Machine Learning as a Tool to Analyze Data from Plasmonic Sensors for Medical Diagnostics

Olivia Jeanne

Master of Engineering

Department of Bioengineering

McGill University

Montréal, Québec, Canada

July 2022

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

Abstract-English

Machine learning has great potential to overcome the complexity and heterogeneity of biological data. The development of high-throughput multiplexed sensors now allows for the acquisition of various biological data in ever-increasing quantities. As a result, more precise diagnostic and monitoring of disease can be achieved. However, the data needs to be processed and sifted in order to find the most relevant features within the data, which poses a huge challenge for humans, due to both the amount of data available, and its heterogeneous and complex nature. The development of data science and machine learning can help scientists and clinicians to interpret the data, and make medical diagnostic and monitoring easier, high-throughput and user friendly, with potential to act as a first, low-cost point of triage to screen patients, in settings where high-end medical facilities and machines are in limited supply. Machine learning algorithms are incredibly versatile and can be used with multiple data types, generated by very different sensors. Here, a type of supervised machine learning algorithm called Support Vector Machines (SVM) is applied to analyze the data from two optical biosensors. The first one is a nanostructured Surface Enhanced Raman Spectroscopy (SERS) platform, which allows entrapment of sub-cellular sized biological material, such as Extracellular Vesicles (EVs) secreted by cells. SERS is an optical characterization method providing information on the molecular composition of analytes. The main challenge SERS faces as a readout is the often subtle changes in the spectra generated by heterogeneous biological samples, which makes traditional analysis difficult and time-consuming. In this work, EVs derived from Glioblastoma Multiforme (GBM) cell lines, liquid biopsies of GBM patients, and controls are studied. SERS is performed on the individual EVs, and the spectra thus collected generate a database of single-EV spectra. These spectra are then analyzed by the SVM, which is used to classify them into the correct mutated cell line. This is successfully done, with an accuracy of 70.04% on the individual mutations. Then, SVM is applied to determine whether

or not it is possible to differentiate GBM patients from healthy controls, based on the analysis of the EV-derived SERS spectra, yielding 95% accuracy in classification. The second sensor used in combination with Machine Learning is a colorimetric platform for rapid and point-of-care viral detection. Clinical samples are collected and subjected to on-chip reverse transcriptase loop-mediated isothermal amplification (RT-LAMP), generating a color change in the presence of viral genetic material. In order to obtain a quick and reliable diagnostic, images of the colorimetric device are collected and analyzed via SVM, reaching a success rate in the classification of healthy vs sick patients of 94% after a 10-minute incubation time, enabling fast screening operations to take place.

Keywords: Machine Learning, SVM, Plasmonics, Diagnostics, Sensors

Abstract-Français

L’apprentissage automatique a un grand potentiel pour surmonter la complexité et l'hétérogénéité des données biologiques. Le développement de capteurs multiplexés à haut débit permet aujourd'hui l'acquisition de données biologiques variées en quantité toujours croissante. En conséquence, un diagnostic et un suivi plus précis de la maladie peuvent être réalisés. Cependant, les données doivent être traitées et passées au crible afin de trouver les caractéristiques les plus pertinentes et de déterminer les relations au sein des données, ce qui représente un énorme défi pour les humains, en raison à la fois de la quantité de données disponibles et de leur nature hétérogène et complexe. Le développement de la science des données et de l'apprentissage automatique peut aider les scientifiques et les praticiens de santé à interpréter les données et à rendre le diagnostic et le suivi médical plus facile pour les utilisateurs, tout en ayant un débit important, avec le potentiel d'agir comme un point de triage à faible coût pour dépister les patients, dans des environnements où les installations médicales

et les machines haut de gamme sont rares. Les algorithmes d'apprentissage automatique sont incroyablement polyvalents et peuvent être utilisés avec plusieurs types de données, générés par des capteurs très différents. Ici, un type d'algorithme d'apprentissage automatique supervisé appelé Support Vector Machine (SVM) est appliqué pour analyser les données de deux biocapteurs optiques. La première est une plateforme nanostructurée de spectroscopie Raman améliorée de surface (SERS), qui permet le piégeage de matériel biologique de taille subcellulaire, comme les Vésicules Extracellulaires (VE) sécrétées par les cellules. La SERS est une méthode de caractérisation optique fournissant des informations sur la composition moléculaire des analytes. Le principal défi auquel SERS est confronté est les changements souvent subtils dans les spectres générés par des échantillons biologiques hétérogènes, ce qui rend l'analyse traditionnellement difficile et chronophage. La SERS est effectué sur les VE individuelles, et les spectres ainsi collectés génèrent une base de données sur le cancer du glioblastome multiforme (GBM), composée d'ensembles de données de spectres obtenues à partir d'uniques VEs dérivées de lignées cellulaires, de biopsies liquides de patients atteints de GBM et de témoins. Ces spectres sont ensuite analysés par le SVM, qui est utilisé pour les classer dans la bonne lignée cellulaire. Ceci est fait avec succès, avec une précision de 70,4% sur les mutations individuelles. Ensuite, le SVM est appliquée pour déterminer s'il est possible ou non de différencier les patients atteints de GBM des témoins sains, sur la base de l'analyse des spectres SERS, permettant d'obtenir une précision de 95 % dans la classification. Le deuxième capteur utilisé en combinaison avec l'apprentissage automatique est une plate-forme colorimétrique pour la détection rapide et au point de service de virus. Des échantillons de salive et nasopharyngés sont prélevés et soumis à une « reverse transcriptase loop-mediated isothermal amplification » (RT-LAMP), générant un changement de couleur en présence de matériel génétique viral. Afin d'obtenir un diagnostic rapide et fiable, les images du dispositif colorimétrique sont collectées et analysées via le SVM, atteignant un taux de réussite dans la

classification des patients sains vs malades de 94% après 10 minutes d'incubation, permettant des opérations de dépistage rapides.

Mots-clés: Apprentissage automatique, SVM, Plasmonique, Diagnostic, Détecteurs

Preface and author contribution

Chapter 3 of this thesis is adapted from the manuscript "Support Vector Machine for classification of Surface Enhanced Raman Spectroscopy spectra of single cancerous Extracellular Vesicles" that has been submitted to Nano Letters for publication as it is presented in the thesis (except for the formatting). It is now under review. Olivia Jeanne is to be the first author of this manuscript. In chapter 3, spectra collection was done by Carolina Del Real Mata, Mahsa Jalali and Yao Lu using a Raman microscope from the Siaj lab at the Université du Québec à Montréal (UQAM). Patient samples were obtained by Janusz Rak, Laura Montermini and Kevin Petrecca. Sara Mahshid provided the original idea, Mahsa Jalali and Carolina Del Real Mata provided additional ideas and contributed to the planning and execution of the project.

Chapter 4 of this thesis is adapted from the manuscript under preparation "Nano-plasmonically Boosted Nucleic Acid Amplification Coupled with Support Vector Machine for Minute Colorimetric Classification of Viral RNA", that is to be submitted for publication. Image acquisition was performed by Tamer Abd El Fatah and Haleema Khan at UQAM. Sara Mahshid provided the original idea, Mahsa Jalali and Tamer Abd El Fatah provided additional information on the data collected and contributed to the planning of the project. Carolina Del Real Mata contributed to the execution of the project. All additional material is original, unpublished work by the author Olivia Jeanne.

Acknowledgements

This work would not have been possible without a great number of people that I am deeply grateful for. I would like to thank my supervisor, Prof. Sara Mahshid for providing me with this opportunity to experience the world of academic research. She was unfailingly willing to provide me with whatever resources I needed during my research and to support and promote my work. I want to thank Mahsa Jalali and Tamer Abd El Fatah for explaining their projects to me, and their enthusiasm at integrating my Machine Learning work to their own projects, which often meant additional experiments and work on their part. For this I very warmly thank you. I want to thank Carolina Del Real Mata for her never-ending and unfailing, support, enthusiasm, and encouragement. Thank you for showing me how to perform Raman spectroscopy and for supporting me during the different stages of my project, from inception to manuscript writing, and all the stages in between. Your help consistently improved my work and research. You have been a constant advocate of my work and my needs, and I could never express how deeply grateful I am about this. I also thank all the members of the Mahshid lab for a fun and inclusive work environment, and, although Covid certainly meant that we did not see much of each other in person, we remained in regular contact and kept up the same good humour and gentle support that we experienced in person. I want to thank my partner, Val, for keeping my anxiety down, my optimism up and my work-life balance healthy. Finally, thanks to my roommates, my friends and family, for their support and their guidance and for allowing me to experience these two years in the best way possible.

List of Abbreviations

ANN: Artificial Neural Network

AUC: Area Under the Curve

CNN: Convolutional Neural Network

DFA: Discriminant Function analysis

DNA: Deoxyribonucleic Acid

DT: Decision Tree

EVs: Extracellular Vesicles

GBC: Gaussian Process Classifier

GBM: Glioblastoma Multiforme

k-NN: k-Nearest Neighbours

LDA: Linear Discriminant Analysis

LS-SVM: Least-Square Support Vector Machine

MAE: Mean Absolute Error

ML: Machine Learning

MLR: Multiple Linear Regression

NN: Neural Network

PCA: Principal Component Analysis

PLS: Partial Least Squares

PLS-R: Partial Least Squares Regression

QDA: Quadratic Discriminant Analysis

RBF: Radial Basis Function

RF: Random Forest

RNA: Ribonucleic Acid

ROC: Receiver Operating Curve

RT-LAMP: Reverse Transcription-

SERS: Surface Enhanced Raman Spectroscopy

SVM: Support Vector Machine

TMB: 3,3′,5,5′-tetramethylbenzidine

List of figures

List of tables

Table of contents

Note: As per formatting guidelines on manuscript-bases theses, the references for Chapters 3 and 4 are situated at the end of the chapters. All the other references used in this thesis are listed in the "Master bibliography" at the end of the document.

1 Introduction

The advent of the latest generation of biosensors signifies that scientists have access to clinical information in increasing number, with improved limits of detection and over a wider range of parameters, be they physiological measurements, nucleic acid or proteins identification and concentration, pathogens presence, etc. As a result, more precise diagnosis and monitoring of disease can be achieved. Huge hopes are riding on these new sensors, with the expectation that they can bring about an improved healthcare system, which tackles the blind spots we are faced with today. However, with the growing amount of data available, comes an equally growing need for tools to help organize, process, understand and interpret this data. Machine learning tools are a promising solution to such challenges, and it is hoped that they can help clinicians and researchers improve current medical practices and diagnosis. This could be made possible by using machine learning in conjunction with more recent sensing techniques, that are slow to be adopted due to their complexity, or by improving already existing techniques by automating the analysis step and removing the need for personnel with training and experience. These two possibilities are studied in this thesis, with the use of machine learning both on a new, complex signal: surface-enhanced Raman spectroscopy (SERS) spectra, and on an already well-established detection method: colorimetry. This thesis explores the feasibility of using a machine learning algorithm, support vector machine (SVM), on these two types of readouts in order to reach a medically relevant diagnosis. SVM was chosen as it is a well-established method, that does not require extensive computational power, making it a potential candidate for point-of-care applications, where computing resources are limited. SVM can also perform well even when presented with a limited training set, with observations comprised of a high

number of variables. It is therefore well-suited for clinical settings where collecting data in sufficient amounts is challenging.

First, spectral data from a Glioblastoma Multiforme (GBM) dataset composed of single extracellular vesicles (EVs) spectra derived from cell lines is used. SVM is employed to classify the spectra into their respective classes, ie. the different cell lines. Then, the spectra derived from the EVs of patients suffering from GBM, as well as healthy controls are analyzed through another SVM algorithm to determine whether it is possible to separate patients from healthy donors. To do so, a relative similarity measure is calculated using Mahalanobis distance. Using this relative similarity measurement, we determine the patient's risk of developing cancer: a high similarity value is associated with high risk of cancer, while a low similarity value means a low risk of cancer.

The second application pertains to analyzing a colorimetric platform for rapid SARS-CoV2 detection from patient samples. To do so, multiple timepoints are studied, with an aim of reaching a time of detection under 20 minutes in total. Images are taken at different timepoints and given to a SVM for analysis of whether the patient is Covid-19 positive or not. The images are analyzed in two color spaces: RGB, and L*a*b* and the results for each color space are compared, to find the one on which the SVM performs best. The accuracy of the SVM was obtained for the different timepoints and color spaces studied, in order to determine a compromise between accurate diagnosis and quick response time. The machine learning model could dramatically speed up the process of image analysis and offer a quick diagnostic solution without the need for any trained personnel for the interpretation of the results.

2 Review of the literature: Machine Learning for SERS and colorimetric detection

2.1 Introduction

Machine learning (ML) is a type of artificial intelligence whose particularity is that the rules used to complete a specific task, such as making a prediction, are not explicitly coded. Machine learning is "a field of study that gives computers the ability to learn without being explicitly programmed"[1]. Rather, the algorithm is fed data and learns by itself how to achieve the best prediction, by finding the relevant features in the data and adapting its parameters during what is called the "training process"[2]. ML algorithms can broadly be classified into two categories: supervised or unsupervised. The former means that the data is labeled during training, while it is not in the latter. In the first case, the algorithm will typically give categorical classification, and in the other, clustering of data in order to find patterns is generally achieved[3,4].

