Analyst

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Analyst*, 2023, **148**, 6061

Evaluation of Raman spectroscopy combined with the gated recurrent unit serum detection method in early screening of gastrointestinal cancer†

Kunxiang Liu, [a,b] Bo Liu,[a,b] Yu Wang,[a,b] Qi Zhao,[c,d] Qinian Wu*[e] and Bei Li*[a,b]

Received 24th July 2023,
Accepted 9th October 2023

DOI: 10.1039/d3an01259j

rsc.li/analyst

Gastric and colorectal cancers are significant causes of human mortality. Conventionally, the diagnosis of gastrointestinal tumors has been accomplished through image-based techniques, including endoscopic and biopsy procedures coupled with tissue staining. Most of these methods are invasive. In contrast, Raman spectroscopy has the advantages of being non-invasive and label-free and requiring no additional reagents, making it a potential tool for the detection of serum components. In this study, we collected Raman spectra of serum samples from patients with gastric cancer (n = 93) and colorectal cancer (n = 92) and from healthy individuals (n = 100). Analysis of Raman peak areas revealed that cancer patients had significantly higher peak areas at around 2923 cm^{-1} compared to normal individuals, which corresponded to the presence of lipids and proteins. We successfully achieved the early screening of gastrointestinal tumors using the improved gated recurrent unit (GRU) algorithm and traditional machine learning methods. The accuracy of identifying digestive tract tumors using different recognition models exceeds 84.72%, with support vector machine (SVM) and GRU achieving 100% accuracy. The use of GRU further demonstrated its ability to differentiate subtypes of gastric and colorectal cancers based on the degree of differentiation and stage, with a recognition accuracy exceeding 95%, which is challenging using traditional machine learning methods. Furthermore, our study revealed that principal component analysis (PCA) dimensionality reduction has a limited impact on the recognition results obtained using different recognition models.

Introduction

Globally, gastric cancer represents a frequently diagnosed malignancy, with an annual incidence of about 1 million cases. Due to the advanced stage at which gastric cancer is often detected and its heightened mortality rate, it was projected to account for 769 000 deaths in 2020.[1] Early detection of gastric cancer is pivotal in improving the survival rate and prognosis, as the 5-year survival rate for early gastric cancer can surpass 90% through effective treatment. Therefore, early detection is deemed the most important modality for improving gastric cancer outcomes.[2] Colorectal cancer ranks third among the commonly diagnosed cancers in men and second in women worldwide. With high metastatic potential and poor prognosis, colorectal cancer is also one of the leading causes of death globally.[3]

At present, the main diagnostic methods for colorectal cancer include the fecal occult blood test, blood tests, and colonoscopy for high-risk individuals.[4,5] Colonoscopy coupled with tissue staining is regarded as the gold standard. Additionally, for the diagnosis of gastric cancer, endoscopy coupled with histopathology remains the primary diagnostic approach.[6] Although endoscopy, which is the gold standard for gastrointestinal diagnosis, has reliable accuracy, it is difficult to popularize it to routine screening diagnosis because endoscopy is invasive and affected by patient compliance and operator techniques.[7] Therefore, we need a new technique for practical and rapid serological testing for early screening of gastrointestinal tumors.

Since its discovery by the Indian scientist C. V. Raman in 1928, Raman spectroscopy techniques have undergone significant development and have widespread application across various fields. The Raman scattering between matter and photons can be reflected in Raman spectra, thus visualizing

[a]*Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, P. R. China. E-mail: beili@ciomp.ac.cn; Tel: +0431-86708966*
[b]*University of Chinese Academy of Sciences, Beijing 100049, P. R. China*
[c]*State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong 510060, P. R. China*
[d]*Cancer Microbiome Platform, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou, Guangdong 510060, P. R. China*
[e]*Department of Pathology, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong 510060, P. R. China. E-mail: wuqn@sysucc.org.cn*
† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3an01259j

the molecular motion of matter. Owing to its ability to reveal the biomolecular changes that malignant tumors instigate in the human body, Raman spectroscopy has emerged as a new diagnostic approach for malignant tumors. Recent years have seen successful applications of Raman spectroscopy in the diagnosis of malignant tumors located in diverse parts of the human body, including breast cancer, brain cancer, cervical cancer, gastric cancer, *etc*.[8–13]

Numerous studies have utilized Raman spectroscopy to diagnose gastric cancer based on blood, tissue, cell line, and other samples. Such studies have successfully distinguished between cancer and normal samples while also identifying different stages of gastric cancer.[7,12,14–16] However, these studies generally involved a limited sample size and did not consider the varying degrees of cancer cell differentiation. Furthermore, the analysis of Raman spectra in these studies mostly relied on traditional machine learning methods such as PCA and linear discriminant analysis (LDA) that yielded positive results.[11] Nonetheless, as the number of recognition types and the amount of spectral data continue to increase, simple machine learning methods may become inadequate. Therefore, the incorporation of deep learning spectral processing methods is crucial for tackling complex challenges associated with Raman spectroscopy data.

Convolutional neural networks (CNNs) are the most utilized models in deep learning applications for Raman spectroscopy recognition.[13,17–19] A CNN is capable of comprehensively mining data features that result in highly accurate classifications. Nonetheless, as the amount of data and spectral categories increase, a CNN necessitates the continuous augmentation of network depth, leading to heightened computational complexity and processing time.[20] Additionally, while CNNs excel in processing image data, they are circumscribed in their ability to process one-dimensional spectral data.[21] Recurrent neural networks (RNNs), on the other hand, can effectively utilize entire spectral data and call upon earlier data for subsequent analyses, providing an edge in processing Raman spectral data.[22]

In long-term RNN training, the primary disadvantage is the problem of long-term dependency. To overcome this issue, long short-term memory (LSTM) neural networks have been used to selectively maintain or discard certain information from previous stages using three control gates and additive iteration to circumvent the problem of gradient explosion.[23] However, the drawback of LSTM is that its computational complexity is more than three times that of the original RNN. Therefore, we adopted a GRU with the same functionality and basic structure as LSTM, reducing computational complexity.

