Exploration of Dimensionality Reduction Techniques for Improving Cancer Classification Accuracy: A Machine Learning Approach to Analyzing Single Extracellular Vesicles via Surface-Enhanced Raman Spectroscopy

Yao Lu

Department of Bioengineering

McGill University

Montréal, Québec, Canada

October 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Engineering

Abstract - English

The complex biochemical composition and nanoscale nature of extracellular vesicles (EVs) pose significant challenges in developing efficient and uniform analytical techniques. Surface-enhanced Raman Spectroscopy (SERS) offers a promising solution by providing detailed molecular fingerprints of vesicles in a label-free manner. The MoSERS platform, developed in Mahshid lab, enables the probing of EVs at the single-vesicle level with reduced sample processing and volume requirements. This capability allows for the specific examination of tumor-derived EVs carrying disease biomarkers, amongst a diverse background of biovesicles. However, standardized protocols for decoding SERS spectra are still lacking due to the heterogeneous nature of EVs, resulting in overlapping peaks with thousands of features per spectrum. Machine learning (ML) presents a potential solution, as its statistical models can capture underlying patterns within complex datasets. This study employed a ResNet-based Convolutional Neural Network (CNN) for its natural feature extraction capabilities via convolutional layers. The study was designed to account for the reduced representation of patient samples using a larger number of healthy samples (n = 16) compared to two different types of brain tumor patients (n = 8, each). A SERS data library of single EV spectra was collected in the lab. The feasibility of increasing binary classification accuracy between the two populations through direct adjustments of the sample train/test distribution was explored. An optimal ratio between the healthy and patient training sets was determined for the two cancer paradigms, respectively. Four dimensionality reduction techniques—spectra resampling, bio-inactive region removal, unsupervised Autoencoder (AE) learning, and Gradient-weighted Class Activation Mapping (Grad-CAM)—were investigated to further improve classification accuracy. Ultimately, the resampling approach yielded a significant accuracy improvement of ~10% compared to the baseline for both cancer models. This supports the hypothesis that the height and relative position of SERS peaks are critical in guiding the model's predictions. Additionally, the proposed resampling method proved effective across two cancer models and follows a standardized data treatment protocol, suggesting its potential applicability in other SERS-based EV studies for facilitating cancer diagnosis.

Keywords: Deep Learning, Surface-enhanced Raman Spectroscopy, Dimensionality Reduction, Extracellular Vesicles

Abstract - French

La composition biochimique complexe et la nature nanométrique des vésicules extracellulaires (EVs) posent des défis importants pour le développement de techniques analytiques efficaces et uniformes. La spectroscopie Raman exaltée de surface (SERS) offre une solution prometteuse en fournissant des empreintes moléculaires détaillées des vésicules de manière sans marquage. La plateforme MoSERS, développée dans le laboratoire Mahshid, permet l'exploration des EVs au niveau de la vésicule unique avec une réduction du traitement des échantillons et des volumes requis. Cette capacité permet un examen spécifique des EVs dérivées de tumeurs portant des biomarqueurs de maladies, au sein d'un arrière-plan diversifié de biovésicules. Cependant, les protocoles standardisés pour l'interprétation des spectres SERS font encore défaut en raison de la nature hétérogène des EVs, entraînant des pics qui se chevauchent avec des milliers de caractéristiques par spectre. L'apprentissage automatique (ML) présente une solution potentielle, car ses modèles statistiques peuvent capturer les motifs sous-jacents dans des ensembles de données complexes. Cette étude a utilisé un réseau de neurones convolutifs (CNN) basé sur ResNet pour ses capacités naturelles d'extraction de caractéristiques via des couches convolutives. L'étude a été conçue pour tenir compte de la représentation réduite des échantillons de patients en utilisant un plus grand nombre d'échantillons sains (n = 16) par rapport à deux types différents de patients atteints de tumeurs cérébrales (n = 8 chacun). Une bibliothèque de données SERS des spectres d'EV uniques a été collectée en laboratoire. La faisabilité d'augmenter la précision de la classification binaire entre les deux populations par des ajustements directs de la distribution des échantillons d'entraînement/test a été explorée. Un ratio optimal entre les ensembles d'entraînement sains et patients a été déterminé pour les deux paradigmes du cancer, respectivement. Quatre techniques de réduction de dimensionnalité — rééchantillonnage des spectres, suppression de la région bio-inactive, apprentissage non supervisé par autoencodeur (AE) et cartographie d'activation par classe pondérée par gradient (Grad-CAM) — ont été étudiées pour améliorer davantage la précision de la classification. Finalement, l'approche de rééchantillonnage

a conduit à une amélioration significative de la précision d'environ 10 % par rapport à la ligne de base pour les deux modèles de cancer, soutenant l'hypothèse selon laquelle la hauteur et la position relative des pics SERS sont essentielles pour guider les prédictions du modèle. De plus, la méthode de rééchantillonnage proposée s'est avérée efficace dans deux modèles de cancer et suit un protocole standardisé de traitement des données, suggérant son applicabilité potentielle dans d'autres études sur les EVs basées sur SERS pour faciliter le diagnostic du cancer.

Mots-clés : Apprentissage profond, Spectroscopie Raman exaltée de surface, Réduction de dimensionnalité, Vésicules extracellulaires

Table of Contents

Contribution of Authors

Section 3 of this thesis is adapted from the manuscript "Optimizing Binary Classification of Heterogeneous Raman Spectra: A Comparison of Dimensionality Reduction Techniques on Cancer Paradigm" submitted to Scientific Reports as it is presented in the thesis, except formatting. It is now under review. Yao Lu is the first author of this manuscript. In Section 3, Sara Mahshid, Masha Jalali and Carolina Del Real Mata provided the original idea. Patient samples were obtained by Janusz Rak, Laura Montermini, Marjan Khatami and Livia Garzia. Spectra collection was done by Carolina Del Real Mata and Yao Lu using a Raman microscope from the Siaj lab at the Université du Québec à Montréal (UQAM). Yao Lu performed the data analysis, and Carolina Del Real Mata and Yao Lu contributed to the interpretation of the results. Sara Mahshid supervised the project and contributed to funding acquisition.

Acknowledgement

It has been quite a journey, and it would not have been possible without the great support from the people I met along the way. I would like to first express my gratitude to my supervisor, Dr. Sara Mahshid, for her invaluable guidance and support throughout my research. Her expertise and insightful feedback have been instrumental in constantly providing direction during the project and in the completion of this thesis. I am also immensely grateful to all the members of the MoSERS research group; I am very lucky and honored to be a part of this team.

Special thanks to Mahsa for being the gentlest and kindest person I know and for generously sharing her genius inventions and brilliant ideas with me. I am always amazed by her innovative inspirations, dedication to research, and willingness to answer my endless questions. I also want to give a special thank to Carolina, who has been my role model at work and is always filled with positivity, helping me look at the better side of life. I could not have completed this project without your constant mental support and the numerous little meetings we had. Thank you for guiding me through every step of the project and for always being there to talk. The tenacity and enthusiasm you showed when facing seemingly impossible challenges will forever motivate me. I would also like to thank all the members of the Mahshid Lab for creating a friendly and

warm work environment. As I quote Tamer here: “We are more like friends and family instead of a lab.”

I would like to express my deepest gratitude to my parents and my little cat sister, Moon. The past two years have been quite challenging, and I have encountered many unexpected obstacles not just from the research side. Their patience and love have guided me through. I would also like to thank my friends for the numerous homemade dinners and quality time together on Steam. Lastly, I would like to thank myself for pulling it together and going through this journey, from the beginning to the end.

List of abbreviations:

AI: Artificial Intelligence

ac-EHD: Alternating Current Electrohydrodynamic

AE: Autoencoder

ANN: Artificial Neural Network

ATR: Attenuated Total Reflection

AUC: Area Under the Curve

BIC: Bound States in the Continuum

CCA: Cholangiocarcinoma

CM : Conditioned Medium

CNN: Convolutional Neural Network

DGC: Differentiated Glioblastoma Cells

DiO: Dihexadecyloxacarbocyanine

FDTD: Difference Time-Domain

ELISA: Enzyme-linked Immunoassay

EVs: Extracellular vesicles

GBM: Glioblastoma

GSC: Glioblastoma Stem Cell

HER2: Human Epidermal Growth Factor Receptor 2

HRP: Horseradish Peroxidase

ISEV: International Society for Extracellular Vesicles

KNN: K-nearest neighbour

LDA: Linear discriminant analysis

LOD: Limit of Detection

LSPR: Localized Surface Plasmon Resonance

ML: Machine Learning

MIL: Multiple Instance Learning

MLP: Multi-Layer Perceptron

NPs : Nanoparticles

PBMC: Peripheral Blood Mononuclear Cells

PCTE: Polycarbonate Track-Etched

PSA: Prostate-Specific Antigen

PDMS: Polydimethylsiloxane

PCA: Principal Component Analysis

PC: Principal Component

POC: Point-of-Care

RF: Random Forest

SPR: Surface Plasmon Resonance

SERS: Surface enhanced Raman Spectroscopy

SVM: Support Vector Machine

TMB: Tetramethylbenzidine

TIR: Total Internal Reflection

TOO: Tissue of Origin

VAE: Variational Autoencoder

List of Figures

1 Introduction

The integration of artificial intelligence (AI) and machine learning (ML) into biosensors has revolutionized the field of medical data analysis. By enabling the rapid and accurate processing of vast and complex datasets, these technologies significantly reduce the time and effort traditionally required for diagnosis. Machine learning algorithms, with their ability to learn and adapt to new data in an automatic manner with little-to-no human intervention, can identify subtle patterns and correlations that might be overlooked by human analysts. This capability is particularly valuable in detecting early indicators of disease, predicting patient outcomes, and tailoring treatment plans to individual patients. By leveraging these technologies, biosensors can now provide deeper insights and more robust analyses, enhancing their utility and effectiveness in both research and clinical applications.

This thesis aims to explore the use of ML and dimensionality reduction techniques in analyzing spectral data for single extracellular vesicles (EVs) acquired using surface-enhanced Raman spectroscopy on the patented MoSERS nanostructure platform [1]. Specifically, a ResNet-based convolutional neural network (CNN) will serve as the base algorithm, supplemented by customized dimensionality reduction modules. CNN-based deep learning is widely applied in spectral analysis due to its ability to handle complex datasets and uncover subtle connections among individual peaks. However, with the surge in the number of features per spectrum—approaching 3,000—the CNN model faces challenges from the "curse of dimensionality." These challenges include increased computational resource demands, insufficient generalization across populations, and potential performance degradation in high-dimensional spaces. Therefore, this thesis will investigate how dimensionality reduction techniques can mitigate these issues, improving model performance with limited data and resources.

This thesis begins with a comprehensive literature review on recent developments in commonly employed optical biosensors and the integration of ML in spectral data analysis for biosensors (Section 2). Next, the body of the thesis aims to strengthen the customizing process of ML for clinical applications (Section 3). To simulate a more realistic clinical setup and to better tailor the model for potential future integration into diagnostic workflows, the analysis will first

address data imbalances where healthy samples greatly outnumber patient samples. Techniques to overcome this challenge include optimizing the distribution of training and testing sets and employing the "leave-one-out" cross-validation approach. Following this, four dimensionality reduction methods—spectrum resampling, bio-inactive region removal, Autoencoder (AE), and Gradient-weighted Class Activation Mapping (Grad-CAM)—will be implemented and compared. Their effectiveness will be evaluated based on the improvement in binary classification accuracy between the healthy and diseased groups relative to the baseline. Lastly, a comprehensive discussion of findings and potential future pathways are included in Section 4.

2 Review of Relevant Literature: Optical Sensors and Machine Learning Assisted spectral data processing and analysis for Extracellular Vesicle Characterization

2.1 Introduction

Intercellular communication via the release of membrane-bound particles, also known as extracellular vesicles (EVs), is a relatively new discovery that has revolutionized the field of cell biology [2]. EVs are a heterogenous family of vesicles derived from endosome or plasma membranes and can be found in all biofluids, including blood, saliva, urine, and cerebrospinal fluid. EV is a general name for these vesicles, but based on their structure and biogenesis, EVs can be further categorised into three main types following a decreasing order in average size: apoptotic bodies (50-5000 nm), microvesicles (100-1000 nm), and exosomes (30-100 nm). Detailed information on EV subtypes has been discussed extensively in previous reviews [3], [4], [5], [6], [7]. Briefly, apoptotic bodies are released by dying cells and are formed through a separation of the plasma membrane from the cytoskeleton. In contrast, microvesicles are formed by direct outward budding of the plasma membrane. Exosomes are one of the most studied EV subtypes and typically originate from the inward budding of the membrane of early endosomes. Since exosomes and microvesicles cannot be distinguished by size alone and their overlapping protein densities further complicate the stratification, the general term EVs will be used to describe these populations unless their cells of origin can be defined [3], [8], [9]. The composition of EVs typically includes lipids, proteins, nucleic acids, and biomarkers of certain diseases depending on

their designated pathways. Regardless of the cells of origin, a set of proteins known as “marker proteins” involved in the EV formation process can always be expected to be found and have been used extensively for the efficient targeting of EVs. Furthermore, EVs can effectively shield their cargo from enzymatic degradation by the extracellular environment. Therefore, the molecular cargo carried by EVs may be used as an indicator of the pathological state of parental cells. This prompts their exploration as potential biomarkers for the early detection and monitoring of cancer [10], [11]. The uptake of EVs by the recipient cell can trigger further intercellular signalling, potentially leading to modification of the physiological state or microenvironment of recipient cells [6], [7]. Numerous studies have suggested that EVs play an important role in disease progression, including cancer metastasis. Moreover, EVs also hold therapeutic potential in controlled drug delivery [12], [13]. Thus, further investigation on both the morphology and biochemical compositions of EVs is imperative for advanced clinical applications.

Based on the extensive morphological heterogeneity characteristic of EVs, no single detection technique to date can capture the full-size range of EVs, let alone a single uniform protocol across different studies[9]. A 2016 international survey showed vast differences in techniques used for both isolation and characterization of EVs, where the choice of method may impact the amount, type, and purity of the recovered EVs [14], [15]. In response to these emerging challenges, numerous biosensors such as electrochemical, photoelectrochemical, and optical sensors have been used for studying EVs. In particular, optical sensors offer the advantage of real-time measurement with high sensitivity, stability, and resistance to background noise [12], [13], [16]. The combination with customized nanostructure and machine learning (ML) assisted mass data analysis enables highly sensitive and cost-effective diagnosis based on EVs.

2.2 Optical sensors for EV identification and characterization

Optical sensors can be divided into two subcategories: labelled and label-free. The former requires additional molecular tags such as fluorescent dyes or lipophilic tracer dyes and can be quite effective for the detection of target molecules with low abundance in the sample. However, due to the vast heterogeneity and limited understanding of EVs’ surface chemistry, it can be challenging to find an appropriate label. Moreover, the aggregation of labels due to non-EV

particles may lead to contradictory results. In contrast, label-free optical sensors allow for non-invasive EV probing with greatly reduced wet lab processes and a minimum effect from false-positive results [17], [18], [19].

With the numerous modern techniques developed for EV detection and characterization, a single technique is still not capable of capturing the full landscape of the complete spectrum of EV properties [14]. The rather recent emergence of optical methods for EV detection may shed light onto new pathways for identifying EV surface receptors and membrane proteins, in addition to their internal content with the potential of probing at single-vesicle level. The following section describes the general working principle of two commonly employed label-free EV characterization techniques: Surface Plasmon Resonance (SPR) and Surface-Enhanced Raman Spectroscopy (SERS). Several examples of their implementations in EV studies are provided to demonstrate their wide applications.

2.2.1 Surface Plasmon Resonance (SPR)

Surface plasmon resonance (SPR) occurs when an incident electromagnetic wave (i.e light) propagates from a medium with a relatively higher refractive index (RI) to one with a lower RI. In a commercial configuration, a Kretschmann geometry setup is usually employed using a high-reflective index glass prism with the attenuated total reflection (ATR) method and a thin metal surface with a relatively low RI [20], [21]. Reflection, especially total internal reflection (TIR), tends to occur at the interface of the two media, rather than refraction. At a certain incident angle, a portion of the energy from the fully reflected light will be transferred into the medium with the lower RI and increases the intensity of the local electric field. The resulting collective oscillation of the electrons that propagate along the surface is known as the surface plasmon. The defined SPR angle where resonance occurs is dependent on the refractive index of the medium near the metal surface. Consequently, any small changes in the sensing medium, such as biomolecular binding interactions between an immobilised ligand and analyte, can lead to a shift in the intensity of incident light, which is calculated as a function of the incident angle [22], [23]. Conventional SPR platforms are used to perform bulk measurements and typically have a limit of detection

(LOD) on the order of 10^6–10^7 EVs/mL. Metallic nanostructures or nanoparticles can also enhance SPR signals to reach a LOD as low as 5×10^3 EVs/mL [24].

Localised surface plasmon resonance (LSPR) refers to the special case where the light wave is trapped within conductive nanoparticles (NPs) that are smaller than the wavelength of incident light. This situation leads to coherent localised plasmon oscillations and can be fine-tuned based on the size, structure, and interparticle separation of NPs. For example, Ag NPs can be electrodeposited over PDMS for increased EV detection sensitivity [25]. However, even minimal variations in nanostructure dimensions could potentially affect the optical resonance conditions and thus the sensitivity and reproducibility of the assay outcome. The increase in chip costs due to the fabrication of intricate nanostructures should also be considered, especially during the translation into clinical studies [9], [23].

One of the most renowned EV-SPR studies was carried out by the Im group termed the nano-plasmonic exosome (nPLEX) sensor, where they proposed sensor included periodic nanohole arrays patterned over a gold film [26]. Building upon their previous studies on the nPLEX platform, the same group has further advanced their technology to achieve single-EV resolution utilizing plasmon-enhanced fluorescence detection combined with periodic gold nanowell structures, referred to as FLEX (**Figure 2.1 (A)**) [27]. With their established protocol, the gold wafers were batch fabricated and the optimal diameter for each gold nanowell was determined to be 200 nm for additional long-range surface plasmon resonance enhancement. The resonance wavelengths can be conveniently tuned by adjusting the periodicity of the arrays. When compared to a flat gold surface, FLEX demonstrated a 6-fold increase in the number of sEVs detected from a cholangiocarcinoma (CCA) cell line. Such signal amplification not only facilitated precise EV subpopulation identification but also mitigated potential biases from the analysis of aggregated EVs with more intense signals. Clinical studies were conducted using bile samples from both CCA patients (n = 17) and benign cases (n = 8). A distinct differential pattern from marker-positive tumor-derived EVs was observed between the two groups using three pre-determined biomarkers (EpCAM, MUC1, and EGFR). Another study that utilized similar concepts was produced by Hu's group, where they also applied an antibody microarray printed on gold films for exosome capture and detection in combination with SPR, shown in **Figure 2.1 (B)** [22]. However, the microfluidic

compartment is rather simple and only serves as a sample delivery unit to the SPR platform. Regardless of the setup, this assay was able to identify unpurified EVs from cell culture supernatant (CCS) without enrichment. Moreover, the assay showed a clear distinction between two human hepatocellular carcinoma cell lines with different metastatic potentials based on the abundance of surface proteins and lipid mass.

It is noteworthy that in the SPR biosensors discussed above, the microfluidic components are primarily comprised of a single chamber connected to a straight microchannel, where the sole function of this structure is to deliver pre-processed samples to the sensing platform. While these designs serve the fundamental purpose of delivering samples to the sensor and may increase the multiplicity of the device by incorporating additional channels, they contribute relatively little value to the overall setup. The integration of optical sensors with microfluidic sample processing units is still a largely unexplored area in the field of EV biosensors.

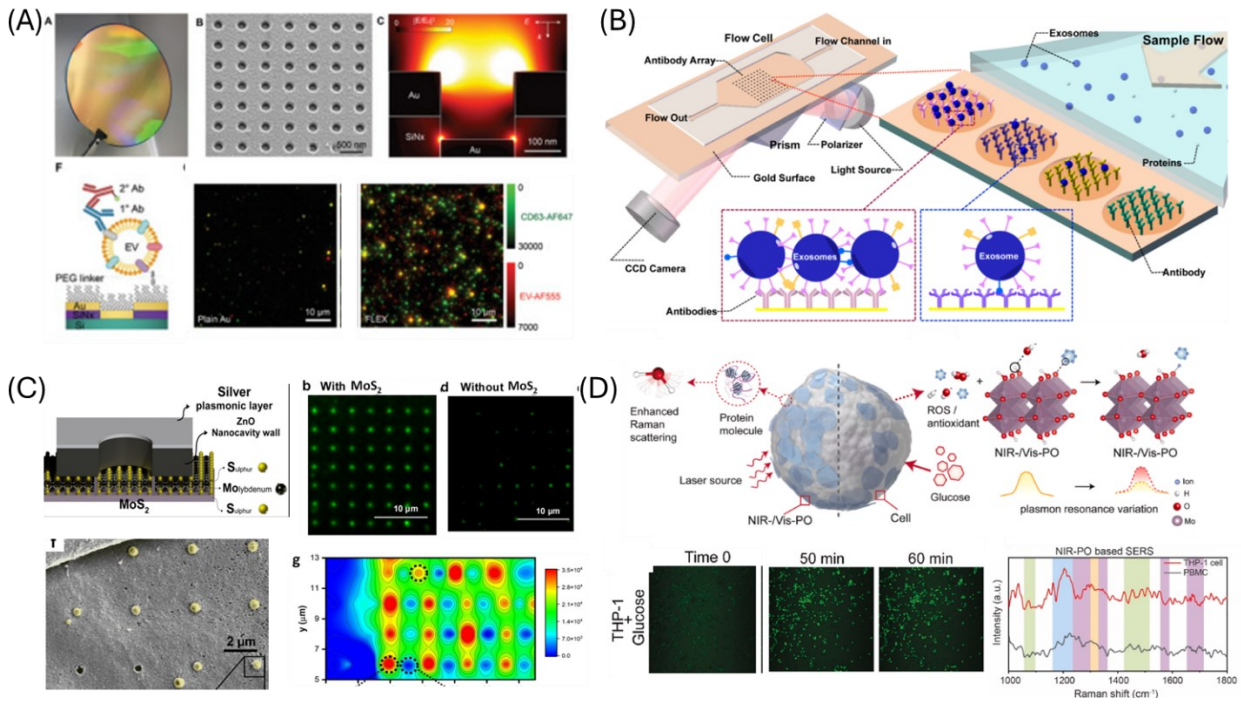

Figure 2.1 Optical platforms for EV profiling and characterization with SPR and SERS.
(A) Fluorescence-amplified extracellular vesicle sensing technology (FLEX) can capture immunolabelled EVs on a plasmonic gold surface. The finite-difference time-domain (FDTD) simulation shows the enhanced electromagnetic fields near the embedded nanowell structure, which leads to amplified fluorescence signals. EpCAM, MUC1, and EGFR were used as Cholangiocarcinoma (CCA) markers. Reproduced with permission from **[27]**. Copyright 2023, Advanced Science. **(B)** Schematic view of SPR imaging platform combined with specific antibody microarrays for the detection of EVs in cell culture supernatant. The antibodies are immobilized over an ultrathin gold film and the optical path is preset at a fixed angle of incidence for capturing changes in the refractive index upon binding. Reproduced with permission from **[28]**. Copyright 2014, Analytical Chemistry **(C)** MoS2-Plasmonic Nanocavities (MoSERS) platform for label-free single EV SERS profiling. The microchip can achieve up to 97% single EV confinement using less than 10µL sample as a combined effect of embedded MoS2 monolayer and optimized nanocavity dimension. SEM and fluorescent micrographs show EV loaded individually into cavities. Combined with machine learning algorithms. Reproduced with permission from **[10]**. Copyright 2023, American Chemical Society. **(D)** Schematic diagram of the proposed PO nanoprobe biosensor, where the cells' production of reactive oxygen species (ROS) and antioxidants under programmed apoptosis leads to shift in surface plasma resonance. Reproduced with permission from **[29]**. Copyright 2022, Biosensors and Bioelectronics.

