Recent Progresses in Machine Learning Assisted Raman Spectroscopy

Yaping Qi, Dan Hu, Yucheng Jiang, Zhenping Wu, Ming Zheng, Esther Xinyi Chen, Yong Liang, Mohammad A. Sadi, Kang Zhang, and Yong P. Chen**

With the development of Raman spectroscopy and the expansion of its application domains, conventional methods for spectral data analysis have manifested many limitations. Exploring new approaches to facilitate Raman spectroscopy and analysis has become an area of intensifying focus for research. It has been demonstrated that machine learning techniques can more efficiently extract valuable information from spectral data, creating unprecedented opportunities for analytical science. This paper outlines traditional and more recently developed statistical methods that are commonly used in machine learning (ML) and ML-algorithms for different Raman spectroscopy-based classification and recognition applications. The methods include Principal Component Analysis, K-Nearest Neighbor, Random Forest, and Support Vector Machine, as well as neural network-based deep learning algorithms such as Artificial Neural Networks, Convolutional Neural Networks, etc. The bulk of the review is dedicated to the research advances in machine learning applied to Raman spectroscopy from several fields, including material science, biomedical applications, food science, and others, which reached impressive levels of analytical accuracy. The combination of Raman spectroscopy and machine learning offers unprecedented opportunities to achieve high throughput and fast identification in many of these application fields. The limitations of current studies are also discussed and perspectives on future research are provided.

1. Introduction

Raman spectroscopy is a sensitive and non-invasive measurement technique that has been extensively used in analytical sciences in the past decades.[1–5] Most commonly, it is based on the interaction between light and the vibration of chemical bonds in materials.[6] On a molecular level, the Raman effect occurs when light interacts with the electron density of the chemical bond, leading to vibrational excitation of the molecule and frequency shift of the light.[7] The Raman effect also occurs when light inelastically scatters and exchanges energy with excitations of materials such as characteristic lattice vibrations of solids. Thus, a vibrational fingerprint intrinsic to a particular molecule or material can be acquired, enabling its identification and characterization.[8–10]

Raman analysis has been increasingly employed in many fields, such as identifying unknown substances in material science, biology, pharmaceutics, and food science.[11–16] Although Raman spectroscopy is a powerful technique, the raw spectral data are often complex and contain a lot of random noise, requiring additional data processing techniques to extract valuable information.[17,18] However,

Y. Qi, D. Hu, Y. Liang, Y. P. Chen
Department of Engineering Science
Faculty of Innovation Engineering
Macau University of Science and Technology
Av. Wai Long, Macau SAR 999078, China
E-mail: ypqi@must.edu.mo; yongchen@purdue.edu

Y. Qi, Y. P. Chen
Advanced Institute for Materials Research (WPI-AIMR)
Tohoku University
Sendai 980–8577, Japan

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adom.202203104

DOI: 10.1002/adom.202203104

Y. Jiang
Jiangsu Key Laboratory of Micro and Nano Heat Fluid Flow Technology and Energy Application
School of Physical Science and Technology
Suzhou University of Science and Technology
Suzhou, Jiangsu 215009, China

Z. Wu
State Key Laboratory of Information Photonics and Optical Communications & School of Science
Beijing University of Posts and Telecommunications
Beijing 100876, China

M. Zheng
School of Materials Science and Physics
China University of Mining and Technology
Xuzhou 221116, China

E. X. Chen, K. Zhang
Faculty of Medicine
Macau University of Science and Technology
Av. Wai Long, Macau SAR China

conventional experimental and computational approaches can take a long processing time, and various qualitative evaluations [19] can lead to erroneous analysis results. They cannot keep up with the rapidly growing demands in multi-domain research of Raman spectroscopy.

With the advancement of artificial intelligence (AI) in recent years, machine learning (ML) has become an effective tool in analytical sciences.[20–22] ML is a part of AI that makes computers able to learn, make decisions or predictions, without the need to be explicitly programmed. ML algorithms iterate continuously on a given dataset to find a function that solves a specific task. AI is achieved by efficiently learning from a pre-labeled large amount of data to generate reasonable predictions on new sets of data, significantly speeding up experimental analysis and computation.[23,24] As a result, the use of ML has effectively contributed to a wide range of research, including the assisted processing of Raman spectroscopy data and its practical applications in a variety of fields.[25] For instance, ML can be applied to analyze complex and large amounts of Raman spectroscopy datasets, identify relationships, patterns, and connections in the datasets, as well as perform classifications.

To understand better about ML, it is necessary to know the major differences between ML and traditional chemometrics and statistical analysis. The latter (traditional methods) depend on statistical and mathematical models to analyze data and generate results or predictions. Traditional chemometrics and statistical analysis are usually utilized to perform qualitative and quantitative analysis of specific functional groups or chemical compounds and carry out feature selection, and data preprocessing in Raman spectra.[26,27] There are two main differences between ML algorithms and traditional chemometrics and statistical analysis methods. First, ML in many cases could be more efficient and capable than traditional chemometrics and statistical analysis methods in analyzing complex, large, and high-dimensional datasets as well as in identifying complicated patterns and connections in data, even without knowing or limiting to specific functional groups or chemical compounds.[26–29] Moreover, ML in most cases can perform task/data analysis without much experience and prior knowledge of the studied system, while traditional chemometrics and statistical analysis methods in general need prior knowledge and enough previous experience of the corresponding chemical system, and knowing which specific functional groups or chemical compounds are to be analyzed.[26,28]

Many ML algorithms have been developed and reported, some examples include decision trees, random forests, support vector machines, and artificial neural networks. Generally speaking, machine learning applied to chemical data such as Raman spectra may also be considered as a subset and more recent addition to chemometrics (which has a long history and many traditional techniques developed or employed predating machine learning). We note that machine learning is a method of teaching computers to learn from data, while traditional techniques such as principal component analysis (PCA), linear regression (LR), partial least squares (PLS) regression, least square (LS), linear or quadratic discriminant analysis (LDA, QDA), spectral preprocessing techniques including smoothing, baseline correction, normalization, etc are mainly based on mathematical and statistical models that are not necessarily learned from data. For example, PCA is extensively used for data dimensionality reduction, by constructing a set of principal components of orthogonal bases, taking only the first few principal components as input, and ignoring the rest. LS is a mathematical optimization technique that finds the best fit for a set of data by minimizing the sum of squares of the errors between each data point and a curve; LDA maximizes the axial component of interclass differentiation by projecting the feature space into a subspace of smaller dimensionality while maintaining information about the different classes. LDA can also be used for dimensionality reduction. These techniques have been an important part of traditional statistical and data analysis, while they can still be adapted and useful in the more recently developed ML methods (for example help to preprocess data, e.g. reducing dimensionality).

ML algorithms and some traditional data analysis methods including PCA, LR, PLS, LS, LDA, QDA, and spectral preprocessing techniques have been utilized to classify the spectra of unknown substances automatically.[30,31] These algorithms together with deep learning techniques provide new ideas for the classification and recognition of Raman spectra and have been subject to intense research in recent years.[32,33] Although there are extensive review papers on the application of machine learning in various areas, fewer reviews focus on the applications of machine learning in Raman spectroscopy. Therefore, this review does not only aim to summarize several traditional chemometrics and statistical methods that are commonly used in ML and some prevalent ML techniques applied in conjunction with Raman spectroscopy in diverse fields but is also committed to providing general background knowledge for readers interested in AI-assisted data analysis for Raman spectroscopy. Finally, current challenges and insights into the directions for future research in this area are presented.

2. Machine Learning for Raman Spectra Analysis

As a method to achieve AI, ML learns from a large amount of known data and then generates some reasonable predictions to expedite the process of experimental analysis and computation and save human resources, allowing ML to be widely used in a myriad of scientific fields.[34–36] To model and analyze Raman spectral data, ML algorithms are typically used in the classification and identification of Raman spectra. The study by Guo et al. provided good guidance from performing Raman spectroscopy to implementing ML modeling.[37] Spectral pre-processing and feature extraction are common steps in traditional ML analysis of Raman spectra.[19,38] Deep learning, a branch of ML research, is based on neural networks that rely on training data to learn and improve their accuracy.[39,40] Deep neural networks have the advantage of potentially avoiding pre-processing stage.[41]

M. A. Sadi, Y. P. Chen
Department of Physics and Astronomy and Elmore Family School of Electrical and Computer Engineering and Birck Nanotechnology Center and Purdue Quantum Science and Engineering Institute
Purdue University
West Lafayette, IN 47907, USA

Y. P. Chen
Institute of Physics and Astronomy and Villum Center for Hybrid Quantum Materials and Devices
Aarhus University
Aarhus-C 8000, Denmark

Table 1. Tools (terms) and descriptions that are commonly used in AI-assisted data analysis of Raman spectroscopy.

Tools (terms)	Descriptions
K-Nearest Neighbor (KNN)	A classifier that finds the K closest sample categories to a data point by calculating the distance between the data point and the known samples, and then obtains the final category output according to the majority voting.
Decision Tree (DT)	A tree structure in which the non-leaf nodes are the features selected for classification and the leaf nodes are the decision results.
Random Forest (RF)	Ensemble of decision trees, using multiple trees to train and predict the samples, and finally, output the results by voting or taking the average.
Support Vector Machine (SVM)	Find the hyperplane that distinguishes the different categories with maximum margins and separate the dataset into different categories by selecting appropriate support vectors.
Bayesian	A classification technique based on the Bayes theorem, where the prior probability distribution is selected and then updated to obtain the posterior distribution.
Artificial Neural Network (ANN)	A mathematical model that simulates the brain's neuronal activity as a set of connected input/output units, where each connection has a weight associated with it.
Convolutional Neural Network (CNN)	A class of feedforward neural networks with convolutional computation and deep structures. It is usually applied to analyze visual imagery.
Recurrent Neural Network (RNN)	A class of neural networks with short-term memory, suitable for processing a range of time-series related problems such as text.
Probabilistic Neural Network (PNN)	A neural network technique based on the Bayesian decision rule that is widely used in classification problems.
Generative Adversarial Network (GAN)	A novel adversarial generative model architecture that learns to generate new data with the same statistics as the training set.

Compared with traditional ML methods, deep learning has a highly promising learning ability and low generalization error.[42] In the following two subsections, the key properties and applications of some traditional chemometrics and statistical analysis methods that are also commonly used in ML, as well as ML-based algorithms, and deep learning methods will be introduced in detail to provide an overview of different tools, methods that have been used in this area, and to summarize the advantages and disadvantages of each technique. Some ML algorithms commonly used to assist Raman spectral analysis are shown in **Table 1**.

2.1. Examples of Traditional Statistical Analysis Methods Commonly used in ML and ML-Based Algorithms

Although some traditional chemometrics and statistical methods are not considered as ML algorithms because they are based on a set of fixed assumptions and parameters, they are still useful tools commonly used in ML. For example, the LR model is a commonly used analysis technique in chemometrics, including the more recently developed ML based methods.[43] It is primarily used to solve linear problems and can also be used for classification, with the LS method being a popular algorithm.[44] However, when the number of sample points is less than the number of variables, it is necessary to use the PLS method,[45] which enables regression modeling in the presence of multiple correlations in the independent variables. Moreover, LS/PLS method is also an optimization ideology, and it is usually used in combination with other algorithms in practical applications.[46–48] Another basic nonlinear classifier, the Bayesian classifier, solves classification problems with the Bayesian formula, which determines the sample as the class with the highest posterior probability and can handle multi-classification issues.[49–52]

Another example PCA is a commonly used technique for dimensionality reduction in ML.[53,54] It constructs a set of principal components of orthogonal bases from the original space by computing, where only the first few principal components are significant as input, and the rest are ignored.[55] Thus, the characteristics of the data are presented in a lower dimensional space using the new variables. The maximum variance theory is used to maximize the retention of the interpretation of original data.[56] ML models are constructed in most cases by starting with dimensionality reduction to reduce computational complexity. The dimensionality reduction maps high-dimensional data onto low-dimensional data, ensuring that it succinctly conveys similar information.[57] As decisions could not be made by PCA based on the data, it is instead a tool that could be utilized in combination with other ML techniques.[58] Therefore, PCA is just a supporting technique (previously and commonly used in classical statistical analysis) that can also be applied in ML. However, it is worth noting that in some cases PCA could be used as a model (i.e., in process analytical technology applications).[59]

A third example LDA projects the data to the hyperplane, which requires the introduction of labeled categories first to achieve dimensionality reduction while also classifying the data.[60] The goal of LDA is to maximize the axial component of interclass differentiation by projecting the feature space into a subspace of smaller dimensionality while maintaining information about the different classes. LDA and PCA are both commonly used methods for dimensionality reduction.[61] However, PCA is used primarily to find a better projection from the perspective of feature covariance, while LDA takes more into account the categorical label information to choose the direction with the best classification performance.[62]

2.2. ML and ML-Based Algorithms

On the other hand, a real machine learning model learns from data and can improve its performance over time and can adapt itself to the new data. Below are several examples of ML algorithms. First, K-nearest neighbor (KNN) is a nonlinear classifier and one of the most straightforward and intuitive classification algorithms.[63] It has no training process, instead it calculates the distance between samples using a distance metric formula and maps them to n-dimensional space.[64] KNN selects the K data points closest to the unknown samples, and votes on their categories, grouping the one with the highest percentage of categories among the K nearest neighbor samples.[65] Another classical algorithm, support vector machine (SVM), is widely regarded as one of the best classification algorithms,[66,67] which computationally solves the partitioning hyperplane that can correctly partition the training data set with the most significant geometric interval and separates the data set into different categories by selecting the appropriate support vectors.[68] However, for SVM, a model must first be trained on the training set before being used to classify the test set directly. Third, decision tree (DT),[69] as the term implies, is based on a tree structure to make decisions, similar to the mechanism of human decision making.[70] Based on a set of nested decision rules, the decision features are used to learn the rules in the dataset to partition the unknown dataset.[71] This algorithm can handle unrelated features in the dataset well but is prone to over-fitting.[72] On the other hand, the RF model based on the DTs has resistance to overfitting, as it randomly samples data with replacement, selects features, and ranks the importance of features by order of nodes, thus improving the resistance of the algorithm to interference.[73] It is, however, a more complex and computationally costly model than DT.

2.3. Deep Learning-Based Algorithms

Machine learning is a broad field that encompasses a wide range of techniques including DTs, random forests (RFs), SVMs and artificial neural networks (ANN), while deep learning is a specific type of artificial neural networks characterized by deep architectures. Deep learning is based on neural network models,[74] which do not require knowledge of the relationship between inputs and outputs or many parameters but instead learn the characteristics of sample data to obtain intrinsic information,[75] making the technique particularly suitable for processing fuzzy, stochastic, and nonlinear data.[76] As the basis of the neural network, ANN is a network structure composed of numerous interconnected processing units.[77] The network is trained through an iterative learning process to adjust and change the connection weights of neurons, process information, and simulate the relationship between inputs and outputs.[78] The subsequent series of deep learning algorithms are essentially based on the ANN derivatives, such as probabilistic neural network (PNN), convolutional neural network (CNN), recurrent neural network (RNN), generative adversarial network (GAN), etc.[79–85]

Deep learning algorithms based on neural networks, such as CNN and RNN, can learn features autonomously by optimizing the weights of each network layer.[80–83] In contrast, traditional chemometrics and statistical methods commonly used in ML such as PCA, PLS, and LDA require manual feature definition and input into the algorithm to obtain classification results. Such feature selection based on human experience is likely to miss some information, resulting in errors. The most significant distinction between deep learning and those traditional chemometrics and statistical methods used in ML is that performance improves as data size increases. This distinction also implies that when sample data is limited, traditional chemometrics, and statistical methods applied in ML may outperform deep learning algorithms.[84] However, with recent data augmentation methods, such as GAN,[85] it is possible to apply deep learning to obtain better results with limited samples. Furthermore, deep learning shows improved performance in completing complex tasks including natural language processing, image and speech recognition, and drug discovery. Therefore, for different algorithms, selecting the appropriate one based on the specific problem is critical.

3. Application of ML & Raman Spectroscopy in Materials Science

Raman spectroscopy reflects the molecular bonding in materials, making it a potent instrument for studying material structures in materials science.[86,87] It is frequently used for determining the composition and rapid classification of materials.[88] Combining the latest technological innovations in computer science with current methods of materials synthesis and characterization could significantly save costs and time for research and development in industry and academia.[89,90] The convergence of computing techniques and materials research,[91] and some AI algorithms, such as ML and deep learning techniques, can aid in identifying materials and comprehending material behaviors and properties more efficiently.[92] In a study by Boonsit et al.,[93] CNN can identify materials with an accuracy of up to 96.7%, even with low-resolution Raman spectra. Pan et al.[94] proposed an algorithmic model for multi-label classification based on deep learning to recognize complex mixture materials. Xie et al.[95] proposed an algorithm for automatic material identification using ML methods for Raman spectroscopy, which can potentially be embedded into the operating software of Raman spectrometers for practical applications. **Table 2** summarizes some examples demonstrating the application of ML algorithms to Raman spectroscopy of materials.

