Grade Diagnosis of Human Glioma Based on Fingerprint and Artificial Neural Network

Wenyu Peng
Xi'an Jiaotong University
Shuo Chen
National Innovation Institute of Defense Technology
Dongsheng Kong
Chinese People's Liberation Army General Hospital
Xiaojie Zhou
Shanghai Advanced Research Institute, Chinese Academy of Science
Xiaoyun Lu
Xi'an Jiaotong University
Chao Chang (changc@xjtu.edu.cn)
National Innovation Institute of Defense Technology

Research

Keywords: glioma, grade, Fourier transform infrared (FTIR), artificial neural network (ANN), fingerprint, principal component analysis-linear discriminate analysis (PCA-LDA)

Posted Date: February 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-166932/v1

Grade diagnosis of human glioma based on fingerprint and artificial neural network

Wenyu Peng[1,2*], Shuo Chen[2*], Dongsheng Kong[3], Xiaojie zhou[4], Xiaoyun Lu[1#], and Chao Chang[1,2 #]

[1]*Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science, Xi'an Jiaotong University, Xi'an 710049, China*

[2]*Innovation Laboratory of Terahertz Biophysics, National Innovation Institute of Defense Technology, Beijing 100071, China*

[3]*Department of Neurosurgery, Chinese People's Liberation Army (PLA) General Hospital, Beijing, P.R. China*

[4]*National Facility for Protein Science in Shanghai, Shanghai Advanced Research Institute, Chinese Academy of Science, Shanghai 201210, China.*

[#]Corresponding author: changc@xjtu.edu.cn, luxy05@126.com

[*]Equal first author

Abstract

Background

The World Health Organization (WHO) grade diagnosis of cancer is essential for surgical outcomes and patient treatment. Traditional pathological grading diagnosis depends on dyes or other histological approaches, which are time-consuming (usually 1-2 days), resource-wasting, and labor-intensive. Fourier transform infrared (FTIR) spectroscopy is a rapid and nondestructive technique that has been widely used for detecting the molecular component changes, which relies on the resonant frequencies absorbance of the molecular bonds.

Methods

To overcome the disadvantages of traditional pathological diagnosis, this paper proposed a novel diagnostic method based on FTIR and artificial neural network (ANN). Firstly, the spectra of high- and low-grade human glioma that without dye were collected by FTIR spectrometer, then the raw data preprocessed with baseline correction and amide I (1649 cm^{-1}) normalization before input into the input-layer of the ANN, after the nonlinear conversion of the neurons in the hidden-layers, the categories were presented in the output-layer. Corresponding to the decrease of the loss function, the weights of the net updated continuously, and finally, the optimized model has the power of prediction for new samples.

Results

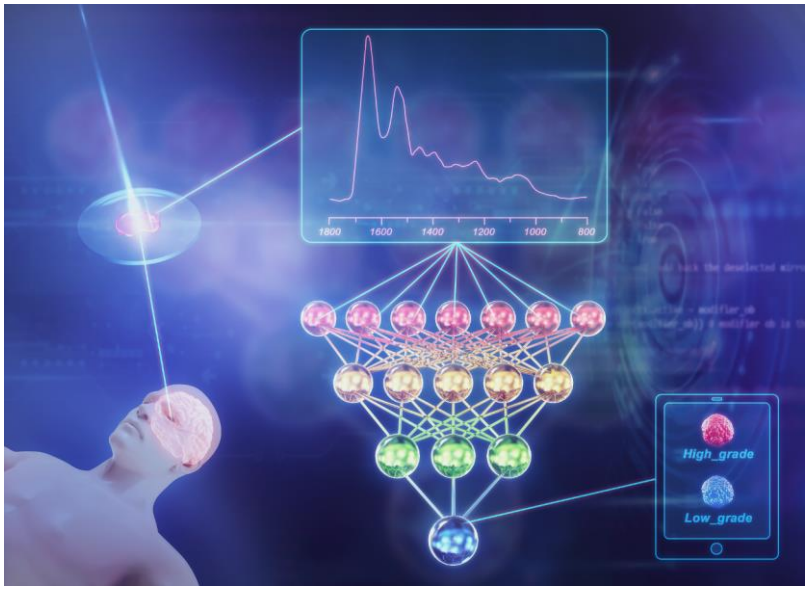
After training on 6225 spectra sourced from 77 glioma patients, the ANN model reached the prediction accuracy, specificity and sensitivity evaluation metrics above 99%, which was much superior to the common classification method of principal component analysis-linear discriminate analysis (PCA-LDA) (the prediction accuracy, specificity and sensitivity are only 87%, 89% and 86%, respectively). Moreover, rather than the lipid range of 2800-3000 cm^{-1}, the ANN learned the fingerprint characteristics of the infrared spectrum to classify the major histopathologic classes of human glioma. Especially, the diagnosis process of the novel method only requires several minutes. Compared to the traditional pathological diagnosis, the efficiency raises almost 500 times.

Conclusions

The infrared range of fingerprint is the major indicator for cancer progression, and the ANN-based diagnosis method can be streamlined, and create a complementary pathway that is independent of the traditional pathology laboratory.

Keywords: glioma, grade, Fourier transform infrared (FTIR), artificial neural network (ANN), fingerprint, principal component analysis-linear discriminate analysis (PCA-LDA)

TOC Graphic



Background

Glioma accounts for almost 30% of all primary brain tumors and 80% of all malignant tumors and is responsible for the majority of deaths from primary brain tumors; furthermore, gliomas can be classified as astrocytomas, oligodendrogliomas, mixed oligoastrocytic gliomas or ependymomas and assigned WHO grades from I to IV on the basis of their histological appearance [1]. The WHO lower grades (I/II) indicate the least malignant behavior. Low-grade cells look less like normal cells and usually grow slowly, but they can grow into nearby brain tissue. Surgery is usually the only treatment for a low-grade tumor, but the tumor is more likely to return after surgery and tends to develop into a malignant tumor. For high-grade (III/IV) glioma, the cells look very abnormal and grow very fast. High-grade cells often return soon after treatment and sometimes spread to other parts of the brain and spinal cord, with treatment involving radiotherapy and chemotherapy being necessary [2]. Historically, histopathology has been the gold standard method for classification diagnosis, with sections deposited on microscopy slides examined based on some key criteria, such as cellular density, nuclear atypia, mitotic activity, necrosis and microvascular proliferation [3]. However, the traditional method also has limitations, such as a tedious experimental process, delayed diagnostic results and the subjectivity of pathologists to some degree. Therefore, a complementary approach based on a rapid, accurate, objective and quantitative analysis tool could provide classification advice for aggressive diseases for clinicians.

Spectroscopy including FTIR, Raman and Terahertz are valuable instruments that have been used for studying molecular component changes in lipids, proteins and nucleic acids in biological samples, such as biological fluids, tissues, and cancer cell lines [4-12]. The commonly method of FTIR mainly relies on the resonant frequencies of the molecular bonds, which result

in some absorption peaks when the transmission electromagnetic waves are collected by a detector of the interferometer [13]. In contrast to those procedures involving dyes and other histological approaches, the FTIR method is rapid and nondestructive, and does not require reagents [14].

Multifactorial statistical analysis methods related to FTIR have been widely used for identifying changes in lipids, proteins, nucleic acids, and carbohydrates, such as principal component analysis (PCA) [15-17] and partial least squares (PLS) [18,19] combined with discriminant analysis (DA), hierarchical cluster analysis (HCA) [20,21], support vector machines (SVMs) [22,23] and random forest (RF) [24]. Smith et al. [25] used the supervised machine learning algorithm of RF as a classifier to separate patients into cancer and noncancer categories based upon the intensities of wavenumbers presented in their spectra and finally achieved a sensitivity and specificity up to 92.8% and 91.5%, respectively. Cameron et al. [26] assessed patients with various brain tumors by using their serum and applied the PLS-DA model to their spectral signatures collected by attenuated-total-reflection FTIR spectroscopy, achieving a sensitivity and specificity greater than 92% in the classification of brain tumors and control patients. Moreover, metastasis *vs.* glioblastoma with the linear SVM reported a 84.3% sensitivity, 96.2% specificity and receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.9, suggesting a high diagnostic capability [26]. As a pattern-recognition-based approach, the artificial neural network (ANN) has been proved to be effective in the analysis of biological specimens [18]. The ANN method consists of many neurons arranged in separate layers and has the capability to transfer the input signal to the output layer of classification through an activation function. A. D. Surowka combined the ANN and synchrotron-radiation-based infrared spectroscopy to study the protein composition of human

glial tumors. After the network was optimized and tested, the standard error of prediction (SEP) was found to be lower than 5% [27]. By using FTIR spectroscopy and the ANN, Argov et al. [28] reported that the method could separate an adenomatous polyp from a malignant cell, with classification percentages of 89%, 81% and 83% for normal, adenomatous polyp, and malignant cells, respectively.

To truly reflect the molecular change during grading, this paper adopts FTIR spectroscopy for tissues instead of serums [29-31]. Additionally, the samples are collected from different patients diagnosed with either low or malignant glioma, and the spectroscopic results are statistically significant. After a comparison of two different supervised machine learning algorithms, i.e., PCA-LDA and ANN, the results demonstrate that the FTIR-ANN method performs better than PCA-LDA. Thus, FTIR-ANN can be a promising clinical diagnostic alternative to histopathology.

Methods

A total of 9360 spectra were collected from 77 patients with different grades of glioma. The subtypes of low-grade glioma (WHO II) are oligodendroglioma (n = 6) and diffuse astrocytoma (n = 15), with 14 males and 7 females, aged 10 to 63 years old, with an average age of 38.3 years old. High-grade glioma (WHO III/IV) includes anaplastic astrocytoma (n = 15), anaplastic oligodendroglioma (n = 10) and glioblastoma (n = 31), covering 36 males and 20 females with ages ranging from 14 to 69 years old and an average age of 48.8 years old. The detailed information is illustrated in Table 1. The experiment was approved by the Institutional Review Board and Research Ethics Committee (S2018-215-01).