2.2.2 Surface enhanced Raman Spectroscopy (SERS)

First discovered by C.V. Raman in 1928, Raman scattering only accounts for a very small fraction of scattered light and occurs when an excitation leads to scattered photons with a different frequency compared to that of the incident photon [30], [31]. This inelastic scattering of laser light

leads to the vibration of chemical bonds in the incident medium, and some photons are scattered with a particular shift in energy levels as a function of both the structural and chemical characteristics of the sample. The frequency can be subsequently recorded and processed into the form of a Raman spectrum, which can be regarded as biochemical "fingerprints" that correspond to the molecular-level composition of the analyte. This technique has been used to characterize the distribution of various cellular components with high spatial resolution [5], [9]. As mentioned previously, the molecular content of EVs may serve as potential biomarkers for liquid biopsy. Thus, the non-destructive and label-free nature of Raman spectroscopy makes it an ideal technique for studying the subtle differences in EV membrane properties and molecular content. Although it is not as commonly used in comparison to previously discussed techniques, Raman spectroscopy has been adapted for characterizing both the soluble and the vesicular components of cell secretions from conditioned medium (CM) [32]. This characterization may provide more insight into the soluble factors that synergistically cooperate with EVs as part of the regenerative effect of CM.

As approximately 1 in 10^7 incident photons will undergo inelastic scattering, Raman spectroscopy is inherently limited by extremely low signal intensity. Consequently, this limitation needs to be balanced with either increased laser power, prolonged scanning duration, or higher sample concentration. Therefore, surface-enhanced Raman spectroscopy (SERS) was developed in 1974, whereby the Raman signal can be enhanced dramatically up to 1014 times [33], [34]. This is achieved by placing the sample molecules between the gaps of certain nanomaterials (also known as "hot spots"). The increase in signal intensity stems from the SPR phenomenon mentioned previously, which is the oscillation of electrons at the vicinity of a metal or semiconductor structure that leads to an enhanced electric field upon interaction with the incident electromagnetic waves. Thus, SERS not only retains the advantages of Raman spectroscopy in that it maintains the vibrational modes of molecules, but also overcomes its aforementioned limitations [35]. This improvement is particularly important for studying EVs as the populations of disease related EVs are scarce compared to healthy ones. When combined with customised platforms or enhancement probes, SERS has the potential to perform EV profiling with single-molecule level resolution and is thus one of the most extensively used techniques in the field of EVs [36].

A recent publication from the Mahshid lab featured a label-free sEV plasmonic assay in combination with SERS, named the MoSERS platform (**Figure 2.1 (C)**) [37]. The platform had a novel structure consisting of a plasmonic silver/ ZnO bilayer with an embedded single-layer MoS_2 floor and plasmonic cavity arrays at the surface. The MoS_2 monolayer provided natural physical attraction forces to the lipid bilayer, specifically Coulomb and van der Waals forces, facilitating the nanoconfinement of EVs without any biological recognition elements. The EV trapping ability of the MoS_2layer was further confirmed by examining fluorescently labelled EVs under TEM, where the fluorescent intensity (i.e. number of captured EVs) was twice the intensity from the nanocavities without MoS_2. The photonic cavities had an optimized diameter at 250 nm to host precisely one single EV, providing sufficient electromagnetic field enhancement for obtaining sEV-resolution SERS spectra. The platform's sEV profiling capability was validated using Glioblastoma (GBM) cell lines, where MoSERS successfully identified the presence of EGFRvIII oncogenic mutation with the detection limit being as low as 1.23%. Furthermore, when interfaced with a convolutional neural network (CNN) ML algorithm, MoSERS achieved an 87% diagnostic accuracy using plasma samples with complex backgrounds from GBM patients (n=12) and healthy controls (n=8), successfully detecting GBM signature mutations in all patients.

Most of the SERS platforms made with noble metals are not tunable due to prefixed morphology, which may lead to low compatibility with ambient refractive index (RI) and other complementary detection techniques such as colorimetry. Meanwhile, certain semiconductors, such as TiO_2 and ZnO, demonstrated significantly improved biocompatibility. However, they are also limited by insufficient electron transfer due to a lack of free ions and require a strong electromagnetic field during operation. The increased laser intensity may also damage the structural integrity of samples [38], [39]. This may be overcome by the development of novel Raman probes. For example, a novel Raman probe was developed using ultra-thin degenerate molybdenum oxide doped with H+, a type of "plasmonic oxide" (PO) material [40] (**Figure 2.1 (D)**). The free electron concentration can be adjusted based on the level of H+ doping, and the subsequent reaction with redox products from specific cancer EVs can lead to an observable variation in spectra intensity. The proposed "nanoflake" structure exhibited a homogeneous increase in plasma resonance at the edge. This system was combined with a serpentine-shaped

microfluidic chip with engraved grooves to hydrodynamically separate individual molecules from the suspension and minimise possible interference from the liquid background. The device successfully differentiated between the THP-1 and HEK-293 cell lines based solely on their SERS patterns with an accuracy of over 90%. However, more advanced studies with clinical samples and rapid on-chip cell immobilization have yet to be conducted.

2.2.3 Other Optical-based Biosensors

Apart from the more commonly studied optical biosensors, nanophotonic resonators follow a similar working principle to SPR but operate using refractometric sensing, where the interactions between the analyte and surface-immobilized capture molecules lead to a change in the dielectric properties of the resonators. This change can be translated into a shift in resonance wavelength and used for the quantification of biological interactions. Jahani et al. addressed the possibility of tracking the red shift of the resonance spectrum upon target binding over a narrow window instead of the entire bandwidth, which eliminated the need for cumbersome and expensive spectrometers (**Figure 2.2 (A)**) [41]. They also employed diatomic metasurfaces with quasi-bound states in the continuum (BICs), where the light wave remains completely localised in the vicinity of the surface. When combined with a customized image processing system, their device detected on average 0.41 nanoparticles/μm^2. Real-time quantification of ovarian cancer-related EV binding was performed with an LOD as low as 1.23×10^8 particles/mL. This study demonstrated a new path for label-free biomolecular studies but focused on the mechanism and setup of the assay, while the detection of EVs was used as a proof-of-concept.

Based on a similar principle of detecting changes in RI, Wang and his group proposed a label-free photonic crystal (PC) biosensor for EV detection shown in **Figure 2.2 (B)** [42]. This biosensor consists of a narrowband optical reflector that reflects a particular wavelength upon excitation. Moreover, the PC surface is patterned with one-dimensional subwavelength gratings, which facilitates phase matchings between the excitation light and the PC resonances and leads to narrowband reflection. The PC detector was integrated with a four-channel microfluidic chip to increase the throughput, where each channel was designed for the host, parasitic, positive, and negative reference samples, respectively. Compared to the conventional SPR devices discussed

previously, PC offers the unique advantage of reduced fabrication cost with comparable sensitivity and does not have the need for signal enhancers due to the naturally narrow line width of the PC resonance. However, current PC studies remain at a preliminary stage since only one type of EV surface protein was used. As a result, future investigations are needed to fully validate its application in clinical settings.

On the other hand, a unique method of sEV entrapment, termed geometry-induced electrohydrodynamic tweezer (GET), was achieved by generating an electrohydrodynamic potential using a finite circular array of gold nanoholes with a void region in the middle [43]. Shown in **Figure 2.2 (C),** the applied opposing alternating current (a.c.) electro-osmotic flows can form a stagnation zone at the central void region with minimum electrohydrodynamic potential. The interaction between the EV's charged double lipid layer and its image charge in the conduction plane contributes to the localization of particles in a parallel manner. Single vesicle entrapment was ensured by finding the optimal A.C. frequency at 3.5 kHz using a void region of 4 μm diameter, where the dipole-dipole repulsion force overcomes the drag force generated by the electro-osmotic flow. In addition, the periodicity of the array induces surface plasmon waves, resulting in a significant enhancement in the fluorescence emission from the single fluorescently labelled EVs. This setup also allows for the facile overlay of both electrohydrodynamic potential and plasmon-enhanced optical trapping potential through the addition of a cavity at the center region. The photothermal heating effect can be effectively mitigated via a sapphire substrate that serves as a heat sink to avoid the burning of samples.

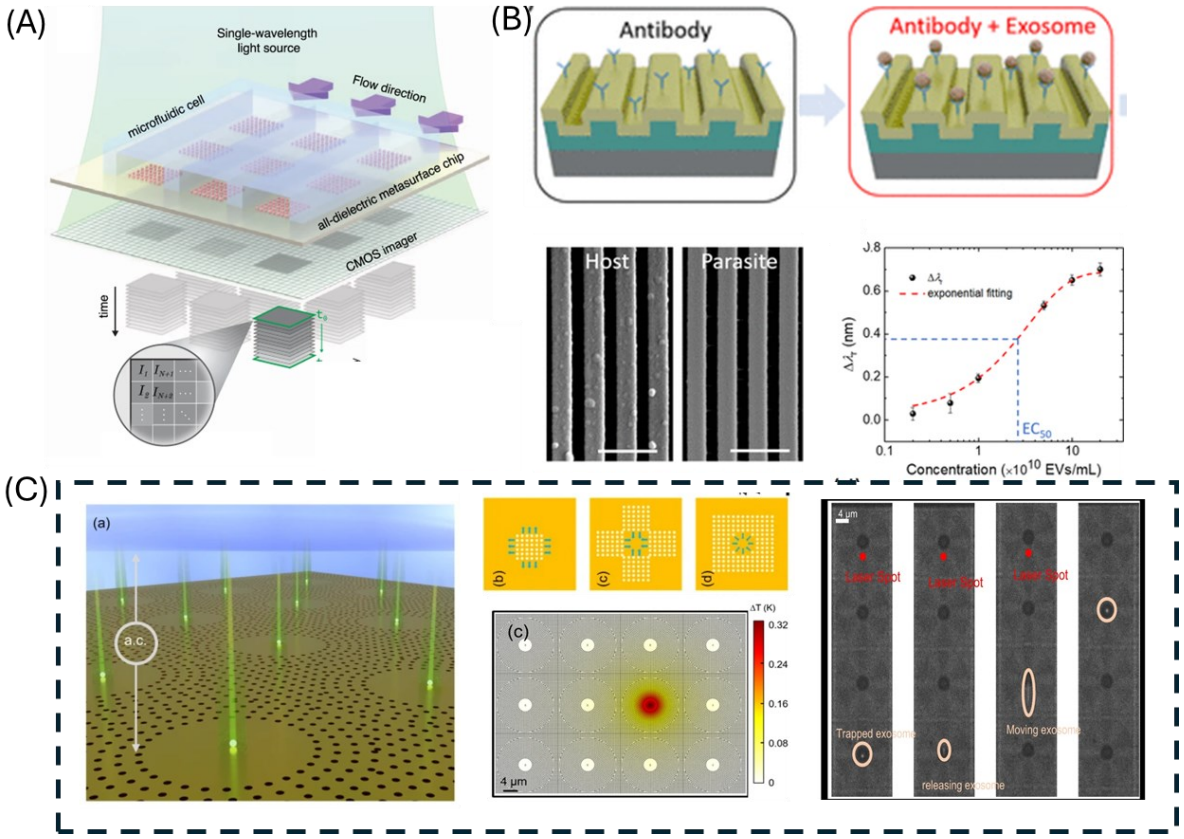

Figure 2.2: Other optical platforms.

(A) Schematic sketch of a real-time in-flow imaging platform with 2D dielectric sensor microarray, where the imaging units are illuminated with single-wavelength light source. A red shift in the transmission spectrum indicates biomarker binding, where the change in intensity can approximated as a linear function of the spectrum shift with a constant of α, which is the slope of the transmission spectrum in the linear region. Reproduced with permission from [41]. Copyright 2021, Nature Communications. **(B)** Schematic diagram of EV isolation from parasite samples and subsequent capturing by the immobilized antibodies, which is again reflected by the shift in collected spectrum. The four channels in the microfluidic unit are designed for the host, parasitic, positive, and negative reference samples, respectively. Reproduced with permission from [42]. Copyright 2018, ACS Sensors. **(C)** Geometry-induced electrohydrodynamic tweezers (GET) utilize electrodynamic potentials for trapping single nanosized EVs. SEM image shows specifically arranged circular arrays of plasmonic nanoholes with a central stagnant region due to the radially outward a.c. electro-osmotic flow. Maximum radiation energy flow for a dipole fluorescence emitter is also observed at the center of the circular region. Both optical trapping potential simulation and recorded trapping stability show stable confinement of a single EV at the center region. Reproduced with permission from [43]. Copyright 2023, Nature Communications.

2.3 Machine learning-assisted cancer biopsy and diagnosis using EV spectral data

With the rapid advancement of EV-characterizing biosensors, the increasing volume and complexity of generated data have posed significant challenges in translating raw data into human-

interpretable information. Meanwhile, the recent boom in artificial intelligence (AI) and machine learning (ML) has provided a convenient method for large-scale data analysis. These technologies can be seamlessly integrated into biosensors to form the new generation of "smart" sensors [44], [45]. ML algorithms can be roughly classified into two main categories: supervised and unsupervised. The former assigns a set of pre-defined labels to the raw dataset and is more commonly used for categorical classification. The latter leverages the machine to draw meaningful and useful interpretations from the data, and the problems can be either classification or regression-based [46]. Regardless of the type, machine learning can be generally summarized as a data-driven pattern recognition approach. Using statistical models designed to capture subtle associations between input data points (features), the algorithm can identify underlying patterns within the population and successfully predict outputs, such as distinguishing between patient and healthy samples or functioning as a discriminative model [47], [48]. Moreover, when facing large amounts of low-resolution data with noisy background, ML enhances the possibility of finding reasonable analytical results and even discover hidden relations between sample parameter with a precise in-built mathematical framework [48]. Given its wide adaptability, natural cost-effectiveness and great freedom in portability, ML has been applied for the interpretation of numerous forms of input data, such as spectral data [49], [50], fluorescent images [51], [52], electrochemical readings [53], and more.

2.3.1 Different Machine Learning Models

Support Vector Machines (SVM) is one of the most seen supervised machine learning algorithms primarily used for classification tasks. Originally proposed in 1963 by Vapnik, the fundamental principle behind SVM is to find the optimal hyperplane that best separates the data into distinct classes [54]. When dealing with relatively simple data that can be linearly separated, the hyperplane is used to maximize the margin distance between the hyperplane and the closest data points from the data with classes assigned, known as support vectors. SVM is also capable of generating nonlinear decision boundaries. By mapping non-linear data into a higher-dimensional space, this implicit transformation enables the SVM to find a linear hyperplane in the higher-

dimensional space using only dot products using preset kernels. Common kernels include linear, polynomial, and radial basis function (RBF) etc. [55], [56], [57]

As the name suggests, linear discriminant analysis (LDA) finds linear combinations of the original features to create new axes for the data projection. Being a derivative of Fisher's discriminant analysis, the core principle of LDA is to project high-dimensional data onto a lower-dimensional space [58]. By maintaining key high dimensional features, the correspondingly projected feature vectors of the assigned classes onto a lower dimensional space can be used for effective stratification. Thus, it can be used for both classification and dimensionality reduction [59]. The two main criteria that LDA follows is to maximize the distance between the means of two classes while minimizing the variance within individual classes. Thus, by computing both in-class and between-class scatter matrices and solving the subsequent eigenvalue problem, an orientation vector (eigenvectors) can be used as the linear discriminants and serve as the basis for the new feature space [58], [60].

The k-Nearest Neighbors (KNN) algorithm is a relatively simple supervised machine learning method used for both classification and regression tasks. Contrary to other ML models, KNN does not necessarily have an in-built model, but rather makes computations at the time of prediction. During classification, KNN identifies certain numbers (k) of closest data points to the new input based on a chosen distance metric, such as Euclidean distance, cosine distance, and hamming distance etc. The class of the new sample is then determined by the majority voting among these “neighbours”. The distances can be regarded as a metric for similarity and need to be chosen carefully as using appropriate similarity measure can effectively enhance the K-nearest neighbor algorithm’s classification accuracy [61], [62], [63].

A decision tree consists of nodes and edges, where each edge represents the making of a decision and each node represents a final class label or a regression value assigned to that sample. The process of splitting dataset and making decisions is applied recursively until the stopping condition is met, such as only leaving certain number of samples in a node. The Random Forest (RF) algorithm consists of multiple decision trees, as described above, where the subset used for each individual tree are randomly selected from the training data via bootstrap aggregating. Thus,

the decisions made at the nodes are based on only on a fraction of the original features, and the final decision is made by majority voting or taking the average among all the decision trees. This structure helps to improve the overall model performance and eliminate overfitting [64], [65].

Similar to the working mechanism of a neuron, where a reaction is only given when the stimulation exceeds a certain threshold, artificial neural networks (ANNs) rely on different activation functions (mostly sigmoid functions) to determine whether or not an output should be produced. ANNs typically consist of an input layer, one or more hidden layers, and an output layer. Each node receives inputs and applies a weighted sum; the result is then passed through an activation function to introduce non-linearity. Interestingly, the hidden layers are often called the "black box" because humans generally are unable to interpret how exactly deep learning algorithm reaches a particular conclusion, while these layers are the key components for capturing complex patterns and revealing hidden features in the data [66], [67]. Weights are the parameters that define the importance of each input feature, and bias provides additional shifting in the activation function for better fitting. During training, these two parameters are continually adjusted to minimize the error between the predicted and actual outputs as a "learning" process of the network. ANNs have been proven to be particularly useful when dealing with wide range of data types with complex features, such as image generation, voice recognition and spectra data processing etc. [68], [69]

Contrary to the models mentioned above, autoencoders (AEs) are a type of unsupervised neural network. One AE can be considered as a basic building blocks of a neural network, and they can be stacked to form hierarchical deep models. It is often used in tasks like dimensionality reduction, non-linear feature learning, and data denoising without any pre-labelled training data [70]. An AE consists of two main components: the encoder and the decoder. The encoder compresses the input data into a lower-dimensional space to obtain meaningful features for effective data representation. This compressed representation, or latent space, is then passed to the decoder, which attempts to reconstruct the original data from the lower-dimensional code. The network is trained to minimize the reconstruction error during this process, the difference between the input and the output is often used as an evaluation metric [71], [72].

2.4 ML-assisted EV spectral data analysis

2.4.1 Deep learning

Deep learning-based algorithms are among the most commonly used models for spectral analysis due to their automatic feature extraction and classification capabilities. These algorithms can extract useful information from the high-dimensional data of spectra and have demonstrated superior performance in analyzing spectral signals [73], [74], [75], [76]. The biomedical field has rapidly advanced over the past decades, with improvements in laboratory instrumentation and analysis techniques leading to an exponential growth in biological databases. These databases often feature a number of variables (features) that exceed the number of observations [77]. Managing such high-dimensional data presents significant challenges for statistical and machine learning methods, making dimensionality reduction essential for alleviating the burden on the network and enhancing model efficiency.

A 2023 study involved 543 patient samples and 210 healthy controls, demonstrating the clinical potential of early-stage cancer diagnosis and tissue of origin (TOO) discrimination among six types of cancers [78]. Using an AuNP-aggregated array chip as a SERS substrate, exosomes isolated via size exclusion chromatography (SEC) from plasma samples were directly dropped onto the dot array in a label-free manner for subsequent spectra collection (**Figure 2.3 (A)**). To account for indistinguishable signals from common impurities such as lipoprotein and calnexin, 100 signals were collected per sample, and the classification results were reported as an averaged prediction per single spectrum based on the multiple instance learning (MIL) concepts. The signals were preprocessed to remove noise, correct the baseline, and eliminate spiked data. Anomalous data were excluded if the intensity of the common band near 860 cm^{-1} did not exceed a manually specified threshold determined through signal comparison. Employing a customized CNN classifier, the decision system first performs a binary classification to determine whether the samples exceed the threshold for classification as a patient with an AUC of 0.97. Subsequent determination of TOO also showed a robust sensitivity of 90.2% with the implementation of a multi-layer perceptron (MLP) network.

