Diagnosis of Colonic Tissue by Spectral Analysis Technology Combined with Artificial Intelligence

Jun Lu[1,a]
[1]Hebei Vocational University of Industry and Technology
College of Intelligent Manufacturing
Shijiazhuang 050000, China
[a]lj19752019@163.com

* Xufeng Chen[2,b]
[2]Hebei Vocational University of Industry and Technology
College of Intelligent Manufacturing
Shijiazhuang 050000, China
* Corresponding author: [b]xufengchen_gyznzz@163.com

Minchai Hao[3,c]
[3]Hebei Vocational University of Industry and Technology
College of Intelligent Manufacturing
Shijiazhuang 050000, China
[c]452713383@qq.com

***Abstract*— Survival for cancer patients primarily depends on the early detection. FT- IR spectroscopy is a very useful and effective tool for investigating chemical changes of biological tissue at molecular level, which make it possible to diagnose tumor. With the development of artificial intelligence, the diagnostic accuracy of spectral analysis technology has been greatly improved. This study This study proposes a method, as evidenced by the experimental results. A method that colon cancer and colitis samples is classified respectively by pathology analysis and spectral analysis technology combined with artificial intelligence. Spectra of patients' colonic tissues are obtained by a FT-IR spectrometer with an attenuated total reflectance sampler. Meanwhile, histological examination for tissue sample is carried out, whose result could be used as reference. Multiplicative scatter correction is performed to correct the baseline and drifting of spectra. Resampling by half-means is utilized to detect the outlying samples. Principal component analysis is utilized to extract useful information and reduce dimensions. Forward-feed artificial neural network is created and trained performs discriminant analysis of colonic tissue spectra. Cross validation is used to evaluate the method. The result shows overall accuracy is 91.8%, therefore demonstrate that spectral analysis technology combined with artificial intelligence could be used to diagnose colon cancer and obtain a result with comparatively high accuracy.**

Keywords-AI; ANN; colon endoscope; colon cancer; colitis.

I. INTRODUCTION

Cancer is the malignant tumor which does great harms to the health of patients. The survival of patients highly depends on the early and accurate detection of cancer. At least 1/3 of the patients could be cured if diagnosed in the early stage [1, 2]. Colon cancer commonly occurs at the junction of rectum and sigmoid colon. The incidence rate is highest in the 40~50 years old group. Its incidence rate ranks 3rd among gastrointestinal tumors. Meanwhile, Chroniccolon diseases are common, and enteritis may develop into carcinoma if the disease is not early diagnosed and closely monitored. Therefore, colon endoscopy and biopsy analysis are frequently performed in clinical detection. It is important to explore a non-invasive, rapid and accurate method to detect malignancy and enteritis in endoscopic biopsies, which will help determine proper medical treatment [3]. Fourier transform infrared spectroscopy is one of the important methods for qualitative analysis of substances. Its analysis could provide a lot of information about functional groups and help to determine some or all molecular types and structures. Thus, it could indicate the changes of biological tissue at the molecular level. In the process of carcinogenesis, nuclear acids, proteins, carbohydrates, and other biomolecules generate significant changes in their molecular structures [4, 5]. Therefore, it is possible for the spectral analysis technology to become a rapid, non-invasive and convenience method to detect tumors. However, with the development of spectrometer manufacturing technology, the amount of spectral data becomes very large. At the same time, these data contain a lot of useless and redundant information. Maybe, at the very beginning, it is possible for an experienced chemist to tell a spectrum of inflammatory tissue from that of tumor. But it has become impossible in nowadays.

Fortunately, artificial intelligence (AI) has made great progress in these years. AI is a new technology science which researches and develops the theory, method, technology and application system for simulating, extending and expanding human intelligence. It attempts to produce a new kind of response and decision-making method similar to human intelligence. The research in this field includes robot, language recognition, image recognition, natural language processing and expert system, etc. Since the concept of artificial intelligence was put forward in 1949, The theory and technology have become progressively more advanced, with an expanding range of applications. In particular, recent years have seen even greater progress [6-8]. Figure 1 shows the development of AI.

It could be believed that if spectral analysis technology, colonoscopy technology and artificial intelligence can be successfully combined, the colonoscopy can be analyzed noninvasively, quickly, conveniently and in vivo. In this paper, AI technology is used to distinguish the spectrum of inflammation and tumor in colon by perform pretreatment, outlier detection, feature extraction, artificial neural network training and prediction. The result shows this technology has a good prospect.

