Data-Driven AI Techniques in Raman Spectroscopy for Biomedical Applications: A Comprehensive Review

Shahbaz Ahmed
Informatik
Universität Koblenz
Koblenz, 56070, Germany
shahbaz03@uni-koblenz.de
ORCID: 0009-0003-7537-9502

Muhammad Raheel Raza
School of Computer Science
The University of Sydney
Camperdown, NSW 2050, Australia
muhammadraheel.raza@sydney.edu.au
ORCID: 0000-0002-6305-2583

Asaf Varol
Dept. of Eng. Management & Tech.
College of Eng. & Comp. Science
Uni. of Tennessee at Chattanooga
Chattanooga, TN, US
asaf-varol@utc.edu

Department of Computer Engineering
College of Eng. & Natural Sciences
Maltepe University
Istanbul, Türkiye
asafvarol@maltepe.edu.tr
ORCID: 0000-0003-1606-4079

2026 14th International Symposium on Digital Forensics and Security (ISDFS) | 979-8-3315-7310-2/26/$31.00 ©2026 IEEE | DOI: 10.1109/ISDFS69419.2026.11458934

***Abstract*—Raman Spectroscopy has emerged as a significant method for biological diagnosis. It emerged because of its ability to provide molecular-level information without the use of labels. The use of machine learning (ML) techniques, especially clustering and deep learning (DL), has strongly improved the analysis of Raman spectra data, thus enhancing the classification and detection of different cancer variants and microbial infections. This study highlights how supervised learning algorithms, such as Support Vector Machines (SVM) and Random Forests, and unsupervised learning algorithms, such as hierarchical, PCA, and K-means clustering, contribute to interpreting and classifying the complex spectral data more accurately. Deep learning models such as CNN is also used to examine their performance in biomarker identification and pattern recognition. This survey paper shows how diagnostic accuracy is improved by merging Raman Spectroscopy with intelligent models, highlighting some key limitations and future directions, including noisy data, high computational demands, and dependence on labeled data. It contributes to a continued development of advanced diagnostic techniques based on spectroscopy and intelligent data analysis.**

Keywords—Raman spectroscopy, Machine Learning, Deep Learning, CNN, PCA, K-means, Pattern recognition.

I. Introduction

Similar to Fourier transform infrared spectroscopy FTIR, Raman spectroscopy is a molecular spectroscopic technique that employs the interaction of light with matter to yield information on the composition or qualities of an item. While IR spectroscopy depends on light absorption, Raman spectroscopy uses a light dispersion technique to offer information. Through the disclosure of information on intramolecular and intermolecular vibrations, Raman spectroscopy has the potential to provide further insight into a process. Both Raman and Fourier transform infrared spectroscopy (FTIR) are effective for identifying a substance because they provide a spectrum that is characteristic of the unique vibrations of a molecule, often known as a "molecular fingerprint." On the other hand, Raman spectroscopy has the ability to offer additional information on lower frequency modes and vibrations, which in turn provide information regarding the structure of the molecular backbone and the crystal lattice [1].

Raman spectroscopy is an advanced molecular spectroscopic analytical method that has become increasingly popular in recent years. As an analytical tool, it can provide detailed molecular information in a nondestructive manner, which makes it a good technique to use in biology, biomedical analysis, and forensics [2, 3]. The use of Raman spectroscopy has had a significant influence on the field of biomedical research, notably in cancer diagnosis, the guidance of surgical procedures, and the identification of microbiological infections [1]. The ability of Raman spectroscopy to diagnose cancer is among its most significant achievements [4]. The technique's capacity to identify molecular alterations in cells and tissues enables early diagnosis and precise characterization of malignant growths. By using technology to differentiate between healthy and sick tissues, surgeons may ensure that malignant cells are precisely removed while maintaining the greatest amount of healthy tissue [2]. This efficient and non-invasive method offers a valuable understanding of the molecular makeup of biological fluids [5]. Specifically, human blood plasma is a multifaceted biological fluid that consists of proteins, lipids, nucleic acids, carbohydrates, and more [6]. Consequently, it serves as a tool for examining the spectral signatures of plasma blood, yielding significant diagnostic insights [7, 8].

The subsequent sections of this study demonstrate the innovations facilitated by Raman Spectroscopy through systematic data-driven and artificial intelligence-focused techniques employed for spectrum interpretation. Some sections focus on combining powerful ML and clustering approaches to obtain biochemical information from intricate Raman signals. Sections 2 and 3 primarily provide a detailed, organized overview of AI-driven approaches for analyzing Raman spectra, especially in biomedical fields such as identifying photogenic bacteria and cancer cells. Section 4 of this study showcases the novelty expressed in its comparative approach, which solidifies previous accomplishments, highlights existing limits, and provides future research paths for data-driven Raman spectroscopy.