WHO grade	Subtypes	Number	Age range/Average	Gender	Number of spectra
II	Oligodendroglioma	6	10-63/38.8	14 males 7 females	4610
II	Diffuse astrocytoma	15			
III	Anaplastic astrocytoma	15	14-69/48.8	36 males 20 females	4750
III	Anaplastic oligodendroglioma	10			
IV	Glioblastoma	31			

Table 1. Detailed sample information according to the WHO classification of high- and low-grade glioma

High- and low-grade glioma formalin-fixed paraffin-embedded (FFPE) tissues from different patients were collected by the General Hospital of the People's Liberation Army (PLAGH). Each specimen was cut into 8 μm thick pieces and carefully spread flat on a barium fluoride (BaF_2) substrate (Fig. 1b). In the subsequent dewaxing step, the tissue slides were immersed in xylene at room temperature for 5 min; this step was repeated twice with fresh xylene. Then, the tissue slides were washed and cleared by immersing them in 100% ethanol for 5 min, which was repeated twice with fresh ethanol. In the last step, these tissue slides were allowed to air dry before the IR spectra were collected [32].

The spectra were detected by FTIR microscopy (Nicolet 6700) at the BL01B beamline of the Shanghai Synchrotron Radiation Facility (SSRF). The absorbance spectra were obtained in transmission mode (Fig. 1a) in the wavenumber range of 800-4000 cm^{-1} at a resolution of 4 cm^{-1} with 16 coadded scans. The aperture size was set to 80 x 80 μm with a step size of 80 μm. The

background spectrum was obtained on the blank area of a 1 mm thick barium fluoride substrate. Each patient contributed approximately 140 spectra to the dataset, and the data were collected and processed by the OMNIC 9.2 software. Data preprocessing included automatic baseline correction and amide I (1649 cm^{-1}) normalization (Fig. 1c and d).

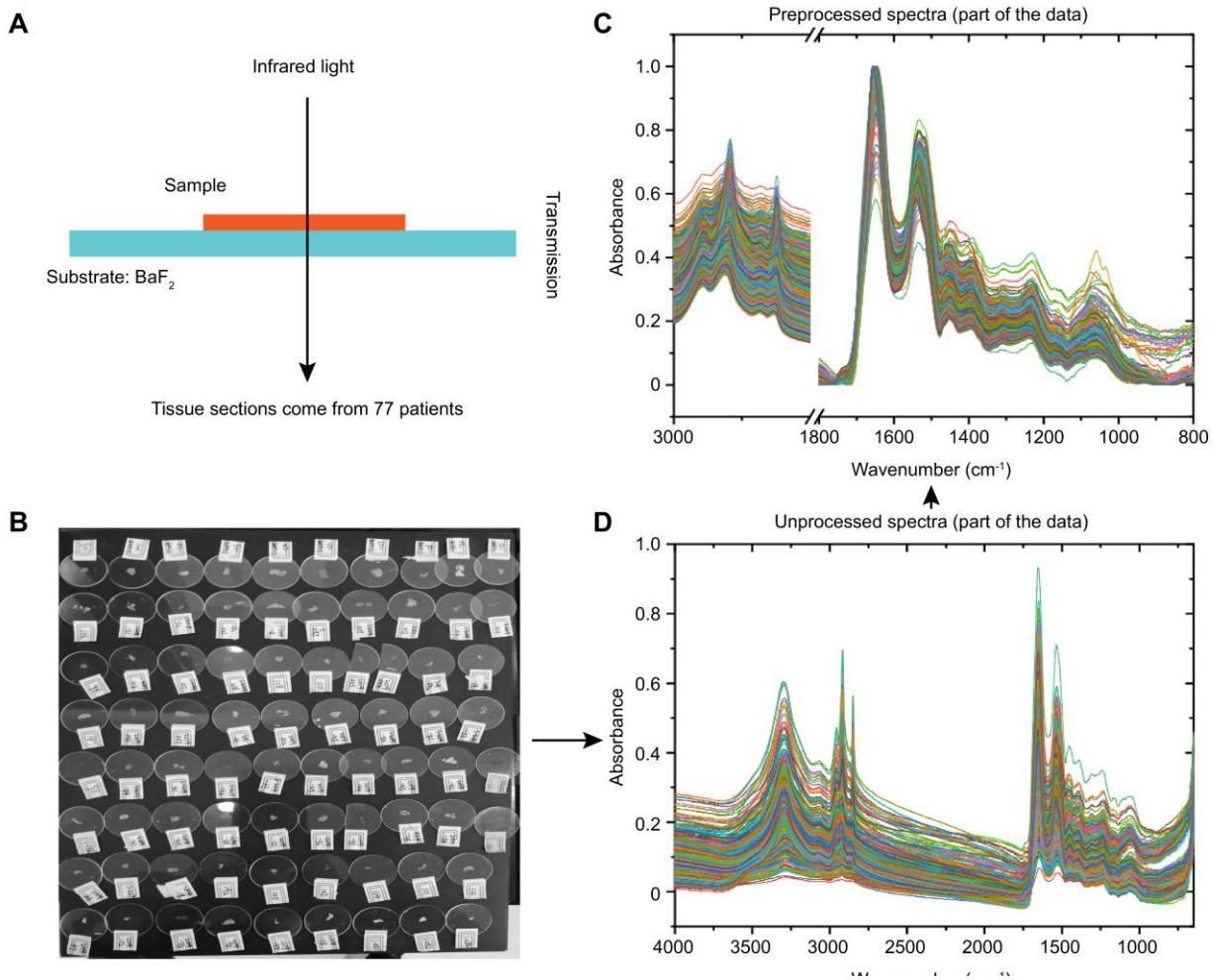


Fig. 1. The glioma tissue was analyzed by FTIR spectroscopy, followed by preprocessing of the resulting data. (a) Spectral collection through transmission mode; (b) 8 µm thick tissue sections from 77 patients diagnosed with high- and low-grade glioma; (c) unprocessed spectra; and (d) spectra after preprocessing, including automatic baseline correction and amide I (1649 cm^{-1}) normalization.

Data Analysis

Both the lipid range of 2800-3000 cm^{-1} and the fingerprint region of 800-1800 cm^{-1} were extracted from the whole spectra, which were preprocessed with baseline correction and amide I normalization. Including high- and low-grade glioma, all of the spectra were randomly divided into a training set (70%) and a test set (30%). The spectral data analysis was performed by using the classification toolbox version 5.4 of Milano Chemometrics and the QSAR Research Group in the MATLAB R2020a environment (MathWorks, Natick, USA). In this research, the PCA-LDA and ANN methods were used.

The PCA-LDA algorithm is a supervised machine learning method that is commonly employed to process IR spectra data. The goal of PCA is to reduce the dimensionality of data and retain as much as possible the variation present in a dataset. Assume the following original space representation:

$$f(m) = a_1 m_1 + a_2 m_2 + ... + a_N m_N \begin{bmatrix} a_1 \\ a_2 \\ ... \\ a_N \end{bmatrix}$$

where m_1, m_2, m_3, ..., m_N is the base in the original n-dimensional space.

The lower-dimensional subspace representation is:

$$f(n) = b_1 n_1 + b_2 n_2 + ... + b_k n_K \begin{bmatrix} b_1 \\ b_2 \\ ... \\ b_K \end{bmatrix}$$

where n_1, n_2, $_{n3}$, ..., n_K is the base in a k-dimensional subspace (K<N).

The information loss is shown below:

$$f(m) - f(n) = b_1 u_1 + b_2 u_2 + ... + b_K u_K = \sum_{i=1}^{K} b_i u_i$$

where k << N.

Then, K is chosen according to the following criterion:

$$\frac{\sum_{i=1}^{K}\lambda_i}{\sum_{i=1}^{N}\lambda_i} > 0.95$$

The goal of LDA is to find directions along which the classes are best separated, taking into consideration the within-classes and between-classes regimes.

$$\max \frac{\left|U^T S_b U\right|}{\left|U^T S_w U\right|}$$

where U is the projection matrix, S_w is the within-class scatter matrix, and S_b is the between-class scatter matrix.

Thus, the following can be concluded:

$$\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_N \end{bmatrix} \rightarrow \text{PCA} \rightarrow \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_K \end{bmatrix} \rightarrow \text{LDA} \rightarrow \begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_{C-1} \end{bmatrix}$$

where C denotes the classes.

In this model, dimension reduction was used for the spectral ranges of 800-1800 cm^{-1} (2076 dimensions) and 2800-3000 cm^{-1} (417 dimensions) and a combination of the two ranges (2493 dimensions). Subsequently, the first 16 PCs for LDA were chosen according to the minimum error rate of fivefold Venetian blind cross-validation. Then, the first two principal components (PCs) were displayed in a scattering graph, which can be clearly visualized.

As a sophisticated computational model based on the nonlinear processing of neurons (nodes), the ANN has been proved to be effective in the analysis of biological specimens [18]. A general ANN consists of two main operation phases: forward propagation for producing the

output results and backward propagation for minimizing the cost value. During the process of error back propagation, the weights are adjusted constantly until the cost value reaches the minimum value with an appropriate learning rate (Fig. 2). Finally, the optimized network has the power of prediction.

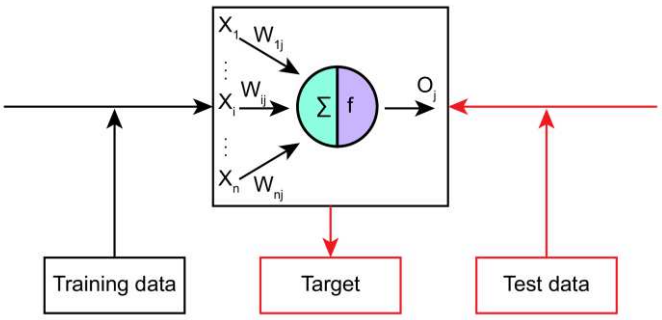


Fig. 2. Schematic of the ANN method.

Assuming that the number of layers in the ANN net is K (K>1), the dimensions of the input and output layers are m_o and m_k, respectively. The output of each layer of the network is expressed as follows:

(1) Input layer

$$Y^{(0)} = \left[Y_1^{(0)}, Y_2^{(0)}, ..., Y_{m_0}^{(0)}\right]^T$$

(2) Output layer

$$Y^{(K)} = f^{(k)}\left[Y_1^{(K)}, Y_2^{(K)}, ..., Y_{m_K}^{(K)}\right]^T$$

(3) Cost function:

$$L = \frac{1}{2}\sum_{i=1}^{m_K}\left(Y_i^{(K)} - T_i\right)^2$$

(4) Weight update:

$$W \leftarrow W - \eta\frac{\partial L}{\partial W}$$

where $f^{(k)}$ is an activation function, η is the learning rate and $\frac{\partial L}{\partial W}$ is the gradient.