Common supervised ML algorithms include discriminant-based algorithms, Random Forests, SVM, and neural networks, which are being increasingly used for regression and classification tasks[5,6]. Some ML methods requires significant amounts of data to in order to be applicable on a large scale, as well as significant computational power. The recent increase in memory size and calculation capabilities of computers, permitting the storage of sufficient data, and the efficient running of powerful algorithms, have allowed them to perform well on a variety of complex tasks, in a great number of fields, from economy to manufacturing, and also in the biology and healthcare sector[7]. In the latter field, researchers and clinicians have taken advantage of ML's ability to use complex and heterogeneous data in order to diagnose patients, using data as varied as images, categorical data, and spectra, among others[8–12]. Machine learning is increasingly seen as an opportunity to improve accuracy in medical surveillance and diagnosis, assisting or even outperforming traditional methods both in the accuracy of the diagnosis and in the high throughput they offer[13,14]. Additionally, the advent of the latest

generation of biosensors is synonymous with an ever-increasing quantity of information available[15], but human limitations hinder the rapid interpretation of large amounts of data, which is often complex, high-dimensional or heterogeneous. Machine learning can help interpret, classify, and identify patterns in the data, and be beneficial both in first-rate clinical or research settings, to interpret complex data generated by last-generation biosensors, or in a low-resource setting, where advanced medical equipment and expert medical judgement are not easily available.

2.2 Supervised methods

Supervised methods, contrary to unsupervised methods, require a training phase in which each datapoint is associated with an output, a category, or a class of interest. It can be a number, or a class name, such as an analyte of interest. During training, multiple instances of each class are given to the algorithm, which then learns to recognize what features make one datapoint belong to its respective class. Training is halted when the algorithm has reached what is considered to be an acceptable performance[22].The results obtained on the training data are then generalized onto the test data, which is given to the trained algorithm, and outputs an answer based on its prior training. Supervised algorithms are used for classification and prediction, but also for regression tasks. Some algorithms, such Random Forests can be used for both supervised and unsupervised tasks[23]. Multiple types of supervised ML algorithms have been developed over the years, and we summarize the main ones below.

2.2.1 Regression-based models

Regression models aim to model an output Y, based on the given data X, in contrast to classification, where the aim is to assign a class to each datapoint without necessarily creating

a model. Some examples of regression-based models include Multiple Linear Regression (MLR), Partial Least Square Regression (PLS-R) and logistic regression. PLS-R is a regression technique that finds a linear expression allowing to map the data space to the output space , by finding the directions in the data space (X space) that explain best the variance in the output space (Y space)[24]. PLS-R enables to use many variables in each observation, including collinear variables, or variables dependent on each other, when a traditional MLR will experience difficulties when doing so[25].

Logistic regression is another type of regression where the output of the model is not the direct outcome, but rather the probability one event has of happening[26]. This is particularly suited for evaluating risks in a situation with only two outcomes[8]. However, unlike PLS, logistic regression needs independence between the variables of an observation and is sensitive to outliers[26]. The data used to build the model must therefore be carefully selected to remove outliers, and the variables given must be chosen so as not to be redundant or dependent on each other.

2.2.2 Discriminant-based methods

Discriminant-based methods such as Linear Discriminant Methods (LDA) or Quadratic Discriminant Methods (QDA) are methods that rely on decision boundaries for classification. The decision boundary is the hyperplane that best separates the classes based on the variance in the training set. This means that the optimal hyperplane is the one that maximizes the distance between the mean of the classes while minimizing the variance within each class[27]. In the case of LDA, it is assumed that all classes have equal covariance, resulting in a linear decision boundary while QDA does not, resulting in a quadratic decision boundary[28]. While LDA performs well on problems with low dimensional data, it encounters difficulties when faced with high-dimensional data, especially when the number of features, ie. the dimension of the

data is lower than the number of observations, or if the dataset only has one instance of a class in the training set[27].

2.2.3 Random Forest (RF)

Random Forests consists of an ensemble of multiple decision trees. Each decision tree is a collection of nodes and leaves. Each node represents a split between two subcategories, and the decision that the algorithm makes depends on the value of the parameters of the datapoint it is testing. The leaf is the final output of the decision tree and corresponds to the desired output, a classification for example, but random trees can also be used for regression[29]. In the case of random forests, multiple decision trees are created, which take into account different parameters of the data, then a majority vote is taken on the outputs of these trees, thus a single output for the random forest is obtained[30]. To create a decision tree to be used in a random forest, a common method is to use bagging, where each tree is trained on a subset of the initial training set. Each subset has the same sample number and each sample is randomly chosen with replacement, making each decision tree unique as it is trained on a unique subset of the training set[30].

2.2.4 Artificial Neural Networks (ANN)

An artificial neural network in its simplest form is composed of one unique artificial neuron. It functions in a similar fashion to a real neuron, in the sense that it receives several weighted inputs, adds them, and applies a non-linear function to the sum. If the result reaches a high enough value, it will activate and give an output, otherwise, it will not[31]. Although the base artificial neuron is simple enough, when multiple neurons are used together, the resulting ANN can be very powerful. During the training phase, the multiple weights are optimized in order to

obtain the most satisfactory outcome possible. ANNs can have several "layers" of neurons, and an ANN with more than three layers is called a "deep" neural network. Various architectures of ANN exist, such as feedforward neural networks, where the data flows unidirectionally[3], convolutional neural networks (CNN), where the architecture consists in one or multiple convolutional layers[32] and Residual neural networks (ResNet)[33], among others. The increase in computational power, memory size, and in the size of the training sets have enabled to build very deep networks, that perform very well on complex datasets such as the ImageNet set[32]. However, other Machine learning models are less demanding in terms of computational power and number of training samples.[31,34,35]

2.2.5 Support Vector Machines (SVM)

Support Vector Machines (SVM) were first introduced by Vapnik in 1995[36]. SVM functions by mapping the data into a high-dimensional space, and then finding the hyperplane that best separates the different classes of data. This is done by using the datapoints closest to the other class and using these as support vectors to determine the hyperplane that maximizes the distance between those support vectors (Figure 1). This is the main difference with Discriminant analysis, where the means and covariance of the different classes in the training data are used to determine the hyperplane.

Consider a binary classification problem. The classes are given by $y = +1$ or $y = -1$

The decision function for this problem is a hyperplane whose equation is given by (1), where **w** and b are determined from the training set.

$$\boldsymbol{w^T x} + b = 0 \qquad (1)$$

This hyperplane separates the data space in two regions of opposite sign. The hyperplane that optimally separates the data is the one the maximizes the margin M =2/||w|| (which is equivalent to minimizing the quantity $\frac{1}{2}$ ||w||^2).

The support vectors x_i will therefore satisfy the condition[37]:

$$\min_i \|\boldsymbol{w^T} x_i + \mathrm{b}\| = 1 \qquad (2)$$

However, in order for this condition to be met, classes must be linearly separable in the space they are plotted in. If this is not the case, it is possible to plot the datapoints in higher dimensional spaces where the classes are linearly separable using a kernel function[37]. Common kernels are the linear, the polynomial and the Radial Basis Function (RBF) kernels, defined by the following equations[38]:

$$\text{Linear: } K(x_i, x_j) = x_i . x_j + 1 \qquad (3)$$

$$\text{Polynomial: } K(x_i, x_j) = (x_i . x_j + 1)^p \qquad (4)$$

$$\text{RBF: } K(x_i, x_j) = e^{\frac{\|x_i - x_j\|^2}{2\sigma^2}} \qquad (5)$$

Binary class SVM can be generalized to multiple classes following the same prinicple[39]. One advantage of SVM over more recent ML techniques, especially neural networks and deep learning algorithms is its ability to work with a reduced number of datapoints, and datasets made of datapoints with a high number of observations, when the more recent techniques need extensive training datasets in order to function at full capacity and with high accuracy[35]. It is also less computationally expensive than most deep learning models[40].

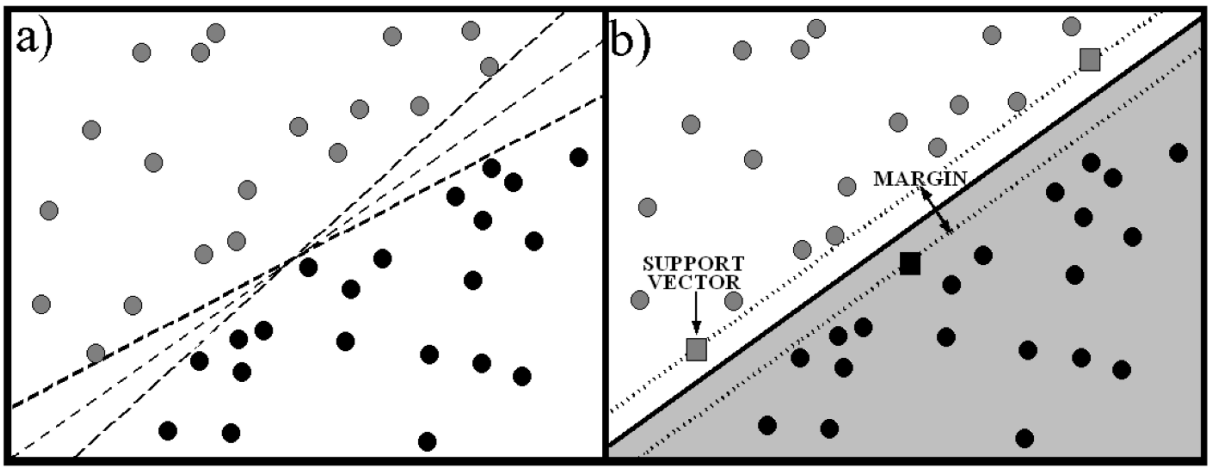

Figure 2.1: SVM hyperplanes (a) Possible separating hyperplanes and (b) optimal separating hyperplane (right) maximising the margin between two classes, the closest samples being indicated as support vectors (square marks). Reproduced with permission from [37]. Original figure available at www.tandfonline.com

2.3 Unsupervised methods

Unsupervised methods in Machine learning are methods where data is analyzed, but is not associated with a label of any kind[4]. This type of algorithm is used to uncover similar occurrences in the given data, to cluster them based on similarity, or to determine relationships and associations within the data. This can be illustrated for example when trying to find the different contexts in which the same word can be used, to group words with similar meanings together[16]. Unsupervised algorithms can also be used for density estimation, or dimension reduction, in the context of data with a high number of dimensions[17].

2.3.1 Dimensionality reduction

Dimensionality reduction can be used for visualisation purposes, outlier detection, and also as a way to simplify the data before using it with a supervised Machine learning algorithm, to decrease computational cost[18]. One of the most common dimensionality reduction algorithms

is the Principal Component Analysis (PCA) algorithm, which projects the data on a space where the first axis corresponds to the direction of highest variance[19]. Other unsupervised algorithms include K-means algorithms, hierarchical clustering, and Gaussian mixture models[17].

2.3.2 k-Nearest Neighbour (kNN)

The k-nearest neighbour algorithm is a fairly intuitive one when it comes to classification. Indeed, when given a test datapoint, the algorithm classifies it depending on its "nearest neighbours" which are the datapoints from the training set closest to the test datapoint. Multiple distances can be used for this algorithm, the most common one being the Euclidean distance. However, other distances such as Hamming, Manhattan or Mahalanobis distances can be used[20]. The class is then attributed to the unknown sample by performing either a majority vote, or a distance-weighted vote[21]. K-Nearest Neighbours is fairly easy to implement, however memory usage increases significantly when dealing with high-dimensional data[3].

2.4 ML in optical readouts

Machine learning has over the years been applied to a wide variety of fields[41,4243] and has also been used in the medical field for cancer and disease diagnosis[44–48]. ML's versatility makes it applicable to most types of readouts, in particular optical sensing through image analysis[12,46,49], and shows promise when used in conjunction with novel optical biosensors that use Raman spectroscopy or colorimetry[50–52]. ML-assisted analysis using Surface Enhanced Raman Spectroscopy spectra and colorimetric readouts have high potential as user-friendly and point-of-care approaches for disease detection, which are reviewed below.

2.4.1 Experimental techniques: SERS and colorimetry

2.4.1.1 SERS

Among multiple optical techniques, Surface Enhanced Raman Spectroscopy (SERS) shows particular promise for small-sized analyte detection and analysis. Raman spectroscopy measures the exchange of energy resulting of the inelastic scattering of photons as it interacts with the analyte. This interaction generates a change in the wavelength of the scattered light, depending on the molecular composition of the sample[53]. This scattered light acts as a fingerprint of the compound of origin. However, Raman scattering is very weak, occurring usually for one in every 10^6-10^8 photons[54]. The high fluorescence of biological samples will also be visible in the spectra, which makes the Raman signals very difficult to distinguish and analyze[54].

SERS is a technique that uses nanostructured metallic surfaces, typically gold or silver, in order to enhance the Raman scattering of a molecule adsorbed on the surface[55]. The signal can be enhanced up to a factor of 10^5-10^6 compared to the theoretical Raman spectrum intensity[55]. The SERS signal thus obtained gives precise information on the chemical structure and composition of the element of interest, while improving the signal-to-noise ratio. It has been used to detect and study a wide range of analytes, including biological materials, using label-free methods, and even achieved single molecule detection[56–60]. Such sensitivity is critical in the detection of early signs of diseases, which are often associated with the modification of the concentration of biomarkers[61].

While SERS provides a wealth of information on the analyte under analysis, it is not without limitations. In single molecule detection, the intensity of Raman spectra can vary depending on the orientation of the molecule on the substrate, and the high number of peaks in the spectra make human analysis difficult[3]. Machine learning however is very well suited to these tasks, as it can capture complex relationships between the data and can extract the features of a spectrum

in order to label it as the spectrum of a specific analyte[3]. Since Machine learning analysis of SERS spectra does not rely on the labeling of specific markers, it can be used for more comprehensive identification and detection of molecules, without being limited to a reduced array of previously identified molecules of interest[62].