3.1. ML-Assisted Raman spectroscopy of Nanomaterials

In this section we review applications of ML-assisted Raman analysis to study nanomaterials, both 2D materials (such as graphene), as well as 1D materials (such as nanotubes).

Graphene, the first 2D material to be discovered, has excellent optical, electrical, and mechanical properties and promising multidisciplinary applications, making it a hotbed of research in recent years.[96–98] A study by Jo et al.[99] concentrated on identifying peak shifts in Raman spectroscopy and combining them with PCA to build a framework that can predict the thickness of a few graphene layers. There is an optimal range of feature sizes where the estimation performs best in this work. Sirico et al.[100] reported a method that combines optical contrast microscopy [101] with an ML algorithm proposed by Lin et al.[102] to determine the thickness of 2D materials. In this experiment, after training the algorithm on a region under the same lighting conditions, it can be applied to the entire sample without additional training, making it ideal for accurately characterizing 2D materials over expansive areas.

Table 2. Latest developments in combining ML methods with Raman spectroscopy for materials research.

Objectives	Methods	Results	References
Identify the number of graphene layers	PCA	Accuracy is below 90% for a small number of features, shows 100% accuracy between 5 and 14 features, and again indicates below 90% using >14 features.	Jo et al.[99] (2018)
Identify the thickness of 2D materials	RF	The classifier provided excellent results in terms of overall accuracy (96.4%) and training speed (8 sec).	Sirico et al.[100] (2021)
Classify graphene Raman spectra according to different charge densities and dielectric environments	CNN	The spectral classification with 99% accuracy using a convolutional neural network (CNN) model.	Chen et al.[105] (2022)
Remove noise from graphene Raman spectra	Supervised/ unsupervised deep neural networks	Compared with traditional methods, supervised and unsupervised models are 50% and 36% better than traditional models.	Machado et al.[106] (2022)
Identify the twist angle of twisted bilayer graphene (tBLG)	PCA, MLR (RF, DT...)	MLRs achieved high R2 scores of 0.98 ± 0.04 and an average root mean square error (RMSE) of twist angle is 0.70° ± 0.93°.	Sheremetyeva et al.[107] (2020)
Identify the twist angle of tBLG	RF	Achieve >99% accuracy in labeling twist angles.	Pablo and Hiroki [108] (2022)
Identify suspended carbon nanotubes (CNTs)	CNN	More than 90% accuracy can be achieved even with a low spectral signal-to-noise ratio.	Zhang et al.[111] (2022)
Distinguish single-layer continuous films and random defect regions of 2D materials	RF	The numerical values of area under the curve (AUC) and average precision (AP) are 0.9852 and 0.9867 for cracks, and 0.9902 and 0.9914 for bilayer.	Mao et al.[112] (2020)
Monolayer detection and 3D characterization of MoS_2	SVM, KNN, RF	The classification accuracy of MoS_2 samples is up to 99.2%.	He et al.[113] (2021)
Distinguish phases of matter	SVM, PCA	For the Orthorhombic-Tetragonal-Cubic and Ferroelectric-Paraelectric phase transition, the best cross-validation accuracy is 98.7% and 99.7%, respectively.	Cui et al.[114] (2019)
Classification of variscite samples from the Gavà mining complex to determine the origin and depth of mining	LR, SVM, LDA, DT, RR	The best result for the accuracy of determining the mine of origin is obtained by SVM (98%).	Díez-Pastor et al.[123] (2020)
Identification of large categories of minerals.	Siamese network	The accuracy of the Siamese network is slightly better than traditional ML algorithms. The accuracy for the Siamese network, SVM, and KNN are 62.27%, 59.75%, and 57.56%, respectively.	Wu et al.[124] (2020)
Recognize minerals	CNN	The max accuracy of the test set is 98.43%, and the average is 97.72%.	Sang et al.[126] (2022)
Classify the plastics	SVM, ANN, PCA	PCA-SVM implementation demonstrated excellent accuracy above 95%. ANN achieved a high accuracy of close to 100% after a few hundred epochs.	Musu et al.[129] (2019)
Raman imaging visualization of microplastics	PCA	PCA can automatically extract and decode the critical information from the spectrum matrix for imaging without referring to standard Raman spectra.	Fang et al.[135] (2022)

(MLR: Machine Learning Regressor, LR: Logistic Regression, RR: Ridge Regression.)

In addition, parameters such as peak position and peak width from Raman spectra of graphene are frequently used to evaluate the strain and charge doping levels.[103,104] Chen et al.[105] used CNN to classify graphene samples with slightly different charge densities or dielectric environments and enhanced the spectra data by adding noise and peak shifting. **Figure 1**a,b,c illustrates the prediction of graphene doping levels using a CNN model, and the flow chart of this experimental design is shown in Figure 1d. Experiments showed that the CNN model can classify the Raman spectra of graphene with different charge doping levels with 99% accuracy and even detect subtle differences in the spectra of graphene on SiO_2 and graphene on silanized SiO_2.[105] Machado et al.[106] proposed two approaches, one is a deep neural network with an autoencoder architecture, and another consists of a fully convolved autoencoder. They were used to remove noise from Raman spectra and improve graphene spectral data quality. These two deep neural network-based methods can significantly improve the conditions for analyzing Raman spectroscopy data from graphene nanosheets.[106] It can also be extended for future use in processing spectral data for various materials.

Sheremetyeva et al.[107] presented a computational framework for identifying the twist angle of twisted bilayer graphene

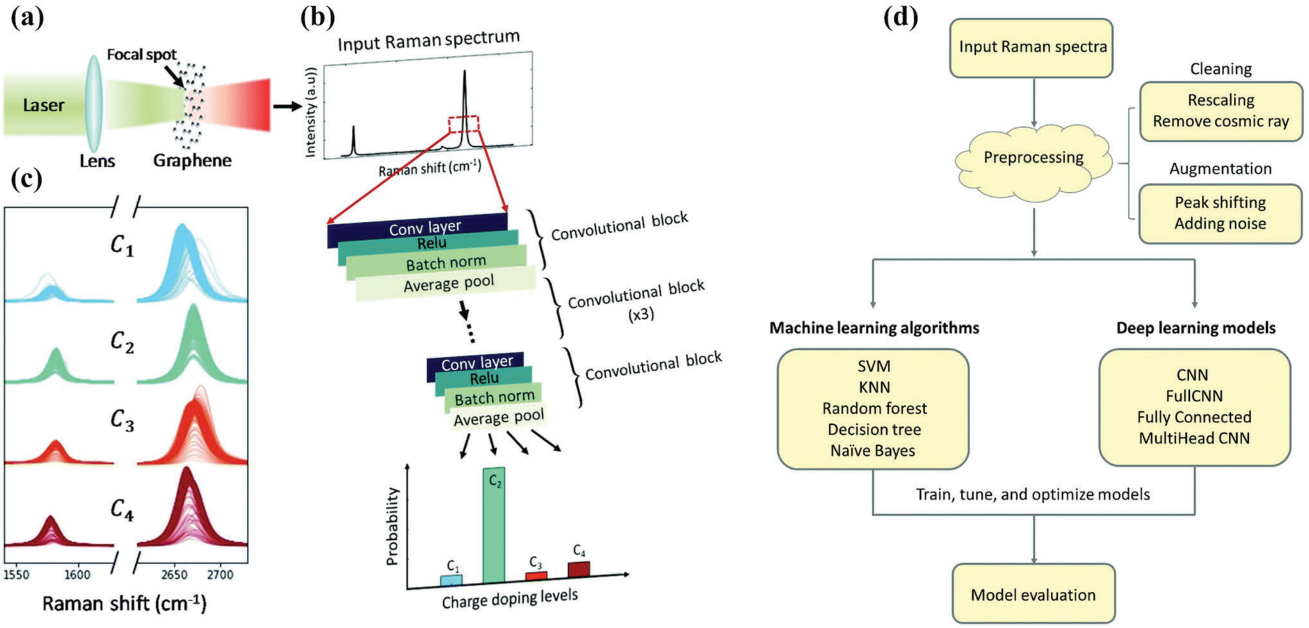


Figure 1. a) Schematic diagram of Raman spectroscopy measurements on graphene samples. b) Representative Raman spectra of graphene. Spectra were fed into a 1D CNN with five convolution blocks and categorized into one of four classes corresponding to four different charge doping levels of the graphene samples. c) Raman spectra of graphene samples with four different charge doping levels. d) Experimental flow chart. After preprocessing, five different (conventional) ML models and four different deep learning models were implemented. Adapted with permission.[105] Copyright 2022, Royal Society of Chemistry.

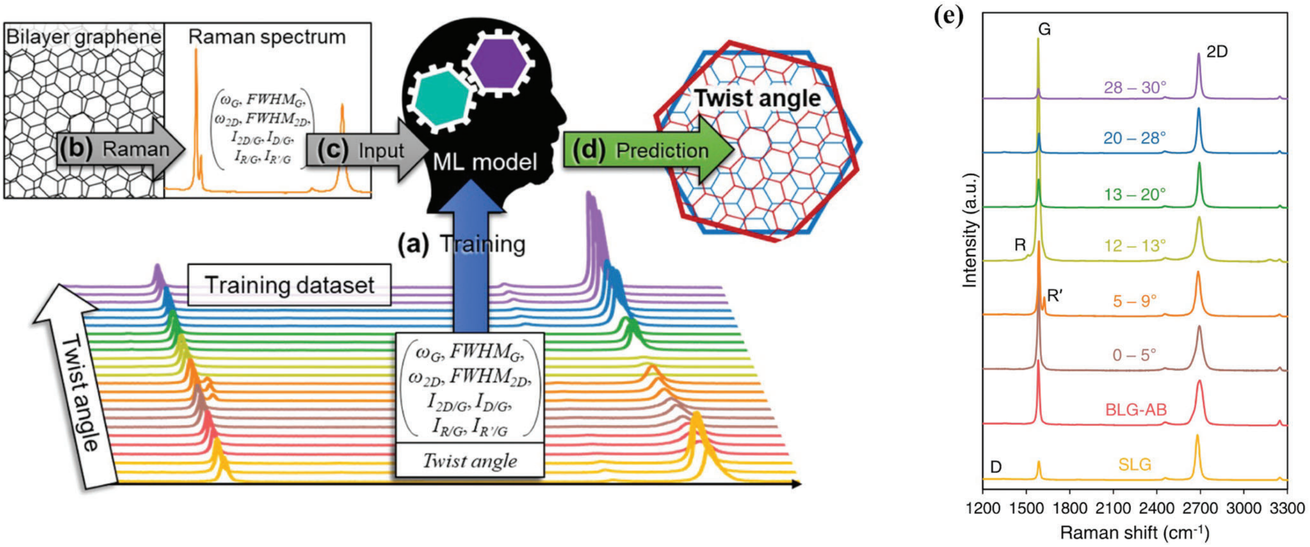


Figure 2. a) ML model is trained using a dataset of Raman spectra with predetermined twist angles. b) Raman features are collected from different tBLG samples to determine their twist angles. c, d) Trained models predict the twist angle of other BLG samples based on their Raman features. e) Average spectra of SLG, BLG-AB, and tBLG with different twist angles. Adapted with permission.[108] Copyright 2022, ACS Publications.

(tBLG) from Raman spectra. After reducing the dimensionality of the data with PCA, machine learning regressors (MLRs) were successfully used to make predictions and compare performance.[107] By analyzing the features learned by the MLRs, the experiments determined that the intensity profile close to the G-band was the most significant feature. Similarly, Solís-Fernández and Ago [108] predicted the twist angle of a tBLG based on its Raman spectrum using an RF algorithm. **Figure 2**a–d depicts the process of determining the BLG twist angle with ML, and Figure 2e illustrates the average spectra at different twist angles. The spectra are normalized to 2D band intensities and shifted vertically, and the Raman spectra are labeled into corresponding bands of twist angle.[108] The proposed method can deliver 99% accuracy in labeling twist angles. Such highly accurate

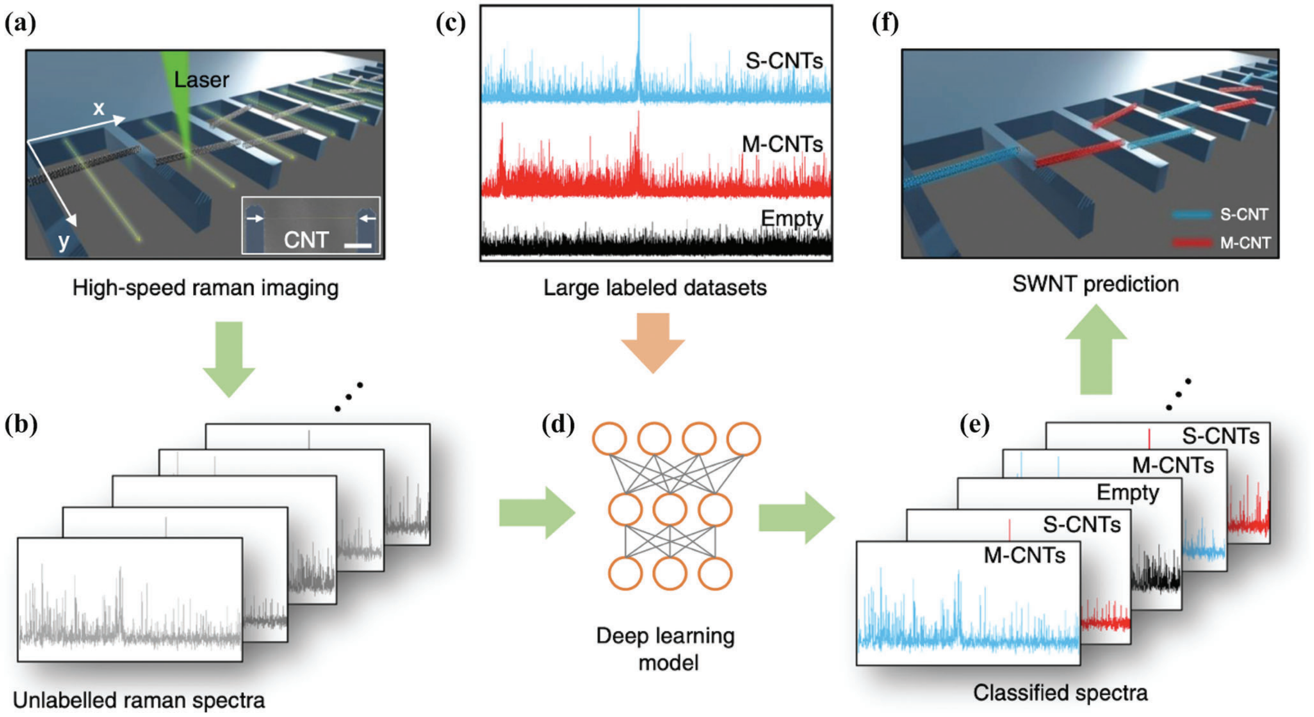


Figure 3. a) High speed Raman spectra collection method for carbon nanotubes (CNTs). b) Generation of unlabeled Raman spectra. c) Labeled datasets are organized into three classes: S-CNTs, M-CNTs, and empty. d) Illustration of a CNN model. e) Identification of unlabeled spectra using the trained model. f) Prediction. Reproduced with permission.[111] Copyright 2022, Springer Nature.

predictions are expected to facilitate the exploration of emerging research areas on stacked and twisted van der Waals heterostructures.

Carbon nanotube (CNT) is a 1D nanomaterial with exceptional mechanical, electrical, and chemical properties, and promising avenues of application.[109,110] In a recent study, Zhang et al.[111] proposed a high-throughput method for rapidly identifying suspended carbon nanotubes (CNTs) employing high-speed Raman imaging and CNN, as depicted in **Figure 3**. This work used a large dataset of CNTs Raman spectra comprising of labels metallic CNTs (M-CNTs), typical metallic CNTs (S-CNTs), and Empty as the source for training sets for CNN models.[111] The trained model was then utilized to predict the label of new unlabeled spectra. Even with a low signal-to-noise ratio, the method can classify the samples with an accuracy of >90%.