In this model, a four-layer perceptron was used for the classification of high- and low-grade glioma. The numbers of nodes in the input layer were the same with wavenumber ranges of 2800-3000 cm^{-1} (417 nodes) and 800-1800 cm^{-1} (2076 nodes) and a combination of the two ranges (2493 nodes). The hidden 1-layer consisted of 50 neurons, with each node receiving all of the nodes from the input layer. The hidden 2-layer included 5 neurons that were also fully connected to the hidden 1-layer and the output layer of one neuron. The activation function was a sigmoidal function, and the learning rate was 0.001. The momentum term alpha was set as 0.5 to cancel the opposing components and enhance the reinforcing components at successive positions. Venetian blind cross-validation (CV) was adopted, and the number of CV groups was five. The model calibration was terminated after 5000 epochs on the training set.

To determine the performance of the models, accuracy, specificity and sensitivity were used as the evaluation metrics. Accuracy represents the ratio of correctly assigned samples. Specificity and sensitivity are the rates of correct identification of negative items and positive items, respectively. The three metrics are expressed below:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Specificity = \frac{TN}{TN + FN}$$

$$Sensitivity = \frac{TP}{TN + FN}$$

where true positive (TP) represents correct classification of the positive class, true negative (TN) describes correct classification of the negative class, false positive (FP) represents incorrect prediction of the positives and false negative (FN) corresponds to incorrect prediction of the

negatives.

Results

The high- and low-grade tissues were predefined based on the histopathologic results before collecting the spectra. Fig. 3 shows the H&E staining of a case of glioblastoma tissue diagnosed with WHO grade IV (a-c) and a case of oligodendroglioma diagnosed with WHO grade II (d-f). Specifically, the left panel presents the H&E staining of a tissue slide under a 10x microscope (a: high, d: low), the middle panel is the tissue morphology under a 32x microscope without dye (b: high degree, e: low degree), and the right panel is the corresponding IR mapping at 1539 cm^{-1} (c: high degree, f: low degree). In addition, Fig. 3(a) presents numerous necrotic foci and blood vessel proliferation, while Fig. 3(b) shows that the cells are characterized richly in some areas, the surrounding nucleus is hollow and the cytoplasm is transparent, with no clear mitosis.

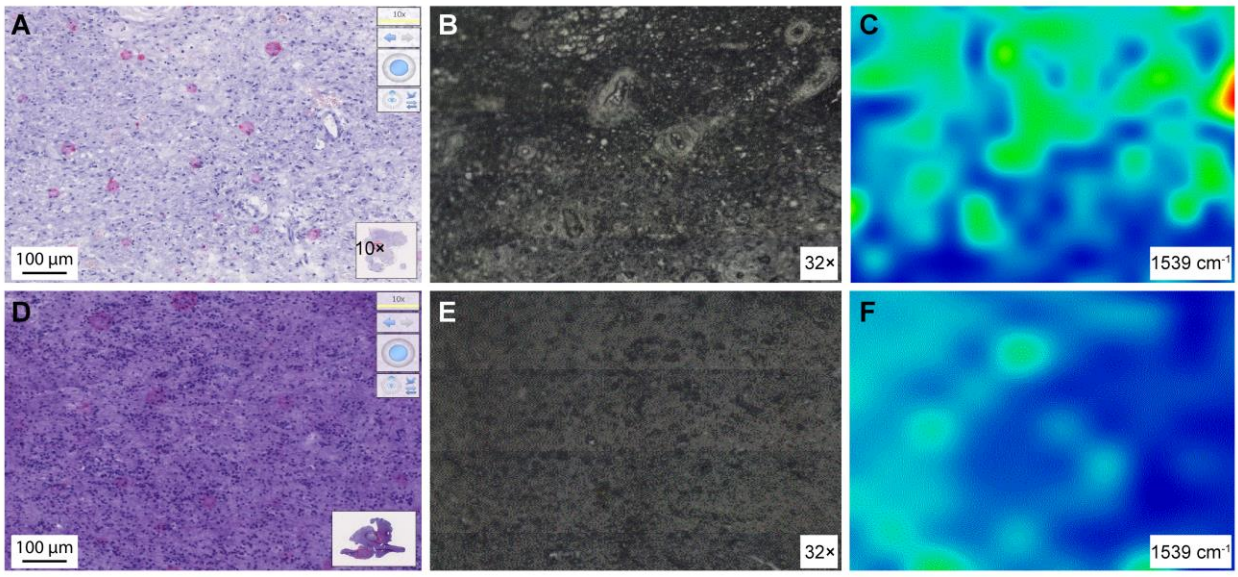


Fig. 3. H&E staining of a tissue slide under a 10x microscope (a: high, d: low), the tissue morphology under a 32x microscope (b: high, e: low), and the corresponding IR mappings at 1539 cm^{-1} (c: high, f: low).

A total of 4610 and 4750 spectra were collected from 21 low-grade and 56 high-grade glioma patients, respectively. The spectra of low- and high-grade gliomas in the ranges of 2800-3000 cm^{-1} and 800-1800 cm^{-1} are shown in Fig. 4(a) and (b). The bands at 2800-3000 cm^{-1} were attributed to lipid absorbance, and the major bands at 2957 cm^{-1}, 2917 cm^{-1} and 2849 cm^{-1} were identified. In detail, 2957 cm^{-1}, 2917 cm^{-1} and 2849 cm^{-1} correspond to the CH_3 asymmetric stretch and the CH_2 asymmetric and symmetric stretching vibrations from lipids, respectively. The range of 800-1800 cm^{-1} represents the fingerprint range, and the bands at 1741 cm^{-1} and 1453 cm^{-1} are assigned to the carbonyl C=O stretch and CH_2 bending stretch from lipids, respectively. There was a phenomenon in the spectra in which the peak at 1741 cm^{-1} was observed only in low-grade tissue and missed in malignant tissue. Thus, the 1741 cm^{-1} peak may be a potential marker of disease progression. Serving as an internal reference normalized in the preprocessed data, the band at 1649 cm^{-1} is the stretching vibration of the C=O groups of the peptide chains from amide I. Amide II at 1539 cm^{-1} belongs to N-H bending and C-N stretching, and 1390 cm^{-1} is attributed to the C=O stretching of COO^- symmetric stretching. Moreover, 1234 cm^{-1} and 1061 cm^{-1} are assigned to the asymmetric and symmetric PO_3^{2-} groups from DNA, RNA and phospholipids, respectively. The IR bands and the corresponding assignments [33] are summarized in Table 2. Statistical analysis was applied to the relative intensity of the bands to obtain semiquantitative information of the graded tissues. Fig. 4(c) shows that there is a considerable difference between the IR spectra of the low and malignant tissues. The intensity in the lipid band of 2917 cm^{-1} for the high grade decreased significantly in contrast to the low grade with a ratio of approximately 0.91 times. According to Student's t-test analysis, the P value was less than 0.01. Fortunately, due to the greater metabolism of cancer cells in disease progression, the protein and nucleic acid levels of the malignant tissues increased significantly, corresponding

to the 1741 cm^{-1} (I_{high}/I_{low}=0.5, P<0.01), 1539 cm^{-1} (I_{high}/I_{low}=1.06, P<0.01), 1453 cm^{-1} (I_{high}/I_{low}=1.15, P<0.01), 1390 cm^{-1} (I_{high}/I_{low}=1.22, P<0.01), 1234 cm^{-1} (I_{high}/I_{low}=1.17, P<0.01) and 1074 cm^{-1} (I_{high}/I_{low}=1.15, P<0.05) bands.

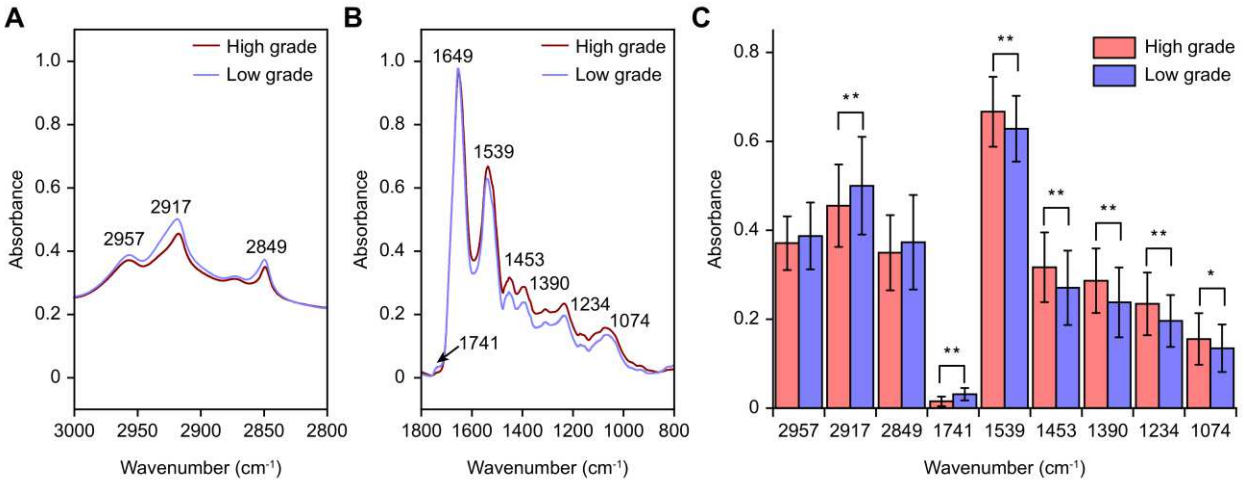


Fig. 4. Average spectra of high and low grades of glioma in the ranges of 2800-3000 cm^{-1} (a) and 800-1800 cm^{-1} (b). (c) The relative intensity of the high-grade *vs.* low-grade bands at 2957 cm^{-1}, 2917 cm^{-1}, 2849 cm^{-1}, 1741 cm^{-1}, 1539 cm^{-1}, 1453 cm^{-1}, 1390 cm^{-1}, 1234 cm^{-1} and 1074 cm^{-1}. ** Mean value is particularly significant (P<0.01). * Mean value is significant (P<0.05). Red and green represent high- and low-grade data, respectively.