2.4.1.2 Colorimetry

Colorimetric assays produce color or a change of color upon interaction of the substrate with the analyte of interest[63]. This can be done either through chemical agents that act as chromogenic indicators, or through surface plasmon resonance that produces a change of color of the substrate[64]. Common chromogenic indicators include redox indicators such as 3,3′,5,5′-tetramethylbenzidine (TMB) and pH indicators such as Cresol red or Thymol Blue[65]. For plasmonic assisted color change, gold or silver nanoparticles are often used, as the color change is driven by their level of aggregation or dispersion, and by the media surrounding them[66].

Colorimetric assays offer the advantage of low cost, easy fabrication, and fast results that are detectable by naked eye or through bright field microscopy[67]. They are therefore particularly suited to low resource settings, and to diagnosis at the point-of-care. They have been widely used in very different settings, from monitoring environment pollution[68] to forensics and chemistry[69], and more specifically in the medical and biomedical fields to detect molecules, pathogens, DNA and RNA and exosomes[64,70–72].

2.4.2 Machine learning applications

2.4.2.1 Machine learning applications using SERS

Multiple Machine learning techniques have been used in conjunction with both Raman and SERS spectra in the medical and biomedical field. They have been successfully used to identify molecules in biological samples such as saliva or urine[34,38], to detect infectious pathogens[47,73,74],

and for cancer detection[51,75–77]. For example, deep learning methods have been developed to identify neurotransmitters in urine[24], to classify multiple cancerous bio-samples according to their type, and identifying lung cancer patients, using exosomes collected from the patient's blood[51,52] (Figure 2.2). K-NN and RF algorithms have been used to distinguish pancreatic cancer, ovarian cancer and pancreatitis patients[75], and RF, SVM and LDA have been used for the detection of Covid-19 in saliva samples[10]. SVM in particular has been widely used to analyze SERS spectra in the medical field. Indeed, SVM has been applied for testing and recognition of drugs in different media (saliva and urine) that were analyzed through SERS [34,78], yielding an accuracy of over 85% and 98% respectively. Bacterial and pathogen analysis and identification has also been carried out through SVM, enabling to detect E. coli, waterborne, and urinary tract pathogens [79–81]. Wang et al. used a linear kernel SVM to detect waterborne pathogens with an accuracy over 95%[80]. Yogesha et al. used a SVM model to detect urinary tract pathogens, but also performed a PCA, as a comparison. PCA was unable to distinguish between all the different bacterial strains that were present in the sample, while SVM could do so, with an overall accuracy of 98,6% on the test set[79]. Finally, multiple diseases, such as liver cirrhosis, cancers such as lung cancer and hepatocellular carcinoma but also gastric diseases or infectious diseases such as dengue fever can be diagnosed using SERS spectra followed by SVM analysis[47,51,77,82–86]. Khan et al. compared three different SVM kernels to analyze human blood serum for dengue fever. The best performing kernel was found to be the order 1 polynomial kernel, achieving 85% accuracy on the test set[47]. Covid-19 patients were identified with 90% accuracy through the analysis of serum samples of sick and healthy patients[73]. Li et al. performed a non invasive diagnosis of Crohn's disease using urine samples of patients and healthy controls with 82.5% accuracy after leave-one-out cross-validation[86]. Liver cirrhosis was studied by Dawuti et al. using a SVM with a RBF kernel, yielding 85.9% accuracy in the

separation between healthy and patients suffering from liver cirrhosis, based on the results from a leave-one-out cross-validation[82].

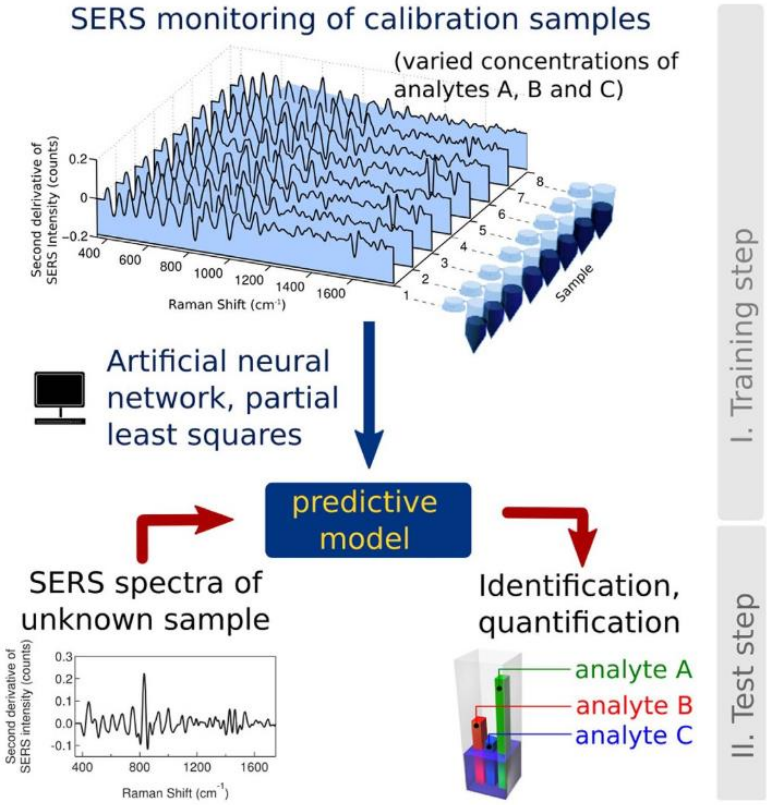

Figure 2.2: Schematic representing the experimental and analytical steps. At first, SERS data is collected from a series of calibration samples with different known concentrations of the analytes. This data is then used to 'train' the data mining methods to build a predictive model. In the next step, SERS spectra of an unknown sample is then collected and analyzed using the predictive model to obtain a result. Reproduced with permission from [24].

2.4.2.2 Machine learning applications to colorimetry

Despite their undeniable ease of use and popularity, colorimetric methods are subject to multiple contingencies, and in particular to human subjectivity. Indeed, color-blindness or different sensitivity to contrast might make colorimetric methods more error-prone, due to

uncertainty in result interpretation[87]. Machine learning algorithms offer the possibility of a more uniform output, as well as an opportunity for entirely automated, high-throughput pipeline, without needing any trained personnel to interpret the results. ML algorithms have been successfully applied with colorimetric readouts for various tasks[88–90]. Solmaz et al. performed quantification of peroxide content comparing both a LS-SVM (a variant of SVM) and a RF algorithm[88]. A grey-world algorithm was also tested to try and mitigate the impact of varied lighting conditions. LS-SVM was determined to work best with the grey-world algorithm, with 87.5% accuracy when classifying the images into 6 classes corresponding to 6 different peroxide concentration ranges. A RF algorithm was used to monitor the levels of bilirubin in urine samples, with an average accuracy of 74.05% , based on 100 repetitions of 5-fold cross-validation.[90]. Regression methods have been used to detect mercury and lead ion concentration[91,92], deep-learning algorithms have been applied to the analysis of organic carbon[93], and pH classification[94]. SVM has also been regularly applied in conjunction with colorimetric assays, for wheat mildew detection, achieving perfect classification rates using a rbf kernel[95], quantification of alcohol in saliva[96] and pH detection[97,98]. In their work, Tania et al. report that LS-SVM had 100% accuracy, outperforming the other approaches, including deep learning approaches[97]. ML algorithms can therefore be used to analyze a wide variety of analytes with high success rates, making it a promising approach to improve colorimetric detection.

2.5 ML for diagnostic purposes

2.5.1 Cancer diagnosis using liquid biopsy through SERS

Cancers and their associated biomarkers are usually analyzed through a traditional tumor biopsy, consisting in sampling part of the tumor and subsequent analysis. However, this technique fails to capture the intratumoral heterogeneity that may arise[99]. Furthermore, due to the invasiveness

of the procedure, a biopsy cannot be performed regularly, and is therefore unable to show tumor evolution. Liquid biopsy is an emerging technique that has been fast-growing in the past decade due to great potential to address these challenges. It consists in the analysis of the biomarkers present in a blood sample to obtain information on a tumor[100]. These biomarkers can be circulating tumor cells, circulating nucleic acids, proteins, metabolites, and also extracellular vesicles (EVs)[101]. The latter are vesicles secreted by all cells, including cancer cells, and are present in most bio-fluids, including blood. They are particularly interesting because they contain cargo (proteins and nucleic acids) that can provide information about their cell of origin[102]. In particular, EVs derived from cancer cells will carry cancer biomarkers[103].

The analysis of EVs through Raman spectroscopy or SERS has captured the interest of the scientific community as it can provide information on their parental cell based on their cargo. Indeed, in the case of cancer, SERS allows the collection of precise information on the composition of the tumor, by analyzing circulating extracellular vesicles and their molecular components, which reflect the cellular environment, therefore permitting to identify patients suffering from cancer, the mutations present in the tumor, thus improving treatment efficiency[104]. Machine learning can help analyze EVs obtained through liquid biopsies in the context of cancer (Table 1). EVs derived from liquid biopsies were used for lung cancer detection, using a CNN, reaching 84% accuracy[51]. Multiple ML algorithms were also used to distinguish between 3 subtypes of breast cancers, using SERS spectra acquired from EVs of three different cell lines[105]. ML has also been used in prostate cancer diagnosis[106], melanoma[56], ovarian cancer[107] and pancreatic cancer[108,109]. It is also possible to separate healthy controls from patients suffering from different cancers, instead of specifically targeting one cancer, or trying to identify the different cancers. A study by Uthamacumaran et al. showed a 90% accuracy in the classification of the spectra derived from EVs of 5 healthy controls and 4 patients suffering from different cancers using SVM[110].

Machine learning for cancer has displayed promising potential, but still needs to overcome some hurdles before it can be a fully translational diagnostic technique. Indeed, it has for now been tested mostly in vitro, using EVs derived from specific cell lines, rather than on EVs extracted from patient samples. Furthermore, the cohorts used are still reduced in size, due to the difficulty of obtaining enough samples, limiting the generalization of ML techniques. Finally, the accuracies obtained with these methods are still somewhat lower than what is obtained using standard biopsy techniques. As a point of comparison, core-needle biopsies for breast cancer detection display 97.7% sensitivity and 100% specificity, and surgical biopsy was found to have a 97.5% sensitivity[111,112]. Thus, in order to be considered as a potential candidate to replace the traditional biopsy, ML-based methods relying on SERS analysis will have to improve their accuracy, sensitivity and specificity. However, as SERS becomes a more widespread technique, the creation of large and varied databases will become easier, enabling ML models to truly show their potential in cancer diagnosis using EVs.

Table 2.1: Table of studies using SERS spectra derived from EVs and ML for cancer diagnosis

Cancer type	ML type	Analyte	Efficiency	Ref
Melanoma	PLS-DA	B16F10 melanoma-derived vesicles and RBC derived vesicles	Acc=93.9% Sens=92% Spe=95.1%	2016[56]
Pancreatic cancer	DFA	EVs of pancreatic cancer cell lines and serum derived EVs	Cell lines Acc=90% Sens=90.6% Spe=97.1% Serum samples Acc=56% Sens=57% Spe=57%	2019[108]
Lung cancer	CNN	EVs of lung cancer cell lines, and plasma derived Evs	Cell lines Acc=94.8% Plasma samples Acc=84% Sens=84% Spe=85%	2020[51]

Prostate cancer	CNN	EVs from healthy blood and from 2 cancerous cell lines	(Results for the full spectra) Processed data: Acc=90.89% Raw data: Acc=95.22%	2020[106]
Breast cancer	SVM (rbf kernel), k-NN, LDA, RF, GBC, DT	EVs from MCF-7, BT-474 and BT-20 cell lines	All algorithms: Acc=100%	2021[105]
Ovarian cancer	Logistic regression	EVs from ovarian cancer cell lines OV-90, OVCAR3, EOC6, EOC18, and ovarian surface epithelial cell line	Healthy vs. OV-90, OVCAR-3: Acc=99.2% Sens=99.0% Spe=99.5% Healthy vs. EOC6: Acc= 99.2% Sens=98.7% Spe=99.5% Healthy vs. EOC18: Acc=99.4% Sens=100% Spe=99.0%	2021[107]
Pancreatic cancer	DT	EVs derived from the serum of patients and healthy control	Sens= 95% Spe = 96%	2021[109]
Multiple cancers	RF, DT, SVM	EVs extracted from serum of patients and healthy controls	RF:Acc=83.33% DT: Acc=100% SVM: Acc=100%	2022[110]

Acc: Accuracy, Spe: Specificity, Sens: Sensitivity, DFA: Discriminant Function analysis, GBC: Gaussian Process Classifier, DT: Decision Tree

2.5.2 Infectious diseases diagnosis through colorimetry

Infectious diseases are to this day a great threat to populations, especially people living in poverty or extreme poverty. Despite the growing arsenal of medical responses, it is not always possible to reach the communities in need of treatment in time to prevent the spread of diseases. Widespread infections can still occur, as has been evidenced in the last years, with SARS, MERS, Ebola, and Covid-19 outbreaks[113]. A solution to limit contagion is to implement testing on a large scale[114]. However, for this to be effective, the testing must be rapid, inexpensive and user friendly. Colorimetry offers good potential to answer all these challenges. Further aided by Machine learning, colorimetric approaches have been deployed recently to diagnose infectious diseases using portable, affordable, and rapid tests. A Random Forest classifier has

been used to detect Tuberculosis-antigen-specific antibodies from the sputum of patients and healthy controls tested with an ELISA platform, reaching 98.4% accuracy, 100% sensitivity and 96.19% specificity in classification[115]. A serology diagnosis for early stage Lyme disease using a multiplexed vertical flow assay and an ANN was also reported (Figure 3)[116]. Finally, a colorimetric test for malaria has recently been developed, using on device LAMP, a paper-based assay and a CNN for the final analysis and detection, with an accuracy over 97%, sensitivity of 90.5% and specificity of 87%[50]. The devices reported here offer good results when compared with existing guidelines for infectious diseases testing devices. For example, the minimal requirements set by the U.S. Food and Drug Administration for devices testing for influenza A and B imposes a 80% sensitivity and 95% specificity when compared to RT-PCR[117]. Machine learning has a high potential for the diagnosis of infectious diseases with high accuracy while ensuring high-throughput and ease of use, and will continue to grow in the following years, as its integration to portable devices such as smartphones becomes more widespread.

