectral data. Therefore, in order to achieve the high diagnostic accuracy, the information besides the sample labels needs to be considered together. For example, unsupervised clustering may play an important role for tissue spectroscopic analysis.

Several clustering methods have been proposed to process the LIBS data, for instance, K-means \cite{13–15}, hierarchical clustering analysis (HCA) \cite{16,17} and Iterative Self-organizing Data Analysis Techniques Algorithm (ISODATA) \cite{16}. These clustering methods distinguish the messy spectral data according to the data distribution. It has always been the focus of research in this field, not only for the preliminary judgment of qualitative analysis, but also for the preliminary processes of quantitative analysis \cite{18–20}. For instance, using clustering algorithms to evaluate the classifiers in ensemble learning can effectively improve quantitative accuracy \cite{20}. However, several types of clustering errors during the analyzing process are also noticed. Among them, there are some typical ones like that, the number of clusters is different from the actual number of sample categories, different kinds of samples are grouped into the same cluster, and the same kind of samples are grouped into different clusters. Each clustering method may generate different clustering errors during the application process. There is also a lack of appropriate evaluation criteria of clustering methods. Usually, the distance between clusters and distance within clusters are used as the evaluation criteria of clustering effect. Among them, the most representative indicator is the Calinski-Harabasz (CH) score \cite{21}. This is an internal quality evaluation standard, the higher the CH score, the better the clustering effect. However, this evaluation method is not comprehensive. Even if the internal quality based on distance is very high, clustering errors still may occur. In addition to internal quality of clustering performance evaluation, external quality should also be considered.

Meanwhile, we often hope to see the distribution of data points intuitively, but static images can only reflect 2D and 3D information. The data distribution of high-dimensional data cannot be visualized and analysis of high-dimensional data takes a long time. In order to reduce the analyzing time and improve the recognizability of data features, some dimensionality reduction methods like principal component analysis (PCA) have been used in the LIBS analysis field \cite{22}. After the dimensionality reduction based on PCA, the first few principal components (PCs) retain most of the original variance information. Meanwhile, after the mapping processing, the data distribution has also been changed. The visual data distributions can be analyzed by selecting the scores from two or three PCs for plotting \cite{23,24}. However, only clustering based on the distribution value ranges of PC scores cannot distinguish all the samples well \cite{25,26}. It partially relies on the operator who set the value ranges. Especially for the non-linear data, the value ranges of PC scores become insufficient for clustering \cite{27}. The PCA method still needs to be combined with other chemometric methods to cluster LIBS data, objectively \cite{28}. Sometimes, during the clustering process, using PC scores as model inputs can show better performance \cite{29}. However, the effect of PCA on each clustering algorithm is still unclear. Furthermore, cluster analysis can only divide samples into different categories, and cannot specifically identify the types of samples. Identification based on supervised learning is still an important analysis step. Combining it with unsupervised clustering learning methods may further improve the sample analysis mechanism. Although some supervised learning methods have been used to further identify the samples in one cluster \cite{30}, the previous method is still not an improvement to the
In this paper, we introduce the Meanshift algorithm to cluster the LIBS spectra of brain tumor and infiltrating tissues. The Meanshift and three other clustering algorithms that have been used are compared based on the original spectral data and spectral PC scores. In order to comprehensively evaluate the performance of each clustering method, we propose an evaluation index as cluster precision (CP) score to use together with CH score. The higher the CP score is, the better the clustering effect will be. The CP score is based on the correctness of the predictive number of classes and the correctness of the sample point clustering results. CH score evaluates from the internal factors of the data, and more consider the influence of the data distribution. While it ignores the influence of other factors on the clustering results besides the data. These external factors may include the degree of adaptation between the data and the clustering algorithm. As a supplement, CP score can evaluate these factors from the perspective of clustering results. Therefore, evaluation based on both CH and CP scores is more conducive to comprehensively measuring the performance of a clustering method. Meanwhile, we evaluated the impact of dimensionality reduction method PCA on the performance of each clustering method. Through the evaluation, we found the best combination of dimensionality reduction and clustering algorithm to do the pre-clustering of spectral data. At the same time, the supervised learning identification of spectral data was performed and the diagnosis accuracy of brain tumor and infiltrating tissues was evaluated on the basis of both clustering and identification. When perform the identification and classification based on LIBS, because of the different choices of classification models, the LIBS identification results may have differences in accuracy for the same spectral data. This reflects the data dependence of various models. Therefore, in different application fields, we usually need to evaluate and select the best model. The application in clinical brain tumor diagnosis is also according to this process. The proposed technology can significantly improve the accuracy of LIBS tumor diagnosis and make the diagnostic accuracy is no longer obviously dependent on the selection of the supervised classification model. In the future, this technology can promote the better application of LIBS in the field of disease diagnosis.