Doctoral fund of Hebei Polytechnic University (bz202205);
Cemented Carbide Cutting Materials Application Technology R & D Center

979-8-3315-0976-7/25/$31.00 ©2025 IEEE

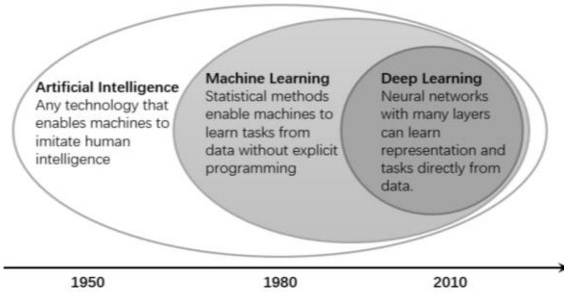
Artificial Intelligence
Any technology that enables machines to imitate human intelligence
Machine Learning
Statistical methods enable machines to learn tasks from data without explicit programming
Deep Learning
Neural networks with many layers can learn representation and tasks directly from data.
1950
1980
2010

Figure 1. Development Process of AI.

II. Methods

A. Patients and materials

Each patient is well informed about this study and all patients have authorized the using of their information and tissue for this study. A total of 91 colon biopsies were collected based on clinical diagnoses, including 44 female and 67 male patients, ranging in age from 22 to 77 years (mean age 56.2 years). Each patient underwent a about 2 mm endoscopic pinch biopsy from the colon. Pathological analysis revealed that the sample set consisted of 42 cases of colitis and 49 cases of cancer.

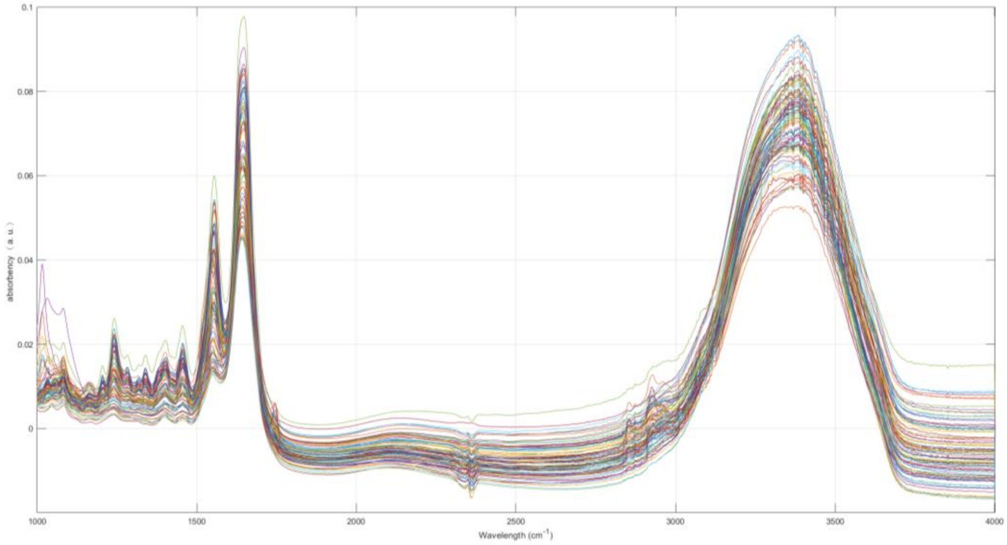

Figure 2. Pretreated Spectra of Patients' Colon.

