

## Classification of Midinfrared Spectra of Colon Cancer Tissue Using a Convolutional Neural Network

In Gyoung Kim<sup>1,2,3</sup>, Changho Lee<sup>2,4</sup>, Hyeon Sik Kim<sup>1</sup>, Sung Chul Lim<sup>5\*</sup>, and Jae Sung Ahn<sup>1\*\*</sup>

<sup>1</sup>Medical & Bio Photonics Research Center, Korea Photonics Technology Institute, Gwangju 61007, Korea

<sup>2</sup>School of Dentistry, Chonnam National University, Gwangju 61186, Korea

<sup>3</sup>Department of Mechanical Engineering, Yonsei University, Seoul 03722, Korea

<sup>4</sup>Department of Nuclear Medicine, Chonnam National University Medical School & Hwasun Hospital, Hwasun 58128, Korea

<sup>5</sup>Department of Pathology, Chosun University Hospital, Gwangju 61453, Korea

(Received September 16, 2021 : revised November 16, 2021 : accepted November 17, 2021)

The development of midinfrared (mid-IR) quantum cascade lasers (QCLs) has enabled rapid high-contrast measurement of the mid-IR spectra of biological tissues. Several studies have compared the differences between the mid-IR spectra of colon cancer and noncancerous colon tissues. Most mid-IR spectrum classification studies have been proposed as machine-learning-based algorithms, but this results in deviations depending on the initial data and threshold values. We aim to develop a process for classifying colon cancer and noncancerous colon tissues through a deep-learning-based convolutional-neural-network (CNN) model. First, we image the midinfrared spectrum for the CNN model, an image-based deep-learning (DL) algorithm. Then, it is trained with the CNN algorithm and the classification ratio is evaluated using the test data. When the tissue microarray (TMA) and routine pathological slide are tested, the ML-based support-vector-machine (SVM) model produces biased results, whereas we confirm that the CNN model classifies colon cancer and noncancerous colon tissues. These results demonstrate that the CNN model using midinfrared-spectrum images is effective at classifying colon cancer tissue and noncancerous colon tissue, and not only submillimeter-sized TMA but also routine colon cancer tissue samples a few tens of millimeters in size.

**Keywords :** Convolution neural network, Hyperspectral imaging, Mid-infrared

**OCIS codes :** (170.3880) Medical and biological imaging; (170.4730) Optical pathology; (170.6510) Spectroscopy, tissue diagnostics; (300.6340) Spectroscopy, infrared

### I. INTRODUCTION

Because many biomolecular materials have absorption peaks in the midinfrared (mid-IR) spectral range, mid-IR hyperspectral imaging enables the identification of sample structure as well as sample chemistry, without staining or prior information about sample materials [1–8]. With the development of quantum cascade lasers (QCLs), it is pos-

sible to measure the mid-IR absorption spectrum faster than with Fourier-transform infrared spectroscopy (FT-IR) [9–11]. Compared to the fastest high-definition FT-IR spectroscopy, QCL-based mid-IR spectroscopy measures a tissue microarray (TMA) much faster, enabling real-time imaging of *Amoeba proteus* [12, 13].

Pathological studies based on mid-IR hyperspectral analysis have been conducted through the analysis of various

\*Corresponding author: sclim@chosun.ac.kr, ORCID 0000-0001-6179-691X

\*\*Corresponding author: jaesung.ahn@kopti.re.kr, ORCID 0000-0002-5107-0848

Color versions of one or more of the figures in this paper are available online.

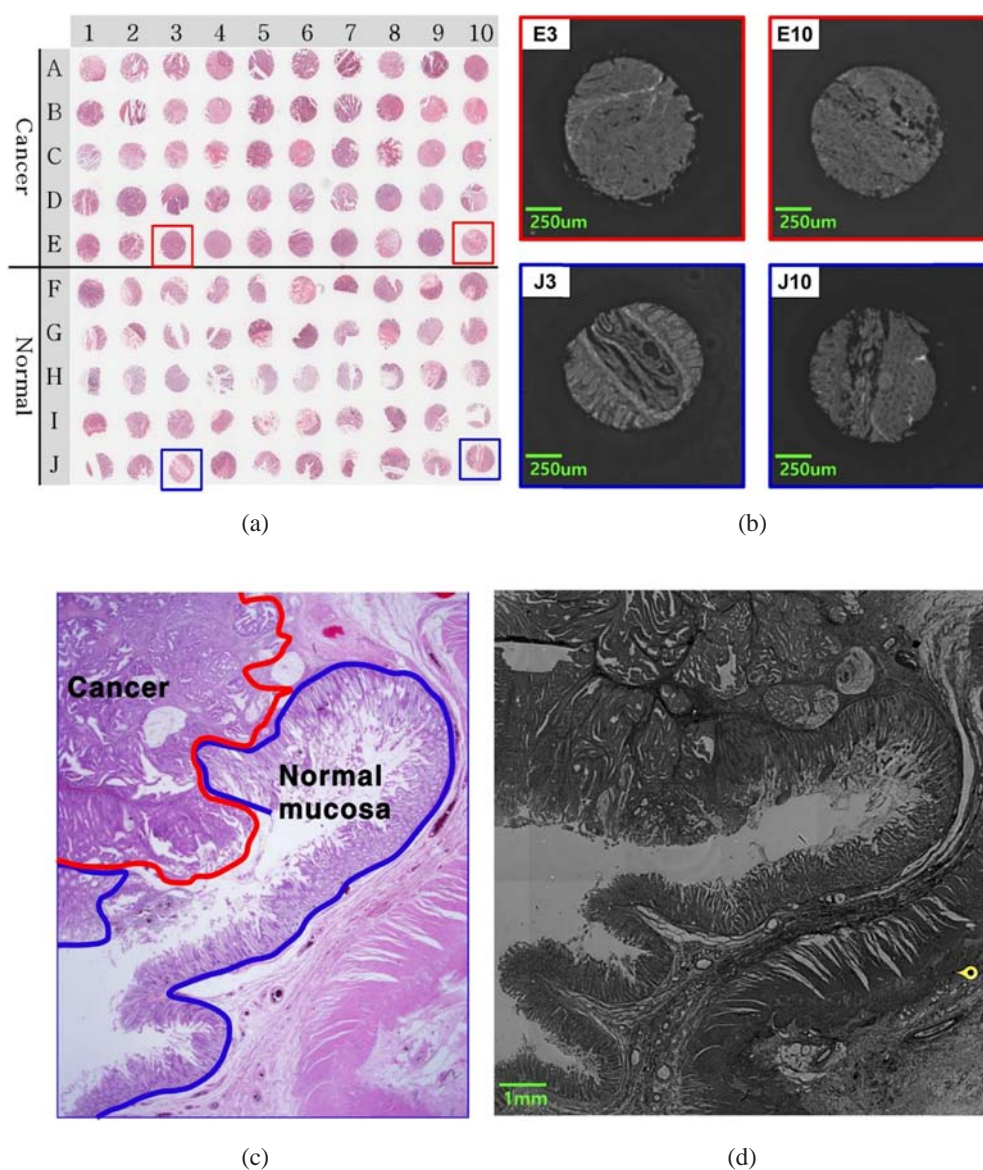


This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © 2022 Current Optics and Photonics