A similar study focused on the stratification between the healthy and diseased population, as well as the classification among three different types of cancers, where a total of 110 clinical samples were involved [79]. Interestingly, before the spectra were fed into a 1D CNN algorithm, they were first augmented to further increase the diversity. Augmentation also helped to maintain the original distribution of each spectrum in the original library while avoiding subsequent overfitting during the training. Baseline changes, random noise addition, and spectra shifts to either the left or right were applied. A linear combination of augmented spectra was used to augment the dataset for 10 folds. Then, general maximum-minimum normalization described before was performed to minimize the effect from the difference in intensity before passing the spectra as inputs of the CNN model. To further demonstrate the superiority of the proposed model with the highest accuracy being 98.27%, the classification results were also compared to traditional machine learning model such as LDA, AlexNet and GoogLeNet-based CNN etc.

2.4.2 Other supervised models

Other supervised models are relatively less common to be used as the principal data analysis method given the inherent complexity of the spectral data unless the sample pool size is limited. Researchers tend to use multiple models simultaneously to find the one with the desired performance. For example, a series of models such as Artificial Neural Network (ANN), SVM, tree classification, RF, AdaBoost, gradient boosting and KNN etc. were applied for the classification among glioblastoma stem cell (GSC) and differentiated glioblastoma cells (DGC) [80]. Standard baseline correction using the rubber band method and vector normalization was applied. Additionally, a negative second derivative of the spectra was calculated using the Savitzky–Golay filter as a denoising method. The dimensionality reduction methods applied are particularly noteworthy. The spectra were first reduced to the fingerprint region from 400 –1800 cm^{-1}, which contains rich bioactive peaks, and then subjected to principal component analysis (PCA). Only 30 principal components (PCs) were selected and used as input for the ML models. Compared to the original data with over 1400 features, this significant reduction in dimensionality helped accelerate computation time by narrowing the focus to a few selected peaks instead of the entire spectrum. Although only a general prediction accuracy of 60.3% was achieved, the ML

models provided a new perspective on population differences from a biological standpoint. Using K-means clustering to assign all single-point measurements into four clusters, the clusters showed distinctive lipid and nucleic acid contents that matched previous reports by identifying characteristic peak positions.

In addition to sample classification and determining the cellular origin of detected EVs, machine learning (ML) can also be applied to predict the density of EVs in sample mixtures using SERS. A RF classifier was initially trained with EV spectra derived from five mixtures of THP-1 and HT-1080 cell lines with varying concentrations, using the combination of concentrations as the class label [81]. A micro/macro-F1 score of over 99% demonstrated the feasibility of detecting sample compositions. Subsequently, an ANN was trained with unmixed HT-1080 and THP-1 EV measurements, while spectra from the previous mixed samples were used solely in the testing sets (**Figure 2.3 (B)**). A clear shift in the tendency of classification as HT-1080 was observed as the concentration of HT-1080 increased. However, the predicted probability of THP-1 decreased rapidly to almost 0 at a 50:50 sample proportion due to the statistical dominance effects. Further analysis involved calculating the Raman shifts using a Gaussian curve fitting algorithm, revealing key wavelengths contributing to the recognition. It is also worth noting that this proof-of-concept study was carried out only with cell line samples, clinical samples with their complex background will pose more challenges to their approach.

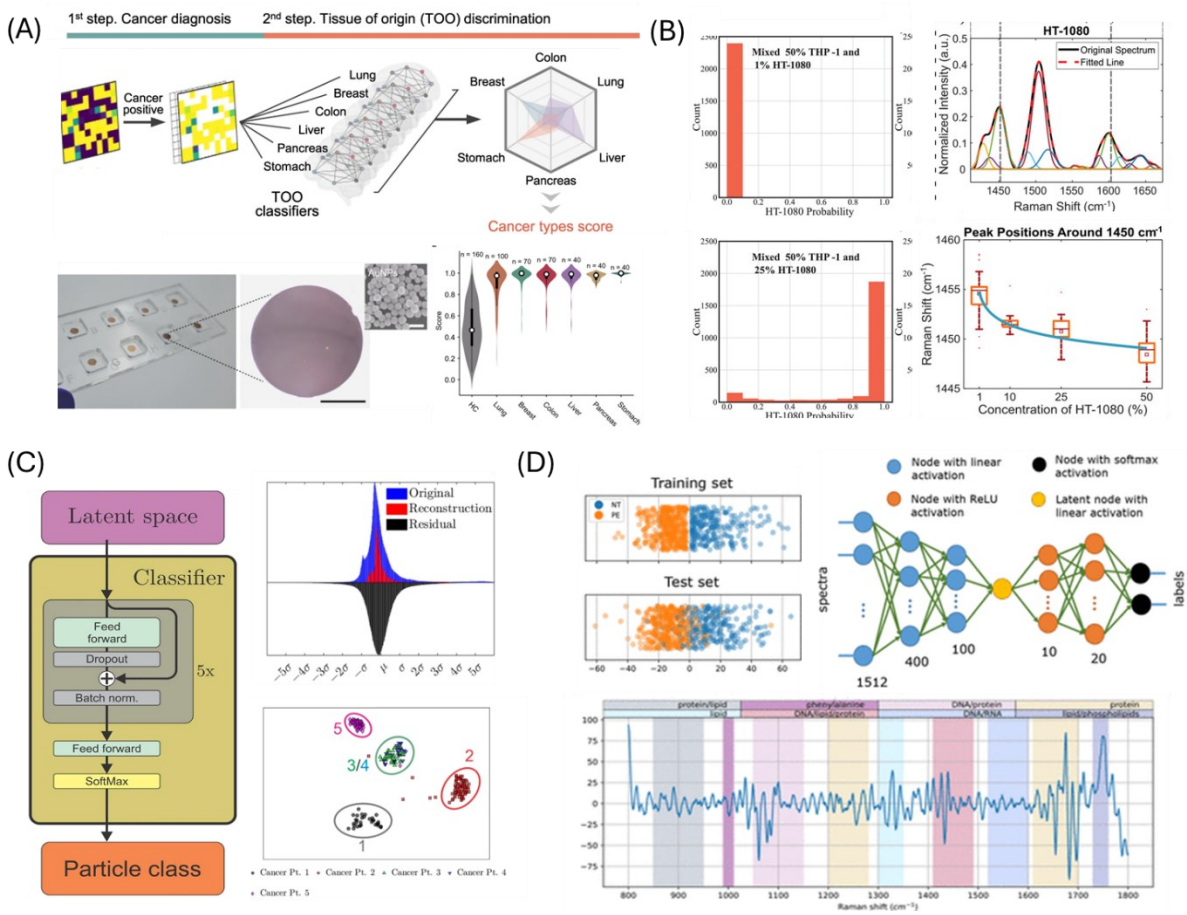

Figure 2.3: ML-assisted data analysis for EV spectra.

(A) Multiclass cancer subtyping first assigns cancer scores as an averaged result to each sample, calculated as the mean values of the multiple instance learning (MIL)-based cancer classifier. The samples deemed as patient will be further examined with pre-trained prediction models, one for each specific cancer. A score tendency by type is produced as model prediction. The plasma samples were directly added onto AuNPs coated on the APTES-functionalized cover glass [78]. Copyright 2023, Nature Communications. **(B)** Artificial neural network algorithm (ANN) can also be applied to profile the EV mixtures concentration, prepared in different ratios quantitatively. The probability of model prediction to be HT-1080 EVs drastically decreased as the concentration reduced from 25% to 1% in the mixed solution. Characteristic peaks were identified around 1450 cm^{-1} and 1600 cm^{-1}, with the peak intensity forming a calibration curve in solution with varying concentrations [81]. Copyright 2023, Small. **(C)** Utilizing Autoencoders (AE) as a dimensionality reduction tool, a classifier is introduced in the latent space. The "condensed" spectra are passed to five blocks of fully connected feed-forward layers with dropout and optional skip connections. Gaussian noise was added to the original dataset prior to being passed into the AE and was significantly reduced in the reconstructed from decoder. Two-dimensional t-SNE projection of latent space EV showed clear distinction among different samples [82]. Copyright 2024, Scientific Reports. **(D)** AE was used in conjunction SERS-active thin films directly formed via femtosecond laser nanopatterning. Projection of training and test set in the one-dimensional latent space demonstrate clear boundaries between the train/test set. Particular directions in Euclidean hyperspace which optimally separates the data can be calculated using the first linear layers, which shed light onto two positive peaks around 1330 and 1745 cm^{-1} that contributes to the classification [83]. Copyright 2022, ACS Sensors.

2.4.3 Autoencoders

Apart from the advantages of using supervised neural networks discussed above, autoencoders often require uniform input data with the same frequency range and resolution for all spectra. This requirement restricts the use of data from multiple sources and introduces issues with calibration drifts during lengthy data collection. These problems usually necessitate extensive data processing, which can be time-consuming. Self-supervised autoencoder learning may offer a solution to these challenges [82]. Using Raman spectra from two laboratories and 13 biological sources, including commercially available cell lines, human blood, and lipoproteins, the autoencoder was first trained in a self-supervised manner (**Figure 2.3 (C)**). The encoder's goal is to extract chemically relevant information from the spectra and pass it to the latent space. To achieve this, the spectra are cut at random locations with Gaussian noise addition and wavelength shifts, making the network robust to inputs with different spectral ranges. The decoder then reconstructs the original spectra using only the essential information in the latent space, employing a Fourier loss function to evaluate the quality of the reconstructed spectra. After validating the autoencoder system, the latent representation of the spectra is used as input for another deep learning network. Due to the extensive pre-processing by the encoder, this subsequent network is relatively small and interacts only with the spectra in the latent space. The test set is also processed through the encoder and then passed to the classifier. This novel system has demonstrated promising results, as t-SNE clustering in the latent space showed clear distinctions, and the system achieved over 90% accuracy in classifying the biological nanoparticles to their correct origins without any preset labels.

Autoencoders are usually used as a dimensionality reduction method, and the decoder output is fed into an additional classifier. However, Kazemzadeh et al. were inspired by the network structure and developed an autoencoder-based neural network classifier [83]. As the pioneering study utilizing femtosecond laser-machined nanoplasmonic surfaces for placental EV SERS applications, spatial K-means clustering was first performed on the normalized spectra of all samples (**Figure 2.3 (D)**). This unsupervised technique automatically labels the samples based on their shapes and peaks, which is particularly useful given the high heterogeneity among EV

populations and within individual samples, making it difficult to assign a uniform label to a single sample. Evident differences in the distribution of cluster centers between EV samples indicated the presence of different mixtures of EV populations in each sample. Following this, a hybrid autoencoder-inspired network and a CNN model were applied for classification tasks. Although the CNN model achieved a significantly higher classification accuracy of 96% compared to the autoencoder, the autoencoder offers the unique advantage of linking any obtained values for each spectrum to directly indicate the presence or absence of specific Raman spectral features. Two distinct positive peaks were observed around 1330 and 1745 cm^{-1}, highlighting the important role of lipids or phospholipids in population stratification in this case.

2.4.4 EV spectral analysis incorporated with dimensionality reduction techniques

The primary goal of dimensionality reduction is to simplify datasets while preserving significant information, thus improving computational efficiency and model performance [84]. Techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have traditionally been used to reduce dimensionality by transforming data into a lower-dimensional space [85], [86]. More recently, AEs have gained popularity for dimensionality reduction due to their ability to capture nonlinear relationships through latent space, as mentioned above. This is particularly important in SERS spectra, where each spectrum contains thousands of features. Subtle data variations are crucial for accurate classification and diagnosis, making the careful selection of dimensionality reduction methods vital, as they may significantly impact the performance of machine learning models.

A recent study investigated different dimensionality reduction approaches of Raman spectra to improve the classification accuracy of intracranial tumors, particularly glioblastomas [87]. Two main methods were explored: PCA and a feature-filtering approach based on biochemical components. PCA is a common technique, while the feature-filtering approach focused on selecting spectral shifts corresponding to statistically significant differences between tissue groups, such as glioblastoma centers and normal white matter. Specifically, the Fisher criterion was calculated for each spectral point across three tissue classes: normal brain tissue,

glioblastoma margin, and glioblastoma center. Spectral points exceeding the critical Fisher value were retained, while others were discarded. A key advantage of feature selection methods is their interpretability, which is valuable for clinicians. After identifying significant features, the study matched them with known biochemical components using a reference library. This approach not only reduced data dimensionality, but also ensured that the retained features were biologically relevant, enhancing tissue classification accuracy. The two techniques were integrated with a support vector machine (SVM) classifier, achieving an accuracy of 83% and 92% specificity in tumor tissue classification.

In some cases, dimensionality reduction methods, such as Independent Component Analysis (ICA), Partial Least Squares (PLS), and LDA, were used as additional modules to further push the limit of classification. Popp et al. first constructed a Raman dataset containing over 2,200 spectra from eight bacterial classes, including various Bacillus species [88]. These proposed dimensionality reduction techniques were combined with a genetic algorithm, inspired by natural evolution, to optimize a merit function related to experimental accuracy. The algorithm employs iterative processes of selection, mating, and mutation to refine a population of vectors, with key steps including constructing the initial population, selecting parents, creating offspring, and replacing less optimal vectors. A supplementary study explored the impact of the number of spectra on accuracy. While more spectra provide critical information that might be lost with fewer scores, increasing their number also raises model complexity and computational time, potentially leading to overfitting. Ultimately, using 35 spectra in the training set was found to balance accuracy and robustness effectively, making it a practical choice when robustness is not the primary concern. The application of the PCA algorithm achieved the highest increase in classification accuracy, improving it by 6%.

2.5 Conclusion

To conclude, the integration of optical sensors and ML presents a promising pathway for the enhanced detection and characterization of EVs, particularly in the context of cancer diagnostics. Label-free optical biosensing techniques, such as SERS, offer significant advantages by eliminating the need for molecular tags, thus minimizing the risk of false positives and non-

specific aggregation. With ML's capacity for processing complex, large-scale data, it has demonstrated wide application and great flexibility in improving the interpretation of high dimensional spectral data, offering insights into the underlying biological patterns. ML-driven analysis enhances the ability to draw meaningful conclusions from noisy or low-resolution data through customized data analysis workflows. Furthermore, the incorporation of dimensionality reduction techniques has been shown to improve model performance, increasing the accuracy of classification outcomes. Ultimately, the combination of advanced optical sensors and ML algorithms holds the potential to revolutionize EV-based diagnostics, providing new opportunities for single-vesicle analysis and improving the accuracy and efficiency of clinical applications.

Potential future pathways include further development of optical biosensor platforms to find an optimal balance between the use of bio labels for specificity and the sensitivity derived from platform structures and materials for molecule-specific information. Another key area for future research is the standardization of data processing workflows across different sensor platforms and spectral data types, as current data preprocessing steps are often customized for specific projects, which can limit reproducibility and broader application in clinical settings. By developing more universal data analysis protocols, researchers can achieve more consistent and reliable outcomes across studies.

3 Optimizing Binary Classification of Heterogeneous Raman Spectra: A Comparison of Dimensionality Reduction Techniques on Cancer Paradigm

Yao Lu[1], Carolina del Real Mata[1], Mahsa Jalali[2], Marjan Khatami[2], Laura Montermini[2], Kevin Petrecca[3], Janusz Rak[2], Livia Garzia[2], Sara Mahshid[1,4]*

[1] Department of Bioengineering, McGill University, Montreal, QC, H3A 0E9 Canada

[2] Research Institute of the McGill University Health Centre (RIMUHC), Montreal, Quebec, H4A 3J1 Canada

[3] Department of Neuropathology, Montreal Neurological Institute-Hospital, McGill University, Montreal, Quebec H3A 2B4, Canada

[4] Division of Experimental Medicine McGill University Montreal, QC H3A 0E9, Canada

*sara.mahshid@mcgill.ca

This manuscript has been submitted for publication to Scientific Reports and is now under review.

Abstract

Surface-enhanced Raman Spectroscopy (SERS) offers label-free molecular fingerprints of extracellular vesicles (EVs) as a promising biomarker for cancer hallmarks. However, effective spectrum interpretation is often hindered by the heterogeneous and nanoscale nature of EVs. Deep learning (DL) models integrated with dimensionality reduction techniques may provide novel insights by detecting subtle patterns and relationships in complex spectral data that are often challenging to identify through traditional analysis methods. Using a SERS library of single EV spectrum, four dimensionality reduction techniques—spectrum resampling, bio-inactive region cutting, Autoencoder (AE), and Gradient-weighted Class Activation Mapping (Grad-CAM)—were evaluated to improve the binary classification accuracy between two different types of cancer patients and healthy controls. Among these methods, the resampling approach yielded a significant accuracy improvement of 10.7% and 8.2% for two cancer models respectively. Resampling simple implementation shows potential for enhancing diagnostic performance using SERS-based EV analysis.

Keywords: Deep Learning, Surface-enhanced Raman Spectroscopy, Dimensionality Reduction, Extracellular Vesicles

3.1 Introduction

The nanoscale sizes and highly heterogeneous composition of extracellular vesicles (EVs) pose significant challenges for a single, effective, and uniform analysis method that can be used for different studies [1]. EVs are heterogeneous, phospholipid membrane-enclosed structures produced by all types of cells. These vesicles play crucial roles in mediating cell-to-cell communication both locally and at a distance [2],[3]. There are various approaches to characterize EVs, traditional approaches, such as antibody/aptamer labeling techniques targeting common surface markers for fluorescence detection, have been widely used [4], [5]. However, the labelling step adds increased complexity to the experiments with additional curing and washing steps. Surface-enhanced Raman Spectroscopy (SERS) may offer an alternative, this label-free approach can capture the entire biochemical signature of the molecules. Each peak in a SERS spectrum corresponds to specific molecular bonds or functional groups based on its Raman shift, while the peak intensity reflects the concentration of the corresponding molecular species [6], [7]. A previous study introduced a unique SERS-assisted nanocavity platform for isolation and label-free fingerprinting of single EVs. The photonic cavity arrays were optimized to host exactly one EV with a 97% confinement efficiency, enabling the detailed analysis of individual EVs by leveraging local electromagnetic field enhancement [8].

Decoding SERS spectrum of simple chemical substrates is straightforward as the peaks are often sparse and well-separated. However, interpreting biomolecules' spectra is challenging due to their heterogeneous nature and complex molecular composition. Often, this complexity is observed in overlapping spectrum peaks, which makes it difficult to determine the true proportion of each characteristic molecular bond. Currently, there are no standardized protocols for decoding SERS spectra [9], [10]. The recent advancements in learning models have provided a convenient yet practical solution to overcome these challenges. These data-driven pattern recognition approaches have built-in statistical models designed to capture subtle associations between input

data points (features) and can identify underlying patterns within the population. In particular, convolutional neural networks (CNN) are one of the most commonly applied deep learning (DL) algorithms for analyzing biomolecular spectral data and can facilitate cancer diagnosis tasks [11], [12], [13], [14]. As the filters from the convolutional layer slide through the input data, they capture both the local patterns and the hierarchical representations of the data, exploiting simple features such as the position of individual peaks to more complex patterns like relationships between peaks as the layer "deepens" [14], [15], [16].

Despite the advantages that CNN offers for spectra analysis, they are still hampered by the inherent complexity of SERS spectra. Firstly, the wavelength shift along the x-axis can be stretched depending on the optical spectrometer and its calibration in each study. Additionally, the intensity along the y-axis is often accompanied by noise [17]. As a result, researchers often develop tailored data treatment protocols specific to the project's needs [18]. This lack of uniformity introduces potential ambiguities and hinders the transfer of learning among studies. Secondly, the dataset used for training CNNs usually requires a balanced representation from each sample population to avoid network bias towards the major class. However, assigning a uniform label to all samples overlooks the heterogeneity of EV populations and leads to decreased model classification performance. Lastly, while the range of SERS shift collection can be personalized, the spectrum grating rate ranges from just over one thousand to approximately three thousand, with various biologically significant peaks shared between healthy and cancer samples. This places considerable pressure on the network to identify the key features that can differentiate between classes.

Unsupervised learning may provide novel insights into sample composition from a different perspective. These algorithms try to find underlying patterns without explicit guidance from humans, thus no labeling of the data is required before training and the results of the unsupervised model may reveal previously undiscovered data relationships [19], [20]. In addition, spectra dimensionality reduction techniques may also help address the aforementioned concerns. By focusing on the most informative features, dimensionality reduction can effectively mitigate the risk of overfitting and enhance model interpretability by mapping high-dimensional data to a

lower-dimensional space. However, an optimal balance point needs to be established between retaining original information and improving model performance [21], [22], [23].

In this work, we have built a sample spectral library consisting of single EV spectra collected using the previously developed nanocavity platform (**Figure 3.1**) [8]. To ensure repeatability and potential for generalization, all proposed data analysis techniques were applied separately to two patient datasets, each belonging to one of the two different cancers available, Medulloblastoma and Glioblastoma. These two cancers were selected as model diseases because they are among the most common brain tumors observed in pediatric and adult patient groups. Investigating the behavior of related EVs crossing the blood-brain barrier and their potential role in EV-facilitated diagnosis could inform future studies on other brain tumors [37], [89]. The goal is to improve the binary classification accuracy between healthy controls and cancerous populations. We began by simulating a realistic clinical setting in which the number of healthy samples (n = 16) is substantially larger than that of patient samples (n = 8) for each cancer. We then explored the feasibility of increasing accuracy through direct adjustments to the sample train/test distribution. After determining the optimal ratio, we applied K-means clustering for sample subgrouping and potential pattern discovery. However, the high heterogeneity of EVs made it difficult to identify a consistent pattern within the same population, leading us to explore dimensionality reduction techniques. Methods such as resampling spectra, removing the bio-inactive region, Autoencoder (AE), and Gradient-weighted Class Activation Mapping (Grad-CAM) were applied.