B. Spectral measurement

Spectra of colonic tissues are measured by a NICOLET FT-IR spectrometer, which manufactured by Thermo Fisher Scientific Inc. The spectrometer is fitted with a mercury cadmium telluride detector, cooled by liquid nitrogen, to enhance both stability and sensitivity. Attenuated total reflectance (ATR) sampler is utilized during measurement to enhanced signal intensity. Attenuated Total Reflectance is a sampling technique paired with IR spectroscopy that allows for direct examination of solid samples without the need for additional preparation. ATR takes advantage of total internal reflection, which generates an evanescent wave. In this process, an infrared light beam is directed through the ATR crystal, reflecting off the internal surface in contact with the sample at least once. This reflection creates the evanescent wave, which extends into the sample. Thus, the signal quality is improved. Specimens are scanned with 32 times coadded at the resolution of 4 cm^{-1} . The scan covers a range from 680 cm^{-1} to 4000 cm^{-1}. It took about two minutes to obtain the spectrum non-invasively. The study lasts for months, the experiments are carried under different measuring condition and spectrometer state. Due to severe drift at the beginning of the spectrum, spectral lines before 1000 cm^{-1} were removed. The baseline, drifting and noise are inevitable. Thus, pretreatment is necessary. Initially, smoothing is applied to filter out high-frequency noise. The CO_2 absorption band and spectral ranges containing little useful information for discriminant analysis are then excluded. Following this, Multiplicative Scatter Correction (MSC) is used to correct for baseline shifts and drift. MSC is a method designed to compensate for both additive and multiplicative effects in spectral data. Originally developed for addressing multiplicative scattering in reflectance spectroscopy, MSC has also proven effective for correcting other issues, such as variations in path length, offset shifts, and interference. The pretreated spectra of patients' tissue are show in Figure 2.

After the spectra of the samples are recorded, the samples are stored in liquid nitrogen and sent for the histological examination as the reference.

Once the spectra of the samples are obtained and saved, they are sent storing in liquid nitrogen for pathological analysis examination and the results would be regarded as the reference.

C. Outlier detection

Developing high-quality multivariate calibration models relies on the completion of several key steps. One crucial step in the calibration process is the detection of outliers because the quality of a calibration model is determined by that of the calibration data. Outlier samples are typically defined as those who deviate significantly from the majority of the data. They may represent variability in the measurement, experimental errors, or indicate something novel. In our study, re-sampling by half-means (RHM) is utilized to detect the outliers in the spectra data set. RHM algorithm based on the estimation of the distribution of spectra by re-sampling. Half of the spectral data set is sampled without replacement for several times [9, 10]. The number of re-sampling experiments ought to be at least double the number of samples. For each re-sampled spectral set, the mean and standard deviation are calculated. Subsequently, all spectra are scaled using the computed mean and standard deviation. Following this, the distances from each spectrum to the mean spectrum are determined. In each re-sampling experiment, a fixed percentage (such as 5%) of the spectra with the farthest distance to the mean spectrum are recorded and score one point. After all the resampling experiments, the spectra with scores extremely higher than others should be considered as outliers and be eliminated.

D. Calibration and Deep Learning

Calibration is the process of train a correlation model between pathological analysis results and spectral data. Then one new unclassified spectrum is obtained, it is imported into the trained model to analyze its properties. The whole process is show in Figure 3.

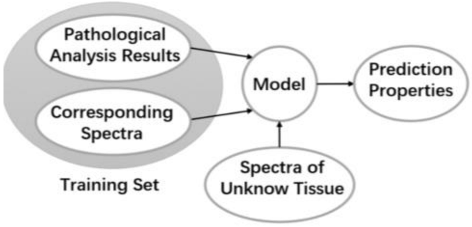

Figure 3. Process of Calibration and Prediction.

Machine learning (ML) employs computational techniques to acquire information directly from data, bypassing the need for predefined equations as models. Deep learning represents a specialized subset of machine learning., which learn from experience just as what comes naturally to humans. In this paper, A feed forward neural network (FNN) is utilized as an instance of deep learning to classify the spectra of colon tissue. FNN is a type of neural network characterized by its straightforward topological structure, where neurons are systematically arranged in layered formations. Thus, its speed is very fast, and it is very suitable for processing spectral cases with large amount of data[11]. Each neuron in a Feedforward Neural Network is exclusively connected to the neurons of the preceding layer, receiving inputs from the former and transmitting outputs to the latter. The absence of feedback loops between layers distinguishes this architecture. As one of the most prevalent and rapidly evolving artificial neural networks, Feedforward Neural Networks have witnessed substantial progress in both theoretical research and practical implementations, currently embodying a high degree of sophistication. FNNs employ a unidirectional, multilayer structure, where each layer comprises multiple neurons. In this neural network paradigm, each neuron receives signals from the preceding layer and generates outputs for the subsequent layer. Specifically, the initial layer is designated as the input layer, the final layer as the output layer, and any intermediary layers are referred to as hidden layers (plural form also applies). The hidden layers can consist of a single layer or multiple layers. Notably, there is no feedback within the entire network; signals propagate unidirectionally from the input layer A typical multilayer feedforward neural network is shown in Figure 4.