clinical samples including respiratory fluid, urine, blood, liver fluid, and biopsy samples [14]. Combined with the statistical pattern recognition of the mid-IR spectrum to reveal the pathological characteristics of prostate tissue, a study on the classification of benign and malignant prostate epithelium was proposed [15]. In addition, a study was proposed in which a label-free spectrum of a lung cancer pathology sample was measured and diagnosed using a multivariate analysis method [7]. By extracting features using various methods, such as principal-component analysis (PCA), linear discriminant analysis (LDA), and hierarchical cluster analysis (HCA), studies to analyze the spectrum of cancer

tissue using machine-learning (ML) -based algorithms, such as support vector machine (SVM) and random forest (RF), have also been proposed [11, 16–22]. Owing to the development of a deep-learning (DL) -based algorithm, Liu *et al.* [23] reported results from a study to detect breast cancer metastasis using a convolutional-neural-network (CNN) [24–27] model using gigapixel pathological images, and Gao *et al.* [28] reported a study that classified cancer molecular subtypes. In addition, studies confirming the results of early lung cancer diagnosis using DL for spectroscopic analysis of circulating exosomes have also been reported [29], and DL studies to analyze gastric cancer tissues using



**FIG. 1.** Measurement samples: (a) H&E stained tissue micro array (TMA) (US Biomax, Maryland, USA). TMA is sliced to minimize the effect of the peak absorption value, by maintaining the sample of the paraffin block at a constant thickness of 4–5  $\mu\text{m}$ . Paraffin with a mid-IR absorption peak is removed by immersing the section in xylene. In addition, serial sections of the sample are compared using H&E staining, to confirm the location of the sample. (b)  $1650\text{-cm}^{-1}$  images of colon cancer tissues E3, E10 and noncancerous colon tissues J3, J10 in deparaffinized TMA (scale bar: 250  $\mu\text{m}$ ). (c) Routine pathological slide diagnosed by pathologists at Chosun University Hospital. (d)  $1650\text{-cm}^{-1}$  image of the deparaffinized routine pathological slide (scale bar: 1 mm).

the difference in spatial characteristics of the spectrum have also been reported [30]. Furthermore, a study was proposed in which electroencephalograms (EEGs) or electrocardiograms (ECGs) were converted into spectrograms and analyzed with CNN [31, 32]. In prior studies, to improve the absorption-spectrum deviation and resonant Mie scattering (RMieS) [33, 34] that occur in biological samples, complex preprocessing, such as PCA [35–41], RMieS correction [33, 34], and secondary differentiation, was required [11]. In addition, ML-based algorithms such as RF yield different results based on the initial value or threshold [42, 43].

In this study, a method of converting the mid-IR spectra of colon cancer tissues into images and classifying them into groups N and T using CNN is proposed. This method simplifies the preprocessing compared to previous studies, reduces the effect of the initial value, and eliminates the need to set a threshold.

## II. METHODS

For the sample, we use a TMA (US Biomax, Maryland, USA) and routine pathological slide for training and verification of the CNN model and the SVM model respectively.

The sample is sliced to minimize the effect of the peak absorption value, by maintaining the sample of the paraffin block at a constant thickness of 4–5  $\mu\text{m}$ . Paraffin with a mid-IR absorption peak is removed by immersing the section in xylene. In addition, serial sections of the sample are compared using hematoxylin and eosin (H&E) staining, to confirm the location of the sample.

All TMA purchased from US Biomax are diagnosed pathologically, as shown in Fig. 1(a). Among the TMA, E (diagnosed as colon cancer tissue) and J (diagnosed as noncancerous colon tissue) are used as test data. The actually measured TMA is deparaffinized and confirmed using  $1650\text{-cm}^{-1}$  absorption images, which are shown in Fig. 1(b). Figure 1(c) is an H&E-stained image with pathological diagnosis of a routine pathological slide at Chosun University Hospital, and Fig. 1(d) shows the  $1650\text{-cm}^{-1}$  absorption image of the deparaffinized routine pathological slide.

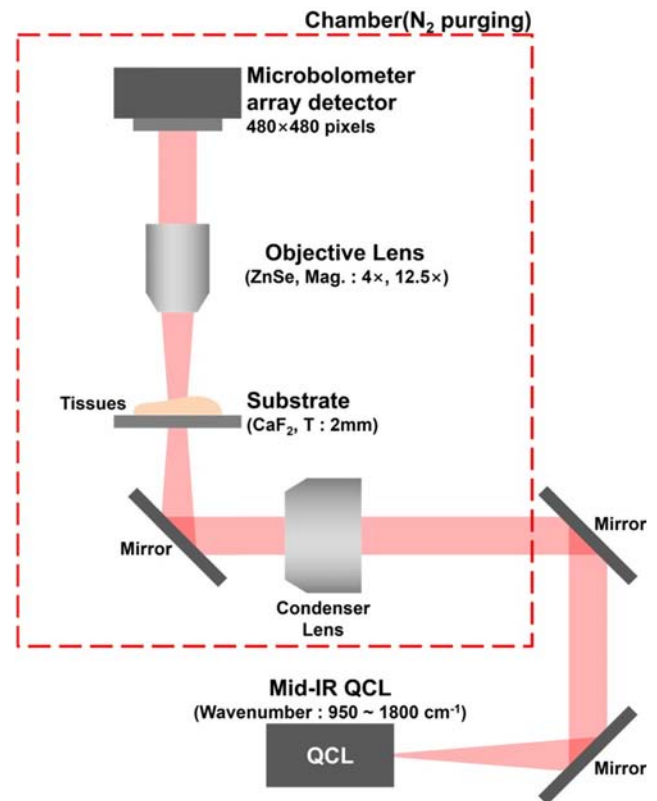
All methods and experimental protocols are approved by the Institutional Review Board of Chosun University Hospital (IRB no. CHOSUN 2021-09-016). Signed informed consent has been waived by the review board. All enrolled patients have provided informed consent. All procedures are conducted in accordance with the approved guidelines and regulations for human experimental research.