2. Materials and methods

2.1. LIBS diagnostic setup

The schematic of the experimental LIBS setup for brain samples measurement has been illustrated in our previous work [31,32]. So we just mention some important experimental details here. A desktop LIBS system was built to measure excised tissue. A homemade He-Ne laser ($\lambda = 632.8$ nm) was used as pointing laser to indicate ablation points. A homemade flash-pumped Q-switched Nd: YAG laser ($\lambda = 1064$ nm, pulse frequency 1 Hz, pulse duration $\tau = 5$ ns, beam diameter Ø6 mm, energy 40 mJ/pulse) was used to excite the sample’s surface. The laser was focused on the sample surface by a convex lens with a focal length of 100 mm. The plasma radiation was collected into the fiber (Ø 600 $\mu$m) through a convex lens with a focal length of 36 mm. The outlet of optical fiber was connected to a two-channel spectrometer (AvaSpec 2048-2-USB2, Avantes). The spectrometer covered a range of 190 nm to 1100 nm with a resolution of 0.2~0.3 nm. External trigger used in the system included a photodetector and a digital delayer (SRS-DG535, Stanford Research System). When the photodetector detected the plasma radiation signal, the spectrometer was triggered by DG535 after a preset delay time. This preset spectral acquisition delay time was 1.29 $\mu$s. The integration time of CCD was 2 ms. A three-dimensional motorized stage was used to adjust the focus position of the laser on the samples. Each shot was focused on a fresh position.
2.2. Sample collection

Human brain tumor tissues were obtained from the Department of Pathology in Sanbo Brain Hospital after routine tumor surgery. Brain tumor and infiltrating tissue samples were collected from the same glioma patient. As the most common primary tumor in the central nervous system, glioma infiltrates from the primary tumor to surrounding healthy tissue [33]. The infiltrating tissue has the similar color and texture compared with both glioma and healthy brain tissue, even the LIBS spectra of them are very similar [31]. Therefore, the glioma and infiltrating tissue are representative to be used as unknown tissue of the identification in brain tumor surgery, and the results can play an evaluation role in the application of clustering algorithms in tissue identification. For each kind of the sample, half was made as paraffin-embedded samples and used for pathological examination, and the other half was prepared freshly on the glass slides for LIBS measurement. All the spectra belong to fresh samples. 1200 original spectra for each kind of sample were collected. In order to reduce spectral fluctuations, the most similar 50 percentage spectra were chosen based on the ascending order of cosine values of the inner angle between each spectrum and the average spectrum according to our prior experience [34]. Then, an average value for every three adjacent spectral data was calculated and we got 200 spectra for each kind of tissue. The average LIBS spectra of glioma and infiltrating boundary tissue are shown in Fig. 1. We calculated the relative standard deviation of each line in the original spectrum and the final spectrum, and took the average value of the RSD of all lines as the RSD of this data type. As shown in Table 1, compared with the original data, the finally obtained spectral data has a higher degree of stability.