Our study analyzed the characteristic peaks of Raman spectra of gastric and colorectal cancer sera and compared the differences in spectra between gastrointestinal tumor serum and normal human serum through statistical analysis of characteristic peak areas. We used an improved GRU network to diagnose gastric and colorectal cancers, identifying different stages and different degrees of differentiation for gastric cancers and different degrees of differentiation for colorectal cancers, and compared the results with those of conventional machine learning recognition methods. Additionally, we investigated the impact of principal component analysis (PCA) on high-dimensional Raman spectroscopy data dimensionality reduction and classification. The workflow of using GRU and Raman spectroscopy to diagnose gastrointestinal tumors and identify tumor subtypes is shown in Fig. 1.

Materials and methods

Collection and preparation of serum samples

We collected serum samples from 93 patients diagnosed with gastric cancer and 92 patients diagnosed with colorectal cancer at Sun Yat-sen University Cancer Center between 2007 and 2013. The eligibility criteria included the diagnosis of gastric cancer by gastroscopy and pathological biopsy, the absence of tumors in other body systems, non-existence of significant cardiac, pulmonary, hepatic, renal, or other organ dysfunction, and no prior surgical or chemotherapy treatments before sample collection. The collection conditions for colorectal cancer samples were comparable to those for gastric cancer samples. We also collected serum samples from 100 non-tumor volunteers without gastrointestinal disease history as a control group at The Third Bethune Hospital of Jilin University.

Following an overnight fast spanning 10 hours, 3 ml of peripheral blood was withdrawn from each subject. After coagulation, centrifugation was carried out at 3000 rpm for 10 minutes. After this, the supernatant serum sample was collected in a specialized cryopreservation tube and maintained in a −80 °C refrigerator until Raman measurements were taken.

Raman measurements

We used a Raman spectroscopy system (R300 (objective lens: Olympus, 100×, NA = 0.8), Hooke Instruments, Changchun, China) with a laser wavelength of 532 nm to collect Raman spectra of serum samples. Raman spectra were obtained for ten different locations of each patient, with three spectra being collected for each location. The three spectra were averaged at each location, resulting in 10 average spectra of serum samples from diverse locations of each patient for the ensuing spectral data analysis. The conditions for Raman spectrum collection were as follows: a grating of 600, a laser power of 5 mW, and an integral time of 3 s. The Raman spectra ranged approximately from 400 cm^{-1} to 3800 cm^{-1}.

Data preprocessing

Data preprocessing can effectively attenuate unnecessary spectral signal changes and interference caused by instrument fluctuations and fluorescent substances.[24] The preprocessing process for the Raman spectra of the serum samples includes cosmic ray removal, filtering, baseline correction, and normalization. By linear fitting the points around the singular values

