Deep Convolutional Neural Networks as a Unified Solution for Raman Spectroscopy-Based Classification in Biomedical Applications

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

13-11-2021 / 24-11-2021

CITATION

Kazemzadeh, Mohammadrahim; Hisey, Colin; Zargar, Kamran; Xu, Peter; Broderick, Neil (2021): Deep Convolutional Neural Networks as a Unified Solution for Raman Spectroscopy-Based Classification in Biomedical Applications. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.17004058.v1>

DOI

[10.36227/techrxiv.17004058.v1](https://doi.org/10.36227/techrxiv.17004058.v1)

Deep Convolutional Neural Networks as a Unified Solution for Raman Spectroscopy-Based Classification in Biomedical Applications

Mohammadrahim Kazemzadeh^{a,b}, Colin L. Hisey^{c,d}, Kamran Zargar-Shoshtari^e, Weiliang Xu^{a,b}, Neil G.R. Broderick^{f,b}

^aDepartment of Mechanical Engineering, University of Auckland, , Auckland, 1010, New Zealand

^bDodd-Walls Centre for Photonic and Quantum Technologies, New Zealand

^cDepartment of Obstetrics and Gynaecology, University of Auckland, , Auckland, 1023, New Zealand

^dHub for Extracellular Vesicle Investigations, University of Auckland, Auckland, 1023, New Zealand

^eDepartment of Surgery, University of Auckland, , Auckland, 1023, New Zealand

^fDepartment of Physics and Photon Factory, University of Auckland, , Auckland, 1010, New Zealand

Abstract

Machine learning has shown great potential for classifying diverse samples in biomedical applications based on their Raman spectra. However, the acquired spectra typically require several preprocessing steps before standard machine learning algorithms can accurately and reliably classify them. To simplify this workflow and enable future growth of this technology, we present a unified solution for classifying biological Raman spectra without any need of preprocessing, including denoising and baseline establishment. This method is developed based on a custom version of a convolutional neural network (CNN) elicited from ResNet architecture, combined with our proposed data augmentation technique. The superiority of this method compared to conventional classification techniques is shown by applying it to Raman spectra of different grades of bladder cancer tissue and surface enhanced Raman spectroscopy (SERS) spectra of various strains of *E. Coli* extracellular vesicles (EVs). These results show that our method is far more robust compared to its conventional counterparts when dealing with the various kinds of spectral baselines produced by different Raman spectrometers.

Keywords: Raman Spectroscopy, SERS, Deep learning, convolutional neural network, extracellular vesicles

1. Introduction

Raman spectroscopy is a powerful technique for fingerprinting the unique bonds within a diverse range of chemical and biological materials [1, 2, 3, 4]. The probability of generating a Raman photon however is very low, with roughly only one out every billion incident photons undergoing inelastic scattering. This makes Raman spectroscopy very sensitive to factors such as fluorescent emissions, thermal noise, the quality of optical filters used, and the calibration precision of the spectrometers. Thus to effectively use the acquired Raman spectra, they must be carefully preprocessed to remove such artifacts. These preprocessing methods include cosmic ray removals[5], spectral denoising [6, 7, 8, 9], and baseline corrections [10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. Currently, human interaction is always needed for successful preprocessing of unknown samples, and is particularly important for biological materials due to the complex nature of their spectral baselines [16].

Various baseline correction techniques have been introduced and applied extensively for Raman spectroscopy applications. These methods can be categorised as polynomial fitting [10,

11, 12], penalized least square [13, 14], wavelet transform [17, 18, 19], morphological operations [15, 16], and windows based methods[20]. Polynomial methods are effective at avoiding Raman signal distortion when dealing with complex baselines, but require choosing the optimum value for the degree of the polynomials. Recently, a least square method called asymmetric least square (AsLS) smoothing was introduced [13] and has been improved by iteratively adapting the weights to reduce the residuals between the estimated baselines and the original signals[14]. The resulting method is known as adaptive iteratively reweighted penalized least squares (airPLS). However, like the polynomial fitting, all the least square methods have some adjustable parameters like the penalized weights and thereby need human interaction to provide an acceptable result.

Mathematical morphology (MOR) has been used previously for feature extraction in image processing, and has even been modified for use in baseline corrections of Raman spectral data[15, 16]. The main shortcoming of MOR[15] and improved morphological [16] method (iMOR) is the presence of jumps in the resulting baselines which are transferred directly into the post-processed Raman spectra. In addition, some parameters like the tolerance and half-window sizes need to be manually adjusted. Sensitive nonlinear iterative peak (SNIP) [20] is another window based method that has been used on spectroscopy data to eliminate unwanted backgrounds. However, this method also suffers from jumps and the necessity of human discretion.

