frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 19 June 2024
DOI 10.3389/fonc.2024.1320220

OPEN ACCESS

EDITED BY
Nicolas A. Karakatsanis,
Cornell University, United States

REVIEWED BY
Bo Zhou,
Yale University, United States
Ivan Bratchenko,
Samara University, Russia

*CORRESPONDENCE
Haishan Zeng
hzeng@bccrc.ca

RECEIVED 12 October 2023
ACCEPTED 23 May 2024
PUBLISHED 19 June 2024

CITATION
Zhao J, Lui H, Kalia S, Lee TK and Zeng H (2024) Improving skin cancer detection by Raman spectroscopy using convolutional neural networks and data augmentation. *Front. Oncol.* 14:1320220. doi: 10.3389/fonc.2024.1320220

Improving skin cancer detection by Raman spectroscopy using convolutional neural networks and data augmentation

Jianhua Zhao[1,2], Harvey Lui[1,2], Sunil Kalia[1,3,4], Tim K. Lee[1,2] and Haishan Zeng[1,2]*

[1]Photomedicine Institute, Department of Dermatology and Skin Science, University of British Columbia and Vancouver Coastal Health Research Institute, Vancouver, BC, Canada, [2]BC Cancer Research Institute, University of British Columbia, Vancouver, BC, Canada, [3]BC Children's Hospital Research Institute, Vancouver, BC, Canada, [4]Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal Health Research Institute, Vancouver, BC, Canada

Background: Our previous studies have demonstrated that Raman spectroscopy could be used for skin cancer detection with good sensitivity and specificity. The objective of this study is to determine if skin cancer detection can be further improved by combining deep neural networks and Raman spectroscopy.

Patients and methods: Raman spectra of 731 skin lesions were included in this study, containing 340 cancerous and precancerous lesions (melanoma, basal cell carcinoma, squamous cell carcinoma and actinic keratosis) and 391 benign lesions (melanocytic nevus and seborrheic keratosis). One-dimensional convolutional neural networks (1D-CNN) were developed for Raman spectral classification. The stratified samples were divided randomly into training (70%), validation (10%) and test set (20%), and were repeated 56 times using parallel computing. Different data augmentation strategies were implemented for the training dataset, including added random noise, spectral shift, spectral combination and artificially synthesized Raman spectra using one-dimensional generative adversarial networks (1D-GAN). The area under the receiver operating characteristic curve (ROC AUC) was used as a measure of the diagnostic performance. Conventional machine learning approaches, including partial least squares for discriminant analysis (PLS-DA), principal component and linear discriminant analysis (PC-LDA), support vector machine (SVM), and logistic regression (LR) were evaluated for comparison with the same data splitting scheme as the 1D-CNN.

Results: The ROC AUC of the test dataset based on the original training spectra were 0.886 ± 0.022 (1D-CNN), 0.870 ± 0.028 (PLS-DA), 0.875 ± 0.033 (PC-LDA), 0.864 ± 0.027 (SVM), and 0.525 ± 0.045 (LR), which were improved to 0.909 ± 0.021 (1D-CNN), 0.899 ± 0.022 (PLS-DA), 0.895 ± 0.022 (PC-LDA), 0.901 ± 0.020 (SVM), and 0.897 ± 0.021 (LR) respectively after augmentation of the training dataset ($p<0.0001$, Wilcoxon test). Paired analyses of 1D-CNN with conventional machine learning approaches showed that 1D-CNN had a 1–3% improvement ($p<0.001$, Wilcoxon test).

Conclusions: Data augmentation not only improved the performance of both deep neural networks and conventional machine learning techniques by 2–4%, but also improved the performance of the models on spectra with higher noise or spectral shifting. Convolutional neural networks slightly outperformed conventional machine learning approaches for skin cancer detection by Raman spectroscopy.

KEYWORDS

Skin cancer detection, Raman spectroscopy, convolutional neural networks (CNN), artificial intelligence (AI), optical diagnosis, data augmentation, machine learning

1 Introduction

Skin cancers including basal cell carcinoma (BCC), squamous cell carcinoma (SCC) and malignant melanoma (MM) are the most common of all types of cancers with an estimate of over 5.4 million new skin cancer cases per year in the US (including 97,610 new melanoma cases) affecting more than 3.3 million patients (1). The incidence in Australia is even higher, with 2/3 of Australians developing skin cancer in their life time (2). Clinical diagnosis of skin cancer is typically based on visual inspection followed by an invasive biopsy of the suspicious lesion. It is invasive, time consuming and costly because the procedures of biopsy involve tissue processing and histology. Biopsies also generate a large number of false negatives and false positives. For example, in a large scale retrospective study of 4741 pigmented skin lesions, it was reported that for each confirmed melanoma, over 20 benign lesions were biopsied (3). Therefore, new techniques to aid skin cancer detection and reduce the misdiagnosis rate are being evaluated. A number of techniques have been proposed and different levels of performance have been demonstrated for skin cancer detection, such as Raman spectroscopy (4–7), dermoscopy (8–10), spectral imaging (11–13), confocal microscopy (14–16), electrical impedance spectroscopy (17, 18), multiphoton microscopy (19–22) and optical coherence tomography (OCT) (23).

Raman spectroscopy is an optical technique that measures the vibrational modes of biomolecules within the tissue. It is very sensitive to biochemical and biological changes associated with pathology. Raman spectroscopy has been investigated extensively for *in vitro* and *in vivo* skin cancer detection (4–7, 24–35). A number of excellent review articles on cancer detection by Raman spectroscopy have been published (36–40). Earlier work on skin cancer detection by Raman spectroscopy was limited either by *ex vivo* biopsied samples or by small number of *in vivo* cases due to long measurement times. For example, Gniadecka et al. (33) measured 223 punch biopsied skin samples by near infrared Fourier transform Raman spectroscopy, in which each spectrum was acquired over approximately 7 minutes. They found that the sensitivity and specificity for diagnosis of melanoma by neural network analysis were as high as 85% and 99%, respectively. Lieber et al. (27) measured 21 lesions and their adjacent normal skin *in vivo* with an integration time of 30 seconds, and reported 100% sensitivity and 91% specificity for discriminating skin lesions from normal skin. We have developed a rapid, real-time Raman spectrometer system for *in vivo* skin measurements that substantially reduced spectral acquisition times to less than a second (41, 42). In a recent study of 518 *in vivo* cases by our group, Lui et al. (4, 5) found that Raman spectroscopy could be used for skin cancer detection with an area under the receiver operating characteristic curve (ROC AUC) as high as 89.6% based on Raman spectrum alone. With feature selection (wavenumber selection) and by incorporating patient demographics into the algorithm, the diagnostic ROC AUC was further improved (6, 7). Very recently, Feng et al. quantified biophysical markers associated with different skin pathologies (24, 25). Bratchenko et al. (30–32) found that by combining Raman and autofluorescence spectra in the near-infrared region using a portable low-cost spectrometer, a reasonable diagnostic accuracy was achieved.

Recently Esteva et al. (43) reported that deep neural networks could improve the performance of skin cancer diagnosis based on color dermoscopic images. It stimulated further studies in artificial intelligence for biomedical image and spectral analysis (44–48). Currently, deep neural networks has been proposed for spectral analysis, such as spectral preprocessing (49–51), spectral classification (31, 52–56), and spectral data highlighting (57). Raman spectroscopy combining with deep neural networks have been reported for detection of breast cancer (biopsied samples, 8 subjects) (54, 58), colon cancer (*ex vivo* samples, 45 subjects) (53), prostate cancer (urine samples, 84 subjects) (55) and liver cancer (serum samples, 66 subjects) (59). All these studies were limited by the small number of cases, which might be over-trained for data-hungry deep neural networks that required a large amount of data to train. Bratchenko et al. (31) reported skin cancer detection using Raman spectroscopy and found that convolutional neural networks substantially improved the ROC AUC from 0.75 for PLS-DA to 0.96 for CNN based on the raw Raman spectra.

The objective of this study is to explore skin cancer detection by analyzing Raman spectra using deep neural networks. Based on clinical interest this study is focused on a dichotomous binary

classification to determine whether a lesion is cancerous. We implemented different data augmentation strategies to increase the training dataset and compared the results of deep neural networks and conventional machine learning techniques with and without data augmentation. The paper is outlined as the following: section 2 described the patient dataset; different data augmentation strategies, in particular the details of one-dimensional generative adversarial networks (1D-GAN) for data augmentation; and the one-dimensional convolutional neural networks (1D-CNN) for spectral classification. Section 3 presented the performance of different data augmentation strategies and the results based on the original training datasets with and without data augmentation. Section 4 summarized the major findings and section 5 concluded the study.

2 Patient and method

2.1 Patient dataset

The dataset used in this study has been reported in a previous publication (7). In total, there were 731 valid lesions from 644 patients, including 326 males and 318 females with a median age of 62 years old (range: 18–94). Of the 731 lesions, 340 cases were cancerous or precancerous lesions (melanomas, basal cell carcinoma, squamous cell carcinoma and actinic keratosis), and 391 cases were benign lesions (atypical nevus, blue nevus, compound nevus, intradermal nevus, junctional nevus and seborrheic keratosis). All these lesions were clinically confirmed by the experienced dermatologists. All of the skin cancer lesions (100%), 29% of the precancer lesions and 34% of the benign lesions were also confirmed by histopathology. This study was approved by the Clinical Research Ethics Board of the University of British Columbia (Vancouver, BC, Canada; protocol C96–0499).

Raman spectra of all the lesions were measured *in vivo* using a custom-build real-time Raman spectrometer system (41, 42). The system contained a 785 nm diode laser, a hand-held Raman probe and a spectrograph equipped with liquid nitrogen cooled back-illumination deep depletion charge coupled device (CCD) detector. The laser was delivered to the Raman probe through a single multimode fiber with core diameter of 100 µm and formed a 3.5 mm diameter spot on the skin target. The Raman signal was collected by the Raman probe and delivered to the spectrograph through a fiber bundle, which consisted of 58 multimode optical fibers with core diameter of 100 µm. The distal end of the fiber bundle was packed into a circular area, and the proximal end connected to the spectrograph was aligned along a specially-designed parabolic line to correct the aberration of the spectrograph. Full-chip vertical hardware binning was achieved after image aberration correction, which improved the signal-to-noise ratio by 16 times (41, 42). The raw Raman signal was filtered by a 5-point box-car smoothing, and the fluorescence background was removed using fifth-order polynomial fitting of the Vancouver Raman Algorithm (60). Most of the lesions (96%) were acquired of a single spectrum; large and inhomogeneous lesions (4%) were acquired of multiple times from different locations within the lesion, and the averaged Raman spectrum was used for analysis. In this study, each individual lesion was considered as an experimental unit for analysis.