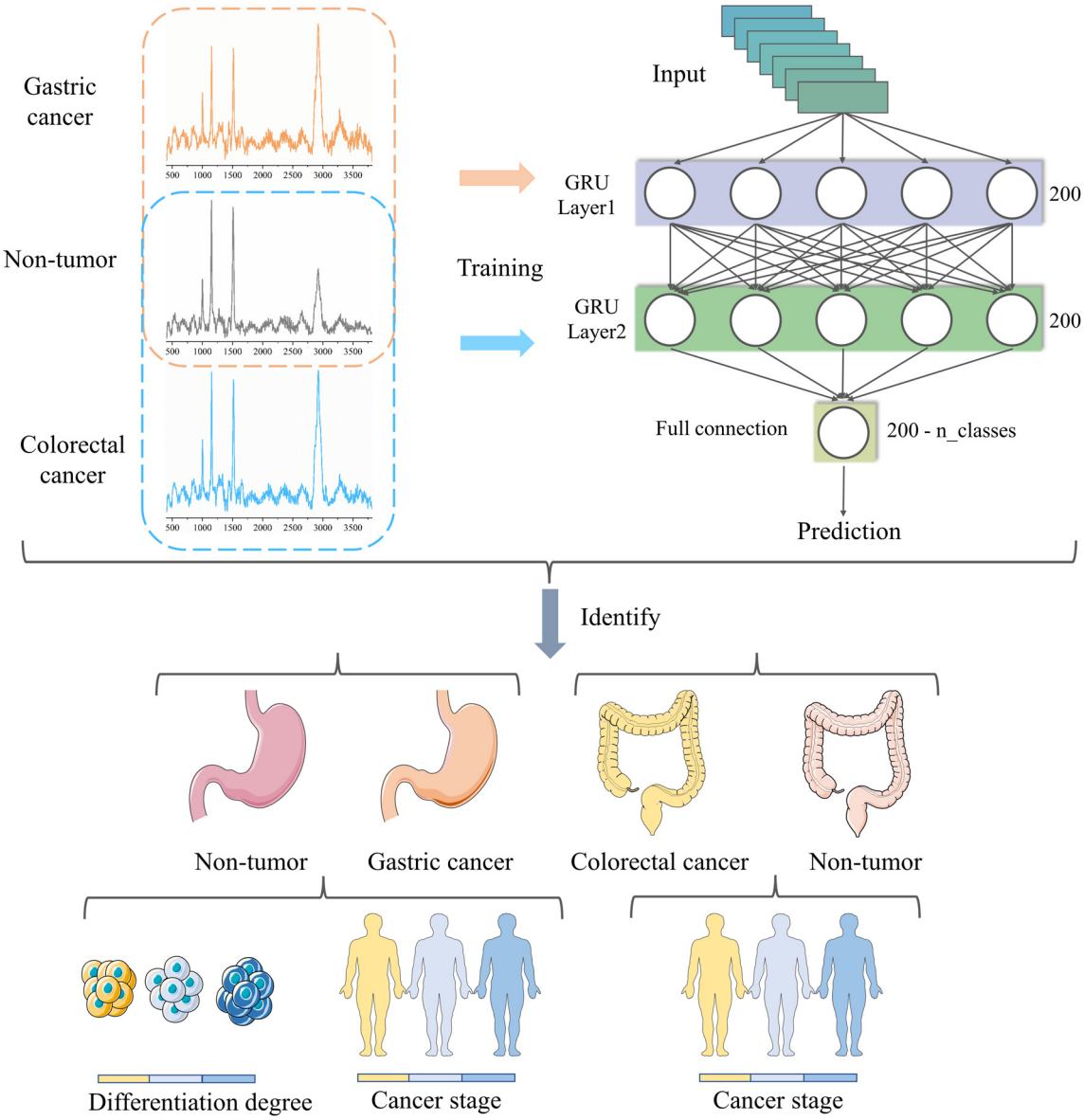

Fig. 1 The workflow of using GRU and Raman spectroscopy to identify gastrointestinal tumors and tumor subtypes.

of the spectral data, we removed the cosmic rays. The parameters were set to filter size = 5 and dynamic factor = 4.5. We filtered the data with a Savitzky–Golay filter with filter window width = 5 and filter fitting order = 3. We used the airPLS algorithm to gradually approximate the Raman spectral baseline with lambda = 100 and maximum number of iterations = 15. Finally, the spectral data were normalized by min–max normalization.[25,26] It is worth mentioning that to handle minor differences in the *x*-axis among the different spectra, we processed the data using cubic spline interpolation. All processing was conducted in Python.

PCA downscaling

Raman spectra have nearly one thousand features, which are typical of high-dimensional feature data. Directly analyzing all features within Raman spectra can generate considerable noise interference, potentially compromising classification accuracy. PCA is one of the most frequently used dimensionality reduction algorithms. PCA works by reconstructing *k*-dimensional features based on the original *n*-dimensional features, with the goal of identifying the direction of maximum variance of the data set as the principal component. PCA algorithms have seen extensive use in data analysis studies revolving around Raman spectroscopy. In each classification assignment within this study, we employed PCA algorithms to extract primary components that could reflect data differences, where the contribution was greater than 95%. We compared the discrepancies between PCA algorithms utilizing various classification algorithms and the direct use of classification algorithms to analyze Raman spectral data.

Machine learning classification methods

We constructed a machine learning classification model utilizing SVM, *k*-nearest neighbor (KNN), and LDA, commonly found within the sklearn machine learning library. For each classification task, we divided the dataset into training and testing sets at a ratio of 4 : 1, further subdividing the training set into a training set and a validation set at a ratio of 4 : 1. During the training process, a 5-fold cross-validation technique was employed, saving the optimal model after five training sessions for predicting the testing dataset. The accuracy of the optimal model in predicting the test set is the outcome presented in the article.

Gated recurrent unit

GRU, a type of RNN, functions similarly to LSTM in controlling the flow and loss of features by introducing a gate mechanism (forgetting gate, memory gate, and output gate) to overcome gradient problems in long-term memory and backpropagation. While performing comparably to LSTM,[27] the GRU[28] possesses fewer output gates and parameters, making it computationally cheaper.

For our experiment, we utilized a two-layer GRU neural network to process Raman spectral data, determining the appropriate number of GRU hidden layer nodes based on statistical feature bandwidth information to extract spectral feature information. We included dropout to the connection layer with a parameter of 0.5 to avoid overfitting the model. Each layer had 200 neurons, producing output for different target categories through the fully connected layer. To address the challenge of Raman spectrum recognition in serum samples of digestive tract tumors, we utilized the GRU to fully learn pre and post-spectrum data, thereby systematically exploring the entire Raman spectrum's characteristics. The activation function for the model was ReLU, and the loss function was the cross-entropy loss function. The ADAM optimizer trained the network, utilizing the following parameters: a learning rate of 0.0001 and exponential decay rates at $\beta\_1 = 0.5$ and $\beta\_2 = 0.999$. We trained the model for 1000 iterations, saving accuracy and loss values for each training session.

Results

Raman analysis of serum from patients with gastrointestinal tumors

We collected a total of 930 Raman spectra (mean spectra) from 93 gastric cancer patients, of which 92 had a clear diagnosis of differentiation degree (20 cases were well and moderately differentiated and 72 cases were poorly differentiated) and 91 had a clear diagnosis of the stage (32 cases of stage I + II and 59 cases of stage III + IV). The specific gender and age distribution are shown in Table 1.

A total of 920 Raman spectra (mean spectra) were collected from 92 patients with colorectal cancer, of which 91 cases had a clear diagnosis of differentiation degree (67 cases with moderately differentiated adenocarcinoma, 10 cases with moderately to poorly differentiated adenocarcinoma, 2 cases with poorly differentiated adenocarcinoma, 2 cases with well-differentiated adenocarcinoma, and 10 cases with mucinous adenocarcinoma) and 91 cases had a clear diagnosis of the stage (including 3 cases of stage I, 25 cases of stage II, 4 cases of stage III, 1 case of stage IIIA, 36 cases of stage IIIB, 6 cases of stage IIIC, and 16 cases of stage IV). Among them, there was only one case of stage IIIA, which was not representative in the spectral analysis, so we removed the data of this case. The specific gender and age distribution are shown in Table 2.