Mao et al.[112] pioneered using ML and Raman spectroscopy to distinguish between single-layer continuous films and random defect regions in 2D materials (**Figure 4**a, b). Based on the position-dependent Raman information of molybdenum disulfide (MoS_2) films, an RF algorithm is used to search for possible hidden correlations between the sample types and the features obtained from the spatial Raman mapping to distinguish the continuous monolayer film from the defective regions [112] (Figure 4c, d). Due to low visual contrast, defects are frequently overlooked in optical microscopy. Therefore, this classification procedure of 2D material defects highlights the advantages and potential of ML as an alternative to traditional analysis methods. Subsequently, He et al.[113] proposed an ML-based method for monolayer detection and 3D characterization of MoS_2. The property that the optical intensity depth of MoS_2 samples captured under a linearly adjustable light source is proportional to the color depth can be utilized to extract the characteristics of questionable monolayer MoS_2 samples effectively.[113] The target value dataset is established based on the Raman spectra of the samples, and SVM is used to classify the monolayer MoS_2 samples, with classification accuracy up to 99.2%.

Cui et al.[114] used the SVM method to mine and learn the behavior vectors of the phonon vibrations in a crystalline lattice from Raman scattering, recognized the orthorhombic, tetragonal, and cubic phases, and constructed the phase diagram in ferroelectric crystals. This study presents a tool commonly used to detect structural properties at the molecular level, which provides the basis for applying generic methods to predict undeveloped structures and materials.

3.2. ML-assisted Raman Spectroscopy of Minerals

Non-destructive Raman spectroscopy has been widely used in mineral analysis.[115–119] Whereas minerals are used as primary structural materials, this makes it also plays a significant role in archaeology.[120,121] Recently, some research has used ML in conjunction with Raman spectroscopy to identify minerals. Carey et al.[122] proposed a full-spectrum matching algorithm for mineral identification and classification in Raman spectroscopy. The experiments demonstrated excellent performance in classification tasks without the need for expensive dimensionality

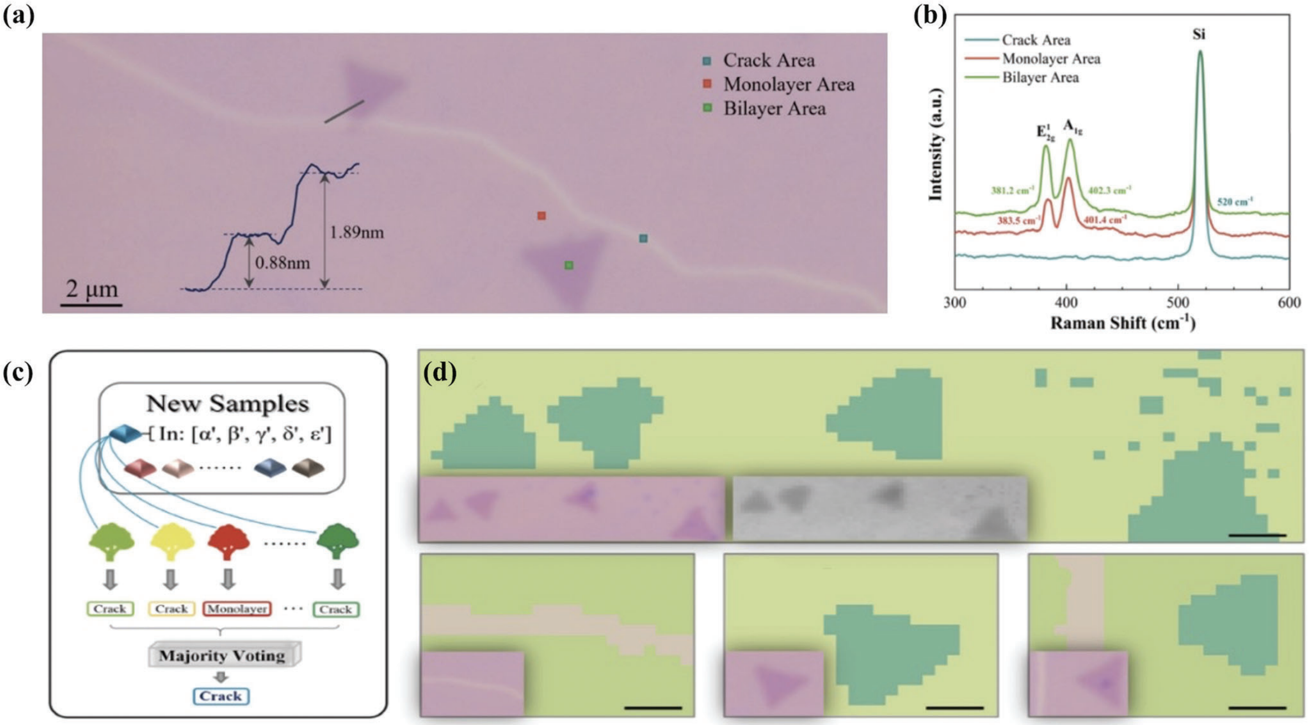


Figure 4. a) Optical image of the MoS_2 sample. The MoS_2 monolayer continuous film has a relatively smooth surface; the line and the dark triangular areas (a cross sectional height profile is shown as inset) are cracks and double layer regions, respectively. b) Raman spectra of the monolayer, crack, and bilayer regions. c) The basic architecture of the prediction procedure in the RF method, each small square represents a spatial measurement point carrying characteristic information extracted from Raman. d) The predicted pictures for different samples with crack(brown), monolayer (grass green), and bilayer (dark green) areas. Reproduced with permission.[112] Copyright 2020, MDPI.

reduction or model training.[122] In another study, ML is used to classify variscite samples, which were once used as gemstones, from the Gavà mining complex to determine their origin and mining depth.[123] The SVM algorithm achieved the highest classification accuracy in this task.

Wu et al.[124] discovered that traditional ML methods perform poorly when dealing with minerals with many categories. To address the issue, they proposed a similarity learning method based on the Siamese network. The Siamese network was optimized using the Hungarian algorithm [125] for negative samples to improve mineral identification accuracy and calculate the similarity between minerals. As a result, it outperforms traditional ML algorithms in terms of robustness.

Sang et al.[126] proposed a 1D deep CNN-based classification model for spectral data that can classify and identify hundreds of mineral categories. For example, the model can identify mineral Raman spectra in the RRUFF dataset.[127] Compared with other traditional ML methods and CNN models, this work has improved accuracy, precision, and recall performance.

3.3. ML-assisted Raman Spectroscopy of Plastics

To reduce the negative impacts of plastic waste on the environment, Raman spectroscopy has been evaluated as a method for identifying some common solid plastics.[128] Musu et al.[129] classified the plastics using SVM and ANN algorithms. The PCA method was employed to reduce the data dimensions of the basic spectral peaks to simplify the computation. In the evaluation phase, PCA-SVM model demonstrated excellent accuracy and robustness, with recognition accuracy remaining above 95% even when the noise was increased to three times the original level.[129] In addition, the ANN model achieves a high recognition accuracy, ≈100%, after a few hundred epoch calculations, but it takes more time than PCA-SVM model.[129] This study assesses the use of Raman spectroscopy-based ML techniques in plastic identification and demonstrates the feasibility of accurate and rapid plastic recycling classification.

Furthermore, microplastics can also cause potential negative impacts on the environment, and the problem has attracted increasing attention.[130] Numerous studies have been initiated to address the issue of microplastic pollution.[131–134] Fang et al.[135] proposed two methods for the visualization of microplastics using Raman imaging, including a logic-based algorithm that combines several images mapped with multiple feature peaks into a single image to improve image determinism. Another method is to decode the spectral matrix using PCA algorithm, which is appropriate for cases with complex samples and a lack of standard spectra.[135] Experimental results indicated that logic-based algorithms may suffer from signal loss when merging multi-peak images, whereas PCA algorithms have the potential to analyze large Raman spectral matrices efficiently.

4. Application of ML & Raman Spectroscopy in Biomedicine

While numerous methods for screening and diagnosing diseases are used in the medical field, clinical requirements for early and precise disease detection are not easily satisfied.[136,137] Raman spectroscopy has recently been applied in vitro and in vivo to solve various biomedical problems.[138–140] As early as decades ago, studies have been done on the potential of Raman spectroscopy as a new tool for biomedical applications, from proof of principle to clinical implementation.[141,13] Raman spectroscopy combined with ML could provide a low-cost, rapid, and noninvasive method for biomedical identification and diagnosis.[142] Recent biomedical applications of this combination are illustrated in **Table 3** and **Table 4**.

4.1. ML-Assisted Raman Spectroscopy for Medical Diagnosis

In a recent study, researchers combined tear-based Raman spectroscopy with ML to diagnose cerebrovascular disease, enabling a rapid, noninvasive classification of cerebral infarction and cerebral ischemia.[143] Analysis of Raman spectroscopy data revealed differences in tyrosine, phenylalanine, and carotenoid levels in the tears of patients with cerebral ischemia and patients with cerebral infarction, which provide a foundation for early screening of patients with cerebrovascular disease. A total of 12 classification models (different combinations of the seven algorithms listed in Table 3) established in the experiment have high accuracy in diagnosis, with PLS-PNN performing the best. This study demonstrates that tear Raman spectroscopy has enormous potential for diagnosing patients with cerebral infarction and ischemia.

Sciortino et al.[144] extracted 2073 Raman spectra from 38 glioma specimens using an eXtreme Gradient Boosting (XGB) and SVM with radial basis function on kernel (RBF-SVM) to classify tumor types. They analyzed the ability of Raman spectra to detect mutation types in unprocessed glioma biopsies, which is essential in surgical guidance as well as intraoperative diagnosis. A recent study deployed Raman spectroscopy with ML algorithms and detected gliomas (**Figure 5**a–d) in surgical scenarios and identified 19 new Raman shifts with known biological significance.[145] According to the findings, Raman spectroscopy combined with supervised ML techniques can distinguish between normal and tumor tissue in fresh samples in vitro. Figure 5e depicts the average spectra and deviations for healthy and tumor patients, with arrows indicating the new Raman peaks. The incorporation of ML further supports the development of real-time tissue analysis using Raman spectroscopy in tumor brain surgery.

Alzheimer's disease (AD) is the most prevalent form of dementia afflicting older adults worldwide. The clinical symptoms of AD have progressed from mild memory loss to severe cognitive impairment,[146] making it one of the most concerning health disorders. The optimal time to diagnose AD is in the preliminary stages of its progression.[147] However, current diagnostic options, such as clinical assessment, are only valid in the late stages of the disease. Therefore, a rapid and effective method for diagnosing AD is urgently needed. Some research attempts to detect AD by analyzing the Raman spectra of various body fluids. Ralbovsky et al.[148] developed a new method for AD diagnosis based on saliva analysis. This study classified saliva samples from normal individuals, patients with AD, and patients with mild cognitive impairment using Raman spectroscopy and genetic algorithms and ANN. The accuracy is 99%, indicating that salivary Raman spectroscopy can be used effectively for early AD diagnosis.

In another work, AD was also diagnosed using near-infrared spectroscopy of cerebrospinal fluid combined with ML analysis.[149] Ralbovsky et al.[150] described a new screening method for determining the risk of AD. The Raman spectroscopy analysis of serum from rats fed on standard and high-fat diets revealed that high-fat diet rats exhibited a pre-AD state. The experimental results demonstrated that the PLS discriminant analysis used for the classification distinguished the two rat groups with 100% accuracy at the donor level.

Furthermore, to address the impact of the intrinsically weak Raman signal on Raman-assisted identification of biomolecules, Huang et al. reported a method to enhance the Raman signal by reducing Raman spectral noise with graphene.[151] The method was applied in a recent study of rapid screening for AD by ML and graphene-assisted Raman spectroscopy.[152] **Figure 6**a illustrates the overall workflow of the analysis. The experiment first collected Raman spectra on brain slices of mice with and without AD (Figure 6b, c) and then used ML to classify the collected spectra. The Raman measurements were performed by contacting a single layer of graphene with the brain slices and thus achieving noise reduction of the Raman spectra of brain tissues to improve the Raman signal-to-noise ratio and increase the accuracy of the ML classification from 77% to 98%.[152] In general, Raman spectroscopy has the potential to be used in the future as a method to identify AD at an early stage of its progression, which has significant implications for the early prevention and treatment of this disease.

The demand for large-scale diagnostic tests has increased in recent years due to the impact of the COVID-19 epidemic.[153] However, the commonly used detection methods are still either labor-intensive or not sufficiently accurate.[154] Rapid and precise detection of COVID-19 is essential for the rational allocation of healthcare resources. Chen et al.[155] constructed a stacked subcode classifier based on eight ML algorithms as a classification tool for serum Raman spectral data to predict the infection status of COVID-19. This classification task achieved an accuracy of 98%. Similarly, Yin et al.[156] collected 177 serum samples from patients with confirmed COVID-19, suspected cases, and healthy individuals for Raman spectroscopy analysis (**Figure 7**a, b). They constructed the corresponding diagnostic algorithm using SVM, and the receiver operating characteristic curve is shown in Figure 7c. This experiment of serum level classification results was correct for all independent test data sets. Ember et al.[157] developed a saliva-based, reagent-free method for detecting COVID-19. They used Raman spectroscopy and ML to detect and analyze changes in the molecular profile of saliva associated with COVID-19 infection.[157] These studies demonstrate the prospect of a rapid Raman spectroscopy screening tool for medical detection.

Much research on Raman spectroscopy and ML in disease has focused on cancer detection.[158] However, the results of many cancer screenings still rely on the expertise of physicians. Researchers have recently attempted to develop more efficient

Table 3. Latest developments in combining ML methods with Raman spectroscopy for medical diagnosis.

Objectives	Methods	Results	References
Classification of cerebral infarction and cerebral ischemia	PCA, PLS, MRMR, SVM, KNN, PNN, DT	The classification accuracy of all models is above 85%. Especially PLS-PNN has achieved 100% accuracy.	Fan et al.[143] (2022)
Classification of glioma types	XGBoost, RBF-SVM	Both the accuracy and precision of distinguishing between IDH-MUT and IDH-WT tumors are 87%.	Sciortino et al.[144] (2021)
Classification of glioma biopsies	RF, GB	The accuracy and precision of distinguishing between tumors and healthy brain tissue are 83% and 82%, respectively.	Riva et al.[145] (2021)
Alzheimer's disease (AD) diagnosis based on saliva analysis	ANN	The classification accuracy of salivary Raman spectroscopy samples from normal, AD, and patients is 99%	Ralbovsky et al.[148] (2019)
Identify the risk of AD	PLS	The experimental results showed that the PLS discriminant analysis is used for the classification with 100% accuracy at the donor level.	Ralbovsky et al.[150] (2021)
Rapid screening of AD	SVM, RF, XGBoost, CatBoost	By contacting the brain slices with a single layer of graphene to improve the Raman signal-to-noise ratio of brain tissue and increased the accuracy of classification from 77% to 98%.	Wang et al.[152] (2022)
Predict the infection status of COVID-19 based on serum Raman spectra	DT-based ExtraTrees, LR, KNN, SVM, AB, GB, RF, MLP	The classification accuracy for 10-fold cross-validated is 98.0%, precision is 98.6%, and recall is 98.5%.	Chen et al.[155] (2021)
Predict the infection status of COVID-19 based on serum Raman spectra	SVM	The classification accuracy between the COVID-19 cases and the suspected cases is 87%, for COVID-19 and the healthy controls is 90%. In comparison, the accuracy between the suspected cases and the healthy control group is 68%.	Yin et al.[156] (2021)
Saliva-based detection of COVID-19 infection	MILES	Even taking into account the gender of the saliva donor, the maximum AUC in the study is 0.80.	Ember et al.[157] (2022)
Classify breast cancer subtypes	PCA-DFA, PCA-SVM	Identification of breast cancer cells and classification of cancer cell subtypes at the single cell level with a classification accuracy of over 97%.	Zhang et al.[159] (2022)
Diagnosis of lung cancer	STFT based CNN	The average accuracy of test groups is 96.5% ± 0.7%.	Qi et al.[160] (2021)
Classification of lung cancer based on serum Raman spectra	Improved ResNeXt	The improved ResNeXt model achieved the best results with accuracy, sensitivity, specificity, and AUC values of 0.968, 0.992, 0.951, and 0.973, respectively.	Leng et al.[161] (2022)
Lung cancer diagnosis based on exosome Raman spectra	ResNet based deep learning model	The model predicted lung cancer with an AUC of 0.912 and an AUC of 0.910 for early-stage patients.	Shin et al.[163] (2020)
Screening of cervical adenocarcinoma and cervical squamous cell carcinoma tissue	airPLS, PLS, PCA, KPCA, KNN, ELM, BPNN, GA-BPNN, LDA…	The airPLS-PLS-KNN algorithm has the highest accuracy rate of 96.3%.	Zhang et al.[164] (2021)
Screening of ovarian cancer	BPNN, PCA	The sensitivity and specificity of cancer detection are 81.0% and 97.3% among normal, cyst, and cancer samples, respectively.	Chen et al.[166] (2022)
Identify blood species	RNN	The recognition accuracy of bidirectional RNN with GRU is 97.7%.	Wang et al.[167] (2021)
Predict gastric cancer	CNN, RF, SVM, KNN	RF had the best performance with an accuracy of 0.928, sensitivity and specificity of 0.947 and 0.908, and AUC of 0.9199.	Li et al.[168] (2021)
Investigate the long-term treatment of G-CSF on colon and breast cancers	PCA, LDA	The classification accuracy is 69.7%±11.8% for 4T1 cells and 67.5%±10.7% for CT26 cells.	Zhang et al.[170] (2021)
Identification of kidney tumor tissue	SVM	The classification accuracy is 92.89%.	He et al.[171] (2021)

(MRMR: Minimum Redundancy Maximum Relevance, XGBoost: eXtreme Gradient Boosting, RBF: Radial Basis Function, GB: Gradient Boosting, CatBoost: Categorical Boosting, AB: Adaptive Boosting, MLP: MultiLayer Perceptron, STFT: Short-time Fourier Transform, ResNet: Residual Neural Network, MILES: Multiple Instance Learning via Embedded Instance Selection, DFA: Discriminant Function Analysis, airPLS: Adaptive Iteratively Reweighted Penalized Least Squares, KPCA: Kernel Principal Component Analysis, ELM: Extreme Learning Machine, BPNN: Backpropagation Neural Network, GA: Genetic Algorithm.)

methods to assist physicians in cancer screening. Zhang et al.[159] adopted Raman spectroscopy combined with PCA–DFA and PCA–SVM to simplify and accelerate the process of distinguishing normal from breast cancer cells, and classifying breast cancer subtypes. The results showed that the proposed algorithm can identify breast cancer cells and classify cancer cell subtypes at a single-cell level with an accuracy of over 97%.