Email addresses: mkaz499@aucklanduni.ac.nz (Mohammadrahim Kazemzadeh), colin.hisey@auckland.ac.nz (Colin L. Hisey), p.xu@auckland.ac.nz (Weiliang Xu), n.broderick@auckland.ac.nz (Neil G.R. Broderick)

Recently, several efforts have been made to either classify the raw Raman spectra directly [21] or automatically denoise and correct their baselines [22] using Convolutional Neural Networks (CNN). In one study [21], different variants of CNN architectures were used to classify the raw Raman spectra of chemical species within minerals. They demonstrated that CNN can even lead to better classification accuracy when trained on raw Raman spectra compared to when the preprocessed data is used for training. They also showed that their proposed CNN classifiers outperform all the conventional classification methods including linear discriminant analysis (LDA), support vector machine (SVM), k-nearest neighbour (KNN), and artificial neural network (ANN). This result is even more significant considering that no human intervention was needed, and thus no bias can be unintentionally included in the analysis. In another study [22], a supervised CNN was developed to correct the baselines and denoise the Raman spectra. They trained the CNN using artificially generated spectra, then applied the trained network to polyethylene, paraffin, and ethanol. They demonstrated successful automated baseline removal that was superior to conventional methods. Recently, researchers have also used U-Net as a variant of CNN to denoise and correct baselines of raw Raman signals [23]. They also used simulated spectra to train their network, then applied it to a number of scenarios including preprocessing of the Raman spectra of pig tissues. LDA classification accuracy of 70.9% between cortical bone and the cancellous bone of the pig was achieved when the raw spectra were preprocessed using their method. They also stated that this result is significantly better than classical methods as no human intervention was needed.

CNN has also been used to classify a variety of different biological samples using their Raman spectroscopy data including different strains of *E. Coli* bacteria cells [24], porcine skin samples which were irradiated by ultraviolet light for different durations [25], and breast cancer tissue samples [26]. In one study [24], the CNN ResNet [27] could successfully classify different strains of *E. Coli* bacteria from highly noisy Raman signals obtained using surface enhanced Raman spectroscopy (SERS) [28]. In another study [25], a custom architecture of single layer multiple kernel-based CNN (SLMK-CNN) was applied to classify preprocessed porcine skin samples. They obtained a classification accuracy of 96.4% and 92.5% from preprocessed and raw spectra, respectively. Baseline corrected breast cancer samples have also been successfully classified using CNN [26]. In that study, they used data augmentation to increase their training set size from 600 to 5000 spectra using methods such as small spectral shifting (2 cm^{-1}), expanding the spectral range from 3000 cm^{-1} to 3600 cm^{-1} , adding Gaussian noise to the spectra, and finally superimposing the spectra within a cluster. These kinds of data augmentation techniques are necessary for the data-hungry neural network models such as CNN to expand the size of the data and reduce overfitting when training the model.

Most recently, CNN has been used in a few instances involving micro and nanoscale lipid-enclosed packages known as extracellular vesicles (EVs). As EVs are produced by all cells and have been found in nearly all biological fluids, there has been

significant interest in characterising them using Raman spectroscopy. In one study, it was applied to both raw and processed conventional Raman spectra from prostate cancer cell line-derived EVs and blood-derived EVs with greater than 90% classification accuracy [29]. In a more recent study, CNN was first trained on processed SERS spectra of EVs from normal and lung cancer cell lines, then applied to processed SERS spectra from lung cancer and healthy patient plasma EVs, again achieving greater than 90% detection accuracy EVs [30]. In this study, we expand on these recent developments and propose a novel architecture of CNN, inspired by ResNet and combined with our data augmentation method, to classify raw conventional Raman and SERS spectra of biological samples with superior accuracy compared to standard machine learning algorithms. Our method does not require any human input and the entire process can be automated. We demonstrate that our method works with both high and low resolution Raman spectra. This is particularly relevant as it enables applications using less expensive Raman spectrometers, encouraging more widespread clinical use. Furthermore, we show that the sensitivity of the proposed method when dealing with spectral shifts, such as the case when the spectrometer is not well-calibrated and either the grating or detector is spatially displaced, is much better than the other machine learning algorithms tested. Lastly, to the best of our knowledge, this is the first example of using CNN to classify EVs using raw SERS spectra and the first application of CNN to classify bacterial EV Raman spectra of any kind.

