Supplementary Material for

Label-free Diagnosis of Lung Cancer by Fourier Transform Infrared Microspectroscopy Coupled with Domain Adversarial Learning

Yudong Tian,[1, †] Xiangyu Zhao,[1, †] Jingzhu Shao,[1] Bingsen Xue,[1] Lianting Huang,[1] Yani Kang,[1] Hanyue Li,[2] Gang Liu,[2] Haitang Yang,[2, *] and Chongzhao Wu[1, *]

[1]Center for Biophotonics, Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China

[2] Department of Thoracic Surgery, Shanghai Chest Hospital, Shanghai Jiao Tong University; Shanghai, China

[†]These authors contributed equally.

** Corresponding author: haitang.yang@shsmu.edu.cn (haitang Yang), czwu@sjtu.edu.cn (Chongzhao Wu)*

Supplementary Material 1: Fourier Transform Infrared Spectroscopy Theory

Infrared spectroscopy is based on the principle of molecular vibration and rotational energy level transitions. When infrared light irradiates a sample, the molecules in the sample absorb specific wavelengths of infrared light, causing changes in the molecular vibration or rotational energy levels. Different chemical bonds and functional groups have specific vibrational frequencies, and thus absorb different wavelengths of infrared light. By measuring the absorption of infrared light by the sample, the infrared spectrum of the sample can be obtained, allowing for the analysis of the molecular structure of the sample.

Fourier Transform Infrared Spectroscopy (FTIR) [1] is an efficient infrared spectroscopy technique. The core of FTIR lies in using the Michelson interferometer to generate an interferogram, which is then converted from the time domain to the frequency domain through Fourier transform, thereby obtaining an infrared spectrum. The Fourier transform equation is defined as:

$$I(v) = \int_{-\infty}^{\infty} I(x) \cdot e^{-i2\pi vx} dx \#(S1)$$

where $I(v)$ is the light intensity at frequency *v*, $I(x)$ is the measured interference signal when the optical path difference is *x*.

Supplementary Material 2: Digital Deparaffinizing

The process of digital deparaffinizing is accomplished by the extended multiplicative signal correction (EMSC) algorithm. The EMSC algorithm is based on the principle of multivariate statistics and achieves correction and standardization of spectral data by modeling and correcting the difference between the sample and reference spectra. The quantitative measurement of vibrational absorption spectra is based on the Lambert-Beer Law, according to which the absorption spectrum is proportional to the effective optical range length, and for transparent samples containing a single light-absorbing chemical, the absorbance $A(\tilde{v})$ can be given by the following equation:

$$A(\tilde{v}) = k(\tilde{v}) \times c \times b \#(S2)$$

where $k(\tilde{v})$ is the characteristic absorbance of a particular component at a particular wavenumber, b is the optical range length, and c is the concentration of the absorbing chemical in the sample. The spectral change caused by a change in optical range length is usually expressed as a “multiplicative” change.

Since biochemical samples are usually very complex in composition, there is a considerable part of the overlap of the absorption characteristics in many different bio-molecular components of the spectra, at this point, Lambert-Beer law can be written as a superposition of the absorbance of several absorbing substances:

$$A(\tilde{v}) = \left(\sum_{i=1}^{n} \left(c_i \times k_i(\tilde{v}) \right) \right) \times b \#(S3)$$

where $k_i(\tilde{v})$ is the single component absorption spectrum, c_i is the concentration of component i, and b is the optical range. Here we assume that the optical range length b is comparable for all components, an assumption that is usually applied to sufficiently homogeneous samples.

Although the concentrations of the various biochemical components of biological tissue are usually

unknown, the overall shape of the IR absorption spectra obtained from biological samples is generally very similar, which means that we can approximate the individual spectra by the average spectrum of the sample as a whole, and thus we can use the average of all the spectra of the sample plus a small deviation to represent the measured absorption spectra $k(\tilde{v})$:

$$k_i(\tilde{v}) = x(\tilde{v}) + \Delta k_i(\tilde{v}) \#(S4)$$

where $x(\tilde{v})$ denotes the average spectra of all the spectra of the tissue sample and $\Delta k_i(\tilde{v})$ denotes the deviation of the individual absorption spectra from the average spectrum. Combining Eq. (S3) with Eq. (S4) we can get:

$$A(\tilde{v}) = \left(\sum_{i=1}^{n} (c_i \times x(\tilde{v}) + \sum_{i=1}^{n} \left(c_i \times \Delta k_i(\tilde{v}) \right) \right) \times b \#(S5)$$

Also, after normalization, the sum of the concentrations of all substances should be 1, so that we can get the final form of the absorption model for Lambert-Beer law:

$$A(\tilde{v}) = \left(x(\tilde{v}) + \sum_{i=1}^{n} \left(c_i \times \Delta k_i(\tilde{v}) \right) \right) \times b \#(S6)$$

To show that a single absorption spectrum is similar to an average spectrum, we can use the following form instead of the above equation:

$$A(\tilde{v}) = x(\tilde{v}) \times b + e(\tilde{v}) \#(S7)$$

where the residual part $e(\tilde{v}) = \sum_{i=1}^{n} (c_i \times \Delta k_i(\tilde{v}))$.