![Fig. 1. The average LIBS spectra of glioma and infiltrating boundary tissue.](image)

| Spectral data types                  | Average RSD values (%) | Original spectra | Final spectra |
|-------------------------------------|------------------------|-----------------|--------------|
| Glioma                              | 19.38                  | 3.44            |
| Infiltrating boundary tissue        | 15.58                  | 3.71            |
2.3. Cluster evaluation criteria and the meanshift clustering method

Usually, the distance between clusters and distance within clusters are used as the evaluation criteria of clustering effect. Among them, the most representative indicator is the Calinski-Harabasz (CH) score [21]. As formula (1) shown, this is an internal quality evaluation standard.

\[
CH = \frac{\text{tr}(B)/(k - 1)}{\text{tr}(W)/(M - k)}
\]

where \(\text{tr}(B)\) is the trace of distance difference matrix between classes, \(\text{tr}(W)\) is the trace of distance difference matrix within one class, \(M\) is the quantity of total spectral data, \(k\) is the predictive number of clusters. When all of the data were predicted as one cluster, the CH score is 0. The higher the CH score is, the better the clustering effect will be.

As formula (2) illustrated, we proposed another evaluation index as cluster precision (CP) score. The CP score measures the accuracy of the clustering from the external clustering results. It takes into account the deviation of the number of clusters and the misclassification of points within the cluster, and supplements the factors that cannot be measured by internal indicators.

\[
CP = 100 \times \frac{\min(k, t)}{\max(k, t)} \sum_{i=1}^{k} \frac{n_i}{M}
\]

where \(t\) is the actual number of clusters, \(i\) is the index of current cluster, and \(n_i\) is the quantity of spectral data of main sample class in current cluster. The total sample amount appears in the denominator, and the number of correctly clustered samples appears in the numerator, so the higher the CP score is, the better the clustering effect is.

The CH score evaluates the distance characteristics of clustering, and the CP score evaluates the clustering accuracy. The comprehensive evaluation of the two criteria can reflect the comprehensive performance of the clustering algorithms. Furthermore, the parameters \(k\) and \(t\) will directly affect the value of the CP score. Because the smaller value of \(k\) and \(t\) is in the numerator and the larger value is in the denominator, no matter whether the number of predicted clusters is larger or smaller than the actual cluster number, the CP score will be reduced proportionally. This shows that no matter if there are more or less predictive clusters, there will be a uniform bias trend. Only when \(k\) is equal to \(t\), the CP score is the highest and shows the better clustering effect. For the influence caused by the points of the clustering error in each cluster, it is reflected by the summation term in the formula.

We introduce the Meanshift algorithm to cluster the LIBS spectra of brain tumor and infiltrating tissues. The Meanshift algorithm was first proposed in 1975 for the pattern recognition [35], and nowadays it has been recognized as a robust approach toward feature space analysis [36]. It has played a good role in image processing fields such as satellite applications [37]. The Meanshift and three other clustering algorithms (K-means, ISODATA and HCA) that have been used in LIBS field are compared based on the spectral data and spectral PC scores, respectively. The flowchart of the Meanshift algorithm is shown in Fig. 2. First, randomly select a point from the unmarked data points as the starting center point. All data points appearing in the area centered on this point with a certain radius \(R\) are classified into one cluster. Calculate the average vector distance between the center point and each point in the surrounding area as the vector \(\text{shift}\). Then, the center moves along the \(\text{shift}\) direction and the moving distance is its absolute value. If the vector \(\text{shift}\) converges to zero, this cluster has been distinguished completely. If the distance between the center and another existing cluster center is less than the threshold, these two clusters will be merged. Repeat this process until there are no un-clustered points.
Fig. 2. The flowchart of Meanshift clustering method (db: the distance between center and another existing cluster center, dt: the distance threshold).
3. Results and discussion

3.1. Predictive data clustering based on the spectra

As shown in Fig. 1, 21 lines and molecular bands can be distinguished and they are listed in Table 2 with their corresponding elements or molecules. The 21 spectral lines and bands in the whole spectral range are related to 8 elements (Mg, Ca, Na, H, N, K, O and C) and 6 molecular band clusters (2 CN clusters and 2 C\textsubscript{2} clusters). The number of emission lines measured by the tumor tissue and the boundary tissue is exactly the same, which may be mainly due to two reasons. One of them is that the two types of samples are still biological brain tissues in nature, with the same element types, only the content of some elements is different. The second is that although the boundary tissue is close to healthy brain tissue, it still contains some cancer cells, which further causes it to resemble tumor tissue. In the tumor tissue, it can be seen that the intensity of the emission line of Ca element is slightly higher than that of the boundary tissue, which is mainly caused by the possible calcification in the tumor area. Although Mg is often associated with tumors, the difference in the spectral lines of Mg in the two types of sample spectra is small, and it is difficult to distinguish by visual observation. Since the spectral difference shown in Fig. 1 is very subtle, a chemometric model or machine learning model should be established for analysis. Using the spectral data as inputs, Meanshift and three other kinds of clustering methods are evaluated. As mentioned before, the clustering models can divide samples into different categories, but cannot specifically identify the types of samples.