2. Methodology

We employed a custom built CNN inspired by ResNet architecture and then trained it using two very different data sets. Our primary data set is from previously obtained Raman spectra of different human bladder tissues [31]. This data set contains 2592 spectra obtained using a portable, low resolution Raman spectrometer (36 cm^{-1}) with three categories: healthy tissue, and either low or high-grade tumors. This data was obtained using various laser powers, therefore it contains Raman signals with the wide range of signal to noise ratios (SNR) and baselines. Our second data set includes the SERS spectra of EVs from different strains (Nissle, K12, UPEC) of *E. Coli* which were cultured in different media (R - RPMI, RF - RPMI + iron) or purified using different methods (SEC - size exclusion chromatography, DG - density gradient ultracentrifugation) [32, 33]. This data is obtained using high resolution (8 cm^{-1}) Raman confocal microscope. SERS peaks of EVs are known to be very weak due to EVs' lack of chromophore molecules, allowing background noises to easily interfere with the EV signals, ultimately requiring explicit human interaction during preprocessing.

These data are then augmented for two main reasons. Firstly, to avoid the overfitting of training and secondly to introduce different scenarios to the proposed network to obtain more reliable accuracy when it's faced with data obtained from different Raman spectrometers or ambient disturbances.

Figure 1: Convolution and Identity blocks used in the proposed ResNet architecture.

2.1. CNN Architecture

We used a custom version of the CNN ResNet architecture [34] which is a feed-forward neural network with added skips that bypass some layers. The proposed version overcomes two main obstacles that occur when implementing a very deep neural network. First, it significantly helps the vanishing gradient problem [35, 36], making it possible to implement a much deeper neural network, and secondly, accuracy saturation and degradation, a problem which arises when more layers are added to a shallow plain neural network [37], can be handled successfully.

The mentioned skips are implemented in two sub-blocks of the whole network called the Identity block and the Convolutional block (Fig. 1). Each of these blocks consists of stacks of convolution, batch normalization, and nonlinear activation layers while the input shortcuts to their output. The only difference between the Identity and Convolutional blocks is the presence of a single convolution and batch normalization layer in the detour pass of the Convolutional block. The added convolutional layer in the bypass of the convolutional block guarantees that the bypass and output of the block are in the same shape and can be added together at the end of the block. In fact, a network that is constructed using only identity blocks cannot have different hyper parameters at each of its blocks, as any change in the filter sizes or introduction of strides can lead to an output with a different shape than its input. Therefore, any time that the block's hyper parameters need to be changed, we have introduced a convolutional block and used identity blocks with the same hyper parameters. Moreover, stride implementation, which can reduce the computational cost of the training and to some degree can avoid the overfitting is only possible through the convolutional block. The parameters n_1^i and n_2^i are the number of filters in the different layers of convolution layers (CONV), while F^i and s^i are the size and stride step defined for

the filters.

The overall architecture of the proposed ResNet is depicted in Fig. 2. The training set for this network is the raw Raman spectra. The data is directly fed into a convolution layer followed by batch normalization and the rectified linear activation layer. The output of the activation is then fed into five consecutive processing steps. Each step consists of a stack of Convolution blocks followed by two Identity blocks, with the parameters of each step being different from the others. Finally, the output of the last step is connected to a fully connected flat layer and the output of the whole network is generated using a softmax activation layer. The network hyper parameters n_1^i , n_2^i , F^i and s^i are chosen to be $\{16, 16, 16, 32, 64\}$, $\{32, 32, 32, 64, 128\}$, $\{30, 15, 7, 4, 3\}$ and $\{2, 2, 2, 2, 2\}$ for each convolutional and identity blocks of Fig. 2, respectively. As shown, the bigger filter sizes F^i are used at the beginning of the network to extract the global features of the data and smaller ones are used toward the last blocks to extract the local features such as noises and peaks shapes and positions.

2.2. Data Augmentation

In order to properly train the neural net, a larger data set than was realistically available was needed, so the experimental data was augmented with artificially generated spectra. This process enlarges the size of the training set from 2073 and 280 to 12000 and 5000 for human bladder cancer and SERS spectra of *E. Coli* EVs, respectively. This was done for two main reasons. First, larger data sets can effectively avoid overfitting which is common with data-hungry methods, and second, the deep learning model is trained in a way which allows it to learn more complex and diverse types of baselines. In fact, we show that this deep learning method is only more effective than the conventional Raman spectroscopy analysis when data augmentation is used.