II. Understanding Machine Learning Framework

Machine learning is a subset of artificial intelligence that concentrates on creating mathematical algorithms and models to allow computers to recognize patterns and make predictions autonomously, without considerable manual programming. Machine learning models can enhance their performance over time through an iterative process that fosters the development of "intelligence." The choice of machine learning models is often determined by criteria such as the nature of the issues, data characteristics, and expected outcomes. Machine learning may be roughly categorized into three types: unsupervised, supervised, and reinforcement learning [9–11].

For unsupervised learning, mathematical algorithms are designed to search for patterns or correlations among data that has not been labeled. To be more specific, the data does not have any predetermined output labels, and the major objective of unsupervised learning is to uncover information that was previously unknown or concealed within the data. K-means, hierarchical, mean shift, and density-based clustering are some examples of the types of algorithms that are utilized in unsupervised learning. Unsupervised learning is frequently utilized for dimensionality reduction tasks, such as principal component analysis (PCA), which is one of the most often utilized approaches in unsupervised learning [13, 14].

In supervised learning, mathematical algorithms are initially trained on a labeled dataset, where the input data corresponds to certain output labels [15, 16]. The algorithms then learn to correlate certain inputs with their corresponding outputs, and their efficiency is assessed based on their ability to predict outputs for novel, unknown data. Supervised learning is generally utilized for classification and regression problems. In both jobs, the input data often comprises many characteristics or qualities. In classification, models strive to discern the decision boundary that differentiates input data into specific classes or categories, resulting in discrete output data. In regression, the algorithms endeavor to discern the fundamental trend within the input data, resulting in continuous output data. Common supervised learning methods encompass linear discriminant analysis (LDA), decision trees, random forests, support vector machines (SVM), partial least-squares regression (PLSR), and partial least-squares discriminant analysis (PLS-DA). LDA is considered a supervised learning technique for dimensionality reduction, effective for addressing classification issues [17].

An illustrative instance is the Support Vector Machine (SVM), which is a widely utilized supervised learning algorithm that effectively addresses intricate data sets and nonlinear relationships [18, 19]. This algorithm demonstrates effectiveness in high-dimensional spaces, operating by identifying the optimal hyperplane that distinguishes various classes within the input space, all while maximizing the separation distance between these classes. The main goal of SVM is to determine a decision boundary that effectively distinguishes the data points into separate classes. In addition to the widely utilized algorithms, various other supervised learning techniques are gaining traction in enhancing cancer discrimination, such as PLSR, PLS-DA, and its orthogonal counterpart OPLS DA [20–25].

While traditional supervised learning models have demonstrated significant promise in aiding cancer diagnosis, there is a growing focus on the development of more sophisticated supervised learning models, especially deep learning models, to overcome certain limitations inherent in the conventional approaches [24–26]. For example, deep learning has the capability to automate specific feature extraction processes, thereby minimizing the manual interventions commonly required in traditional supervised learning. Furthermore, although traditional supervised learning models may face challenges with high-dimensional data, deep learning models are capable of managing such data effectively and capturing the intricate relationships within it through their advanced multilayer network architecture. Deep learning, a subfield of machine learning influenced by the workings of the human brain, emphasizes the creation and application of neural networks. These networks are designed to autonomously learn hierarchical representations of data, enabling the extraction of complex patterns and features. Generally, many of the advanced deep learning models, especially those utilizing artificial neural networks (ANNs) and convolutional neural networks (CNNs), exhibit certain similarities in their structures and operations [27–29]. Artificial Neural Networks (ANNs), recognized as some of the pioneering algorithms in the realm of deep learning, are composed of a network of interconnected nodes. These nodes are systematically arranged into various layers. Although artificial neural networks exhibit considerable adaptability in both classification and regression tasks, convolutional neural networks are particularly proficient in performing specialized tasks related to images, such as object detection and image recognition [30, 31]. The architecture of convolutional neural networks (CNNs) is composed of convolutional layers, pooling layers, and fully connected layers. These distinct layers work in concert to effectively capture local patterns and hierarchies present in grid-like data, facilitate the reduction of spatial dimensions, and adeptly process various features.

While the majority of research has concentrated on unsupervised and supervised learning, recent years have witnessed the emergence of alternative machine learning methodologies, including semi supervised, transfer, and reinforcement learning, to mitigate certain limitations inherent in traditional unsupervised and supervised learning approaches [11–13] Semi supervised learning may be seen as a combination of unsupervised and supervised learning [32, 33]. Semi-supervised learning effectively utilizes a small quantity of labeled data with a substantial volume of unlabeled data, making it advantageous in scenarios where obtaining extensive labeled data is prohibitively expensive, labor-intensive, or time-consuming. In recent years, the exploration of machine learning has gained momentum due to its numerous appealing characteristics, particularly in enhancing cancer diagnosis. Numerous machine learning models are presently being combined with nanomaterial-enabled optical spectroscopy to discern intricate and nuanced molecular and cellular alterations linked to cancer, which could potentially lead to earlier and more precise cancer detection.