The averaged Raman spectra and standard deviation of skin cancers and precancerous, and benign skin lesions were shown in Figure 1. All the spectra were normalized to their respective area under the curve between 500 and 1800 cm^{-1} before being averaged. Major Raman peaks were located around 855, 936, 1,002, 1,271, 1,302, 1,445, 1,655, and 1,745 cm^{-1}. It was noted that all the skin lesions shared similar Raman peaks and bands with different intensities. These differences in intensities provided the diagnostic capability between skin cancers and benign skin lesions. It is difficult if not impossible to identify the peaks that can provide the best discrimination. Features extracted from machine learning techniques (such as principal components) and deep neural networks are used for classification (45), but generally difficult to interpret. A gradient-weighted class activation mapping (Grad-CAM) can be performed to highlight which regions contribute the most to the classification (61).

2.2 Data augmentation strategies

Deep neural networks require a large number of cases for training. Many data augmentation strategies were proposed for image analysis such as flipping, color space, translation, rotation, noise injection, image mixing, random cropping and generative adversarial networks (62–64). Different from images, the intensity of the Raman spectrum is highly dependent on the Raman shift (wavenumbers). Therefore, different data augmentation strategies are needed for one-dimensional spectral analysis. In previous studies, a number of data augmentation strategies for spectral analysis was proposed, including adding random noise (52, 53, 65), spectral shift (52, 53, 65), spectral superimposition (spectral linear combination) (52, 53, 65), offset (66), adding a slope (66), multiplication (66) and generative adversarial networks (GAN) (67). However, not all the data augmentation strategies were applicable for Raman

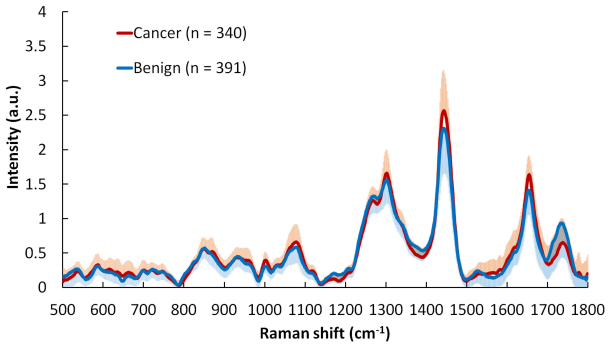

FIGURE 1
Averaged Raman spectra (and standard deviation) of malignant (n=340, including melanoma, basal cell carcinoma, squamous cell carcinoma and actinic keratosis) and benign skin lesions (n=391, including benign nevi and seborrheic keratosis). All the spectra were normalized to their respective areas under the curve between 500 and 1800 cm^{-1} before being averaged. For clarity, standard deviation is shown top half for cancer and bottom half for benign lesions.

spectroscopy. In this study, we proposed the following strategies for data augmentation of Raman spectra, including adding random noise, spectral shift, spectral linear combination, and artificially synthesized spectrum through generative adversarial networks. Note that data augmentation is conducted only for the training dataset.

2.2.1 Data augmentation by addition of random noise

The training dataset can be augmented by adding random noise of different noise levels (Figure 2A), which can be written as

$$S'(v_i) = S(v_i) + N(v_i) \quad (1)$$

Where $S'(v_i)$ is the augmented spectrum, $S(v_i)$ is the original spectrum, and $N(v_i)$ is the random noise. The random noise level is defined as the amplitude of the noise over the maximum peak intensity (I_{max}) of the full training dataset. For example, a k percent of random noise is defined as k percent of the peak intensity, written as $N(v_i) = 2 \times (rand(\ ) - 0.5) \times k / 100 \times I_{max}$, where $rand(\ )$ is a random generator that produces uniformly distributed random numbers within the interval of (0,1).

2.2.2 Data augmentation by spectral shift

The training dataset can be augmented by shifting the spectrum a few pixels (or wavenumbers) (Figure 2B), which can be written as

$$S'(v_i) = S(v_{i \pm m}) + N(v_i) \quad (2)$$

Where $S'(v_i)$ the augmented spectrum at wavenumber v_i, $S(v_{i \pm m})$ is the original spectra at wavenumber $v_{i \pm m}$, m=*1, 2, 3, …* and $N(v_i)$ is the random noise at wavenumber v_i, generated by the same formula as shown in section 2.2.1.

2.2.3 Data augmentation by spectral linear combination

The training dataset can also be augmented by linearly combining two or more sets of spectra. In this study, we implemented data augmentation by linearly combining two sets of spectra (Figure 2C), which can be written as

$$S'(v_i) = rS_1(v_i) + (1 - r)S_2(v_i) + N(v_i) \quad (3)$$

Where $S'(v_i)$ is the augmented spectrum at wavenumber v_i, $S_1(v_i)$ and $S_2(v_i)$ are the two sets of randomly selected original spectra from the training dataset. r is a randomly generated number that is uniformly distributed between 0 and 1, representing the ratio of the two sets of the original spectra. Note that $S_1(v_i)$ and $S_2(v_i)$ are randomly chosen from either the cancer group or the benign group. No attempt is tried to combine the spectra of one from the cancer and the other from the benign groups.

2.2.4 Data augmentation by generative adversarial networks

Another way for data augmentation is using generative adversarial networks, which is far more complicated than the above simple data augmentation techniques. We designed a one-dimensional conditional generative adversarial network (1D-GAN) for Raman spectral generation as shown in Figure 3. It takes the general architecture of a conditional generative adversarial network, which contains two separate networks: a generator and a discriminator. The

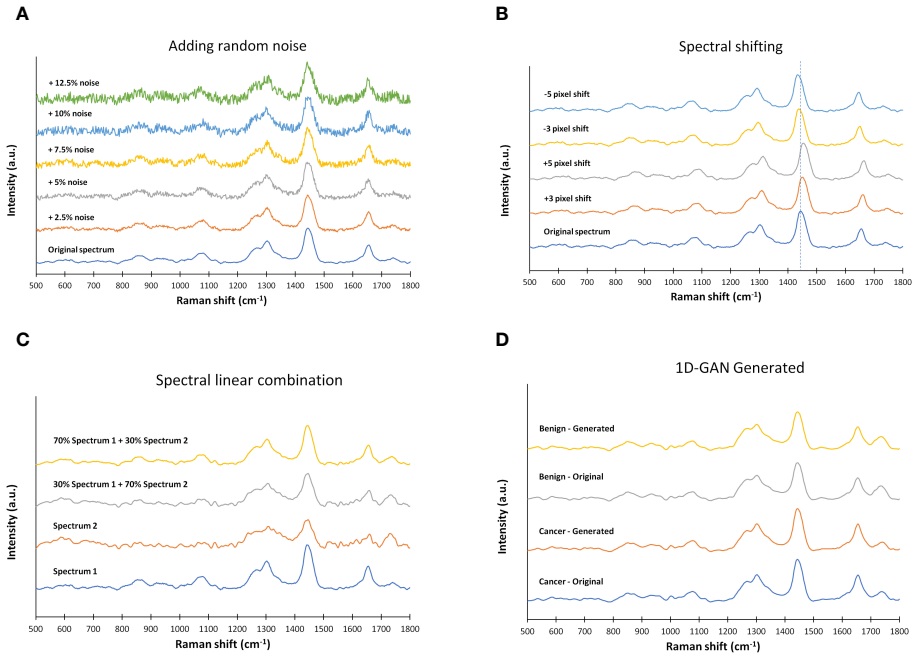

FIGURE 2
Examples of data augmentation for the training dataset of Raman spectra. **(A)** adding random noise of different noise levels, **(B)** spectral shifting, **(C)** spectral linear combination, and **(D)** data augmentation by one dimensional generative adversarial networks (1D-GAN) (averaged spectra are shown).

generator takes random noise and the label (cancer or benign) as input, and generates a synthetic spectrum. The discriminator takes the synthetic spectrum and label, and the real Raman spectra and labels as input, and tries to discriminate the synthetic spectrum from the real spectra. If the discriminator can separate the synthetic spectrum from the real spectra, it will provide a feedback to the generator to modify the parameters in such a way that the synthetic spectrum looks more like a real spectrum (decreasing the loss function). This process is iterative and eventually the generated spectrum looks so close to the real spectrum that the discriminator could not separate the synthetic spectrum from the real spectra. Once the 1D-GAN is trained, the generator can be used to generate arbitrary number of synthetic Raman spectra based on random input (67). The discriminator could sometimes be used directly for image and spectral discrimination purpose as well (68, 69).

The architectures of the generator and discriminator are shown in Figure 3 and the parameters are listed in Table 1. The generator contains 5 transposed convolutional layers (TransConvolution). Each of the first 4 transposed convolutional layers is followed by a batch normalization layer and a regularization layer (ReLu). The output size of each layer is governed by $o = (i\text{-}1)\ s - 2p + k$, where o is the output size, i is the input size, s is the stride, p is the padding size, and k is the kernel size (70). Assuming the input size is 4, the output size of the generator is 619, which is the length of the spectrum in this study within the range of 500 – 1800 cm^{-1}.

The discriminator also contains 5 convolutional layers. Each of the first 4 convolutional layers was followed by a regularization layer (LeakyReLu). The output size of each layer is governed by $o = (i + 2p - k)/s + 1$, where o is the output size, i is the input size, s is the stride, p is the padding size, and k is the kernel size (70). The output size of the discriminator is 1, indicating the input spectrum is either real or synthesized after the discriminator.

The mini batch size was 256. The initial learning rate was 0.0002. The gradient decay factor and the squared gradient decay factor was 0.5 and 0.999 respectively. The total number of epochs was 25,000. With such parameters for the above 1D-GAN architecture, it took about 4.5 hours to complete the training using a mainframe GPU (Advanced Research Computing, University of British Columbia, Sockeye high-performance computing platform). After the 1D-GAN was trained, 5,000 spectra were generated for skin cancers and 5,000 spectra were generated for benign lesions. The average of the 1D-GAN generated spectra and the average of real spectra were shown in Figure 2D.

2.3 One dimensional convolutional neural networks for spectral classification

We developed and tested a number of 1D-CNN architectures for Raman spectral classification, including different number of convolutional layers (1–5); number of kernels (16, 32, 64, 128) for each convolutional layer; kernel sizes (3, 5, 7, 9); mini-batch sizes (16, 32, 64, 128, 256); pooling methods (max pooling and average pooling); and sizes of the fully connected layers (128, 256, 512) (Supplementary Tables S1–S4). The final architecture of the designed 1D-CNN for Raman spectral classification that provided the best performance contained an input layer, 4 convolutional layers, 2 fully connected layers, a softmax layer and an output layer (Figure 4). Each of the four convolutional layers was followed by a batch normalization layer, a regularization layer (ReLu) and an average pooling layer. In total, the 1D-CNN had 21 layers. Note that each layer represents a specific data manipulation. The four convolutional layers had the same kernel (filter) size, padding and stride (kernel size = [3, 1], padding = 'same', and stride = 1), but with different number of kernels (16, 32, 64, and 128 respectively). Zero padding was added to each convolutional layer (padding ='same') so that the output of each convolutional layer had the same size as the input. The size of batch normalization layer was 256 for the original training dataset and 1024 for the augmented training dataset. The four average pooling layers had the same parameters (size = [2,1], stride= [2,1]) so that after each pooling layer the size was reduced by half. The training process was optimized by adaptive moment estimation (adam) (71). The initial learning rate was 0.001, which was dropped by a factor of 0.9 for every 2 period. The training process was monitored through the accuracy and the loss function (cross entropy) of the training and validation datasets.