Qi et al.[160] proposed a new method of short-time Fourier transform (STFT)-based CNN combined with Raman spectroscopy to analyze lung tissues to diagnose lung cancers. When dealing with large data volumes of samples, the STFT-based CNN approach can provide a more accurate means of tissue section inference in pathology. Leng et al.[161] analyzed and verified the changes of substance components in the serum of patients

Table 4. Latest developments in combining ML methods with Raman spectroscopy for pathogens in biomedicine.

Objectives	Methods	Results	References
Analysis of Raman spectra of human and avian viruses	CNN	The accuracy of the binary classification of influenza viruses is 99%, for the four subtypes of influenza A is 96%, and 95% for the enveloped and non-enveloped viruses.	Ye et al.[172] (2022)
Identification of Burkholderia mallei and Related Species	SVM, PCA	Identification accuracy of >90% could be achieved on the spectra level.	Moawad et al.[173] (2019)
Detection of bacteria	CNN	Using a 30-class bacterial isolate dataset for training and testing, the ML model achieved recognition accuracy of ≈86% and recognition speed close to real-time.	Kukula et al.[174] (2021)
Identification of Marine Pathogens	RNN (LSTM)	The classification accuracy of the RNN using LSTM method is over 94%.	Yu et al.[175] (2021)
Detection of food-borne pathogens	KPCA, DT	The classification accuracy is in the range of 87.1%–95.8%.	Yan et al.[176] (2021)
Identification of E. coli strains	ANN, SVM	Even with a limited data set, SVM achieved an average accuracy of 98.8%. When the data set is large enough, ANN can achieve 100% accuracy.	Zahn et al.[177] (2022)
Distinguish between resistant and sensitive E. coli strains	SVM, LDA	For the Raman microspectroscopy data, the accuracy is 75%, correctly classifying 15/20 strains.	Nakar et al.[178] (2022)
Diagnosis of bacterial pathogens at the single-cell level	Neural Network-based DAE	Demonstrated 92% (simple filter using 1 s/cell spectra) and 84% (DAE using 0.1 s cell spectra) identification accuracy.	Xu et al.[183] (2022)
Identification of pathogenic bacteria	CNN	The identification accuracy is 99.7%.	Ho et al.[184] (2019)

(LSTM: Long and Short-term Memory, DAE: Denoising Autoencoders.)

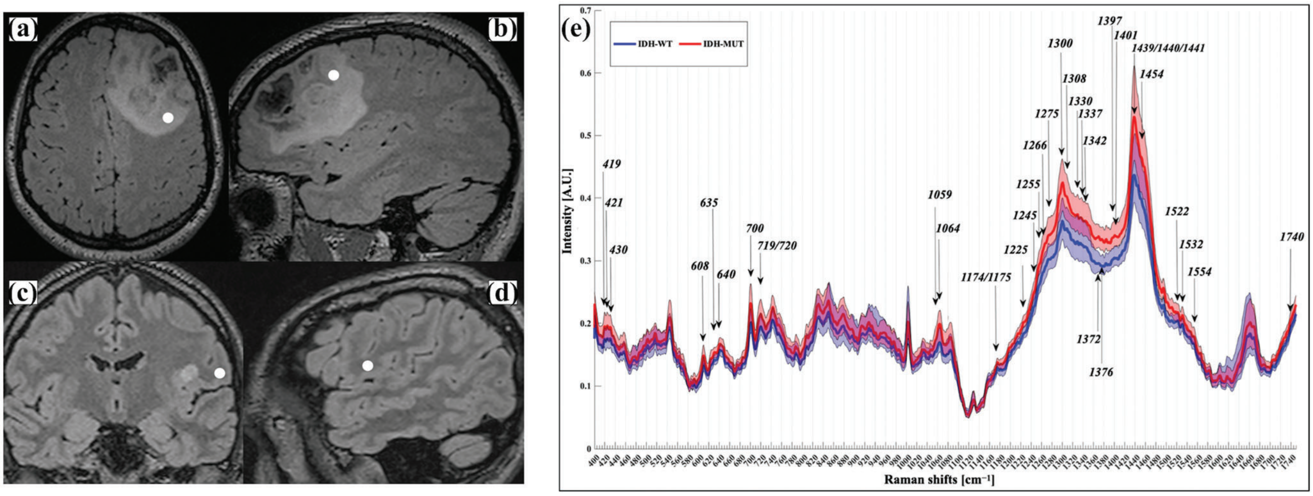


Figure 5. a, b) Axial and sagittal views of preoperative MRI showed tumor consistent with Anaplastic Oligodendroglioma IDH-1 mutant. The white spots showed the intraoperative site of tissue biopsies registered in the neuro navigation system and labeled as tumors. c, d) Preoperative MRI showed the intraoperative location of tissue collection labeled as healthy. e) Normalized mean spectra (curves) and standard deviations (shaded bands) for healthy (blue) and tumor patients (red). The arrows mark the new Raman peaks. Reproduced with permission.[145] Copyright 2021, MDPI.

with lung cancer and found significant differences between lung cancer patients and normal controls in major components in serum such as phenylalanine, β-carotene, and cholesterol. As shown in **Figure 8**c, the peaks of β-carotene at 1157 and 1517 cm^{-1} were significantly lower in lung cancer patients than in controls, while the peak of cholesteryl ester at 1669 cm^{-1} was higher than in controls. This experiment proposes a model based on an improved ResNeXt (Figure 8b) better suited for processing spectral data than the original model (Figure 8a), resulting in an accurate classification of serum Raman spectra of lung cancer patients. In addition, exosomes have been used as promising biomarkers for liquid biopsies.[162] Shin et al. achieved an accurate diagnosis of early-stage lung cancer based on a deep learning method assisted SERS of exosomes.[163] The authors isolated exosomes from cell culture supernatants and human plasma samples and collected their Raman spectra. The deep

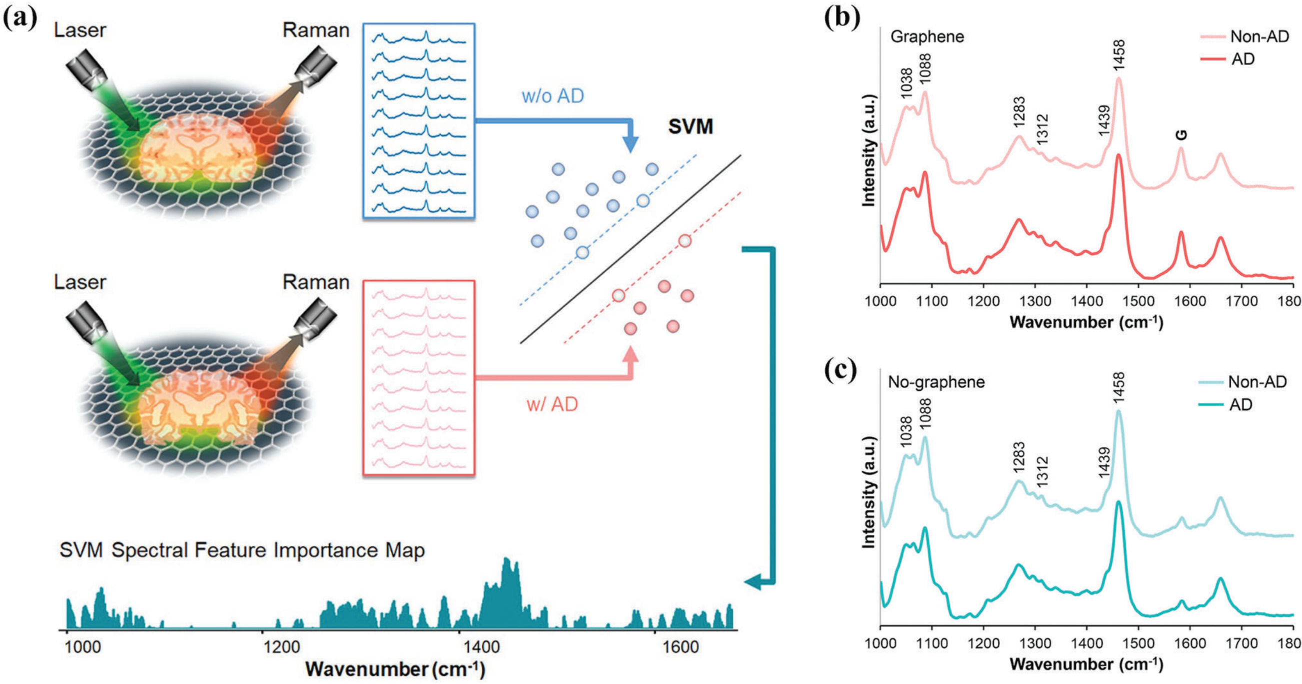


Figure 6. a) Workflow for data collection, pre-processing, ML classification and interpretation of graphene-assisted Raman signals. The demonstrated ML classifier is a linear SVM model for differentiating AD/non-AD Raman spectra. b) Raman spectra in the cortex region of brain slices preprocessed using graphene with and without AD. The G-band of graphene at 1589 cm^{-1} is notated as "G". c). Raman spectra in the cortex region of brain slices preprocessed not using graphene with and without AD. Reproduced with permission.[152] Copyright 2022, ACS Publications.

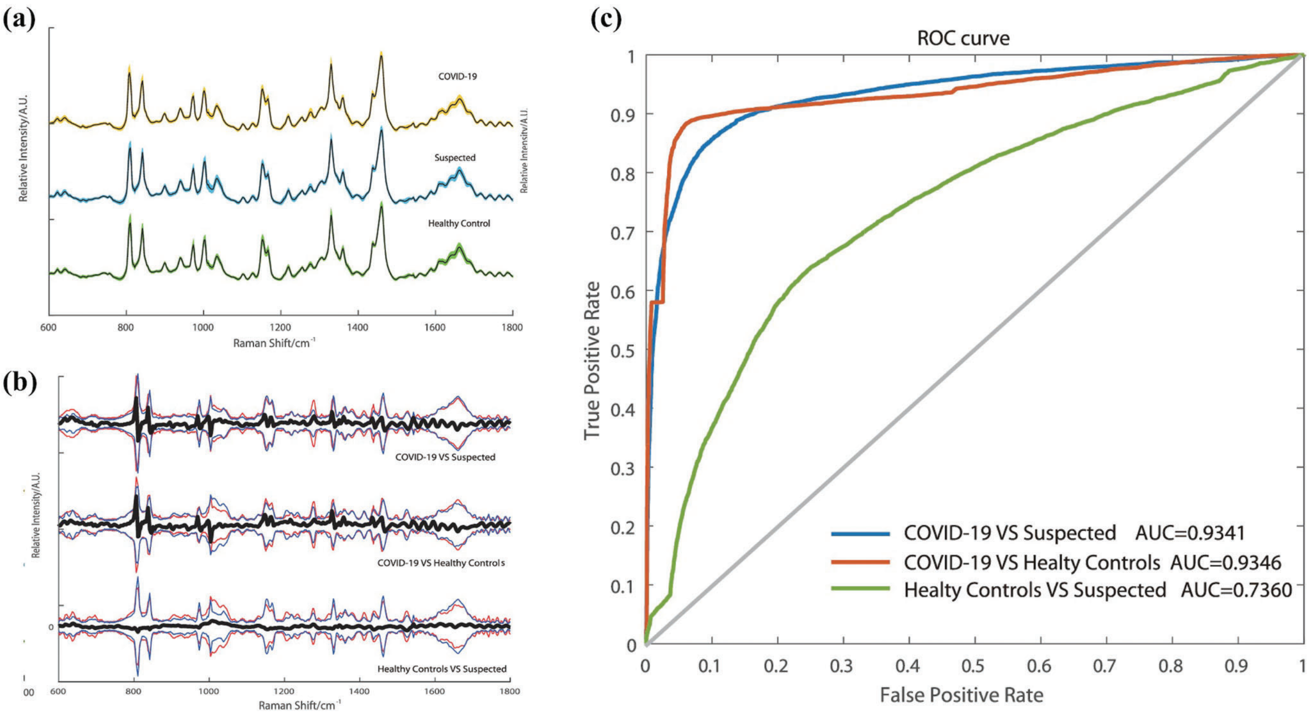


Figure 7. a) Total mean of three Raman spectra of COVID-19, suspected, and healthy groups. b) Raman difference signal between groups (black) and ±2 standard deviations between groups (red and blue). c) The receiver operating characteristic curve of the SVM diagnostic algorithm for the COVID-19 group versus the suspected group, the COVID-19 group versus the healthy control group, and the suspected group versus the healthy control group. Reproduced with permission.[156] Copyright 2021, John Wiley & Sons.

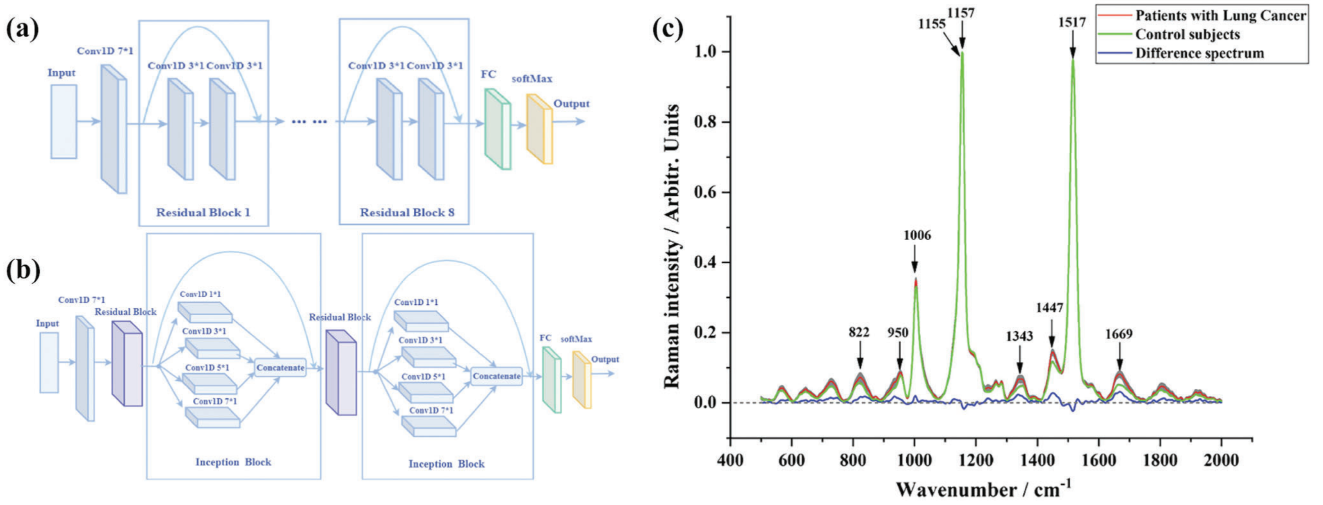


Figure 8. a) ResNet model. b) Improved ResNeXt model. c) Normalized mean spectra and sample standard errors (shaded areas) of control subjects and of lung cancer patients and their difference spectra. Reproduced with permission.[161] Copyright 2022, John Wiley & Sons.

learning algorithm based on ResNet used in the experiment was able to detect the features of exosomes of lung cancer cells and successfully identify patients with early-stage lung cancer.[163]

Early screening is significant for some cancers with long incubation periods, such as ovarian cancer. Zhang et al.[164] proposed a method for early screening of cervical adenocarcinoma and cervical squamous cell carcinoma based on cancerous tissue data collected by Raman spectroscopy and several classification models constructed by ML algorithms. After a comparative study, the airPLS-PLS-KNN algorithm has the highest accuracy rate of 96.3%.