QCL-based mid-IR microscopy (Spero QT; Daylight Solutions Inc., San Diego, USA) is used for mid-IR hyperspectral imaging, as shown in Fig. 2. The hyperspectral image ( $\lambda = 5.5\text{--}10.5\ \mu\text{m}$ ) is measured in  $480 \times 480$  pixels using a microbolometer-array detector within approximately 45 s. The scale can be measured with a low magnification of  $4\times$  [field of view (FOV)  $2 \times 2\ \text{mm}$ , spatial resolution  $12\ \mu\text{m}$ ] and a high magnification of  $12.5\times$  (FOV  $650 \times 650\ \mu\text{m}$ , spatial resolution  $5\ \mu\text{m}$ ).

The mid-IR image of the sample is measured by choosing low magnification in transmission mode. To minimize variation of the spectral values due to the environment, the chamber is purged with nitrogen for 10 minutes, and the background is measured and applied once every 20 minutes.

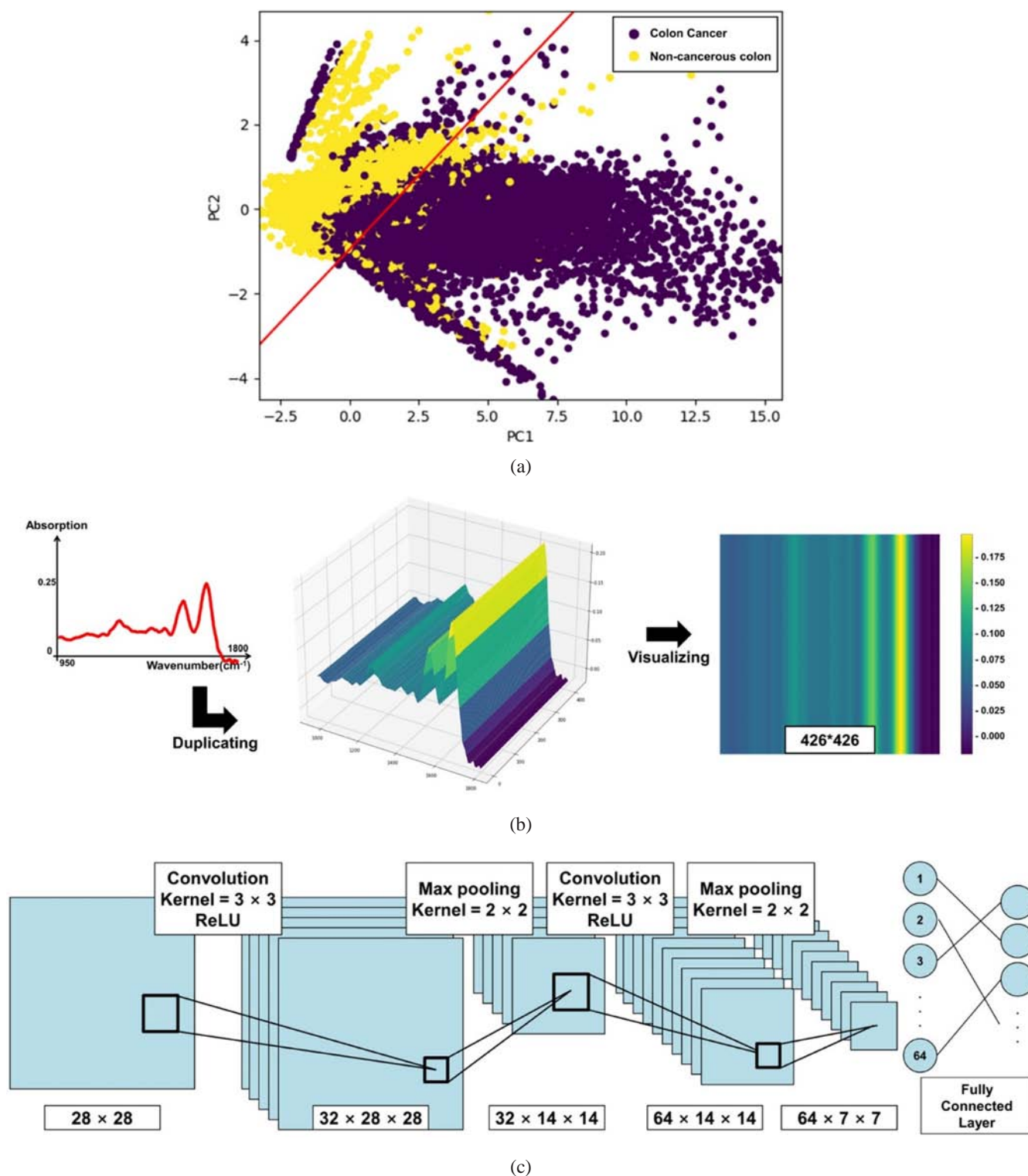
Due to the preparation of the sample while maintaining a constant thickness, the Beer-Lambert Law can be applied, and the influence of the sample thickness on the mid-IR absorption spectrum can be reduced. The Beer-Lambert Law shows the relationship between transmission and light absorption due to light passing through a substance, which causes light attenuation, such as reflection, diffraction, refraction, and scattering. This law states that the absorption of light is proportional to the thickness and the concentration of the sample. The formula used for the Beer-Lambert Law [44] is as follows ( $T$ : transmission rate,  $I_0$ : intensity of incident light,  $I_t$ : intensity of transmission light,  $A$ : absorption of light,  $\alpha$ : absorption factor,  $d$ : thickness of substance):