III. Methodology

The combination of machine learning techniques with nanomaterial-mediated optical spectroscopy presents an exciting opportunity to advance cancer detection. This integration facilitates the identification and characterization of complex spectral patterns that may indicate cancer within intricate optical spectra. As of now, both unsupervised and supervised machine learning techniques have been employed to enhance the efficacy of diverse nanomaterial-mediated optical spectroscopy methods. In particular, the application of

supervised machine learning in conjunction with nanomaterial-mediated optical spectroscopy for the purpose of cancer detection can be achieved through a series of clearly defined steps, as shown in Fig. 1[14]. The application of nanomaterials specifically designed for targeting cancer involves their initial use in labeling particular cancer cells. This process is aimed at enhancing the readout signals emitted by these cells, as illustrated in Fig. 1a. The acquisition of various optical spectra about both normal and nanomaterial-labeled cancer cells is subsequently conducted, as shown in Fig. 1b. After gathering a substantial collection of these optical spectra, they are systematically divided into training and testing data sets. These sets are subsequently utilized to train and evaluate the machine learning models demonstrated in Fig. 1c. The models that have been developed, demonstrating satisfactory and reproducible performance, can subsequently be employed to analyze optical spectral data from patients whose conditions are unknown. This application holds the potential to enhance the processes of cancer diagnosis and prognosis, as stated in Fig. 1d.

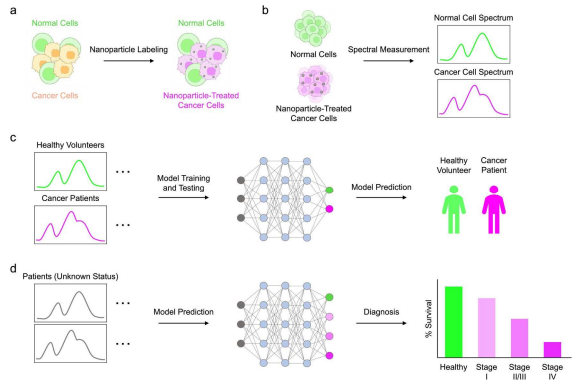

Fig. 1. Workflow overview of machine-learning-assisted optical spectroscopic detection of cancer using nanomaterials [14].

In addition to traditional supervised learning algorithms, there has been an increasing amount of work investigating the application of more sophisticated machine learning techniques, frequently centered on deep multilayer perceptrons (MLPs). This type of representation learning is known as deep learning, and it has the capability to address a range of artificial intelligence tasks [41]. In the context of Raman spectroscopy, deep learning models demonstrate significant utility in both the pre-processing and modeling of data. Theoretically, these advanced techniques can be applied across a diverse array of Raman spectral data. In the presence of a large number of Raman spectra, it is feasible to input them directly into deep learning models, bypassing the need for any pre-processing steps. When employing Raman spectra without the application of pre-processing techniques, it is essential for the deep learning model to inherently perform this function in conjunction with the tasks of classification or regression. In the context of various Raman experiments, it is possible to either retrain the models or utilize them directly for a new experiment or task. In addition, the predominant deep learning algorithms utilized within this domain include convolutional neural networks (CNNs), residual networks (ResNets), recurrent neural networks (RNNs), autoencoders, and generative adversarial networks (GANs) [42].

In general, the steps of data pre-processing, feature extraction (or feature selection), and data modeling are essential components of the analytical process. Conversely, in the discipline of deep learning, the intricate tasks traditionally requiring multiple steps can be efficiently managed by a singular neural network, provided that a sufficient amount of training data is available. According to the various output types, deep learning applications for Raman spectroscopy can be categorized into four primary components: pre-processing, classification, regression, and highlighting, as shown in Fig. 2. Upon the completion of model training utilizing a Raman spectrum as the input, the pre-processing model generates an alternative Raman spectrum, which is typically subjected to filtering or de noising processes. Concurrently, the classification model produces a corresponding label, while the regression model yields a numerical value or a probabilistic estimate. Additionally, the highlighting model systematically segments the input into distinct components, often resulting in the identification of a specific region of interest (ROI) within the 1D spectral data.