Table 1 Detailed information about the gastric cancer patients in this study

	Degree of differentiation (n = 92)		Stages of gastric cancer (n = 91)	
	Well and moderately differentiated (n = 20)	Poorly differentiated (n = 72)	I + II (n = 32)	III + IV (n = 59)
Male	16 (80%)	41 (56.94%)	25 (78.125%)	31 (52.54%)
Female	4 (20%)	31 (43.06%)	7 (21.875%)	28 (47.46%)
Age (<56)	9 (45%)	41 (56.94%)	14 (43.75%)	35 (59.32%)
(>56)	11 (55%)	31 (43.06%)	18 (56.25%)	24 (40.68%)

The 100 healthy individuals used as controls included 50 males and 50 females, with a mean age of 48.13 ± 15.2890.

Raman spectroscopy, as a new optical diagnostic method, can qualitatively and quantitatively reflect the biochemical components of biological samples, such as proteins, lipids, and nucleic acids. We collected a total of 930 Raman spectra for gastric cancer, 920 Raman spectra for colorectal cancer, and 999 Raman spectra for non-tumor control and analyzed the differences of Raman spectra between sera of gastric and colorectal cancer patients and non-tumor individuals, respectively.

Fig. 2a shows the differences between the average Raman spectra of gastric cancer and non-tumor groups. The main characteristic peaks are located at 857, 1000, 1152, 1275, 1332, 1435, 1515, 1602, and 2923 cm^{-1}. Fig. 2b shows the difference between the average Raman spectra of the intestinal cancer group and the non-tumor group. The main characteristic peaks are located at 857, 952, 1000, 1152, 1275, 1332, 1435, 1515, 1602, and 2923 cm^{-1}. By searching the literature on the application of Raman spectroscopy, it was found that these characteristic peaks are mainly related to proteins, lipids, amide III, and nucleic acids. The specific Raman peak position comparison is shown in Table S1.†

To have a more visual and macroscopic view of the variation in the content of the components corresponding to the above Raman feature peak positions, we calculated the peak areas of the 10 feature peaks for each Raman spectrum and plotted the heat map (Fig. S1†). It is worth mentioning that to better remove the Raman spectral baseline and to better compare the differences in peak areas between spectra, in this process, we used the airPLS algorithm to gradually approximate the Raman spectral baseline with lambda = 100 and maximum number of iterations = 15. From the heat map, we can visualize

Table 2 Detailed information about the colorectal cancer patients in this study

Degree of differentiation (n = 91)					
	Moderately differentiated (n = 67)	Moderately to poorly differentiated (n = 10)	Poorly differentiated (n = 2)	Well differentiated (n = 2)	Mucinous adenocarcinoma (n = 10)
Male	42 (62.69%)	7 (70%)	2 (100%)	1 (50%)	9 (90%)
Female	25 (37.31%)	3 (30%)	0	1 (50%)	1 (10%)
Age	53.55 ± 11.8325	51.3 ± 9.8899	68 ± 15	56 ± 15	56.4 ± 10.8738

Stages of colorectal cancer (n = 91)							
	I (n = 3)	II (n = 25)	III (n = 4)	IIIA (n = 1)	IIIB (n = 36)	IIIC (n = 6)	IV (n = 16)
Male	2 (66.67%)	17 (68%)	4 (100%)	0	25 (69.44%)	4 (60%)	9 (56.25%)
Female	1 (33.33)	8 (32%)	0	1 (100%)	11 (30.56%)	2 (40%)	7 (43.75%)
Age	48.67 ± 10.1434	52.68 ± 11.4916	44.25 ± 6.2998	77	56.19 ± 11.8786	51.17 ± 10.1229	54.125 ± 11.8895

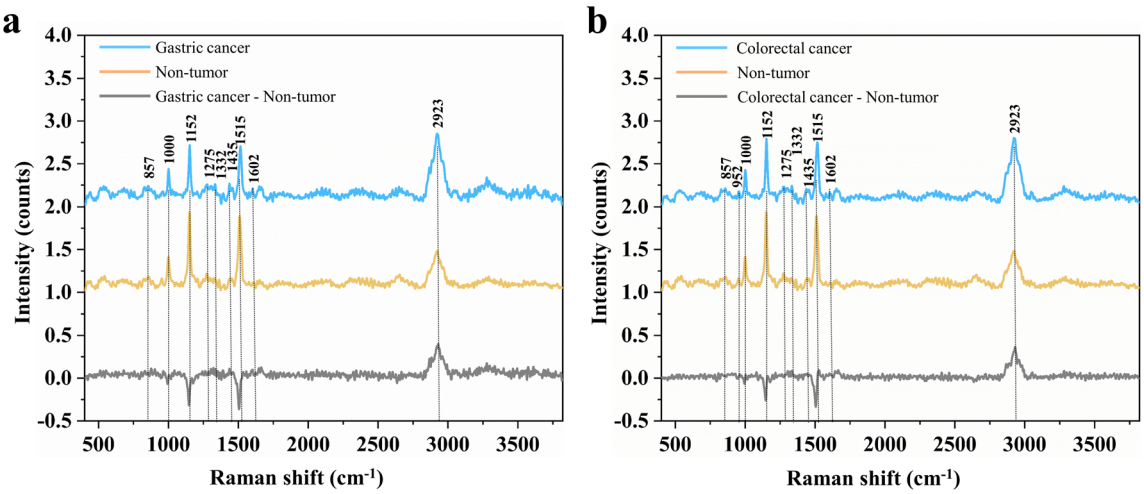

Fig. 2 Differences in Raman spectra between the gastric cancer and the non-tumor group, and the colorectal cancer and the non-tumor group.

the differences in intensity between gastric cancer and non-tumor groups at 1000, 1152, 1435, 1515, 1602, and 2923 cm^{-1}, especially at 2923 cm^{-1}. The differences in intensity between colorectal cancer and non-tumor groups are observed at 1000, 1152, 1515, 1602, and 2923 cm^{-1}. It is obvious that the intensity of Raman peaks at the 2923 cm^{-1} position is significantly higher in both gastric cancer and colorectal cancer patients than in non-tumor individuals, and the C–H peaks represented at this position are mainly related to the lipids and proteins in biological samples.