| Elements and molecular fragments | Line/Band Wavelength(nm) |
|----------------------------------|--------------------------|
| Mg II                            | 279.1, 279.8             |
| Ca II                            | 393.4, 396.8             |
| Ca I                             | 422.7                    |
| Na I                             | 589.0, 589.6, 818.3, 819.5|
| H I                              | 656.3                    |
| K I                              | 766.5, 769.9             |
| O I                              | 777.4                    |
| CN (1,0)                         | 358.4                    |
| CN (0,0)                         | 384.6, 385.7, 386.5, 388.0|
| C\textsubscript{2} (0,0)         | 516.8, 517.9             |
| C\textsubscript{2} (0,1)         | 558.5                    |

As shown in Fig. 3, the clustering result of each method is different, and the number of clusters is also different. Through the Meanshift algorithm, the two kinds of tissues were identified into one cluster. The radius $R$ in the Meanshift algorithm was determined as the decile of average distance between each data point and the first given center. The threshold for cluster merging is 0 by default. Meanwhile, the ISODATA clustered the samples as four types. Different with three other kinds of clustering methods, the K-means algorithm need to preset the cluster numbers. In our analytical process, according to the number of known sample types, the number of clusters is preset to 2. So, the CP score cannot reflect the automatic clustering accuracy. The presetting process brings inconvenience to its practical application.

Evaluated by the CH score and the proposed CP score, the results are listed in Table 3. The Meanshift cannot cluster the spectral data as two types. Due to this reason, it is apparently not a good method to do the predictive data clustering. With the highest CH and CP scores, the HCA
Fig. 3. The clustering results of K-means, Meanshift, ISODATA and HCA.

is the best clustering method. Meanwhile, the two indicators reveal the effect of clustering from
different aspects, but for different clustering methods, the shown trends are similar.

Table 3. The evaluation results of clustering abilities based on spectral data

| Cluster methods | K-means  | Meanshift | ISODATA | HCA     |
|-----------------|----------|-----------|---------|---------|
| CH score        | 102.33   | 0         | 42.70   | 123.63  |
| CP score(%)     | 94.75    | 25.00     | 49.50   | 99.00   |

3.2. Predictive data clustering based on the PCA scores

In order to further evaluate the impact of the dimensionality reduction method PCA on the
clustering algorithms, we preform the PCA of spectral data and using PC scores for the clustering.

The cumulative variances of different number of principle components (PCs) are illustrated in
Fig. 4. In many situations, the first several PCs can already represent a large part of cumulative
variances. However, for the spectral data of brain tumor tissues, it is different. The information
distribution is relatively even. The first three PCs only represent a cumulative variance at 60.65%.
The first 399 PCs represent the whole 100% of variances. The uniform distribution of PCA data
for brain tumors is related to the nature of samples. The content of various ions in biological
tissues is related. The balance of element ions reflects the physiological state and has a high
degree of correlation with each other. Therefore, the correlation of variables in the spectral data
is also relatively high. The PCA score distribution is relatively uniform.

By choosing 2 or 3 PCs for visual data representation, we can visually observe the clustering
data distribution in Fig. 5. Through the two PCs plotting, we can roughly see the clusters, but the
boundary still partially relies on subjective judgment by the operator. Through the three PCs
plotting, it can be found that the increasing of PC dimensions does not significantly improve the intuitive clustering effect. The scores of first two PCs are used in the further clustering analysis. The results of PC score clustering based on the same clustering methods are shown in Fig. 6. We can find that the clustering result of Meanshift shows a great improvement but the effect on the ISODATA is non-obvious.