$$T = \frac{I_t}{I_0}, \quad (1)$$

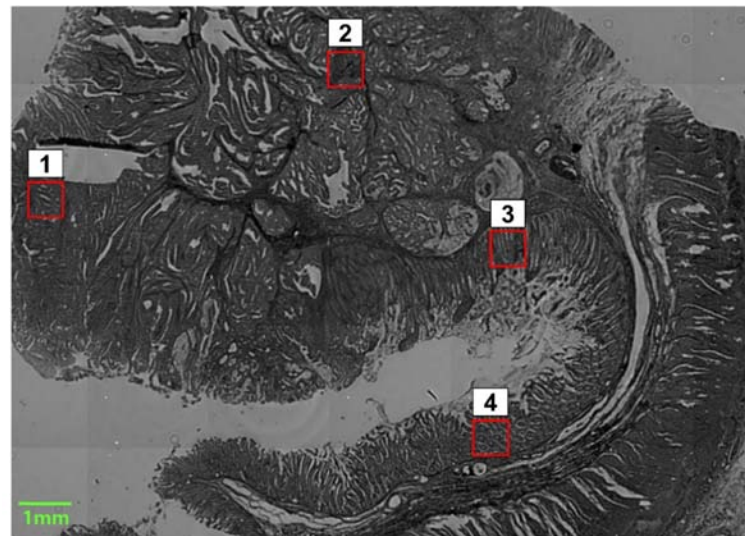


**FIG. 2.** Schematic of the experimental setup. The mid-IR hyperspectral image is measured at a wavelength from 5.5 to  $10.5\ \mu\text{m}$  in  $480 \times 480$  pixels using a microbolometer-array detector within approximately 45 s. The sample is measured with a low magnification of  $4\times$  [field of view (FOV)  $2 \times 2\ \text{mm}$ , spatial resolution  $12\ \mu\text{m}$ ] in transmission mode.

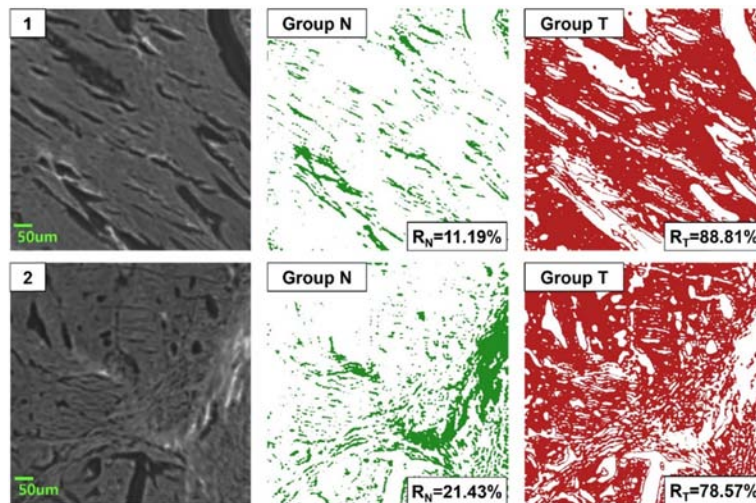




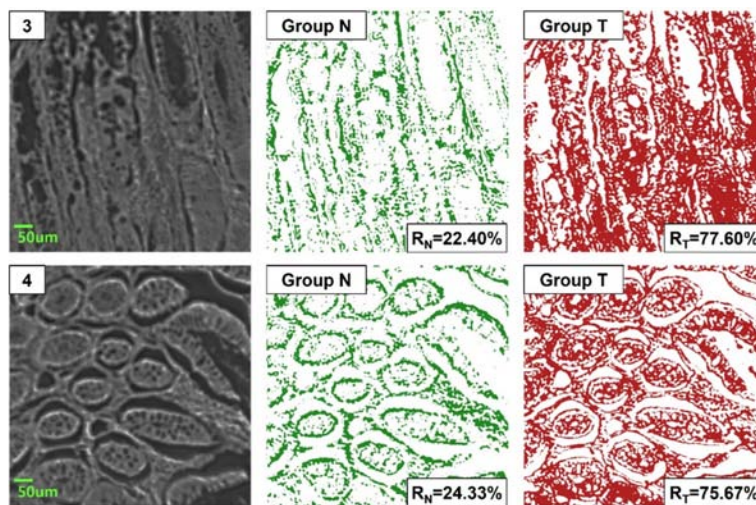
**FIG. 3.** Preprocessing and structure of the convolutional-neural-network (CNN) model: (a) the classification vector of the support vector machine (SVM) model in the feature graph of primary principal component (PC) and secondary PC, (b) To train the CNN model, which is an image-based deep-learning (DL), the mid-IR spectrum is converted into an image. The number of duplicate spectra is set to 426, which is the number of measured wave numbers, because the input data of the CNN model must be square, and (c) to reduce the image size and training time, the data size is converted from  $426 \times 426$  to  $28 \times 28$  and entered into a CNN model, consisting of a convolution and maxpooling repetitive layer and a fully connected layer.



(a)



(b)



(c)

**FIG. 4.** Classification results for the support vector machine (SVM) model for colon cancer and noncancerous colon tissue in a routine pathological slide: (a) s 1 and 2 (diagnosed as cancer) and 3 and 4 (diagnosed as normal mucosa) in a routine pathological slide (scale bar: 1 mm), (b) classification result for the SVM model of 1 and 2, diagnosed as cancer (scale bar: 50  $\mu\text{m}$ ), and (c) classification result for the SVM model of 3 and 4, diagnosed as normal mucosa (scale bar: 50  $\mu\text{m}$ ).

$$A = \log\left(\frac{I_0}{I_t}\right) = -\log T = \alpha d. \quad (2)$$

At the time of measurement, the absorption peak's position of Amide I (a type of protein) could be shifted by  $\pm 2\text{--}4\text{ cm}^{-1}$ . For better comparison of the characteristics of the spectra, the spectra of all pixels must be aligned so that Amide I peak is located at  $1650\text{ cm}^{-1}$ . To remove substrate-only parts without tissue samples from 230,400 pixels, only pixels with a peak value at  $1650\text{ cm}^{-1}$  between 0 and 2 are included in the training and evaluation data. The target data required for training data is set to group T to 1 and group N to 0. Group T is set to the TMA that is pathologically diagnosed with colon cancer, and group N is set to the TMA that is pathologically diagnosed with noncancerous colon tissue.

The preprocessing images are divided into a training dataset, test dataset, and validation dataset. The training dataset is the dataset used to train the model; the test dataset is the dataset that extracts 20% of random data from the training dataset, to check whether the training is performing well; and the validation dataset, called test data, is a separate dataset from the training dataset.