An example of the training process, including the accuracy and the loss function of the training and validation processes are shown in

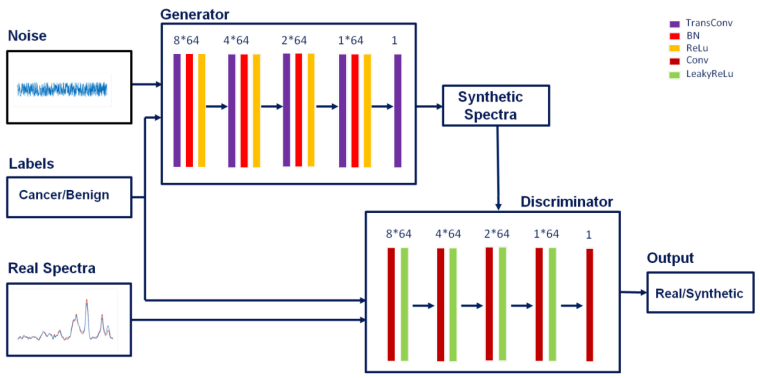

FIGURE 3
One-dimensional generative adversarial networks (1D-GAN) for data augmentation. 8*64, 4*64, 2*64, 1*64 and 1 are the number of kernels of each convolutional layer (and transposed convolutional layer for the generator).

TABLE 1 Parameters for the generator and discriminator of the one-dimensional generative adversarial networks (1D-GAN) for data augmentation.

Network	Input size	Layer number	Kernel size	Number of kernels	Stride	Cropping	Output size
Generator	4	1	[3,1]	8 * 64	[1,1]	[0,0]	6
	6	2	[7,1]	4 * 64	[4,1]	[1,0]	25
	25	3	[7,1]	2 * 64	[3,1]	[1,0]	77
	77	4	[7,1]	1 * 64	[4,1]	[1,0]	309
	309	5	[5,1]	1	[2,1]	[1,0]	619
Discriminator	619	1	[5,1]	8 * 64	[2,1]	[1,0]	309
	309	2	[7,1]	4 * 64	[4,1]	[1,0]	77
	77	3	[7,1]	2 * 64	[3,1]	[1,0]	25
	25	4	[7,1]	1 * 64	[4,1]	[1,0]	6
	6	5	[6,1]	1	[1,1]	[0,0]	1

The input size of the generator is designed as 4 and the output size of the generator is 619.

Figure 5. Here the original training dataset was used, which contained 512 cases. The mini-batch size was 256, and the maximum number of epochs was 100. Therefore, there were 2 iterations per epoch and a maximum of 200 iterations per training process in this example. The accuracy for the training dataset was defined as the ratio of the correctly classified cases to the total cases of the training dataset in the mini-batch (n=256). The accuracy for the validation dataset was defined as the ratio of the correctly classified cases to the total cases of the full validation dataset (n=73). It was noted that the accuracy of the model was improving at the beginning after the network training process started. However, the model may be over-trained if the training process could not be terminated at appropriate training stage. To prevent over-training, we implemented a strategy to stop the training process if the accuracy of the validation dataset was not improving for specific iterations. Usually the number of iterations for terminating the training process could be set between 10 and 50. If the number was too small, the trained model might be premature; while if the number was too large, the model might be over-trained. Figure 5A shows that the model was terminated at 76 iterations (as shown by the arrow) because the accuracy of validation dataset was not improving for 50 iterations. The parameters of last training process were used as the parameters for the final model. Figure 5B shows the cross entropy loss of the training and validation process. Similarly to the accuracy, the cross entropy loss was calculated on mini-batch for the training process and on the full validation dataset for the validation process. It could be seen that the loss was initially decreased and then was leveled off if the model kept training. The arrow indicated a possible stopping stage based on the strategy for the training process to prevent over-training. In our experiment, we terminated the training process if the accuracy for the validation dataset was not improving for 50 iterations.

2.4 Conventional machine learning approaches

Conventional machine learning approaches are sometimes called chemometrics or multivariate statistical analyses (48, 72).

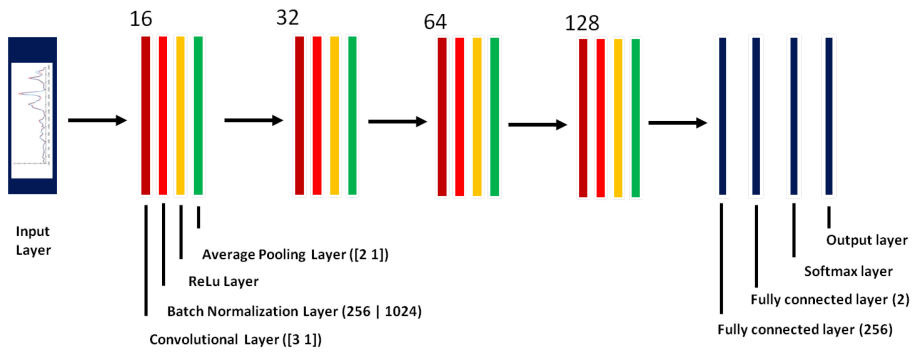

FIGURE 4
Architecture and parameters of one-dimensional convolutional neural networks (1D-CNN) for Raman spectral classification. The number of kernels for each convolutional layer was 16, 32, 64 and 128 with kernel size = [3,1] for all the convolutional layers. The mini-batch size was 256 for the training dataset without augmentation and 1024 for the training dataset with augmentation. The pooling size was [2,1] with stride [2,1] for each average pooling layer. The size for the first fully connected layer was 256 and for the second fully connected layer was 2.

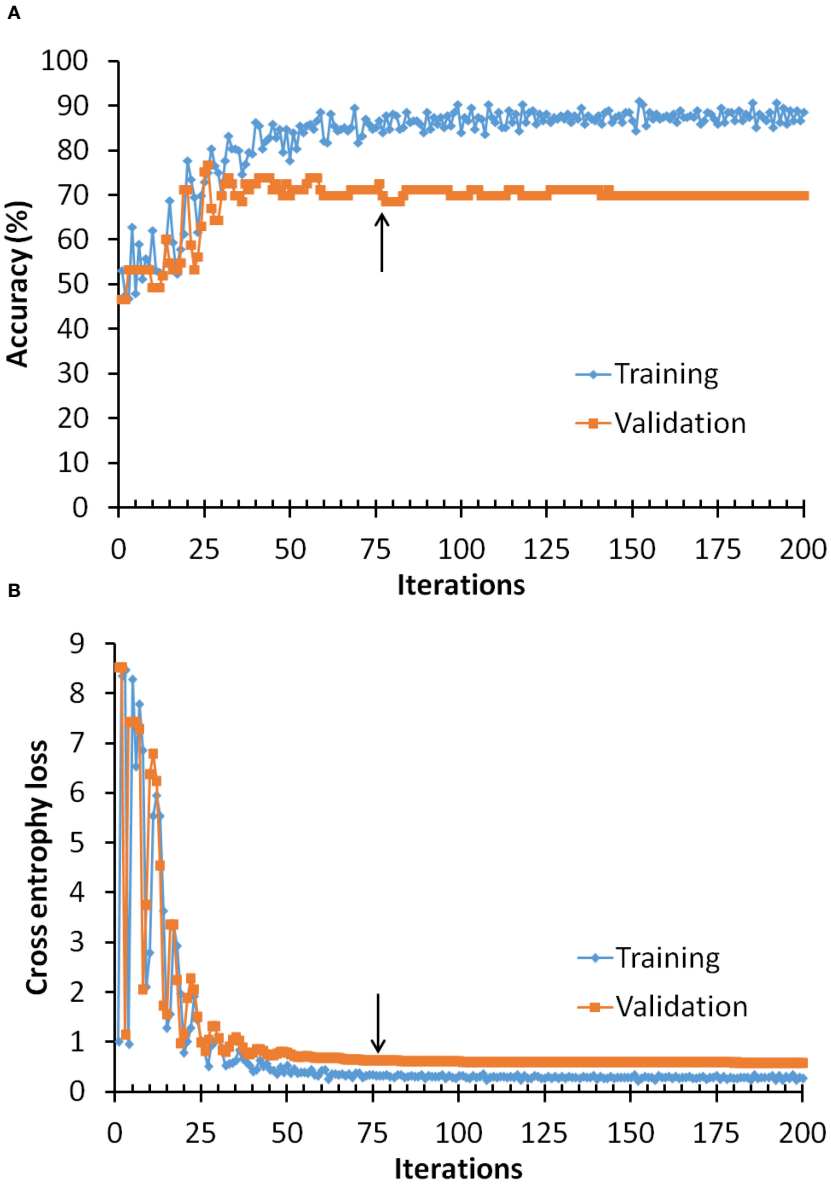

FIGURE 5
Example of the training process of the 1D-CNN for Raman spectral classification. Arrows showed the performance of the validation process no longer improving over at least 50 iterations, a possible stopping stage to prevent over-training. **(A)** Accuracy of the training and validation process. **(B)** Cross entropy loss of the training and validation process.

For comparison purpose, conventional machine learning techniques, including partial least squares for discriminant analysis (PLS-DA), principal component and linear discriminant analysis (PC-LDA), support vector machine (SVM), and logistic regression (LR) were implemented.

2.5 Dataset split and data augmentation

The dataset split and data augmentation procedures were illustratively shown in Figure 6. For analyses without data augmentation, the stratified original dataset (n=731) was randomly divided into training (70%, i.e. 70% of cancerous and 70% of benign cases), validation (10%) and test set (20%). For analyses with data augmentation, only the training set was augmented. Data augmentation of the training set was implemented after random split of the original dataset. Therefore, the cases in the training, validation and test datasets were the same as the analyses without data augmentation. The same data split scheme and augmentation were used for the 1D-CNN and all the conventional machine learning analyses. All the above models including 1D-CNN, PLS-DA, PC-LDA, SVM and LR were implemented using parallel computing on UBC ARC (Advanced Research Computing, University of British Columbia) Sockeye high-performance computing platform. The random split was repeated in parallel 56 times and the mean was reported (parallel computing requires multiple of 8). All the programs were implemented using Matlab (Version 2021a, Mathworks, Natick, MA, USA).

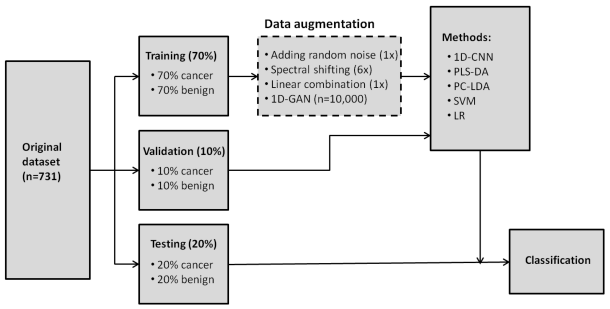

FIGURE 6
Random split of original dataset into training (70%), validation (10%) and testing (20%) datasets with and without data agumentation (dashed frame). Note that for analyses with augmentation, it was implemented only to the training dataset, not the validation and testing datasets. 1D-CNN, one-dimensional convolutional neural networks; PLS-DA, partial least squares for discriminant analysis; PC-LDA, pricipal component and linear discriminant analysis; SVM, support vector machine; LR, logistic regression.