The multiplicative signal correction (MSC) model is an extension of the Lambert-Beer model shown in Eq. (S7), which represents the absorption spectrum as a linear sum of a constant baseline a and Eq. (S7). The basic MSC model is as follows:

$$A(\tilde{v}) = a + bx(\tilde{v}) + e(\tilde{v}) \#(S8)$$

The unknown parameters in Eq. (S8) can be estimated by least squares regression. To estimate the parameters a and b, the spectra are corrected as follows:

$$A_{corr}(\tilde{v}) = (A(\tilde{v}) - a)/b \#(S9)$$

The above equation approximates each spectrum by averaging the spectrum and a constant baseline, but in biological samples, baseline variations in the spectra cannot generally be represented by a constant straight line. EMSC is an improvement on the MSC method by expanding the baseline with a baseline of arbitrary slope, a quadratic term, or a term of higher polynomial order, so that nonlinear effects present in the spectra are included in the calculations, and the EMSC model equation is formulated as follows:

$$A(\tilde{v}) = a + x(\tilde{v}) \times b + d_1 \times \tilde{v} + d_2 \times \tilde{v}^2 + \cdots + d_p \times \tilde{v}^p + e(\tilde{v}) \#(S10)$$

Based on Eq. (S10), we add the absorption spectra of undesired contaminants in the sample to obtain the final spectral model:

$$A(\tilde{v}) = a + x(\tilde{v}) \times b + c \sum_{j=1}^{m} k_j(\tilde{v}) \times c_j + d_1 \times \tilde{v} + d_2 \times \tilde{v}^2 + \cdots + d_p \times \tilde{v}^p + e(\tilde{v}) \#(S11)$$

Similarly, we can estimate the unknown parameters in Eq. (S11) by least squares regression, and the final corrected spectrum is:

$$A_{corr}(\tilde{v}) = \frac{A(\tilde{v}) - a - d_1 \times \tilde{v} - d_2 \times \tilde{v}^2 - \cdots - d_p \times \tilde{v}^p - c\sum_{j=1}^{m} k_j(\tilde{v}) \times c_j}{b} \#(S12)$$

In this way, we can exclude the effects of baseline drift and contaminants such as paraffin on the spectra. At the same time, by fitting the obtained coefficients we can exclude the background spectra mixed in the sample and avoid their influence on the subsequent modeling.

In this work, we chose a baseline polynomial fit with a term number of p=4 and employed spectra from the regions of pure paraffin for building the spectral contaminant model k_j. The principal components analysis (PCA) was used to decompose the models of paraffin. Fig. S1(a) and Fig. S1(b) present the spectral dataset obtained from the pure paraffin region and the paraffin model (principal component) derived using PCA decomposition [2]. Component 1 carries more information of wavenumbers greater than 1500 cm^{-1}, while component 2 carries more information of wavenumbers which are lower than 1500 cm^{-1}. The spectra before and after digital deparaffinizing are illustrated in Fig. S1(c) and Fig. S1(d), respectively. The results indicate that the paraffin bands near 1473, 1462, and 1373 cm^{-1} in the spectra, along with other spectral interferences, are mostly eliminated.