Figure 2: Overall architecture of the proposed CNN ResNet network used for Raman spectroscopy classification of biological samples. Convolution and Identity blocks are depicted in more detail in Fig. 1.

Figure 3: Proposed methods of the data augmentation showing an example of an (a) original spectra, (b) added noise, (c) spectral shift, and (d) baseline augmentation.

Otherwise, even slight deviations from the available baselines can affect the performance of the deep network.

To create the augmented data we started with the experimental spectra and applied baseline changes, the addition of random noise, spectral shifting to either the left or right, and finally a linear combination of previously augmented spectra. A representative example of the methods that were used for data augmentation are depicted in Fig. 3. For the baseline changes, we restricted ourselves to adding a randomly chosen linear and quadratic polynomials with the additional constraint that resulting spectra would always be positive. Also, the Gaussian noise with the mean of zero and variance proportional to the amplitude of the signal has been used for this data augmentation.

To show the effect of the data augmentation over the per-

formance of the proposed convolutional deep neural network we considered three different training scenarios including training over the original data, data augmentation with linear combination of the available baselines in the data set, and data augmentation with second degree polynomials and the original baselines. The preprocessed bladder cancer data, using airPLS baseline correction for these three scenarios is depicted in Fig. 4. Principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) [38], and uniform manifold approximation and projection (UMAP)[39] have been applied for the data visualization in lower dimensions. In fact, these data visualization techniques show the classification possibility over the data as they are unsupervised machine learning, and any separation in their embedding space is due solely to the difference in their hyper dimensional space and not from their labels. We also see that the UMAP has the better preservation of the continuity of the data within a cluster in comparison to t-SNE. This phenomena has been described in detail in [39] and [40] for various types of data including visualization of single cell data.

Interestingly, the border between different kinds of cancer and healthy samples are more mixed in UMAP projection of the dataset which is augmented by polynomial baselines than of the native baseline augmentation (Fig 4). It shows that the baseline correction can statistically bias the data when it deals with more complex baselines. This is the main reason that the accuracy of the conventional machine learning is lower when the test dataset is obtained using a different Raman spectrometer than that used for the training dataset, as any change in the components of the Raman spectrometer can lead to the considerable changes in the quality of the signals and baselines.

A summary of the entire process of training and testing the proposed CNN and conventional machine learning is shown in Fig. 5. For the test branch, a block was added to simulate the condition in which the Raman spectra are obtained using a

Figure 4: Original (top), augmented with its native baselines (middle), and augmented with its native baselines and second degree polynomials (bottom) Raman spectral data of bladder cancer tissues used for training purpose. For each dataset, the preprocessed Raman spectra with airPLS baseline correction and their two dimensional projections using PCA, t-SNE, and UMAP are depicted.

different machine. This block alters the spectral baselines and is capable of shifting the spectra to the left or right. The effect of this block is explained in detail in the next section.

2.3. Network Training

The proposed network was implemented using TensorFlow and keras in python, and as this is a classification problem, categorical cross entropy was used for the loss function. Adam optimiser[41] as variant of stochastic gradient descending algorithm was also used for training the proposed network while learning rates, β_1 , β_2 and ϵ were chosen to be 10^{-3} , 0.9, 0.999 and 10^{-8} , respectively. To avoid the overfitting, we have considered the Least Absolute Shrinkage and Selection Operator (lasso) as the kernel regularizer known also as the L2 regularizer at each convolutional layer. We also monitored the both training and validation loss through the network training to mitigate overfitting. The training was terminated based on our defined criteria of validation loss (smaller than 0.1). For all scenarios and datasets, the training was completed in less than

100 epochs and took less than 1 hour with our mid range GPU (NVIDIA QUADRO P620).

3. Results and Discussion

We compared the classification ability of the proposed method to a number of established machine learning algorithms and evaluated their ability to cope with changes in the spectroscopy setup or weak calibrations. This performance test was carried out using two vastly different biological datasets and spectrometers: human bladder cancer tissue Raman spectra from a portable spectrometer and *E. Coli* EV SERS spectra from a confocal Horiba LabRAM Evolution Raman microscope.