2.6 Extended test dataset

In order to evaluate the models for situations that are slightly out of the original scope, such as lower spectral quality, we introduced random noise and spectral shift to the test dataset, hereafter referred as extended test dataset, in addition to the original test dataset. Similar to sections 2.2.1 and 2.2.2, up to 12.5% of random noise and 6-pixel shift were applied to the original test dataset to generate the extended test dataset.

2.7 Statistical analysis

Paired analysis (Wilcoxon test) of the test set between 1D-CNN and PLS-DA, PC-LDA, SVM and LR, and the test of the above models between using original and augmented training set were performed (GraphPad, Boston, MA, USA). A p-value of less than 0.05 ($p<0.05$) was regarded as statistically significant.

3 Results

3.1 Evaluation of augmentation parameters

We first evaluated the optimal parameters for data augmentation, such as the level of random noise, the range of the spectral shift, the number of linearly combined spectra, and the number of synthesized spectra generated by the generative adversarial networks.

3.1.1 Level of random noise

The hypothesis for data augmentation by adding random noise is that the augmented spectra are measured by systems of different signal to noise ratios. We evaluated data augmentation by adding different levels of random noise from 1% to 12.5% following Equation 1 (Figure 7A). As expected, the performances of 1D-CNN, PLS-DA and PC-LDA were immune to noise levels. Surprisingly, the performances of SVM and LR were all improved. The performance of SVM after augmentation with high level of noise (i.e. >7.5%) was even better than any other techniques. Because the number of cases was less than the number of variables for the original training set, data augmentation by adding random noise improved the performance of LR. However, data augmentation by solely adding random noise could not solve the number of case versus variable issues (full rank issues). Its performance was still the lowest compared with other techniques.

3.1.2 Range of spectral shifting

The hypothesis for data augmentation by spectral shifting is that the augmented spectra are measured from different systems of variable qualities. We evaluated data augmentation by spectral shifting of 1 to 6 pixels following Equation 2 without addition of random noise (Figure 7B). It was found that the performance of 1D-CNN, PLS-DA and PC-LDA were independent of spectral shifting, while the performance of SVM was decreased monotonically. Surprisingly, it was found that data augmentation by spectral shifting was particularly useful for LR. Its performance was equivalent to 1D-CNN, PLS-DA and PC-LDA, and even better than SVM after data augmentation with large spectral shifting.

3.1.3 Number of cases by spectral linear combination

The hypothesis for data augmentation by spectral linear combination is that the augmented spectrum is measured from a lesion that mimics two measured lesions. We evaluated data augmentation by spectral linear combination in multiples of the training dataset following Equation 3 (Figure 7C). It was found that the performance of spectral linear combination was the worst compared to data augmentation by adding random noise or spectral shifting, indicating that it was very unlikely that a lesion would have the properties of two measured lesions. Data augmentation by solely spectral linear combination did not improve the performance of 1D-CNN, PLS-DA, PC-LDA, SVM or LR.

3.1.4 Number of synthesized spectra by 1D-GAN

The hypothesis of data augmentation by 1D-GAN is that the properties of a lesion that is not in the original dataset can be synthesized. The beauty of 1D-GAN is that once it is trained, it can be used to generate any number of synthesized spectra. To determine the optimal number of synthesized spectra by 1D-GAN, models with the original spectra and synthesized spectra of n=500, 1,000, 2,000, 5,000, 10,000, 20,000 and 30,000 were evaluated (Figure 7D). It was found that the performance of 1D-CNN, PLS-DA, PC-LDA, SVM and LR were all improved with augmentation by 1D-GAN generated spectra. However, PLS-DA, PC-LDA and SVM were not dependent on the number of synthesized cases. 1D-CNN was slightly improved with the number of synthesized spectra until it reached plateau at around n=10,000, while for LR no matter how many cases were synthesized, it was not sufficient.

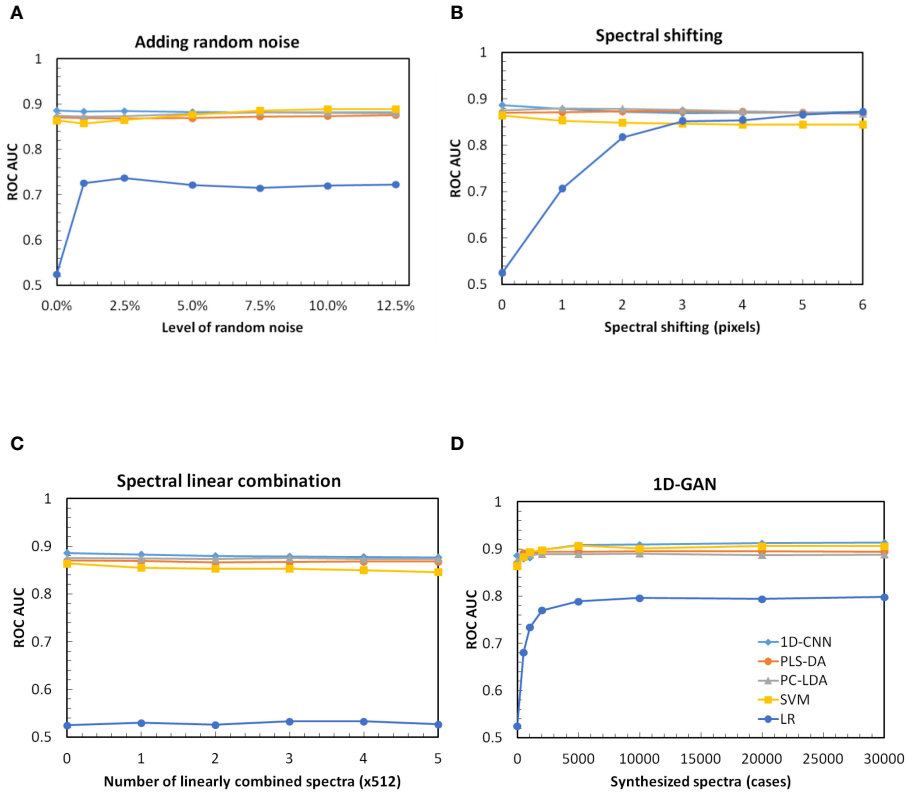

FIGURE 7
ROC AUC of the test dataset of 56 random repetitions based on the original training dataset and different augmentation parameters, **(A)** adding random noise, **(B)** spectral shifting, **(C)** spectral linear combination, and **(D)** synthesized spectra by 1D-GAN. 1D-CNN, one-dimensional convolutional neural networks; PLS-DA, partial least squares for discriminant analysis; PC-LDA, principal component and linear discriminant analysis; SVM, support vector machine; LR, logistic regression.

3.1.5 Optimal parameters for data augmentation

Based on the above evaluation, the following parameters were selected for data augmentation of the training set in this study: random noise level of 5% (n=512), spectral shift of 1–3 pixels (n=512x6), spectral linear combination (n=512) and 1D-GAN synthesized spectra (n=10,000). To prevent collinearity, a 5% random noise was applied to the spectrally shifted and linearly combined spectra.

3.2 Diagnosis based on original training dataset without data augmentation

The ROC curves for the training, validation and test dataset based on 1D-CNN, PLS-DA, PC-LDA, SVM and LR for one of the 56 random splits are shown in Figure 8 (top row). It could be seen that the performance of the training set were always better than the validation and the test sets, indicating that the models were slightly over-trained, particularly for SVM. LR failed based on the original training set in this example, because the number of variables (619) was more than the number of cases (512).

The ROC AUCs of the 56 calculations for the original test datasets based on the original training datasets without augmentation were shown in Figure 9. The training set had higher ROC AUCs than the validation and test datasets on average (Table 2). The averaged ROC AUCs of the original *test dataset* based on the original training spectra without augmentation were 0.886±0.022 (1D-CNN), 0.870±0.028 (PLS-DA), 0.875±0.033 (PC-LDA), 0.864±0.027 (SVM), and 0.525±0.045 (LR), respectively. Paired analyses showed that 1D-CNN outperformed conventional machine learning approaches by 1–3% including PLS-DA, PC-LDA and SVM based on the original spectra (Wilcoxon, $p<0.001$). 1D-CNN also had the smallest standard deviation of the ROC AUCs.

3.3 Diagnosis based on original training dataset with augmentation

There were 14,608 cases in the training dataset after augmentation, consisting of the original spectra (n=512) plus spectra that were augmented by adding random noise (5% noise, n=512), spectral shifting (1–3 pixel shifting, n = 512 x 6), combining spectra linearly (n=512), and 1D-GAN (n=10,000). The ROC curves based on same random split of the original dataset after augmentation for 1D-CNN, PLS-DA, PC-LDA, SVM and LR are shown in Figure 8 (bottom row). It showed that all the models for the training dataset were also over-trained, particularly for SVM. Surprisingly, it was found that after data augmentation, LR performed very well since the number of cases was now larger

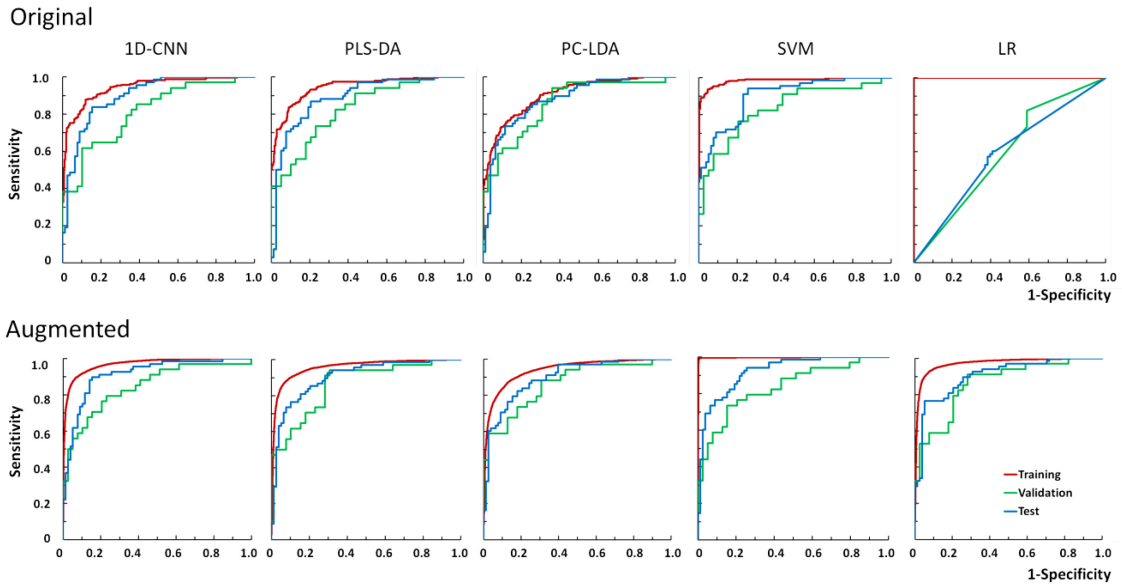

FIGURE 8
Example of the ROC curves of the training, validation (n=73) and test (n=146) datasets. Top row is based on the original training spectra (n=512), and bottom row is based on the augmented spectra (n=14,608). 1D-CNN, one-dimensional convolutional neural networks; PLS-DA, partial least squares for discriminant analysis; PC-LDA, principal component and linear discriminant analysis; SVM, support vector machine; LR, logistic regression.

than the number of variables. As expected data augmentation also resulted in smoother training ROCs.