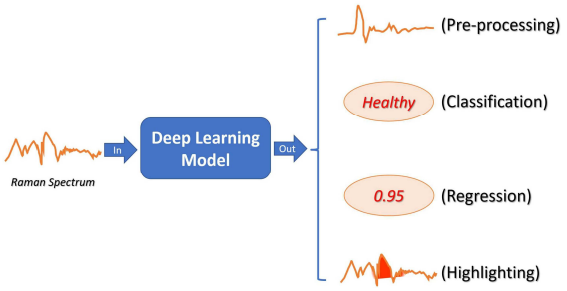

Fig. 2. Four categories of deep learning applications in Raman spectroscopy.

IV. Results and Discussion

A variety of unsupervised machine learning algorithms have been employed in combination with nanomaterial-mediated optical spectroscopy for the purpose of detecting cancer biomarkers. This encompasses PCA along with its numerous variants. For instance, a modified principal component analysis incorporating feature selection was integrated with antibody-conjugated CdSe/ZnS quantum dots to enhance the fluorescence detection of various gastrointestinal tumor markers, particularly CA-125, CA-19-9, cancer embryonic antigen, and alpha-fetoprotein [34]. The PCA-based approach that was reported demonstrated significantly enhanced precision, achieving rates of 99.52% for colon tumors and 99.03% for gastric cancer. Additionally, it exhibited superior accuracy, with figures of 94.86% for colon tumors and 94.2% for gastric cancer, in comparison to alternative algorithms.

To enhance its performance, Principal Component Analysis (PCA) can be integrated with various machine learning algorithms, one notable example being the supervised Linear Discriminant Analysis (LDA). An integrated PCA-LDA method has been employed to improve both the sensitivity and specificity of an SERS-enabled blood test utilized for the diagnosis of nasopharyngeal cancer [35]. The enhancement of Raman scattering in biomolecules present within blood plasma was achieved through the incorporation of silver nanoparticles. The analysis revealed distinct cancer-specific variations, characterized by elevated levels of nucleic acids, collagen, phospholipids, and phenylalanine, alongside reduced concentrations of amino acids and saccharides. These findings were derived from the SERS spectra, which were subsequently categorized into clusters representing healthy volunteers and cancer patients through the application of PCA. The implementation of LDA resulted in an enhancement of sensitivity to 90.7% and attained a specificity of 100%,

thereby effectively differentiating SERS spectra between individuals diagnosed with cancer and healthy volunteers. In a comparable manner, the PCA-LDA algorithm has been employed to improve the detection of colorectal cancer through the application of Au nanoparticle-enabled SERS spectroscopy [36]. The assignments of Raman bands in SERS spectra revealed a reduction in saccharides and proteins, while an increase in nucleic acids was observed in patients with cancer. The diagnostic algorithms that utilize PCA-LDA demonstrated an impressive sensitivity of 97.4% and a perfect specificity of 100% in the classification of SERS spectra, effectively differentiating cancer samples from their normal counterparts. In addition to Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) has been integrated with various algorithms, such as Partial Least Squares Discriminant Analysis (PLS-DA), to enhance the capability of Surface-Enhanced Raman Spectroscopy (SERS) in cancer delineation, as demonstrated in a recent study [37]. This study presents the integration of Principal Component Analysis (PCA) with Partial Least Squares Discriminant Analysis (PLSDA) to enhance the identification of bladder cancer through the analysis of urine samples utilizing gold nanoparticle- based Surface-Enhanced Raman Spectroscopy (SERS). The dimensionality of the Raman spectra underwent reduction via Principal Component Analysis (PCA), subsequently leading to the classification of noncancerous and cancerous groups through Partial Least Squares Discriminant Analysis (PLS-DA).