Fig. S2† shows the average Raman spectra of patients with different differentiation and different stages of gastric and colorectal cancers; just from the spectra, it is difficult for us to visualize the differences between the spectra with the naked eye. So, we need to identify them with the help of statistical or artificial intelligence-related methods.

Distinguishing gastric cancer and colorectal cancer based on Raman spectroscopy and deep learning

We implemented the recognition of Raman spectra of gastric and colorectal cancers using an improved GRU network with 1000 training cycles. A comparison with commonly used machine learning models was also made. We validated the effectiveness of PCA-SVM, SVM, PCA-KNN, KNN, PCA-LDA, LDA, PCA-GRU, and GRU to identify Raman spectra. The results of different methods to identify serum Raman spectra of patients with gastric cancer (n = 930) *versus* the non-tumor group (n = 999) were 100%, 100%, 99.48%, 100%, 98.19%, 84.72%, 100%, and 100%. The results of different methods to identify colorectal cancer patients (n = 920) *versus* non-tumor individuals (n = 999) were 100%, 100%, 99.22%, 99.74%, 98.70%, 86.46%, 100%, and 100%. The specific results are shown in Table 3.

According to the results, machine learning methods such as SVM, KNN and LDA can identify gastric or colorectal cancer, and most of the identification results can reach over 98% accuracy. Among them, SVM is known as the "king of binary classification", so it is the best in this recognition, but the improved GRU is not inferior either. The confusion matrix, receiver operating characteristic (ROC), training accuracy and loss values of gastrointestinal tumor spectral data identified by GRU and PCA-GRU (contribution >95%) are shown in Fig. 3. In conclusion, Raman spectroscopy combined with an AI algor-

Table 3 Identification results of different classification algorithms for distinguishing gastric and colorectal cancers

	PCA-SVM	SVM	PCA-KNN	KNN	PCA-LDA	LDA	PCA-GRU	GRU
GC & NT	100	100	99.48	100	98.19	84.72	100	100
CC & NT	100	100	99.22	99.74	98.70	86.46	100	100

Abbreviations: GC, gastric cancer; NT, non-tumor; and CC, colorectal cancer.