Approximate Wavenumbers	Assignments
2957	CH_3 asymmetric stretching from lipids
2917	CH_2 asymmetric stretching vibrations from lipids
2849	CH_2 symmetric stretching vibrations from lipids
1741	C=O vibrations from lipids (absent in high-grade gliomas)
1649	C=O stretching vibrations of the peptide chains from amide I
1539	Mainly N-H bending and C-N stretching from protein amide II
1453	Protein CH_2 and CH_3 bending of methyl, and CH_2 bending from
1390	C=O stretch of COO^- symmetric stretch
1234	Asymmetric PO_3^{-2} group from DNA, RNA and phospholipids

Table 2. Absorbance bands observed in the spectra and the corresponding assignments

PCA-LDA

The PCA-LDA method was applied to three wavenumber ranges, which are denoted as ranges 1, 2, and 3. Ranges 1 and 2 correspond to the fingerprint range of 800-1800 cm^{-1} and lipid range of 2800-3000 cm^{-1}, while range 3 combines both 1 and 2. In this model, the first 16 PCs attributed to 100% variance were used for the linear discriminate analysis. Fig. 5(a), (c), and (e) illustrate the score plots between PC1 and PC2 for high and low grades in the ranges of 1, 2, and 3, respectively, and (b), (d), and (f) are the corresponding loadings of PC1. The purple symbols represent malignant glioma, and the pink samples are low-grade ones. The dot symbols correspond to the training set, and the star symbols are samples in the test set. The dotted line is the classification boundary. As shown in Fig. 5(a), (c) and (e), PC1 accounts for 97.6%, 99.53% and 97.37% in the three different spectral ranges, respectively, originating mainly from amide I (1649 cm^{-1}) and amide II (1539 cm^{-1}) (b, f) and CH_2 asymmetric from lipids (2917 cm^{-1}) (d). The results demonstrate that the PCA-LDA model cannot separate the grades clearly by PC1 and PC2 on the training set (dot symbols) with fivefold Venetian blind cross-validation, as well as on the test set (star symbols).

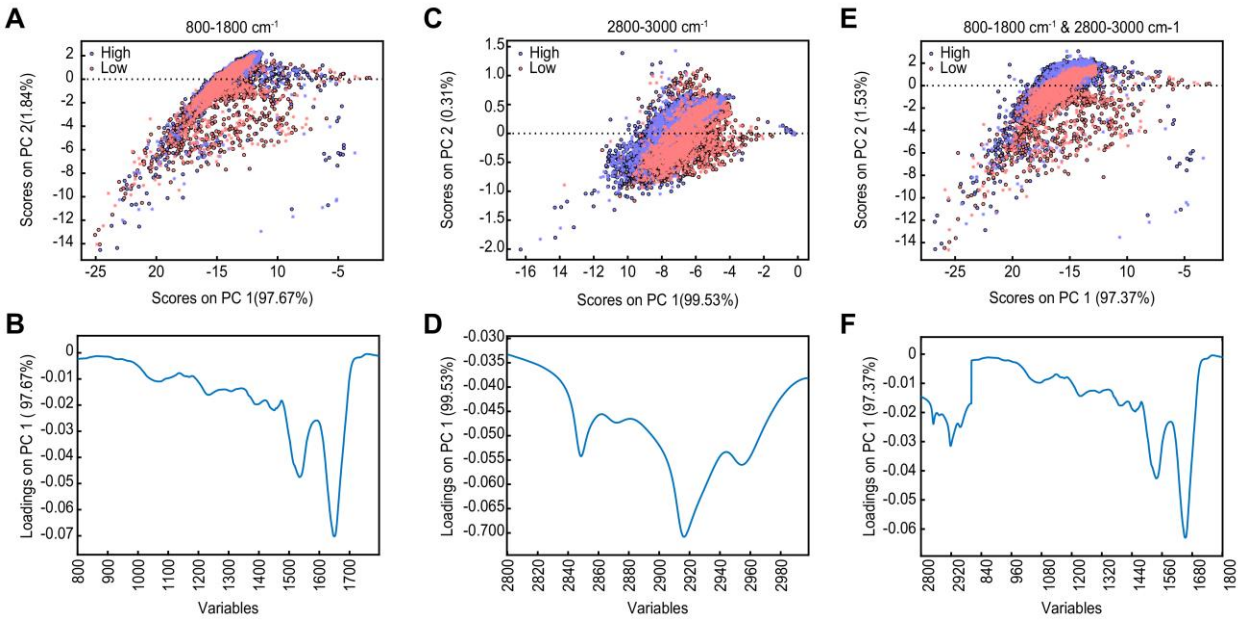


Fig. 5. PCA-LDA results. (a), (c), and (e) are the score plots between PC1 and PC2 for high-grade *vs.* low-grade glioma in wavenumber ranges 1, 2, and 3. (b), (d), and (f) are the loadings of PC1 corresponding to (a), (c), and (e), respectively. Ranges 1, 2, and 3 represent the wavenumber ranges of 800-1800 cm^{-1}, 2800-3000 cm^{-1}, and both 800-1800 cm^{-1} and 2800-3000 cm^{-1}, respectively.

ANN

A total of 6552 spectra, including high- and low-grade spectra, were used as the training set to train the ANN net with Venetian blind cross-validation, while the remaining 2808 spectra were used as the test set to evaluate the model. The classification outputs based on ranges 1, 2, and 3 are shown in Fig. 6(a)-(c), respectively. Similar to Figs.5 uses dots and stars to indicate the training and test sets, respectively. As shown in Fig. 6(a) and (c), the grades could be separated clearly in ranges 1 and 3, and the related accuracy, sensitivity and specificity were all greater than 0.98 (f, j). However, for range 2, the ANN network exhibited lower performance with an

accuracy of 0.90, sensitivity of 0.91, and specificity of 0.91 (g). As a graph of the true positive rate *vs.* false positive rate, the ROC curve also represents the performance of the classification model at all classification thresholds. The AUC represents the area under the ROC curve integrated from (0, 0) to (1, 1). It is an attractive indicator with scale invariance and classification-threshold invariance. The AUC ranges from 0 to 1, with higher values indicating better performance of the model. As shown in Fig. 4(d) and (h), the AUC values related to ranges 1 and 3 can reach 1, while the value is 0.98 for range 2 (Fig. 4(f)).

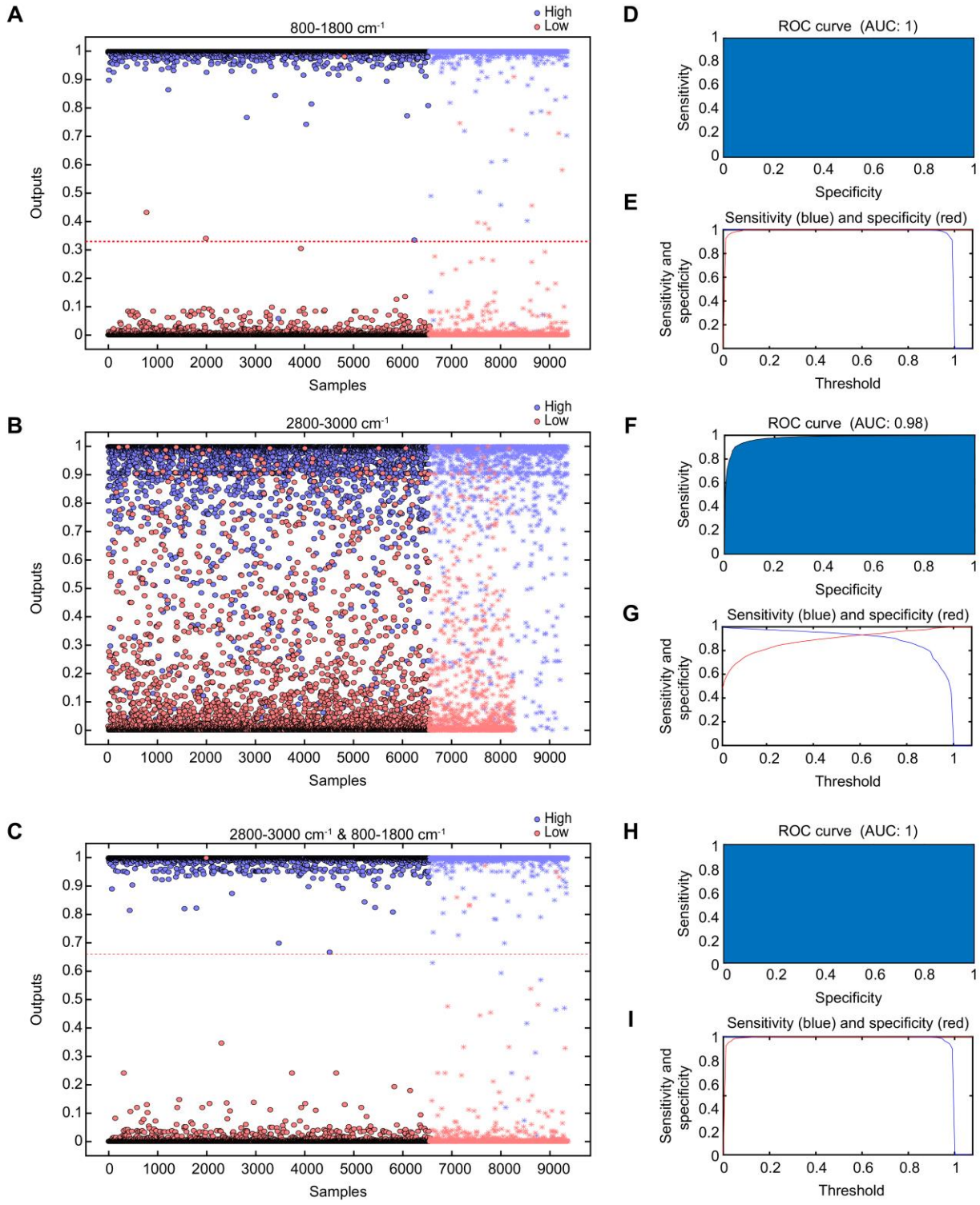


Fig. 6. Corresponding to ranges 1, 2, and 3, respectively, (a), (b), and (c) are the classification results of the ANN output, (d), (f), and (h) are the ROC results, and (e), (g), and (i) are the related

sensitivity (blue) and specificity (red) plots. Purple dots represent the high grade and pink dots represent the low grade in the training set. Purple stars indicate the high grade and pink stars indicate the low grade in the test set. Ranges 1, 2, and 3 denote the wavenumber ranges of 800-1800 cm^{-1}, 2800-3000 cm^{-1}, and both 800-1800 cm^{-1} and 2800-3000 cm^{-1}, respectively.