The ROC AUCs of the 56 calculations for the original test datasets based on the original training datasets with augmentation are shown in Figure 9. The ROC AUCs of the training dataset were much higher than the validation and test datasets (Table 2).The ROC AUCs of the *original test dataset* based on the original training dataset with augmentation were 0.909±0.021 (1D-CNN), 0.899±0.022 (PLS-DA), 0.895±0.022 (PC-LDA), 0.901±0.020 (SVM), and 0.897±0.021 (LR). It showed that after data augmentation, 1D-CNN slightly out-performed all the conventional machine learning techniques by 1–2%, including PLS-DA, PC-LDA, SVM and LR ($p<0.001$, Wilcoxon test).

Because only the training datasets were augmented, the test datasets were the same for 1D-CNN and conventional machine learning approaches. Paired analyses demonstrated that models based on the original training datasets with augmentation (n=14,608) significantly improved the diagnostic ROC AUCs of the original test datasets by 2–4% compared with models based on the original training datasets without augmentation (n=512) for both 1D-CNN and conventional machine learning methods (PLS-DA, PC-LDA, and SVM) ($p<0.0001$, Wilcoxon test). Augmentation was particularly useful for LR when the number of cases was smaller than the number of variables, which was improved by 71%.

3.4 Diagnosis based on different augmentation strategies

We also calculated the performance of models based on different augmentation strategies to the training datasets. The split scheme of the original spectra was the same as section 3.2 and section 3.3. Validation and test datasets were not augmented. All the models were repeated 56 times.

3.4.1 Augmented spectra without original

When 1D-CNN and the conventional machine learning approaches were trained on the augmented spectra without the original training datasets, here data augmentation included adding random noise (n=512), spectral shifting (n = 512 x 6), combining spectra linearly (n=512), and 1D-GAN (n=10,000), the ROC AUCs of the *original test datasets* were found to be 0.908±0.021 (1D-CNN), 0.899±0.022 (PLS-DA), 0.893±0.023 (PC-LDA), 0.906±0.021 (SVM), and 0.897±0.022 (LR) (Table 2), almost identical to the results based on the original training datasets with augmentation (section 3.3), indicating that the contribution of the original training datasets may be negligible after data augmentation.

3.4.2 Augmented spectra by 1D-GAN only

When 1D-CNN and the conventional machine learning approaches were trained on the augmented spectra synthesized by 1D-GAN only (n=10,000), the ROC AUCs of the *original test datasets* were found to be 0.907±0.021 (1D-CNN), 0.893±0.023 (PLS-DA), 0.886±0.025 (PC-LDA), 0.905±0.021 (SVM), and 0.813±0.033 (LR) (Table 2). The results were inferior to the models based on the original training datasets with augmentation (section 3.3) and augmented spectra without original (section 3.4.1), but still better than the models based on the original training datasets without augmentation (section 3.2). However, the performance based on augmented spectra by 1D-GAN only was not sufficient for LR, indicating that other augmentation strategies were still needed.

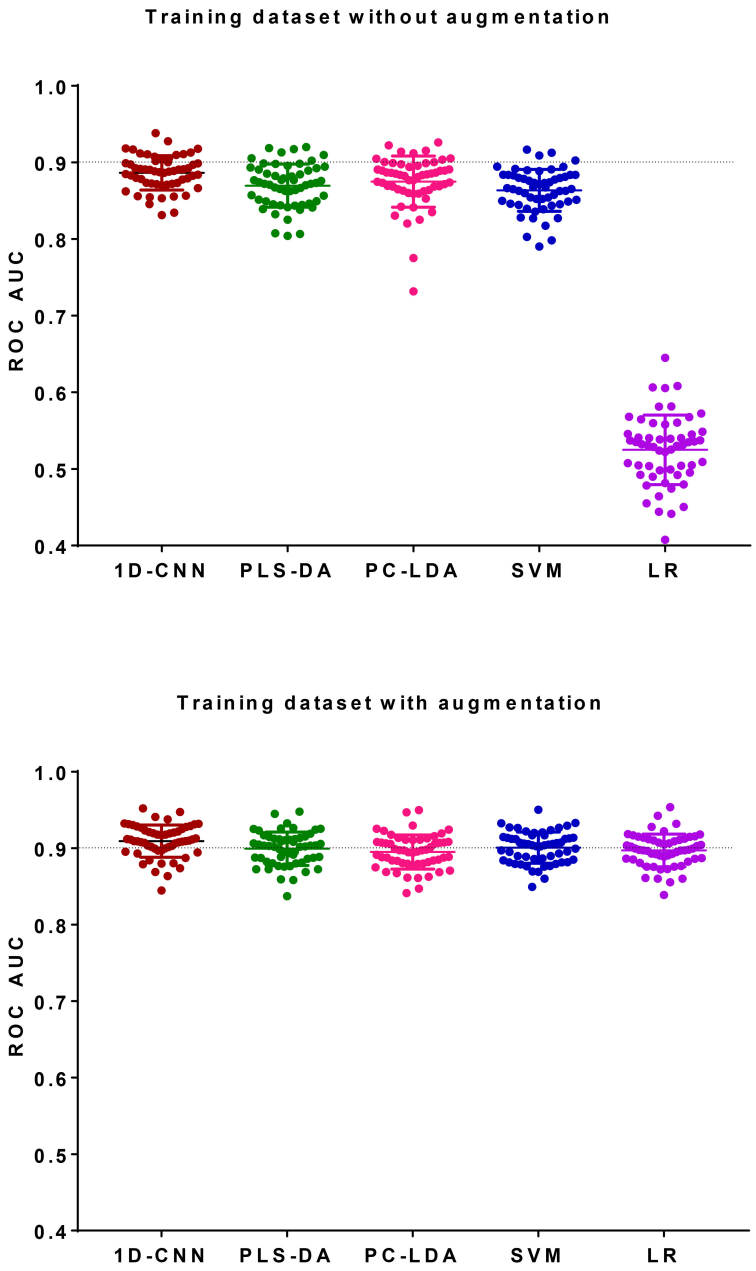

FIGURE 9
ROC AUC of the test dataset of 56 random repetitions based on the original training dataset without augmentation and the original training dataset with augmentation. Bar shows the mean and standard deviation.

3.4.3 Original and augmentation without 1D-GAN

When the models were trained on the original training datasets with augmentation by adding random noise, spectral shifting and linear combination without 1D-GAN (n=4,608), the ROC AUCs of the *original test datasets* were found to be 0.879±0.022 (1D-CNN), 0.884±0.023 (PLS-DA), 0.880±0.026 (PC-LDA), 0.877 ±0.024 (SVM), and 0.875±0.022 (LR) (Table 2). The results indicated that data augmentation by adding random noise, spectral shifting and linear combination without 1D-GAN worked well for conventional machine learning techniques, particularly LR, but not 1D-CNN.

3.4.4 Augmentation without 1D-GAN

When the models were trained on the augmented datasets (by adding random noise, spectral shifting and linear combination without 1D-GAN) without the original training datasets (n=4,096), the ROC AUCs of the *original test datasets* were found to be 0.876±0.022 (1D-CNN), 0.884±0.023 (PLS-DA), 0.880±0.026

TABLE 2 ROC AUCs (mean ± standard deviation) of the training, validation (n=73) and test dataset (n=146) of 56 random split schemes based on models of different training datasets.

Training Dataset	Dataset	1D-CNN	PLS-DA	PC-LDA	SVM	LR
Original without augmentation	Training	0.941±0.005	0.911±0.030	0.910±0.025	0.980±0.006	0.999±0.001
	Validation	0.882±0.041	0.889±0.039	0.900±0.036	0.853±0.043	0.541±0.058
	Test	**0.886±0.022**	**0.870±0.028**	**0.875±0.033**	**0.864±0.027**	**0.525±0.045**
Original with augmentation	Training	0.989±0.005	0.961±0.002	0.949±0.011	0.999±0.000	0.980±0.002
	Validation	0.903±0.042	0.895±0.036	0.909±0.035	0.893±0.039	0.889±0.037
	Test	**0.909±0.021**	**0.899±0.022**	**0.895±0.022**	**0.901±0.020**	**0.897±0.021**
Augmented spectra without original	Training	0.989±0.005	0.963±0.002	0.949±0.015	0.999±0.000	0.981±0.002
	Validation	0.902±0.040	0.895±0.037	0.909±0.035	0.897±0.039	0.889±0.037
	Test	**0.908±0.021**	**0.899±0.022**	**0.893±0.023**	**0.906±0.021**	**0.897±0.022**
Augmented spectra by 1D-GAN only	Training	1.000±0.001	0.993±0.000	0.980±0.021	1.000±0.000	1.000±0.000
	Validation	0.902±0.038	0.890±0.040	0.912±0.033	0.903±0.034	0.810±0.047
	Test	**0.907±0.021**	**0.893±0.023**	**0.886±0.025**	**0.905±0.021**	**0.813±0.033**
Original and augmentation without 1D-GAN	Training	0.946±0.010	0.931±0.007	0.905±0.016	0.995±0.001	0.955±0.006
	Validation	0.876±0.043	0.881±0.039	0.895±0.037	0.866±0.043	0.870±0.041
	Test	**0.879±0.022**	**0.884±0.023**	**0.880±0.026**	**0.877±0.024**	**0.875±0.022**
Augmentation without 1D-GAN	Training	0.947±0.011	0.931±0.007	0.905±0.017	0.992±0.002	0.956±0.006
	Validation	0.873±0.043	0.881±0.039	0.894±0.038	0.872±0.043	0.871±0.041
	Test	**0.876±0.022**	**0.884±0.023**	**0.880±0.026**	**0.880±0.023**	**0.876±0.022**

Original without augmentation: original training spectra without augmentation (n=512); Original with augmentation: original training spectra with augmentation including adding random noise, spectral shifting, spectral linear combination and 1D-GAN (n=14,608); Augmented spectra without original: augmentation including adding random noise, spectral shifting, spectral linear combination and 1D-GAN but without original spectra (n=14,096); Augmented spectra by 1D-GAN only: augmented training spectra by 1D-GAN only (n=10,000); Original and augmentation without 1D-GAN (n=4608): Original training spectra with augmentation including adding random noise, spectral shifting, spectral linear combination but without 1D-GAN; Augmentation without 1D-GAN (n=4096): augmented spectra including adding random noise, spectral shifting, spectral linear combination but without 1D-GAN. Boldface highlights the results of the test dataset.