3.1. Low Resolution Raman Spectra of Bladder Cancer Tissue

The achieved accuracy and the confusion matrix for the proposed CNN trained by the augmented data compared to the results of conventional machine learning algorithms using prepro-

Figure 5: Overall process of training and testing for all machine learning and proposed deep learning algorithms

Figure 6: Confusion matrix obtained using LDA, KNN, SVM, ANN, RF, and our proposed CNN on raw data.

processed augmented data are shown in Fig. 6. For this preprocessing, we used airPLS algorithms to remove the baseline and the parameters for this algorithm were tuned manually to achieve the best possible result. Interestingly, the result obtained from CNN are almost as accurate as the best machine learning algorithms, even though it was fed solely with raw data.

To clearly demonstrate the effects of data augmentation, we tested all the classification algorithms with simulated test sets to check their robustness against changes in baselines and spec-

tral shifts. We considered the 20th order polynomial, sinusoidal baselines (Equation 1), and consistent shifts of the spectra to a higher or lower wavenumber.

$$baseline = A \sin\left(\alpha \frac{\pi(x - x_0)}{5L}\right) + a\pi + b \quad (1)$$

Here, A is a random number between 0 and 5000, and x_0 and L are the minimum and the difference between maximum and minimum wave numbers of investigated spectra, respectively. α is a coefficient used in Fig. 7 and Fig. 8 and b is chosen in a way to make sure that the spectra value is more than zero for all investigated wavenumbers while placing the spectra randomly between 0 and 400 above the zero level. Finally, a is another random number between -1 and 1 which is the phase of the simulated sinusoidal baseline.

For all machine learning algorithms, the test sets were pre-processed with airPLS while raw simulated data was used for CNN. This was done by training over a non-augmented training data set and an augmented training data set. The result of this test is depicted in Fig. 7, showing that the proposed deep learning method trained with the original data has an acceptable accuracy while only dealing with the original baselines of the data (spectral shift=0), in cases where $\alpha = 0$ for the simulation of sinusoidal baselines and zero order polynomials for the simulation of polynomial baselines (which are both equal to the corrected baselines, see Equations 1). In other words, the deep learning only understood a way to remove the original baselines, and if the same data with a slight change of the baseline or spectral shift is tested, the result is unacceptable. Another interesting finding is the high sensitivity of the trained deep learning over spectral shifts, indicating that the classification is based on the positions of Raman peaks and not their baselines.

As shown in Fig. 7, the data augmentation drastically enhances the performance of the proposed deep learning method. It demonstrates that the proposed method, trained with augmented data, outperforms the conventional methods for all the simulation scenarios while even better results are achieved when second degree polynomial baselines are used for data augmentation. This simulation also demonstrates the ability of the

Figure 7: Accuracy of the machine learning and the proposed deep learning trained using various data augmentation methods. Different test sets were simulated by adding different polynomials and sinusoidal baselines to the data. Also, different amounts of spectral shifts were simulated and tested on the trained methods. All the data used for machine learning algorithms are preprocessed using airPLS algorithm.

Figure 8: Comparison between the accuracy of various machine learning algorithms and the proposed deep learning when the data for machine learning algorithms are preprocessed using AsLS, iMOR and SNIP baseline correction methods.

trained network to effectively deal with the baselines as they do not adversely affect the classification. The sensitivity of the classification accuracy over the spectral shifts also shows the classification based on the peaks' positions and shapes. However, as the extracted features of the deep learning are mostly unreadable, one does not know exactly how it manages to re-

move the baseline or which peak or peaks plays the most important roles in the classification. This can be considered as the downfall of the presented method against the conventional methods (PCA/SVM) in which the principal components can be easily extracted and interpreted by the user.

To demonstrate the effects of different baseline correction

techniques, the same experiments from Fig. 7 were repeated using AsLS, iMOR, and NIP. As shown in Fig. 8, iMOR and SNIP performed better than airPLS and AsLS for the several cases of the simulated baselines (e.g. higher order of polynomials), but the results are still far weaker compared to the proposed method.

Figure 9: Two raw and preprocessed SERS spectra of Nissle-R-SEC EVs. The heterogeneity due to the different biochemical compositions, possibly causing the presence or absence of some peaks, and differences in Raman intensity, possibly resulting from different sizes or physical placements on the substrate, are clearly illustrated.