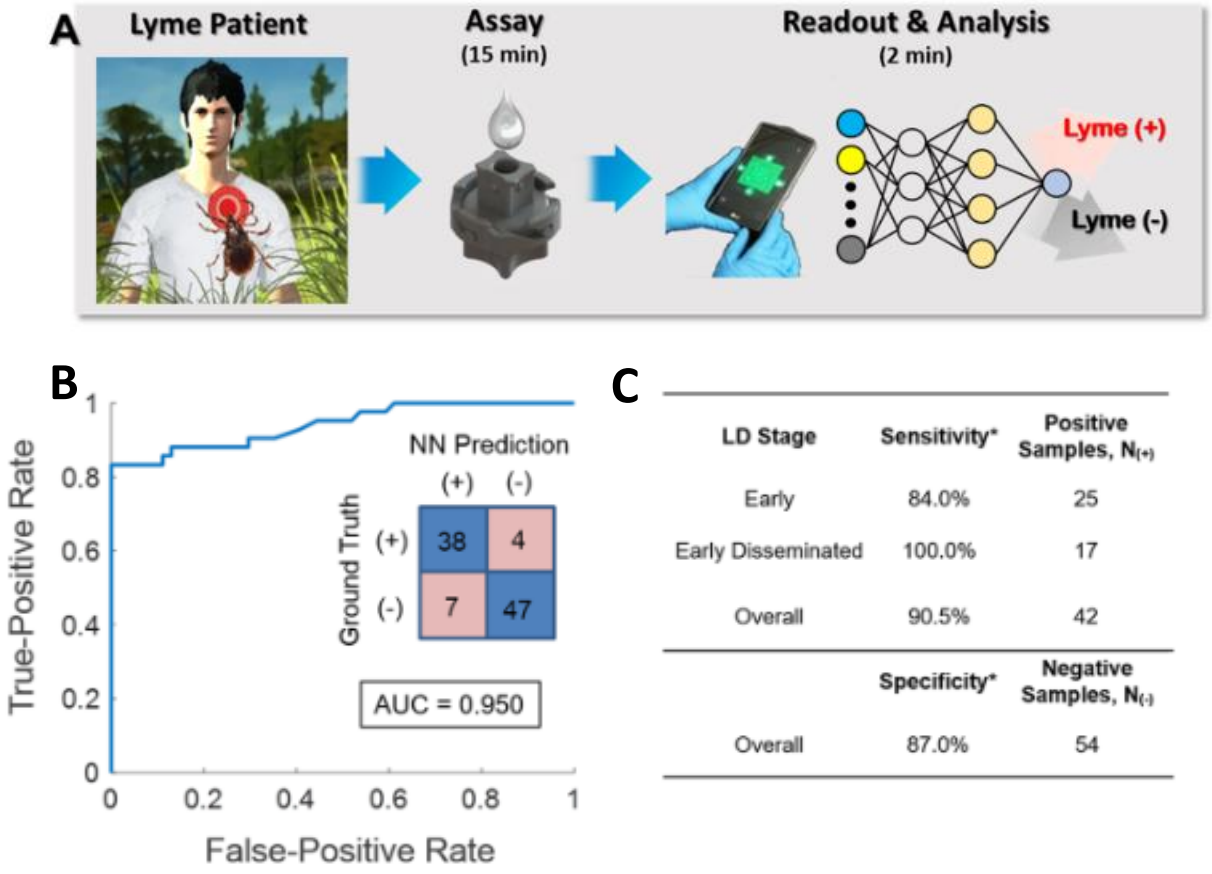

Figure 2.3: Machine learning for Lyme disease diagnostic. A) Overview of POC Lyme disease diagnostic testing using xVFA and machine learning. B) ROC curve for the blind testing data (NTest = 96) as output from the neural network from the training set (NTrain = 100). The inset shows the confusion matrix and area under the ROC curve (AUC). C) The table to the right summarizes the performance over the blindly tested LD human serum samples with respect to the two-tier testing method. Adapted with permission from[116]. Copyright 2020 American Chemical Society.

2.6 Conclusion

The development of Machine learning offers great opportunities to researchers and clinicians to further enhance the performance of new generation sensors. The creation of open access libraries that enable new users to code ML algorithms resulted in its increased use, even in fields of research that did not necessarily heavily rely on computer science before. The

integration of ML to analysis and diagnosis pipelines is set to become more popular as time goes on. Deeper and more complex architectures will grow in popularity as the size of datasets and computing power increases, improving their runtime and performance. Lighter architectures, both in the short and long term are also sure to stay popular, due to their ease of implementation and their smaller requirements in data and computing power, making them suitable for point-of-care settings. SVM in particular remains a popular choice, able to give accurate predictions using small but highly complex training datasets. There is great potential for the application of Machine learning algorithms in cancer detection using EVs obtained from liquid biopsies, and in the diagnosis of infectious diseases through colorimetric methods.

In the following chapters, we explore the application of SVM on SERS spectra for Glioblastoma (GBM) diagnosis, and for colorimetric Covid-19 diagnosis, as a proof that it can be successfully used on both of these datasets, which, to the best of our knowledge, have not been studied through SVM before. The contents of Chapter 3 have been submitted to Nano Letters as shown in this thesis, except for the formatting.

3 Support Vector Machine for classification of Surface Enhanced Raman Spectroscopy spectra of cancerous single Extracellular Vesicles

Olivia Jeanne[1], Carolina del Real Mata[1], Mahsa Jalali[1], Laura Montermini[2], Yao Lu[1], Kevin Petrecca[3], Janusz Rak[2], Sara Mahshid*

[1] *Department of Bioengineering, McGill University Montreal, QC, Canada*

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

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

*sara.mahshid@mcgill.ca

This manuscript has been submitted for publication to Nano Letters and is now under review.

List of Figures

List of Tables

3.1 Abstract

Machine learning can potentially overcome the complexity of biological data and classify large datasets like spectra of heterogeneous biological samples generated by surface-enhanced Raman spectroscopy (SERS). Here, we use a support vector machine (SVM) to analyze the SERS spectra from single extracellular vesicles (EVs), a new biomarker continuously released from cancer cells into biofluids. The heterogeneity and intrinsic complexity of cancerous EVs are challenges in liquid biopsy. We developed a SERS-assisted nanocavity microchip that allows the isolation of single EVs and outputs information on the biomarker status. Using the microchip, we generated a database of single-EV spectra derived from glioblastoma multiforme (GBM) cell lines, GBM patient samples, and controls. We demonstrate that SVM can analyze spectral data at the single-EV resolution and achieve 70,4% accuracy in the detection of specific GBM mutations from different cell lines and 95% accuracy in binary classification of real samples into GBM-positive and GBM-negative.

Keywords: Machine Learning, SVM, SERS, Plasmonics, Diagnostics, sensors

3.2 Introduction

Machine learning algorithms have proven particularly helpful for the organization and analysis of growing amounts of data in all fields. Medical data is no exception, as the latest generation of sensors produces a vast amount of data for continuous monitoring of health parameters[1]. Machine learning's main advantages are its versatility, as it is compatible with numerous input formats like images, spectra, and categorical data[2–4]; its competence to swiftly process large amounts of data, and its capability to identify and use complex relationships within the given data[5,6]. In the context of health and medical care, using machine learning analysis can provide

solutions to challenges such as personalised medical care, early and accurate monitoring of disease onset and progression, as well as treatment quality improvement[7]. An emerging trend is the coupling of machine learning with new highly sensitive sensors for improved detection and diagnosis. Sensors using an optical readout are an attractive option for the detection of biological analytes due to their high sensitivity and good signal-to-noise ratio[8,9].

The race for developing improved point-of-care sensors which are easy to operate, robust, with a long shelf life, and are adapted for low-resource settings results in an increasing number of sensors trying to bypass the need for a labeling agent, such as 3,3′,5,5′-tetramethylbenzidine (TMB) or fluorophores[8,10]. This can be achieved, among other techniques, by using surface plasmons, which occurs when the incident light causes the free electrons present in specific metals to oscillate[11]. An optical technique harnessing this principle is surface-enhanced Raman spectroscopy (SERS). SERS is possible because of the presence of nanostructures on a plasmonic metallic surface, which dramatically enhances the intensity of the Raman scattering of a molecule adsorbed on this nanostructured surface [12]. By exploiting SERS capabilities, plasmonic sensors can achieve single-molecule detection[13]. The spectrum generated by SERS is then a "fingerprint" of the biological analyte, which enables the obtention of precise structural information on its components[14]. Additionally, it can be used in a relatively low resource and point-of-care context, with the advent of handheld Raman probes[15,16]. However, while SERS spectra give a wealth of information on the component under study, they are complex and difficult to interpret. This is further complicated by the heterogeneity of the spectra caused by the various orientations of the molecule on the SERS surface[17], hampering the application of SERS readout for clinical purposes. A solution to these drawbacks is using machine learning approaches for analysis and diagnosis. Indeed, machine learning can be used to interpret this rich signal information, therefore enhancing clinicians' abilities to diagnose diseases and monitor the health status of individuals in a highly personalized manner. Multiple machine

learning algorithms have been used to analyze SERS spectra, like Discriminant Analysis[18–20], k-nearest neighbors[21], neural networks[22,23], and support vector machine (SVM)[24,25]. The latter, SVM, is a supervised machine learning algorithm that classifies the data based on their position with respect to a separating hyperplane, determined during the training phase[18]. SVM has the ability to achieve high accuracy despite having a reduced number of training points and small datasets comprised of observations that are integrated by a high number of variables[27]. SVM algorithms have been applied over the years to a multitude of data in the medical and biomedical fields, including Raman and SERS spectra with high success. Existing technology applying SVM to Raman and SERS spectra is summarised in Table 1. Among the varied biological material available for analysis, extracellular vesicles (EVs) contain comprehensive information that can be used as cancer biomarkers[28]. EVs are nanosized vesicles secreted by cells that carry molecular cargo such as proteins and nucleic acids, representative of their cell of origin[29]. Consequentially, EVs originating from cancer cells will have different cargo than EVs from healthy cells and, thus, different spectral fingerprints. The study of EVs at the single-level is hypothesized to hold great potential, as EVs' heterogeneity caused by inter and intracellular variability, presents a challenge for their batch-analysis [30,31].

In this work, the SERS spectra database is comprised of single EVs of glioblastoma (GBM) cells with a single EV resolution. Machine learning analysis is used to classify spectra obtained from different mutated cell lines, and to diagnose patients with GBM. We have applied an SVM to assist in the interpretation of cancerous SERS spectra and perform the analysis (Figure 1). Machine learning algorithms have the potential to be used where machines and personnel are lacking, making it a suitable candidate for low-resource, point-of-care medicine. The proposed work paves the way for a robust method to classify and analyze SERS spectra for applications in cancer diagnosis.

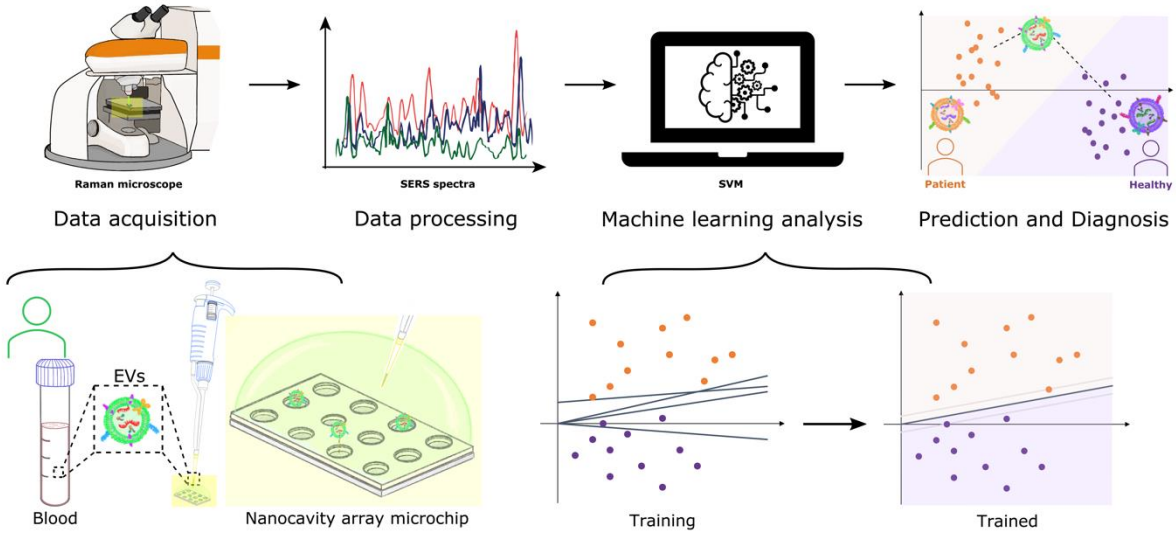

Figure 3.1: Schematic of the data collection and Machine learning processing. Cancerous single EV SERS spectra are collected for diagnosis from a blood liquid biopsy, processed, and loaded into our microchip. The spectra are collected from a Raman microscope and pre-processed for analysis. SVM is used to analyze the data collected, during training, SVM finds the hyperplane that best separates the training data, and then uses this hyperplane to classify test data. Resulting in predictions for the differentiation of cell lines, and of the health status of the patient.