(PC-LDA), 0.880±0.023 (SVM), and 0.876±0.022 (LR) (Table 2). These results were quite similar to the models trained on the original and augmentation without 1D-GAN (section 3.4.3), indicating that data augmentation by adding random noise, spectral shifting and linear combination without 1D-GAN covers the original training datasets. It is interesting to note that data augmentation without 1D-GAN worked better for LR than augmented spectra by 1D-GAN only ($p<0.0001$, Wilcoxon); while it was the opposite for 1D-CNN, PLS-DA, PC-LDA and SVM where augmented spectra by 1D-GAN only worked better ($p<0.0001$, Wilcoxon) (section 3.4.2).

3.5 Diagnosis on extended test dataset

3.5.1 Models based on the original training dataset without augmentation

When applying the models based on the original training dataset without augmentation (models in section 3.2) to the extended test dataset, the ROC AUCs of the extended test datasets are shown in Figure 10 (top row). It is found that all the models including 1D-CNN, PLS-DA, PC-LDA and SVM perform the best on the original test dataset. They all can tolerate 2.5% of random noise or ±1 pixel shift; and the performance starts to deteriorate with further added noise or spectral shift. In terms of random noise, PC-LDA and PLS-DA perform better than 1D-CNN and SVM at high noise levels; while in terms of spectral shifting, 1D-CNN performs better than PC-LDA, PLS-DA and SVM across all the situations. It is also noticed that the effect for spectral shifting is not symmetric. The performance drops faster when the spectrum is shifted to lower wavenumbers.

3.5.2 Models based on the original training dataset with augmentation

When applying the models based on the original training dataset with augmentation (models in section 3.3) to the extended test dataset, the ROC AUCs of the extended test datasets are shown in Figure 10 (bottom row). Again, it is found that all the models including 1D-CNN, PLS-DA, PC-LDA and SVM perform the best on the original test dataset. The performance for models trained on the augmented training dataset was much better than the models trained on the original training dataset (Supplementary Figures S1, S2). In terms of random noise, PC-LDA and PLS-DA still perform better than 1D-CNN and SVM at high noise levels; while in terms of spectral shifting, 1D-CNN and SVM perform better than PC-LDA and PLS-DA. 1D-CNN still performs the best across all the situations.

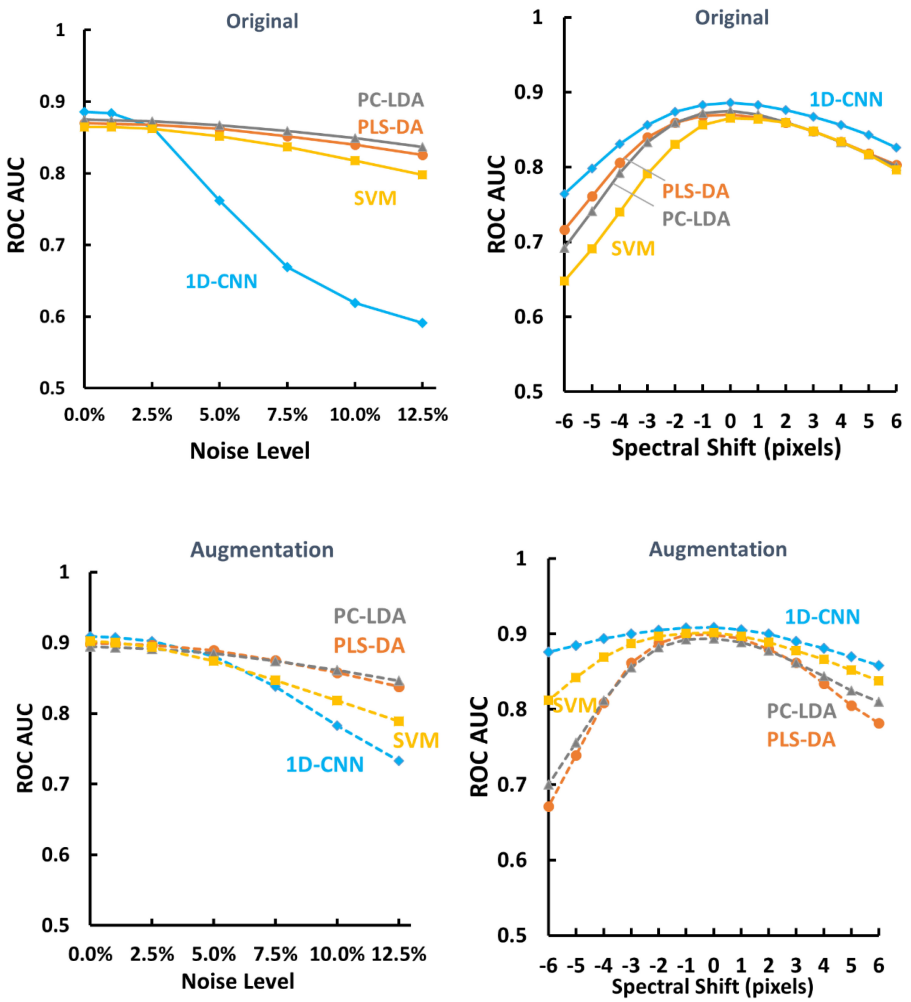

FIGURE 10
ROC AUC of the extended test dataset by adding random noise or spectral shift to the original test dataset for models based on the original training dataset without augmentation (top row) and models based on the original training dataset with augmentation (bottom row). Data shown are the mean of 56 random repetitions using parallel computing. 1D-CNN, one-dimensional convolutional neural networks; PLS-DA, partial least squares for discriminant analysis; PC-LDA, principal component and linear discriminant analysis; SVM, support vector machine.

4 Discussions

4.1 Data augmentation improved diagnostic performance of 1D-CNN and conventional machine learning techniques

Adding random noise, spectral shifting, linear combination of the spectra (65), and generative adversarial networks (67) were proposed in previous studies of deep neural networks, and none of them implemented all data augmentation strategies presented in this study. We applied all the data augmentation strategies to both deep neural networks and conventional machine learning techniques, and found that data augmentation not only improved the performance of deep neural networks, but also the conventional machine learning techniques. Particularly, data augmentation improved the performance of LR when the number of cases was less than the number of variables.

None of the previous studies provided details of how the data augmentation was implemented or evaluated the effect of different data augmentation strategies. In this study, we systematically investigated the effect of different data augmentation strategies and found that different augmentation strategies have variable contribution to the improvement of deep neural networks and conventional machine learning techniques. For example, augmentation by adding random noise improved SVM and LR, but not PLS-DA, PC-LDA and 1D-CNN; augmentation by spectral shifting improved LR but not PLS-DA, PC-LDA, SVM and1D-CNN; augmentation by spectral linear combination has almost no contribution to all the techniques (Figure 7); and augmentation by 1D-GAN improved the performance of 1D-CNN, PLS-DA, PC-LDA and SVM more than LR (Table 2). The best performance was achieved when all these strategies were combined (Figure 9).When all the augmentation strategies were combined, it improved the performance of 1D-CNN, PLS-DA, PC-LDA and SVM by 2–4%, and LR by 71%.

4.2 1D-CNN outperforms conventional machine learning approaches

Convolutional neural networks contain multiple hierarchical layers with each layer representing a specific data manipulation. It contains many parameters to train and thus very versatile. The versatility of convolutional neural networks brings both advantage and disadvantage, in that an optimal model can always be found by tuning the parameters, while in the meantime there is no standard architecture that can fit all situations. For example, in this study we tried 1D-CNN architectures of multiple convolutional layers with different number of kernels, kernel sizes, pooling methods, mini-batch sizes and sizes of the fully connected layer (section 2.3, Supplementary Tables S1-S4), and eventually found that 1D-CNN with 4 convolutional layers, with each convolutional layer having 16, 32, 64 and 128 kernels, kernel size = [3,1], mini-batch size = 256, and average pooling had the best performance (Figure 4). With the optimal 1D-CNN architecture and parameters, we found that 1D-CNN out-performed other machine learning techniques by 1–3% based on the original Raman spectra. After data augmentation, 1D-CNN outperformed other machine learning techniques by 1–2% (Figure 9).

Different from deep neural networks, conventional machine learning techniques were not as versatile as 1D-CNN, and thus there were less parameters to train. Therefore, in designing CNN, the performance of conventional machine learning techniques could be used as baseline for benchmarking convolutional neural networks. 1D-CNN outperforms conventional machine learning techniques, with the cost of longer training and more effort to find the optimal network architectures and parameters.

4.3 Parallel computing provides advantages for both 1D-CNN and conventional machine learning techniques

Cross validation is commonly used in conventional machine learning techniques (PLS-DA, PC-LDA, SVM and LR). The most commonly used cross validation techniques are leave-one-out cross-validation (LOO-CV) and K-fold cross-validation. For LOO-CV, usually a case or patient is left out for testing and the remaining n-1 cases or patients are used for training. The procedure is repeated n times so that every case or patient is tested once. The results are the average of the n models. Similar to LOO-CV, K-fold cross-validation is to divide the stratified cases or patients into K groups; one group of the cases or patients is left out for testing and the remaining K-1 groups are used for training. The procedure is repeated K times so that every case or patient is tested once. The results are the average of the K models. LOO-CV can be regarded as a special case of K-fold cross-validation where K=n.

Deep neural networks often require the dataset being split into training, validation and test datasets. The model is generated from the training dataset; the hyperparameters are fine-tuned from the validation dataset; and the test set provides unbiased evaluation of the final model. Sometimes the validation dataset can serve as the test dataset when the original dataset is divided into two subsets. If early stopping is needed for model training, an independent validation dataset is usually needed. LOO-CV and K-fold cross-validation have been used in deep neural networks for Raman spectrum classification (52, 54, 65). Some authors implemented K-fold cross-validation on the training dataset and tested on a holdout test set. When the test set is large enough and representative to the distribution of the dataset, it provides an unbiased evaluation of the model (31, 59, 73). However, as Khristoforova et al (74) pointed out, the major drawbacks of previous publications on Raman spectroscopy and chemometrics were insufficient sample size, lack of cross-validation, and/or incorrect division of the data into subsets. In this study, the original dataset was randomly split into training, validation and test datasets, and early stopping criteria were used during the model training process. This process was repeated 56 times, taking the advantage of parallel computing, and thus prevented over-fitting and bias.

4.4 Data augmentation improves model performance on extended test dataset

Although there are standard protocols for measurement and data processing of Raman spectra, including wavelength calibration, intensity calibration, fluorescence background removal, and/or normalization, it is still difficult to evaluate the performance of models across multiple systems (72, 75, 76). In this paper we proposed a simple method to evaluate the models on different conditions by adding random noise or spectral shift to the original test dataset to generate the extended test dataset, mimicking the situation of multiple systems. We evaluated the models trained on the original training dataset (section 3.2) and models trained on the augmented training dataset (section 3.3) to the original test dataset (Figure 9) and the extended test dataset (Figure 10). It was found that the models based on the original training dataset without augmentation performed well only on situations with similar spectral quality (i.e. original test dataset), and the performance became deteriorated when the spectral quality was compromised (i.e. the extended test dataset with increased random noise or spectral shifting) (Figure 10). However, the models based on the augmented training dataset not only improved the performance on the original test dataset (Figures 9, 10), but also had higher tolerance on low spectral quality, i.e. the situations with increased random noise or spectral shifting (Figure 10). The models based on the augmented training dataset could perform even better on the extended test dataset (with up to 5% increase of random noise or 3-pixel spectral shifting) than the models based on the original training dataset applied to the original test dataset (Supplementary Figures S1, S2), indicating that data augmentation could improve the applicability of the models trained on high quality data to situations of low spectral quality.