In order to quantify the impact of dimensionality reduction method PCA on the clustering method, we still calculated the CH score representing internal information and the CP score representing external information. As listed in Table 4, for the CH score, all four kinds of clustering methods show an increasing, which indicates that the PCA can improve the data distribution to some extent. However, the effects on the CP score are different. The PCA process changes the data distribution, but the clustering results do not only rely on the data distribution. The clustering result is also related to the principle of the adopted clustering algorithm. Cluster analysis calculation methods can mainly be divided into the following types: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and
model-based methods. Among them, the most widely used is partitioning methods like K-means and ISODATA, hierarchical methods like HCA and density-based methods like Meanshift. Compared with the other two types, the density-based method is more affected by the data distribution.

Table 4. The evaluation results of clustering abilities based on spectral PCA scores

| Cluster methods | PCA-K-means | PCA-Meanshift | PCA-ISODATA | PCA-HCA |
|-----------------|-------------|---------------|-------------|---------|
| CH score        | 264.40      | 311.80        | 268.80      | 297.68  |
| CP score(%)     | 94.75       | 100.00        | 49.38       | 98.25   |

Comparing the original spectra clustering and PC scores clustering together in Fig. 7, we can see that the PCA improves both CH and CP scores of Meanshift greatly. The PCA-Meanshift shows the best clustering performance among all the methods. This may be due to the principle of Meanshift is mainly based on the data distribution. PCA improves the distribution of data, while Meanshift mainly clusters different categories based on data distribution. Compared with the other three algorithms, the Meanshift algorithm assumes that the data sets of different clusters conform to different probability density distributions, and it finds the same kind of sample points through the fastest direction of the increase of the sample point density. It can handle clusters of any shape, and the clustering results are relatively stable. Compared with the similarity, which is the basis of HCA, the sample point density is more affected by the data distribution, so the change in the data space has a greater impact on it. Compared with the K-means clustering algorithm, the Meanshift algorithm does not need to set the number of clusters, making analysis and application more convenient. Similarly, Meanshift requires fewer parameters than ISODATA.
3.3. Predictive data clustering combined with supervised learning methods

The CP score of PCA-Meanshift is exactly 100%, which means that all data are clustered well. Combined with the supervised subsequent classifiers, this predictive data clustering method can be used to improve the identification accuracy. In this work, the data of the training set and the test set are collected from different areas of the tissue, and do not contain the same data each other. As commonly used simple chemometric classification models in LIBS field, the SVM and k-NN models for the identification of glioma and infiltrative have been built by our group based on the limited spectra [31]. However, the highest accuracies of k-NN and SVM are only 89% and 95%, respectively. Although the diagnostic accuracy is higher than the existing methods, we hope that the diagnosis accuracy can be higher for the diagnosis during the surgery. Therefore, for the same dataset, we perform the PCA-Meanshift for clustering first. Then, in each cluster, all data are labeled based on the majority of predicted labels based on supervised methods in this cluster. The whole process about combining predictive data clustering with supervised learning methods is shown in Fig. 8. In all the measured raw data, 75% of the data are divided into the
training set and 25% were divided as testing set. Then the trained supervised classifiers are used to predict the sample labels of testing set data. When combining predictive data clustering with supervised classifiers, there’s no need to train models again. We just perform predictive clustering analysis on the testing data set. Then, the predicted labels of the supervised classifier and the results of the predictive clustering are analyzed comprehensively. The label with the largest number in each cluster is used as the label of all sample points in the cluster.

As shown in Fig. 9, the diagnostic accuracy achieves 100% for both SVM and k-NN. Compared with the highest accuracy without the predictive clustering, the proposed method increases the diagnostic accuracy to an ideal level. Furthermore, we also analyze other supervised learning methods like Random Forests (RF) and Simca which are usually used in the LIBS analysis. Compared with SVM and k-NN, these two kinds of models have lower diagnostic accuracies at 84% and 81%, respectively. Even though, the diagnostic accuracies can be improved to 100% through the proposed pre-clustering method, which are illustrated in Fig. 9.