3.3 Results and discussion

Major improvements in cancer treatment could stem from earlier and less invasive diagnosis methods. In recent years liquid biopsy has been positioned as an attractive alternative candidate due to its less invasive and simpler way of performing diagnosis[32]. Liquid biopsy is a type of biopsy where analysis is performed to study biomarkers present in patients' blood. EVs are a biomarker existent in most body fluids[33] and can be obtained through a simple liquid biopsy[34]. EVs are secreted by all cells, including tumour cells, and carry molecular cargo (protein and nucleic acid) and surface proteins that allow identification of their cell of origin. The analysis of EVs, and of their cargo through SERS makes it possible to identify the cell line and mutation

of the cell it originated from[30,31]. However, in order to fully harness the potential of EVs for diagnosis, it is necessary to analyze individual EVs, rather than signals obtained by averaging over multiple EVs. Previous work from this group has enabled the detection of individual EVs, as well as distinguish EVs from multiple GBM cell lines[35,36]. GBM was chosen as the most common and deadly form of brain cancers, with an average time of survival no longer than 18 months after the initial diagnosis[37]. The gold standard diagnosis method used today is the analysis of a tumour biopsy, an invasive procedure which remains prone to undesirable side-effects. SERS spectra collected from single EVs via liquid biopsy have the potential to be used as a promising diagnostic tool to help determine a patient's cancer status.

Table 3.1: Table of studies using SVM in the biomedical field, and their efficiency metrics

Category	Aim	Analyte	Readout	Efficiency	Ref.
Cancer detection	Lung cancer detection	EVs of lung cancer cell lines, and plasma derived exosomes	SERS	EVs of lung cancer cell lines: *Acc>94.8%* Plasma derived exosomes: *AUC= 0.76*	[36]
	Non small cell lung cancer identification	NSCLC cell lines H1229, H460 and A549, and healthy leukocytes	SERS	4 cell types *Acc=88.75%* Leukocytes vs. cancer cells *Acc=100%*	[37]
	Glioma detection	Supernatant from brain tissue	SERS	*Acc=97.9%* *Sens=96%* *Spe=100%*	[38]
Disease detection	Covid-19 detection	Serum	Raman	*Acc=90%* *Sens=89%* *Spe=93%*	[23]
	Covid-19 detection using NS1 un saliva	Saliva of subjects Covid-19 positive, negative and having had a past Covid-19 infection	Raman	*Acc=78%* *Sens=78%* *Spe=89%*	[19]
	Dengue infection detection	Blood samples of infected and healthy subjects	Raman	L-SVM: *Acc=82%* *Sens=71%* *Spe=91%* Poly-SVM of order 1:	[39]

				Acc=85% *Sens=73%* *Spe=93%* Rbf-SVM (2) : *Acc=82%* *Sens=72%* *Spe=78%*	
	Crohn's disease	Urine samples	SERS	*Acc=82.5%* *Sens=82.1%* *Spe=83.3%* *F-1 score=86.2%*	[40]
Bacterial detection	Staphylococci strain detection	Bacteria from 16 staphylococci strains	Raman	IPF background removal: *Acc=98.8* RCF background removal: *Acc=98.9*	[18]
	Identification of pathogenic bacteria	30 bacterial and yeast strains	Raman	30 strains *Acc=74.9%*	[41]
	Bacterial identification	9 E.coli strains	SERS	Fingerprint region L-SVM: *Acc=90.1%* Rbf-SVM: *Acc=91.0%* Fingerprint+CH-stretching region: L-SVM: *Acc=91.5%* Rbf-SVM: *Acc=92.6%*	[42]
Drug testing	Drug recognition	Urine spiked with MDMA and MAMP at different concentrations	SERS	*Acc=97.76%*	[43]
	Direct testing of drugs in urine	Urine spiked with MAMP, and real urine samples	SERS	Spiked urine C-SERS: *Acc=85%* *Sens=84.5%* *Spe=85.5%* Spiked urine D-SERS: *Acc=96.1%* *Sens=96.1%* *Spe=96.1%* Real urine samples (D-SERS): *Acc>90%*	[24]
	Testing of illicit drugs	Illicit drugs (oxycodone, cocaine, heroin, THC) dissolved in water. Saliva spiked with cocaine	SERS	Drug identification: *Acc=100%* Drug quantification *Acc=98.3%*	[44]

3.3.1 Cell classification and mutation identification via SVM

The main strength of ML and SVM is their ability to identify unique features in the data that is given while leveraging this knowledge to classify unknown test data. Thus, slight differences in the spectra, deriving from molecular modifications of the analyte, can be identified and used by the SVM to accurately classify the spectra into a specific class. Identifying specific mutations in GBM is central for effective diagnosis. We analyze EVs derived from normal glial cells, glioma cell lines, and glioma stem cell lines (Table 2). One molecular alteration is EGFRvIII which leads to the expression of the mutated EGFRvIII protein, a known marker of cancer. Indeed, this mutated gene in glioma cells leads to increased survival, invasion, and proliferation rates, which are all factors for tumorigenesis[47]. The mutated EGFRvIII protein is found in EVs secreted by the molecularly altered cells, allowing for the mutation monitoring through the analysis of the EVs[48]. We also studied cells displaying upregulated O^6-methylguanine DNA methyltransferase (MGMT), a marker that indicates resistance to temozolomide (TMZ), used in GBM chemotherapy treatment for GBM[49]. While EGFRvIII is useful in determining cancer onset and progression, MGMT expression can give an interesting insight into the susceptibility of the tumour to therapy, and risks of relapse[49]. To this end, we include both the mutated and wild-type glioma cell lines, to determine if it is possible to separate these mutations in cell lines with high accuracy. Namely, U87 and U373 cell lines had been engineered to express the EGFRvIII oncogene. Patient derived glioma stem cell lines GSC83 and GSC1005 naturally express the EGFRvIII mutation and were compared against GSC83 and GSC1005 in which the EGFRvIII had been knocked-out through CRISPR-Cas9. Finally, MGMT expression was studied through a series of GSC1123 cell lines, some naturally

expressing MGMT, while the others did not. Normal Human Astrocytes (NHA) were used as a healthy control.

Table 3.2: Cell lines used, with the mutation resulting in modified protein expression.

Cell type	Cell line	Mutated cell lines	Mutation type
Glial cell	NHA	NHA	*Control glial cell*
Glioma cell	U87	U87 Parental	*No EGFRvIII expression*
		U87 EGFRvIII	*EGFRvIII expression*
	U373	U373 Parental	*No EGFRvIII expression*
		U373 EGFRvIII	*EGFRvIII expression*
Glioma stem cell	GSC83	GSC83 wt	*Natural EGFRvIII expression*
		GSC83 ko	*No EGFRvIII expression*
	GSC1005	GSC1005 wt	*Natural EGFRvIII expression*
		GSC1005 ko	*No EGFRvIII expression*
	GSC1123	GSC1123	*No MGMT expression*
		GSC1123 MGMT	*MGMT expression*

The spectra obtained and analyzed by the SVM are shown in Figure 3.2b. For illustration purposes, we show the average spectra with 1 standard deviation, for each mutated cell line. The spectra display common peaks around 520cm^{-1},1000cm^{-1}, 1240-1330cm^{-1}, 1600cm^{-1} corresponding to SS disulfide bridge, phenylananin, amide, C-H bond from proteins and tyrosine respectively[50–52]. As can be expected, there is some heterogeneity between the spectra, making their analysis and classification difficult. This is the reason for turning to SVM, as a means to better analyze and classify them.

The SERS spectra acquired were separated according to each mutation of each cell line, for a total of 11 classes, and then analyzed by SVM. The single-EV resolution data was separated into training and testing sets, with 70% of the data used for training and 30% for testing, the results of the testing set are shown in Figure 3.2c-d. The overall accuracy, considering all the classes of the model is 70.04%, with most of the errors being between the same cell line but

different mutations, especially for GSC1005 and GSC1123, as shown in the confusion matrix (Figure 3.2c). The receiver operating characteristic (ROCs) and area under the curve (AUC) values also support that the SVM has the ability to differentiate mutations using the SERS spectra, as AUC values are systematically over 0.75, with an average of 0.94 (Figure 3.2d).

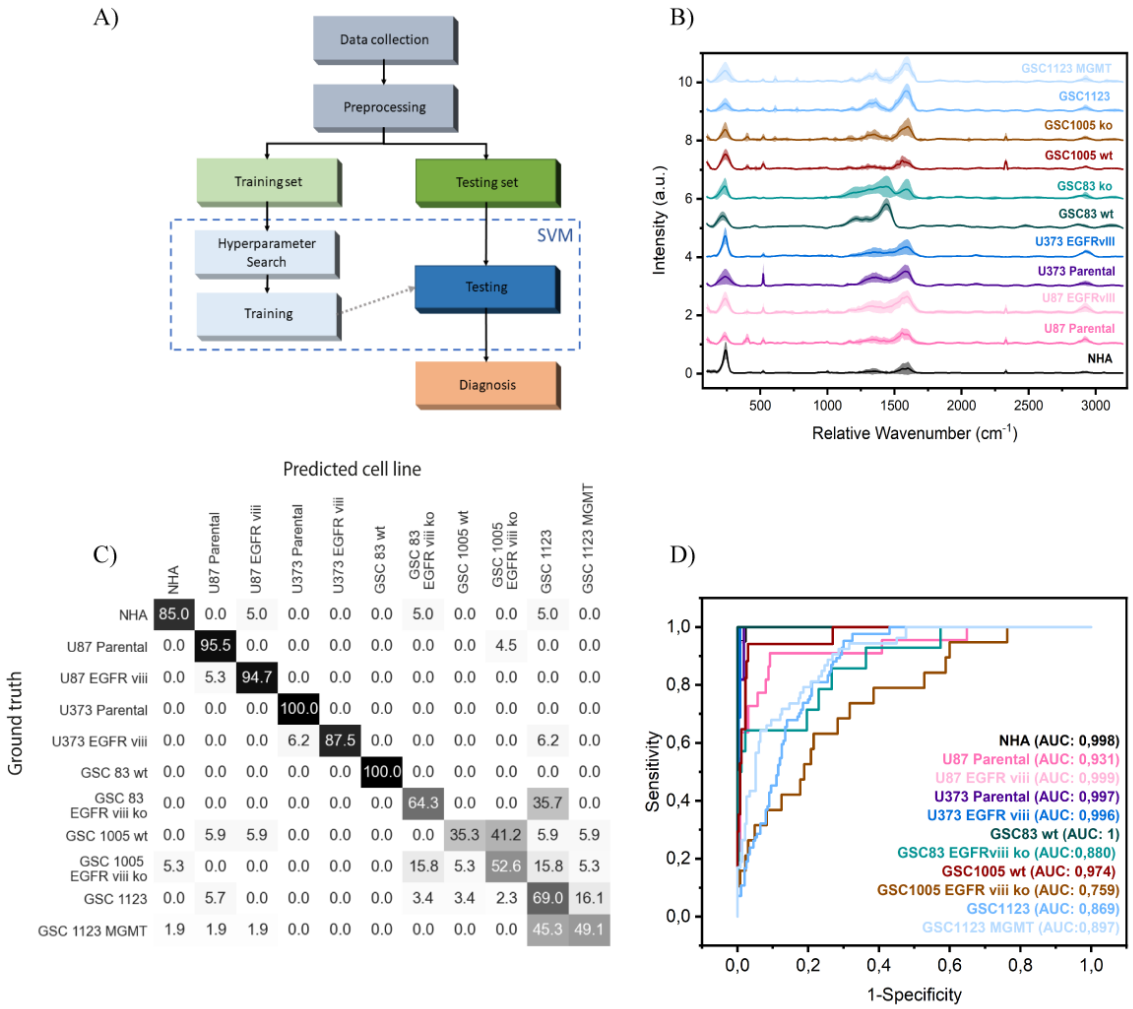

Figure 3.2: Cell line and mutation analysis using SVM. A) Flowchart representing the different steps in this work, from the data collection to the final diagnosis. Training is first performed on the training set, then the trained SVM algorithm is tested on the unseen test data. B) Spectra collected from each cell line and mutation. The colored line is the average spectra, with 1 standard deviation shown. C) Confusion matrix showing the accuracy of the prediction for each class (showing each cell line, and mutation that was introduced to the cell line), normalized along the rows. D) ROC of the 11 classes, calculated in a one-vs-all approach.

When grouping the results by cell line, the SVM can classify any given spectra within its correct cell line with at least 72% true positive rate (TPR), with the TPR being over 90% for U87, U373 and GSC1123 (Table 3.3). The true positive rate is calculated for each cell line in the following manner:

$$TPR = \frac{TP}{TP+FN} \qquad (1)$$

TP is the number of true positives, the number of spectra that were correctly predicted as belonging to their rightful class, and FN, is the number of false negatives, which are the number of spectra from a cell line that were incorrectly predicted as belonging to another cell line. The AUC for each cell lines is also calculated, which emphasizes that the cell line classification is extremely successful, with AUC systematically higher than 0.94. This further demonstrates that most of the errors done by the SVM result from misclassification within one cell line.