		ANN			PCA-LDA		
Dataset	Index	800-1800	2800-3000	Combination	800-1800	2800-300	Combination
CV	Sensitivity	**0.99**	0.92	0.99	0.87	0.83	0.90
	Specificity	**0.99**	0.92	0.98	0.89	0.85	0.89
	Accuracy	**0.99**	0.92	0.99	0.88	0.84	0.89
TS	Sensitivity	**0.99**	0.90	0.98	0.86	0.83	0.89
	Specificity	**1**	0.91	1	0.89	0.86	0.89
	Accuracy	**0.99**	0.91	0.99	0.87	0.85	0.89

Table 3. Performance of PCA-LDA *vs.* that of the ANN in different spectral ranges

Discussion

FTIR spectroscopy has emerged in recent years as an analysis method of choice that has several clear advantages over histology, such as no specific reagents, rapid measurement, operational simplicity and repetitive analysis. In this research, we collected 9360 FTIR spectra from a total of 77 patients with 56 high-grade (WHO III/IV) and 21 low-grade (WHO II) glioma cases. The lack of grade I data was due to a few clinical samples that could not be analyzed statistically. As shown in the average spectra, high-grade glioma presents an overall decrease in lipid levels and content growth of proteins and nucleic acids. Specifically, the related band intensities of 1539/1649, 1453/1649, 1390/1649, 1234/1649, and 1074/1649 are enhanced, and the ratio of 2917/1649 is diminished. The growth of high-grade N-H bending and C-N stretching from protein amide II may contribute to the breakage of intramolecular C=O···H-N hydrogen bonds, leading to rotation of the CN bond and rearrangement of the entire system of hydrogen bonds in malignant tumor proteins [32]. According to Kar et al. [34] and Sitnikova et al. [35], breast

cancer presents a significant increase in the intensity ratio of amide I, amide II, and nucleic acid [28], and the difference may be attributed to a greater metabolic activity of cancer cells in disease progression [36]. The ratios of the intensities of the 2849 (CH_2) and 2917 (CH_3) bands are diminished, indicating a large number of methylene groups in malignant tissue. Moreover, the band at 1741 cm^{-1} could be observed in the low-grade tissue but not in the spectra of malignant tissue. Thus, the 1741 cm^{-1} band could serve as a marker that distinguishes the grades. In research on the chemical changes in healthy brain tissues and glioblastoma tumor tissues, Depciuch et al. [7] found that compared to control brain cancer, FTIR spectra of cancer brain tissue showed a significant difference in chemical composition; hence, they assumed that lipids could be a spectroscopic marker for brain tumors. Combined with the multifactorial statistical analysis of PCA-LDA and the ANN in the ranges of 800-1800 cm^{-1} and 2800-3000 cm^{-1} and a combination of the two ranges, the results (Table 3) demonstrate that the ANN algorithm operating within 800-1800 cm^{-1} achieves the best performance, with an accuracy, a specificity and a sensitivity all reaching 99% on the training set and a prediction accuracy, specificity and sensitivity are above 99% on the test set, which is much superior to the PCA-LDA, which the prediction accuracy, specificity and sensitivity are only 87%, 89% and 86%, respectively. Therefore, it can be concluded that the infrared range of 800-1800 cm^{-1} is the major indicator for cancer progression, and the ANN-based method could be established as a promising diagnostic tool in clinic. Although the use of IR spectra can be seen as a promising method for the detection of cancer progression, some weaknesses still exist, such as the low signal on aqueous samples with strong absorbance. Furthermore, contamination in a sample will affect the spectral data and lead to incorrect interpretation [32], and the data preprocessing procedure should also be standardized to avoid different interpretations of the results. When the process is standardized

and unified, the method can be used as an alternative approach for the clinical grade diagnosis of human glioma.

Conclusions

In this study, we report an alternative workflow that combines the Fourier transform infrared (FTIR) spectroscopy and artificial neural network (ANN) to predict diagnosis the grade of human glioma in a fast (within several minutes, the efficiency raises almost 500 times), accurate (overall accuracy, specificity and sensitivity evaluation metrics can reach above 99%), and without reagent way, this method is much superior to the common classification method of principal component analysis-linear discriminate analysis (PCA-LDA) (the prediction accuracy, specificity and sensitivity are only 87%, 89% and 86%, respectively). The ANN mainly learned the fingerprint characteristics in the infrared spectrum to classify the major histopathologic classes of human glioma. These results demonstrate that grade diagnosis of human glioma can be streamlined, and create a complementary pathway that are independent of the traditional pathology laboratory.

Abbreviations: FTIR - Fourier Transform Infrared; ANN - Artificial Neural Network; PCA - Principal Component Analysis; LDA - Linear Discriminate Analysis; CV - Cross Validation; TS - Test Set; BaF_2 - Barium Fluoride; FFPE – Formalin-Fixed Paraffin Embedded; H&E - Hematoxylin and Eosin; ROC - Receiver Operating Characteristic; AUC - Area under the ROC Curve

Declarations

Ethics approval and consent to participate

The experiment was approved by the Institutional Review Board and Research Ethics Committee (S2018-215-01).

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare no conflict of interest for this article.

Funding

There was no funding support.

Authors' contribution

Wenyu Peng and Shuo Chen performed main experiments and primary data analysis and wrote the manuscript. Dongsheng Kong provided the clinical tumor samples of glioma. Xiaojie Zhou provided technical supports in the data collection. Xiaoyun Lu and Chao Chang conceived ideas and supervised the research progress. All the authors have read and approved the final manuscript.

Acknowledgements

We thank the staff from BLO1B beamline of National Center for Protein Science Shanghai (NSPSS) at Shanghai Synchrotron Radiation Facility, for assistance during data collection, and

we thank the Department of Neurosurgery of Chinese People’s Liberation Army (PLA) General Hospital, for the support of clinical tumor samples.

References

[1] M. Weller, W. Wick, K. Aldape, M. Brada, M. Berger, S.M. Pfister, R. Nishikawa, M. Rosenthal, P.Y. Wen, R. Stupp, G. Reifenberger, Glioma, Nat. Rev. Dis. Primers 1 (2015) 15017. https://doi.org/10.1038/nrdp.2015.40.

[2] M.J. van den Bent, M. Weller, P.Y. Wen, J.M. Kros, K. Aldape, S. Chang, A clinical perspective on the 2016 WHO brain tumor classification and routine molecular diagnostics, Neuro-Oncology 19 (2017) 614–624. https://doi.org/10.1093/neuonc/now277.

[3] A. Perry, P. Wesseling, Histologic classification of gliomas, Handb. Clin. Neurol. 134 (2016) 71–95. https://doi.org/10.1016/B978-0-12-802997-8.00005-0.

[4] A. Beljebbar, S. Dukic, N. Amharref, M. Manfait, Screening of biochemical/histological changes associated to C6 glioma tumor development by FTIR/PCA imaging, Analyst 135 (2010) 1090–1097. https://doi.org/10.1039/b922184k.

[5] A. Sala, K.E. Spalding, K.M. Ashton, R. Board, H.J. Butler, T.P. Dawson, D.A. Harris, C.S. Hughes, C.A. Jenkins, M.D. Jenkinson, D.S. Palmer, B.R. Smith, C.A. Thornton, M.J. Baker, Rapid analysis of disease state in liquid human serum combining infrared spectroscopy and “digital drying”, J. Biophotonics 13 (2020) e202000118. https://doi.org/10.1002/jbio.202000118.

[6] H.J. Butler, B.R. Smith, R. Fritzsch, P. Radhakrishnan, D.S. Palmer, M.J. Baker, Optimised spectral pre-processing for discrimination of biofluids via ATR-FTIR

spectroscopy, Analyst 143 (2018) 6121–6134. https://doi.org/10.1039/c8an01384e.

[7] J. Depciuch, B. Tołpa, P. Witek, K. Szmuc, E. Kaznowska, M. Osuchowski, P. Król, J. Cebulski, Raman and FTIR spectroscopy in determining the chemical changes in healthy brain tissues and glioblastoma tumor tissues, Spectrochim. Acta A Mol. Biomol. Spectrosc. 225 (2020) 117526. https://doi.org/10.1016/j.saa.2019.117526.

[8] M.M. Grzelak, P.M. Wróbel, M. Lankosz, Z. Stęgowski, Ł. Chmura, D. Adamek, B. Hesse, H. Castillo-Michel, Diagnosis of ovarian tumour tissues by SR-FTIR spectroscopy: a pilot study, Spectrochim. Acta A Mol. Biomol. Spectrosc. 203 (2018) 48–55. https://doi.org/10.1016/j.saa.2018.05.070.

[9] W. Shi, Y. Wang, L. Hou, C, Ma, Detection of living cervical cancer cells by transient terahertz spectroscopy, J. Biophotonics 14 (2021) e202000237. https://doi.org/10.1002/jbio.202000237.

[10] L. Hou, W. Shi, C. Dong, et al. , Spectrochim. Acta A Mol. Biomol. Spectrosc. (2020), https://doi.org/10.1016/j.saa.2020.119044.

[11] Y. Peng, C. Shi, Y. Zhu, et al. , Terahertz spectroscopy in biomedical field: a review on signal-to-noise ratio improvement. PhotoniX 1, 12 (2020). https://doi.org/10.1186/s43074-020-00011-z.

[12] Y. Peng, C. Shi, X. Wu, et al. , Terahertz Imaging and Spectroscopy in Cancer Diagnostics: A Technical Review, BME Frontiers, vol. 2020, Article ID 2547609, 11 pages, 2020. https://doi.org/10.34133/2020/2547609

[13] A.A. Bunaciu, H.Y. Aboul-Enein, Ş. Fleschin, Vibrational spectroscopy in clinical

analysis, Appl. Spectrosc. Rev. 50 (2015) 176–191. https://doi.org/10.1080/05704928.2014.955582.