![Fig. 9. The diagnostic accuracies of four supervised learning models with or without predictive data clustering.](image)

In this proposed hybrid analytical process, the predictive data clustering plays a role in rigorous clustering based on data characteristics. Afterwards, specific type labels are assigned to each cluster based on the supervised learning methods. Different from directly determining the specific label of each cluster through some known label data in the cluster, the entire hybrid analytical process is completely through model prediction. Once the model is established, the analysis can be completed independently without the need for known data to participate in clustering or classification.

4. Conclusions

Discrimination of the infiltrative boundary of glioma is an important issue that plagues clinical medicine. In this work, a predictive clustering method is proposed to improve the accuracy supervised classification models for identifying glioma and boundary tissue. The predictive data clustering process is carried out individually with the supervised diagnostic process. The result of pre-clustering is used to modify the diagnosis result of supervised learning models. The hybrid method combining the predictive clustering and supervised learning can improve diagnosis accuracy to an ideal level as 100%, and the proposed pre-clustering method greatly reduces the
dependence of diagnostic accuracy on the supervised models. The predictive clustering has great potential in clinical tissue distinguishing.

At the same time, in this work, we propose a new index CP score for evaluating clustering methods. Combining CH and CP scores can evaluate clustering performance more comprehensively and select the best predictive data clustering method. Then, we evaluate the impact of PCA coordinate transformation on different clustering methods. The dimensionality reduction method PCA has different effects on different clustering methods. It depends on the dependence of the clustering algorithm on the data distribution. The density-based clustering methods like Meanshift are more affected by the data distribution. Therefore, PCA-Meanshift shows the best clustering effect among all compared methods.

Limited by the difficulty of accumulating patient samples, this study is still a preliminary conclusion. As the number of patient cases increases, the model needs to be further optimized. In the future, we should do more intraoperative experimental verification based on multi-center, large case samples. Furthermore, this method should be extended to more situations with unknown tissue types.

**Funding.** National Natural Science Foundation of China (62075011, 61775017); Graduate Technological Innovation Project of Beijing Institute of Technology (2019C2X20026).

**Disclosures.** The authors declare no conflicts of interest.

**Data availability.** Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the corresponding author upon reasonable request.