In SVM [45–48] model training, 10 principal components (PCs) extracted from the preprocessed data through PCA [35–41] are input. The SVM model is classified into two groups: colon cancer tissue group T and noncancerous colon tissue group N. As a result, the classification vector of the SVM model appears as Fig. 3(a) in the feature graph of primary PC and secondary PC. PCA is performed using the `sklearn.decomposition.PCA` function provided by Python (Python Software Foundation, DE, USA), and the SVM model is created using the `sklearn.svm.SVC` function provided by Python.

To train the CNN model, which is an image-based DL, the mid-IR spectrum is converted into an image. To classify the hyperspectral mid-IR image of the tissue pixel by pixel, the mid-IR spectrum of each pixel is duplicated several times to create an image. The number of duplicate spectra is set to 426, which is the number of measured wave numbers, because the input data of the CNN model must be square.

To train with large amounts of image data, the image size is converted from  $426 \times 426$  to  $28 \times 28$  using the `resize` function of `OpenCV`. The layer of the CNN model is designed to repeat convolution and max-pooling, and the feature nodes are connected as a fully connected layer; it is constructed using the `Conv2D`, `MaxPooling`, `Flatten`, `Dropout`, and `Dense` functions provided by `Keras` (Keras, CA, USA). All convolution layers consist of  $3 \times 3$  kernels [stride = 1, padding = 1, and Rectified Linear Unit (ReLU)], and all pooling layers consist of  $2 \times 2$  kernels.

### III. RESULTS

The SVM model and CNN model are trained with 2,521,326 input data. The trained SVM and CNN models

are tested using TMA and a routine pathological slide. Among routine pathological slides, 1 and 2 of Figs. 4 and 5 (diagnosed as cancer) and 3 and 4 of Figs. 4 and 5 (diagnosed as normal mucosa) are tested. The formula used for the spectral classification result of the hyperspectral mid-IR image is as follows ( $R_c$ : ratio of cancer tissue,  $R_n$ : ratio of noncancerous tissue,  $N_c$ : number of pixels corresponding to cancer tissue,  $N_n$ : number of pixels corresponding to noncancerous tissue,  $N_t$ : number of total pixels):

$$R_c(R_n) = \frac{N_c(N_n)}{N_t} \times 100. \quad (3)$$

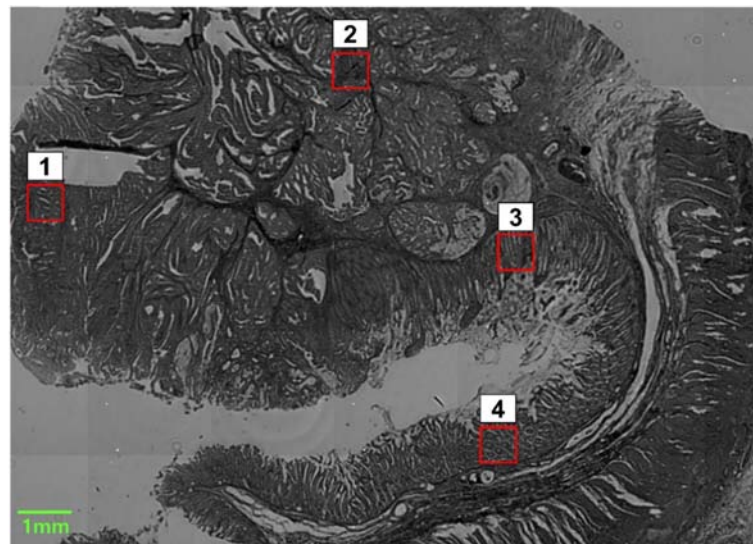
In the SVM model, 89.27% of the 367,780 pixels of tissues diagnosed with colon cancer among TMA are classified into group T, and 18.95% of the 310,001 pixels of tissues diagnosed with noncancerous colon tissue among TMA are classified as group N (Fig. 6). In addition, of the 668,947 pixels in the routine pathological slide, 80.16% are classified as group T and 19.84% are classified as group N. As a result, the test results of the SVM model tend to be biased toward group T regardless of the pathology diagnosis, for both TMA and routine pathological slide. The SVM regularization parameter and the type and coefficient of the SVM kernel are not optimized, so it is predicted that the SVM model will have difficulty in classifying PCs.

In the CNN model, 63.29% of the 367,780 images of tissues diagnosed with colon cancer among TMA are classified as group T. Of the 310,001 images of tissues diagnosed with noncancerous colon tissue among TMA, 80.17% are classified into group N (Fig. 7). In addition, of the 668,947 images of a routine pathological slide, 35.20% are classified as group T and 64.80% as group N. Unlike the test results for the ML-based SVM model, the CNN model confirms that colon cancer and noncancerous colon tissues can be classified, and there is no need to optimize parameters and kernels, so it is more effective in classifying.

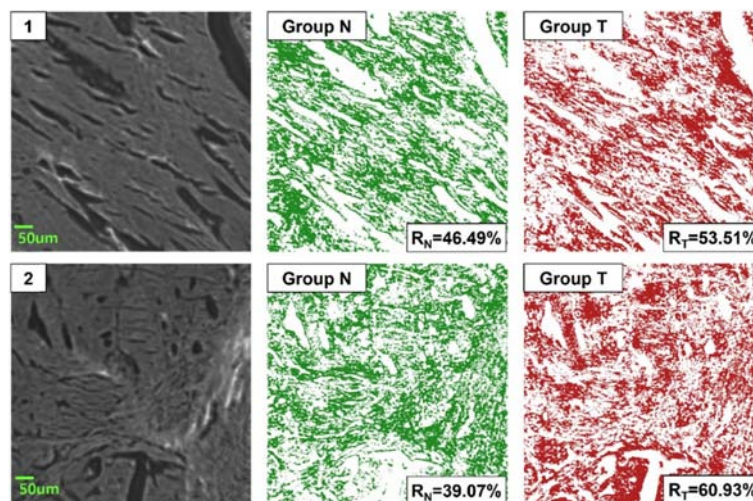
### IV. DISCUSSION

In prior studies, complex preprocessing such as PCA, RMieS [33, 34] calibration, and secondary differentiation were required to improve RMieS [33, 34] and absorption-spectrum deviations occurring in biological samples [11]. In addition, for ML-based algorithms such as RF, the resulting value is based on the initial value or threshold [42, 43]. Therefore, a method for converting the mid-IR spectrum of colon cancer into an image and classifying it into group N or T using CNN has been proposed. The preprocessing is simpler than that in previous studies, and the influences of the initial value and threshold are less. The CNN model was evaluated based on TMA and a routine pathological slide. As a result of the CNN model in this study classifying TMA images by pixel, 63.29% of pixels were classified as Group T in cancer tissue TMA and 80.27% as Group N in normal tissue TMA. Therefore, the CNN model of