[14] A.A. Bunaciu, V.D. Hoang, H.Y. Aboul-Enein, Vibrational micro-spectroscopy of human tissues analysis: review, Crit. Rev. Anal. Chem. 47 (2017) 194–203. https://doi.org/10.1080/10408347.2016.1253454.

[15] R.S. Uysal, I.H. Boyaci, Authentication of liquid egg composition using ATR‐FTIR and NIR spectroscopy in combination with PCA, J. Sci. Food Agric. 100 (2020) 855‐862. https://doi.org/10.1002/jsfa.10097.

[16] E. Kaznowska, J. Depciuch, K. Łach, M. Kołodziej, A. Koziorowska, J. Vongsvivut, I. Zawlik, M. Cholewa, J. Cebulski, The classification of lung cancers and their degree of malignancy by FTIR, PCA-LDA analysis, and a physics-based computational model, Talanta 186 (2018) 337–345. https://doi.org/10.1016/j.talanta.2018.04.083.

[17] E. Kaznowska, J. Depciuch, K. Szmuc, J. Cebulski, Use of FTIR spectroscopy and PCA-LDC analysis to identify cancerous lesions within the human colon, J. Pharm. Biomed. Anal. 134 (2017) 259–268. https://doi.org/10.1016/j.jpba.2016.11.047.

[18] P. Barmpalexis, A. Karagianni, I. Nikolakakis, K. Kachrimanis, Artificial neural networks (ANNs) and partial least squares (PLS) regression in the quantitative analysis of cocrystal formulations by Raman and ATR-FTIR spectroscopy, J. Pharm. Biomed. Anal. 158 (2018) 214–224. https://doi.org/10.1016/j.jpba.2018.06.004.

[19] Y. Kou, Q. Li, X. Liu, R. Zhang, X. Yu, Efficient detection of edible oils adulterated with used frying oils through PE-film-based FTIR spectroscopy combined with DA and PLS, J.

Oleo Sci. 67 (2018) 1083–1089. https://doi.org/10.5650/jos.ess18029.

[20] N. Cebi, C.E. Dogan, A.E. Mese, D. Ozdemir, M. Arıcı, O. Sagdic, A rapid ATR-FTIR spectroscopic method for classification of gelatin gummy candies in relation to the gelatin source, Food Chem. 277 (2019) 373–381. https://doi.org/10.1016/j.foodchem.2018.10.125.

[21] G.C. Andrade, C.M. Medeiros Coelho, V.G. Uarrota, Modelling the vigour of maize seeds submitted to artificial accelerated ageing based on ATR-FTIR data and chemometric tools (PCA, HCA and PLS-DA), Heliyon 6 (2020) e03477. https://doi.org/10.1016/j.heliyon.2020.e03477.

[22] Y. Li, F. Li, X. Yang, L. Guo, F. Huang, Z. Chen, X. Chen, S. Zheng, Quantitative analysis of glycated albumin in serum based on ATR-FTIR spectrum combined with SiPLS and SVM, Spectrochim. Acta A Mol. Biomol. Spectrosc. 201 (2018) 249–257. https://doi.org/10.1016/j.saa.2018.05.022.

[23] Y.Y. Wang, J.Q. Li, H.G. Liu, Y.Z. Wang, Attenuated total reflection-fourier transform infrared spectroscopy (ATR-FTIR) combined with chemometrics methods for the classification of Lingzhi species, Molecules 24 (2019) 2210. https://doi.org/10.3390/molecules24122210.

[24] H.Z. Chen, G.Q. Tang, W. Ai, L.L. Xu, K. Cai, Use of random forest in FTIR analysis of LDL cholesterol and tri-glycerides for hyperlipidemia, Biotechnol. Prog. 31 (2015) 1693–1702. https://doi.org/10.1002/btpr.2161.

[25] B.R. Smith, K.M. Ashton, A. Brodbelt, T. Dawson, M.D. Jenkinson, N.T. Hunt, D.S.

Palmer, M.J. Baker, Combining random forest and 2D correlation analysis to identify serum spectral signatures for neuro-oncology, Analyst 141 (2016) 3668–3678. https://doi.org/10.1039/c5an02452h.

[26] J.M. Cameron, C. Rinaldi, H.J. Butler, M.G. Hegarty, P.M. Brennan, M.D. Jenkinson, K. Syed, K.M. Ashton, T.P. Dawson, D.S. Palmer, M.J. Baker, Stratifying brain tumour histological sub-types: the application of ATR-FTIR serum spectroscopy in secondary care, Cancers 12 (2020) 1710. https://doi.org/10.3390/cancers12071710.

[27] A.D. Surowka, D. Adamek, M. Szczerbowska-Boruchowska, The combination of artificial neural networks and synchrotron radiation-based infrared micro-spectroscopy for a study on the protein composition of human glial tumors, Analyst 140 (2015) 2428–2438. https://doi.org/10.1039/c4an01867b.

[28] S. Argov, J. Ramesh, A. Salman, I. Sinelnikov, J. Goldstein, H. Guterman, S. Mordechai, Diagnostic potential of Fourier-transform infrared microspectroscopy and advanced computational methods in colon cancer patients, J. Biomed. Opt. 7 (2002) 248–254. https://doi.org/10.1117/1.1463051.

[29] F. Elmi, A.F. Movaghar, M.M. Elmi, H. Alinezhad, N. Nikbakhsh, Application of FT-IR spectroscopy on breast cancer serum analysis, Spectrochim. Acta A Mol. Biomol. Spectrosc. 187 (2017) 87–91. https://doi.org/10.1016/j.saa.2017.06.021.

[30] D. Bury, C.L.M. Morais, M. Paraskevaidi, K.M. Ashton, T.P. Dawson, F.L. Martin, Spectral classification for diagnosis involving numerous pathologies in a complex clinical setting: a neuro-oncology example, Spectrochim. Acta A Mol. Biomol. Spectrosc. 206

(2019) 89–96. https://doi.org/10.1016/j.saa.2018.07.078.

[31] J.R. Hands, G. Clemens, R. Stables, K. Ashton, A. Brodbelt, C. Davis, T.P. Dawson, M.D. Jenkinson, R.W. Lea, C. Walker, M.J. Baker, Brain tumour differentiation: rapid stratified serum diagnostics via attenuated total reflection Fourier-transform infrared spectroscopy, J. Neuro-Oncol. 127 (2016) 463–472. https://doi.org/10.1007/s11060-016-2060-x.

[32] M.J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H.J. Butler, K.M. Dorling, P.R. Fielden, S.W. Fogarty, N.J. Fullwood, K.A. Heys, C. Hughes, P. Lasch, P.L. Martin-Hirsch, B. Obinaju, G.D. Sockalingum, J. Sulé-Suso, R.J. Strong, M.J. Walsh, B.R. Wood, P. Gardner, F.L. Martin, Using Fourier transform IR spectroscopy to analyze biological materials, Nat. Protoc. 9 (2014) 1771–1791. https://doi.org/10.1038/nprot.2014.110.

[33] A.C.S. Talari, M.A.G. Martinez, Z. Movasaghi, S. Rehman, I.U. Rehman, Advances in Fourier transform infrared (FTIR) spectroscopy of biological tissues, Appl. Spectrosc. Rev. 52 (2017) 456–506. https://doi.org/10.1080/05704928.2016.1230863.

[34] S. Kar, D.R. Katti, K.S. Katti, Fourier transform infrared spectroscopy based spectral biomarkers of metastasized breast cancer progression, Spectrochim. Acta A Mol. Biomol. Spectrosc. 208 (2019) 85–96. https://doi.org/10.1016/j.saa.2018.09.052.

[35] V.E. Sitnikova, M.A. Kotkova, T.N. Nosenko, T.N. Kotkova, D.M. Martynova, M.V. Uspenskaya, Breast cancer detection by ATR-FTIR spectroscopy of blood serum and multivariate data-analysis, Talanta 214 (2020) 120857. https://doi.org/10.1016/j.talanta.2020.120857.

[36] S. Kumar, C. Desmedt, D. Larsimont, C. Sotiriou, E. Goormaghtigh, Change in the

microenvironment of breast cancer studied by FTIR imaging, Analyst 138 (2013) 4058–4065. https://doi.org/10.1039/c3an00241a.