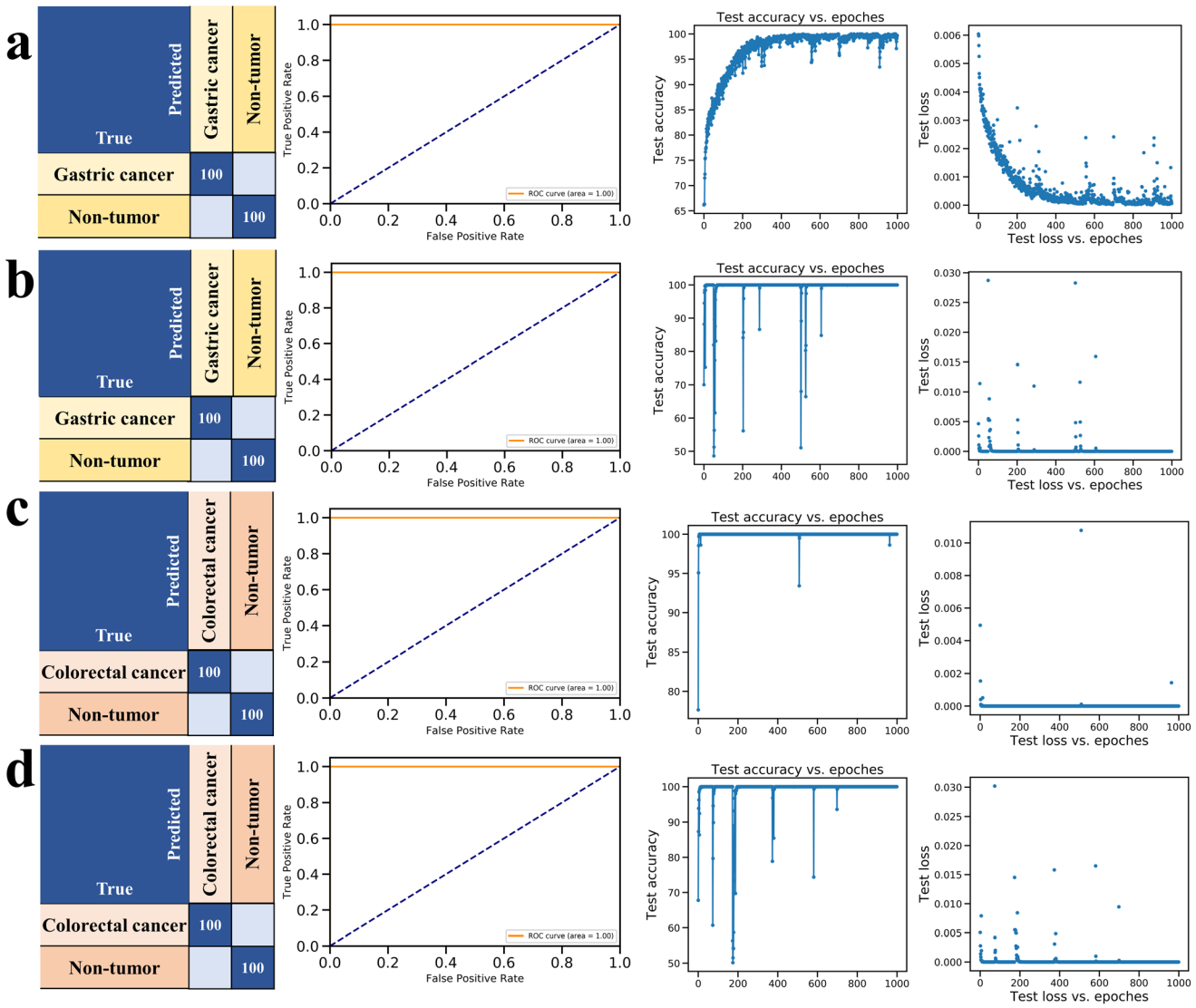

Fig. 3 The confusion matrix, ROC, training accuracy and loss values of gastrointestinal tumor spectral data identified by GRU and PCA-GRU (contribution >95%).

ithm can distinguish gastric and colorectal cancers using serum samples.

Identification of gastric and intestinal cancer subtypes based on Raman spectroscopy and deep learning

It is difficult for us to visually detect differences in different degrees of differentiation or different subtypes of gastric or colorectal cancer by Raman spectroscopy, so we need to use artificial intelligence algorithms to assist in identification. The accuracy of our Raman spectra using PCA-SVM, SVM, PCA-KNN, KNN, PCA-LDA, LDA, PCA-GRU and GRU to distinguish well and moderately differentiated gastric cancer (n = 200) and poorly differentiated (n = 720) gastric cancer was 85.87%, 80.98%, 91.85%, 91.85%, 82.61%, 82.61%, 95.70% and 95.70%. The accuracy of identifying gastric cancer of stage I + II (n = 320) and stage III + IV (n = 590) was 75.27%, 79.67%,

Table 4 Accuracy of different classification methods for identifying subtypes of gastric and colorectal cancers

	PCA-SVM	SVM	PCA-KNN	KNN	PCA-LDA	LDA	PCA-GRU	GRU
DG (%)	85.87	80.98	91.85	91.85	82.61	82.61	95.70	95.70
SG (%)	75.27	79.67	87.91	89.56	68.13	73.08	95.60	93.40
SC (%)	68.39	84.48	91.95	90.23	67.24	68.39	96.60	98.30

Abbreviations: DG, differentiation degree of gastric cancer; SG, stages of gastric cancer; and SC, stages of colorectal cancer.