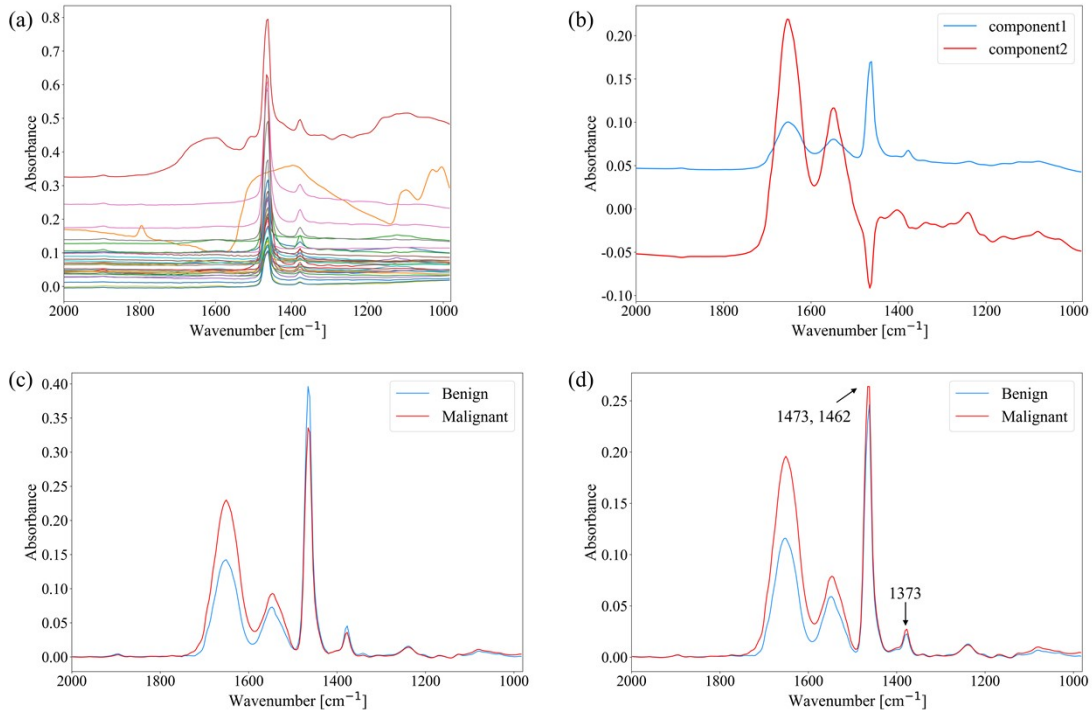

Fig. S1. Results of paraffin modeling and digital deparaffinizing. (a) The absorption spectra from the pure paraffin region; (b) Paraffin models calculated with PCA; (c) The average spectra plotted from raw data; (d) The average spectra plotted from the digitally deparaffinizing data where the paraffin peaks at 1473 cm^{-1}, 1462 cm^{-1} and 1373 cm^{-1} were mostly neutralized.

Supplementary Material 3: IRS-DANN Model Development

The objective functions for the label classification part and domain identification part of the model are respectively the cross-entropy loss function and focal loss. The hyperparameters α and γ of the focal loss are 0.6 and 2 respectively. Minimization of the objective function was achieved by the Adam optimizer, with the initial learning rate to be 1×10^{-3}, and the weight decay to be 1×10^{-6}.Both the label classifier and domain discriminator are a three-layer MLP with ReLU function and LayerNorm. All layers in the ISDANN were initialized randomly. In this study, we chose three structures of CNN,

Bi-LSTM, and Transformer backbone for the IRS-DANN model.

A. CNN-based Encoder

The backbone of the CNN structure is a Resnet network [3] containing 8 residual blocks. In this study, the first step is to map the FTIR spectrum into a one-dimensional array containing 782 points. Second, after a convolution of 64 kernels with a size of 7×1 and a 3×1 maximum pooling with a step size of 2, the dimensionality of the feature map is reduced to 391. Third, 8 residual modules are used for feature extraction (R1-R8 are used here to represent), where R1 and R2 have 64 convolution kernels of size 3 × 1, R3 and R4 have 128 convolution kernels of size 3 × 1, R5 and R6 have 256 convolution kernels of size 3 × 1, R7 and R8 have 512 convolution kernels of size 3 × 1, R3, R5 and R7 uses 2× stride to reduce the size of feature map. The feature map is finally transformed into a feature vector of length 512 by an adaptive mean pooling and flattening operation.