Figures

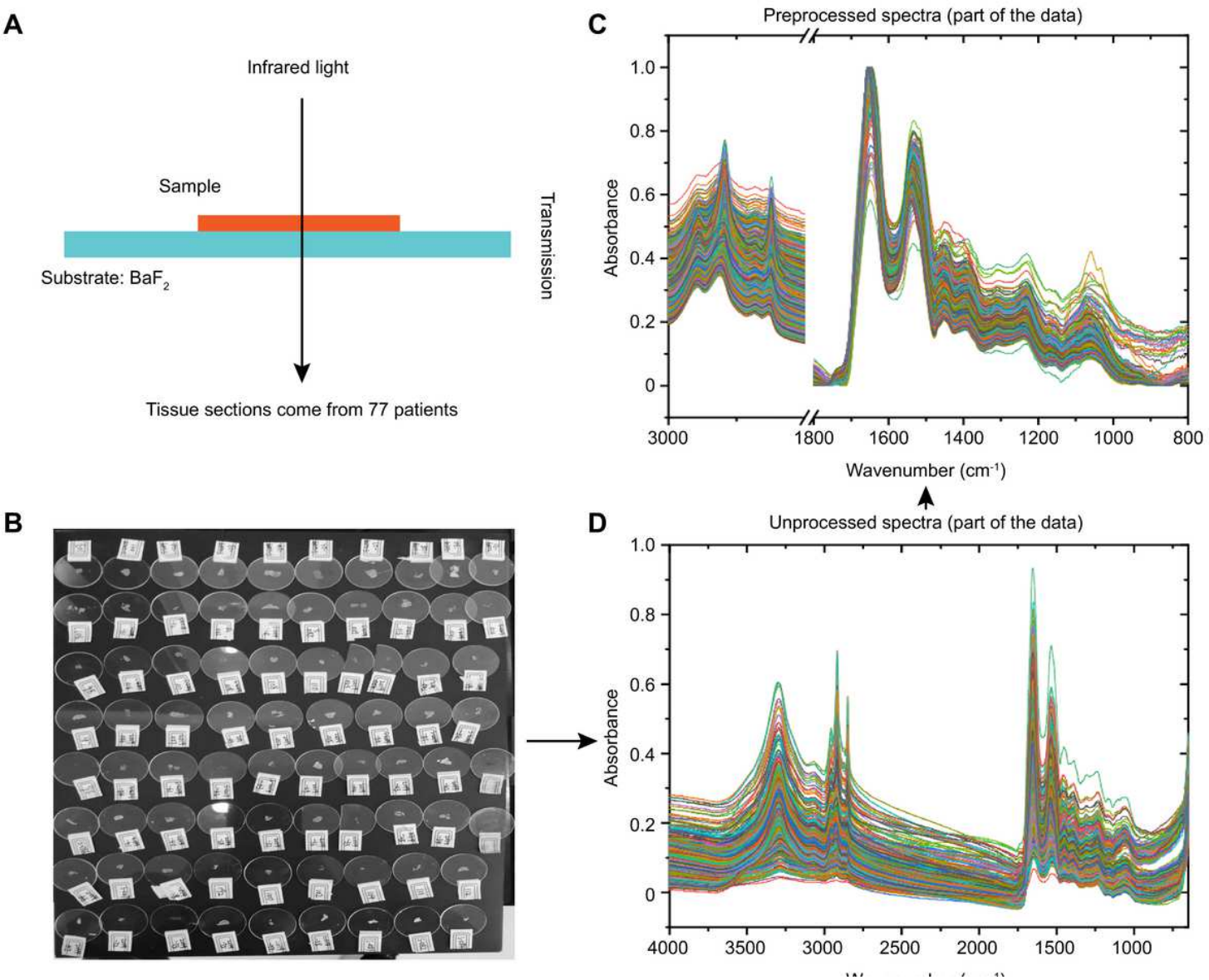


Figure 1

The glioma tissue was analyzed by FTIR spectroscopy, followed by preprocessing of the resulting data. (a) Spectral collection through transmission mode; (b) 8 μm thick tissue sections from 77 patients diagnosed with high- and low-grade glioma; (c) unprocessed spectra; and (d) spectra after preprocessing, including automatic baseline correction and amide I (1649 cm-1) normalization.

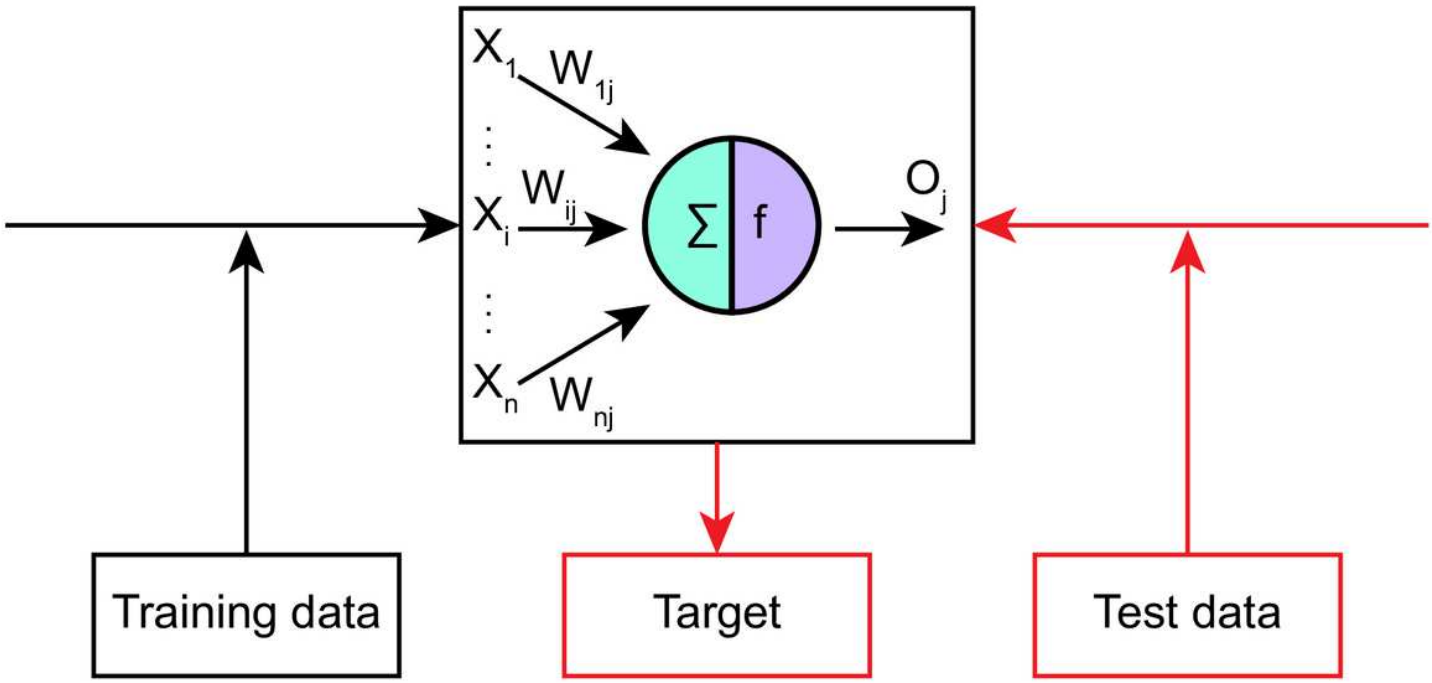


Figure 2

Schematic of the ANN method.

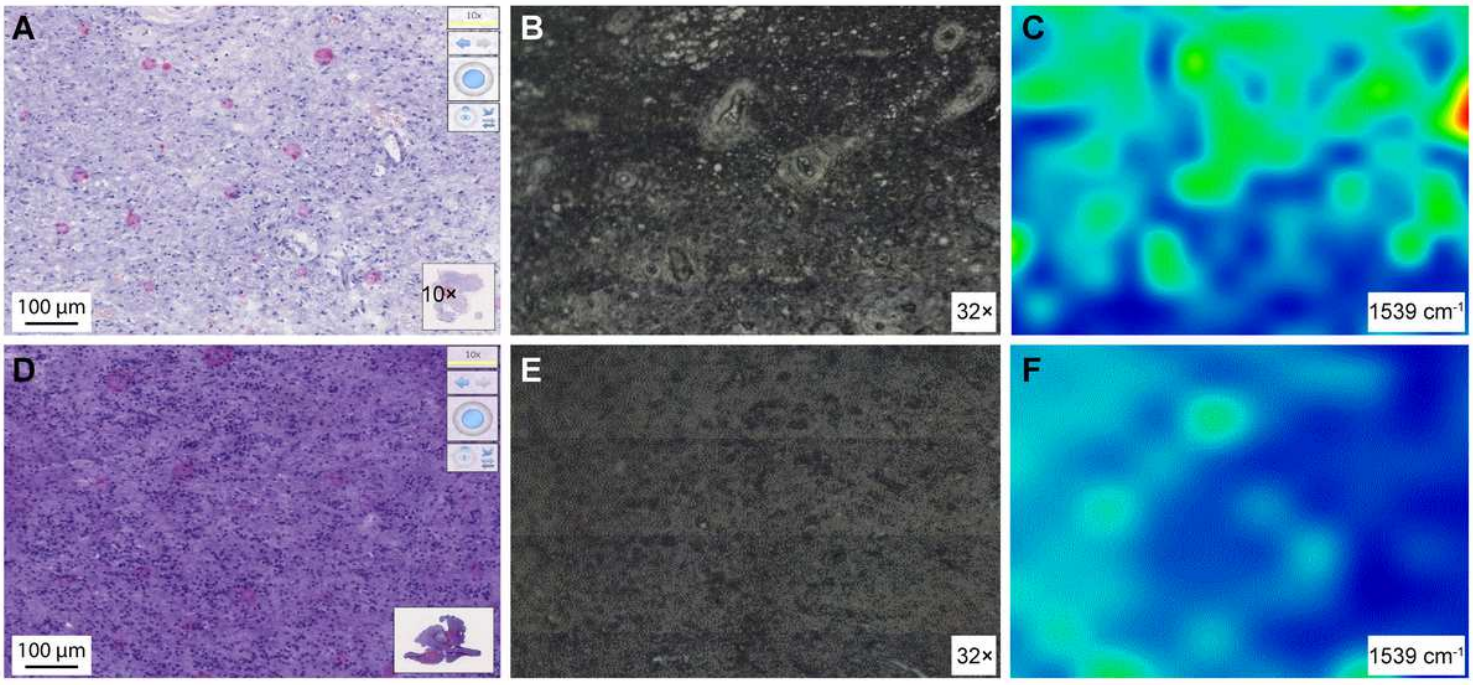


Figure 3

H&E staining of a tissue slide under a 10x microscope (a: high, d: low), the tissue morphology under a 32x microscope (b: high, e: low), and the corresponding IR mappings at 1539 cm-1 (c: high, f: low).