Table 3.3: True positive rate and AUC for each cell line

	TPR	AUC
NHA	85	0.99697
U87	97.561	0.99368
U373	96.2963	0.99323
GSC83	81.8182	0.94436
GSC1005	72.7273	0.95845
GSC1123	90.099	0.97438

3.3.2 Implementation of SVM in clinical study

SVM has previously demonstrated ability at binary classification for the diagnosis of patients[40,42]. Here, we implemented SVM to diagnose patients suffering from GBM using EVs obtained through liquid biopsy. Using spectra taken from individual EVs from both patient

samples and healthy controls, we classify them into a healthy class (GBM-negative), or a patient class (GBM-positive). The training set used contains both spectra taken from the EVs of cultured cells and patient and healthy samples. This is in order to obtain a more representative and comprehensive training set that includes EVs excreted by cells from the entire human body, which will be found in both patients and healthy samples. A held-out test set composed of the spectra from 12 patients and 8 healthy controls are used to evaluate the potential of the SVM algorithm as a diagnostic tool. EVs were collected pre-operatively from confirmed GBM patients.

Each spectrum is predicted individually by the algorithm before additional analysis is done to obtain a diagnosis of the general health state of a patient. The global accuracy of the test set is 77,98%. An explanation for this outcome could be attributed to GBM patients being able to have cells that secrete healthy EVs, creating false negatives, as they will be labeled as GBM-positive, even if the actual phenotype is GBM-negative. Also, false positives can be encountered due to random mutations in healthy controls which can make cells secrete EVs having the same phenotype as patient EVs. Figure 3a shows the score assigned to each spectrum, represented as a data point, of the entire test set, divided into healthy and patient spectra. A positive score indicates a SERS spectrum is classified as GBM-positive (ie a "patient" spectra), while a negative score indicates a SERS spectrum classified as GBM-negative (ie a "healthy" spectra). The average score of the spectra obtained from healthy controls is negative (-0.735), while the average score of the patient-derived spectra is positive (0.748), showing a statistically significant difference in the average scores, with $p<0.001$, demonstrating a successful individual spectra classification by the SVM. However, it is important to determine whether or not this ML algorithm can help to distinguish between healthy and patient samples by examining each clinical sample individually, instead of each spectrum. Tables 3.4 and 3.5 show the average score for each patient and healthy control. All the patients have a positive average

score (Table 3.4) and all the healthy controls except one have a negative average score (Table 3.5), exhibiting an almost perfect accuracy when determining the health status of a clinical sample. To simplify the reading during the discrimination process and to take into account that multiple ML algorithms could be used in the future to reach this diagnosis, thus leading to different outputs, we have calculated the average relative similarity of each clinical sample (Figure 3.3b). The ROC curve of the SVM used also shows an AUC of 0.84 (Figure 3.3c), further confirming the efficient performance of the algorithm.

Table 3.4: SVM score obtained for each patient sample. Each score is obtained by averaging the score of the spectra belonging to the sample.

	Patient											
Number	1	2	3	4	5	6	7	8	9	10	11	12
SVM score	1.32	0.23	1.61	1.36	0.73	0.28	0.45	0.88	0.13	0.65	0.86	0.97

Table 3.5: SVM score for each healthy sample.

	Healthy							
Number	1	2	3	4	5	6	7	8
SVM score	-1.8	-2	-0.1	-0.1	0.21	-1.1	-0.9	-0.8

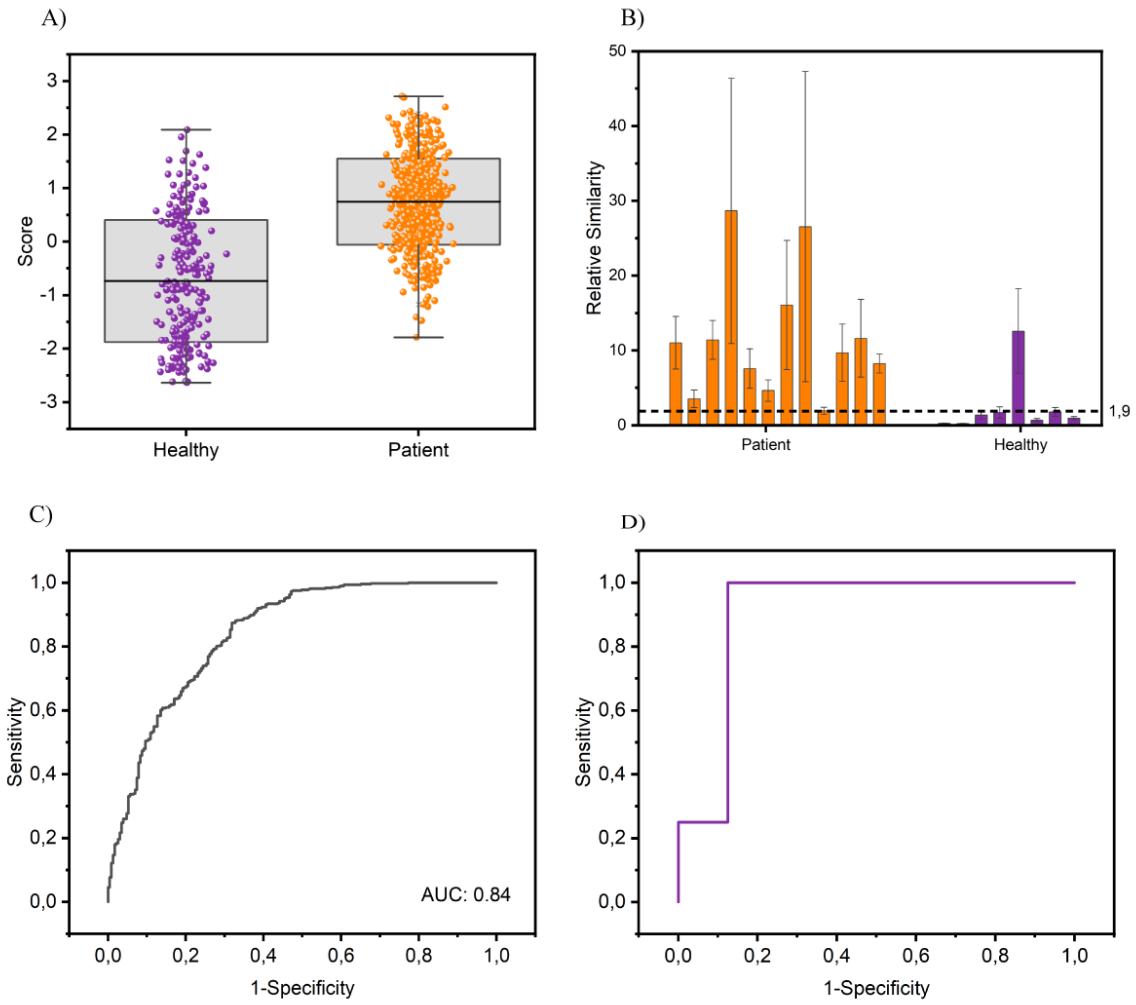

Figure 3.3: Diagnosis of GBM in patients using SVM. A) Score and whisker plot of each of the tested spectra after classification by SVM, separated into healthy and patient-derived spectra. A negative score indicates a spectrum classified as healthy, and positive, as a patient suffering from GBM. The box shows the mean, SD, and the whiskers the minimum and maximum values. One-way ANOVA confirmed that the means value difference was statistically significant ($p<0.001$). B) Average relative similarity, calculated for each patient and healthy sample. A high average relative similarity means a high risk of cancer, and a low one means a low risk of having cancer. The threshold of 1.9 is the cut-off value that allows for classifying patients and healthy controls with the best specificity and sensitivity. C) ROC curve for the SVM classification, with an AUC of 0.84. D) ROC curve obtained from the average similarity of the samples. It is used to determine the optimal threshold that offers the best compromise between sensitivity and specificity when classifying patient samples.

A low average relative similarity means a low risk of cancer (GBM-negative sample) and a high average relative similarity means a high risk of cancer (GBM-positive sample). For each sample, Mahalanobis distance was calculated between every spectrum and the patient and healthy spectra of the training set[53,54]. The Mahalanobis distance enables the calculation of the similarity while taking into account different standard deviations along different axes. The ratio of each distance is then calculated to obtain the relative similarity of each spectrum. At last, all the relative similarities are averaged to obtain the average relative similarity for the sample. The cut-off point to distinguish between GBM-negative and GBM-positive samples is established by calculating the point which minimizes the distance to the point (0;1) on the ROC obtained with the average relative similarity of the samples (Figure 3.3d), allowing for discrimination between the two with maximum sensitivity and specificity. The threshold is determined to be at 1.9, with close to perfect classification (95% accuracy, 100% sensitivity, and 88% specificity). SVM is a powerful tool to aid in the identification of SERS spectral features critical to distinguish cell lines and mutations from healthy controls, with applications for cancer diagnosis based on liquid biopsy.

3.4 Conclusion

Our study using SVM on SERS spectra has proven successful in the identification of mutations in GBM-associated cell lines, and the diagnosis of GBM in patients. We performed the analysis of single EV spectra collected via a nanocavity array microchip and SERS. An accuracy of 70,4% was reached for mutation and cell line classification, and 95% accuracy was attained in patient diagnosis. For the latter, this analysis was done with clinical samples collected by the minimal invasive liquid biopsy, which is easier to acquire and process than the traditional invasive method, tumor biopsy. Furthermore, blood samples can be repeatedly taken over a long period, making it possible to monitor the evolution of the patient in a minimally invasive and dynamic

manner. This additionally, enables overseeing the molecular evolution or response to treatment through this analysis pipeline. SVM has proven successful in analyzing complex SERS spectra to give a clinically relevant medical diagnosis, harnessing the high quantity of information present in such spectra in a label-free manner. This could further encourage the use of SERS for sensing uses without resorting to time-consuming analysis of the spectra. SVM is a simpler and less demanding type of algorithm when compared to deep learning methods, both in terms of computing power and training data. Although deep learning approaches tend to replace classical ML approaches such as SVM due to their higher accuracy, and the growing availability of both training data and computing power (or cloud platforms), allowing deep learning implementation, SVM is still able to predict the presence of GBM in a clinically relevant manner with 95% accuracy. This is of particular importance when studying diseases where it is difficult to acquire a high amount of data, due to the rarity of the disease, or the difficulty of collecting relevant biological material. It must also be noted that SVM requires much less computational power than deep learning algorithms to be operated, making it easier to deploy in settings where computing infrastructure or energy availability is scarce. Overall, coupling SVM with an integrated nanocavity array and SERS could prove to have great diagnostic potential in point-of-care settings.

3.5 Acknowledgements

The authors thank the Faculty of Engineering at McGill University, the Canadian Cancer Society (255878 CCSRI), Natural Science and Engineering Research Council of Canada (NSERC, G247765), New Frontiers in Research Fund (250326), Canada Foundation for Innovation (CFI, G248924) for financial support. CdRM thanks McGill Engineering Award (MEDA and Fonds du Recherche du Quebec (FRQnet).

3.6 References

1. Fang, R., Pouyanfar, S., Yang, Y., Chen, S.-C. & Iyengar, S. S. Computational Health Informatics in the Big Data Age: A Survey. *ACM Comput. Surv.* **49**, 1–36 (2016).
2. Mor-Yosef, S. Ranking the Risk Factors for Cesarean: Logistic Regression Analysis of a Nationwide Study. *Obstetrics and Gynecology* 944–947 (1990).
3. Yang, J. *et al.* Deep learning for vibrational spectral analysis: Recent progress and a practical guide. *Analytica Chimica Acta* **1081**, 6–17 (2019).
4. Titano, J. J. *et al.* Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med* **24**, 1337–1341 (2018).
5. Zitnik, M. *et al.* Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion* **50**, 71–91 (2019).
6. Rajkomar, A., Lingam, S., Taylor, A. G., Blum, M. & Mongan, J. High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. *J Digit Imaging* **30**, 95–101 (2017).
7. Archenaa, J. & Anita, E. A. M. A Survey of Big Data Analytics in Healthcare and Government. *Procedia Computer Science* **50**, 408–413 (2015).
8. Cooper, M. A. Optical biosensors in drug discovery. *Nat Rev Drug Discov* **1**, 515–528 (2002).
9. Damborský, P., Švitel, J. & Katrlík, J. Optical biosensors. *Essays in Biochemistry* **60**, 91–100 (2016).
10. Zanchetta, G., Lanfranco, R., Giavazzi, F., Bellini, T. & Buscaglia, M. Emerging applications of label-free optical biosensors. *Nanophotonics* **6**, 627–645 (2017).
11. de Mol, N. J. & Fischer, M. J. E. Surface Plasmon Resonance: A General Introduction. in *Surface Plasmon Resonance* (eds. Mol, N. J. & Fischer, M. J. E.) vol. 627 1–14 (Humana Press, 2010).
12. Moskovits, M. Surface-enhanced spectroscopy. *Rev. Mod. Phys.* **57**, 783–826 (1985).
13. Nie, S. & Emory, S. R. Probing Single Molecules and Single Nanoparticles by Surface-Enhanced Raman Scattering. *Science* **275**, 1102–1106 (1997).
14. Feliu, N. *et al.* SERS Quantification and Characterization of Proteins and Other Biomolecules. *Langmuir* **33**, 9711–9730 (2017).
15. Rickard, J. J. S. *et al.* Rapid optofluidic detection of biomarkers for traumatic brain injury via surface-enhanced Raman spectroscopy. *Nat Biomed Eng* **4**, 610–623 (2020).
16. Jahn, I. J. *et al.* Surface-enhanced Raman spectroscopy and microfluidic platforms: challenges, solutions and potential applications. *Analyst* **142**, 1022–1047 (2017).
17. Lussier, F., Thibault, V., Charron, B., Wallace, G. Q. & Masson, J.-F. Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering. *TrAC Trends in Analytical Chemistry* **124**, 115796 (2020).
18. Rebrošová, K. *et al.* Rapid identification of staphylococci by Raman spectroscopy. *Sci Rep* **7**, 14846 (2017).
19. Carlomagno, C. *et al.* COVID-19 salivary Raman fingerprint: innovative approach for the detection of current and past SARS-CoV-2 infections. *Sci Rep* **11**, 4943 (2021).
20. Koster, H. J. *et al.* Surface enhanced Raman scattering of extracellular vesicles for cancer diagnostics despite isolation dependent lipoprotein contamination. *Nanoscale* **13**, 14760–14776 (2021).