**References**

1. J. El Haddad, L. Canioni, and B. Bousquet, “Good practices in LIBS analysis: review and advices,” Spectrochim. Acta, Part B 101, 171–182 (2014).
2. V. Pulleschi, “Laser-induced breakdown spectroscopy: principles of the technique and future trends,” ChemTexts 6(2), 18 (2020).
3. Y. Gimenez, B. Bousquet, F. Trichard, A. Kulesza, J. M. Laurent, V. Zaun, F. Lux, J. M. Benoit, G. Panczer, P. Dugourd, O. Tillement, F. Pelascini, L. Sancey, and V. Motto-Ros, “3D Imaging of Nanoparticle Distribution in Biological Tissue by Laser-Induced Breakdown Spectroscopy,” Sci Rep 6(1), 29936 (2016).
4. H. Wei, Z. Zhao, Q. Lin, and Y. Duan, “Study on the molecular mechanisms against human breast cancer from insight of elemental distribution in tissue based on laser induced breakdown spectroscopy (LIBS),” Biol. Trace Elem. Res. 199(5), 1686–1692 (2021).
5. Y. Chu, Z. Zhang, Q. He, F. Chen, Z. Sheng, D. Zhang, H. Jin, F. Jiang, and L. Guo, “Half-life determination of inorganic-organic hybrid nanomaterials in mice using laser-induced breakdown spectroscopy,” J. Adv. Res. 24, 353–361 (2020).
6. Z. A. Abdel-Salam, V. Pulleschi, and M. A. Harith, “Study of the feeding effect on recent and ancient bovine bones by nanoparticle-enhanced laser-induced breakdown spectroscopy and chemometrics,” J. Adv. Res. 17, 65–72 (2019).
7. R. Gaudioso, E. Ewusi-Annan, W. Xia, and N. Melikechi, “Diagnosis of Alzheimer’s disease using laser-induced breakdown spectroscopy and machine learning,” Spectrochim. Acta, Part B 171, 105931 (2020).
8. J. H. Han, Y. Moon, J. J. Lee, S. Choi, Y.-C. Kim, and S. Jeong, “Differentiation of cutaneous melanoma from surrounding skin using laser-induced breakdown spectroscopy,” Biomed. Opt. Express 7(1), 57–66 (2016).
9. Y. Chu, F. Chen, Z. Sheng, D. Zhang, S. Zhang, W. Wang, H. Jin, J. Qi, and L. Guo, “Blood cancer diagnosis using ensemble learning based on a random subspace method in laser-induced breakdown spectroscopy,” Biomed. Opt. Express 11(8), 4191–4202 (2020).
10. J. Wang, L. Li, P. Yang, Y. Chen, Y. Zhu, M. Tong, Z. Hao, and X. Li, “Identification of cervical cancer using laser-induced breakdown spectroscopy coupled with principal component analysis and support vector machine,” Lasers Med Sci 33(6), 1381–1386 (2018).
11. J. Vařek, E. Képes, L. Duponchel, V. Motto-Ros, C. Fabre, S. Conmennam, F. Schreckenberg, P. Prasse, D. Riebe, R. Junjuri, M. K. Gundawar, X. Tan, P. Polžíka, and J. Kaiser, “Classification of challenging LIBS soil sample data - EMSLIBS contest,” Spectrochim. Acta, Part B 169, 105872 (2020).
12. G. Feng, Q. Wang, J. Kong, L. Dong, X. Cui, W. Liu, K. Wei, and W. Xiangli, “Extending the spectral database of laser-induced breakdown spectroscopy with generative adversarial nets,” Opt. Express 27(5), 6958–6969 (2019).
13. Q. Wang, L. He, Y. Zhao, Z. Peng, and L. Liu, “Study of cluster analysis used in explosives classification with laser-induced breakdown spectroscopy,” Laser Phys. 26(6), 065605 (2016).
14. Y. Guo, Y. Tang, Y. Du, S. Tang, L. Guo, X. Li, Y. Lu, and X. Zeng, “Cluster analysis of polymers using laser-induced breakdown spectroscopy with K-means,” Plasma Sci. Technol. 20(6), 065505 (2018).
15. Y. Tang, Y. Guo, Q. Sun, S. Tang, J. Li, L. Guo, and J. Duan, “Industrial polymers classification using laser-induced breakdown spectroscopy combined with self-organizing maps and K-means algorithm,” Optik 165, 179–185 (2018).
Research Article