5 Conclusion

In summary, we designed a one-dimensional convolutional neural networks (1D-CNN) for skin cancer detection by Raman spectroscopy and compared with conventional machine learning

techniques (PLS-DA, PC-LDA, SVM and LR). We proposed and evaluated different data augmentation strategies including adding random noise, spectral shifting, spectral linear combination and synthetic spectra by 1D-GAN. Each augmentation strategy had different performance, but when all the augmentation strategies were combined, it substantially improved the performance of 1D-CNN and all the conventional machine learning analyses by 2–4% ($p<0.0001$, Wilcoxon). We also found that a well-designed 1D-CNN outperformed conventional machine learning techniques by 1–3% using original dataset and by 1–2% after data augmentation. Data augmentation is a simple but an effective way to improve the performance of deep neural networks and machine learning techniques. Models trained on augmented training dataset not only perform better, but also have higher tolerance on spectral quality of the test dataset.

Data availability statement

Data can be made available to researchers upon application and subject to the approval of the Clinical Research Ethics Board of the University of British Columbia. Enquiries regarding data access can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by the Clinical Research Ethics Board of the University of British Columbia. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

JZ: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. HL: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. SK: Investigation, Writing – review & editing. TL: Methodology, Writing – review & editing. HZ: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This project was supported by grants from the Canadian Cancer Society (#011031 and #015053), the Canadian Institutes of Health Research (#PP2-111527), the Canadian Dermatology Foundation, the VGH & UBC Hospital Foundation In It for Life Fund and the BC Hydro Employees Community Service Fund. This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of British Columbia.

Acknowledgments

The authors greatly appreciate the volunteer patients for participating in this study. The authors also acknowledge David McLean, Zhiwei Huang, Iltefat Hamzavi, Abdulmajeed Alajlan, Hana Alkhayat, Ahmad Al Robaee, Wei Zhang, Michelle Zeng, Youwen Zhou, Laurence Warshawski, David Zloty, Bryce Cowan for their contributions in patient recruiting, spectral measurements and data analyses.

Conflict of interest

The authors and the BC Cancer Agency hold several patents for Raman spectroscopy that have been licensed to Vita Imaging Inc San Jose, California.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2024.1320220/full#supplementary-material

References

1. Cancer facts & Figures 2023. Available at: www.cancer.org (Accessed September 13, 2023).

2. Olsen CM, Pandeya N, Green AC, Ragaini BS, Venn AJ, Whiteman DC. Keratinocyte cancer incidence in Australia: a review of population-based incidence trends and estimates of lifetime risk. *Public Health Res Pract.* (2022) 32:1–8. doi: 10.17061/phrp3212203

3. English DR, Del Mar C, Burton RC. Factors influencing the number needed to excise: excision rates of pigmented lesions by general practitioners. *Med J Aust.* (2004) 180:16–9. doi: 10.5694/j.1326-5377.2004.tb05766.x

4. Lui H, Zhao J, McLean D, Zeng H. Real-time Raman spectroscopy for in *vivo* skin cancer diagnosis. *Cancer Res.* (2012) 72:2491–500. doi: 10.1158/0008-5472.CAN-11-4061

5. Zhao J, Zeng H, Kalia S, Lui H. Using Raman spectroscopy to detect and diagnose skin cancer in vivo. *Dermatol Clin*. (2017) 35:495–504. doi: 10.1016/j.det.2017.06.010

6. Zhao J, Zeng H, Kalia S, Lui H. Wavenumber selection based analysis in Raman spectroscopy improves skin cancer diagnostic specificity. *Analyst*. (2016) 141:1034–43. doi: 10.1039/C5AN02073E

7. Zhao J, Zeng H, Kalia S, Lui H. Incorporating patient demographics into Raman spectroscopy algorithm improves *in vivo* skin cancer diagnostic specificity. *Trans Biophotonics*. (2019) 1:e201900016. doi: 10.1002/tbio.201900016

8. Massone C, Di Stefani A, Soyer HP. Dermoscopy for skin cancer detection. *Curr Opin Oncol*. (2005) 17:147–53. doi: 10.1097/01.cco.0000152627.36243.26

9. Menzies SW, Bischof L, Talbot H, Gutenev A, Avramidis M, Wong L, et al. The performance of SolarScan: an automated dermoscopy image analysis instrument for the diagnosis of primary melanoma. *Arch Dermatol*. (2005) 141:1388–96. doi: 10.1001/archderm.141.11.1388

10. Wolner ZJ, Yélamos O, Liopyris K, Rogers T, Marchetti MA, Marghoob AA. Enhancing skin cancer diagnosis with dermoscopy. *Dermatol Clin*. (2017) 35:417–37. doi: 10.1016/j.det.2017.06.003

11. Moncrieff M, Cotton S, Claridge E, Hall P. Spectrophotometric intracutaneous analysis: a new technique for imaging pigmented skin lesions. *Br J Dermatol*. (2002) 146:448–57. doi: 10.1046/j.1365-2133.2002.04569.x

12. Monheit G, Cognetta AB, Ferris L, Rabinovitz H, Gross K, Martini M, et al. The performance of MelaFind: a prospective multicenter study. *Arch Dermatol*. (2011) 147:188–94. doi: 10.1001/archdermatol.2010.302

13. Emery JD, Hunter J, Hall PN, Watson AJ, Moncrieff M, Walter FM. Accuracy of SIAscopy for pigmented skin lesions encountered in primary care: development and validation of a new diagnostic algorithm. *BMC Dermatol*. (2010) 10:9. doi: 10.1186/1471-5945-10-9

14. Rajadhyaksha M, Marghoob A, Rossi A, Halpern AC, Nehal KS. Reflectance confocal microscopy of skin in *vivo*: From bench to bedside. *Lasers Surg Med*. (2017) 49:7–19. doi: 10.1002/lsm.22600

15. Farnetani F, Scope A, Braun RP, Gonzalez S, Guitera P, Malvehy J, et al. Skin cancer diagnosis with reflectance confocal microscopy: reproducibility of feature recognition and accuracy of diagnosis. *JAMA Dermatol*. (2015) 151:1075–80. doi: 10.1001/jamadermatol.2015.0810

16. Ahlgrimm-Siess V, Laimer M, Rabinovitz HS, Oliviero M, Hofmann-Wellenhof R, Marghoob AA, et al. Confocal microscopy in skin cancer. *Curr Dermatol Rep*. (2018) 7:105–18. doi: 10.1007/s13671-018-0218-9

17. Malvehy J, Hauschild A, Curiel-Lewandrowski C, Mohr P, Hofmann-Wellenhof R, Motley R, et al. Clinical performance of the Nevisense system in cutaneous melanoma detection: an international, multicentre, prospective and blinded clinical trial on efficacy and safety. *Br J Dermatol*. (2014) 171:1099–107. doi: 10.1111/bjd.13121

18. Braun RP, Mangana J, Goldinger S, French L, Dummer R, Marghoob AA. Electrical impedance spectroscopy in skin cancer diagnosis. *Dermatol Clin*. (2017) 35:489–93. doi: 10.1016/j.det.2017.06.009

19. Zhao J, Zhao Y, Wu Z, Tian Y, Zeng H. Nonlinear optical microscopy for skin in *vivo*: Basics, development and applications. *J Innovative Optical Health Sci*. (2023) 16:2230018. doi: 10.1142/S179354582230018X

20. Paoli J, Smedh M, Wennberg AM, Ericson MB. Multiphoton laser scanning microscopy on non-melanoma skin cancer: morphologic features for future non-invasive diagnostics. *J Invest Dermatol*. (2008) 128:1248–55. doi: 10.1038/sj.jid.5701139

21. Dimitrow E, Ziemer M, Koehler MJ, Norgauer J, Konig K, Elsner P, et al. Sensitivity and specificity of multiphoton laser tomography for in *vivo* and ex vivo diagnosis of Malignant melanoma. *J Invest Dermatol*. (2009) 127:1752-1758. doi: 10.1038/jid.2008.439

22. Seidenari S, Arginelli F, Dunsby C, French PM, König K, Magnoni C, et al. Multiphoton laser tomography and fluorescence lifetime imaging of melanoma: morphologic features and quantitative data for sensitive and specific non-invasive diagnostics. *PloS One*. (2013) 8:e70682. doi: 10.1371/journal.pone.0070682

23. Levine A, Wang K, Markowitz O. Optical coherence tomography in the diagnosis of skin cancer. *Dermatol Clin*. (2017) 35:465–88. doi: 10.1016/j.det.2017.06.008

24. Feng X, Moy AJ, Nguyen HT, Zhang J, Fox MC, Sebastian KR, et al. Raman active components of skin cancer. *BioMed Opt Express*. (2017) 8:2835–50. doi: 10.1364/BOE.8.002835

25. Feng X, Moy AJ, Nguyen HT, Zhang Y, Zhang J, Fox MC, et al. Raman biophysical markers in skin cancer diagnosis. *J BioMed Opt*. (2018) 23:057002. doi: 10.1117/1.JBO.23.5.057002

26. Lieber CA, Majumder SK, Billheimer D, Ellis DL, Mahadevan-Jansen A. Raman microspectroscopy for skin cancer detection in vitro. *J BioMed Opt*. (2008) 13:024013. doi: 10.1117/1.2899155

27. Lieber CA, Majumder SK, Ellis DL, Billheimer D, Mahadevan-Jansen A. *In vivo* nonmelanoma skin cancer diagnosis using Raman microspectroscopy. *Lasers Surg Med*. (2008) 40:461–7. doi: 10.1002/lsm.20653

28. Nunes LD, Martin AA, Silveira L, Zampieri M. FT-Raman spectroscopy study for skin cancer diagnosis. *Spectroscopy-an Int J*. (2003) 17:597–602. doi: 10.1155/2003/104696

29. Sigurdsson S, Philipsen PA, Hansen LK, Larsen J, Gniadecka M, Wulf HC. Detection of skin cancer by classification of Raman spectra. *IEEE Trans Biomed Eng*. (2004) 51:1784–93. doi: 10.1109/TBME.2004.831538

30. Bratchenko IA, Bratchenko LA, Moryatov AA, Khristoforova YA, Artemyev DN, Myakinin OO, et al. *In vivo* diagnosis of skin cancer with a portable Raman spectroscopic device. *Exp Dermatol*. (2021) 30:652–63. doi: 10.1111/exd.14301

31. Bratchenko IA, Bratchenko LA, Khristoforova YA, Moryatov AA, Kozlov SV, Zakharov VP. Classification of skin cancer using convolutional neural networks analysis of Raman spectra. *Comput Methods Programs Biomedicine*. (2022) 219:106755. doi: 10.1016/j.cmpb.2022.106755