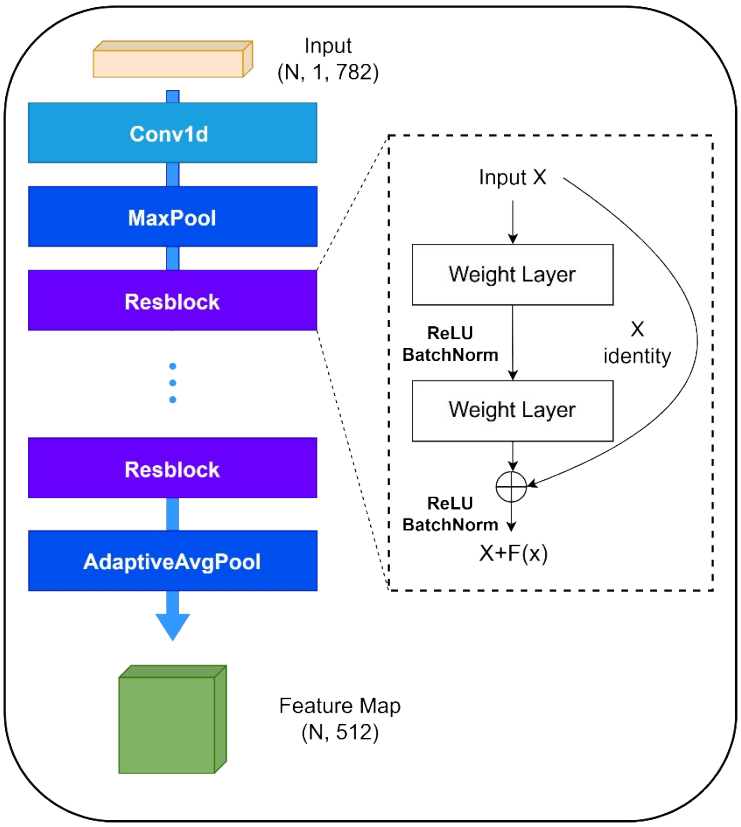

Fig. S2. The network structure of CNN-based backbone.

B. Bi-LSTM-based Encoder

As shown in Fig. S3, the Bi-LSTM-based encoder consists of two convolution layers and a Bi-LSTM layer. First, the FTIR spectrum is mapped into a one-dimensional vector with a length of 782. Second, a convolutional layer is then used to expand the number of channels of the input spectrum to 64, where the size of the convolution kernel is 3 and the size of the padding is 1. Third, the feature maps output from the convolutional layer are transposed and fed into a one-layer bidirectional LSTM network. The hidden layer size of the Bi-LSTM neural network is 100. The output of the BILSTM network is then transposed and the number of channels is reduced to 1 by a convolution with kernel size of 3 and padding size of 1.

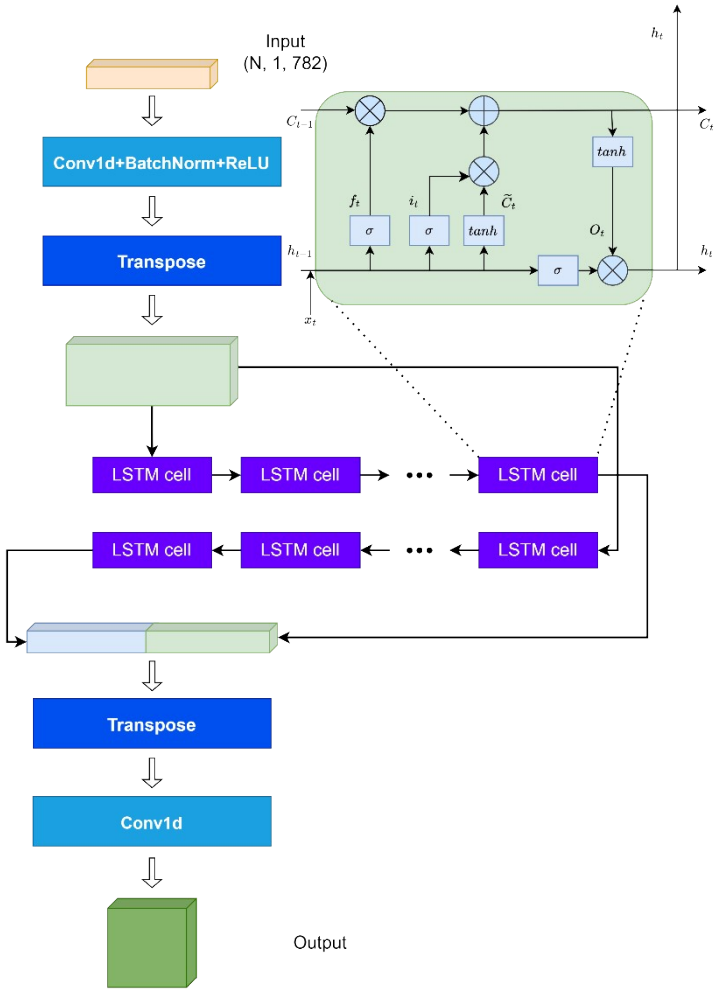

Fig. S3. The network structure of Bi-LSTM-based backbone.

C. Transformer-based Encoder

As shown in Fig. S4, the main component of the Transformer-based encoder is a 4-layer transformer model with residual structures [4]. Transformer is a deep learning model architecture suitable for sequence-to-sequence tasks, and its key structure lies in the multi-head self-attention mechanism, where multiple self-attention layers are stacked and integrated. Fig. S6 illustrates the process of the self-attention module in transformers. The Conv Fusion module is a residual structure for connecting the outputs of different transformer layers, and the corresponding process is shown in Fig. S5. Essentially, it is a 2d convolution, in which the outputs of different layers are stacked and then convolved into the residual output with the number of channels minus 1 and the size remains unchanged. In the transformer-based encoder, the FTIR spectrum is first expanded to 64 channels by a convolutional block with a convolutional kernel size of 3 and a padding size of 1. After transposition, the convolved spectrum is embedded in a fully connected layer with the number of units in the fully connected layer being 128. The embedded spectrum is then fed into the residual transformer model for feature encoding. The encoded output is passed through an average pooling layer to obtain the final feature vector used for classification.