Blood-based Raman spectroscopy can realize simple, minimally invasive, and efficient cancer detection, making it promising for detecting ovarian cancer.[165] For example, in a study on ovarian cancer,[166] a total of 174 blood samples were collected from 95 patients with initial suspicion of ovarian cancer, of which 62 patients were diagnosed with ovarian cancer and 33 with ovarian cysts after further diagnosis. Based on these two types of samples with a control group consisting of 79 normal blood samples, a triple classification function of cancer, cysts, and normal cases was achieved.[166] Even though the classification results could be improved, it was still able to demonstrate the diagnostic potential of plasma-based Raman spectroscopy with ML for ovarian cancer. In addition, Wang et al.[167] proposed a method for identifying blood species by applying RNN to Raman spectroscopy and achieved discrimination of 20 blood species, including humans and different animals, in which the recognition accuracy of bidirectional RNN with gate recurrent unit (GRU) was 97.7%. The experiments demonstrate the potential of this method in practical application scenarios such as customs inspection and medical or forensic identification.

Gastric cancer (GC) is a widespread malignant tumor of the digestive tract. Li et al.[168] measured the serum Raman spectra of GC patients and healthy controls and used 1D CNN, RF, SVM, and KNN to diagnose and predict gastric cancer. The experimental results indicated that the RF algorithm has the best performance. A study showed granulocyte colony stimulating factor (G-CSF) was directly related to gastric cancer metastasis.[169] Then Zhang et al.[170] investigated the long-term treatment of cancer cells by G-CSF. PCA was used to determine the Raman spectrum band with the most significant difference between normal control cells and cancer cells treated with G-CSF. To evaluate the progress of G-CSF treated cells, the concept of aggression score was derived using a posteriori probability based on the linear discriminant function.[170] This work may lead to identifying new targets for cancer treatment.

Early diagnosis of renal cell carcinoma can significantly benefit patients. He et al.[171] trained more than 3000 Raman spectra obtained from the normal kidney, fat, and tumor tissue from 77 patients to build an SVM model based on Raman spectroscopy. The trained SVM model can achieve in vitro identification of kidney tumor tissue with an accuracy of 92.89%. This study demonstrates the potential of Raman spectroscopy-based ML models in the rapid clinical diagnosis of kidney cancer.

4.2. ML-assisted Raman Spectroscopy of Pathogen

In clinical diagnosis, infections due to bacterial pathogens are widespread, and severe acute bacterial infections can even be fatal. Therefore, accurate and rapid identification of bacterial infections is critical in medicine. Ye et al.[172] used a Raman dataset collected from various viruses to train CNN models capable of being highly accurate and sensitive to identify viruses imaged by Raman spectroscopy. The experiments accurately classified different types of human and avian viruses. Moawad et al.[173] applied the SVM model to identify Burkholderia mallei and related species and demonstrated the potential of ML-based Raman spectroscopy as a bacterial bio-diagnostic tool. Kukula et al.[174] used a four-layer CNN architecture to achieve the classification of 30 classes of bacteria and achieved an accuracy of approximately 86%. Yu et al.[175] demonstrated the accurate identification of eight strains isolated from the marine organism Urechis unicinctus in the marine environment by Raman spectroscopy with the RNN model employing long and short-term memory (LSTM) (**Figure 9**a, b). The authors evaluated the performance

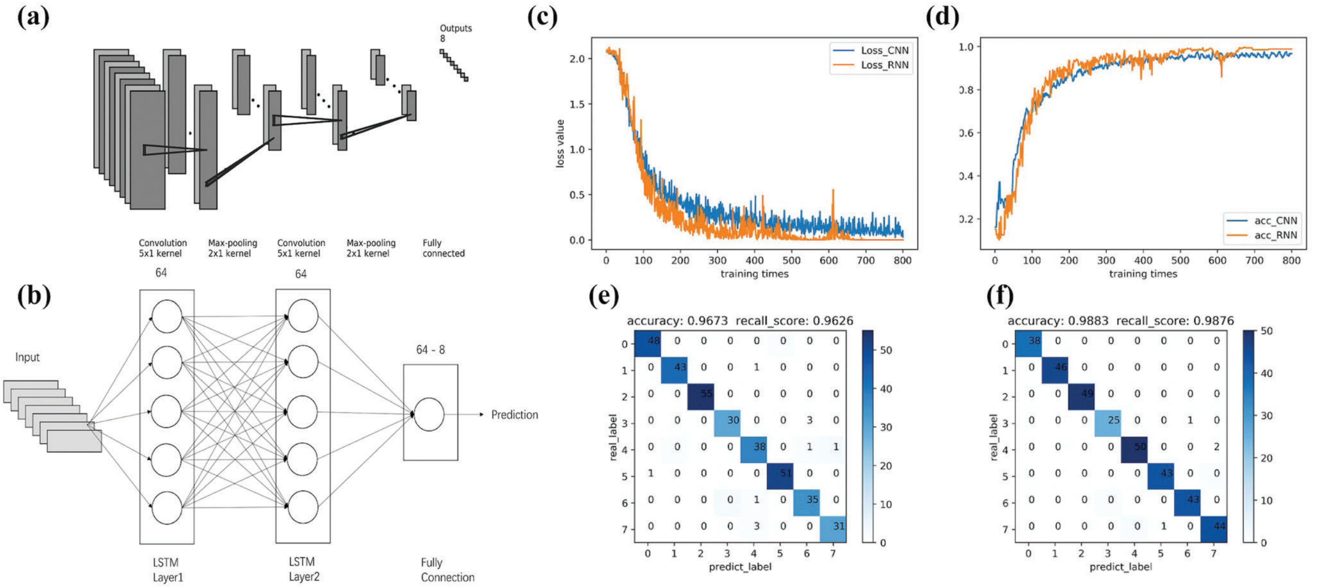


Figure 9. a)A CNN that consists of three layers was designed for detection of marine pathogens. The first two layers are used to extract the features of the Raman data, and each layer is a combination of a convolution layer and a pooling layer. The last layer is a fully connected neural network layer for classification. b) RNN model uses the LSTM method and consists of three layers with 64 neurons in each layer, yielding a final output of eight dimensions through the fully connected layer. c) Change in the loss value was recorded when the CNN and RNN models were trained. Although the two models dealt with the same Raman data, the training efficiency of the RNN model was higher, and the decrease in the loss value was lower. d) Accuracy results of the CNN and RNN models for the test data set. e, f) Final classification results of the two models. The RNN model has a significantly improved accuracy rate compared to the CNN model. Reproduced with permissio.[175] Copyright 2021, ACS Publications.

of the proposed RNN approach with LSTM versus CNN.[175] The training results (Figure 9c–f) showed that the RNN model in this study has significantly higher accuracy than the CNN model, with an accuracy higher than 94%.

Yan et al.[176] collected single-cell Raman spectra of food-borne pathogens from seven common genera (i.e., *E. coli, Listeria monocytogenes*, etc.) The data were processed by Kernel PCA combined with a decision tree (KPCA-DT). KPCA extracts the nonlinear features of the raw data and evaluates and discriminates individual bacterial cells at the serotype level by a DT algorithm with classification accuracy in the range of 87.1%–95.8%.[176] Zahn et al.[177] used ANN and SVM models to distinguish the Raman spectra of 11 E. coli strains, respectively. They experimentally demonstrated that both methods could potentially analyze complex spectral datasets for biomedical applications, with SVM showing better results on small datasets and ANN with relatively complex structures showing better results on large datasets.[177] Nakar et al.[178] first used Raman spectroscopy to successfully differentiate between resistant and sensitive strains of E. coli without exposure to antibiotics. Previous experiments revealed that resistant strains had a higher ratio of nucleic acid to protein.[179,180] Nakar et al. subtracted the mean spectrum of sensitive strains from the mean spectrum of the resistant strains to obtain the different spectra, which were then classified using the ML algorithm.[178] In conclusion, the classification method employing the Raman spectrum combined with ML methods can successfully match the corresponding strain types according to the Raman spectrum of a single cell. It can assist in preventing the rise of antimicrobial resistance to improve medical diagnosis.

In addition, the weak signals of Raman spectra under natural conditions make data set construction and accurate identification challenging.[181,182] Xu et al.[183] proposed to use signal-to-noise ratio (SNR) as an evaluation metric for Raman by analyzing single-cell Raman spectra (SCRS) with short acquisition times (and low SNR) to obtain more spectral data. Then 11141 Raman spectra from nine strains were used for bacterial identification using two ML methods, one is a simple filter, and another is a neural network-based denoising autoencoder. Similarly, to address the problem of weak Raman signals, Ho et al.[184] generated a broad range of datasets of bacterial Raman spectra and applied the deep learning method CNN to accurately identify 30 common bacterial pathogens. The experimental flow is shown in **Figure 10**. Experiments were conducted to validate the results from clinical isolates from 50 patients, only using ten bacterial spectra per patient isolate, with a recognition accuracy of 99.7%.

5. Application of ML & Raman Spectroscopy in Food Science

As people become more health conscious, they realize that the problem of food adulteration is seriously damaging the interests of consumers.[185,186] In pursuit of higher commercial value, some producers and sellers use cheaper substitutes for the original ingredients they should use to make illegal profits. Therefore, efficient and accurate technology is needed to provide precise information about food ingredients and detect food adulteration. The fast, reagent-free, and non-destructive characteristics of Raman spectroscopy make it a standard method in food identification.[15,187] In addition, considering the large amount of data that spectroscopy usually generates, Raman spectroscopy is often combined with ML, which enables more advanced data processing.

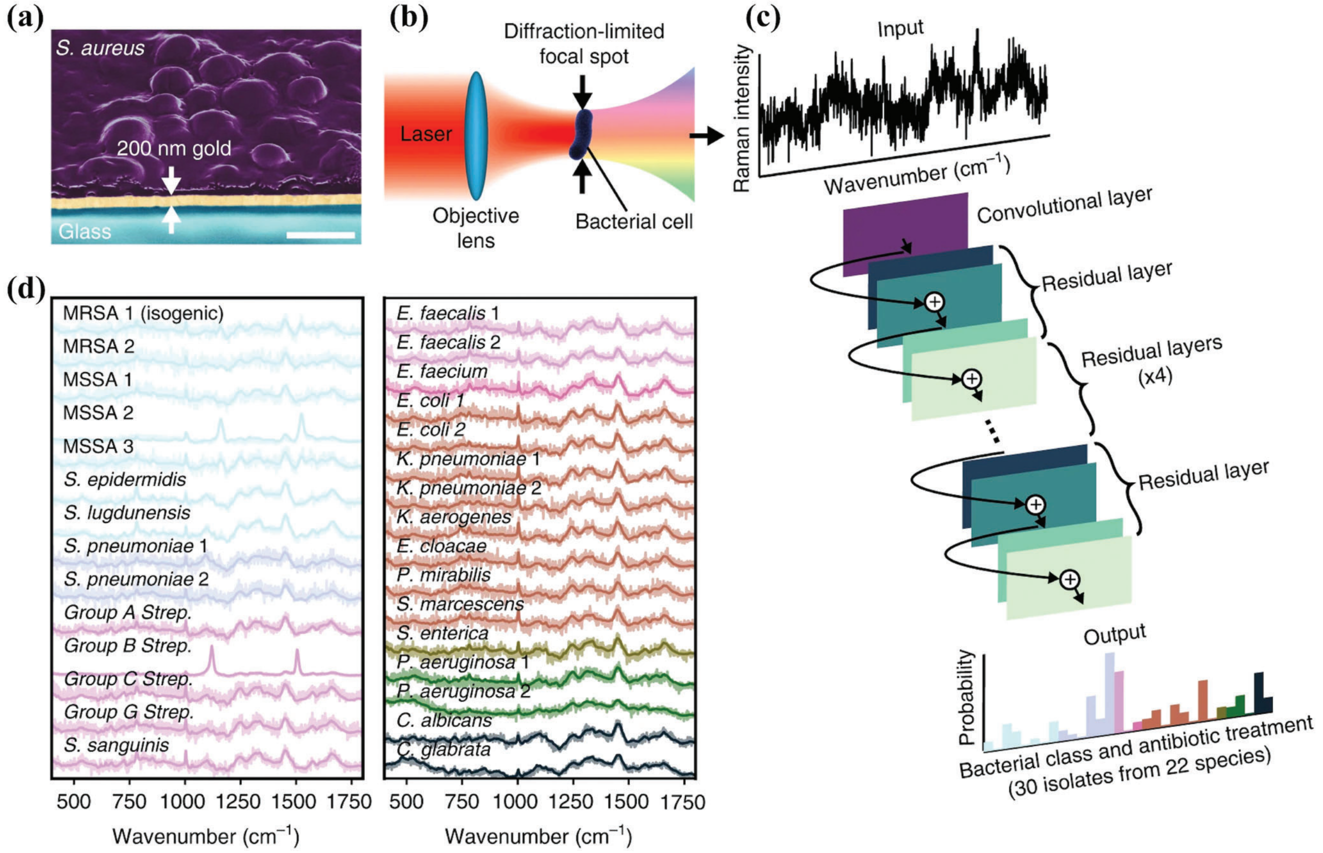


Figure 10. a) To build a training dataset of Raman spectra, deposit bacterial cells onto gold-coated silica substrates and collect spectra from 2000 bacteria over monolayer regions for each strain. An SEM cross section of the sample is shown (gold coated to allow for visualization of bacteria under electron beam illumination). b) Focusing the excitation laser source to a diffraction-limited spot size, Raman signal from single cells can be acquired. c) Using a 1D residual network with 25 total convolutional layers, low-signal Raman spectra are classified as one of 30 isolates, which are then grouped by empiric antibiotic treatment. d) Averages of 2000 spectra from 30 isolates are shown in bold and overlaid on representative examples of noisy single spectra for each isolate. Spectra are color-grouped according to antibiotic treatment. Adapted with permission.[184] Copyright 2019, Springer Nature.

Table 5 shows the application of Raman spectroscopy and ML to food safety.

Chen et al.[188] developed an ML model based on the MSC-GA-KM-Cubist method to detect whether the Atlantic salmon was adulterated by analyzing the component functional groups corresponding to the Raman peaks from the fat of two fish types. While the Raman spectra of fat observed when Atlantic salmon was adulterated with different proportions of rainbow trout were similar, there is a linear relationship between Atlantic salmon's adulteration rate and the Raman spectrum's intensity.

As early as 2005, Ellis et al.[189] first attempted vibrational spectroscopy techniques, including Raman spectroscopy analysis of the molecular structure and intergroup relationships of muscle products combined with PCA-DFA methods to study muscle foods closely related to poultry species such as chicken and turkey. Robert et al.[190] used Raman spectroscopy combined with three classification techniques, PCA, partial least square discriminant analysis (PLS-DA), and SVM, to identify different species of red meat with a short analysis time of 15 s and high accuracy. Therefore, ML combined with Raman spectroscopy can be used as an alternative technique in complete meat identification. In other words, since we can identify meat categories and poultry products that are similar in chemical composition, we can also apply this to identify counterfeit meat.