21. Banaei, N. *et al.* Machine learning algorithms enhance the specificity of cancer biomarker detection using SERS-based immunoassays in microfluidic chips. *RSC Adv.* **9**, 1859–1868 (2019).
22. Lee, W., Lenferink, A. T. M., Otto, C. & Offerhaus, H. L. Classifying Raman spectra of extracellular vesicles based on convolutional neural networks for prostate cancer detection. *J Raman Spectrosc* **51**, 293–300 (2020).
23. Liu, J. *et al.* Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *Analyst* **142**, 4067–4074 (2017).
24. Yin, G. *et al.* An efficient primary screening of COVID-19 by serum Raman spectroscopy. *J Raman Spectrosc* **52**, 949–958 (2021).
25. Dong, R., Weng, S., Yang, L. & Liu, J. Detection and Direct Readout of Drugs in Human Urine Using Dynamic Surface-Enhanced Raman Spectroscopy and Support Vector Machines. *Anal. Chem.* **87**, 2937–2944 (2015).
26. Noble, W. S. What is a support vector machine? *Nat Biotechnol* **24**, 1565–1567 (2006).
27. Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: supervised methods. *Nat Methods* **15**, 5–6 (2018).
28. Zhu, Q. *et al.* Metabolomic analysis of exosomal-markers in esophageal squamous cell carcinoma. *Nanoscale* **13**, 16457–16464 (2021).
29. Valadi, H. *et al.* Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nat Cell Biol* **9**, 654–659 (2007).
30. Lee, C. *et al.* 3D plasmonic nanobowl platform for the study of exosomes in solution. *Nanoscale* **7**, 9290–9297 (2015).
31. Yu, X. *et al.* An aptamer-based new method for competitive fluorescence detection of exosomes. *Nanoscale* **11**, 15589–15595 (2019).
32. Ignatiadis, M., Sledge, G. W. & Jeffrey, S. S. Liquid biopsy enters the clinic — implementation issues and future challenges. *Nat Rev Clin Oncol* **18**, 297–312 (2021).
33. Vlassov, A. V., Magdaleno, S., Setterquist, R. & Conrad, R. Exosomes: Current knowledge of their composition, biological functions, and diagnostic and therapeutic potentials. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1820**, 940–948 (2012).
34. Zhou, B. *et al.* Application of exosomes as liquid biopsy in clinical diagnosis. *Sig Transduct Target Ther* **5**, 144 (2020).
35. Jalali, M. *et al.* Plasmonic nanobowtiefluidic device for sensitive detection of glioma extracellular vesicles by Raman spectrometry. *Lab Chip* **21**, 855–866 (2021).
36. Hosseini, I. I. *et al.* Nanofluidics for Simultaneous Size and Charge Profiling of Extracellular Vesicles. *Nano Lett.* **21**, 4895–4902 (2021).
37. André-Grégoire, G. & Gavard, J. Spitting out the demons: Extracellular vesicles in glioblastoma. *Cell Adhesion & Migration* **11**, 164–172 (2017).
38. Shin, H. *et al.* Early-Stage Lung Cancer Diagnosis by Deep Learning-Based Spectroscopic Analysis of Circulating Exosomes. *ACS Nano* **14**, 5435–5444 (2020).
39. Zhang, Y. *et al.* Identification and distinction of non-small-cell lung cancer cells by intracellular SERS nanoprobes. *RSC Adv.* **6**, 5401–5407 (2016).
40. Sun, J. *et al.* Detection of glioma by surface-enhanced Raman scattering spectra with optimized mathematical methods. *J Raman Spectrosc* **50**, 1130–1140 (2019).
41. Khan, S. *et al.* Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM). *Biomed. Opt. Express* **7**, 2249 (2016).
42. Li, B. *et al.* Non-invasive diagnosis of Crohn's disease based on SERS combined with PCA-SVM. *Anal. Methods* **13**, 5264–5273 (2021).
43. Ho, C.-S. *et al.* Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nat Commun* **10**, 4927 (2019).

44. Walter, A., März, A., Schumacher, W., Rösch, P. & Popp, J. Towards a fast, high specific and reliable discrimination of bacteria on strain level by means of SERS in a microfluidic device. *Lab Chip* **11**, 1013 (2011).
45. Weng, S. *et al.* Deep learning networks for the recognition and quantitation of surface-enhanced Raman spectroscopy. *Analyst* **145**, 4827–4835 (2020).
46. Dies, H., Raveendran, J., Escobedo, C. & Docoslis, A. Rapid identification and quantification of illicit drugs on nanodendritic surface-enhanced Raman scattering substrates. *Sensors and Actuators B: Chemical* **257**, 382–388 (2018).
47. Gan, H. K., Kaye, A. H. & Luwor, R. B. The EGFRvIII variant in glioblastoma multiforme. *Journal of Clinical Neuroscience* **16**, 748–754 (2009).
48. Choi, D. *et al.* The Impact of Oncogenic EGFRvIII on the Proteome of Extracellular Vesicles Released from Glioblastoma Cells. *Molecular & Cellular Proteomics* **17**, 1948–1964 (2018).
49. Garnier, D. *et al.* Divergent evolution of temozolomide resistance in glioblastoma stem cells is reflected in extracellular vesicles and coupled with radiosensitization. *Neuro-Oncology* **20**, 236–248 (2018).
50. Moreno, M. *et al.* Raman spectroscopy study of breast disease. *Theor Chem Acc* **125**, 329–334 (2010).
51. Rygula, A. *et al.* Raman spectroscopy of proteins: a review. *J. Raman Spectrosc.* **44**, 1061–1076 (2013).
52. Wang, P. *et al.* Giant Optical Response from Graphene–Plasmonic System. *ACS Nano* **6**, 6244–6249 (2012).
53. De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* **50**, 1–18 (2000).
54. Mahalanobis, P. C. On the Generalized Distance in Statistics. *Proceedings of the National Institute of Science of India* **12**, 49–55 (1936).

3.7 Supporting information: Support Vector Machine for classification of Surface Enhanced Raman Spectroscopy spectra of Cancerous Single Extracellular Vesicles

Olivia Jeanne[1], Carolina del Real Mata[1], Mahsa Jalali[1], Laura Montermini[2], Yao Lu[1], Kevin Petrecca[3], Janusz Rak[2], Sara Mahshid*

[1] *Department of Bioengineering, McGill University Montreal, QC, Canada*

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

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

sara.mahshid@mcgill.ca

3.7.1 Experimental Methods

Data collection

The datasets used here are integrated by single-EV spectra collected in the Mahshid lab, the spectra were obtained and processed as previously described in Jalali et al[1]. Briefly, EVs were isolated from a non-cancerous Normal Human Astrocytic cell line (NHA), two glioma cell lines (U87, U373), and three glioma stem cell lines (GSC83, GSC1005 and GSC1123). Additionally, real human samples EVs were also isolated from the blood of healthy donors and confirmed glioblastoma patients' plasma, prior to undergoing surgery. The EVs were loaded on a nanostructured microchip previously developed in Mahshid lab, containing a SERS-active patterned nanocavity array which employs the synergy of materials and structures to ensure a single EV spectrum acquisition. The SERS spectra were then preprocessed by means of baseline subtraction, smoothing, and normalization, to obtain the datasets used by the ML algorithms, made of one-dimensional vectors of 1245 observations corresponding to the intensity value as a function of the wavelength (Figure 2a).

Cell line classification

To explore the potential of the SVM algorithm on our datasets, two studies were performed using SERS spectra, the first studied different cell lines using a multiclass SVM, and the second used a binary SVM to classify patient samples. For the SVM multiclass classification, the library comprised 6 different cell lines (Glioma U87, Glioma U373, GSC1005, GSC83, Glioma 1123, and NHA) and various mutations on these cell lines. The dataset of each cell line type is then formed from a total of 946 single EVs collected spectra. The 11 cell lines (considering the mutations cell lines) correspond respectively to the 11 classes that integrate the dataset used for this study. The SVM is then trained, using 70% of the previously mentioned datasets as its training set.

Patient classification

For the patient classification another SVM algorithm, this time performing binary classification, is trained and tested on a clinically relevant dataset, integrated by 1267 spectra from plasma samples of glioma patients and healthy donors. This dataset is generated by EVs samples derived from plasma of eight healthy donors and ten diagnosed patients' plasma. The training set consists of 80% of the spectra acquired from patient samples 3, 5, and 8, from healthy donors 1,2,3, and 4, and the U373 viii, U87 viii, and MGMT 1123-9 cell lines. The remaining patient and healthy spectra are used for testing. One-way ANOVA is used in the binary classification to verify that the difference between the mean scores of healthy and patient spectra is statistically significant at the level of 0.001 ($F(1,711)=398$, $p=10^{-70}$).

Machine learning

In order to achieve the best results possible, hyperparameter searches were performed using Bayesian search[2]. A 5-fold cross validation (CV) was performed on the training set to avoid overfitting, meaning that the training set was divided into 5 subsets and 4 were used for training while the fifth was used as a validation set. This process is repeated with each subset successively held out and used as a validation set. For the 11-class classification, the search space was: C and gamma between 10^{-6} and 100, with a log-uniform distribution, and a choice of kernel of either rbf or linear. The hyperparameters thus obtained are a linear kernel and C: 1.81. For the binary classification, the search space was C and gamma between 10^{-6} and 100, with a radial basis function kernel. The hyperparameters obtained are a radial basis function (rbf) kernel and C:100, gamma: 0.13. All the spectra preprocessing is done using Origin Pro 2019b software, WiRE 5.1 and Python. The SVM is coded using the scikit-learn module of Python[3].

3.7.2 Additional references:

1. Jalali, M. *et al.* Plasmonic nanobowtiefluidic device for sensitive detection of glioma extracellular vesicles by Raman spectrometry. *Lab Chip* **21**, 855–866 (2021).
2. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE* **104**, 148–175 (2016).
3. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 6 (2011).

Intermediate conclusion and transition

SVM has shown its potential in analyzing and leveraging high dimensional and complex SERS spectra to obtain a clinical diagnosis of GBM patients. It is also able to distinguish between multiple mutations occurring in a variety of cells. This work opens the way for the analysis of complex tumors, with multiple mutations, using EVs extracted from the blood of suspected patients. It also has potential for diagnosis of GBM, as a complement to be used in conjunction with the traditional clinical techniques. Using SVM, analyses could also be performed repeatedly over an extended period of time, enabling to follow tumoral evolution. This study holds promise for future automated analysis and interpretation of SERS spectra in a clinical setting.

Machine learning is known notably for its extreme versatility, allowing it to analyze very different data sets. It is therefore fair to ask if SVM can be used in another setting, on another type of optical readout with similar success, or if the success we have observed will be limited to highly complex SERS data. This is what we explore in the next chapter, where we use SVM for colorimetric detection and analysis. We want to determine whether SVM can be helpful in the context of rapid tests, by analyzing images of a colorimetric platform for infectious disease detection. In our case, the disease we are analyzing is Covid-19. Colorimetric RT-LAMP is performed on a plasmonic color sensitive platform, driving the color change that is being analyzed. In this colorimetric study, both the type of disease investigated and the data used for the analysis are very different than what has been studied in the previous chapter, as we observe an infectious viral disease, using images instead of spectra. We analyze clinical samples from 34 patients, divided in 15 naso-pharyngeal swabs and 18 saliva samples, and 15 negative controls. We determine the parameters necessary to obtain optimal results using this testing platform by experimenting on two different color spaces and multiple timepoints.

4 Nano-plasmonically Boosted Nucleic Acid Amplification Coupled with Support Vector Machine for Minute Colorimetric Classification of Viral RNA

Olivia Jeanne[1], Carolina del Real Mata[1], Tamer Abd El Fatah[1], Mahsa Jalali[1], Haleema Khan[1], Sara Mahshid[1]*

[1] *Department of Bioengineering, McGill University Montreal, QC, Canada*

**sara.mahshid@mcgill.ca*

List of figures

List of tables

4.1 Abstract

Data science and machine learning's rapid development over the past years have allowed scientists to analyze medical and clinical data in a user-friendly, higher-throughput, and less personnel-extensive way, assisting clinicians in diagnosing and monitoring health issues. Machine learning has the potential for accurate, low-cost sensing and diagnosis with the possibility of being deployed in environments where high-end medical facilities are unavailable, making it an ideal candidate for point-of-care testing. Among readout techniques, colorimetry is a promising low-cost and simple sensing method that can be coupled with machine learning for diagnostic applications. Here, a support vector machine (SVM) is used to analyze the data from a plasmonic color sensitive chip for rapid and point-of-care detection of viral RNA. We illustrate its diagnostic capability using a paradigm of viral respiratory infection such as Covid-19. The device uses a highly sensitive RT-LAMP assay for the detection of viral RNA monitored through color change. In order to obtain a rapid and accurate diagnosis, images of the plasmonic color sensitive device are collected with a basic imaging system and analyzed via SVM, achieving a 94% success rate in the classification of healthy vs sick patients after 10 minutes. This point-of-care system would help to prevent the fast spread of infectious diseases through rapid screening operations.