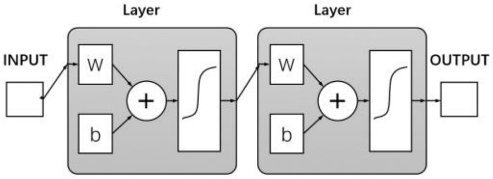

Figure 4. Typical Structure of FNN.

III. RESULTS

When pretreatment is completed, RHM is applied on the spectral data set. 500 resample experiments are carried out. The RHM scores of sample 3, 9, 14, 22, 25, 46 are extremely high compare to other sample. Their scores are show in Table 1. Other samples' RHM scores are all less than 30 points. Therefore, it is speculated that these samples are seriously affected by the unstable measurement environment and spectrometer state during the spectral measuring. Then, there 4 samples are determined as singular values and eliminated.

TABLE I. HIGH RHM SCORES OF OUTLIERS

RHM Scores of Outliers Samples						
Sample No.	3	9	14	22	25	46
RHM Scores	388	274	500	300	281	500

The rest 85 samples include 38 cases of colitis and 47 cases of cancer. PCA is applied on this data set to perform feature extraction and dimension reduction. Figure 5 shows the information ratio explained by principal components (PCs).

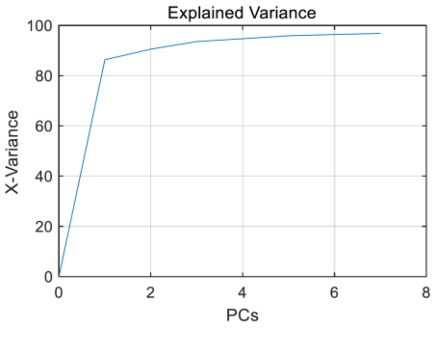

Figure 5. information ratio explained by PCs.

The first five variables explain more than 95% information. Thus, reserve the first five variables for next step and the rest PCs are eliminated. A two-layer feed-forward network is created for classification. The network has 18 neurons in its hidden layer for a given matrix of input vectors of 5 PCs. Leave 1 out cross validation is used to evaluate the discrimination results. Take one of these training processes for instance, the convergence process is shown in Figure 6.

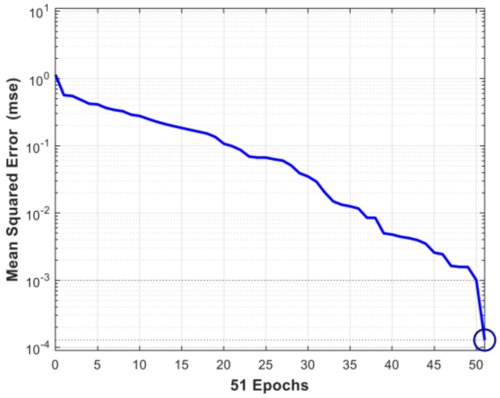

Figure 6. Convergence process of training network.

It could be seen that the training of this network is very fast. The training error converge in 51st epoch and takes less than 1 second. Table 2 shows the deep learning discrimination results of leave-one-out cross validation.

TABLE II. COMPARISON OF DEEP LEARNING RESULTS WITH PATHOLOGICAL ANALYSIS

Deep Learning **Pathology**	**Colon cancer**	**Colitis**
Colon cancer	43	4
Colitis	3	35

Among the 38 cases of colitis samples, 35 cases are correctly distinguished while 3 samples are mis-judged; among the 47 cases of cancer samples, 43 cases are well-judged while 4 cases is mis-judged. Pathological analysis result is taken as the true value and calculate the statistical results of colon biopsies diagnosis using Deep learning are show in table 3.And the total correctness is (43+35)/(43+35+3+4)=91.8%.

TABLE III. STATISTICAL RESULTS OF DISCRIMINATION RESULTS

Item	**Value**
Positive predict value	43/(43+3)=93.5%
Negativ predict value	35/(35+4)=89.7%
Sensitivity	43/(43+4)=91.5%
Specificity	35/(35+3)=92.1%