Vol. 12, No. 7 / 1 July 2021 / Biomedical Optics Express

16. L. He, Q. Wang, Y. Zhao, L. Liu, and Z. Peng, “Study on cluster analysis used with laser-induced breakdown spectroscopy,” Plasma Sci. Technol. 18(6), 647–653 (2016).
17. T. A. Teklemariam and J. Gotera, “Application of laser induced breakdown spectroscopy in food container glass discrimination,” Spectrochim. Acta, Part B 155, 34–43 (2019).
18. M. Konstantinidis, K. Cote, E. A. Lalla, G. Zhang, M. G. Daly, X. Gao, and P. Dietrich, “On the application of a novel linear mixture model on laser-induced breakdown spectroscopy: implications for Mars,” J. Chemometr. 33(10), e3174 (2019).
19. C. Wu, D. Sun, M. Su, Y. Yin, W. Han, Q. Lu, and C. Dong, “Quantitative analysis of Pb in soil samples by laser-induced breakdown spectroscopy with a simplified standard addition method,” J. Anal. At. Spectrom. 34(7), 1478–1484 (2019).
20. X. Li, J. Yang, F. Chang, X. Zheng, and X. He, “LIBS quantitative analysis for vanadium slags based on selective ensemble learning,” J. Anal. At. Spectrom. 34(6), 1135–1144 (2019).
21. V. M. Vergara, M. Salman, A. Abrol, F. A. Espinoza, and V. D. Calhoun, “Determining the number of states in dynamic functional connectivity using cluster validity indexes,” J. Neurosci. Meth. 337, 108651 (2020).
22. P. Polžíka, J. Klus, E. Képeš, D. Procházka, D. W. Hahn, and J. Kaiser, “On the utilization of principal component analysis in laser-induced breakdown spectroscopy data analysis, a review,” Spectrochim. Acta, Part B 148, 65–82 (2018).
23. N. Baskali-Bouregaa, M.-L. Milliand, S. Mauffrey, E. Chabert, M. Forrestier, and N. Gilon, “Tea geographical origin explained by LIBS elemental profile combined to isotopic information,” Talanta 211, 120674 (2020).
24. Q. Wang, G. Teng, C. Li, Y. Zhao, and Z. Peng, “Identification and classification of explosives using semi-supervised learning and laser-induced breakdown spectroscopy,” J. Hazard. Mater. 369, 423–429 (2019).
25. H. A. Harun and R. Zainal, “Evaluation of a thermoelectric cooler as a sample pre-treatment method for laser-induced breakdown spectroscopy (LIBS) analysis of liquid samples,” Appl. Spectrosc. 74(3), 285–294 (2020).
26. Y. Zhao, Q. Wang, X. Cui, G. Teng, K. Wei, and H. Liu, “Discrimination of hazardous bacteria with combination laser-induced breakdown spectroscopy and statistical methods,” Appl. Opt. 59(5), 1329–1337 (2020).
27. L. Brunnbauer, S. Larisegger, H. Lohninger, M. Nehiebel, and A. Limbeck, “Spatially resolved polymer classification using laser induced breakdown spectroscopy (LIBS) and multivariate statistics,” Talanta 209, 120572 (2020).
28. M. Dong, L. Wei, J. J. González, D. Oropeza, J. Chirinos, X. Mao, J. Lu, and R. E. Russo, “Coal discrimination analysis using tandem laser-induced breakdown spectroscopy and laser ablation inductively coupled plasma time-of-flight mass spectrometry,” Anal. Chem. 92(10), 7003–7010 (2020).
29. W. Liao, Q. Lin, S. Xie, Y. He, Y. Tian, and Y. Duan, “A novel strategy for rapid detection of bacteria in water by the combination of three-dimensional surface-enhanced Raman scattering (3D SERS) and laser induced breakdown spectroscopy (LIBS),” Anal. Chim. Acta 1043, 64–71 (2018).
30. H. Peng, G. Chen, X. Chen, Z. Lu, and S. Yao, “Hybrid classification of coal and biomass by laser-induced breakdown spectroscopy combined with K-means and SVM,” Plasma Sci. Technol. 21(3), 034008 (2019).
31. G. Teng, Q. Wang, H. Zhang, W. Xiangli, H. Yang, X. Qi, X. Cui, B. S. Idrees, K. Wei, and M. N. Khan, “Discrimination of infiltrative glioma boundary based on laser-induced breakdown spectroscopy,” Spectrochim. Acta, Part B 165, 105787 (2020).
32. G. Teng, Q. Wang, H. Yang, X. Qi, H. Zhang, X. Cui, B. S. Idrees, W. Xiangli, K. Wei, and M. N. Khan, “Pathological identification of brain tumors based on the characteristics of molecular fragments generated by laser ablation combined with a spiking neural network,” Biomed. Opt. Express 11(8), 4276–4289 (2020).
33. C. Velásquez, S. Mansouri, C. Mora, F. Nassiri, S. Suppiah, J. Martino, G. Zadeh, and J. L. Fernández-Luna, “Molecular and clinical insights into the invasive capacity of glioblastoma cells,” J. Oncol. 2019, 1–16 (2019).
34. Q. Wang, G. Teng, X. Qiao, Y. Zhao, J. Kong, L. Dong, and X. Cui, “Importance evaluation of spectral lines in Laser-induced breakdown spectroscopy for classification of pathogenic bacteria,” Biomed. Opt. Express 9(11), 5837–5850 (2018).
35. K. Fukunaga and L. D. Hostetler, “The estimation of the gradient of a density function, with applications in pattern recognition,” IEEE Trans. Inf. Theory 21(1), 32–40 (1975).
36. D. Comuniciu and P. Meer, “Mean shift: a robust approach toward feature space analysis,” IEEE Trans. Pattern Anal. Machine Intell. 24(5), 603–619 (2002).
37. X. Li, L. Zhang, L. Wang, and X. Wan, “Effects of BOW model with affinity propagation and spatial pyramid matching on polarimetric SAR image classification,” IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 10(7), 3314–3322 (2017).