Significantly, through the integration of the PCAPLS-DA model with SERS spectroscopy, the research demonstrated that bladder cancer in its early and polyp stages could be diagnosed with an impressive accuracy of 99.6%. Numerous research studies employ unsupervised PCA to enhance the cancer diagnostic efficacy of nanomaterial-mediated optical spectroscopy, while various other machine learning algorithms, including supervised random forest and SVM, are also under active investigation. In a recent study, a random forest classifier was utilized in conjunction with a SERS detection platform that included immuno-magnetic capture probes and Ag nanoparticle-based immuno-SERS reporters to analyze various extracellular vesicle proteins and differentiate breast cancer subtypes from single-cell SERS spectra [38]. The random-forest-based model that was developed demonstrated a notable classification accuracy of 87.5% for metastatic breast cancer cells, while achieving an impressive 100% accuracy for nonmetastatic breast cancer cells. Both CD9 and focal adhesion kinase have been identified and suggested as potential biomarkers for the differentiation of metastatic breast cancer cells from their nonmetastatic counterparts. In a distinct approach, an Ag nanowire-enhanced surface-enhanced Raman scattering (SERS) technique was integrated with supervised support vector machine (SVM) analysis. This combination was employed to differentiate various cancer types from healthy controls by analyzing the Raman spectra obtained from serum samples [39]. The developed model utilizing support vector machines demonstrated an impressive accuracy of 95.81%, alongside a sensitivity of 95.87% and a specificity of 95.40% in the classification of pan-cancer groups as compared to the healthy group. Moreover, the strategy that has been reported demonstrates the capability to differentiate between gastric, liver, lung, and colorectal cancers and noncancerous diseases originating from the same organs. The specific delineations of gastric cancer and hepatocellular carcinoma were accomplished with an impressive accuracy of 93.33% and 92.31%, respectively. In a further illustration, a technique utilizing Ag nanorods for surface-enhanced Raman scattering (SERS) was combined with orthogonal partial least squares discriminant analysis (OPLS-DA) to examine the Raman spectra of human serum samples, thereby aiding in the detection and staging of lung adenocarcinoma [40]. The application of the OPLS-DA algorithm demonstrated an excellent efficacy in classifying the Raman spectra obtained from lung adenocarcinoma and control groups, achieving a specificity of 97.6% and a sensitivity of 98.1%. Furthermore, the staging of lung adenocarcinoma demonstrates an accuracy of 84.3% for stage I, 93.3% for stage II, and 86.5% for stages III and IV cancer. In a particular study, artificial neural networks (ANN) were combined with various methods such as logistic regression, decision trees, random forests, and support vector machines (SVM) to enhance the fluorescence detection of ovarian cancer in patient serums using single-walled carbon nanotubes [43]. The described method exhibited a sensitivity of over 80% and a specificity surpassing 90% in identifying high-grade serous ovarian cancer in symptomatic patients.

CNN is a prevalent deep learning technique utilized for cancer diagnosis using nanomaterial-mediated optical spectroscopy. This architecture has been employed to assess the SERS spectra of clinical samples, including patient blood, mRNA, and exosomes, to differentiate various cancers, such as prostate, liver, head and neck, and lung cancer [44-47], with excellent sensitivity and specificity. A CNN architecture utilizing a residual neural network (ResNet) was created for the early diagnosis of lung cancer through the analysis of SERS spectra from exosomes derived from human plasma [47]. The SERS spectra were initially obtained from the exosomes of both normal and malignant lung cells, as well as from the exosomes of plasma from healthy volunteers and lung cancer patients. A ResNet-based model comprising a convolutional layer, a pooling layer, several sequential basic blocks, and fully connected layers was subsequently built to identify lung cancer exosomes and determine cancer staging. To assess the efficacy of the created model in identifying lung cancer, the similarity across several groups of exosomes was initially determined using PCA, followed by a distance-based similarity analysis. The comparative similarity of different human plasma exosomes was further assessed in relation to the exosomes derived from lung cancer cells, revealing a clear correlation with the stages of lung cancer. Moreover, lung cancer in all patients was predicted with an area under the curve (AUC) of 0.912, whereas an AUC of 0.910 was achieved for predicting stage I lung cancer patients. The new ResNet-based CNN architecture surpassed traditional machine learning algorithms and other deep learning models in the classification of cellular exosomes and the prediction of cancer development [14].

Recently, several CNN-based deep learning models, including the background removal network, data augmentation network, and prostate cancer detection network, were utilized in conjunction with a hyperspectral SERS-mediated technique to characterize the molecular phenotypes of biofluids [48]. The integrated technique, enhanced by deep learning, facilitated the differentiation of prostate cancer from benign prostatic hyperplasia in human serum, achieving an overall accuracy of 80.8%, significantly surpassing that of the conventional prostate-specific antigen test and the prostate imaging reporting and data system. A CNN-based technique

was created to facilitate nanomaterial-mediated fluorescence detection and diagnosis of breast cancer [49]. A collection of fluorescent probes, namely P1 to P12, was initially synthesized and utilized to label exosomes obtained from various malignant breast epithelial cell lines and one normal breast epithelial cell line. Latent Dirichlet Allocation (LDA) was then employed to categorize the fluorescence spectra of distinct exosome groups, confirming the method's high precision. A multitude of fluorescence spectra were subsequently obtained from the exosomes of breast cancer patients and healthy donors. The unprocessed fluorescence spectra were analyzed to create multichannel Fmaps containing 1562 fluorescence feature points. Detailed analyses of the resulting Fmaps demonstrated that different exosome samples exhibited distinct Fmaps. The Fmaps of patients exhibited a significant degree of resemblance, contrasting sharply with those of healthy participants. This enhances the precision of breast cancer patient screening.