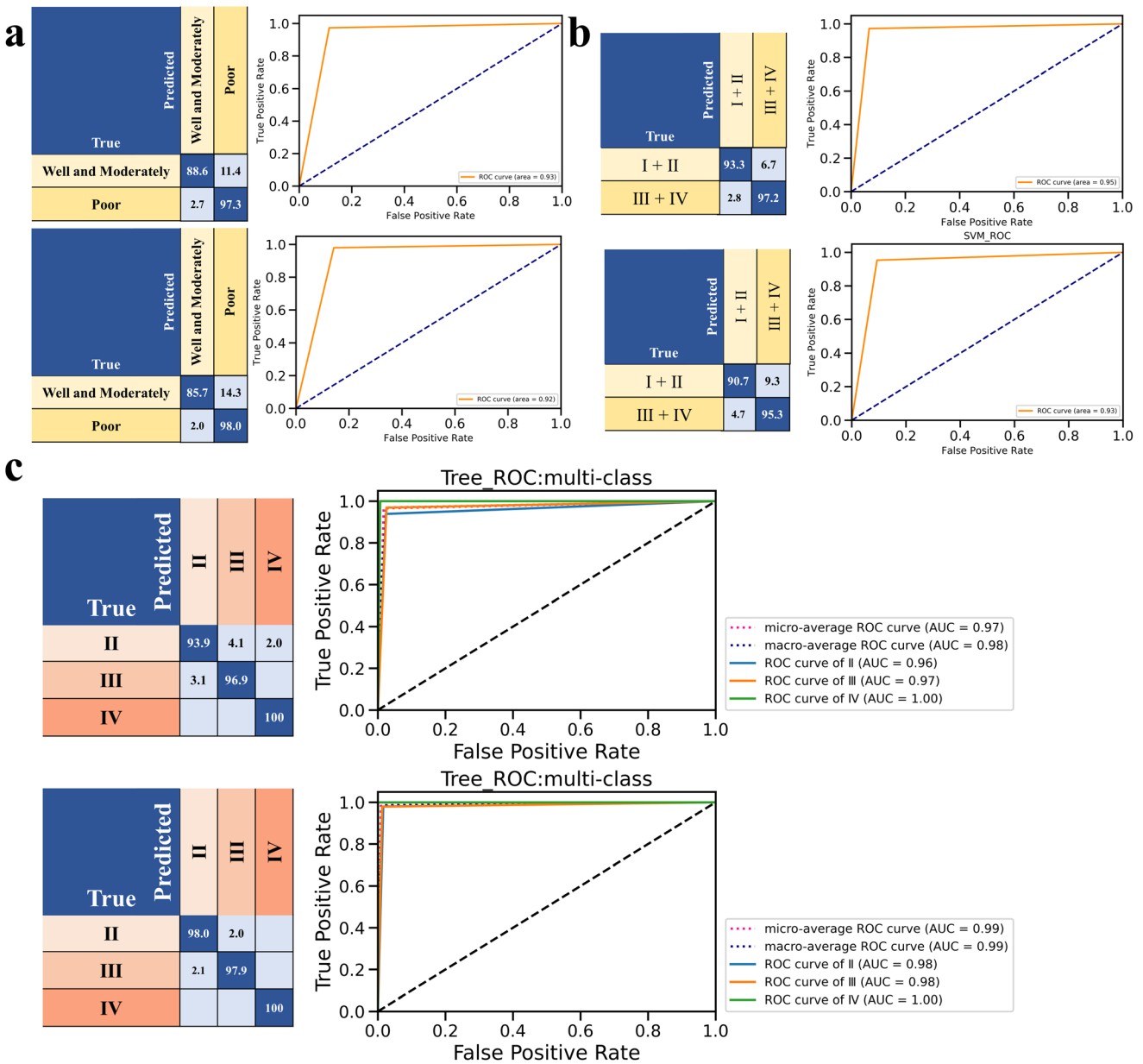

Fig. 4 Confusion matrix and ROC for GRU identification of gastric and colorectal cancers with different degrees of differentiation and different stages. (a) Confusion matrix and ROC for GRU differentiation of Raman spectra of well and moderately differentiated (n = 200) and poorly differentiated (n = 720) gastric cancers. (b) Confusion matrix and ROC for GRU differentiation of Raman spectra of gastric cancers of stage I + II (n = 320) and stage III + IV (n = 590). (c) Confusion matrix and ROC for GRU differentiation of Raman spectra of colorectal cancers of stage II (n = 250), III (n = 460), and IV (n = 160).

87.91%, 89.56%, 68.13%, 73.08%, 95.60% and 93.40%, respectively. To ensure a roughly balanced sample size between the different kinds of colorectal cancer during the identification of colorectal cancer subtypes, we will only differentiate between the different stages of colorectal cancer here and regard all stages III, IIIB, and IIIC of colorectal cancer as stage III for the purpose of identification. The accuracy of identifying stage II, III (III + IIIB + IIIC), and IV colorectal cancers was 68.39%, 84.48%, 91.95%, 90.23%, 67.24%, 68.39%, 96.60% and 98.30%, respectively. The specific results are shown in Table 4. According to the results, PCA can indeed be used as a data processing method to reduce the amount of Raman spectral data and has little effect on the identification results.

The GRU network showed higher accuracy than other methods in identifying Raman spectra with small differences and when there were more categories. The accuracy of GRU reached more than 95% when performing detailed classification of gastric and colorectal cancers. The confusion matrix and ROC of spectral data for GRU and PCA-GRU to identify gastric cancer and colorectal cancer subtypes are shown in Fig. 4. In conclusion, Raman spectroscopy combined with an artificial intelligence algorithm can enable us to distinguish gastric and colorectal cancer subtypes using serum samples.

Discussion

Raman spectra can reflect the biochemical composition of biological samples and reflect the compositional differences between gastric or colorectal cancer serum and non-tumor serum. By analyzing the Raman spectra of gastric and colorectal cancers, we found that the area of the Raman characteristic peak near 2923 cm^{-1} in the serum of patients with gastric or colorectal cancer was generally higher than that of healthy human serum. This peak was mainly associated with C–H peaks in lipids and proteins.

We also improved and built a GRU network for identifying gastrointestinal cancer Raman spectral data and obtained good identification results. Compared with machine learning algorithms such as SVM, KNN and LDA, GRU performs well on both binary and multi-classification problems. In distinguishing gastric cancer from non-tumor and colorectal cancer from non-tumor, the lowest accuracy of 84.72% was recognized by ordinary machine learning methods, and SVM and GRU were able to achieve 100% accuracy. GRU also showed good recognition ability when identifying Raman spectra of cancer subtypes with smaller differences, with recognition accuracy >95%. Therefore, Raman spectroscopy combined with deep learning can enable us to distinguish gastric and colorectal cancers, as well as to identify cancer subtypes. Besides, we also found that PCA dimensionality reduction has little effect on the recognition of Raman spectra by the classification algorithm. The GRU-based Raman spectral recognition method is a general spectral classification algorithm that can be applied not only to gastrointestinal tumor early screening, but also to other different types of spectral recognition.

Raman spectroscopy is, after all, a new technique for optical tumor early screening, and there is no clear standard for the interpretation of certain peaks and contents in Raman spectra. In the present study, we only analyzed biochemical components in gastrointestinal tumor serum samples qualitatively and semi-quantitatively. In the next step, we will combine liquid chromatography-mass spectrometry and other means to assist in the verification of the compositional differences reflected by Raman spectroscopy. In addition, we will also implement and refine Raman spectroscopy-based cancer diagnosis using other samples such as serum and tissues.

Conclusions

We analyzed and compared Raman spectra of sera from patients with gastrointestinal tumors with those of normal subjects and found that gastrointestinal tumor sera exhibited a significantly higher peak area at the C–H position at 2923 cm^{-1}, reflecting metabolic differences in proteins and lipids between cancer and normal sera. By combining Raman spectroscopy with a deep learning spectral recognition model, we have successfully achieved the early screening of gastric cancer and colorectal cancer. Ordinary machine learning methods have a minimum accuracy of 84.72%, while SVM and GRU can achieve 100% accuracy. We used the improved GRU network algorithm to successfully distinguish gastric cancers with different degrees of differentiation and different stages, as well as colorectal cancers with different stages, all with an accuracy rate of more than 95%, which is difficult to achieve with ordinary machine learning methods. Besides, we also found that the PCA dimensionality reduction method has little effect on the recognition accuracy of different classification models. Overall, Raman spectroscopy can reflect the differences between gastric and colorectal cancer sera and normal human serum, and the use of Raman spectroscopy and the GRU network can enable the early screening of gastrointestinal tumors as well as subtype identification.