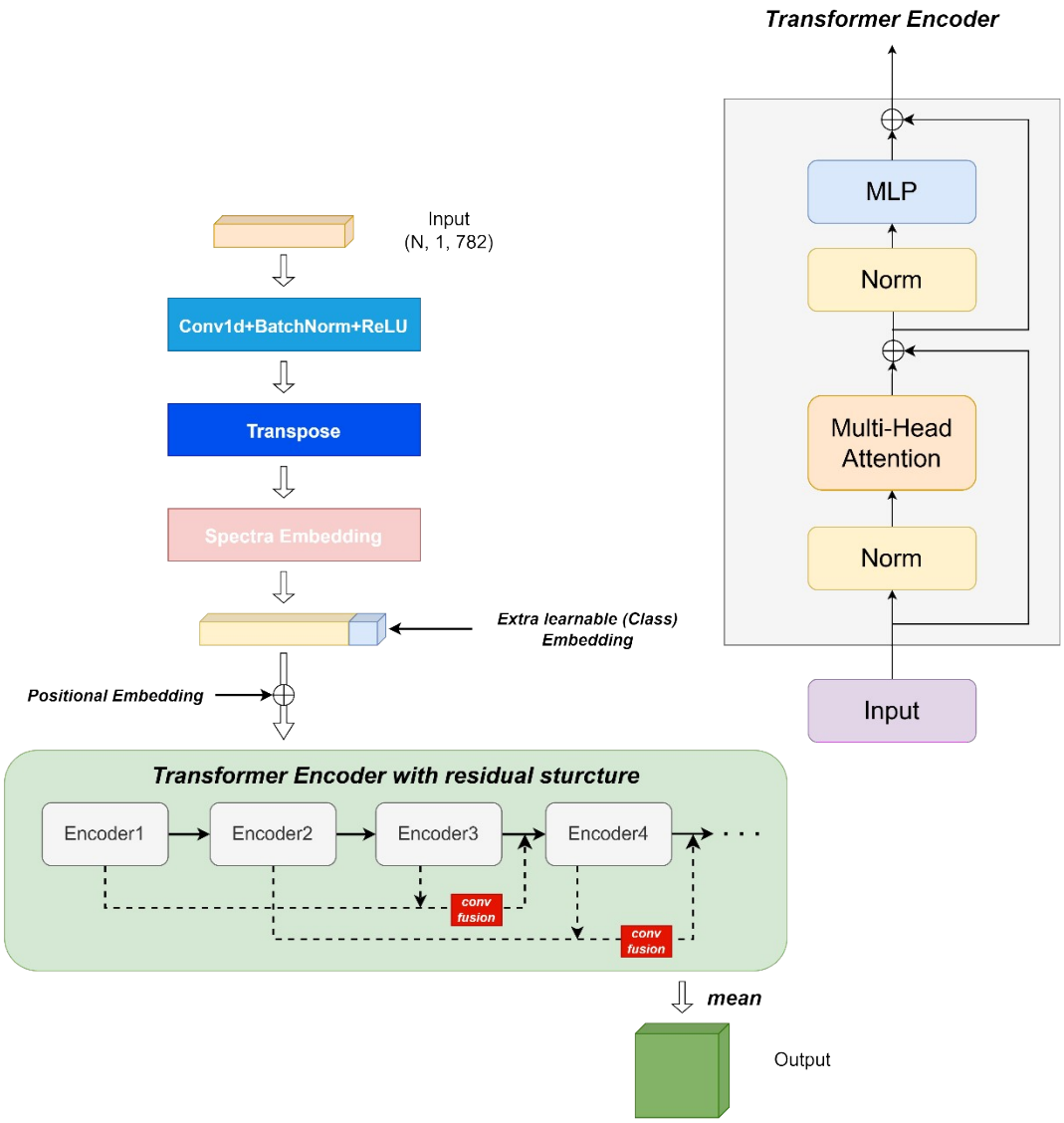

Fig. S4. The network structure of Transformer-based backbone.