TABLE I. SUMMARY OF CANCER DETECTION STUDIES USING RAMAN SPECTROSCOPY COMBINED WITH ML AND DL TECHNIQUES.

Technique	Spectroscopy	Cancer Type	Performance	Ref.
PCA + Fea-	Quantum dots	Colon,	Precision: 99.52%	[34]
ture Selection	(CdSe/ZnS)	Gastric	(colon), 99.03%	
			(gastric); Accuracy:	
			94.86%, 94.2%	
PCA + LDA	SERS (Ag NPs,	Naso-	Sensitivity: 90.7%,	[35]
	blood plasma)	pharyngea	Specificity: 100%	
PCA + LDA	SERS (Au	Colorectal	Sensitivity: 97.4%,	[35]
	NPs, serum)		Specificity: 100%	
PCA+PLS-	SERS (gold	Bladder	Accuracy: 99.6%	[36]
DA	NPs, urine)			
Random Forest	SERS (Ag NPs,	Breast	Accuracy: 87.5%	[37]
	EV proteins)	types)	(metastatic), 100%	
			(nonmetastatic)	
SVM	SERS (Ag	Pancancer	Accuracy: 95.81%,	[38]
	nanowires, serum)		Sensitivity: 95.87%,	
			Specificity: 95.40%	
OPLS-DA	SERS (Ag	Lung adeno	Sensitivity: 98.1%,	[39]
	nanorods, serum)	carcinoma	Specificity: 97.6%;	
			Staging Accuracy:	
			84.3%-93.3%	
ANN +other	Fluorescence	Ovarian	Sensitivity: >80%,	[43]
models	(CNTs)		Specificity: >90%	
ResNet (CNN)	SERS (plasma	Lung	AUC: 0.912 (all),	[47]
	exosomes)		0.910 (Stage I)	
CNN (multi-	Hyperspectral			
ple networks)	SERS (biofluids)	Prostate	Accuracy: 80.8%	[48]
CNN+LDA	Fluorescence	Breast	High precision, clear	[49]
	(exosomes)		separation of Fmaps	

V. CONCLUSION AND FUTURE WORK

Raman spectroscopy, when combined with clustering-based AI models, has played an important role in the advancement of early cancer detection and microbial diagnosis. SVM, PCA, LDA, and PLSDA techniques have proven high accuracy. Whereas handling complex spectral data efficiently is now possible using deep learning models such as CNNs. Although these models are efficient, they still have some limitations, such as complex pre-processing, a need for huge amounts of labeled data, and difficulties in understanding how these models make decisions. Additionally, CNN exhibited diminished effectiveness as CNN-based methodologies often require extensive, well-annotated datasets to develop robust feature hierarchies, which are seldom available in Raman studies. Furthermore, Raman spectra are one-dimensional signals with slight peak shifts, whereas CNN architectures are mainly intended for spatial feature extraction in images. This restricts the efficiency of convolutional operations without substantial preprocessing or architectural alterations. Since PCA can effectively reduce high-dimensional, highly correlated spectral data into a small set of orthogonal components that retain the most informative variance while mitigating noise and redundancy, it performs well in Raman spectroscopy studies because biomedical Raman datasets are frequently small and subject to experimental variability. PCA is especially well-suited for them, improving class separability and generalization.

Future techniques can be made aiming to fill the gaps in existing techniques by adopting hybrid methods that integrate the strengths of DL, SL, and clustering. This will not only enhance the accuracy of cancer detection but also improve the clinical acceptance rate by giving a clear understanding of the decision-making process. Furthermore, machine learning and deep learning techniques have revealed a great achievement in detecting cancer using spectroscopy, but these techniques still have limitations. Some of the traditional methods mostly depend on manual preprocessing techniques like principal component analysis and Least Discriminant Analysis. These techniques fight with noise, high-dimensional data, and zero shifts that are very common in Raman spectroscopy. Models that use deep learning techniques require larger datasets with labels and significant computing power, which is like a "black box". These factors made it difficult to understand how conclusions are made. Moreover, these existing techniques also lack the capacity to simplify across a variety of data types or experimental setups. In order to overcome these complications, an innovative approach might be used by bringing together the advantages of both methodologies, with unsupervised feature learning merging with interpretable AI models. The performance, robustness, and transparency of the technique can be made better by using an explainable deep learning model, which may use attention-based approaches or a fusion model with both explainable deep learning techniques and signal processing. Furthermore, their practical worth in clinical diagnostics might be greatly enhanced by creating models that respond to new data with minimal training and need less preprocessing.