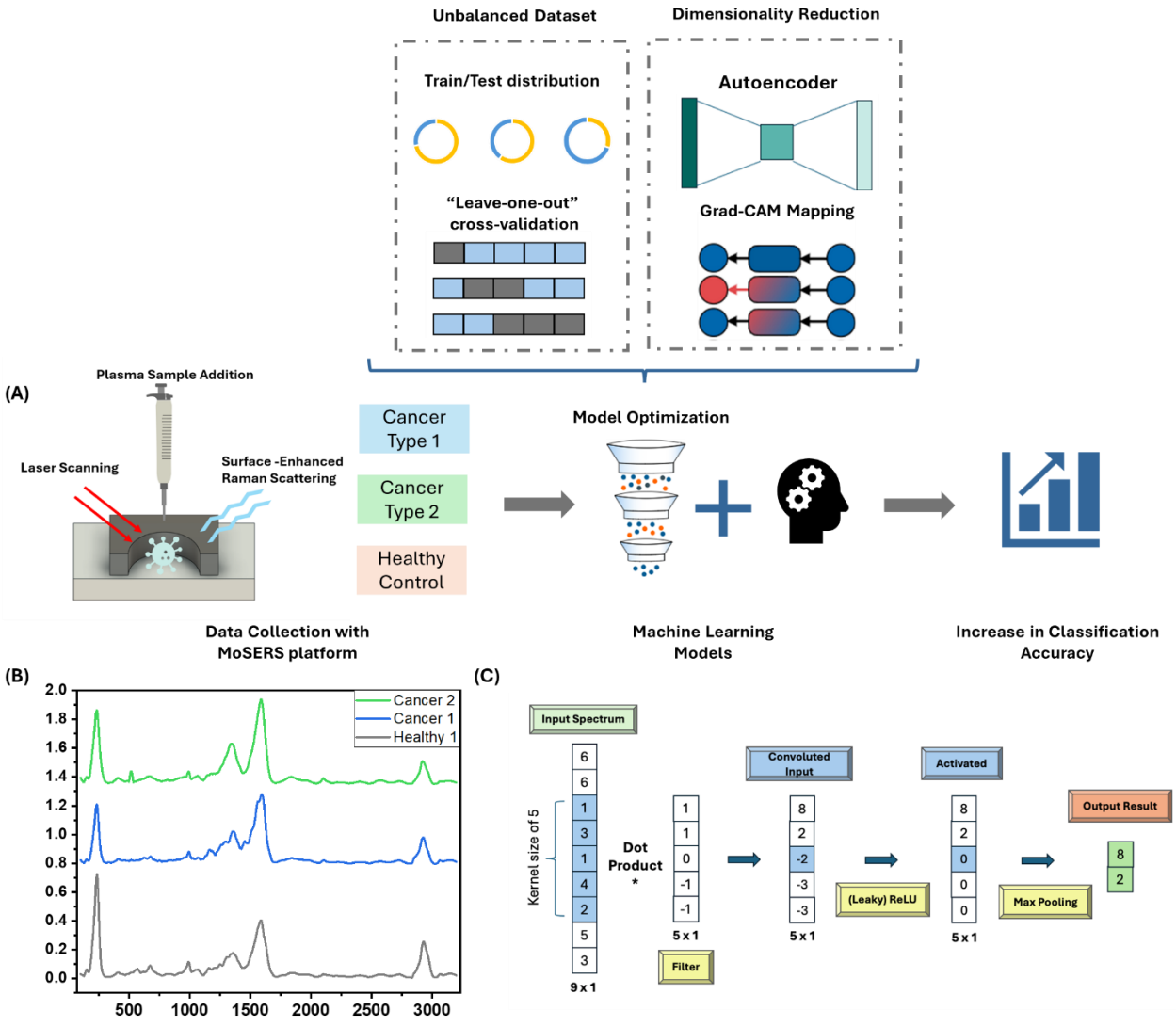

Figure 3.1: General Schematics.

(A) After single EV SERS spectra for cancer patients were collected using the nanocavity platform in a label-free manner, the data first underwent standard cleaning procedures such as baseline subtraction, normalization and smoothing. To mimic a realistic clinical setting, the number of patient samples (n=8) was set to be only half of the healthy controls (n =16), and an optimal train/set data distribution was first discussed. Then, dimensionality reduction algorithms, including an AE and Grad-CAM, were compared and studied to find the highest increase in binary population classification compared to the base accuracy from the original dataset. **(B)** Averaged spectrum from the cancerous and healthy populations. The peak positions are difficult to discern, where the main differences come from the height of the peaks. The high EV heterogeneity within each sample should also be considered in addition to averaged results. **(C)** General schematics for the structure of a CNN and simplified working principle of the convolutional layer.

3.2 Methods

3.2.1 Data Collection

The datasets used were single-EV spectra collected using the nanocavity platform. Detailed experiment procedures and data-cleaning protocols before DL analysis have been described in previous publications [37]. Briefly, EVs were first isolated from the blood-plasma samples of Medulloblastoma (MB) patients (n = 8), Glioblastoma (GBM) patients (n = 8), and healthy controls (n = 16) via IZON columns. MB is addressed as "Cancer Type 1" and GBM is addressed as "Cancer Type 2" in this study. Using an InVia Raman microscope, the isolated EVs samples were loaded into the nanocavity platform in a minute amount. Each spectrum was collected between wavelengths 100–3200 cm^{-1}, with each dataset containing approximately 70 single EV spectra.

GBM samples were provided by our collaborator Dr. Petrecca and were collected under approval from the Neurosciences Panel of the MUHC Research Ethics Board (REB: IRB00010120). The MB samples and healthy controls were obtained through our collaborators from a biobank (MP-37-2017-3256). All human samples were collected with consent from the subjects or legal guardians.

3.2.2 Data Cleaning

Prior to saving each spectrum, its baseline was corrected using an in-built function from WiRE 5, where a linear baseline was automatically fit in the collected spectral range. The data files were then compiled via MATLAB R2022a to ease batch processing. Subsequently, the dataset undergoes normalization and smoothing by Savitzky-Golay with a second-order polynomial fit performed using Origin Pro 2019b software. The data is now ready for DL analysis.

3.2.3 CNN Model Construction and Evaluation

The base CNN model was adapted from Ho et al. [90], written in Python (Python Software Foundation. Python Language Reference, version 3.8.8. Available at http://www.python.org) in Spyder 5.4.2 and utilizing the Pytorch package 2.0 [91]. The network is based on a ResNet-based

structure with an initial convolutional layer, followed by one residual layer, and a final fully connected layer. The initial convolutional layer has 64 filters, and the hidden residual layer contains 20 filters. The residual layer includes 4 convolutional layers, bringing the total depth of the network to 6 layers. During each epoch of the training process, 10% of the training set was randomly selected and used as the validation set to evaluate the robustness of the model. The performance of each CNN model was evaluated based on the classification accuracy calculated from the confusion matrix using Equation 1 below. True positives (TP) are the number of class 1 spectra correctly predicted as their given label of 1 matches the predicted one. True negatives (TN) are the number of class 0 spectra correctly predicted, with a label of 0. False negatives (FN) represent the number of class 1 spectra misclassified as class 0, and false positives (FP) are the number of class 0 spectra misclassified as class 1.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (1)$$

3.2.4 Autoencoder

An autoencoder (AE) is a type of unsupervised neural network designed to extract features from unlabeled data by compressing the input information into a lower-dimensional latent space [25], [26]. The encoder was constructed with an input layer that feeds into two sequential encoder layers, where each encoder layer includes a batch normalization and a rectified linear activation function (ReLU) to introduce non-linearity and stabilize the learning process. The output is a condensed representation in the latent space. The dimensionality was set to be reduced to half of the original input features. The structure was followed by two decoder layers similar to the encoder but in reversed order and with 1 output layer. A linear activation function was again included to ensure continuous output values. Finally, the AE was compiled using the Adam optimizer and mean squared error (MSE) loss function for the quantification of the reconstruction error between the input and the output data. The Adam optimizer uses the exponential moving averages of both the gradients and their squared values as momentum terms, dynamically adjusting the learning rates for each parameter until reaching optimal performance [27], [28].

3.2.5 Gradient-weighted Class Activation Mapping

Gradient-weighted Class Activation Mapping (Grad-CAM) uses gradients that flow into the final convolutional layer and produce a coarse localization map highlighting the important regions used during classification [29]. Starting with the forward pass, the input was passed through the CNN model, and feature maps were generated at each convolutional layer following the normal order progression of the model. Then in the backward pass, the gradients of the output for the feature maps were calculated as a reversed process of the forward pass. As convolutional features naturally retain spatial information, which is lost in the fully connected layers, we capture the feature map generated by the last convolutional layer. The gradient of the score for a class (class $\boldsymbol{c}$) was first calculated using Equation 2 below, where $\boldsymbol{y^c}$ is the score of class $\boldsymbol{c}$, and $\boldsymbol{A^k}$ is the feature map of a convolutional layer. The neuron importance weight is noted using α_k^c.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2}$$

The gradients were then pooled to obtain a single importance weight for each feature map channel. Each feature map was multiplied by its corresponding importance weight, emphasizing the regions of the feature maps that are most important for predicting the target class. A ReLU was applied to the linear combination of maps because in our case only features that have a positive influence on the classification are of interest.

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{3}$$

Finally, the weighted feature maps were summed along the channel dimension to produce a single importance map, which was normalized to a range between 0 and 1 [30], [31].

3.3 Results and Discussion

3.3.1 Unbalanced Data Library

A CNN model was selected as the base algorithm for this study due to its natural feature extraction capability via convolutional layers. Each neuron in a convolutional layer is connected to a local region of the input spectrum, allowing the CNN to focus on smaller regions where important features may reside [32], [33]. This corresponds to the nature of SERS spectra, where key information comes from peaks' position, intensity, and shapes. Additionally, instead of the conventional pooling layer, the network was implemented with stridden convolution, which introduces a stride variable that controls the step size of the filter as it moves over the input. This effectively preserves the location of the SERS peaks while reducing computational complexity when dealing with large datasets [15], [34].

As a general consideration, the number of spectra from each population used to train the CNN should be balanced. This helps ensure the network produces unbiased and generalized results, avoiding favoring the dominant class and misclassifying the minority class [35]. However, this principle might not apply when using CNN to facilitate cancer diagnosis in a clinical setting. In EV-based studies, the diverse heterogeneity in structure and composition of these biovesicles from the plasma samples also poses great challenges to DL analysis [8], [36]. It is suspected that only around 3% of the plasma-derived EVs carry brain tumor mutation signatures, given the difficulty of crossing the blood-brain barrier [37]. Thus, the signals from cancerous EVs can easily be diluted against a vast background of non-cancer-related EVs shared between both the healthy and diseased groups. This questions the notion of an even class balance for training a model, as an underrepresentation is expected from one of the classes. Furthermore, as many cancer-related EVs are still understudied, there is currently no standard library available to represent the SERS spectra of disease signal-bearing EVs to the best of our knowledge.

3.3.1.1 Adjusting Ratio between Healthy and Patient Train Set

To address these concerns, we propose a simple approach, reducing the number of spectra used in training from each healthy sample to a certain percentage of the patient group before

applying other advanced data analysis techniques. This method reassesses weights during training, as the majority class dominates the loss function. As gradients and weights are updated during backpropagation, the network prioritizes minimizing the loss for the majority class. By adjusting the network to assign greater weight to the minority class, such as patient spectra, the model can improve its accuracy in recognizing these less common cases, which are critical in real-world scenarios [38], [39]. This strategy assumes that healthy samples are generally more uniform, thereby reducing concerns about generalization to new healthy samples. However, finding an optimal balance is essential and will be discussed further later in the text.

A total of 8 patient samples from both Cancer 1 and Cancer 2, and 16 healthy controls were used in this study to simulate a more realistic clinical background with an unmatched number of samples available. As shown in **Figure 3.2 (A)** and **Figure 3.2 (E)**, the confusion matrix for the original dataset was obtained with an equal number of spectra from both populations used for training, using 60% of each healthy sample and 70% of each patient sample. Since more healthy spectra were set aside for testing than patients, the global accuracy calculated with the complete test set may be biased (**Figure S1**). To provide a more accurate measure, local accuracy is reported here by randomly sampling the healthy test set to match the remaining patient spectra. The model achieved an average classification accuracy of 57.0% for Cancer 1 versus healthy and 76.8% for Cancer 2 versus healthy, establishing the baseline accuracy for the study. Notably, while nearly all healthy spectra were correctly classified, a significant portion of patient spectra were misclassified as false negatives. This aligns with earlier observations that the algorithm overemphasized the healthy population when trained on a balanced dataset, due to the dominance of healthy samples.

By adjusting the ratio between the training set of the two groups, an optimal value was found by setting the ratio to 0.6 for Cancer 1 (number of spectra in the healthy train set = number of spectra in the patient train set * 0.6) and 0.8 for Cancer 2. It led to the highest increase in average classification accuracy compared to the other ratio (**Figure S2**). For Cancer 1, accuracy improved by 6.3%, where the evident redistribution of the patient spectra confirms that an uneven training set ratio compels the network to focus more on patient signatures (**Figure 3.2 (B)**). The same trend was reflected in **Figure 3.2 (C)**, as the data point distribution of the patient spectra shows a clearer focus. Cancer 2 followed a similar trend, with a moderate accuracy increase of 2.6%. Although the

improvement in classification probability was less pronounced, the patient spectra in **Figure 3.2 (F)** showed a clearer distribution with significantly fewer false negative cases. Also, the distribution became more uniform with reduced extreme values, indicating that fewer spectra were misclassified as healthy (**Figure 3.2 (G)**). Overall, this adjustment in the training set ratio resulted in a moderate enhancement in model performance, demonstrating the effectiveness of this strategy in improving classification accuracy for the minority class.

3.3.1.2 Leave-One-Out Cross-Validation

Leave-One-Out Cross-Validation (LOOCV) is an established method for estimating the performance of machine learning algorithms where one data point is excluded from the training and the model is trained on the rest of the dataset. The excluded data is then included in the testing set, and the same process will be repeated for each data point in the dataset [40], [41]. In this study, where the number of patient samples is only half that of the healthy controls, a certain portion of healthy samples can be excluded from the training set and later used entirely in the testing set. This approach not only improves the balance of group representation in training but also provides additional means of evaluating the robustness of the model. The average change in classification accuracy for Cancer 1 versus healthy is shown in **Figure 3.2 (D)**, with different numbers of healthy samples excluded from the training set noted on the x-axis. Excluding three healthy samples from the training set yielded the highest accuracy increase of 3.7%, while excluding more led to reduced accuracy compared to the baseline. Listed in **Table S2**, we evaluated how well the model performed when healthy samples were partially included in the training set (Healthy #1-3) and entirely used in the testing set (Healthy #7-8), using one model with six healthy samples allocated completely for testing. The CNN model performed relatively poorly when encountering unseen samples, as their probability of being classified as patients outweighed the correct classification, leading to a decrease in overall accuracy. This also suggests that the healthy population also exhibits certain heterogeneity, as the model could not generalize learned patterns to a relatively large number of new samples.

For Cancer 2 versus healthy, no significant accuracy increase was observed and excluding more than four healthy samples decreased overall model performance (**Figure 3.2 (H)**). This trend

presumably originated from adjusting the percentage of each healthy sample used for training. As more healthy samples were allocated entirely for testing, a higher percentage of the remaining ones were used for training to balance the number of spectra from both populations. This enabled the model to better learn healthy signatures, improving performance when fewer unknown samples were introduced. However, to achieve better generalization, a larger sample pool is necessary. Thus, at this step, we report that using a 0.6 ratio for Cancer 1 and a 0.8 ratio for Cancer 2 produced the highest accuracy increase for our dataset. All subsequent studies in this report were conducted using these two ratios respectively for the training sample distribution.

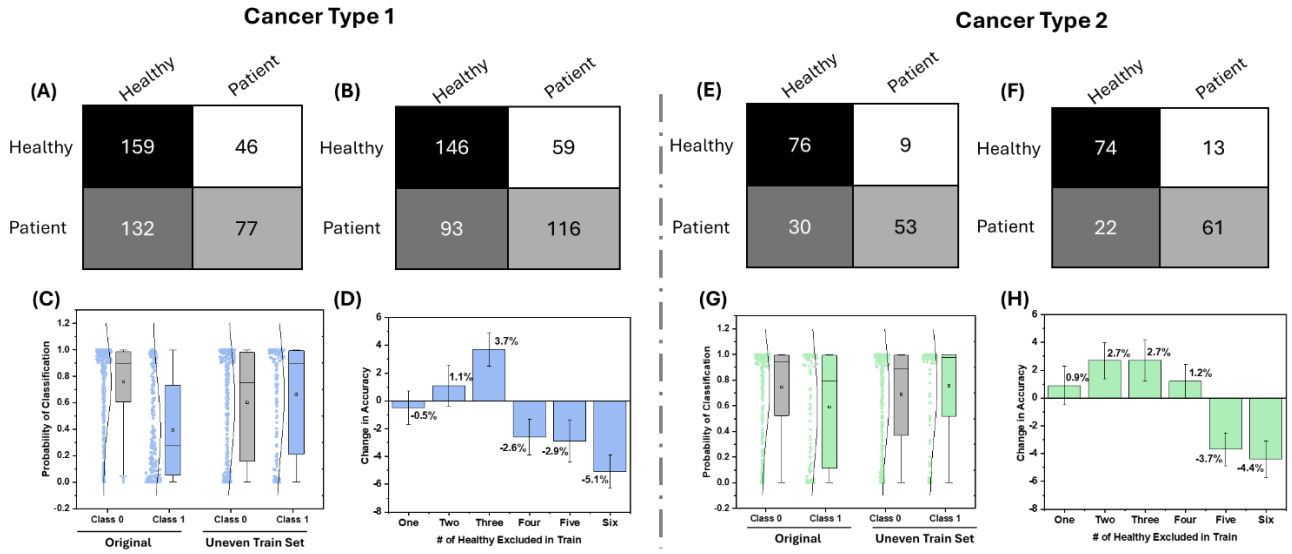

Figure 3.2 Classification accuracy changes with varying train/test distribution.

(A) to (D) for Cancer 1. (A) Confusion matrix for the original dataset with an averaged classification accuracy of 57.0%, which serves as the base accuracy. **(B)** Confusion matrix for the model with 0.6 ratio in the training set. **(C)** Box plot for the distribution of classification accuracy for the model in panel **B (D)** The average change in classification accuracy for varying numbers of healthy samples excluded in Leave-One-Out Cross-Validation (LOOCV) algorithm. **(E) to (H) for Cancer 2**. **(E)** Confusion matrix for the original dataset with an averaged classification accuracy of 76.8 %. **(F)** Confusion matrix for the model with a training set ratio of 0.8. **(G)** Box plot for the distribution of classification accuracy for the model in panel F. **(H)** Averaged change in accuracy for LOOCV models.

3.3.2 Unsupervised Learning with K-means clustering

As mentioned, a predetermined single label is always assigned to all spectra from a sample group since a CNN was the supervised learning model employed for the classification tasks in this

study. The model compares the predicted outcomes with the assigned labels to calculate its accuracy and loss values, and "learns" to perform better over time. Nevertheless, the nanocavity platform used for the spectra collection is a label-free technique and EVs exhibit exceedingly high heterogeneity, even within the same sample. Therefore, assigning the same label across an entire healthy or patient sample's EV population may lead to decreased classification performance. This is speculated to be a result of the model's struggles to identify a uniform and generalized pattern that accurately summarizes an entire population.

K-means clustering is a common unsupervised method where data points are assigned into K groups based on their distance from each group's centroid. The data points closest to a given centroid are clustered into the same category, allowing the identification of natural groupings within the data [20]. Based on previous observations that lipoproteins usually constitute a large portion of the EV signals collected, the number of groups was set to three [42]. The remaining signals from patients and healthy individuals were assumed to form one cluster each. Spatial K-means clustering was performed for all samples, with three examples each from the healthy and the two cancer populations shown in **Figure 3.3 (A)**. The clustering results did not demonstrate a coherent pattern within the groups, as varying levels of spectra were assigned to the three clusters (**Fig. S3**). Increased cluster numbers were also tested (**Fig. S4**), but no evident pattern could be determined. The spectra clustering visualization for Healthy #1 showed a clear distinction among the K-means group within the same sample (**Figure 3.3 (B)**). The cluster centers of the proposed three clusters in both **Figure 3.3 (C)** and **Figure 3.3 (D)** showed relatively similar peak positions across two cancer models, with the main differences coming from the height of the peaks. This incoherence among samples suggests additional data treatment methods are needed to enhance the model's performance.

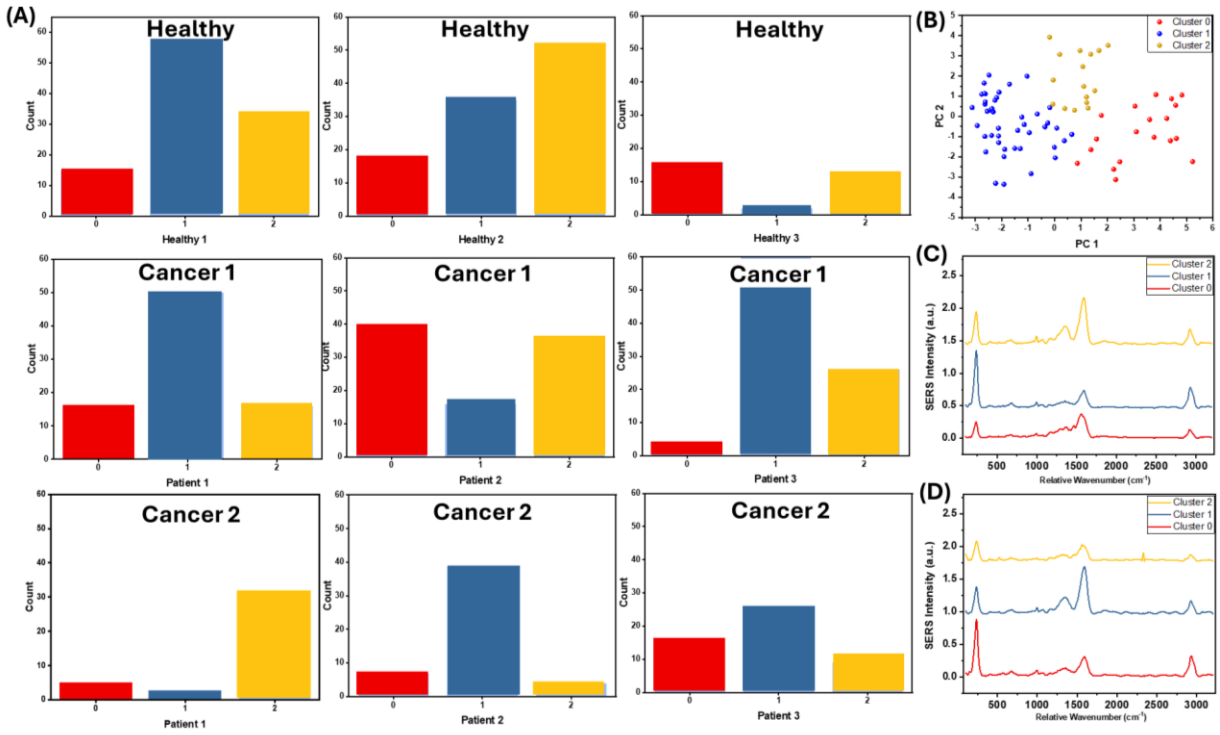

Figure 3.3 K-means clustering for unsupervised subgroup identifying.