32. Bratchenko IA, Artemyev DN, Myakinin OO, Khristoforova YA, Moryatov AA, Kozlov SV, et al. Combined Raman and autofluorescence ex vivo diagnostics of skin cancer in near-infrared and visible regions. *J BioMed Opt*. (2017) 22:027005–5. doi: 10.1117/1.JBO.22.2.027005

33. Gniadecka M, Philipsen PA, Sigurdsson S, Wessel S, Nielsen OF, Christensen DH, et al. Melanoma diagnosis by Raman spectroscopy and neural networks: structure alterations in proteins and lipids in intact cancer tissue. *J Invest Dermatol*. (2004) 122:443–9. doi: 10.1046/j.0022-202X.2004.22208.x

34. Gniadecka M, Wulf HC, Mortensen NN, Nielsen OF, Christensen DH. Diagnosis of basal cell carcinoma by Raman spectroscopy. *J Raman Spectrosc*. (1997) 28:125–9. doi: 10.1002/(SICI)1097-4555(199702)28:2/3<125::AID-JRS65>3.3.CO;2-R

35. Philipsen PA, Knudsen L, Gniadecka M, Ravnbak MH, Wulf HC. Diagnosis of Malignant melanoma and basal cell carcinoma by in *vivo* NIR-FT Raman spectroscopy is independent of skin pigmentation. *Photochem Photobiol Sci*. (2013) 12:770–6. doi: 10.1039/c3pp25344a

36. Santos IP, Barroso EM, Schut TCB, Caspers PJ, van Lanschot CG, Choi D-H, et al. Raman spectroscopy for cancer detection and cancer surgery guidance: translation to the clinics. *Analyst*. (2017) 142:3025–47. doi: 10.1039/C7AN00957G

37. Austin LA, Osseiran S, Evans CL. Raman technologies in cancer diagnostics. *Analyst*. (2016) 141:476–503. doi: 10.1039/C5AN01786F

38. Kong K, Kendall C, Stone N, Notingher I. Raman spectroscopy for medical diagnostics—From in-vitro biofluid assays to in-vivo cancer detection. *Adv Drug Delivery Rev*. (2015) 89:121–34. doi: 10.1016/j.addr.2015.03.009

39. Auner GW, Koya SK, Huang C, Broadbent B, Trexler M, Auner Z, et al. Applications of Raman spectroscopy in cancer diagnosis. *Cancer Metastasis Rev*. (2018) 37:691–717. doi: 10.1007/s10555-018-9770-9

40. Wang W, Zhao J, Short M, Zeng H. Real-time in *vivo* cancer diagnosis using raman spectroscopy. *J Biophotonics*. (2015) 8:527–45. doi: 10.1002/jbio.201400026

41. Huang Z, Zeng H, Hamzavi I, McLean DI, Lui H. Rapid near-infrared Raman spectroscopy system for real-time in *vivo* skin measurements. *Opt Lett*. (2001) 26:1782–4. doi: 10.1364/OL.26.001782

42. Zhao J, Lui H, McLean DI, Zeng H. Integrated real-time Raman system for clinical in *vivo* skin analysis. *Skin Res Technol*. (2008) 14:484–92. doi: 10.1111/j.1600-0846.2008.00321.x

43. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. (2017) 542:115–8. doi: 10.1038/nature21056

44. Pradhan P, Guo S, Ryabchykov O, Popp J, Bocklitz TW. Deep learning a boon for biophotonics? *J Biophotonics*. (2020) 13:e201960186. doi: 10.1002/jbio.201960186

45. Luo R, Popp J, Bocklitz T. Deep learning for Raman spectroscopy: a review. *Analytica*. (2022) 3:287–301. doi: 10.3390/analytica3030020

46. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer collaboration for skin cancer recognition. *Nat Med*. (2020) 26:1229–34. doi: 10.1038/s41591-020-0942-0

47. He H, Yan S, Lyu D, Xu M, Ye R, Zheng P, et al. *Deep learning for biospectroscopy and biospectral imaging: state-of-the-art and perspectives*. *Anal. Chem*. (2021) 93:3653–3665. doi: 10.1021/acs.analchem.0c04671

48. Meza Ramirez CA, Greenop M, Ashton L, Rehman IU. Applications of machine learning in spectroscopy. *Appl Spectrosc Rev*. (2021) 56:733–63. doi: 10.1080/05704928.2020.1859525

49. Shen J, Li M, Li Z, Zhang Z, Zhang X. Single convolutional neural network model for multiple preprocessing of Raman spectra. *Vibrational Spectrosc*. (2022) 121:103391. doi: 10.1016/j.vibspec.2022.103391

50. Wahl J, Sjödahl M, Ramser K. Single-step preprocessing of raman spectra using convolutional neural networks. *Appl Spectrosc*. (2020) 74:427–38. doi: 10.1177/0003702819888949

51. Gebrekidan MT, Knipfer C, Braeuer AS. Refinement of spectra using a deep neural network: Fully automated removal of noise and background. *J Raman Spectrosc*. (2021) 52:723–36. doi: 10.1002/jrs.6053

52. Ma D, Shang L, Tang J, Bao Y, Fu J, Yin J. Classifying breast cancer tissue by Raman spectroscopy with one-dimensional convolutional neural network. *Spectrochimica Acta Part A: Mol Biomolecular Spectrosc*. (2021) 256:119732. doi: 10.1016/j.saa.2021.119732

53. Wu X, Li S, Xu Q, Yan X, Fu Q, Fu X, et al. Rapid and accurate identification of colon cancer by Raman spectroscopy coupled with convolutional neural networks. *Japanese J Appl Phys*. (2021) 60:067001. doi: 10.35848/1347-4065/ac0005

54. Kothari R, Jones V, Mena D, Bermúdez Reyes V, Shon Y, Smith JP, et al. Raman spectroscopy and artificial intelligence to predict the Bayesian probability of breast cancer. *Sci Rep*. (2021) 11:6482. doi: 10.1038/s41598-021-85758-6

55. Chen S, Zhang H, Yang X, Shao X, Li T, Chen N, et al. Raman spectroscopy reveals abnormal changes in the urine composition of prostate cancer: an application of

an intelligent diagnostic model with a deep learning algorithm. *Advanced Intelligent Syst*. (2021) 3:2000090. doi: 10.1002/aisy.202000090

56. Zhang X, Lin T, Xu J, Luo X, Ying Y. DeepSpectra: An end-to-end deep learning approach for quantitative spectral analysis. *Anal Chim Acta*. (2019) 1058:48–57. doi: 10.1016/j.aca.2019.01.002

57. Fukuhara M, Fujiwara K, Maruyama Y, Itoh H. Feature visualization of Raman spectrum analysis with deep convolutional neural network. *Anal Chim Acta*. (2019) 1087:11–9. doi: 10.1016/j.aca.2019.08.064

58. Kothari R, Fong Y, Storrie-Lombardi MC. Review of laser Raman spectroscopy for surgical breast cancer detection: stochastic backpropagation neural networks. *Sensors*. (2020) 20:6260. doi: 10.3390/s20216260

59. Meng C, Li H, Chen C, Wu W, Gao J, Lai Y, et al. Serum Raman spectroscopy combined with Gaussian—convolutional neural network models to quickly detect liver cancer patients. *Spectrosc Lett*. (2022) 55:79–90. doi: 10.1080/00387010.2022.2027988

60. Zhao J, Lui H, McLean DI, Zeng H. Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. *Appl Spectrosc*. (2007) 61:1225–32. doi: 10.1366/000370207782597003

61. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. (2017). Grad-cam: Visual explanations from deep networks *via* gradient-based localization, in: *Proceedings of the IEEE international conference on computer vision*, Venice, Italy: Piscataway, NJ, IEEE pp. 618-626. doi: 10.1109/ICCV.2017.74

62. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J big Data*. (2019) 6:1–48. doi: 10.1186/s40537-019-0197-0

63. Wang J, Perez L. The effectiveness of data augmentation in image classification using deep learning. *arXiv:1712.04621*. (2017) 11:1–8. doi: 10.48550/arXiv.1712.04621

64. Mikołajczyk A, Grochowski M. (2018). Data augmentation for improving deep learning in image classification problem, in: *2018 international interdisciplinary PhD workshop (IIPhDW)*, Świnoujście, Poland: IEEE pp. 117-122. Piscataway, NJ. doi: 10.1109/IIPHDW.2018.8388338

65. Liu J, Osadchy M, Ashton L, Foster M, Solomon CJ, Gibson SJ. Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *Analyst*. (2017) 142:4067–74. doi: 10.1039/C7AN01371J

66. Bjerrum EJ, Glahder M, Skov T. Data augmentation of spectral data for convolutional neural network (CNN) based deep chemometrics. *arXiv: 1710.01927*. (2017). doi: 10.48550/arXiv.1710.01927

67. Wu M, Wang S, Pan S, Terentis AC, Strasswimmer J, Zhu X. Deep learning data augmentation for Raman spectroscopy cancer tissue classification. *Sci Rep*. (2021) 11:23842. doi: 10.1038/s41598-021-02687-0

68. Yu S, Li H, Li X, Fu YV, Liu F. Classification of pathogens by Raman spectroscopy combined with generative adversarial networks. *Sci Total Environ*. (2020) 726:138477. doi: 10.1016/j.scitotenv.2020.138477

69. Qin Z, Liu Z, Zhu P, Xue Y. A GAN-based image synthesis method for skin lesion classification. *Comput Methods Programs Biomedicine*. (2020) 195:105568. doi: 10.1016/j.cmpb.2020.105568

70. Dumoulin V, Visin F. A guide to convolution arithmetic for deep learning. *arXiv: 1603.07285*. (2016). doi: 10.48550/arXiv.1603.07285

71. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv: 1412.6980*. (2014). doi: 10.48550/arXiv.1412.6980

72. Guo S, Popp J, Bocklitz T. Chemometric analysis in Raman spectroscopy from experimental design to machine learning–based modeling. *Nat Protoc*. (2021) 16:5426–59. doi: 10.1038/s41596-021-00620-3

73. Acquarelli J, van Laarhoven T, Gerretzen J, Tran TN, Buydens LM, Marchiori E. Convolutional neural networks for vibrational spectroscopic data analysis. *Anal Chim Acta*. (2017) 954:22–31. doi: 10.1016/j.aca.2016.12.010

74. Khristoforova Y, Bratchenko L, Bratchenko I. Combination of Raman spectroscopy and chemometrics: A review of recent studies published in the Spectrochimica Acta, Part A: Molecular and Biomolecular Spectroscopy Journal. *arXiv: 2210.10051*. (2022). doi: 10.48550/arXiv.2210.10051

75. Butler HJ, Ashton L, Bird B, Cinque G, Curtis K, Dorney J, et al. Using Raman spectroscopy to characterize biological materials. *Nat Protoc*. (2016) 11:664–87. doi: 10.1038/nprot.2016.036

76. Bocklitz T, Walter A, Hartmann K, Rösch P, Popp J. How to pre-process Raman spectra for reliable and stable models? *Anal Chim Acta*. (2011) 704:47–56. doi: 10.1016/j.aca.2011.06.043