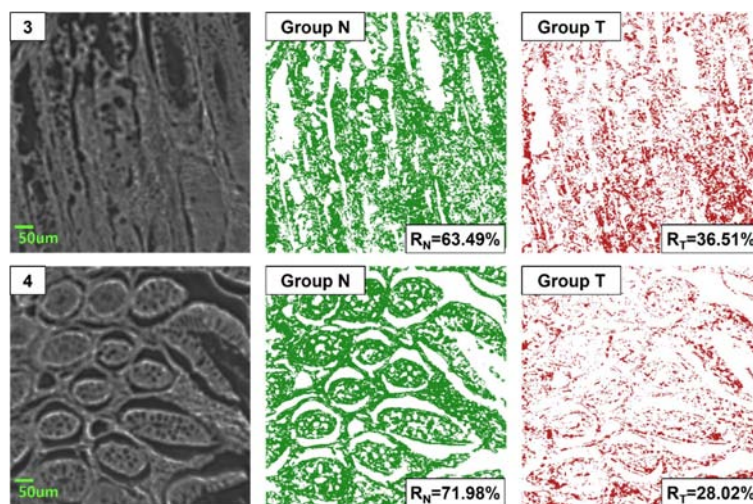




(a)

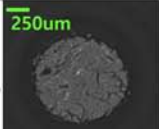


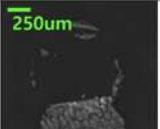





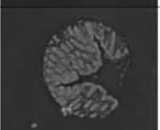


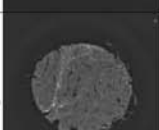
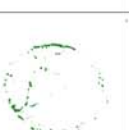










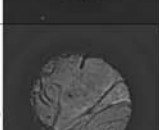





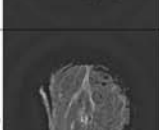


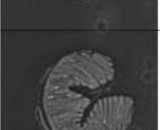


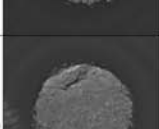

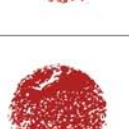
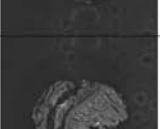



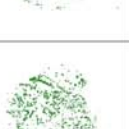
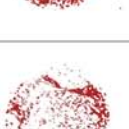
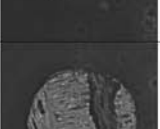


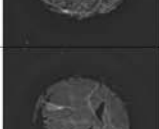













(b)



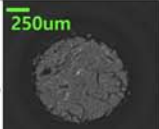


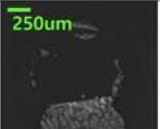


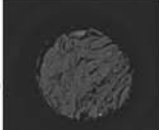


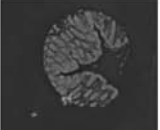


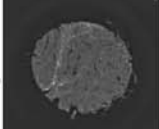
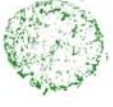










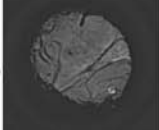


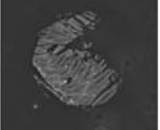











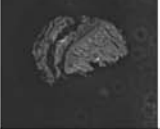

















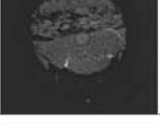


(c)

**FIG. 5.** Classification results for the convolutional-neural-network (CNN) model for colon cancer and noncancerous colon tissue in a routine pathological slide: (a) locations 1 and 2 (diagnosed as cancer) and 3 and 4 (diagnosed as normal mucosa) in a routine pathological slide (scale bar: 1 mm), (b) classification result for the CNN model of 1 and 2, diagnosed as cancer (scale bar: 50  $\mu\text{m}$ ), and (c) classification result for the CNN model of 3 and 4, diagnosed as normal mucosa (scale bar: 50  $\mu\text{m}$ ).

Colon Cancer Tissues					Normal Colon Tissues				
TMA (TNM)	Amide I (@1650cm <sup>-1</sup> )	Group N	Group T	Ratio(T)	TMA (TNM)	Amide I (@1650cm <sup>-1</sup> )	Group N	Group T	Ratio(N)
E1 (T4N2M0)				93.65%	J1				6.88%
E2 (T4N1M0)				93.15%	J2				18.26%
E3 (T3N0M0)				93.50%	J3				15.45%
E4 (T3N0M0)				96.88%	J4				14.78%
E5 (T3N1M0)				97.14%	J5				5.04%
E6 (T3N1M0)				90.75%	J6				26.25%
E7 (T3N1M0)				95.62%	J7				25.50%
E8 (T3N0M0)				56.49%	J8				33.36%
E9 (T3N2M0)				83.36%	J9				28.32%
E10 (T4N2M0)				77.27%	J10				15.63%
Total				89.27%	Total				18.95%

**FIG. 6.** Classification results for the support vector machine (SVM) model for colon cancer and noncancerous colon tissue in tissue micro array (TMA): left, colon cancer tissue (E1:10); right, noncancerous colon tissue (J1:10) (scale bar: 250 μm).



Colon Cancer Tissues					Non-cancerous Colon Tissues				
TMA (TNM)	Amide I (@1650cm <sup>-1</sup> )	Group N	Group T	Ratio(T)	TMA (TNM)	Amide I (@1650cm <sup>-1</sup> )	Group N	Group T	Ratio(N)
E1 (T4N2M0)				61.26%	J1				85.12%
E2 (T4N1M0)				48.02%	J2				83.81%
E3 (T3N0M0)				69.66%	J3				85.80%
E4 (T3N0M0)				48.83%	J4				84.29%
E5 (T3N1M0)				74.70%	J5				86.09%
E6 (T3N1M0)				64.14%	J6				71.90%
E7 (T3N1M0)				70.25%	J7				67.14%
E8 (T3N0M0)				51.45%	J8				68.59%
E9 (T3N2M0)				71.14%	J9				89.97%
E10 (T4N2M0)				73.41%	J10				78.98%
Total				63.29%	Total				80.17%