(A) Label distributions in K-Means clustering when K set to 3 for sample healthy controls and two cancer groups. **(B)** K-means clustering visualization for Healthy #1 using principal components in 2D dimension. **(C) and (D)** Clusters' centers for each K-means group with matching colors to panel A for Cancer 1 and Cancer 2 respectively.

3.3.3 Spectral Dimensionality Reduction

3.3.3.1 Data Condensation with AE and Resampled Data

In our dataset, a single Raman spectrum consists of 2924 data points, resulting in 2924 features for machine learning models. Such high-dimensional data often contains certain irrelevant information, which can increase complexity and reduce the model's predictive power as it becomes difficult to find generalized patterns for the assigned classes. Moreover, the "curse of dimensionality" may lead to an increased risk of overfitting, which occurs when a model becomes overly tailored to the training data, capturing noises that do not generalize well to unseen data [9], [43], [44]. To mitigate these challenges, dimensionality reduction is an avenue worth exploring.

AE is a type of unsupervised DL model commonly employed for dimensionality reduction. AEs usually consist of two main parts: an encoder and a decoder, it can be imagined as having a funnel structure (**Figure 3.4 (A)**). The encoder transforms the original data into a latent space, preserving only the features that help identify complex patterns. The decoder then reconstructs the original data using these preserved features, ideally closely matching the input data. The model is typically trained with the purpose of minimizing the reconstruction error, with the key lying in the lower-dimensional representation of the input features in the latent space [26]. Since the process of data condensing is unsupervised, the machine may be able to discover previously hidden patterns without human bias [45]. Thus, we propose training the CNN classifier in this latent space, leveraging the reduced dimensionality and condensed feature mappings for improved performance. By focusing on the most relevant features, the AE may help the CNN avoid overfitting and enhances its ability to generalize to new data.

To further enhance the robustness of the model, Gaussian noise was added to the original spectra before they were transformed into the latent space via the encoder. This is a common practice for autoencoders, as adding noise to the previously cleaned dataset forces the decoder to focus on and capture the most essential features during the reconstruction process [46], [47]. By learning to reconstruct the original data from noisy inputs, the model becomes more resilient to variations and noise, which are often encountered in real-world applications. As shown in **Figure 3.4 (B)**, the spectrum reconstructed by the decoder exhibited a reduced noise level and more closely resembled the original spectra compared to the input that was fed into the encoder, proving that key features were retained in the latent space. To gain further insights into the behavior of data condensation, the data distribution in the latent space was visualized (**Figure 3.4 (C)**). The visualization revealed that most spectra from the two classes largely overlapped, which reflects the shared characteristics of these biovesicles. Only a small portion of the data deviated from the main overlapped area, potentially containing the key information necessary for differentiating the samples.

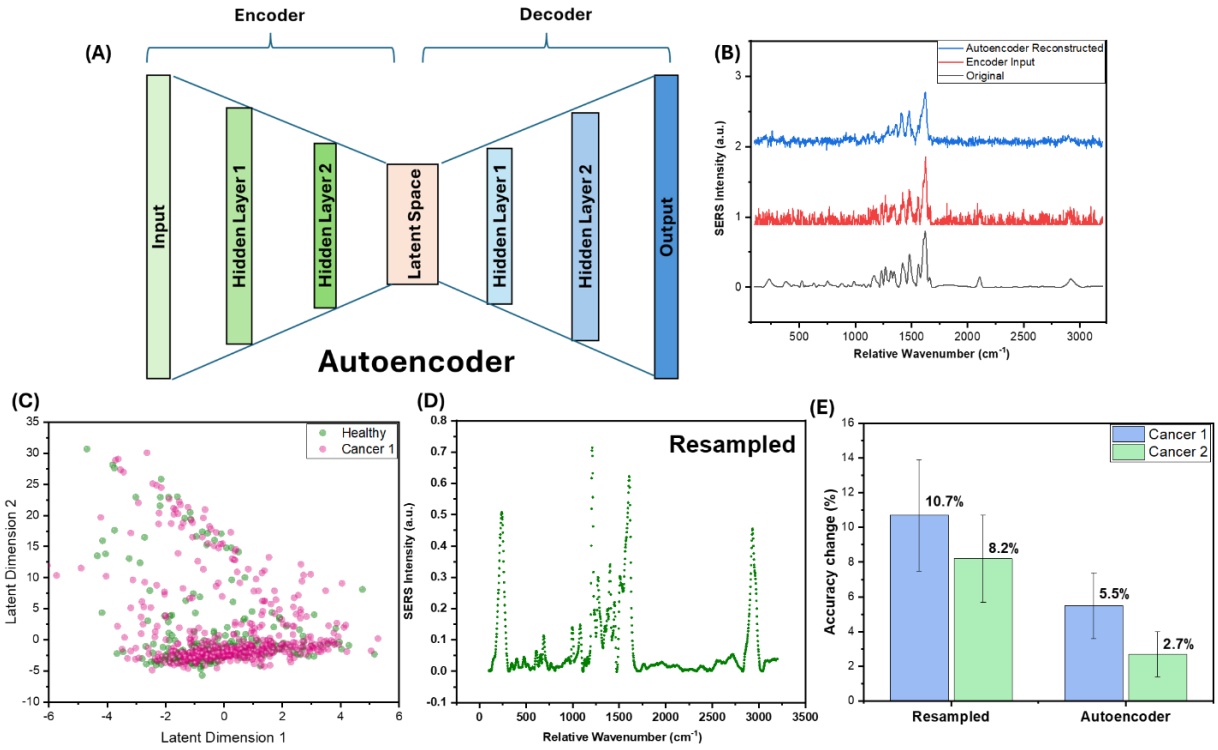

Figure 3.4 Data Condensation with Autoencoder and Resampled Data.

(A) The CNN model was interfaced with an Autoencoder (AE), where subsequent classification occurred within the latent space, and the features were condensed in an unsupervised manner. **(B)** Gaussian noise was added to the original spectra before passing into the AE to reinforce the model's capability of filtering important features. The reconstructed spectra showed significantly reduced noise level. **(C)** Data visualization in the latent space using the first two dimensions. **(D)** A resampled spectrum consisting of 1245 features. **(E)** Classification accuracy change for the two cancer models. Both resampled dataset showed significant increase in classification accuracy, while the improvement from AE is less evident.

On the other hand, resampling is a technique that selects data points from the original samples to form new datasets and was designed for class imbalance problems [48], [49]. In the case of Raman spectra, the key information lies in the position and intensity of the peaks in a consecutive format, meaning no single data point determines the final classification. Therefore, by taking a spin on the traditional resampling methods, the spectra can be resampled using a set of predetermined wavelengths that selectively omit intermediate points on the x-axis. Specifically, from an original dataset containing 2924 features, the resampling process reduces the dataset to 1245 features, cutting the feature count by more than half. As shown in **Figure 3.4 (D)**, the spacing between each data point is sparser compared to the original data since only half of the features

were retained, but it preserves the general shape and location of the peaks which can greatly facilitate effective classification.

The effectiveness of the two applied data condensation techniques was evaluated by comparing their increase in accuracy relative to the base accuracy (**Figure 3.4 (E)**). The models trained with the resampled data exhibited a significant improvement, achieving a 10.7% and 8.2% increase in accuracy for Cancer 1 and Cancer 2, respectively. Thus, by focusing on a subset of relevant data points, resampling has proven effective in mitigating the effects of the "curse of dimensionality". This technique ensures that the essential characteristics, like the height and position of the SERS spectra, are retained while eliminating unnecessary data points, leading to more accurate predictions with reduced computation power. However, when the AE was integrated into the model, the improvement was not as evident. This outcome suggests potential limitations of using AE for this specific dataset. The data condensation in the latent space may have resulted in the loss of critical information, particularly the location and height of peaks, which are essential for distinguishing between classes. While the AE is effective at reducing dimensionality, it might have oversimplified the data, leading to the omission of subtle yet important spectral features. In complex datasets like SERS spectra, where classification depends on these fine-grained details, such losses can significantly impact model performance.

3.3.3.2 Data Selection with Bio-silent Region Cut and Grad-CAM

There is a "silent region" on the SERS spectra, ranging from 1800 cm^{-1} to 2800 cm^{-1} there are no biologically relevant peaks [50], [51]. In EV-based studies, removing this section from the spectra while preserving the remaining features for DL analysis may serve as an effective method of dimensionality reduction. However, this approach does not necessarily guarantee an increase in classification accuracy. DL models, often regarded as "black boxes", are challenging to decompose into intuitive components that humans easily understand. As a result, the primary evaluation metric for assessing a model's performance is classification accuracy, which can make it difficult to pinpoint where the network may have erred or how to enhance its accuracy more effectively.

To address this challenge, the Grad-CAM algorithm was developed in 2017 [29]. Originally designed for interpreting images, Grad-CAM identifies the final convolutional layer in

the network and calculates the gradient information flowing into that layer (**Figure 3.5 (A)**). The feature maps that are generated at each convolutional layer capture various patterns and spatial hierarchies in the data, the gradients reflect feature importance by indicating how much each feature in a particular layer contributes to the final class score. This process produces a coarse localization map that highlights important regions in the image for predicting a specific class.

We propose applying a similar concept to spectra data. After training the model on spectra with the bio-silent region removed, we can backtrack the gradient maps. These gradients typically averaged over all spatial locations, yield a set of importance weights that indicate the overall significance of each feature map channel for the target class prediction. Regions with higher weights contribute more to the final feature importance mapping [30], [31], [52]. This map can then be used to identify which regions of the spectra the CNN network considers most important for predicting the target class. To test this approach, we retrained a new CNN model using only these identified regions of the spectra, as a method of both feature mapping and dimensionality reduction, to evaluate whether this improves the accuracy. A sample spectrum is shown in **Figure 3.5 (B)**, where the highlighted regions indicate the areas on the original spectrum that align with the Grad-CAM-generated importance map. These peaks were deemed more significant for sample classification by the previously trained CNN model. Given the high heterogeneity of EVs, the averaged spectra may not serve as the best representation. Therefore, 20 spectra from each sample were randomly selected, and their feature maps were calculated to identify the common wavelength regions deemed important by the machine. **Figure 3.5 (C)** shows the SERS intensity at the five most common peak points identified by Grad-CAM, revealing locations that differentiate among the three populations. For instance, at 1492 cm^{-1}, the healthy controls exhibited relatively uniform intensity, whereas the two cancer groups displayed greater variability. Conversely, at 1544 cm^{-1}, the cancer populations showed more uniform intensity, while the healthy group exhibited more variation.

An average accuracy increases of 3.7% was observed when the biosilent region was removed from the Cancer 1 dataset, suggesting a modest improvement despite the significant reduction of 800 features. Similarly, the Cancer 2 dataset showed a 2.6% increase in accuracy (**Figure 3.5 (E)**). However, this improvement is not particularly substantial, indicating that the

removed region was not critical for classification. This is further supported by **Figure 3.5 (B)**, where no prominent feature importance peaks appear within the region between 2000 cm^{-1} and 2800 cm^{-1}. A similar trend is observed in **Figure 3.5 (D)**, where the biosilent region appears relatively inactive, lacking significant Raman peaks that could aid in distinguishing between different classes. When Grad-CAM mapping was applied, the model for Cancer 1 experienced a slight decrease in average performance, while a 2.1% accuracy increase was noted for the Cancer 2 dataset. These results suggest that the selection of wavelength regions based on feature importance may not have adequately captured the complex and diverse nature of EVs, leading to an incomplete or biased representation of the data. Additionally, the process of manually selecting important regions and training the model exclusively on these areas may have distorted the relative positions of SERS peaks, which are crucial for accurate classification. This distortion could have resulted in the loss of important contextual information necessary for distinguishing between different classes, ultimately reducing performance.

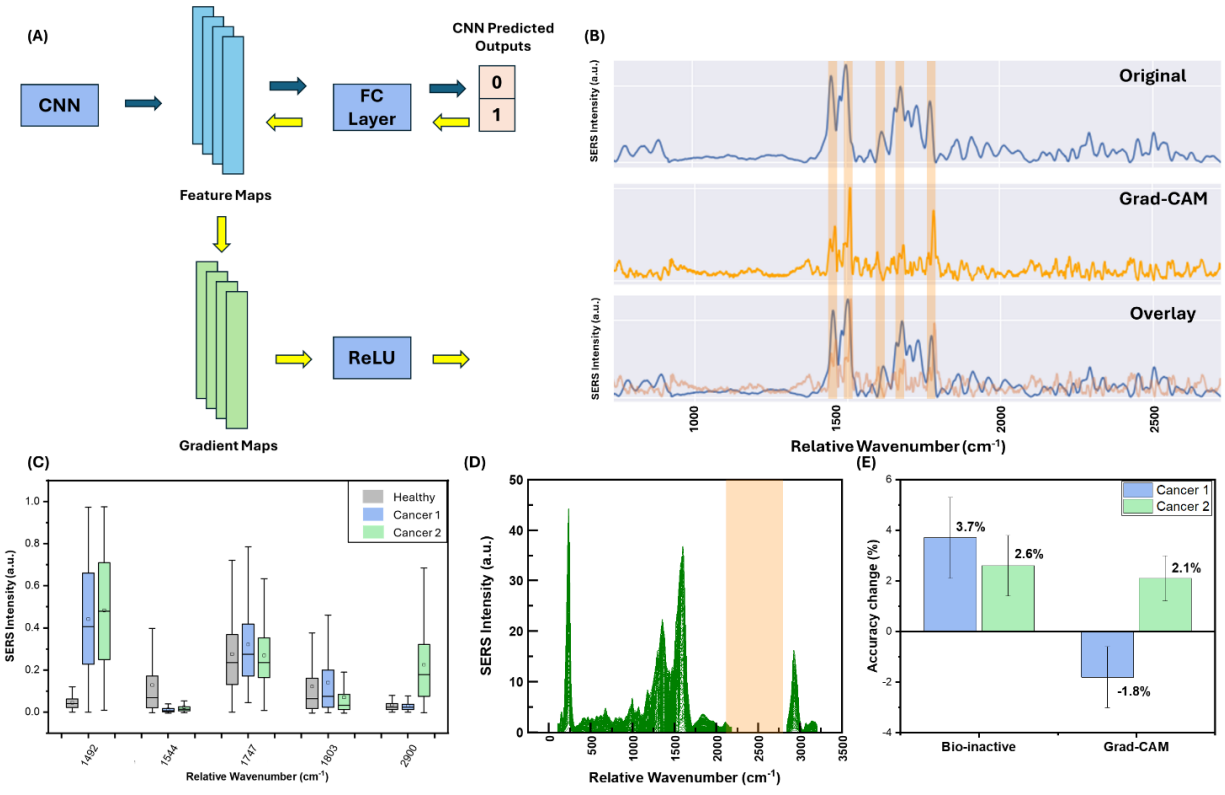

Figure 3.5 Data Selection with Bio-silent Region Cut and Grad-CAM.

(A) Grad-CAM backtracks the gradient distribution in the last convolutional layers of the CNN model and generates a feature map that highlights key sections of the spectra that the model considered important for classification. **(B)** Grad-CAM feature importance mapping for one sample spectrum. The overlapping region indicates the features considered important by the model during classification. **(C)** Boxplot of SERS intensities at five most common wavelengths with high Grad-CAM importance. **(D)** Bio-silent region is highlighted with no biologically active peaks and was removed in the training set as an effective method of feature reduction. **(E)** Classification accuracy change for the two cancer models. The removal of bio-inactive regions did not lead to significant change in accuracy, while the manual selection of key regions based on Grad-CAM caused distortion of the relative position of the peaks with reduced model performance.

3.4 Conclusion

The study investigates several methods for enhancing the binary classification accuracy of a CNN model using single EV Raman spectra datasets. Initially, by optimizing the training set ratio between the two classes, the network was compelled to place more emphasis on the minority class. Such improvement was also achieved without the addition of any complex algorithms to the original network. Subsequently, we compared four dimensionality reduction techniques, finding that resampled data led to the most significant improvement. This outcome aligns with previous studies demonstrating that resampling is one of the most effective approaches for constructing classifiers from imbalanced datasets. In addition, the proposed resampling method was proven effective across two cancer models and follows a standardized data treatment protocol, making it potentially transferable to other studies. The findings show the importance of both class balance in training and careful feature selection in improving the robustness and accuracy of DL models in this context.

For future studies, it will be important to expand the focus beyond binary classification to explore multiclass scenarios, which can reveal more detailed patterns within the data. Multiclass classification also offers deeper insights into the varying degrees of disease states and pushes the boundary of DL model performance to more real-life applications. Additionally, increasing the sample pool is critical for enhancing the model's generalization capabilities. A larger dataset would not only improve the robustness of the model but also help minimizing overfitting when the number of spectra sample surpasses the spectrum dimension, leading to more reliable predictions across diverse populations.

Acknowledgment

Authors thank the Engineering Faculty at McGill University, the Canadian Cancer Society (255878 CCSRI), and the Charles Bruneau Foundation/ The Research Institute of the McGill University Health Centre (RI-MUCH) 9094 (259300). S.M. acknowledges financial support from the Canada Research Chairs Program. Y.L acknowledges financial support from the Faculty of Engineering for the McGill Engineering Undergraduate Student Masters Award (MEUSMA). C.d.R.M. acknowledges the support from Fonds de Recherche du Québec - Nature et Technologies (FRQNT) doctoral fellowship and the Faculty of Engineering for the McGill Engineering Doctoral Award (MEDA) award. MJ acknowledges the Child Health Research Excellence postdoctoral scholarship at RI-MUCH and the Canadian Institutes of Health Research (CIHR) postdoctoral fellowship.

Author Contributions

S.M., Y.L., and C.d.R.M. contributed to the idea conception. Y. L., C.d.R.M, and M.J. contributed to the design and planning of the experiments. L.M., M.K., J.R., and L.G. contributed to obtaining and preparing human samples. Y. L., C.d.R.M, and M.J. contributed to data collection. Y.L. contributed to data analysis. Y.L. and C.d.R.M. contributed to the interpretation of results. Y.L., C.d.R.M., M.J., and S.M. contributed to the preparation of the manuscript with the support of all co-authors. S.M. supervised the project and contributed to funding acquisition.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

Code available at: https://github.com/saramahshid/Dimensionality-reduction

The patient spectra library that supports the findings of this study is not openly available due to reasons of sensitivity and is available from the corresponding author upon reasonable request. Data are located in controlled access data storage at McGill University.

References

[1] Z. Zhao, H. Wijerathne, A. K. Godwin, and S. A. Soper, “Isolation and analysis methods of extracellular vesicles (EVs),” Extracell. Vesicles Circ. Nucleic Acids, vol. 2, pp. 80–103, 2021, doi: 10.20517/evcna.2021.07.

[2] E. I. Buzas, “The roles of extracellular vesicles in the immune system,” Nat. Rev. Immunol., vol. 23, no. 4, pp. 236–250, Apr. 2023, doi: 10.1038/s41577-022-00763-8.

[3] G. van Niel, D. R. F. Carter, A. Clayton, D. W. Lambert, G. Raposo, and P. Vader, “Challenges and directions in studying cell–cell communication by extracellular vesicles,” Nat. Rev. Mol. Cell Biol., vol. 23, no. 5, pp. 369–382, May 2022, doi: 10.1038/s41580-022-00460-3.

[4] “New Technologies for Analysis of Extracellular Vesicles | Chemical Reviews.” Accessed: Aug. 05, 2024. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.chemrev.7b00534

[5] Y. Couch et al., “A brief history of nearly EV-erything – The rise and rise of extracellular vesicles,” J. Extracell. Vesicles, vol. 10, no. 14, p. e12144, 2021, doi: 10.1002/jev2.12144.

[6] M. T. Yaraki, A. Tukova, and Y. Wang, “Emerging SERS biosensors for the analysis of cells and extracellular vesicles,” Nanoscale, vol. 14, no. 41, pp. 15242–15268, 2022, doi: 10.1039/D2NR03005E.

[7] H. Shin, D. Seo, and Y. Choi, “Extracellular Vesicle Identification Using Label-Free Surface-Enhanced Raman Spectroscopy: Detection and Signal Analysis Strategies,” Molecules, vol. 25, no. 21, Art. no. 21, Jan. 2020, doi: 10.3390/molecules25215209.

[8] M. Jalali et al., “MoS2-Plasmonic Nanocavities for Raman Spectra of Single Extracellular Vesicles Reveal Molecular Progression in Glioblastoma,” ACS Nano, vol. 17, no. 13, pp. 12052–12071, Jul. 2023, doi: 10.1021/acsnano.2c09222.