In another study, a method based on the rapid processing of Raman spectra using ML algorithms to certify edible oils was successfully developed.[191] A total of 36 samples obtained from seven categories of oils through different extraction procedures and mixed in different modes were selected for the study.[191] A classification learner constructed by ML algorithms was used to identify the most relevant oil classification model, realize adulteration detection, and preliminarily estimate its magnitude. Another investigation was conducted on adulterating extra virgin olive oil (EVOO) with inexpensive edible oil.[192] Raman spectra were obtained from binary, ternary, and quaternary mixtures of oil samples using non-negative least squares to obtain the relative concentrations of EVOO and other cheap oils to determine the purity of EVOO. This research demonstrates the potential of Raman spectroscopy in determining the purity of edible oils. Zhao et al.[193] discovered that Raman spectroscopy based on fatty acid composition effectively classified edible oils. ML algorithms significantly improved the accuracy of Raman spectroscopy analysis, and experimental results showed that PCA with RF model is the best algorithm for Raman

Table 5. Latest developments in combining ML methods with Raman spectroscopy for food science.

Objectives	Methods	Results	References
Identification of rainbow trout adulteration in Atlantic salmon	MSC, GA, KM, Cubist	The determination coefficient (R2) and root mean square error of prediction sets (RMSEP) in the test sets are 0.87 and 10.93, respectively.	Chen et al.[188] (2019)
Identification of muscle foods	PCA, DFA	It is possible to discriminate qualitatively between all four muscle groups and find relevant wavenumbers.	Ellis et al.[189] (2005)
Identification of intact beef, venison and lamb	PCA, PLS-DA, SVM	The accuracy for predicting unknown samples exceeded 80% (PLS-DA) and 92% (SVM). SVM and PLS-DA models perform best in predicting venison samples with sensitivities of 100%.	Robert et al.[190] (2021)
Evaluation of edible oils	ML in Matlab (KNN, PCA...)	The model that works well (accuracy 88.9%) is a subspace KNN when the PCA is disabled.	Berghian-Grosan and Magdas [191] (2020)
Analysis of the authenticity and concentration of extra virgin olive oil	PLS	The purity of the spiked extra virgin olive oil can be determined, although it is mixed with one or cheaper oils.	Duraipandian et al.[192] (2019)
Detection of edible oils type and adulteration	PCA, CNN, RF...	Several ML algorithms are time efficient and 100% accurate in classifying edible oils based on an acid composition by gas chromatography.	Zhao et al.[193] (2022)
Detection of adulterated honey	PLS, LDA	The overall accuracy rate in detecting authentic and adulterated honey is 96.54%.	Oroian et al.[194] (2018)
Detection of adulterated Suichang native honey	SVM, PNN, CNN	The overall accuracy of CNN, PNN, and SVM models are 100%, 100%, and 99.75%, respectively.	Hu et al.[195] (2022)
Detection of fruit distillates	DT, DA, SVM, KNN, Ensemble classifiers	Trademark fingerprint identification obtained a model accuracy of 95.5% (only one sample was misclassified). For the geographical distinction of fruit wines, the accuracy is 90.9%.	Grosan et al.[196] (2020)

(MSC: Multiple Scattering Correction, KM: K-means Clustering, DA: Discriminant Analysis.)

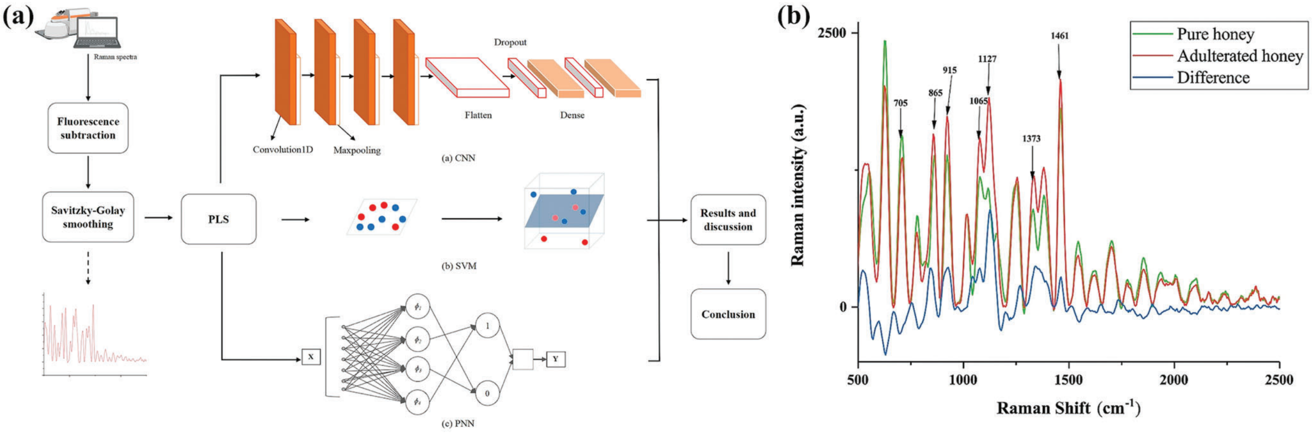


Figure 11. a) Experimental process for classifying pure honey. b) Mean spectra of samples. Reproduced with permission.[195] Copyright 2022, Springer Nature.

spectroscopy-based edible oil classification. This study illustrates the potential of ML-assisted Raman spectroscopy for the rapid identification and detection of food products.

ML combined with Raman spectroscopy has been used to detect and classify honey. In 2018, Oroian et al.[194] used partial least squares linear discriminant analysis (PLS-LDA) to detect honey adulterated with fructose, glucose, transformed sugar, hydrolyzed inulin syrup, and wort, and achieved an overall accuracy of 96.54% in detecting authentic honey from adulterated honey. A recent study applies SVM, PNN, and CNN to Raman spectroscopy to classify pure honey and adulterated honey samples.[195] The structure of this work is depicted in **Figure 11**a, and the mean spectra of honey samples are shown in Figure 11b. The overall accuracy of the CNN, PNN, and SVM models is 100%, 100%, and 99.75%, respectively, when sensitivity, specificity, and accuracy are considered. The results indicated that Raman spectroscopy combined with ML algorithms can accurately detect low-concentration adulterated honey. Falsely declaring the origin and source of a product is also a form of food adulteration. Grosan et al.[196] combined five predictive modeling approaches with Raman spectra of fruit distillates in a specific spectral range to distinguish the trademark, geographical and botanical origin.

Table 6. Latest developments in combining ML methods with Raman spectroscopy for other fields.

Objectives	Methods	Results	References
Classification of disposable masks	PCA, SVM, Bayesian, BPNN	The accuracy of Bayesian discriminant model has reached 100.0%, which can be used as the best model for mask classification.	Liu et al.[197] (2021)
Estimation of holocellulose content of poplar clones	SVR, DT, RF, GBM…	All models successfully predicted the whole cellulose content, with the advanced GBM algorithm outperforming all models during training and testing.	Gao et al.[199] (2022)
Identification of handmade paper	PCA, LS, SVM, KNN, RF	PCA-LR had the highest classification and prediction accuracy (R2 = 1).	Yan et al.[200] (2022)
Identification of white mineral pigments	Deep CNN	Achieved an accuracy of up to 98.7%.	Qi et al.[201] (2022)

(GBM: Gradient Boosting Machine, SVR: Support Vector Regression.)

6. Application of ML & Raman in Other Fields

As a powerful and fast analysis method, ML assisted Raman spectroscopy is also frequently applied to study objects commonly used in life, such as masks, artifacts, paints, etc.[197–201] (**Table 6**). Disposable masks rose significantly under the lasting effects of Covid-19, which increased the probability of their presence at crime scenes. Liu et al.[197] analyzed the Raman spectra of mask samples from 37 different cities and manufacturers. They proposed a classification and recognition method for disposable masks based on feature extraction and multi-model optimization. The authors divided the mask categories by PCA, compared feature peaks in Raman spectrum, and then constructed disposable mask classification and recognition models based on SVM, Bayesian discriminant analysis, and backward propagation neural network (BPNN). The training and testing accuracy of the Bayesian discriminant model has reached 100.0%, making it the optimal model for mask classification and recognition, which may play a significant role in identifying material evidence in courts.

In addition, Raman spectroscopy combined with ML has the potential to be applied to tree breeding programs. Gao et al. used Raman spectroscopy ML to predict the lignin content in poplar trees.[198] In another new study,[199] they recently presented nine ML models built based on features extracted from Raman spectra that successfully predicted the holocellulose content of poplar trees. Yan et al.[200] focused on identifying handmade paper. They measured 18 kinds of handmade paper samples using Raman spectroscopy. They constructed five ML models, PCA-LS, PLS-LS, SVM-LS, KNN, and RF, to assess the impact and effect of data processing and to classify and predict the samples. Among the different models, PCA-LR algorithm has the highest classification and prediction accuracy. Qi et al.[201] proposed a deep CNN-based method to automatically identify the Raman spectra of white mineral pigments, achieving accuracy up to 98.7%. Experimental results showed that the method can analyze the composition of white mineral pigments efficiently and non-destructively, and the proposed method exhibits superior performance compared to traditional learning algorithms such as PCA-DNN and SVM. The non-destructive nature of these works makes the Raman spectroscopy with ML approach valuable for artifact research, archaeology, etc., and can stimulate further research on related cultural relics.

7. Conclusion; Challenges and Perspectives for ML-Assisted Analysis of Raman Spectral Data

The adoption of ML has effectively contributed to a wide range of research involving processing of Raman spectroscopy data and practical applications in many fields, such as material classification, biomedicine, food science, and other areas. In this review, we first introduced the concepts of machine learning by comparing ML-methods with traditional chemometrics and statistical techniques. Then, we introduced the application of different methods in ML in analyzing Raman spectroscopy data in different areas, summarized advantages and disadvantages, results achieved and experiments or data or techniques that are still lacking. Overall, our various examples showed that it is promising to solve many research problems more effectively in various fields using ML combined with Raman spectroscopy than using the traditional chemometrics and statistical techniques only. Furthermore, deep learning related methods further improved the ability for computers to analyze larger and complex data including Raman spectroscopy data as well as to finish more difficult tasks.

During the process of using various traditional chemometrics and statistical techniques and ML-methods for the analysis of Raman spectroscopy data, many different challenges could be tackled, which could not be comprehensively covered here. Hence, we only selected two specific examples including data enhancement for Raman spectra and anomaly detection for Raman spectra because they are not only widely studied and important problems in ML, but also especially related to data analysis of Raman spectroscopy.

7.1. Data Enhancement for Raman Spectra

The size of training data is a significant issue for deep learning and ML.[75,202] It directly impacts the performance of deep learning. It is possible to overfit when deep learning is used on a small dataset.[203] The biggest challenge for multi-classification studies of Raman spectra utilizing deep learning techniques, is that the amount of Raman data available to researchers is often minimal in practice. Therefore, solving the problem of insufficient Raman spectral data is urgent.

As a solution to the problem, a hierarchical pipeline of data enhancement steps reinforced with the GAN method was presented by Frischia et al.[204] The data augmentation pipeline began with a natural individual Raman spectrum. It progressed to include several signal processing procedures or algorithms to artificially increase the number of spectra in the original dataset, such as by adding white Gaussian noise to the original signal, applying baseline removal algorithms, noise reduction filtering, clustering, shifting, and merging the data.[204] The dataset is then enhanced further using the GAN framework. A discriminator is trained to identify fake data (artificial Raman spectra) from actual samples of the original training dataset, while the generator is trained to generate Raman spectra.[204] When using deep learning models (e.g., CNN), the GAN-enhanced dataset exhibits a distinctly positive behavior with significantly higher accuracy.

7.2. Anomaly Detection for Raman Spectra

Classification problems have typically attempted to identify two or more classes. However, suppose only one type of data is available. In that case, the goal is generally to test the new data and compare it to the original training data, which is often used for anomaly detection.[205] Anomalistic samples will inevitably appear in Raman spectroscopy measurement, as this technique is susceptible to environmental and other influences. From an ML perspective, anomaly detection will significantly affect the predictive model performance.[206]

Hofer et al.[207] proposed a one-class anomaly detector based on Raman spectral autoencoder. Using a biological application as an example, Hofer et al. measured and trained a single-class model for an average pre-transfected cell class and made it learn regular class features by minimizing reconstruction error for a given loss function.[207] When using the learned encoding to reconstruct spectra, the samples are considered abnormal if they exceed a standard deviation threshold.[207] The results demonstrated that anomaly detection can aid in the reconstruction of Raman spectra and lead to promising classification results.

Although there is no doubt about the great potential of using ML in assisting the analysis of Raman spectroscopy data, there are still quite many research directions that need to be further studied and developed in this area. First, the limited data size of Raman spectroscopy should be solved urgently, only data enhancement is not enough, the main reason for limited data size is that Raman spectroscopy data produced in different labs are obtained under different conditions and those data in general not useful for another lab, if a public database could be established to store all the Raman spectroscopy data from all the labs in the world and a standard normalization or preprocessing method or an enhanced ML method capable of doing this task is available one day, it is expected that data size would no longer be a problem. Second, currently, it takes quite a long time to obtain an image of Raman spectroscopy, with the further advancement for the efficiency of Raman spectrometer, increased number of Raman spectroscopy images is expected. Third, most Raman spectrometers are still quite big in size, if their size could be further decreased, portable Raman spectrometers may be possible in the near future, by then, together with more advanced ML techniques, it is expected that they will become more useful in medicine, diagnose ,and sample analysis. Fourth, more research is needed to use AI-assisted data analysis of Raman spectroscopy on soil, plant, rock, and food samples, results of which would provide a lot of useful information for sample identification, classification and prediction of their origins, distribution, characteristics and properties, breakthrough results are expected in these previously less investigated areas. One intriguing direction can be to combine ML and Raman spectroscopy to guide more efficient screening and discovery of materials with better functional properties (such as those dependent on phonons, for example, thermal conductivity, thermoelectrics, or even superconductivity). Finally, although several standard ML methods have been combined with Raman spectroscopy and applied in several fields, including deep learning algorithms, it is still of significant interest to develop new and simple ML methods based on different research requirements for devising new classification and recognition modes based on Raman spectra. The emerging field of combining ML with Raman spectroscopy can contribute, along with other established techniques, to expedite material characterization, showing advantages in terms of manpower requirement, equipment and time cost, and accuracy, in a myriad of sectors.

Acknowledgements

The author acknowledge partial support of this work by Macau Science and Technology Development Fund (FDCT Grants 0106/2020/A3 and 0031/2021/ITP) and Macau University of Science and Technology (FRG-21-034-MISE). The author also acknowledge earlier support and inspiration from Purdue University Discovery Park Big Idea Challenge program. The author thank all reviewers' valuable comments to improve this manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Y.Q., D. H. equally contributed to this work. Y.Q. performed conceptualization, funding acquisition, supervision, writing-Original draft, writing-reviewing, and Editing Y.C. performed conceptualization, funding acquisition, supervision, writing-reviewing, and editing. D.H. wrote-original draft along with reviewing and editing. Y.J., Z.W., M.Z., E.X.C., Y.L., M.A.S., K.Z. wrote along with reviewing and editing.

Keywords

artificial intelligence, deep learning, machine learning, material science, Raman spectroscopy

Received: December 25, 2022
Revised: March 11, 2023
Published online:

[1] J. M. Chalmers, P. R. Griffiths, *Handbook of Vibrational Spectroscopy*, Wiley, Hoboken, NJ, USA **2002**.
[2] N. B. Colthup, H. D. Lawrence, E. W. Stephen, *Introduction to Infrared and Raman Spectroscopy*, Elsevier, San Diego, CA, USA **2012**.
[3] M. Fan, G. F. Andrade, A. G. Brolo, *Anal. Chim. Acta* **2011**, *693*, 7.