REFERENCES

[1] Mettler-Toledo, "Raman Spectroscopy | Instrumentation, Introduction & Principle," Mettler-Toledo. [Online]. Available: https://www.mt.com/au/en/home/applications/L1_AutoChem_Applications/Raman-Spectroscopy.html?GLO_YT_Autochem_OTH_Youtube_Autochem,

[2] E. S. Allakhverdiev, et al., "Spectral insights: Navigating the frontiers of biomedical and microbiological exploration with Raman spectroscopy," Journal of Photochemistry and Photobiology B: Biology, vol. 252, p. 112870, 2024.

[3] M. Deluca, H. Hu, M. N. Popov, J. Spitaler, and T. Dieing, "Advantages and developments of Raman spectroscopy for electroceramics," Communications Materials, vol. 4, no. 1, p. 78, 2023.

[4] H. Hano, et al., "Power of light: Raman spectroscopy and machine learning for the detection of lung cancer," ACS Omega, vol. 9, no. 12, pp. 14084-14091, 2024.

[5] D. Cialla-May, et al., "Raman spectroscopy and imaging in bioanalytics," Analytical Chemistry, vol. 94, no. 1, pp. 86-119, 2021.

[6] C. Sole, et al., "The circulating transcriptome as a source of cancer liquid biopsy biomarkers," in Seminars in Cancer Biology, vol. 58, pp. 100-108, 2019.

[7] N. Kuhar, S. Sil, T. Verma, and S. Umapathy, "Challenges in application of Raman spectroscopy to biology and materials," RSC Advances, vol. 8, no. 46, pp. 25888-25908, 2018.

[8] C. Chen, et al., "Rapid diagnosis of lung cancer and glioma based on serum Raman spectroscopy combined with deep learning," Journal of Raman Spectroscopy, vol. 52, no. 11, pp. 1798-1809, 2021.

[9] S. Dhoble, T.-H. Wu, and Kenry, "Decoding nanomaterial-biosystem interactions through machine learning," Angewandte Chemie International Edition, vol. 63, no. 16, p. e202318380, 2024.

[10] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," SN Computer Science, vol. 2, no. 3, p. 160, 2021.

[11] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, "A guide to machine learning for biologists," Nature Reviews Molecular Cell Biology, vol. 23, no. 1, pp. 40-55, 2022.

[12] C. Sahli and Kenry, "Enhancing nanomaterial-based optical spectroscopic detection of cancer through machine learning," ACS Materials Letters, vol. 6, no. 10, pp. 4697-4709, 2024.

[13] R. Bro and A. K. Smilde, "Principal component analysis," Analytical Methods, vol. 6, no. 9, pp. 2812-2831, 2014.

[14] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150202, 2016.

[15] A. Singh, et al., "A review of supervised machine learning algorithms," in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1310-1315, 2016.

[16] T. Jiang, et al., "Supervised machine learning: a brief primer," Behavior Therapy, vol. 51, no. 5, pp. 675-687, 2020.

[17] R. Graf, et al., "Comparing linear discriminant analysis and supervised learning algorithms for binary classification—A method comparison study," Biometrical Journal, vol. 66, no. 1, p. 2200098, 2024.

[18] J. Nalepa and M. Kawulok, "Selecting training sets for support vector machines: a review," Artificial Intelligence Review, vol. 52, no. 2, pp. 857-900, 2019.

[19] R. Rodríguez-Pérez and J. Bajorath, "Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery," Journal of Computer-Aided Molecular Design, vol. 36, no. 5, pp. 355-362, 2022.

[20] A.-L. Boulesteix and K. Strimmer, "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data," Briefings in Bioinformatics, vol. 8, no. 1, pp. 32-44, 2007.

[21] X. Li, T. Yang, and S. Li, "Discrimination of serum Raman spectroscopy between normal and colorectal cancer using selected parameters and regression-discriminant analysis," Applied Optics, vol. 51, no. 21, pp. 5038-5043, 2012.

[22] M. Roman, et al., "Exploring subcellular responses of prostate cancer cells to X-ray exposure by Raman mapping," Scientific Reports, vol. 9, no. 1, p. 8715, 2019.

[23] A. O. Tokareva, et al., "Feature selection for OPLS discriminant analysis of cancer tissue lipidomics data," Journal of Mass Spectrometry, vol. 55, no. 1, p. e4457, 2020.

[24] A. Kleppe, et al., "Designing deep learning studies in cancer diagnostics," Nature Reviews Cancer, vol. 21, no. 3, pp. 199-211, 2021.

[25] K. A. Tran, et al., "Deep learning in cancer diagnosis, prognosis and treatment selection," Genome Medicine, vol. 13, no. 1, p. 152, 2021.

[26] R. Adam, et al., "Deep learning applications to breast cancer detection by magnetic resonance imaging: a literature review," Breast Cancer Research, vol. 25, no. 1, p. 87, 2023.