**FIG. 7.** Classification results for the convolutional-neural-network (CNN) model for colon cancer and noncancerous colon tissue in tissue micro array (TMA): left, colon cancer tissue (E1:10); right, noncancerous colon tissue (J1:10) (scale bar: 250  $\mu$ m).

this study has demonstrated that both TMA and a routine pathological slide can be classified as colon cancer or non-cancerous colon tissue. In addition, the ML-based SVM model showed results skewed toward group T, and most of the mid-IR spectra of J (diagnosed as noncancerous colon tissue) were classified as group T. On the contrary, it was confirmed that the CNN model has relatively high accuracy in the classification of colon cancer and noncancerous colon tissue, compared to the SVM model. However, the accuracy of the CNN model is still not sufficient, because the number of epochs based on the model training time is insufficient compared to the number of data, and the target data are comprehensive because there is no pixel-by-pixel diagnosis for colon cancer tissue. To increase the accuracy of the model, pixel-by-pixel annotation of the colon cancer tissue by the pathologist is required, and either the number of epochs or the depth of the model should be increased by using a deeper CNN algorithm.

## V. CONCLUSION

In conclusion, a method for converting the mid-IR spectrum of colon cancer to an image and classifying it into group N or T using CNN was proposed, and the model was evaluated using TMA and a routine pathological slide. We visualized mid-IR spectra into a 2D image, for classification of colon cancer tissue using a CNN model. As a result, the complex preprocessing was simplified and the deviation in the results that was observed in previous studies was reduced. This study showed that the CNN model could classify colon cancer and noncancerous colon tissue. In addition, the ML-based SVM model showed results that were skewed toward group T, and it was confirmed that the mid-IR spectra of J (diagnosed with noncancerous colon tissue) were mostly classified as group T. Compared to the SVM model, the CNN model has relatively high accuracy in the classification of colon cancer and noncancerous colon tissue. We expect that in future works the sensitivity, specificity, and accuracy of the CNN model could be assessed through comparison with the pathologist's decision data with pixel-by-pixel target data. Thus, it is expected that the CNN model of this study will be helpful in pathological diagnosis through the mid-IR absorption spectrum, which is an objective indicator.

## FUNDING

Korean Medical Device Development Fund grant funded by the Korean government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, and the Ministry of Food and Drug Safety) (Project Number: 1711137874, KMDF\_PR\_20200901\_0008).

## ACKNOWLEDGMENT

This work was supported by the Korean Medical Device Development Fund grant funded by the Korean government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, and the Ministry of Food and Drug Safety) (Project Number: 1711137874, KMDF\_PR\_20200901\_0008), and was also partially supported by a grant from the "HPC Support" Project, supported by the "Ministry of Science and ICT" and NIPA.

## DISCLOSURES

The authors declare no conflicts of interest.

## DATA AVAILABILITY

Data underlying the results presented in this paper are not publicly available at the time of publication, which may be obtained from the authors upon reasonable request.

## REFERENCES

1. H. H. Mantsch and D. Chapman, *Infrared spectroscopy of bio-molecules* (Wiley-Liss, USA, 1996).
2. 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, and F. L. Martin, "Using Fourier transform IR spectroscopy to analyze biological materials," *Nat. Protoc.* **9**, 1771–1791 (2014).
3. K. B. Beć, J. Grabska, and C. W. Huck, "Biomolecular and bioanalytical applications of infrared spectroscopy—A review," *Anal. Chim. Acta* **1133**, 150–177 (2020).
4. K. Z. Liu, M. Xu, and D. A. Scott, "Biomolecular characterization of leucocytes by infrared spectroscopy," *Br. J. Haematol.* **136**, 713–722 (2007).
5. M. Diem, M. Romeo, S. Boydston-White, M. Miljković, and C. Matthäus, "A decade of vibrational micro-spectroscopy of human cells and tissue (1994–2004)," *Analyst* **129**, 880–885 (2004).
6. M. Hermes, R. B. Morrish, L. Huot, L. Meng, S. Junaid, J. Tomko, G. R. Lloyd, W. T. Masselink, P. Tidemand-Lichtenberg, C. Pedersen, F. Palombo, and N. Stone, "Mid-IR hyperspectral imaging for label-free histopathology and cytology," *J. Opt.* **20**, 023002 (2018).
7. B. Bird, M. Miljković, S. Remiszewski, A. Akalin, M. Kon, and M. Diem, "Infrared spectral histopathology (SHP): a novel diagnostic tool for the accurate classification of lung cancer," *Lab. Invest.* **92**, 1358–1373 (2012).
8. B. Bird and M. J. Baker, "Quantum cascade lasers in biomedical infrared imaging," *Trends Biotechnol.* **33**, 557–558 (2015).
9. P. Bassan, A. Sachdeva, J. H. Shanks, M. D. Brown, N. W. Clarke, and P. Gardner, "Whole organ cross-section chemical