[9] A. Plante et al., “Dimensional reduction based on peak fitting of Raman micro spectroscopy data improves detection of prostate cancer in tissue specimens,” J. Biomed. Opt., vol. 26, no. 11, p. 116501, Nov. 2021, doi: 10.1117/1.JBO.26.11.116501.

[10] Q. Zhang, T. Ren, K. Cao, and Z. Xu, “Advances of machine learning-assisted small extracellular vesicles detection strategy,” Biosens. Bioelectron., vol. 251, p. 116076, May 2024, doi: 10.1016/j.bios.2024.116076.

[11] M. Erzina et al., “Precise cancer detection via the combination of functionalized SERS surfaces and convolutional neural network with independent inputs,” Sens. Actuators B Chem., vol. 308, p. 127660, Apr. 2020, doi: 10.1016/j.snb.2020.127660.

[12] S. Yan et al., “Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level,” Talanta, vol. 226, p. 122195, May 2021, doi: 10.1016/j.talanta.2021.122195.

[13] C.-C. Xiong et al., “Rapid and precise detection of cancers via label-free SERS and deep learning,” Anal. Bioanal. Chem., vol. 415, no. 17, pp. 3449–3462, Jul. 2023, doi: 10.1007/s00216-023-04730-7.

[14] H. Shin et al., “Single test-based diagnosis of multiple cancer types using Exosome-SERS-AI for early stage cancers,” Nat. Commun., vol. 14, no. 1, Art. no. 1, Mar. 2023, doi: 10.1038/s41467-023-37403-1.

[15] C.-S. Ho et al., “Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning,” Nat. Commun., vol. 10, p. 4927, Oct. 2019, doi: 10.1038/s41467-019-12898-9.
[16] Y. Qi et al., “Recent Progresses in Machine Learning Assisted Raman Spectroscopy,” Adv. Opt. Mater., vol. 11, no. 14, p. 2203104, 2023, doi: 10.1002/adom.202203104.
[17] M. N. Jensen et al., “Identification of extracellular vesicles from their Raman spectra via self-supervised learning,” Sci. Rep., vol. 14, no. 1, p. 6791, Mar. 2024, doi: 10.1038/s41598-024-56788-7.
[18] F. Cui, Y. Yue, Y. Zhang, Z. Zhang, and H. S. Zhou, “Advancing Biosensors with Machine Learning,” ACS Sens., vol. 5, no. 11, pp. 3346–3364, Nov. 2020, doi: 10.1021/acssensors.0c01424.
[19] N. Verbeeck, R. M. Caprioli, and R. Van de Plas, “Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry,” Mass Spectrom. Rev., vol. 39, no. 3, pp. 245–291, 2020, doi: 10.1002/mas.21602.
[20] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” Inf. Sci., vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.
[21] G. T. Reddy et al., “Analysis of Dimensionality Reduction Techniques on Big Data,” IEEE Access, vol. 8, pp. 54776–54788, 2020, doi: 10.1109/ACCESS.2020.2980942.
[22] H. S. Obaid, S. A. Dheyab, and S. S. Sabry, “The Impact of Data Pre-Processing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning,” in 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON), Mar. 2019, pp. 279–283. doi: 10.1109/IEMECONX.2019.8877011.
[23] W. Jia, M. Sun, J. Lian, and S. Hou, “Feature dimensionality reduction: a review,” Complex Intell. Syst., vol. 8, no. 3, pp. 2663–2693, Jun. 2022, doi: 10.1007/s40747-021-00637-x.
[24] A. Paszke et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” Dec. 03, 2019, arXiv: arXiv:1912.01703. doi: 10.48550/arXiv.1912.01703.
[25] S. Chen and W. Guo, “Auto-Encoders in Deep Learning—A Review with New Perspectives,” Mathematics, vol. 11, no. 8, Art. no. 8, Jan. 2023, doi: 10.3390/math11081777.
[26] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, “Autoencoders and their applications in machine learning: a survey,” Artif. Intell. Rev., vol. 57, no. 2, p. 28, Feb. 2024, doi: 10.1007/s10462-023-10662-6.
[27] S. Y. ŞEN and N. ÖZKURT, “Convolutional Neural Network Hyperparameter Tuning with Adam Optimizer for ECG Classification,” in 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Oct. 2020, pp. 1–6. doi: 10.1109/ASYU50717.2020.9259896.
[28] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Jan. 29, 2017, arXiv: arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980.
[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” Int. J. Comput. Vis., vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.
[30] C.-C. Xiong et al., “Rapid and precise detection of cancers via label-free SERS and deep learning,” Anal. Bioanal. Chem., vol. 415, no. 17, pp. 3449–3462, Jul. 2023, doi: 10.1007/s00216-023-04730-7.

[31] G. Shi et al., “1D Gradient-Weighted Class Activation Mapping, Visualizing Decision Process of Convolutional Neural Network-Based Models in Spectroscopy Analysis,” Anal. Chem., vol. 95, no. 26, pp. 9959–9966, Jul. 2023, doi: 10.1021/acs.analchem.3c01101.
[32] L. Alzubaidi et al., “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” J. Big Data, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.
[33] J. M. Vaz and S. Balaji, “Convolutional neural networks (CNNs): concepts and applications in pharmacogenomics,” Mol. Divers., vol. 25, no. 3, pp. 1569–1584, 2021, doi: 10.1007/s11030-021-10225-3.
[34] R. Riad, O. Teboul, D. Grangier, and N. Zeghidour, “Learning strides in convolutional neural networks,” Feb. 03, 2022, arXiv: arXiv:2202.01653. doi: 10.48550/arXiv.2202.01653.
[35] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” Insights Imaging, vol. 9, no. 4, Art. no. 4, Aug. 2018, doi: 10.1007/s13244-018-0639-9.
[36] G. Bordanaba-Florit, F. Royo, S. G. Kruglik, and J. M. Falcón-Pérez, “Using single-vesicle technologies to unravel the heterogeneity of extracellular vesicles,” Nat. Protoc., vol. 16, no. 7, Art. no. 7, Jul. 2021, doi: 10.1038/s41596-021-00551-z.
[37] M. J. Y. Ang et al., “Deciphering Nanoparticle Trafficking into Glioblastomas Uncovers an Augmented Antitumor Effect of Metronomic Chemotherapy,” Adv. Mater., vol. 34, no. 3, p. 2106194, 2022, doi: 10.1002/adma.202106194.
[38] L. N. Smith, “A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay,” Apr. 24, 2018, arXiv: arXiv:1803.09820. doi: 10.48550/arXiv.1803.09820.
[39] M. V. Narkhede, P. P. Bartakke, and M. S. Sutaone, “A review on weight initialization strategies for neural networks,” Artif. Intell. Rev., vol. 55, no. 1, pp. 291–322, Jan. 2022, doi: 10.1007/s10462-021-10033-z.
[40] C. Castellani et al., “Circulating extracellular vesicles as non-invasive biomarker of rejection in heart transplant,” J. Heart Lung Transplant., vol. 39, no. 10, pp. 1136–1148, Oct. 2020, doi: 10.1016/j.healun.2020.06.011.
[41] S. Bates, T. Hastie, and R. Tibshirani, “Cross-Validation: What Does It Estimate and How Well Does It Do It?,” J. Am. Stat. Assoc., vol. 119, no. 546, pp. 1434–1445, Apr. 2024, doi: 10.1080/01621459.2023.2197686.
[42] J. B. German, J. T. Smilowitz, and A. M. Zivkovic, “Lipoproteins: When size really matters,” Curr. Opin. Colloid Interface Sci., vol. 11, no. 2–3, pp. 171–183, Jun. 2006, doi: 10.1016/j.cocis.2005.11.006.
[43] I. Romanishkin et al., “Differentiation of glioblastoma tissues using spontaneous Raman scattering with dimensionality reduction and data classification,” Front. Oncol., vol. 12, Sep. 2022, doi: 10.3389/fonc.2022.944210.
[44] V. Berisha et al., “Digital medicine and the curse of dimensionality,” Npj Digit. Med., vol. 4, no. 1, pp. 1–8, Oct. 2021, doi: 10.1038/s41746-021-00521-5.
[45] J. Brownlee, “Autoencoder Feature Extraction for Classification,” MachineLearningMastery.com. Accessed: Jul. 22, 2024. [Online]. Available: https://machinelearningmastery.com/autoencoder-for-classification/

[46] J. Xie, J. Fang, C. Liu, and L. Yang, “Unsupervised Deep Spectrum Sensing: A Variational Auto-Encoder Based Approach,” IEEE Trans. Veh. Technol., vol. 69, no. 5, pp. 5307–5319, May 2020, doi: 10.1109/TVT.2020.2982203.
[47] C. He et al., “Accurate Tumor Subtype Detection with Raman Spectroscopy via Variational Autoencoder and Machine Learning,” ACS Omega, vol. 7, no. 12, pp. 10458–10468, Mar. 2022, doi: 10.1021/acsomega.1c07263.
[48] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
[49] S. Tyagi and S. Mittal, “Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning,” in Proceedings of ICRIC 2019, P. K. Singh, A. K. Kar, Y. Singh, M. H. Kolekar, and S. Tanwar, Eds., Cham: Springer International Publishing, 2020, pp. 209–221. doi: 10.1007/978-3-030-29407-6_17.
[50] D. Cui, L. Kong, Y. Wang, Y. Zhu, and C. Zhang, “In situ identification of environmental microorganisms with Raman spectroscopy,” Environ. Sci. Ecotechnology, vol. 11, p. 100187, Jul. 2022, doi: 10.1016/j.ese.2022.100187.
[51] M. Z. Vardaki, V. G. Gregoriou, and C. L. Chochos, “Biomedical applications, perspectives and tag design concepts in the cell – silent Raman window,” RSC Chem. Biol., vol. 5, no. 4, pp. 273–292, 2024, doi: 10.1039/D3CB00217A.
[52] M. Kazemzadeh, C. L. Hisey, K. Zargar-Shoshtari, W. Xu, and N. G. R. Broderick, “Deep convolutional neural networks as a unified solution for Raman spectroscopy-based classification in biomedical applications,” Opt. Commun., vol. 510, p. 127977, May 2022, doi: 10.1016/j.optcom.2022.127977.

3.5 Supplementary Information: Optimizing Binary Classification of Heterogeneous Raman Spectra: A Comparison of Dimensionality Reduction Techniques on Cancer paradigm

Yao Lu[1], Carolina del Real Mata[1], Mahsa Jalali[2], Marjan Khatami[2], Laura Montermini[2], Januzs Rak[2], Livia Garzia[2], Sara Mahshid[1,3]*

[1] Department of Bioengineering, McGill University, Montreal, QC, H3A 0E9 Canada

[2] Research Institute of the McGill University Health Centre (RIMUHC), Montreal, Quebec, H4A 3J1 Canada

[3] Division of Experimental Medicine McGill University Montreal, QC H3A 0E9, Canada

*sara.mahshid@mcgill.ca

Table S1: List of Abbreviations

Extracellular Vesicle	EV
Surface Enhanced Raman Spectroscopy	SERS
Deep Learning	DL
Autoencoder	AE
Gradient-weighted Class Activation Mapping	Grad-CAM
Convolutional Neural Network	CNN
Leave-One-Out Cross-Validation	LOOCV

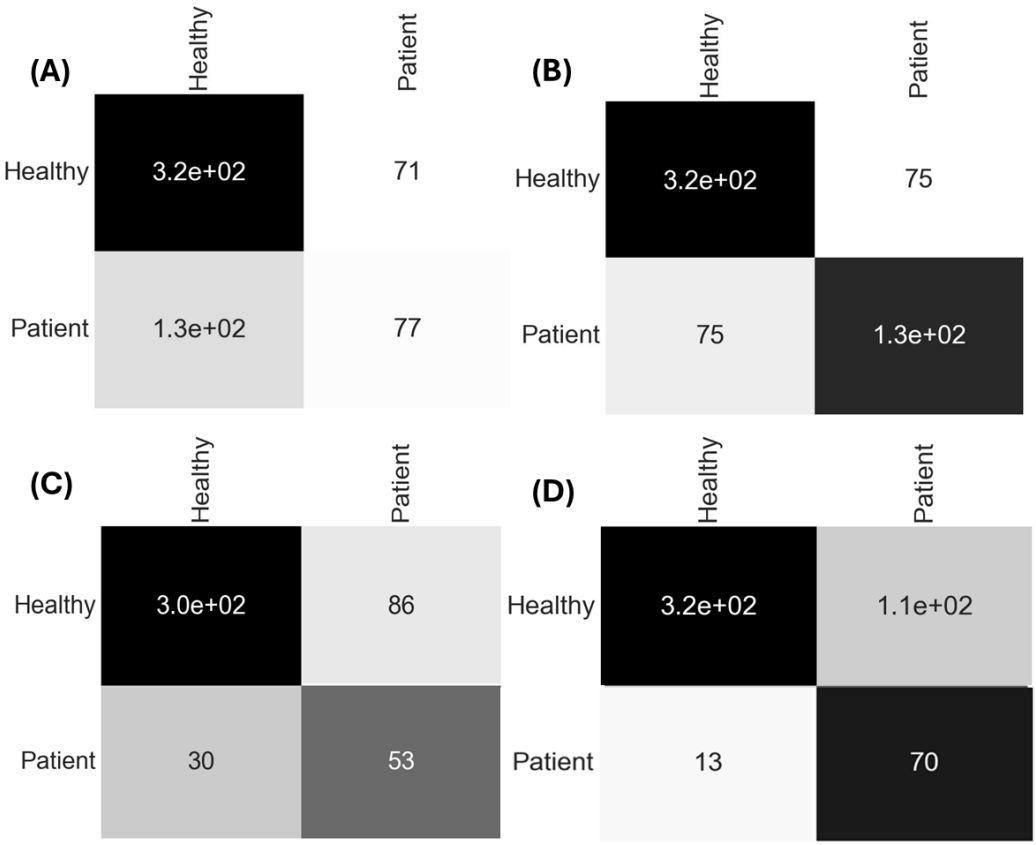

Fig S1: **(A)** Global accuracy of the confusion matrix for Cancer 1 versus healthy controls using the original dataset. **(B)** Global accuracy of the confusion matrix for Cancer 1 versus healthy controls using a 0.6 ratio for the training set. **(C)** Global accuracy of the confusion matrix for Cancer 2 versus healthy controls using the original dataset. **(D)** Global accuracy of the confusion matrix for Cancer 2 versus healthy controls using a 0.8 ratio for the training set.

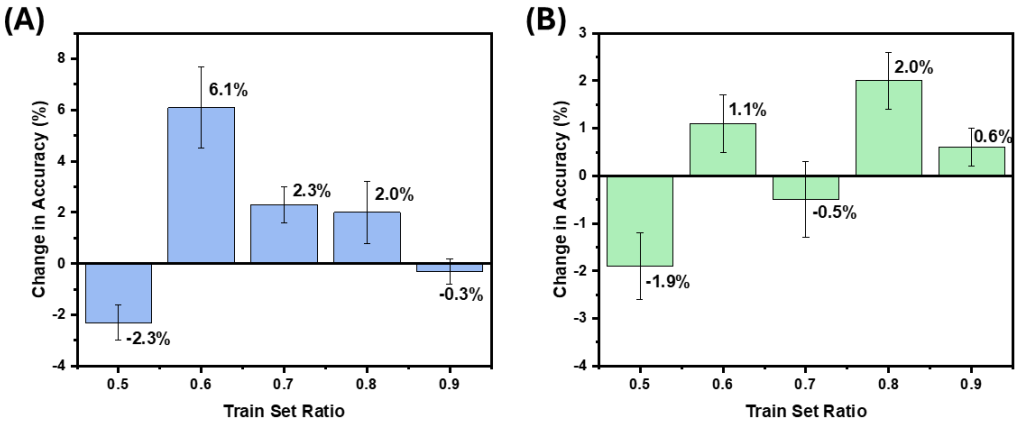

Fig S2: **(A)** Change in accuracy for different healthy/cancer proportions in the train set for Cancer 1. **(B)** Change in accuracy for different healthy/cancer proportions in the train set for Cancer 2.

Table S2: Classification probability for healthy samples included in the training set (#1 - #3) and used completely in test (#7 - #9)

	Probability of classified as Healthy	**Probability of classified as Patient**
Healthy #1	*0.63*	*0.37*
Healthy #2	*0.78*	*0.22*
Healthy #3	*0.64*	*0.36*
Healthy #7	*0.23*	*0.77*
Healthy #8	*0.43*	*0.57*
Healthy #9	*0.53*	*0.47*

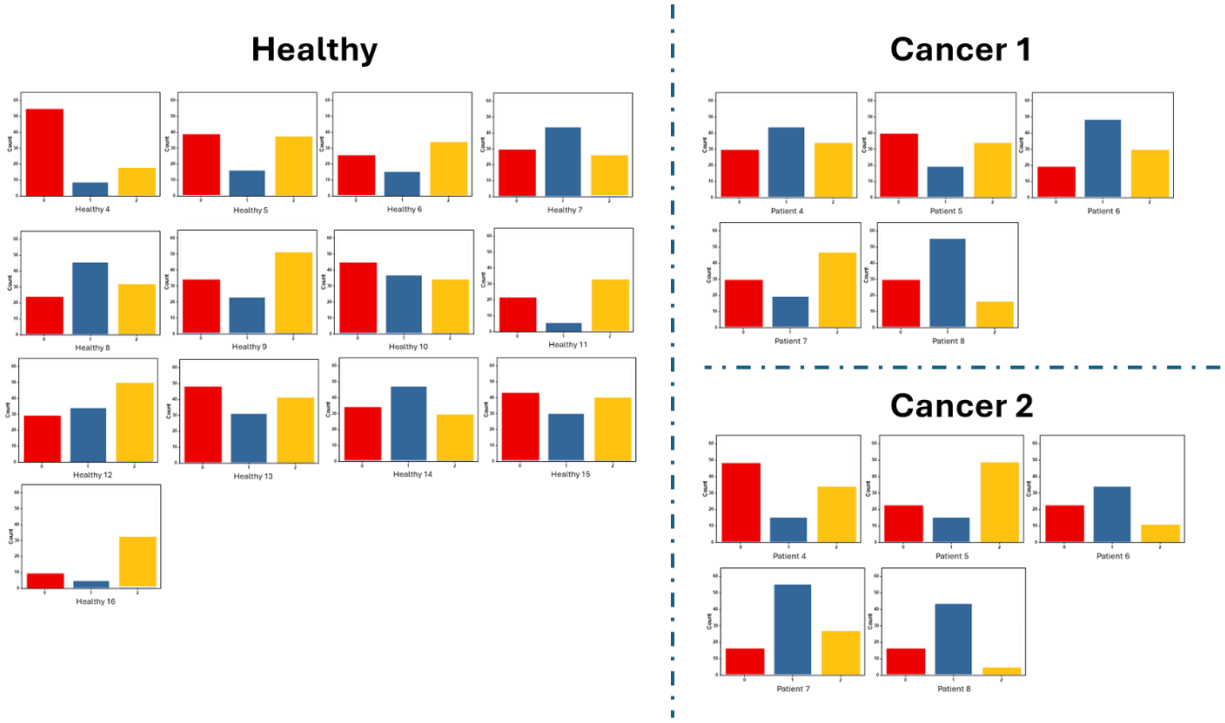

Fig S3: K-means clustering (k =3) for the remaining Cancer 1, and Cancer 2 patients and healthy samples.

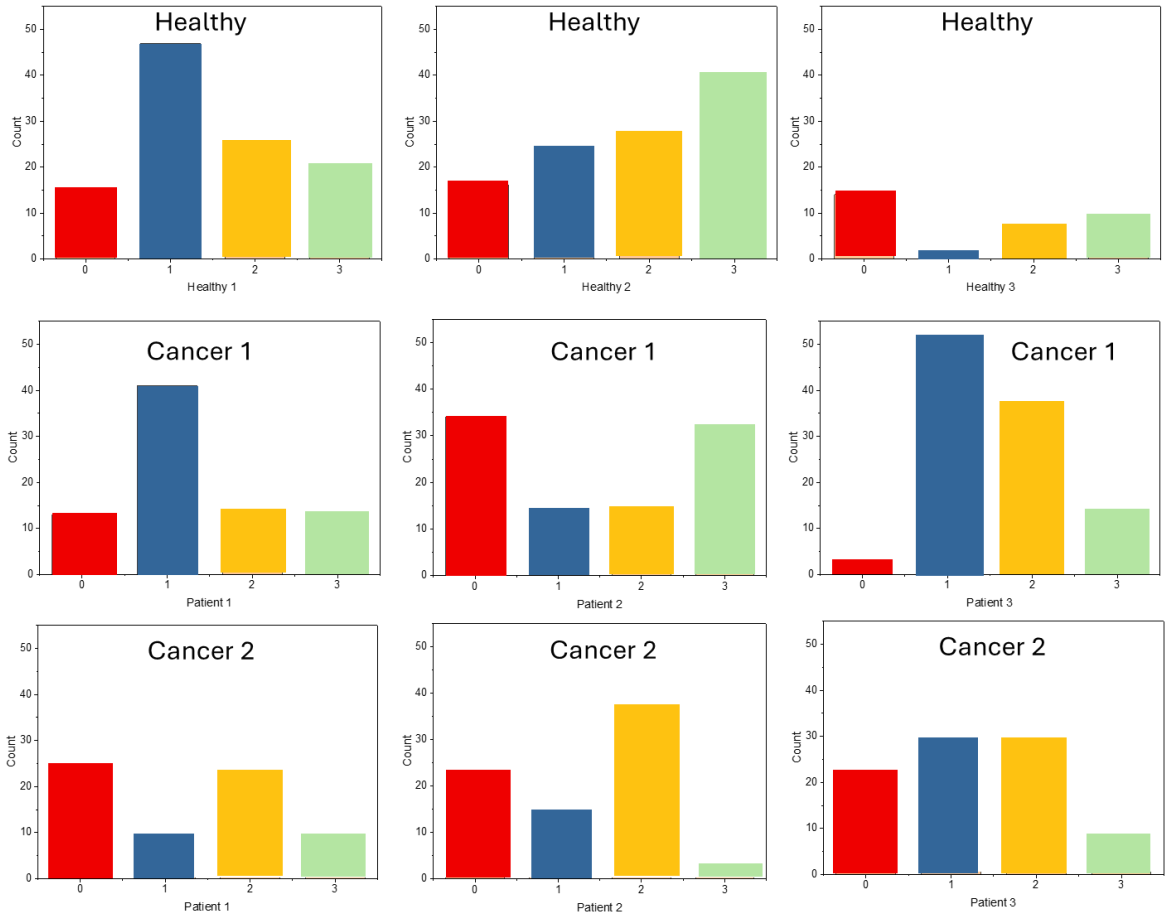

Fig S4: K-means clustering with 4 clusters for selected Cancer 1 and Cancer 2 patients, and healthy controls.