3.2. High Resolution *E. Coli* EV SERS Spectra

The second dataset used to test the proposed CNN includes SERS spectra obtained using a high resolution confocal Raman microscope. SERS substrates used in this research [42] were designed and fabricated based on notion of transformation optics [43, 44, 45, 46] and a combination of soft and nanoparticle lithography. We demonstrated its highly reproducible for chemicals and biological materials [42] while its enhancement factor is improved by 20 times compared to its traditional counterparts. EVs from different strains of *E. Coli* bacteria that were cultured in different media or purified using different methods were used [47]. The conventional preprocessig and machine learning classification of these same SERS spectra were recently reported [33].

Although commercial confocal Raman microscopes are much less sensitive to weak calibration and reproducibility issues related to spectral baselines compared to custom-built low resolution Raman spectrometers, there are still some major challenges regarding SERS spectra of EVs. Firstly, due to heterogeneous nature of EVs [48] and their lack of chromophore molecules [49], their SERS spectra are very weak and diverse and can be easily mixed with the short band Raman spectra

Figure 10: UMAP projection of original (top) and augmented (bottom) *E. Coli* EV SERS data.

of plasmonic nanoparticles in SERS. Another major challenge is the high variance of the EVs' SERS intensities across the sensor's surface, potentially caused by the wide EV size distribution (from roughly 30 to hundreds of nanometers depending on their subtype) and natural tendency to aggregate in patches. Thereby, an extensive baseline corrections and denoising is needed (Fig. 9) for each one the spectra individually, requiring excessive amounts of time and effort which is still ultimately vulnerable to unintended bias.

To address these issues, we applied the mentioned deep learning network on augmented data of EV's SERS spectra to automatically classify them. To the best of our knowledge, this is the first reported successful attempt to classify EV SERS spectra using raw spectral data. The UMAP projection of the preprocessed training dataset and its augmented version is depicted in Fig. 10. Also, the confusion matrix obtained using proposed method is depicted in Fig. 11. Comparing this result with the best result previously obtained by machine learning technique, (96% using Gaussian process classifier) [33], we can see that the fully automated deep learning provides very acceptable results. However, to further improve this area the size of EV spectral database must be expanded and further optimisation is needed to tune this method purposely for SERS application in EVs. These efforts will be expanded in future work given the promising results reported herein.

Figure 11: Confusion matrix obtained using proposed deep learning method for *E. Coli* EV classification.

4. conclusion

In this study, we demonstrate the utility of a novel architecture of CNN, inspired by ResNet, combined with our data augmentation method to classify raw Raman spectra of biological samples with superior accuracy compared to standard machine learning algorithms. The reported CNN does not require any human input, can be completely automated, and can be utilized using much smaller data sets than ResNet. We further demonstrated that it can handle extreme baseline scenarios that cause failures when using other techniques, which could be particularly relevant for encouraging more widespread clinical use of Raman spectroscopy. Furthermore, we show that the sensitivity of the the proposed method when dealing with spectral shifts is much better than other conventional machine learning algorithms. We anticipate that these findings will encourage and enable further developments of Raman spectroscopy in a diverse range of biomedical applications.