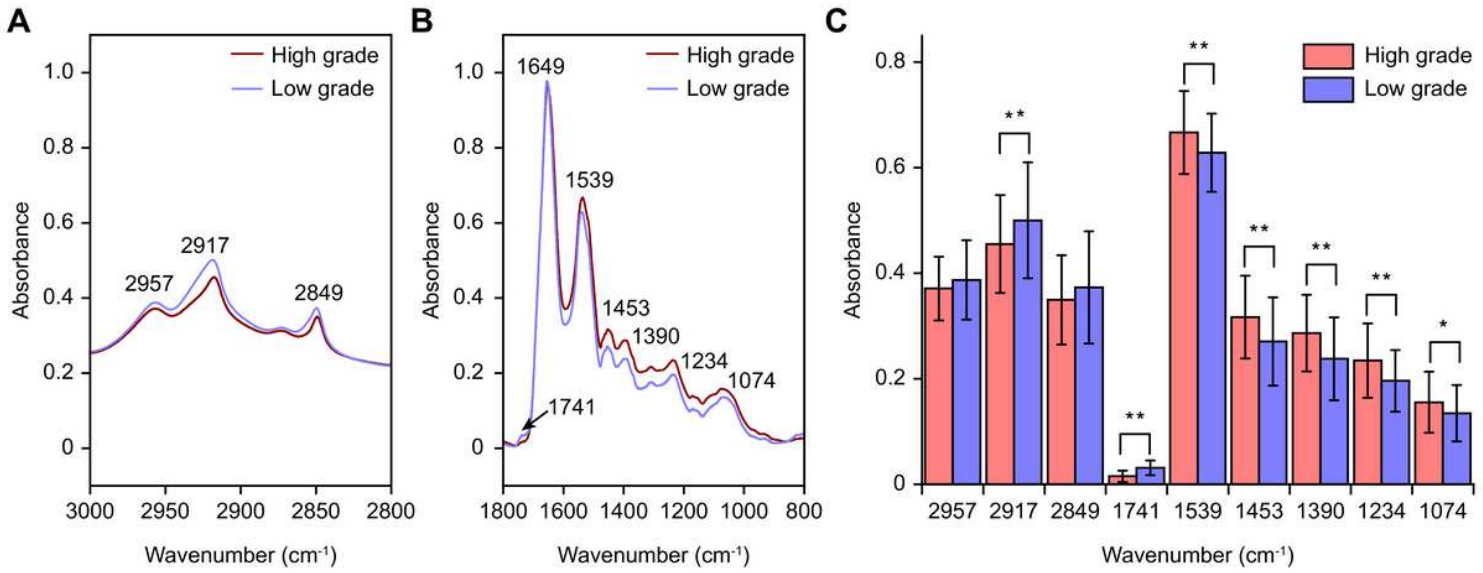


Figure 4

Average spectra of high and low grades of glioma in the ranges of 2800-3000 cm-1 (a) and 800-1800 cm-1 (b). (c) The relative intensity of the high-grade vs. low-grade bands at 2957 cm-1, 2917 cm-1, 2849 cm-1, 1741 cm-1, 1539 cm-1, 1453 cm-1, 1390 cm-1, 1234 cm-1 and 1074 cm-1. ** Mean value is particularly significant (P<0.01). * Mean value is significant (P<0.05). Red and green represent high- and low-grade data, respectively.

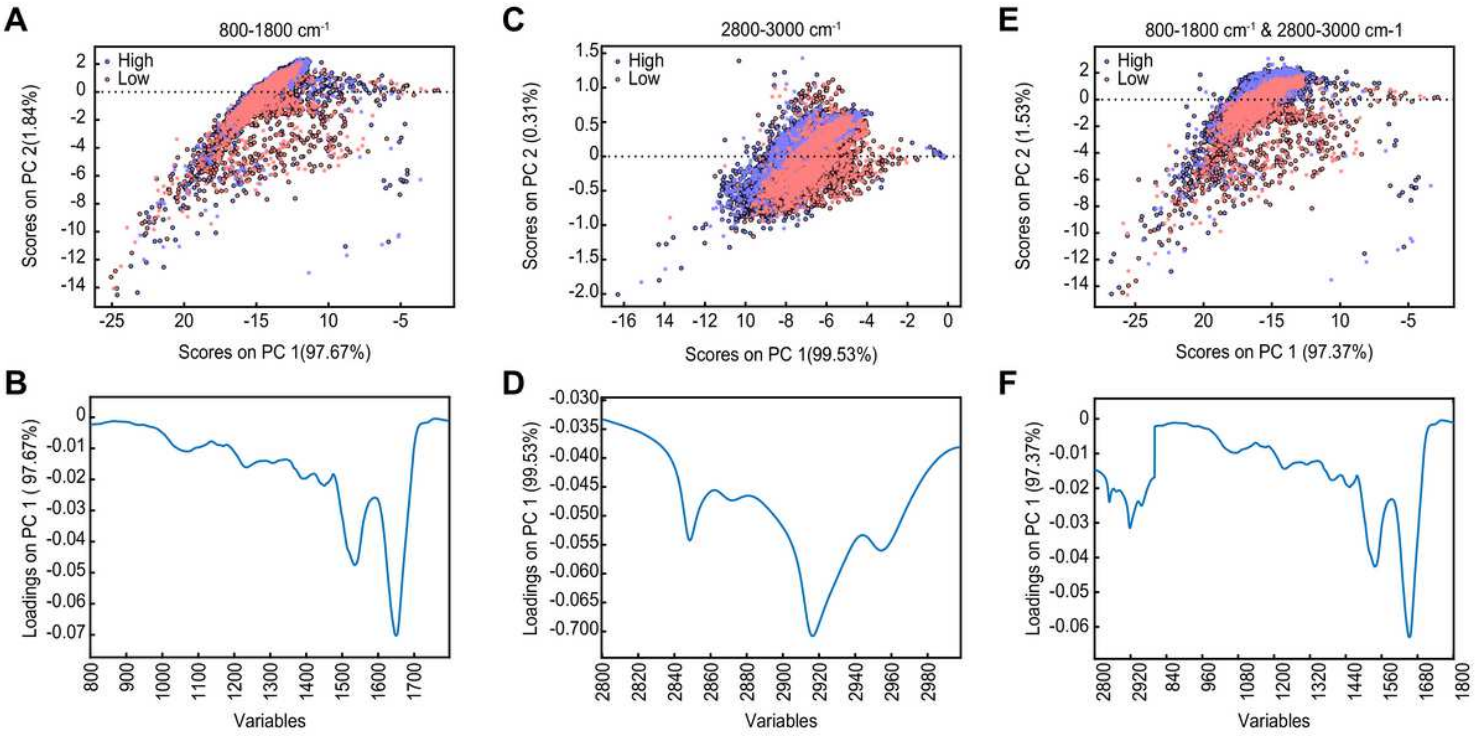


Figure 5

PCA-LDA results. (a), (c), and (e) are the score plots between PC1 and PC2 for high-grade vs. low-grade glioma in wavenumber ranges 1, 2, and 3. (b), (d), and (f) are the loadings of PC1 corresponding to (a), (c), and (e), respectively. Ranges 1, 2, and 3 represent the wavenumber ranges of 800-1800 cm-1, 2800-3000 cm-1, and both 800-1800 cm-1 and 2800-3000 cm-1, respectively.

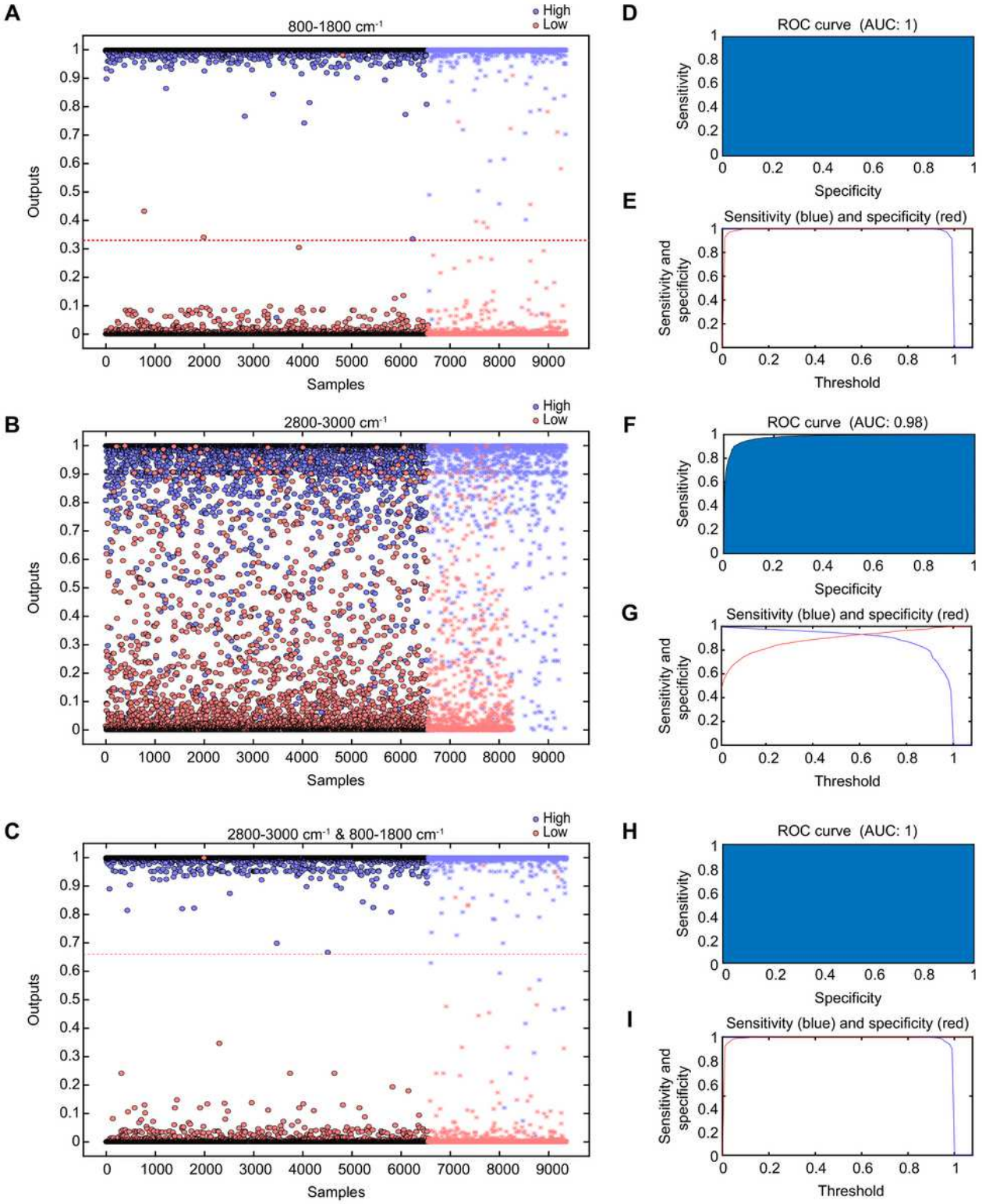


Figure 6

Corresponding to ranges 1, 2, and 3, respectively, (a), (b), and (c) are the classification results of the ANN output, (d), (f), and (h) are the ROC results, and (e), (g), and (i) are the related sensitivity (blue) and specificity (red) plots. Purple dots represent the high grade and pink dots represent the low grade in the training set. Purple stars indicate the high grade and pink stars indicate the low grade in the test set. Ranges 1, 2, and 3 denote the wavenumber ranges of 800-1800 cm-1, 2800-3000 cm-1, and both 800-1800 cm-1 and 2800-3000 cm-1, respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Onlinefloatimage1.png