4 Comprehensive discussion of findings

4.1 Single EV Technologies

EVs play crucial roles in mediating cell-to-cell communication both locally and at a distance, given their capacity to navigate the extracellular matrix where most cells are embedded [92]. This discovery has also challenged the long-held hypothesis that direct cell-to-cell contact is the primary communication method between cancer cells and surrounding stromal cells, highlighting the broader and more versatile roles that EVs play in this complex network [93], [94]. However, the majority of currently available EV characterization techniques rely on bulk analysis, where the signals obtained are averaged across a population of EVs. This approach, while practical, overlooks the variations among individual vesicles and was historically considered necessary due to the nanoscale size of EVs, but presents significant technical difficulties for single-vesicle profiling [24], [95], [96].

The highly heterogeneous nature of EVs, both in terms of their biophysical properties and molecular composition, further complicates the bulk analysis, as multiple biomarkers are often needed to capture relevant molecules of interest. This reliance on multiple markers can dilute both the sensitivity and specificity of diagnostic results, as the signals may become overwhelmed by background noise or irrelevant information. Additionally, assays are frequently constrained by the current state of biomarker discovery, leaving many EV subtypes underexplored or entirely unstudied. The scarcity of tumor-shed EVs carrying critical disease-indicating biomarkers amidst a vast population of irrelevant healthy EVs shows the need for more precise analytical techniques. This makes sEV analysis a critical advancement for early diagnosis and disease progression monitoring. The precise identification of interested vesicles may significantly increasing the likelihood of detecting the faint yet crucial signals that indicate the presence of disease. Single EV profiling technology holds great potential to be the key to a more thorough understanding of the biological functions and pathways of previously undiscovered EV subtypes, addressing technological restrictions in disease development. Recent studies have reported several single-EV analysis techniques using bead-based microwell arrays combined with tyramide signal amplification droplet-based digital assays [97], or single molecule imaging achieved via total

internal reflection fluorescence microscopy [98], [99]. These developments have highlighted the exciting potential of sEV analysis to unlock new frontiers in EV research and diagnostic applications. However, when it comes to plasmonic platforms, the field of sEV profiling remains relatively understudied due to challenges associated with both single EV entrapment and signal amplification, which are critical to the success of nanofabrication techniques. Despite these challenges, the miniature size, ease of device setup, and the highly adaptable nature of plasmonic surface designs present significant opportunities for their future application. These platforms could eventually be integrated into clinical settings to facilitate more accurate and rapid diagnostics. Beyond the recently developed MoSERS platform in the Mahshid lab [37], recent innovations from the Im group have introduced periodic gold nanowell structures that utilize plasmon-enhanced fluorescence detection, capable of profiling EVs with single-vesicle resolution [27]. This demonstrates the growing interest in plasmonic technologies and their potential for advancing EV research.

Despite the significant potential, label-free single EV SERS platforms, like MoSERS, face several challenges in their integration into routine cancer diagnostics. One of the most prominent obstacles is the construction of a comprehensive spectral library that can reliably identify key biomarkers. Even though plasma samples are concentrated using size-exclusion chromatography (SEC), the overall EV concentration in solution often remains relatively low, and high-resolution spectra are not guaranteed. Additionally, the label-free nature of the technique introduces another layer of complexity, as it provides no prior information about the source of the detected signal. While it may be possible to distinguish between patient-derived and healthy control samples, there is no definitive assurance that a given signal carries a tumor-specific EV signature, and the natural heterogeneity present even among healthy EVs, further complicates this process. These factors contribute to the slower development and clinical translation of label-free technologies, despite their clear advantages in non-invasive diagnostics. Potential solutions to these challenges will be discussed in greater detail in the subsequent ***'Next Steps'*** section.

4.2 Choice of Step Size in Dimensionality Reduction and ML Model Optimization

The choice of step size, often referred to as the learning rate, is a critical parameter that significantly influences the performance of ML models. It dictates how much the model updates its parameters during each iteration as it works to minimize the loss function, directly affecting how the model navigates the complex, high-dimensional data space[100], [101]. In ML models, it is usually based on gradient descent method or the steepest descent algorithm [102]. In the context of dimensionality reduction, the learning rate can be interpreted as the degree of reduction applied to the original dataset and the step size used for retaining relevant data points. A well-tuned step size ensures efficient capture of essential features while discarding irrelevant information. Finding this optimal balance is especially important for spectral data, where subtle differences at specific wavelengths are key for accurate classification.

The selection of the ML model is not the primary focus of this study, as it has been extensively discussed in previous works [37], [103]. Simpler models, including Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM), were tested on the same dataset but yielded unsatisfactory results. Given the frequent application of deep learning, particularly CNNs, in analyzing high-dimensional Raman spectroscopy data, a ResNet-based CNN was adapted from the work of Ho et al. as the foundational architecture for this project [90].

Section 3 focused on comparing the effects of dimensionality reduction techniques on improving binary classification accuracy. However, the role of step size during this process may be further investigated. In Section Section3, the dimensionality was greatly reduced to approximately half of the original set, preserving around 1400 data points for each of the four techniques. This number was determined based on the resampling method, where each sample was filtered based on the nearest matching wavelength compared to a predefined set of wavelengths. As 1245 data points were ultimately preserved, the step sizes of the other three techniques were adjusted to match this number, allowing for a more uniform comparison of the effectiveness of the techniques. Nonetheless, careful consideration should be given to the selection of step size. If set too high, the model risks overshooting the optimal solution, bypassing the most accurate reduced-

dimensional representation of the data. Conversely, a small step size can slow convergence and increase the number of training iterations needed to find the optimal solution. This is particularly important in dimensionality reduction, as it is often a precedent step prior to the actual ML model. Effective tuning of the step size requires experimentation within a set range. Although this is usually not explicitly mentioned, previous studies has applied similar methods for finding their optimal step size [104], [105].

4.3 Evaluation Metric Selection for Single EV Data with High Spectral Heterogeneity

As this study was based on single-EV spectra, there is a critical need to develop novel analysis methods and adopt more appropriate evaluation metrics tailored to the complexity of EV compositions. Traditional approaches usually assess accuracy on a per-spectrum basis using the confusion matrix (**Figure 4.1**). While this is convenient for initial evaluations, it may fail to capture the heterogeneous nature of EVs since each spectrum represents only a small snapshot of the overall EV population of a sample. Thus, relying on per-spectrum classification may lead to misleading results, particularly in terms of false negatives, where a single misclassified spectrum could disproportionately affect the overall accuracy.

Future work should consider evaluating accuracy on a per-sample basis by considering a contributed result from each spectrum in that sample. This shift in perspective recognizes that not every signal within a sample needs to be perfectly classified as a "patient" signal to yield clinically valuable insights. It may also better reflect the inherent biological variability of EV populations, accounting for the subtle differences and overlapping features that exist between healthy and cancerous samples. By focusing on the combined accuracy of entire samples, this approach can provide a more realistic and comprehensive understanding of the model's performance. Such an evaluation would not only help in reducing false negatives but also offer a clearer indication of the model's diagnostic potential in real-world clinical settings, where variability within the population is expected.

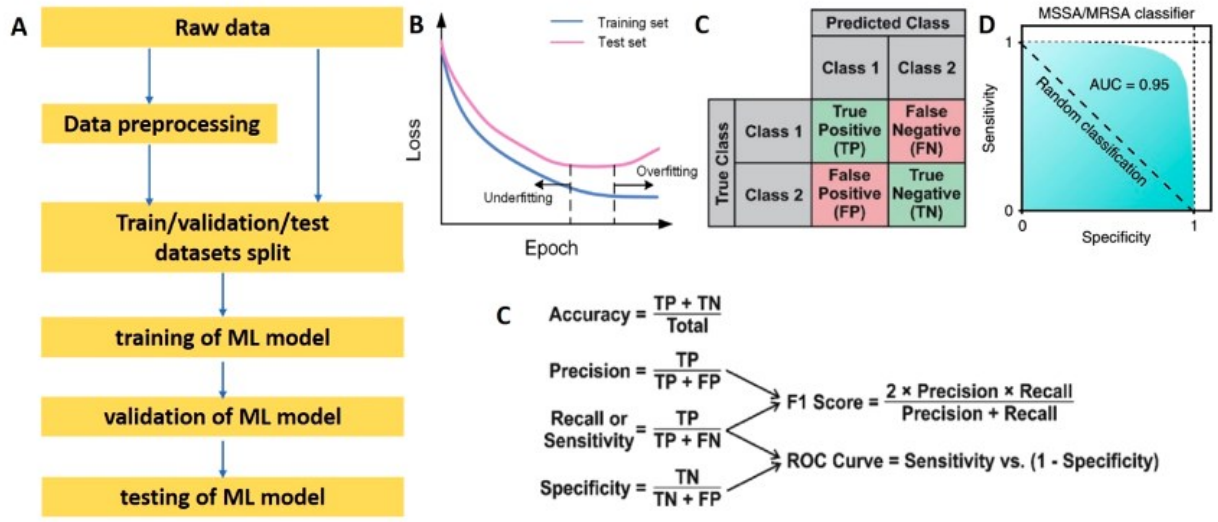

Figure 4.1: Common evaluation metrics for ML-based spectra analysis.

(A) General workflow of ML-based data analysis. **(B)** Loss curve in the training stage. Overfitting will occur when the loss of the test set increases. **(C)** The representation of a confusion matrix represents the prediction summary in matrix form, where the true positive and true negatives on the diagonal squares. **(D)** Representation of a ROC curve, which shows the performance of a classification model at all classification thresholds. AUC can be used as a metric to assess the performance of a classification model. Reproduced with permission from ref [48]. Copyright 2020, ACS Sensors.

4.4 Spectral Data Pre-Processing

As briefly discussed in Section 3, the spectral data processing and analysis steps are often highly specific to the device and data type, with customized workflows designed for each project. Common preprocessing techniques include denoising, baseline drift elimination, Fourier transform, and normalization [106], [107], [108]. These preprocessing steps have a direct impact on the overall performance of ML models, as they shape the quality and consistency of the data fed into the algorithms. For SERS spectral data, each spectrum typically undergoes smoothing using the Savitzky–Golay filter to reduce noise, followed by background subtraction to eliminate unwanted signals, and min-max normalization, which scales the data between 0 and 1 for better comparability across spectra [45], [48], [109]. This workflow is also reflected in Section 2 and Section 3.2, where the reviewed studies generally follow these preprocessing steps, though the specific details regarding the order and method of implementation are often not mentioned. This lack of transparency can hinder the reproducibility and transferability of findings across studies. While this variability can partly be attributed to the unique characteristics of each study, device,

and experimental setup, it also highlights the need for more standardized and generalized data-cleaning methods in the field. Given that SERS-based bioanalysis is still in its developmental stage, establishing clearer protocols would not only enhance reproducibility but also lead to broader adoption of these techniques across various research and clinical applications.

4.5 Next Steps

Ultimately, the envisioned use of the MoSERS platform in clinical settings is to serve as a supplementary tool for cancer diagnosis, rather than as a standalone diagnostic standard. Its primary role would be to screen patients suspected of having early-stage tumors that are difficult to detect using conventional imaging or biopsy techniques. Thus, future studies will need to involve much larger sample pools, as well as extend beyond binary classification to investigate cancer subtypes and metastatic stages. Incorporating established cell lines into the spectral library may also be valuable addition, as they offer a more consistent and prominent model of patient signatures. This would enhance the reproducibility of results and provide a more reliable reference for comparison. This expansion may eventually provide a more comprehensive understanding of the platform's diagnostic potential in diverse clinical scenarios. Another potential modification to maintain the platform's label-free nature would be the integration of an EV concentration step prior of adding the sample. By targeting commonly shared surface biomarkers such as CD9, CD63, and CD81, this step could help eliminate lipoproteins, which are often present in large quantities in samples and may interfere with accurate detection [110], [111].

Furthermore, as we were challenged by the complex sample compositions where the specific EV compositions may be difficult to control, unsupervised learning could be the key for interpreting these unknown mixtures. More advanced models like variational autoencoders could offer promising solutions, enabling deeper insights into the underlying structures and distributions within the data [112]. By leveraging these techniques, we can improve the platform's ability to classify and analyze heterogeneous EV populations, ultimately making it more adaptable and effective in clinical applications.

5 Final Conclusion and Summary

In conclusion, the integration of label-free SERS platforms and ML-assisted spectral data analysis presents a powerful approach for enhancing EV characterization, particularly in cancer diagnostics. ML enables the processing of complex, high-dimensional spectral data, improving the interpretation of underlying biological patterns and facilitating meaningful conclusions from even noisy data. This study highlighted the importance of optimizing data processing workflows and ensuring class representation balance in training sets to improve model robustness and performance. The resampling technique was particularly effective in addressing imbalanced datasets, demonstrating significant improvements in classification accuracy and potential transferability to other cancer paradigms. Moreover, these findings emphasize the critical role of both feature selection and model optimization in enhancing the accuracy of deep learning models.

For future studies, expanding beyond binary classification to multiclass scenarios is essential for gaining deeper insights into tumor metastatic states and pushing the boundaries of ML in clinical applications. Future research should also focus on increasing sample sizes, as larger datasets are crucial for improving model generalization and reducing overfitting. Additionally, developing more standardized data analysis protocols will promote reproducibility and facilitate broader application of these techniques across different studies and clinical settings. By continuing to refine these technologies, the combination of optical sensors and ML holds great promise for revolutionizing EV-based diagnostics, ultimately improving the detection early-stage cancer in an efficient and cost-effective manner.

Master Bibliography

[1] "Innovation – Mahshid Research Group." Accessed: Nov. 30, 2024. [Online]. Available: https://mahshidlab.com/innovation/

[2] S. Cheng *et al.*, "Advances in microfluidic extracellular vesicle analysis for cancer diagnostics," *Lab. Chip*, vol. 21, no. 17, pp. 3219–3243, Aug. 2021, doi: 10.1039/D1LC00443C.

[3] L. M. Doyle and M. Z. Wang, "Overview of Extracellular Vesicles, Their Origin, Composition, Purpose, and Methods for Exosome Isolation and Analysis," *Cells*, vol. 8, no. 7, p. 727, Jul. 2019, doi: 10.3390/cells8070727.

[4] M. LeClaire, J. Gimzewski, and S. Sharma, "A review of the biomechanical properties of single extracellular vesicles," *Nano Sel.*, vol. 2, no. 1, pp. 1–15, 2021, doi: 10.1002/nano.202000129.

[5] H. Shin, D. Seo, and Y. Choi, "Extracellular Vesicle Identification Using Label-Free Surface-Enhanced Raman Spectroscopy: Detection and Signal Analysis Strategies," *Molecules*, vol. 25, no. 21, Art. no. 21, Jan. 2020, doi: 10.3390/molecules25215209.

[6] S. EL Andaloussi, I. Mäger, X. O. Breakefield, and M. J. A. Wood, "Extracellular vesicles: biology and emerging therapeutic opportunities," *Nat. Rev. Drug Discov.*, vol. 12, no. 5, Art. no. 5, May 2013, doi: 10.1038/nrd3978.

[7] R. Kalluri and V. S. LeBleu, "The biology, function, and biomedical applications of exosomes," *Science*, vol. 367, no. 6478, p. eaau6977, Feb. 2020, doi: 10.1126/science.aau6977.

[8] G. Li, W. Tang, and F. Yang, "Cancer Liquid Biopsy Using Integrated Microfluidic Exosome Analysis Platforms," *Biotechnol. J.*, vol. 15, no. 5, p. 1900225, 2020, doi: 10.1002/biot.201900225.

[9] M. Imanbekova, S. Suarasan, Y. Lu, S. Jurchuk, and S. Wachsmann-Hogiu, "Recent advances in optical label-free characterization of extracellular vesicles," *Nanophotonics*, vol. 11, no. 12, pp. 2827–2863, Jun. 2022, doi: 10.1515/nanoph-2022-0057.

[10] C. del Real Mata, O. Jeanne, M. Jalali, Y. Lu, and S. Mahshid, "Nanostructured-Based Optical Readouts Interfaced with Machine Learning for Identification of Extracellular Vesicles," *Adv. Healthc. Mater.*, vol. 12, no. 5, p. 2202123, 2023, doi: 10.1002/adhm.202202123.

[11] K. P. De Sousa, I. Rossi, M. Abdullahi, M. I. Ramirez, D. Stratton, and J. M. Inal, "Isolation and characterization of extracellular vesicles and future directions in diagnosis and therapy,"

WIREs Nanomedicine Nanobiotechnology, vol. 15, no. 1, p. e1835, 2023, doi: 10.1002/wnan.1835.

[12] M. He and Y. Zeng, "Microfluidic Exosome Analysis toward Liquid Biopsy for Cancer," *J. Lab. Autom.*, vol. 21, no. 4, pp. 599–608, Aug. 2016, doi: 10.1177/2211068216651035.

[13] S. M. Mousavi *et al.*, "Microfluidics for detection of exosomes and microRNAs in cancer: State of the art," *Mol. Ther. Nucleic Acids*, vol. 28, pp. 758–791, Apr. 2022, doi: 10.1016/j.omtn.2022.04.011.

[14] C. Théry *et al.*, "Minimal information for studies of extracellular vesicles 2018 (MISEV2018): a position statement of the International Society for Extracellular Vesicles and update of the MISEV2014 guidelines," *J. Extracell. Vesicles*, vol. 7, no. 1, p. 1535750, Dec. 2018, doi: 10.1080/20013078.2018.1535750.

[15] C. Gardiner *et al.*, "Techniques used for the isolation and characterization of extracellular vesicles: results of a worldwide survey," *J. Extracell. Vesicles*, vol. 5, p. 10.3402/jev.v5.32945, Oct. 2016, doi: 10.3402/jev.v5.32945.

[16] A. Meggiolaro *et al.*, "Microfluidic Strategies for Extracellular Vesicle Isolation: Towards Clinical Applications," *Biosensors*, vol. 13, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/bios13010050.

[17] Y.-T. Chen *et al.*, "Review of Integrated Optical Biosensors for Point-of-Care Applications," *Biosensors*, vol. 10, no. 12, Art. no. 12, Dec. 2020, doi: 10.3390/bios10120209.

[18] S. Gholizadeh *et al.*, "Microfluidic approaches for isolation, detection, and characterization of extracellular vesicles: Current status and future directions," *Biosens. Bioelectron.*, vol. 91, pp. 588–605, May 2017, doi: 10.1016/j.bios.2016.12.062.

[19] M. T. Yaraki, A. Tukova, and Y. Wang, "Emerging SERS biosensors for the analysis of cells and extracellular vesicles," *Nanoscale*, vol. 14, no. 41, pp. 15242–15268, Oct. 2022, doi: 10.1039/D2NR03005E.

[20] M. Piliarik, H. Vaisocherová, and J. Homola, "Surface Plasmon Resonance Biosensing," in *Biosensors and Biodetection*, A. Rasooly and K. E. Herold, Eds., in Methods in Molecular Biology™. , Totowa, NJ: Humana Press, 2009, pp. 65–88. doi: 10.1007/978-1-60327-567-5_5.

[21] Y. Tang, X. Zeng, and J. Liang, "Surface Plasmon Resonance: An Introduction to a Surface Spectroscopy Technique," *J. Chem. Educ.*, vol. 87, no. 7, pp. 742–746, Jul. 2010, doi: 10.1021/ed100186y.

[22] J. Roberto, K. L. Poulin, R. J. Parks, and P. O. Vacratsis, "Label-free quantitative proteomic analysis of extracellular vesicles released from fibroblasts derived from patients with spinal

muscular atrophy," *PROTEOMICS*, vol. 21, no. 13–14, p. 2000301, 2021, doi: 10.1002/pmic.202000301.

[23] V. Yesudasu, H. S. Pradhan, and R. J. Pandya, "Recent progress in surface plasmon resonance based sensors: A comprehensive review," *Heliyon*, vol. 7, no. 3, p. e06321, Mar. 2021, doi: 10.1016/j.heliyon.2021.e06321.

[24] L. K. Chin *et al.*, "Plasmonic Sensors for Extracellular Vesicle Analysis: From Scientific Development to Translational Research," *ACS Nano*, vol. 14, no. 11, pp. 14528–14548, Nov. 2020, doi: 10.1021/acsnano.0c07581.

[25] D. Raju, S. Bathini, S. Badilescu, R. J. Ouellette, A. Ghosh, and M. Packirisamy, "LSPR detection of extracellular vesicles using a silver-PDMS nano-composite platform suitable for sensor networks," *Enterp. Inf. Syst.*, vol. 14, no. 4, pp. 532–541, Apr. 2020, doi: 10.1080/17517575.2018.1526326.

[26] H. Im *et al.*, "Label-free detection and molecular profiling of exosomes with a nano-plasmonic sensor," *Nat. Biotechnol.*, vol. 32, no. 5, pp. 490–495, May 2014, doi: 10.1038/nbt.2886.

[27] M. H. Jeong *et al.*, "Plasmon-Enhanced Single Extracellular Vesicle Analysis for Cholangiocarcinoma Diagnosis," *Adv. Sci.*, vol. 10, no. 8, p. 2205148, Jan. 2023, doi: 10.1002/advs.202205148.