References

- [1] A. S. Haka, K. E. Shafer-Peltier, M. Fitzmaurice, J. Crowe, R. R. Dasari, M. S. Feld, Diagnosing breast cancer by using raman spectroscopy, *Proceedings of the National Academy of Sciences* 102 (35) (2005) 12371–12376.
- [2] M. Vendrell, K. K. Maiti, K. Dhaliwal, Y.-T. Chang, Surface-enhanced raman scattering in cancer detection and imaging, *Trends in biotechnology* 31 (4) (2013) 249–257.
- [3] Z. Movasaghi, S. Rehman, I. U. Rehman, Raman spectroscopy of biological tissues, *Applied Spectroscopy Reviews* 42 (5) (2007) 493–541.
- [4] H. Chen, X. Li, N. Broderick, Y. Liu, Y. Zhou, J. Han, W. Xu, Identification and characterization of bladder cancer by low-resolution fiber-optic raman spectroscopy, *Journal of biophotonics* 11 (9) (2018) e201800016.
- [5] H. G. Schulze, R. F. Turner, A two-dimensionally coincident second difference cosmic ray spike removal method for the fully automated processing of raman spectra, *Applied spectroscopy* 68 (2) (2014) 185–191.
- [6] H. Chen, W. Xu, N. Broderick, J. Han, An adaptive denoising method for raman spectroscopy based on lifting wavelet transform, *Journal of Raman Spectroscopy* 49 (9) (2018) 1529–1539.
- [7] F. Ehrentreich, L. Stimmchen, Spike removal and denoising of raman spectra by wavelet transform methods, *Analytical chemistry* 73 (17) (2001) 4364–4373.
- [8] B. Manifold, E. Thomas, A. T. Francis, A. H. Hill, D. Fu, Denoising of stimulated raman scattering microscopy images via deep learning, *Biomedical optics express* 10 (8) (2019) 3860–3874.
- [9] Z. Fang, Y. Tao, W. Wang, W. Zhang, L. Duan, Y. Liu, C. Yan, L. Qu, C. Han, Joint sparse representation and denoising method for raman spectrum, *Journal of Raman Spectroscopy* 49 (12) (2018) 1972–1977.
- [10] P. Cadusch, M. Hlaing, S. Wade, S. McArthur, P. Stoddart, Improved methods for fluorescence background subtraction from raman spectra, *Journal of Raman Spectroscopy* 44 (11) (2013) 1587–1595.
- [11] J. Liu, J. Sun, X. Huang, G. Li, B. Liu, Goldindc: a novel algorithm for raman spectrum baseline correction, *Applied spectroscopy* 69 (7) (2015) 834–842.
- [12] C. Gallo, V. Capozzi, M. Lasalvia, G. Perna, An algorithm for estimation of background signal of raman spectra from biological cell samples using polynomial functions of different degrees, *Vibrational Spectroscopy* 83 (2016) 132–137.
- [13] J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, J. Tan, Asymmetric least squares for multiple spectra baseline correction, *Analytica chimica acta* 683 (1) (2010) 63–68.
- [14] Z.-M. Zhang, S. Chen, Y.-Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, *Analyst* 135 (5) (2010) 1138–1146.
- [15] R. Perez-Pueyo, M. J. Soneira, S. Ruiz-Moreno, Morphology-based automated baseline removal for raman spectra of artistic pigments, *Applied spectroscopy* 64 (6) (2010) 595–600.
- [16] Y. Chen, L. Dai, An automated baseline correction method based on iterative morphological operations, *Applied spectroscopy* 72 (5) (2018) 731–739.
- [17] J. Li, L.-P. Choo-Smith, Z. Tang, M. G. Sowa, Background removal from polarized raman spectra of tooth enamel using the wavelet transform, *Journal of Raman Spectroscopy* 42 (4) (2011) 580–585.
- [18] Y. Liu, W. Cai, X. Shao, Intelligent background correction using an adaptive lifting wavelet, *Chemometrics and Intelligent Laboratory Systems* 125 (2013) 11–17.
- [19] D. Chen, Z. Chen, E. Grant, Adaptive wavelet transform suppresses background and noise for quantitative analysis by raman spectrometry, *Analytical and bioanalytical chemistry* 400 (2) (2011) 625–634.
- [20] M. Morháč, An algorithm for determination of peak regions and baseline elimination in spectroscopic data, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 600 (2) (2009) 478–487.
- [21] J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, S. J. Gibson, Deep convolutional neural networks for raman spectrum recognition: a unified solution, *Analyst* 142 (21) (2017) 4067–4074.
- [22] J. Wahl, M. Sjödaahl, K. Ramser, Single-step preprocessing of raman spectra using convolutional neural networks, *Applied spectroscopy* 74 (4) (2020) 427–438.
- [23] M. T. Gebrekidan, C. Knipfer, A. S. Braeuer, Refinement of spectra using a deep neural network: Fully automated removal of noise and background, *Journal of Raman Spectroscopy* 52 (3) (2021) 723–736.
- [24] C.-S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. Saleh, S. Ermon, J. Dionne, Rapid identification of pathogenic bacteria using raman spectroscopy and deep learning, *Nature communications* 10 (1) (2019) 1–8.
- [25] W. B. Sohn, S. Y. Lee, S. Kim, Single-layer multiple-kernel-based convolutional neural network for biological raman spectral analysis, *Journal of Raman Spectroscopy* 51 (3) (2020) 414–421.
- [26] D. Ma, L. Shang, J. Tang, Y. Bao, J. Fu, J. Yin, Classifying breast cancer tissue by raman spectroscopy with one-dimensional convolutional neural network, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 256 (2021) 119732.
- [27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [28] P. L. Stiles, J. A. Dieringer, N. C. Shah, R. P. Van Duyne, Surface-enhanced raman spectroscopy, *Annu. Rev. Anal. Chem.* 1 (2008) 601–626.
- [29] W. Lee, A. T. Lenferink, C. Otto, H. L. Offerhaus, Classifying raman