IV. CONCLUSIONS

According our previous research, different spectral characteristics could be found between colitis and malignant colon samples. These spectral characteristics are intimately linked to alterations in the structure and composition of biological molecules within tissue cells. Specifically, the C=O band observed near 1743 cm^{-1} is indicative of fat content in tissues. Additionally, the C-H stretching vibration bands situated at approximately 2966 cm^{-1}, 2927 cm^{-1}, and 858 cm^{-1} are associated with lipid and fat levels. Notably, these bands tend to diminish or even disappear in the spectra of malignant tissues, due to the consumption of fat in such tissues to meet the heightened nutritional and energy demands that accompany carcinoma development. The absorption peak at ~1643 cm^{-1} corresponds to both the amide I band of proteins and the H-O-H deformation vibration of water. Meanwhile, the absorption peak at ~1550 cm^{-1} is attributed to the amide II band of proteins. In the spectra of malignant colon tissues, the relative intensity of the amide II band compared to the ~1643 cm^{-1} absorption peak is reduced compared to that observed in colitis biopsies. Furthermore, in the spectra of cancerous samples, the intensity of the peak at ~1460 cm^{-1} is weaker than that of the peak at ~1400 cm^{-1}.

The peak at ~1 460 cm^{-1} is stronger.

In summary, the results indicate that colon cancer could be distinguished from colitis with high accuracy using FT-IR spectroscopy combined with artificial intelligence. FT-IR spectroscopy of biopsy samples taken at the time of endoscopy can provide an accurate, immediate diagnostic differentiation between enteritis and colon cancer. But for practical application, there are still some limitations. With the development of endoscope and optical fiber manufacturing technology, it could be expected that spectral analysis technology combined with artificial intelligence would become a real-time, non-invasive, accurate and in vivo diagnosis. Also, this technology would be practical for immediate diagnosis at endoscopy.

REFERENCES

[1] Beata Brozek-Pluska, Adam Dziki, Halina Abramczyk. Virtual spectral histopathology of colon cancer - biomedical applications of Raman spectroscopy and imaging, Journal of Molecular Liquids, 2020 (303): 112676.

[2] Rui Yang, Gang Zheng, Defa Ren, Chunzhou Chen, Cheng Zeng, Wei Lu, Hua Li. The clinical significance and biological function of tropomyosin in colon cancer. iomedicine & Pharmacotherapy. 2018 (101):1-7.

[3] Ng Kheng-Seong,West Nicholas P,Scott Nigel,Holzgang Melanie,Quirke Phil, Jayne David G. What factors determine specimen quality in colon cancer surgery A cohort study.[J]. International journal of colorectal disease, 2020,35(5).

[4] C. Conti, P. Ferraris, E. Giorgini, C. Rubini, S. Sabbatini, G. Tosi, J. Anastassopoulou, P. Arapantoni, E. Boukaki, S. Konstadoudakis, T. Theophanides, C. Valavanis. FT-IR microimaging spectroscopy: A comparison between healthy and neoplastic human colon tissues. Journal of Molecular Structure. 2008 (881): 46-51.

[5] Jie Liu, Hong Cheng, Xiaoyi Lv, Zhaoxia Zhang, Xiangxiang Zheng, Guohua Wu, Jun Tang, Xiaorong Ma, Xiaxia Yue. Use of FT-IR spectroscopy combined with SVM as a screening tool to identify invasive ductal carcinoma in breast cancer. Optik. 2020 (204):164225.

[6] Filipp Fabian V. Opportunities for Artificial Intelligence in Advancing Precision Medicine.[J]. Current genetic medicine reports, 2019,7(4): 208-213.

[7] Hironobu Fujiyoshi, Tsubasa Hirakawa, Takayoshi Yamashita. Deep learning-based image recognition for autonomous driving. IATSS Research. 2019, 43 (4): 244-252.

[8] Leijin Xiong, Dingli Zhang, Yu Zhang. Water Leakage Image Recognition of Shield Tunnel via Learning Deep Feature Representation. Journal of Visual Communication and Image Representation. 2019: 102708.

[9] Leo H. Chiang, Randy J. Pell, Mary Beth Seasholtz. Exploring process data with the use of robust outlier detection algorithms. Journal of Process Control, 2003(13), 5: 437-449.

[10] Yan Wang, Qingfen Liu, Han-dan Hou, Seungmin Rho, Brij Gupta, Ying-xin Mu, Wei-zheng Shen, Big data driven outlier detection for soybean straw near infrared spectroscopy. Journal of Computational Science. 2018 (26): 178-189.

[11] H.-C. Shin, M. Orton, D.J. Collins, S. Doran, M.O. Leach. Chapter 7 - Organ Detection Using Deep Learning. Medical Image Recognition, Segmentation and Parsing, Academic Press, 2016: 123- 153.