- imaging using label-free mega-mosaic FTIR microscopy,” *Analyst* **138**, 7066–7069 (2013).
10. P. Bassan, M. J. Weida, J. Rowlette, and P. Gardner, “Large scale infrared imaging of tissue micro arrays (TMAs) using a tunable quantum cascade laser (QCL) based microscope,” *Analyst* **139**, 3856–3859 (2014).
  11. C. Kuepper, A. Kallenbach-Thieltges, H. Juette, A. Tannapfel, F. Großerueschkamp, and K. Gerwert, “Quantum cascade laser-based infrared microscopy for label-free and automated cancer classification in tissue sections,” *Sci. Rep.* **8**, 7717 (2018).
  12. K. Yeh, S. Kenkel, J.-N. Liu, and R. Bhargava, “Fast infrared chemical imaging with a quantum cascade laser,” *Anal. Chem.* **87**, 485–493 (2015).
  13. K. Haase, N. Kröger-Lui, A. Pucci, A. Schönhals, and W. Petrich, “Real-time mid-infrared imaging of living microorganisms,” *J. Biophotonics* **9**, 61–66 (2016).
  14. A. Schwaighofer, M. Brandstetter, and B. Lendl, “Quantum cascade lasers (QCLs) in biomedical spectroscopy,” *Chem. Soc. Rev.* **46**, 5903–5924 (2017).
  15. D. C. Fernandez, R. Bhargava, S. M. Hewitt, I. and W. Levin, “Infrared spectroscopic imaging for histopathologic recognition,” *Nat. Biotechnol.* **23**, 469–474 (2005).
  16. A. Kallenbach-Thieltges, F. Großerüschkamp, A. Mosig, M. Diem, A. Tannapfel, and K. Gerwert, “Immunohistochemistry, histopathology and infrared spectral histopathology of colon cancer tissue sections,” *J. Biophotonics* **6**, 88–100 (2013).
  17. F. Chu and L. Wang, “Applications of support vector machines to cancer classification with microarray data,” *Int. J. Neural Syst.* **15**, 475–484 (2005).
  18. C. Leslie, E. Eskin, and W. S. Noble, “The spectrum kernel: a string kernel for SVM protein classification,” in *Proc. Pacific Symposium on Biocomputing* (Kauai, Hawaii, USA, Jan. 2002), pp. 564–575.
  19. K. Ali, Y. Lu, U. Das, R. K. Sharma, S. Wiebe, K. Meguro, V. Sadanand, D. R. Fourney, A. Vitali, M. Kelly, T. May, J. Gomez, and E. Pellerin, “Biomolecular diagnosis of human glioblastoma multiforme using Synchrotron mid-infrared spectromicroscopy,” *Int. J. Molec. Med.* **26**, 11–16 (2010).
  20. G. L. Owens, K. Gajjar, J. Trevisan, S. W. Fogarty, S. E. Taylor, B. Da Gama-Rose, P. L. Martin-Hirsch, and F. L. Martin, “Vibrational biospectroscopy coupled with multivariate analysis extracts potentially diagnostic features in blood plasma/serum of ovarian cancer patients,” *J. Biophotonics* **7**, 200–209 (2014).
  21. L.-W. Shang, D.-Y. Ma, J.-J. Fu, Y.-F. Lu, Y. Zhao, X.-Y. Xu, and J.-H. Yin, “Fluorescence imaging and Raman spectroscopy applied for the accurate diagnosis of breast cancer with deep learning algorithms,” *Biomed. Opt. Express* **11**, 3673–3683 (2020).
  22. O. Collier, V. Stoven, and J.-P. Vert, “LOTUS: A single-and multitask machine learning algorithm for the prediction of cancer driver genes,” *PLoS Comput. Biol.* **15**, e1007381 (2019).
  23. Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, J. D. Hipp, L. Peng, and M. C. Stumpe, “Detecting cancer metastases on gigapixel pathology images,” arXiv 1703.02442 (2017).
  24. S.-C. B. Lo, S.-L. A. Lou, J.-S. Lin, M. T. Freedman, M. V. Chien, and S. K. Mun, “Artificial convolution neural network techniques and applications for lung nodule detection,” *IEEE Trans. Med. Imaging* **14**, 711–718 (1995).
  25. B. B. Traore, B. Kamsu-Foguem, and F. Tangara, “Deep convolution neural network for image recognition,” *Ecol. Inform.* **48**, 257–268 (2018).
  26. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, P. Bartlett, Eds. (Neural Information Processing Systems Foundation, USA. 2012), vol. 25, pp. 1097–1105.
  27. S.-C. B. Lo, H.-P. Chan, J.-S. Lin, H. Li, M. T. Freedman, and S. K. Mun, “Artificial convolution neural network for medical image pattern recognition,” *Neural Netw.* **8**, 1201–1214 (1995).
  28. F. Gao, W. Wang, M. Tan, L. Zhu, Y. Zhang, E. Fessler, L. Vermeulen, and X. Wang, “DeepCC: a novel deep learning-based framework for cancer molecular subtype classification,” *Oncogenesis* **8**, 44 (2019).
  29. H. Shin, S. Oh, S. Hong, M. Kang, D. Kang, Y.-G. Ji, B. H. Choi, K.-W. Kang, H. Jeong, Y. Park, S. Hong, H. K. Kim, and Y. Choi, “Early-stage lung cancer diagnosis by deep learning-based spectroscopic analysis of circulating exosomes,” *ACS nano* **14**, 5435–5444 (2020).
  30. B. Hu, J. Du, Z. Zhang, and Q. Wang, “Tumor tissue classification based on micro-hyperspectral technology and deep learning,” *Biomed. Opt. Express* **10**, 6370–6389 (2019).
  31. L. Yuan and J. Cao, “Patients’ EEG data analysis via spectrogram image with a convolution neural network,” in *Intelligent Decision Technologies 2017* (Springer, Switzerland, 2018), pp. 13–21.
  32. J. Huang, B. Chen, B. Yao, and W. He, “ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network,” *IEEE Access* **7**, 92871–92880 (2019).
  33. P. Bassan, A. Kohler, H. Martens, J. Lee, H. J. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke, and P. Gardner, “Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples,” *Analyst* **135**, 268–277 (2010).
  34. B. Bird, M. Miljkovic, and M. Diem, “Two step resonant Mie scattering correction of infrared micro-spectral data: human lymph node tissue,” *J. Biophotonics* **3**, 597–608 (2010).
  35. S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemom. Intell. Lab. Syst.* **2**, 37–52 (1987).
  36. M. E. Tipping and C. M. Bishop, “Probabilistic principal component analysis,” *J. R. Stat. Soc. Series B Stat. Methodol.* **61**, 611–622 (1999).
  37. M. Ringnér, “What is principal component analysis?,” *Nat. Biotechnol.* **26**, 303–304 (2008).
  38. M. Richardson, (2009). Principal component analysis [Online], Available: <http://www.dsc.ufcg.edu.br/~hmg/disciplinas/pos-graduacao/rm-copin-2014.3/material/SignalProcPCA.pdf>



39. H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Comput. Stat.* **2**, 433–459 (2010).
40. R. Bro and A. K. Smilde, "Principal component analysis," *Anal. Methods* **6**, 2812–2831 (2014).
41. J. Shlens, "A tutorial on principal component analysis," arXiv:1404.1100 (2014).
42. L. Breiman, "Random forests," *Mach. Learn.* **45**, 5–32 (2001).
43. A. Liaw and M. Wiener, "Classification and regression by random Forest," *R news* **2**, 18–22 (2002).
44. D. F. Swinehart, "The beer-lambert law," *J. Chem. Educ.* **39**, 333–335 (1962).
45. J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.* **9**, 293–300 (1999).
46. D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," *Neurocomputing* **55**, 169–186 (2003).
47. W. S. Noble, "What is a support vector machine?," *Nat. Biotechnol.* **24**, 1565–1567 (2006).
48. A. Widodo and B.-S. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," *Mech. Syst. Signal Process.* **21**, 2560–2574 (2007).