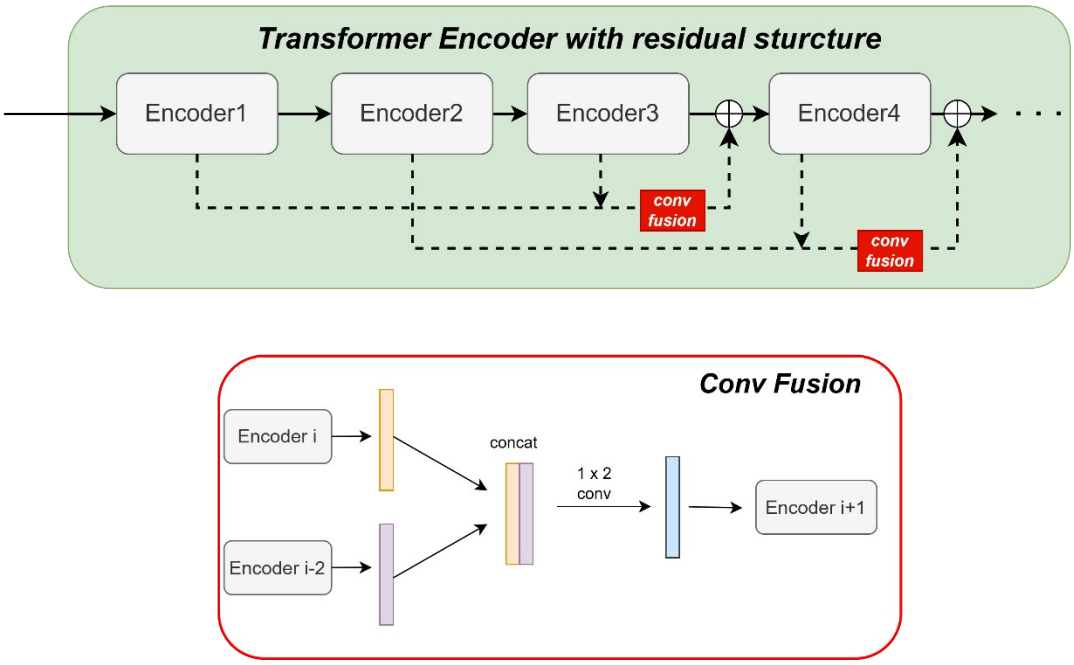

Fig. S5. Diagram of the convolution fusion module in the Transformer backbone.

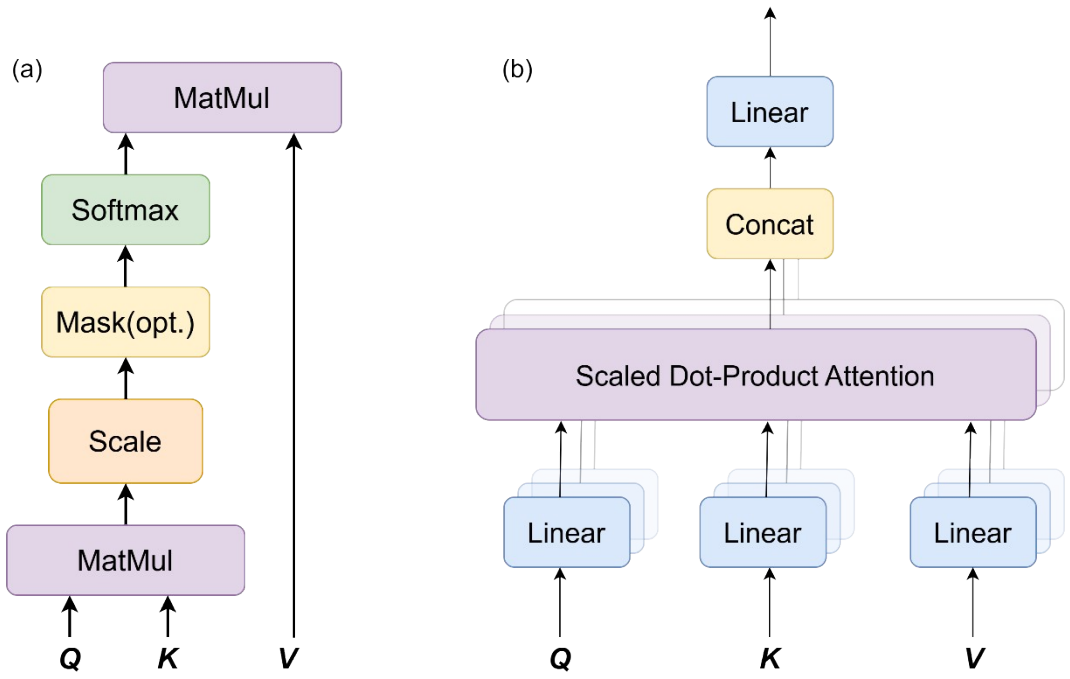

Fig. S6. Illustration of the attention mechanism in transformers. (a) Self-attention module. (b) Multi-head attention mechanism.