[27] L. Alzubaidi, et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," Journal of Big Data, vol. 8, no. 1, p. 53, 2021.

[28] I. H. Sarker, "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions," SN Computer Science, vol. 2, no. 6, pp. 1-20, 2021.

[29] A. Goel, A. K. Goel, and A. Kumar, "The role of artificial neural network and machine learning in utilizing spatial information," Spatial Information Research, vol. 31, no. 3, pp. 275-285, 2023.

[30] R. Yamashita, et al. "Convolutional neural networks: an overview and application in radiology," Insights into Imaging, vol. 9, no. 4, pp. 611-629, 2018.

[31] S. Abut, H. Okut, and K. J. Kallail, "Paradigm shift from Artificial Neural Networks (ANNs) to deep Convolutional Neural Networks (DCNNs) in the field of medical image processing," Expert Systems with Applications, vol. 244, p. 122983, 2024.

[32] G. Yu, et al., "Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images," Nature Communications, vol. 12, no. 1, p. 6311, 2021.

[33] R. Jiao, et al., "Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation," Computers in Biology and Medicine, vol. 169, p. 107840, 2024.

[34] G. Saren, L. Zhu, and Y. Han, "Quantitative detection of gastrointestinal tumor markers using a machine learning algorithm and multicolor quantum dot biosensor," Computational Intelligence and Neuroscience, vol. 2022, p. 9022821, 2022.

[35] S. Feng, et al., "Nasopharyngeal cancer detection based on blood plasma surface-enhanced Raman spectroscopy and multivariate analysis," Biosensors and Bioelectronics, vol. 25, no. 11, pp. 2414-2419, 2010.

[36] D. Lin, et al., "Colorectal cancer detection by gold nanoparticle-based surface-enhanced Raman spectroscopy of blood serum and statistical analysis," Optics Express, vol. 19, no. 14, pp. 13565-13577, 2011.

[37] S. Lee, M. Jue, K. Lee, B. Paulson, J. Oh, M. Cho, and J. K. Kim, "Early-stage diagnosis of bladder cancer using surface-enhanced Raman spectroscopy combined with machine learning algorithms in a rat model," Biosensors and Bioelectronics, vol. 246, p. 115915, 2024.

[38] J. Wang, L. Cong, W. Shi, W. Xu, and S. Xu, "Single-cell analysis and classification according to multiplexed proteins via microdroplet-based self-driven magnetic surface-enhanced Raman spectroscopy platforms assisted with machine learning algorithms," Analytical Chemistry, vol. 95, no. 29, pp. 11019-11027, 2023.

[39] S. Dong, et al., "Early cancer detection by serum biomolecular fingerprinting spectroscopy with machine learning," ELight, vol. 3, no. 1, p. 17, 2023.

[40] K. Liu, et al., "Label-free surface-enhanced Raman spectroscopy of serum based on multivariate statistical analysis for the diagnosis and staging of lung adenocarcinoma," Vibrational Spectroscopy, vol. 100, pp. 177-184, 2019.

[41] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[42] R. Luo, J. Popp, and T. Bocklitz, "Deep learning for Raman spectroscopy: A review," Analytica, vol. 3, no. 3, pp. 287-301, 2022.

[43] M. Kim, et al., "Detection of ovarian cancer via the spectral fingerprinting of quantum-defect-modified carbon nanotubes in serum by machine learning," Nature Biomedical Engineering, vol. 6, no. 3, pp. 267-275, 2022.

[44] X. Shao, et al., "Deep convolutional neural networks combine Raman spectral signature of serum for prostate cancer bone metastases screening," Nanomedicine: Nanotechnology, Biology and Medicine, vol. 29, p. 102245, 2020.

[45] N. Cheng, et al., "An antibody-free liver cancer screening approach based on nanoplasmonics biosensing chips via spectrum-based deep learning," NanoImpact, vol. 21, p. 100296, 2021.

[46] J. Q. Li, et al., "Machine learning using convolutional neural networks for SERS analysis of biomarkers in medical diagnostics," Journal of Raman Spectroscopy, vol. 53, no. 12, pp. 2044-2057, 2022.

[47] H. Shin,et al., "Early-stage lung cancer diagnosis by deep learning-based spectroscopic analysis of circulating exosomes," ACS Nano, vol. 14, no. 5, pp. 5435-5444, 2020.

[48] X. Bi, et al., "SERSomes for metabolic phenotyping and prostate cancer diagnosis," Cell Reports Medicine, vol. 5, no. 6, 2024.

[49] Y. Jin, et al., "Fluorescence analysis of circulating exosomes for breast cancer diagnosis using a sensor array and deep learning," ACS Sensors, vol. 7, no. 5, pp. 1524-1532, 2022.