[4] A. Kudelski, *Talanta* **2008**, *76*, 1.
[5] L. B. Ayres, F. J. Gomez, J. R. Linton, M. F. Silva, C. D. Garcia, *Anal. Chim. Acta* **2021**, *1161*, 338403.
[6] D. A. Long, *Vibrational Spectroscopy of Molecules on Surfaces. Methods of Surface Characterization*, Vol. 13 (Eds: J. T. Yates, T. E. Madey), Springer, Boston, MA, USA **1977**.
[7] W. Ahmed, M. J. Jackson, J. M. Jackson, *Emerging Nanotechnologies for Manufacturing*, William Andrew, Burlington, MA, USA **2009**.
[8] R. S. Das, Y. K. Agrawal, *Vib. Spectrosc.* **2011**, *57*, 163.
[9] N. P. Pieczonka, R. F. Aroca, *Chem. Soc. Rev.* **2008**, *37*, 946.
[10] E. Garcia-Rico, R. A. Alvarez-Puebla, L. Guerrini, *Chem. Soc. Rev.* **2018**, *47*, 4909.
[11] K. Nakamoto, *Infrared and Raman Spectra of Inorganic and Coordination Compounds, Part B: Applications in Coordination, Organometallic, and Bioinorganic Chemistry*, John Wiley & Sons, Hoboken, New Jersey, USA **2009**.
[12] S. Y. Ding, J. Yi, J. F. Li, B. Ren, D. Y. Wu, R. Panneerselvam, Z. Q. Tian, *Nat. Rev. Mater.* **2016**, *1*, 1.
[13] Z. Movasaghi, S. Rehman, I. U. Rehman, *Appl. Spectrosc. Rev.* **2007**, *42*, 493.
[14] T. Vankeirsbilck, A. Vercauteren, W. Baeyens, G. Van der Weken, F. Verpoort, G. Vergote, J. P. Remon, *Trends Analyt. Chem.* **2002**, *21*, 869.
[15] E. C. Y. Li-Chan, *Trends Food Sci. Technol.* **1996**, *7*, 361.
[16] A. P. Craig, A. S. Franca, J. Irudayaraj, *Annu. Rev. Food Sci. Technol.* **2013**, *4*, 369.
[17] M. J. Pelletier, *Appl. Spectrosc.* **2003**, *57*, 20A.
[18] R. Gautam, S. Vanga, F. Ariese, S. Umapathy, *EPJ Tech. Instrum.* **2015**, *2*, 8.
[19] H. J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N. J. Fullwood, B. Gardner, P. L. Martin-Hirsch, M. J. Walsh, *Nat. Protoc.* **2016**, *11*, 664.
[20] E. Mjolsness, D. DeCoste, *Science* **2001**, *293*, 2051.
[21] S. Zhong, K. Zhang, M. Bagheri, J. G. Burken, A. Gu, B. Li, X. Ma, B. L. Marrone, Z. J. Ren, J. Schrier, W. Shi, *Environ. Sci. Technol.* **2021**, *55*, 12741.
[22] Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, X. Liu, Y. Wu, F. Dong, C. W. Qiu, J. Qiu, *The Innovation* **2021**, *2*, 100179.
[23] T. M. Mitchell, *Machine Learning*, McGraw-hill, New York, USA **1997**.
[24] I. H. Sarker, *SN Comput. Sci.* **2021**, *2*, 160.
[25] F. Lussier, V. Thibault, B. Charron, G. Q. Wallace, J. F. Masson, *Trends Analyt. Chem.* **2020**, *124*, 115796.
[26] R. Houhou, T. Bocklitz, *Anal. Sci. Adv.* **2021**, *2*, 128.
[27] W. F. D. C. Rocha, C. B. D. Prado, N. Blonder, *Molecules* **2020**, *25*, 3025.
[28] F. Cui, Y. Yue, Y. Zhang, Z. Zhang, H. S. Zhou, *ACS Sens.* **2020**, *5*, 3346.
[29] R. Helin, U. G. Indahl, O. Tomic, K. H. Liland, *J. Chemom.* **2022**, *36*, e3374.
[30] L. Pan, P. Zhang, C. Daengngam, S. Peng, M. Chongcheawchamnan, *J. Raman Spectrosc.* **2022**, *53*, 6.
[31] R. Han, R. Ketkaew, S. Luber, *J. Phys. Chem. A* **2022**, *126*, 801.
[32] M. H. Mozaffari, L. L. Tay, (Preprint) arXiv:2006.10575, submitted: Jun **2020**.
[33] R. Luo, J. Popp, T. Bocklitz, *Analytica* **2022**, *3*, 287.
[34] G. R. Schleder, A. C. Padilha, C. M. Acosta, M. Costa, A. Fazzio, *J. Phys. Mater.* **2019**, *2*, 032001.
[35] D. M. Dimiduk, E. A. Holm, S. R. Niezgoda, *Integr. Mater. Manuf. Innov.* **2018**, *7*, 157.
[36] Y. K. Dwivedi, L. Hughes, E. Ismagilova, G. Aarts, C. Coombs, T. Crick, Y. Duan, R. Dwivedi, J. Edwards, A. Eirug, V. Galanos, *Int. J. Inf. Manag.* **2021**, *57*, 101994.
[37] S. Guo, J. Popp, T. Bocklitz, *Nat. Protoc.* **2021**, *16*, 5426.
[38] M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys, C. Hughes, *Nat. Protoc.* **2014**, *9*, 1771.
[39] M. A. Nielsen, *Neural Networks and Deep Learning*, Determination press, San Francisco, CA, USA **2015**.
[40] J. Schmidhuber, *Neural Netw.* **2015**, *61*, 85.
[41] J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, S. J. Gibson, *Analyst* **2017**, *142*, 4067.
[42] Y. Lai, *J. Phys. Conf. Ser.* **2019**, *1314*, 012148.
[43] D. Maulud, A. M. Abdulazeez, *J. Appl. Sci. Technol. Trends.* **2020**, *1*, 140.
[44] X. Yan, X. Su, *Linear Regression Analysis: Theory and Computing*, World Scientific, Singapore **2009**.
[45] P. Geladi, B. R. Kowalski, *Anal. Chim. Acta* **1986**, *185*, 1.
[46] Y. Li, X. Shao, W. Cai, *Talanta* **2007**, *72*, 217.
[47] F. Chauchard, R. Cogdill, S. Roussel, J. M. Roger, V. Bellon-Maurel, *Chemometr. Intell. Lab Syst.* **2004**, *71*, 141.
[48] D. Wu, Y. He, S. Feng, D. W. Sun, *J. Food Eng.* **2008**, *84*, 124.
[49] I. Kononenko, *In Machine Learning—EWSL-91: European Working Session on Learning Porto, Portugal*, Springer, Berlin, Heidelberg **1991**.
[50] K. M. Leung, *Naive bayesian classifier, Polytechnic University Department of Computer Science/Finance and Risk Engineering* **2007**, *2007*, 123.
[51] P. Langley, W. Iba, K. Thompson, *Aaai* **1992**, *90*, 223.
[52] J. Luo, C. M. Vong, P. K. Wong, *IEEE Trans. Neural Netw. Learn. Syst.* **2013**, *25*, 836.
[53] S. Wold, K. Esbensen, P. Geladi, *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37.
[54] H. Abdi, L. J. Williams, *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433.
[55] J. E. Jackson, *A user's guide to principal components*, John Wiley & Sons, New York, USA **2005**.
[56] I. K. Fodor, *A survey of dimension reduction techniques*, Lawrence Livermore National Lab, CA, USA **2002**.
[57] C. J. Burges, *Found. Trends Mach. Learn.* **2010**, *2*, 275.
[58] I. T. Jolliffe, J. Cadima, *Philos. Trans., Math. Phys. Eng. Sci.* **2016**, *374*, 20150202.
[59] J. M. Vargas, S. Nielsen, V. Cárdenas, A. Gonzalez, E. Y. Aymat, E. Almodovar, G. Classe, Y. Colón, E. Sanchez, R. J. Romañach, *Int. J. Pharm.* **2018**, *538*, 167.
[60] A. Tharwat, T. Gaber, A. Ibrahim, A. E. Hassanien, *AI Commun.* **2017**, *30*, 169.
[61] O. C. Hamsici, A. M. Martinez, *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 647.
[62] S. Balakrishnama, A. Ganapathiraju, *IEEE Trans. Signal Inf. Process.* **1998**, *18*, 1.
[63] E. Fix, J. L. Hodges, *Int. Stat. Rev.* **1989**, *57*, 238.
[64] T. N. Phyu, in *Proceedings of the International Multiconference of Engineers and Computer Scientists*, Citeseer, New Jersey, USA **2009**.
[65] A. Mucherino, P. J. Papajorgji, P. M. Pardalos, in *Data Mining in Agriculture* (Eds: A. Mucherino, P. J. Papajorgji, P. M. Pardalos)., Springer, New York, USA **2009**.
[66] C. Cortes, V. Vapnik, *Int. J. Mach. Learn. Cybern.* **1995**, *20*, 273.
[67] W. S. Noble, *Nat. Biotechnol.* **2006**, *24*, 1565.
[68] P. H. Chen, C. J. Lin, B. Schölkopf, *Appl. Stoch. Models Bus. Ind.* **2005**, *21*, 111.
[69] J. R. Quinlan, *Int. J. Mach. Learn. Cybern.* **1986**, *1*, 81.
[70] J. R. Quinlan, *IEEE Trans. Syst. Man Cybern* **1990**, *20*, 339.
[71] S. R. Safavian, D. Landgrebe, *IEEE Trans. Syst. Man Cybern.* **1991**, *21*, 660.
[72] W. Tong, H. Hong, H. Fang, Q. Xie, R. Perkins, *J. Chem. Inf. Model.* **2003**, *43*, 525.

[73] J. Ali, R. Khan, N. Ahmad, I. Maqsood, *Int. J. Comput. Sci. Appl.* **2012**, *9*, 272.
[74] A. Bashar, *Artif. Intell.* **2019**, *1*, 73.
[75] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[76] D. Floreano, C. Mattiussi, *Bio-inspired Artificial Intelligence: Theories, Methods, and Technologies*, MIT press, Cambridge, MA, USA **2008**.
[77] A. K. Jain, J. Mao, K. M. Mohiuddin, *Computer* **1996**, *29*, 31.
[78] A. D. Dongare, R. R. Kharde, A. D. Kachare, *Int. J. Eng. Technol. Innov.* **2012**, *2*, 189.
[79] D. F. Specht, *Neural Netw.* **1990**, *3*, 109.
[80] J. Bouvrie, *Notes on convolutional neural networks*, Citeseer, New Jersey, USA **2006**.
[81] K. O'Shea, R. Nash, (Preprint) arXiv:1511.08458, v2, submitted: Dec **2015**.
[82] L. R. Medsker, L. C. Jain, *Des. Appl.* **2001**, *5*, 64.
[83] A. Shrestha, A. Mahmood, *IEEE Access.* **2019**, *7*, 53040.
[84] P. Wang, E. Fan, P. Wang, *Pattern Recognit. Lett.* **2021**, *141*, 61.
[85] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville, Y. Bengio, *Commun. ACM.* **2020**, *63*, 139.
[86] D. R. Neuville, D. de Ligny, G. S. Henderson, *Rev. Mineral. Geochem.* **2014**, *78*, 509.
[87] J. F. Li, Y. F. Huang, Y. Ding, Z. L. Yang, S. B. Li, X. S. Zhou, F. R. Fan, W. Zhang, Z. Y. Zhou, D. Y. Wu, B. Ren, Z. L. Wang, Z. Q. Tian, *Nature* **2010**, *464*, 392.
[88] H. Vašková, *Int. J. Math. Model. Methods Appl. Sci.* **2011**, *5*, 1205.
[89] D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. A. Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, A. G. A, *Nat. Rev. Mater.* **2018**, *3*, 5.
[90] W. S. Leong, G. Arrabito, G. Prestopino, *Crystals* **2020**, *10*, 308.
[91] S. Ramakrishna, T. Y. Zhang, W. C. Lu, Q. Qian, J. S. C. Low, J. H. R. Yune, D. Z. L. Tan, S. Bressan, S. Sanvito, S. R. Kalidindi, *J. Intell. Manuf.* **2019**, *30*, 2307.
[92] B. Ryu, L. Wang, H. Pu, M. K. Chan, J. Chen, *Chem. Soc. Rev.* **2022**, *51*, 1899.
[93] S. Boonsit, P. Kalasuwan, P. van Dommelen, C. Daengngam, *J. Phys. Conf. Ser.* **2021**, *1719*, 012081.
[94] L. Pan, P. Pipitsunthonsan, C. Daengngam, S. Channumsin, S. Sreesawet, M. Chongcheawchamnan, *IEEE Sens. J.* **2021**, *21*, 10834.
[95] Y. Xie, Q. B. You, P. Y. Dai, S. Y. Wang, P. Y. Hong, G. K. Liu, J. Yu, X. L. Sun, Y. M. Zeng, *Acta A Mol. Biomol. Spectrosc.* **2019**, *222*, 117086.
[96] A. K. Geim, *Science* **2009**, *324*, 1530.
[97] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, A. A. Firsov, *Science* **2004**, *306*, 666.
[98] K. S. Novoselov, L. Colombo, P. R. Gellert, M. G. Schwab, K. Kim, *Nature* **2012**, *490*, 192.
[99] M. K. Jo, U. Ravaioli, presented at IEEE 13th Nanotechnology Materials and Devices Conf. (NMDC), xx, Portland, USA **2018**.
[100] D. G. Sirico, G. Acampora, P. Maddalena, F. Gesuele, in *CLEO: Science and Innovations (pp. JTh3A-8)*, Optica Publishing Group, Washington, D.C. **2021**.
[101] H. Li, J. Wu, X. Huang, G. Lu, J. Yang, X. Lu, Q. H. Xiong, H. Zhang, *ACS Nano* **2013**, *7*, 10344.
[102] X. Lin, Z. Si, W. Fu, J. Yang, S. Guo, Y. Cao, J. Zhang, X. Wang, P. Liu, K. Jiang, W. Zhao, *Nano Res.* **2018**, *11*, 6316.
[103] J. E. Lee, G. Ahn, J. Shim, Y. S. Lee, S. Ryu, *Nat. Commun.* **2012**, *3*, 1024.
[104] S. Berciaud, S. Ryu, L. E. Brus, T. F. Heinz, *Nano Lett.* **2009**, *9*, 346.
[105] Z. Chen, Y. Khaireddin, A. K. Swan, *Analyst* **2022**, *147*, 1824.
[106] L. R. P. Machado, M. O. S. Silva, J. L. E. Campos, D. L. Silva, *J. Raman Spectrosc.* **2022**, *53*, 863.
[107] N. Sheremetyeva, M. Lamparski, C. Daniels, B. V. Troeye, V. Meunier, *Carbon* **2020**, *169*, 455.
[108] P. Solís-Fernández, H. Ago, *ACS Appl. Nano Mater.* **2022**, *5*, 1356.
[109] P. Avouris, M. Freitag, V. Perebeinos, *Nat. Photon.* **2008**, *2*, 341.
[110] A. Srivastava, O. N. Srivastava, S. Talapatra, R. Vajtai, P. M. Ajayan, *Nat. Mater.* **2004**, *3*, 610.
[111] J. Zhang, M. L. Perrin, L. Barba, J. Overbeck, S. Jung, B. Grassy, A. Agal, R. Muff, R. Brönnimann, M. Haluska, C. Roman, C. Hierold, M. Jaggi, M. Calame, *Microsyst. Nanoeng.* **2022**, *8*, 19.
[112] Y. Mao, N. Dong, L. Wang, X. Chen, H. Wang, Z. Wang, I. M. Kislyakov, J. Wang, *Nanomaterials* **2020**, *10*, 2223.
[113] Y. He, Y. Ju, Q. Wang, *Appl. Surf. Sci.* **2021**, *565*, 150530.
[114] A. Y. Cui, K. Jiang, M. H. Jiang, L. Y. Shang, L. Q. Zhu, Z. G. Hu, G. S. Xu, J. H. Chu, *Phys. Rev. Appl.* **2019**, *12*, 054049.
[115] W. P. Griffith, *Nature* **1969**, *224*, 264.
[116] S. Potgieter-Vermaak, N. Maledi, N. Wagner, J. H. P. Van Heerden, R. Van Grieken, J. H. Potgieter, *J. Raman Spectrosc.* **2011**, *42*, 123.
[117] R. L. Frost, M. L. Weier, P. A. Williams, P. Leverett, J. T. Kloprogge, *J. Raman Spectrosc.* **2007**, *38*, 574.
[118] S. Andò, E. Garzanti, *Geol. Soc. Publ.* **2014**, *386*, 395.
[119] S. Das, M. J. Hendry, *Chem. Geol.* **2011**, *290*, 101.
[120] P. Vandenabeele, H. G. Edwards, L. Moens, *Chem. Rev.* **2007**, *107*, 675.
[121] D. Bersani, P. P. Lottici, *J. Raman Spectrosc.* **2016**, *47*, 499.
[122] C. Carey, T. Boucher, S. Mahadevan, P. Bartholomew, M. D. Dyar, *J. Raman Spectrosc.* **2015**, *46*, 894.
[123] J. F. Díez-Pastor, S. E. Jorge-Villar, Á. Arnaiz-González, C. I. García-Osorio, Y. Díaz-Acha, M. Campeny, J. Bosch, J. C. Melgarejo, *J. Raman Spectrosc.* **2020**, *51*, 1563.
[124] C. W. Wu, R. J. Shi, W. D. Zeng, *Laser Optoelectron. Prog.* **2020**, *57*, 093301.
[125] H. W. Kuhn, *Nav. Res. Logist. Q.* **1955**, *2*, 83.
[126] X. Sang, R. G. Zhou, Y. Li, S. Xiong, *Neural Process. Lett.* **2022**, *54*, 677.
[127] B. Lafuente, R. T. Downs, H. Yang, N. Stone, *Highlights in mineralogical crystallography*, De Gruyter (O), Berlin, Germany **2015**.
[128] F. Vilaplana, S. Karlsson, *Macromol. Mater. Eng.* **2008**, *293*, 274.
[129] W. Musu, A. Tsuchida, H. Kawazumi, N. Oka, *International Conference on Cybernetics and Intelligent System*, IEEE, New York **2019**.
[130] S. Karbalaei, P. Hanachi, T. R. Walker, M. Cole, *Environ. Sci. Pollut. Res.* **2018**, *25*, 36046.
[131] A. B. Silva, A. S. Bastos, C. I. L. Justino, J. P. da Costa, A. C. Duarte, T. A. P. Rocha-Santos, *Anal. Chim. Acta* **2018**, *1017*, 1.
[132] C. F. Araujo, M. M. Nolasco, A. M. P. Ribeiro, P. J. A. Ribeiro-Claro, *Water Res.* **2018**, *142*, 426.
[133] Z. Sobhani, X. Zhang, C. Gibson, R. Naidu, M. Megharaj, C. Fang, *Water Res.* **2020**, *174*, 115658.
[134] D. Schymanski, C. Goldbeck, H. U. Humpf, P. Fürst, *Water Res.* **2018**, *129*, 154.
[135] C. Fang, Y. L. Luo, X. Zhang, H. P. Zhang, A. Nolan, R. Naidu, *Chemosphere* **2022**, *286*, 131736.
[136] D. Shen, G. Wu, H. I. Suk, *Annu. Rev. Biomed. Eng.* **2017**, *19*, 221.
[137] R. Etzioni, N. Urban, S. Ramsey, M. McIntosh, S. Schwartz, B. Reid, J. Radich, G. Anderson, L. Hartwell, *Nat. Rev. Cancer* **2003**, *3*, 243.
[138] S. Laing, L. E. Jamieson, K. Faulds, D. Graham, *Nat. Rev. Chem.* **2017**, *1*, 0060.
[139] P. Crow, N. Stone, C. A. Kendall, J. S. Uff, J. A. M. Farmer, H. Barr, M. P. J. Wright, *Br. J. Cancer* **2003**, *89*, 106.
[140] P. J. Caspers, G. W. Lucassen, R. Wolthuis, H. A. Bruining, G. J. Puppels, *Biopolymers* **1998**, *4*, S31.
[141] L. P. Choo-Smith, H. G. M. Edwards, H. P. Endtz, J. M. Kros, F. Heule, H. Barr, J. S. Robinson Jr, H. A. Bruining, G. J. Puppels, *Biopolymers* **2002**, *67*, 1.
[142] N. M. Ralbovsky, I. K. Lednev, *Chem. Soc. Rev.* **2020**, *49*, 7428.
[143] Y. Fan, C. Chen, X. Xie, B. Yang, W. Wu, F. Yue, X. Lv, C. Chen, *Lasers Med. Sci.* **2022**, *37*, 417.