Supplementary Material 4: Grad-weighted Class Activation Mapping

The Grad-weighted class activation mapping (Grad-CAM) technique can be leveraged to measure the importance of different spectral locations for the prediction. Based on the improvement of the original CAM, Grad-CAM obtains the weights by solving for the bias derivative of the category confidence of the network output on the feature map. The steps to generate our spectral Grad-CAM are

as follows:

1. For a given spectral input into the network, forward propagation is performed to obtain the feature map A^k of the last convolutional layer, k represents the channel index.
2. Perform backpropagation to obtain the gradient $\frac{\partial y^c}{\partial A^k}$ of the probability y^c of the category c of the network output with respect to A^k.
3. Calculate weights based on gradient: $\alpha_k^c = \frac{1}{N}\sum_i\sum_j \frac{\partial y^c}{\partial A_{i,j}^k}$.
4. Calculate the Grad-CAM: $G_{Grad-CAM}^{\ c} = interpolate(ReLU(\sum_k \alpha_k^c A^k))$.

Supplementary Material 5: T-distributed Stochastic Neighbor Embedding

T-Distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction algorithm, primarily used to map high-dimensional data into a low-dimensional space for visualization and analysis. Its core idea is to maintain the local similarity of data points in both high-dimensional and low-dimensional spaces, while minimizing the distortion of global structure.

In high-dimensional space, t-SNE measures the similarity by calculating the conditional probability between data points. The similarity between data points x_i and x_j is defined as:

$$P_{j|i} = \frac{\exp\left(-\| x_i - x_j \|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} exp\left(-\| x_i - x_k \|^2 / 2\sigma_i^2\right)} \#(S13)$$

where σ_i, which is controlled by the perplexity parameter, is the bandwidth of the Gaussian kernel for the data point x_i and $P_{j|i}$ represents the probability of point x_j being a neighbor of point x_i.

To simplify the calculation, t-SNE symmetrizes the conditional probabilities into joint probabilities to compute the similarity distribution P_{ij}:

$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2n} \#(S14)$$

where *n* is the total number of data points.

In the low-dimensional space, t-SNE uses the t-distribution to measure the similarity between data points. For low-dimensional points y_i and y_j, their similarity can be defined as:

$$Q_{ij} = \frac{\left(1 + \| y_i - y_j \|^2\right)^{-1}}{\sum_{k \neq l}\left(1 + \| y_k - y_l \|^2\right)^{-1}} \#（S15）$$

The goal of t-SNE is to make the similarity distribution P in the high-dimensional space and the similarity distribution Q in the low-dimensional space as consistent as possible. The difference between these two distributions can measured using the Kullback-Leibler (KL) divergence:

$$KL(P \| Q) = \sum_{i \neq j} P_{ij} log \frac{P_{ij}}{Q_{ij}} \#(S16)$$

By employing the gradient descent method to optimize the positions of points in the low-

dimensional space and gradually reduce the KL divergence, the final projection of high-dimensional data into the low-dimensional space can be obtained.

Supplementary Material 6: T-SNE Visualization of Benchmark Models

Fig. S7 shows the 2D t-SNE visualization of the output features of the benchmark models' encoder. It can be seen that the data still exhibits a tendency to cluster according to the source domain, with some overlap observed among samples of different categories.