[28] L. Zhu *et al.*, "Label-Free Quantitative Detection of Tumor-Derived Exosomes through Surface Plasmon Resonance Imaging," *Anal. Chem.*, vol. 86, no. 17, pp. 8857–8864, Sep. 2014, doi: 10.1021/ac5023056.

[29] Z. Han *et al.*, "Integrated microfluidic-SERS for exosome biomarker profiling and osteosarcoma diagnosis," *Biosens. Bioelectron.*, vol. 217, p. 114709, Dec. 2022, doi: 10.1016/j.bios.2022.114709.

[30] A. Orlando *et al.*, "A Comprehensive Review on Raman Spectroscopy Applications," *Chemosensors*, vol. 9, no. 9, Art. no. 9, Sep. 2021, doi: 10.3390/chemosensors9090262.

[31] Y. Zhang, H. Hong, and W. Cai, "Imaging with Raman Spectroscopy," *Curr. Pharm. Biotechnol.*, vol. 11, no. 6, pp. 654–661, Sep. 2010.

[32] C. Carlomagno, C. Giannasi, S. Niada, M. Bedoni, A. Gualerzi, and A. T. Brini, "Raman Fingerprint of Extracellular Vesicles and Conditioned Media for the Reproducibility Assessment of Cell-Free Therapeutics," *Front. Bioeng. Biotechnol.*, vol. 9, 2021, Accessed: Apr. 14, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbioe.2021.640617

[33] S. Kumar *et al.*, *Surface-Enhanced Raman Scattering: Introduction and Applications*. IntechOpen, 2020. doi: 10.5772/intechopen.92614.

[34] A. I. Pérez-Jiménez, D. Lyu, Z. Lu, G. Liu, and B. Ren, "Surface-enhanced Raman spectroscopy: benefits, trade-offs and future developments," *Chem. Sci.*, vol. 11, no. 18, pp. 4563–4577, May 2020, doi: 10.1039/D0SC00809E.

[35] "A Review on Surface-Enhanced Raman Scattering - PMC." Accessed: Apr. 15, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6627380/

[36] M. Jalali *et al.*, "Plasmonic nanobowtiefluidic device for sensitive detection of glioma extracellular vesicles by Raman spectrometry," *Lab. Chip*, vol. 21, no. 5, pp. 855–866, Mar. 2021, doi: 10.1039/D0LC00957A.

[37] M. Jalali *et al.*, "MoS2-Plasmonic Nanocavities for Raman Spectra of Single Extracellular Vesicles Reveal Molecular Progression in Glioblastoma," *ACS Nano*, vol. 17, no. 13, pp. 12052–12071, Jul. 2023, doi: 10.1021/acsnano.2c09222.

[38] Y. S. Rim, "Review of metal oxide semiconductors-based thin-film transistors for point-of-care sensor applications," *J. Inf. Disp.*, vol. 21, no. 4, pp. 203–210, Oct. 2020, doi: 10.1080/15980316.2020.1714762.

[39] I. Şerban and A. Enesca, "Metal Oxides-Based Semiconductors for Biosensors Applications," *Front. Chem.*, vol. 8, p. 354, May 2020, doi: 10.3389/fchem.2020.00354.

[40] B. Y. Zhang *et al.*, "Highly accurate and label-free discrimination of single cancer cell using a plasmonic oxide-based nanoprobe," *Biosens. Bioelectron.*, vol. 198, p. 113814, Feb. 2022, doi: 10.1016/j.bios.2021.113814.

[41] Y. Jahani *et al.*, "Imaging-based spectrometer-less optofluidic biosensors based on dielectric metasurfaces for detecting extracellular vesicles," *Nat. Commun.*, vol. 12, no. 1, Art. no. 1, May 2021, doi: 10.1038/s41467-021-23257-y.

[42] Y. Wang, W. Yuan, M. Kimber, M. Lu, and L. Dong, "Rapid Differentiation of Host and Parasitic Exosome Vesicles Using Microfluidic Photonic Crystal Biosensor," *ACS Sens.*, vol. 3, no. 9, pp. 1616–1621, Sep. 2018, doi: 10.1021/acssensors.8b00360.

[43] C. Hong and J. C. Ndukaife, "Scalable trapping of single nanosized extracellular vesicles using plasmonics," *Nat. Commun.*, vol. 14, no. 1, p. 4801, Aug. 2023, doi: 10.1038/s41467-023-40549-7.

[44] "Machine Learning-Reinforced Noninvasive Biosensors for Healthcare - Zhang - 2021 - Advanced Healthcare Materials - Wiley Online Library." Accessed: Jul. 09, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/adhm.202100734

[45] X. Bi, L. Lin, Z. Chen, and J. Ye, "Artificial Intelligence for Surface-Enhanced Raman Spectroscopy," *Small Methods*, vol. 8, no. 1, p. 2301243, 2024, doi: 10.1002/smtd.202301243.

[46] H. Raji, M. Tayyab, J. Sui, S. R. Mahmoodi, and M. Javanmard, "Biosensors and machine learning for enhanced detection, stratification, and classification of cells: a review," *Biomed. Microdevices*, vol. 24, no. 3, p. 26, Aug. 2022, doi: 10.1007/s10544-022-00627-x.

[47] D. Cakmakci *et al.*, "Machine learning assisted intraoperative assessment of brain tumor margins using HRMAS NMR spectroscopy," *PLoS Comput. Biol.*, vol. 16, no. 11, p. e1008184, Nov. 2020, doi: 10.1371/journal.pcbi.1008184.

[48] F. Cui, Y. Yue, Y. Zhang, Z. Zhang, and H. S. Zhou, "Advancing Biosensors with Machine Learning," *ACS Sens.*, vol. 5, no. 11, pp. 3346–3364, Nov. 2020, doi: 10.1021/acssensors.0c01424.

[49] O. Guselnikova *et al.*, "Label-free surface-enhanced Raman spectroscopy with artificial neural network technique for recognition photoinduced DNA damage," *Biosens. Bioelectron.*, vol. 145, p. 111718, Dec. 2019, doi: 10.1016/j.bios.2019.111718.

[50] M. Erzina *et al.*, "Precise cancer detection via the combination of functionalized SERS surfaces and convolutional neural network with independent inputs," *Sens. Actuators B Chem.*, vol. 308, p. 127660, Apr. 2020, doi: 10.1016/j.snb.2020.127660.

[51] M. Xiao, Z. Liu, N. Xu, L. Jiang, M. Yang, and C. Yi, "A Smartphone-Based Sensing System for On-Site Quantitation of Multiple Heavy Metal Ions Using Fluorescent Carbon Nanodots-Based Microarrays," *ACS Sens.*, vol. 5, no. 3, pp. 870–878, Mar. 2020, doi: 10.1021/acssensors.0c00219.

[52] Y. Liu *et al.*, "Fluorescent Microarrays of in Situ Crystallized Perovskite Nanocomposites Fabricated for Patterned Applications by Using Inkjet Printing," *ACS Nano*, vol. 13, no. 2, pp. 2042–2049, Feb. 2019, doi: 10.1021/acsnano.8b08582.

[53] M. Taniguchi, "Combination of Single-Molecule Electrical Measurements and Machine Learning for the Identification of Single Biomolecules," *ACS Omega*, vol. 5, no. 2, pp. 959–964, Jan. 2020, doi: 10.1021/acsomega.9b03660.

[54] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, in COLT '92. New York, NY, USA: Association for Computing Machinery, Jul. 1992, pp. 144–152. doi: 10.1145/130385.130401.

[55] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," *Methods Mol. Biol. Clifton NJ*, vol. 609, pp. 223–239, 2010, doi: 10.1007/978-1-60327-241-4_13.

[56] S. Winters-Hilt, A. Yelundur, C. McChesney, and M. Landry, "Support Vector Machine Implementations for Classification & Clustering," *BMC Bioinformatics*, vol. 7, no. Suppl 2, p. S4, Sep. 2006, doi: 10.1186/1471-2105-7-S2-S4.

[57] S. HUANG, N. CAI, P. P. PACHECO, S. NARANDES, Y. WANG, and W. XU, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," *Cancer Genomics Proteomics*, vol. 15, no. 1, pp. 41–51, Dec. 2017, doi: 10.21873/cgp.20063.

[58] A. Sharma and K. K. Paliwal, "Linear discriminant analysis for the small sample size problem: an overview," *Int. J. Mach. Learn. Cybern.*, vol. 6, no. 3, pp. 443–454, Jun. 2015, doi: 10.1007/s13042-013-0226-9.

[59] R. Graf, M. Zeldovich, and S. Friedrich, "Comparing linear discriminant analysis and supervised learning algorithms for binary classification—A method comparison study," *Biom. J.*, vol. 66, no. 1, p. 2200098, 2024, doi: 10.1002/bimj.202200098.

[60] D. M. Witten and R. Tibshirani, "Penalized classification using Fisher's linear discriminant," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 73, no. 5, pp. 753–772, Nov. 2011, doi: 10.1111/j.1467-9868.2011.00783.x.

[61] C. Feng, B. Zhao, X. Zhou, X. Ding, and Z. Shan, "An Enhanced Quantum K-Nearest Neighbor Classification Algorithm Based on Polar Distance," *Entropy*, vol. 25, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/e25010127.

[62] H. S. Mondal, K. A. Ahmed, N. Birbilis, and M. Z. Hossain, "Machine learning for detecting DNA attachment on SPR biosensor," *Sci. Rep.*, vol. 13, no. 1, p. 3742, Mar. 2023, doi: 10.1038/s41598-023-29395-1.

[63] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Ann. Transl. Med.*, vol. 4, no. 11, p. 218, Jun. 2016, doi: 10.21037/atm.2016.03.37.

[64] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[65] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Syst. Appl.*, vol. 237, p. 121549, Mar. 2024, doi: 10.1016/j.eswa.2023.121549.

[66] S.-H. Han, K. W. Kim, S. Kim, and Y. C. Youn, "Artificial Neural Network: Understanding the Basic Concepts without Mathematics," *Dement. Neurocognitive Disord.*, vol. 17, no. 3, pp. 83–89, Sep. 2018, doi: 10.12779/dnd.2018.17.3.83.

[67] I. A. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," *J. Microbiol. Methods*, vol. 43, no. 1, pp. 3–31, Dec. 2000, doi: 10.1016/S0167-7012(00)00201-3.

[68] Z. Ballard, C. Brown, A. M. Madni, and A. Ozcan, "Machine learning and computation-enabled intelligent sensor design," *Nat. Mach. Intell.*, vol. 3, no. 7, pp. 556–565, Jul. 2021, doi: 10.1038/s42256-021-00360-9.

[69] W. Lee, J. J. Seong, B. Ozlu, B. S. Shim, A. Marakhimov, and S. Lee, "Biosignal Sensors and Deep Learning-Based Speech Recognition: A Review," *Sensors*, vol. 21, no. 4, Art. no. 4, Jan. 2021, doi: 10.3390/s21041399.

[70] S. Chen and W. Guo, "Auto-Encoders in Deep Learning—A Review with New Perspectives," *Mathematics*, vol. 11, no. 8, Art. no. 8, Jan. 2023, doi: 10.3390/math11081777.

[71] P. Li, Y. Pei, and J. Li, "A comprehensive survey on design and application of autoencoder in deep learning," *Appl. Soft Comput.*, vol. 138, p. 110176, May 2023, doi: 10.1016/j.asoc.2023.110176.

[72] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, "Autoencoders and their applications in machine learning: a survey," *Artif. Intell. Rev.*, vol. 57, no. 2, p. 28, Feb. 2024, doi: 10.1007/s10462-023-10662-6.

[73] W. Lee, A. T. M. Lenferink, C. Otto, and H. L. Offerhaus, "Classifying Raman spectra of extracellular vesicles based on convolutional neural networks for prostate cancer detection," *J. Raman Spectrosc.*, vol. 51, no. 2, pp. 293–300, 2020, doi: 10.1002/jrs.5770.

[74] R. Luo, J. Popp, and T. Bocklitz, "Deep Learning for Raman Spectroscopy: A Review," *Analytica*, vol. 3, no. 3, Art. no. 3, Sep. 2022, doi: 10.3390/analytica3030020.

[75] E. Ryzhikova *et al.*, "Raman spectroscopy of blood serum for Alzheimer's disease diagnostics: specificity relative to other types of dementia," *J. Biophotonics*, vol. 8, no. 7, pp. 584–596, 2015, doi: 10.1002/jbio.201400060.

[76] C. Chen *et al.*, "Rapid diagnosis of lung cancer and glioma based on serum Raman spectroscopy combined with deep learning," *J. Raman Spectrosc.*, vol. 52, no. 11, pp. 1798–1809, 2021, doi: 10.1002/jrs.6224.

[77] C. O. S. Sorzano, J. Vargas, and A. P. Montano, "A survey of dimensionality reduction techniques," Mar. 12, 2014, *arXiv*: arXiv:1403.2877. doi: 10.48550/arXiv.1403.2877.

[78] H. Shin *et al.*, "Single test-based diagnosis of multiple cancer types using Exosome-SERS-AI for early stage cancers," *Nat. Commun.*, vol. 14, no. 1, Art. no. 1, Mar. 2023, doi: 10.1038/s41467-023-37403-1.

[79] C.-C. Xiong *et al.*, "Rapid and precise detection of cancers via label-free SERS and deep learning," *Anal. Bioanal. Chem.*, vol. 415, no. 17, pp. 3449–3462, Jul. 2023, doi: 10.1007/s00216-023-04730-7.

[80] L. M. Wurm *et al.*, "Rapid, label-free classification of glioblastoma differentiation status combining confocal Raman spectroscopy and machine learning," *Analyst*, vol. 148, no. 23, pp. 6109–6119, 2023, doi: 10.1039/D3AN01303K.

[81] U. Parlatan *et al.*, “Label-Free Identification of Exosomes using Raman Spectroscopy and Machine Learning,” *Small*, vol. 19, no. 9, p. 2205519, 2023, doi: 10.1002/smll.202205519.

[82] M. N. Jensen *et al.*, “Identification of extracellular vesicles from their Raman spectra via self-supervised learning,” *Sci. Rep.*, vol. 14, no. 1, p. 6791, Mar. 2024, doi: 10.1038/s41598-024-56788-7.

[83] M. Kazemzadeh *et al.*, “Classification of Preeclamptic Placental Extracellular Vesicles Using Femtosecond Laser Fabricated Nanoplasmonic Sensors,” *ACS Sens.*, vol. 7, no. 6, pp. 1698–1711, Jun. 2022, doi: 10.1021/acssensors.2c00378.

[84] G. T. Reddy *et al.*, “Analysis of Dimensionality Reduction Techniques on Big Data,” *IEEE Access*, vol. 8, pp. 54776–54788, 2020, doi: 10.1109/ACCESS.2020.2980942.

[85] M. F. Kabir, T. Chen, and S. A. Ludwig, “A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction,” *Healthc. Anal.*, vol. 3, p. 100125, Nov. 2023, doi: 10.1016/j.health.2022.100125.

[86] S. Velliangiri, S. Alagumuthukrishnan, and S. I. Thankumar joseph, “A Review of Dimensionality Reduction Techniques for Efficient Computation,” *Procedia Comput. Sci.*, vol. 165, pp. 104–111, Jan. 2019, doi: 10.1016/j.procs.2020.01.079.

[87] I. Romanishkin *et al.*, “Differentiation of glioblastoma tissues using spontaneous Raman scattering with dimensionality reduction and data classification,” *Front. Oncol.*, vol. 12, Sep. 2022, doi: 10.3389/fonc.2022.944210.

[88] W. Schumacher, S. Stöckel, P. Rösch, and J. Popp, “Improving chemometric results by optimizing the dimension reduction for Raman spectral data sets,” *J. Raman Spectrosc.*, vol. 45, no. 10, pp. 930–940, 2014, doi: 10.1002/jrs.4568.

[89] J. Y. Choi, “Medulloblastoma: Current Perspectives and Recent Advances,” *Brain Tumor Res. Treat.*, vol. 11, no. 1, pp. 28–38, Jan. 2023, doi: 10.14791/btrt.2022.0046.

[90] C.-S. Ho *et al.*, “Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning,” *Nat. Commun.*, vol. 10, p. 4927, Oct. 2019, doi: 10.1038/s41467-019-12898-9.

[91] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” Dec. 03, 2019, *arXiv*: arXiv:1912.01703. doi: 10.48550/arXiv.1912.01703.

[92] G. van Niel, D. R. F. Carter, A. Clayton, D. W. Lambert, G. Raposo, and P. Vader, “Challenges and directions in studying cell–cell communication by extracellular vesicles,” *Nat. Rev. Mol. Cell Biol.*, vol. 23, no. 5, pp. 369–382, May 2022, doi: 10.1038/s41580-022-00460-3.

[93] W.-H. Chang, R. A. Cerione, and M. A. Antonyak, "Extracellular Vesicles and Their Roles in Cancer Progression," in *Cancer Cell Signaling: Methods and Protocols*, M. Robles-Flores, Ed., New York, NY: Springer US, 2021, pp. 143–170. doi: 10.1007/978-1-0716-0759-6_10.

[94] A. Möller and R. J. Lobb, "The evolving translational potential of small extracellular vesicles in cancer," *Nat. Rev. Cancer*, vol. 20, no. 12, pp. 697–709, Dec. 2020, doi: 10.1038/s41568-020-00299-w.

[95] H. Shao, H. Im, C. M. Castro, X. Breakefield, R. Weissleder, and H. Lee, "New Technologies for Analysis of Extracellular Vesicles," *Chem. Rev.*, vol. 118, no. 4, pp. 1917–1950, Feb. 2018, doi: 10.1021/acs.chemrev.7b00534.

[96] S. Ferguson, K. S. Yang, and R. Weissleder, "Single extracellular vesicle analysis for early cancer detection," *Trends Mol. Med.*, vol. 28, no. 8, pp. 681–692, Aug. 2022, doi: 10.1016/j.molmed.2022.05.003.

[97] S. Ansaryan *et al.*, "High-throughput spatiotemporal monitoring of single-cell secretions via plasmonic microwell arrays," *Nat. Biomed. Eng.*, vol. 7, no. 7, Art. no. 7, Jul. 2023, doi: 10.1038/s41551-023-01017-1.

[98] G. Wu *et al.*, "Single-cell extracellular vesicle analysis by microfluidics and beyond," *TrAC Trends Anal. Chem.*, vol. 159, p. 116930, Feb. 2023, doi: 10.1016/j.trac.2023.116930.

[99] S. H. Hilton and I. M. White, "Advances in the analysis of single extracellular vesicles: A critical review," *Sens. Actuators Rep.*, vol. 3, p. 100052, Nov. 2021, doi: 10.1016/j.snr.2021.100052.

[100] S. Sun, Z. Cao, H. Zhu, and J. Zhao, "A Survey of Optimization Methods From a Machine Learning Perspective," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3668–3681, Aug. 2020, doi: 10.1109/TCYB.2019.2950779.

[101] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, "Don't Decay the Learning Rate, Increase the Batch Size," arXiv.org. Accessed: Sep. 07, 2024. [Online]. Available: https://arxiv.org/abs/1711.00489v2

[102] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," arXiv.org. Accessed: Sep. 07, 2024. [Online]. Available: https://arxiv.org/abs/1611.03530v2

[103] O. Jeanne, "Machine learning as a tool to analyse data from plasmonic sensors for medical diagnostics." Accessed: Dec. 08, 2024. [Online]. Available: https://escholarship.mcgill.ca/concern/theses/0g354m91t

[104] Q. Tong, G. Liang, and J. Bi, “Calibrating the adaptive learning rate to improve convergence of ADAM,” *Neurocomputing*, vol. 481, pp. 333–356, Apr. 2022, doi: 10.1016/j.neucom.2022.01.014.

[105] J. M. Ede and R. Beanland, “Adaptive learning rate clipping stabilizes learning,” *Mach. Learn. Sci. Technol.*, vol. 1, no. 1, p. 015011, Apr. 2020, doi: 10.1088/2632-2153/ab81e2.

[106] H. Qian *et al.*, “Diagnosis of urogenital cancer combining deep learning algorithms and surface-enhanced Raman spectroscopy based on small extracellular vesicles,” *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.*, vol. 281, p. 121603, Nov. 2022, doi: 10.1016/j.saa.2022.121603.

[107] F. Ahmadi *et al.*, “Integrating machine learning and digital microfluidics for screening experimental conditions,” *Lab. Chip*, vol. 23, no. 1, pp. 81–91, 2023, doi: 10.1039/D2LC00764A.

[108] K. J. Kobayashi-Kirschvink *et al.*, “Prediction of single-cell RNA expression profiles in live cells by Raman microscopy with Raman2RNA,” *Nat. Biotechnol.*, pp. 1–9, Jan. 2024, doi: 10.1038/s41587-023-02082-2.

[109] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, “A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science,” in *Supervised and Unsupervised Learning for Data Science*, M. W. Berry, A. Mohamed, and B. W. Yap, Eds., Cham: Springer International Publishing, 2020, pp. 3–21. doi: 10.1007/978-3-030-22475-2_1.

[110] J. D. Spitzberg, S. Ferguson, K. S. Yang, H. M. Peterson, J. C. T. Carlson, and R. Weissleder, “Multiplexed analysis of EV reveals specific biomarker composition with diagnostic impact,” *Nat. Commun.*, vol. 14, no. 1, p. 1239, Mar. 2023, doi: 10.1038/s41467-023-36932-z.

[111] K. Ekström, R. Crescitelli, H. I. Pétursson, J. Johansson, C. Lässer, and R. Olofsson Bagge, “Characterization of surface markers on extracellular vesicles isolated from lymphatic exudate from patients with breast cancer,” *BMC Cancer*, vol. 22, no. 1, p. 50, Jan. 2022, doi: 10.1186/s12885-021-08870-w.

[112] D. P. Kingma and M. Welling, “An Introduction to Variational Autoencoders,” *Found. Trends® Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019, doi: 10.1561/2200000056.