Data availability

The code supporting the results of this study is available upon request from the corresponding author. These data cannot be made publicly available due to privacy or ethical constraints. The code can be accessed through the GitHub link (**https://github.com/Kunxiang-Liu/Raman-gastric-cancer.git**).

Ethical approval statement

This retrospective study was reviewed and approved by the Institutional Review Board of Sun Yat-Sen University Cancer Center (approval no. B2022-664-01) and The Third Bethune Hospital of Jilin University (approval no. 2023020713), and the requirement to obtain informed written consent was waived.

Author contributions

Kunxiang Liu: formal analysis, writing – original draft, visualization, and implementation. Bo Liu: methodology. Yu Wang: methodology. Qi Zhao: formal analysis. Qinian Wu: writing – review & editing. Bei Li: writing – review & editing.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1 F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre and A. Jemal, *CA Cancer J. Clin.*, 2018, **68**, 394–424.
2 J. Nallala, C. Gobinet, M. D. Diebold, V. Untereiner, O. Bouche, M. Manfait, G. D. Sockalingum and O. Piot, *J. Biomed. Opt.*, 2012, **17**, 116013.
3 P. Laissue, *Mol. Cancer*, 2019, **18**, 5.
4 J. Bjork, *EPMA J.*, 2010, **1**, 513–521.
5 N. Chapelle, M. Martel, E. Toes-Zoutendijk, A. N. Barkun and M. Bardou, *Gut*, 2020, **69**, 2244–2255.
6 A. R. Hatfield, G. Slavin, A. W. Segal and A. J. Levi, *Gut*, 1975, **16**, 884–886.
7 M. Li, H. He, G. Huang, B. Lin, H. Tian, K. Xia, C. Yuan, X. Zhan, Y. Zhang and W. Fu, *Front. Oncol.*, 2021, **11**, 665176.
8 A. S. Haka, Z. Volynskaya, J. A. Gardecki, J. Nazemi, R. Shenk, N. Wang, R. R. Dasari, M. Fitzmaurice and M. S. Feld, *J. Biomed. Opt.*, 2009, **14**, 054023.
9 A. Mizuno, H. Kitajima, K. Kawauchi, S. Muraishi and Y. Ozaki, *J. Raman Spectrosc.*, 1994, **25**, 25–29.
10 L. E. Kamemoto, A. K. Misra, S. K. Sharma, M. T. Goodman, H. Luk, A. C. Dykes and T. Acosta, *Appl. Spectrosc.*, 2010, **64**, 255–261.
11 K. Liu, Q. Zhao, B. Li and X. Zhao, *Front. Bioeng. Biotechnol.*, 2022, **10**, 856591.
12 K. Liu, B. Liu, Y. Zhang, Q. Wu, M. Zhong, L. Shang, Y. Wang, P. Liang, W. Wang, Q. Zhao and B. Li, *Comput. Struct. Biotechnol. J.*, 2023, **21**, 802–811.
13 L. Huang, H. Sun, L. Sun, K. Shi, Y. Chen, X. Ren, Y. Ge, D. Jiang, X. Liu, W. Knoll, Q. Zhang and Y. Wang, *Nat. Commun.*, 2023, **14**, 48.
14 M. S. Bergholt, W. Zheng, K. Lin, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. So and Z. Huang, *Biosens. Bioelectron.*, 2011, **26**, 4104–4110.
15 S. Feng, J. Pan, Y. Wu, D. Lin, Y. Chen, G. Xi, J. Lin and R. Chen, *Sci. China: Life Sci.*, 2011, **54**, 828–834.
16 S. K. Teh, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh and Z. Huang, *J. Biomed. Opt.*, 2008, **13**, 034013.
17 C. S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. E. Saleh, S. Ermon and J. Dionne, *Nat. Commun.*, 2019, **10**, 4927.
18 A. K. Boardman, W. S. Wong, W. R. Premasiri, L. D. Ziegler, J. C. Lee, M. Miljkovic, C. M. Klapperich, A. Sharon and A. F. Sauer-Budge, *Anal. Chem.*, 2016, **88**, 8026–8035.
19 B. Liu, K. Liu, N. Wang, K. Ta, P. Liang, H. Yin and B. Li, *Talanta*, 2022, **244**, 123383.
20 S. Y. Hsu, L. R. Yeh, T. B. Chen, W. C. Du, Y. H. Huang, W. H. Twan, M. C. Lin, Y. H. Hsu, Y. C. Wu and H. Y. Chen, *Molecules*, 2020, **25**, 4792.
21 H. Li, J. Zhou, Y. Zhou, Q. Chen, Y. She, F. Gao, Y. Xu, J. Chen and X. Gao, *Front. Physiol.*, 2021, **12**, 655556.
22 H. Sak, A. Senior and F. Beaufays, arXiv preprint arXiv:1402.1128, 2014, DOI: **10.48550/arXiv.1402.1128**.
23 B. Liu, K. Liu, J. Sun, L. Shang, Q. Yang, X. Chen and B. Li, *J. Biophotonics*, 2023, **16**, e202200270.
24 Y. J. Liu, M. Kyne, C. Wang and X. Y. Yu, *Comput. Struct. Biotechnol. J.*, 2020, **18**, 2920–2930.
25 P. A. Gorry, *Anal. Chem.*, 2002, **62**, 570–573.
26 Z. M. Zhang, S. Chen and Y. Z. Liang, *Analyst*, 2010, **135**, 1138–1146.
27 S. Hochreiter and J. Schmidhuber, *Neural Comput.*, 1997, **9**, 1735–1780.
28 K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734, DOI: **10.3115/v1/D14-1179**.