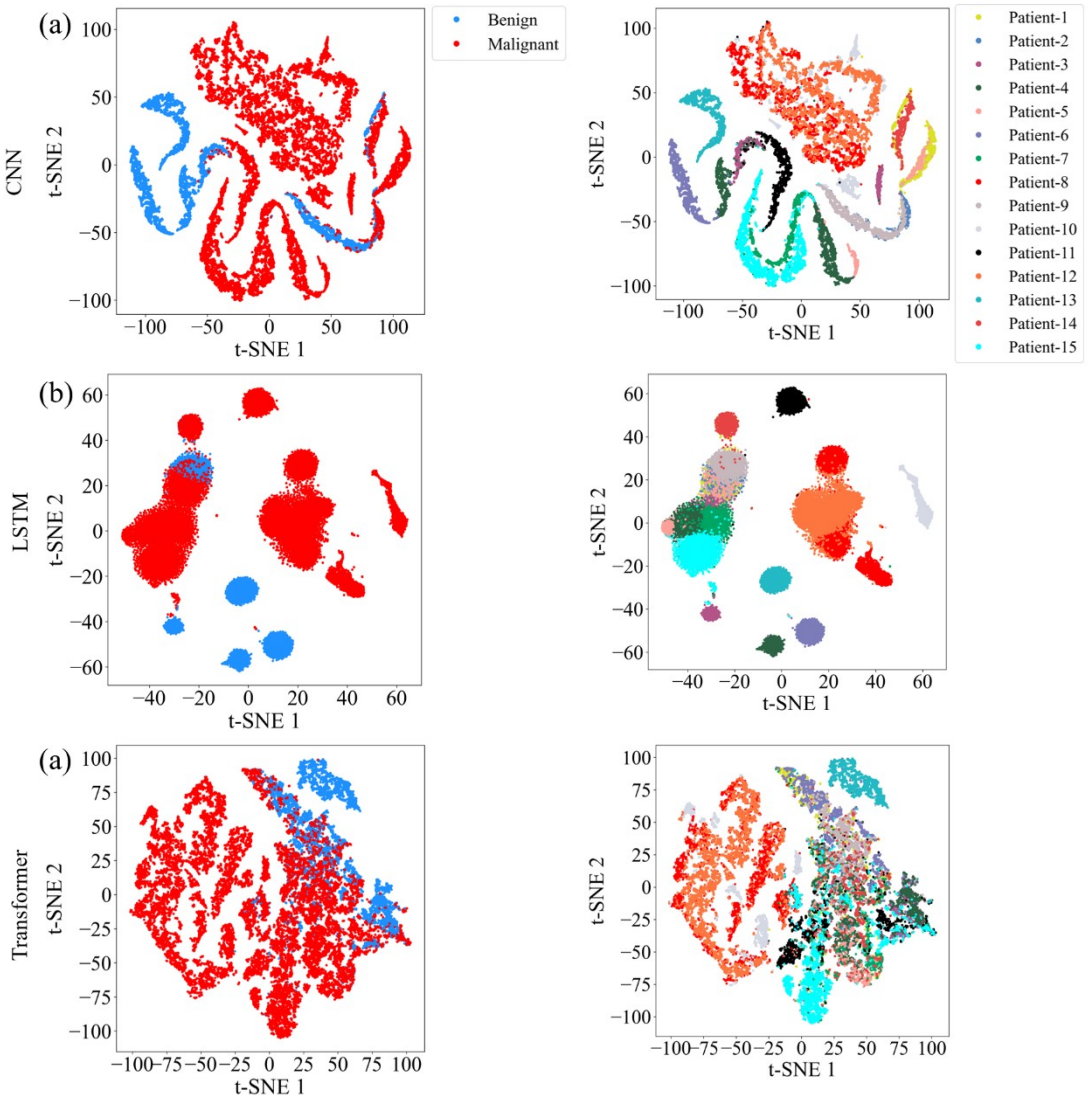

Fig. S7. 2D visualization of the output features of the benchmark model's encoder using t-SNE. Scatter plots are labeled for different categories (left) and patients (right). Each row from top to bottom represents the CNN model, the LSTM model, and the Transformer model respectively.

Supplementary Table 1:

Table S1. FTIR Spectra Dataset Information

Patients Index	Sample diagnosis	Number of points	Histopathologic classification
1	Malignant	448	invasive adenocarcinoma
2	Malignant	491	invasive adenocarcinoma

3	Malignant	1126	invasive adenocarcinoma
4	Malignant	4500	invasive adenocarcinoma
5	Malignant	2642	adenocarcinoma in situ
6	Malignant	1500	microinvasive adenocarcinoma
7	Malignant	1500	invasive adenocarcinoma
8	Malignant	666	invasive adenocarcinoma
9	Malignant	4500	invasive adenocarcinoma
10	Benign	1412	invasive adenocarcinoma
11	Benign	1500	chronic granulomatous inflammation
12	Benign	1500	chronic granulomatous inflammation
13	Malignant	840	invasive adenocarcinoma
	Benign	449	
14	Malignant	392	invasive adenocarcinoma
	Benign	702	
15	Malignant	1406	invasive adenocarcinoma
	Benign	900	

References

1. A. Dutta, "Chapter 4 - Fourier Transform Infrared Spectroscopy," in *Spectroscopic Methods for Nanomaterials Characterization*, S. Thomas, R. Thomas, A. K. Zachariah, and R. K. Mishra, eds., Micro and Nano Technologies (Elsevier, 2017), pp. 73–93.
2. F. A. de Lima, C. Gobinet, G. Sockalingum, S. B. Garcia, M. Manfait, V. Untereiner, O. Piot, and L. Bachmann, "Digital de-waxing on FTIR images," Analyst **142**, 1358–1370 (2017).
3. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.
4. D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "SpectralFormer: Rethinking Hyperspectral Image Classification With Transformers," IEEE Trans. Geosci. Remote Sensing **60**, 1–15 (2022).