[144] T. Sciortino, R. Secoli, E. d'Amico, S. Moccia, M. C. Nibali, L. Gay, M. Rossi, N. Pecco, A. Castellano, E. D. Momi, B. Fernandes, M. Riva, L. Bello, *Cancers* **2021**, *13*, 4196.

[145] M. Riva, T. Sciortino, R. Secoli, E. D'Amico, S. Moccia, B. Fernandes, M. C. Nibali, L. Gay, M. Rossi, E. D. Momi, L. Bello, *Cancers* **2021**, *13*, 1073.

[146] A. Ward, S. Tardiff, C. Dye, H. M. Arrighi, *Dement. Geriatr. Cogn. Disord. Extra.* **2013**, *3*, 320.

[147] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, L. Beckett, *Alzheimers Dement.* **2005**, *1*, 55.

[148] N. M. Ralbovsky, L. Halámková, K. Wall, C. Anderson-Hanley, I. K. Lednev, *J. Alzheimer's Dis.* **2019**, *71*, 1351.

[149] E. Ryzhikova, N. M. Ralbovsky, V. Sikirzhytski, O. Kazakov, L. Halamkova, J. Quinn, E. A. Zimmerman, I. K. Lednev, *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2021**, *248*, 119188.

[150] N. M. Ralbovsky, G. S. Fitzgerald, E. C. McNay, I. K. Lednev, *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2021**, *254*, 119603.

[151] S. Huang, R. Pandey, I. Barman, J. Kong, M. Dresselhaus, *ACS Photonics.* **2018**, *5*, 2978.

[152] Z. Y. Wang, J. R. Ye, K. Y. Zhang, L. Ding, T. Granzier-Nakajima, J. C. Ranasinghe, Y. Xue, S. Sharma, I. Biase, M. Terrones, S. H. Choi, C. Z. Ran, R. E. Tanzi, S. X. Huang, C. Zhang, S. X. Huang, *ACS Nano* **2022**, *16*, 6426.

[153] T. R. Mercer, M. Salit, *Nat. Rev. Genet.* **2021**, *22*, 415.

[154] S. Iravani, *Mater. Adv.* **2020**, *1*, 3092.

[155] D. Chen, *Appl. Artif. Intell.* **2021**, *35*, 1147.

[156] G. Yin, L. T. Li, S. Lu, Y. Yin, Y. Z. Su, Y. L. Zeng, M. Luo, M. H. Ma, H. Y. Zhou, L. Orlandini, D. Z. Yao, G. Liu, J. Y. Lang, *J. Raman Spectrosc.* **2021**, *52*, 949.

[157] K. Ember, F. Daoust, M. Mahfoud, F. Dallaire, E. Z. Ahmad, T. Tran, A. Plante, M. K. Diop, T. Nguyen, A. St-Georges-Robillard, N. Ksantini, J. Lanthier, A. Filiatrault, G. Sheehy, G. Beaudoin, C. Quach, D. Trudel, F. Leblond, *J. Biomed. Opt.* **2022**, *27*, 025002.

[158] G. W. Auner, S. K. Koya, C. H. Huang, B. Broadbent, M. Trexler, Z. Auner, A. Elias, K. C. Mehne, M. A. Brusatori, *Cancer Metastasis Rev.* **2018**, *37*, 691.

[159] L. Zhang, C. Li, D. Peng, X. Yi, S. He, F. Liu, X. Zheng, W. Huang, L. Zhao, X. Huang, *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2022**, *264*, 120300.

[160] Y. Qi, L. Yang, B. Liu, L. Liu, Y. Liu, Q. Zheng, D. Liu, J. Luo, *Anal. Chim. Acta* **2021**, *1179*, 338821.

[161] H. Leng, C. Chen, R. Si, C. Chen, H. Qu, X. Lv, *J. Raman Spectrosc.* **2022**, *53*, 1302.

[162] B. Zhou, K. Xu, X. Zheng, T. Chen, J. Wang, Y. Song, Y. Shao, S. Zheng, *Signal Transduct. Target Ther.* **2020**, *5*, 144.

[163] H. Shin, S. Oh, S. Hong, M. Kang, D. Kang, Y. Ji, B. Choi, K. Kang, H. Jeong, Y. Park, S. Hong, *ACS Nano* **2020**, *14*, 5435.

[164] H. Zhang, C. Chen, R. Gao, Z. Yan, Z. Zhu, B. Yang, C. Chen, X. Lv, H. Li, Z. Huang, *Photodiagnosis Photodyn. Ther.* **2021**, *33*, 102104.

[165] C. Ciobanu, K. J. Ember, B. J. Nyíri, S. Rajan, V. Chauhan, F. Leblond, S. Murugkar, *IEEE Instrum. Meas. Mag.* **2022**, *25*, 62.

[166] F. Chen, C. Sun, Z. Yue, Y. Zhang, W. Xu, S. Shabbir, L. Zou, W. Lu, W. Wang, Z. Xie, L. Zhou, *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2022**, *265*, 120355.

[167] P. Wang, L. Guo, Y. Tian, J. Chen, S. Huang, C. Wang, P. Bai, D. Chen, W. Zhu, H. Yang, W. Yao, *OSA Contin.* **2021**, *4*, 672.

[168] M. Li, H. He, G. Huang, B. Lin, H. Tian, K. Xia, C. Yuan, X. Zhan, Y. Zhang, W. Fu, *Front. Oncol.* **2021**, *11*, 665176.

[169] F. Qeadan, P. Bansal, J. A. Hanson, E. J. Beswick, *J. Transl. Med.* **2020**, *18*, 137.

[170] W. Zhang, I. Karagiannidis, E. D. S. Van Vliet, R. Yao, E. J. Beswick, A. Zhou, *Analyst* **2021**, *146*, 6124.

[171] C. He, X. Wu, J. Zhou, Y. Chen, *J. Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2021**, *252*, 119520.

[172] J. Ye, Y. T. Yeh, Y. Xue, Z. Wang, N. Zhang, H. Liu, K. Zhang, R. Ricker, Z. Yu, A. Roder, N. P. Lopez, *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2118836119.

[173] A. A. Moawad, A. Silge, T. Bocklitz, K. Fischer, P. Rösch, U. Roesler, M. C. Elschner, J. Popp, H. Neubauer, *Molecules* **2019**, *24*, 4516.

[174] K. Kukula, D. Farmer, J. Duran, N. Majid, C. Chatterley, J. Jessing, Y. Li, *Computing and Communication Workshop and Conference*, IEEE, New York **2021**.

[175] S. Yu, X. Li, W. Lu, H. Li, Y. V. Fu, F. Liu, *Anal. Chem.* **2021**, *93*, 11089.

[176] S. Yan, S. Wang, J. Qiu, M. Li, D. Li, D. Xu, D. Li, Q. Liu, *Talanta* **2021**, *226*, 122195.

[177] J. Zahn, A. Germond, A. Y. Lundgren, M. T. Cicerone, *J. Biophotonics.* **2022**, *15*, e202100274.

[178] A. Nakar, A. Pistiki, O. Ryabchykov, T. Bocklitz, P. Rösch, J. Popp, *Anal. Bioanal. Chem.* **2022**, *414*, 1481.

[179] A. Walter, M. Reinicke, T. Bocklitz, W. Schumacher, P. Rösch, E. Kothe, J. Popp, *Anal. Bioanal. Chem.* **2011**, *400*, 2763.

[180] A. Germond, T. Ichimura, T. Horinouchi, H. Fujita, C. Furusawa, T. M. Watanabe, *Commun. Biol.* **2018**, *1*, 85.

[181] N. Kuhar, S. Sil, T. Verma, S. Umapathy, *RSC Adv.* **2018**, *8*, 25888.

[182] C. Zong, M. Xu, L. J. Xu, T. Wei, X. Ma, X. S. Zheng, R. Hu, B. Ren, *Chem. Rev.* **2018**, *118*, 4946.

[183] J. Xu, X. Yi, G. Jin, D. Peng, G. Fan, X. Xu, X. Chen, H. Yin, J. M. Cooper, W. E. Huang, *ACS Chem. Biol.* **2022**, *17*, 376.

[184] C. S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. Saleh, S. Ermon, J. Dionne, *Nat. Commun.* **2019**, *10*, 4927.

[185] A. Choudhary, N. Gupta, F. Hameed, S. Choton, *Int. J. Chem. Stud.* **2020**, *8*, 2564.

[186] H. Kendall, B. Clark, C. Rhymer, S. Kuznesof, J. Hajslova, M. Tomaniova, P. Brereton, L. Frewer, *Trends Food Sci. Technol.* **2019**, *94*, 79.

[187] A. M. Herrero, *Food Chem.* **2008**, *107*, 1642.

[188] Z. Chen, T. Wu, C. Xiang, X. Xu, X. Tian, *Molecules* **2019**, *24*, 2851.

[189] D. I. Ellis, D. Broadhurst, S. J. Clarke, R. Goodacre, *Analyst* **2005**, *130*, 1648.

[190] C. Robert, S. J. Fraser-Miller, W. T. Jessep, W. E. Bain, T. M. Hicks, J. F. Ward, C. R. Craigie, M. Loeffen, K. C. Gordon, *Food Chem.* **2021**, *343*, 128441.

[191] C. Berghian-Grosan, D. A. Magdas, *Talanta* **2020**, *218*, 121176.

[192] S. Duraipandian, J. C. Petersen, M. Lassen, *Appl. Sci.* **2019**, *9*, 2433.

[193] H. Zhao, Y. Zhan, Z. Xu, J. J. Nduwamungu, Y. Zhou, R. Powers, C. Xu, *Food Chem.* **2022**, *373*, 131471.

[194] M. Oroian, S. Ropciuc, S. Paduret, *Food Anal. Methods.* **2018**, *11*, 959.

[195] S. Hu, H. Li, C. Chen, C. Chen, D. Zhao, B. Dong, X. Lv, K. Zhang, Y. Xie, *Sci. Rep.* **2022**, *12*, 3456.

[196] C. Berghian-Grosan, D. A. Magdas, *Sci. Rep.* **2020**, *10*, 21152.

[197] J. Liu, C. Li, H. Lü, W. Kong, W. Sun, G. Zhang, *Laser Optoelectron. Prog.* **2021**, *58*, 1630004.

[198] W. Gao, L. Zhou, S. Liu, Y. Guan, H. Gao, B. Hui, *Bioresour. Technol.* **2022**, *348*, 126812.

[199] W. Gao, L. Zhou, S. Liu, Y. Guan, H. Gao, J. Hu, *Carbohydr. Polym.* **2022**, *292*, 119635.

[200] C. Yan, Z. Cheng, S. Luo, C. Huang, S. Han, X. Han, Y. Du, C. Ying, *J. Raman Spectrosc.* **2022**, *53*, 260.

[201] W. Qi, T. Mu, S. Chen, Y. Wang, *J. Raman Spectrosc.* **2022**, *53*, 746.

[202] S. Dargan, M. Kumar, M. R. Ayyagari, G. A. Kumar, *Arch. Comput. Methods Eng.* **2020**, *27*, 1071.

[203] C. Shorten, T. M. Khoshgoftaar, *J. Big Data* **2019**, *6*, 60.

[204] S. Di Frischia, P. Giammatteo, F. Angelini, V. Spizzichino, E. De Santis, L. Pomante, presented at IEEE International Conference on Big Data, 2891, Atlanta, GA, USA **2020**.

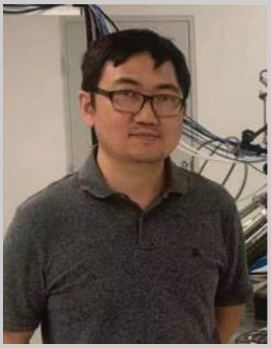
[205] Z. S. PAN, B. CHEN, Z. M. MIAO, G. Q. NI, *Dianzi Xuebao* **2009**, *37*, 2496.

[206] A. B. Nassif, M. A. Talib, Q. Nasir, F. M. Dakalbab, *IEEE Access* **2021**, *9*, 78658.

[207] K. Hofer-Schmitz, P. H. Nguyen, K. Berwanger, presented at ESANN 2018 Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 189, Bruges, Belgium **2018**.

Yaping Qi is a Postdoctoral Researcher at Macau University of Science and Technology and a visiting scholar at Aarhus University, Denmark. She obtained Ph.D. degree in Physics in 2019 at the University of Hong Kong. She was a Research Fellow in Prof. David A. Weitz's group at Harvard University from 2017 to 2018 and Research Associate at Purdue University from 2019 to 2021. She was a visiting postdoc at Aarhus University, Denmark from October to December 2021, and in the Department of Physics and Astronomy, at the University of British Columbia in Canada from January to March 2022. She is currently working on 2D materials, thin films, and AI.

Dan Hu is currently a master's student in Prof. Yong Chen's group in intelligent technology at the Macau University of Science and Technology. She received her B.S. degree in Science (Computer Technology and Application) from Macau University of Science and Technology. She is interested in artificial intelligence technologies, including machine learning and deep learning, and their applications in multi-domain research.



Yong P. Chen is Karl Lark-Horovitz Professor of Physics and Astronomy and Professor of Electrical and Computer Engineering, Director of Purdue Quantum Science and Engineering Institute at Purdue University. He is also a Villum Investigator and Professor at Aarhus University (Demark) and Principal Investigator in WPI (World Premier International Research Center)-AIMR (Advanced Institute for Materials Research) at Tohoku University (Japan) and MUST Chair Professor at Macau University of Science and Technology. He received his M.Sc. at MIT and Ph.D. at Princeton University. His research interests include nano/solid state physics (graphene & 2D materials, topological insulators) and atomic/molecular physics (Bose-Einstein condensates, polar molecules).