- spectra of extracellular vesicles based on convolutional neural networks for prostate cancer detection, *Journal of raman spectroscopy* 51 (2) (2020) 293–300.
- [30] H. Shin, S. Oh, S. Hong, M. Kang, D. Kang, Y.-g. Ji, B. H. Choi, K.-W. Kang, H. Jeong, Y. Park, et al., Early-stage lung cancer diagnosis by deep learning-based spectroscopic analysis of circulating exosomes, *ACS nano* 14 (5) (2020) 5435–5444.
 - [31] H. Chen, X. Li, N. Broderick, Y. Liu, Y. Zhou, J. Han, W. Xu, Identification and characterization of bladder cancer by low-resolution fiber-optic raman spectroscopy, *Journal of biophotonics* 11 (9) (2018) e201800016.
 - [32] J. Hong, P. Dauros-Singorenko, A. Whitcombe, L. Payne, C. Blenkiron, A. Phillips, S. Swift, Analysis of the escherichia coli extracellular vesicle proteome identifies markers of purity and culture conditions, *Journal of extracellular vesicles* 8 (1) (2019) 1632099.
 - [33] M. Kazemzadeh, C. Hisey, P. Dauros Singorenko, S. Swift, K. Zargar, P. Xu, N. Broderick, Label-free classification of bacterial extracellular vesicles by combining nanoplasmonic sensors with machine learning, *TechRxiv* (2021).
 - [34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
 - [35] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, *IEEE transactions on neural networks* 5 (2) (1994) 157–166.
 - [36] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feed-forward neural networks, in: *Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings*, 2010, pp. 249–256.
 - [37] K. He, J. Sun, Convolutional neural networks at constrained time cost, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5353–5360.
 - [38] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., *Journal of machine learning research* 9 (11) (2008).
 - [39] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, *arXiv preprint arXiv:1802.03426* (2018).
 - [40] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, E. W. Newell, Dimensionality reduction for visualizing single-cell data using umap, *Nature biotechnology* 37 (1) (2019) 38–44.
 - [41] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
 - [42] M. Kazemzadeh, C. L. Hisey, A. Artuyants, C. Blenkiron, L. W. Chamley, K. Zargar-Shoshtari, W. Xu, N. G. Broderick, Space curvature-inspired nanoplasmonic sensor for breast cancer extracellular vesicle fingerprinting and machine learning classification, *Biomedical Optics Express* 12 (7) (2021) 3965–3981.
 - [43] M. Kazemzadeh, N. Broderick, W. Xu, K. Z. Shoshtari, Design and optimization of broadband and transparent sers based on transformation optics, in: *SPIE Micro+ Nano Materials, Devices, and Applications 2019*, Vol. 11201, International Society for Optics and Photonics, 2019, p. 112011L.
 - [44] M. Kazemzadeh, W. Xu, N. G. Broderick, Faster and more accurate time domain electromagnetic simulation using space transformation, *IEEE Photonics Journal* 12 (4) (2020) 1–13.
 - [45] M. Kazemzadeh, N. Broderick, W. Xu, Novel time-domain electromagnetic simulation using triangular meshes by applying space curvature, *IEEE Open Journal of Antennas and Propagation* 1 (2020) 387–395.
 - [46] P. Mao, C. Liu, G. Favraud, Q. Chen, M. Han, A. Fratalocchi, S. Zhang, Broadband single molecule sers detection designed by warped optical spaces, *Nature communications* 9 (1) (2018) 1–8.
 - [47] J. Hong, P. Dauros-Singorenko, A. Whitcombe, L. Payne, C. Blenkiron, A. Phillips, S. Swift, Analysis of the escherichia coli extracellular vesicle proteome identifies markers of purity and culture conditions, *Journal of Extracellular Vesicles* 8 (1) (2019) 1632099.
 - [48] H. Shin, D. Seo, Y. Choi, Extracellular vesicle identification using label-free surface-enhanced raman spectroscopy: Detection and signal analysis strategies, *Molecules* 25 (21) (2020) 5209.
 - [49] L.-J. Xu, C. Zong, X.-S. Zheng, P. Hu, J.-M. Feng, B. Ren, Label-free detection of native proteins by surface-enhanced raman spectroscopy using iodide-modified nanoparticles, *Analytical chemistry* 86 (4) (2014) 2238–2245.