Keywords: Machine learning, SVM, Colorimetry, RT-LAMP, Diagnostics, Sensors

4.2 Introduction

Machine learning analysis offers a solution to some challenges faced today by modern medicine[1]. Enabling automated detection of diseases and health parameter surveillance can improve treatment quality and pave the way towards personalized medical care[1]. Machine learning algorithms have demonstrated their value in classifying and interpreting clinically

relevant data, especially considering the increasing amount of information being generated as sensing technologies progresses[2]. It also offers high throughput processing capabilities, high versatility, and has a strong ability to uncover, and use complex non-linear relationships within the given data[3,4]. Also, these algorithms allow to obtain a satisfactory analysis of low-resolution or noisy data, which would have been discarded otherwise[5]. A common practice is to couple machine learning models with state-of-the-art and high-sensitivity sensors. Among the possible readouts, optical techniques offer high sensitivity and specificity, and a good signal-to-noise ratio making them an attractive option[6,7].

Support vector machine (SVM), a type of supervised machine learning, has the valuable ability to achieve high accuracy during classification despite being trained on a reduced number of data points. This holds true even when the dataset is made of instances that are composed of a high number of variables compared to its size[8]. SVM algorithms have been applied over the years to a multitude of data, including colorimetric images in the medical field (Table 1). For its various advantages, including its simplicity, SVM is an ideal candidate to be incorporated into the analysis of sensing applications.

The translation of optical readouts into highly sensitive and low-cost point-of-care applications is challenged by their frequent need for bulky and expensive equipment. However, colorimetry has multiple advantages such as rapidity, cost-effectiveness, and easy integration into portable point-of-care settings. In this context, the color change in colorimetric assays is a suitable readout as it can be detected by the naked eye or using a basic bright-field microscope [9,10]. While a colorimetric readout is simple to interpret, the accurate identification of the color change is susceptible to factors such as the time of observation and personnel skills, which hinders the reliability of the diagnosis. This is where the potential of machine learning can best be leveraged, for improving the detection methods of infectious diseases, an asset crucial in the

context of global pandemics. We demonstrate its applicability using a paradigm of Covid-19 respiratory infection. The ongoing Covid-19 pandemic has exposed the vulnerabilities in healthcare systems. SARS-CoV-2 high infection rate is attributed to asymptomatic and presymptomatic spread and airborne transmission through both saliva droplets and aerosols[11,12]. Reverse transcription polymerase chain reaction (RT-PCR) is the gold standard method for SARS-CoV-2 detection. However, RT-PCR can only be performed in centralized facilities which results in delays for the processing and analysis of the samples [13]. A rapid and early diagnosis is as crucial to controlling the spread of SARS-CoV-2 as it is challenging[14]. In consequence, there is a pressing need for portable, affordable, high throughput point-of-care devices to perform reliable Covid-19 diagnosis.

Here, we employ a SVM algorithm with a radial-basis function (rbf) kernel to assist in the interpretation of images of a colorimetric assay (Figure 4.1). The images were collected using a plasmonic color sensitive device developed in Mahshid Lab (AbdElFatah et al, manuscript under preparation) which addresses the challenges of detection methods of infectious diseases, previously described. The device uses an isothermal reverse transcriptase loop-mediated isothermal amplification (RT-LAMP) for the detection of viral RNA via a colorimetric readout. The plasmonic nano-surface of the device allows plasmonic-induced enhancement of the RT-LAMP reaction through surface free electron injection from the surface to the media, dramatically reducing detection time. The nucleic acid amplification from the RT-LAMP releases H^+ ions, which are then detected using a pH-sensitive dye, phenol red[15]. This approach is coupled with a machine learning analysis to generate a, in 10 minutes. The database used for the SVM analysis is integrated by images taken from clinical samples of 34 patients admitted with Covid-19 symptoms, and healthy controls. The proposed practical colorimetric SVM supports the capabilities of the analysis approach to be of assistance in the diagnosis of

infectious diseases, paving the way for machine learning-assisted diagnostics in low resource and point-of-care settings, where speed and high throughput are a priority.

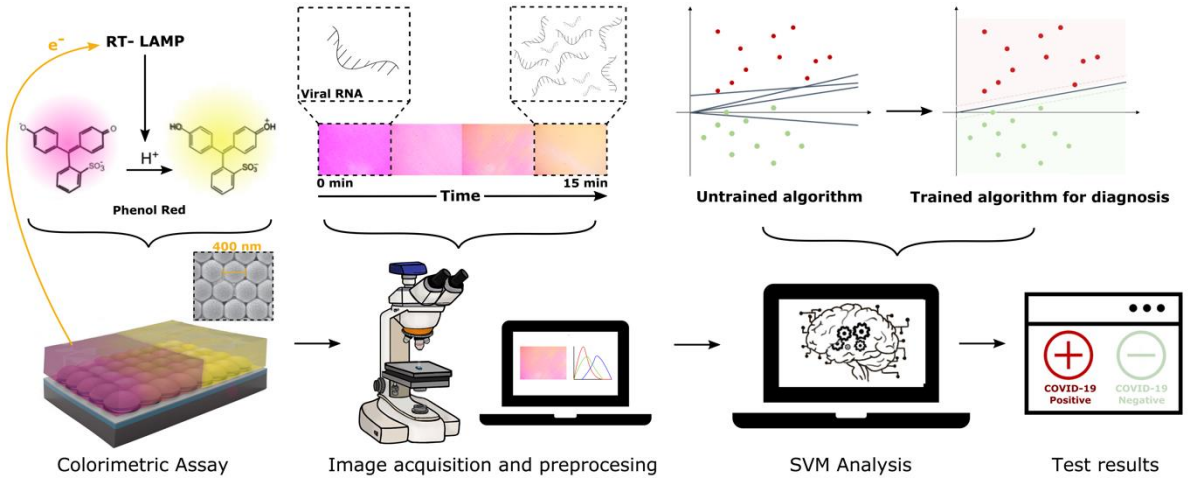

Figure 4.1: Workflow of the viral RNA detection. The sample is loaded onto the plasmonic color sensitive device where isothermal reverse transcriptase loop-mediated isothermal amplification (RT-LAMP) of the viral RNA is performed. The amplification is enhanced by the injection of free electrons originating at the plasmonic nano-surface of the device. RT-LAMP releases H+ which allows for its monitoring through the color change of a pH-sensitive dye, Phenol Red. The colorimetric assay is surveilled by image acquisition using a bright field microscope. An SVM algorithm, previously trained is used to analyze the data collected and generate a prediction of the health status of the patient: positive or negative to Covid-19.

4.3 Methods

4.3.1 Data collection

The data is obtained from analyzing 34 patients samples and 15 healthy controls. The samples are loaded onto the plasmonic color sensitive device which is then maintained at 65°C to enable the plasmonically enhanced RT-LAMP reaction, during which the colorimetric change is monitored. Images of the platform are acquired regularly over 15 minutes, at 1, 3, 7, 10, 15 min, to monitor the color change as a function of time. The preprocessing starts with the images

being cropped to 80% of the initial size, to remove areas potentially impacted by the "coffee ring effect". Pixels with a hue value between 85 and 140 (blue range) are removed and replaced by the mean value of the rest of the image. The image is then thresholded by removing the 25% least saturated parts of the image and substituting them with the mean value of the rest of the image. Finally, the image is divided into 20 smaller images from which we extract 6 features: mean color value, standard deviation, mode, skew, energy, and entropy in each color channel r, g, and b, for a grand total of 18 values. When studying the L*a*b* color space, the same procedure is performed using the L*, a*, and b* channels. The formulas used for each feature are adapted from Sergyan[16]. The grey-scale intensity was replaced with the corresponding intensity for the red, green, and blue (and L*, a*, and b*) channels successively.

The dataset used is integrated by the 34 patient samples and 15 healthy controls divided into a training and a testing set. For every sample, 9 images are acquired per timepoint. The training set consists of 2 thirds of the vectors from patients 1, 7, 9, 11, 15, 19, 21, 23, 25, 29, 31 and negatives 2, 3, 4, 5, 6, 7, 8, 9, 11, 13 and 15; the test set integrates the remaining vectors. The data is divided into 2 classes: healthy (negative) and sick (positive).

When studying the performance of the algorithm to monitor the color change as a function of time, each timepoint is evaluated independently using the data only from that corresponding timepoint for both training and testing.

4.3.2 Machine learning

An SVM with a rbf kernel is used to obtain a prediction for each vector of the test set, either as healthy or as sick. The hyperparameters C and gamma are determined through a Bayesian search[17], the search space being C and gamma both ranging from 10^{-4} to 100 in a log-uniform

distribution. The absence of overfitting is validated using 5-fold cross-validation. The entire analysis pipeline is summarized in Figure 4.2a.

To analyze patients' samples, all the vectors derived from all images of one microscopy time point are tested in one pool. The probability of being determined as Covid-19 positive is obtained for all the vectors, and the average probability for each patient is then calculated. The probabilities are derived from the SVM scores using Platt scaling[18].

Another method was tested, where instead of considering each vector as an independent observation, a result for each entire image is calculated. The vectors from a single of the 9 images in a timepoint are tested, then a prediction for the entire image is obtained. The same procedure is followed with the other 8 images of the timepoint and finally an average is calculated over all the images for one patient. However, this method was abandoned for being more computationally expensive to implement and not yielding significantly better results.

All image processing is done in Python using the cv2 and scikit-image modules. The SVM is coded using the scikit-learn module of Python[19].

4.4 Results and discussion

Colorimetric methods present a number of advantages such as rapidity and user-friendliness, making them an ideal candidate for point-of-care settings. They can be implemented using various methods, such as paper-based colorimetric tests, or in solutions, such as for Enzyme-Linked Immunosorbent Assays, but also using nanoparticles and their plasmonic properties that lead to color generation[9,20,21]. We have chosen a colorimetry assay and a plasmonic nano-surface and combined them to form plasmonic color sensitive device, offering the additional advantages of being rapid, portable and cost-effective.

Images of the platform were taken at multiple timepoints, preprocessed, and analyzed via an SVM algorithm. Examples of the images taken for monitoring the color change are shown in Figure 2b. The sensing chamber hosts a platform which has an initial intense pink color for both the negative and the positive samples due to the presence of the pH-sensitive dye phenol red. Then, for the positive sample, as the incubation time increases, nucleic acids are amplified through the RT-LAMP reaction and further enhanced by the plasmonic nano-surface. This results in the release of H^+ ions which change the pH of the solution, to which phenol red reacts producing a color change in the solution[15]. Thus the color gradually changes to orange, while the sensing chamber of the negative sample stays pink. However, it is difficult to distinguish by naked eye at what point the platform starts changing color in comparison to the negative sample. Consequently, we have used an SVM with a rbf kernel as a way to solve this challenge.

Table 4.1: Machine Learning models used in medical analysis with colorimetric readout

Study	Analyte	Color Assay	ML Type	Efficiency	Ref
Determination of glucose in saliva using TMB, KI+Chitosan, KI	Glucose	Paper based microfluidic device	LDA, GBC, RF	For TMB, KI+Chi, KI: LDA: Acc= 98.24%; 75.45%; 74.13% GBC: Acc= 93.30%; 83.04%;75.98% RF: Acc= 96.46%; 78.63%; 76.83%	[22]
Determination of bilirubin in urine	Bilirubin	Chloroauric acid immobilized on a paper strip	RF	Acc=74.05%	[23]
Classification of a TB antigen test	TB antigen	ELISA	k-means, QDA, k-NN, RF, ANN	k-means: Acc=68.1% QDA: Acc=93.7% RF: Acc=96.1% k-NN: Acc=97.6% ANN: Acc=99.2%	[24]
Determine alcohol concentration in saliva	Alcohol	Paper test strip	LDA, SVM, ANN	LDA: PPV=95% SVM: PPV=99% ANN: PPV=98%	[25]

				(Standard concentrations in L*a*b* color space)	
Monitoring skin pH	Skin moisture	pH-responsive hydrogels	Linear regression	r = 0.93; P< 0.01 MAE = 0.27	[26]
Serodiagnosis of early-stage Lyme disease	Anti-Lyme human IgG and IgM	Paper based vertical flow assay	Neural Network	AUC= 0.963 Acc= 91.7% Sens=85.7% Spe=96.3%	[27]
Malaria detection	DNA	Paper-based microfluidic lateral flow DNA molecular assay	CNN	Acc= 97.83%	[28]
Classification of SARS-CoV-2 saliva samples	SARS-CoV-2	RT-LAMP	SVM	Acc=94% (Patient samples)	This work

TMB: 3,3' ,5,5' tetramethylbenzidine, KI: potassium iodide, LDA: Linear Discriminant Analysis, GBC: Gradient Boosting Classifier, RF: Random Forest, ELISA: Enzyme Linked Immunosorbent Assay, QDA: Quadratic Discriminant Analysis, k-NN: k-Nearest Neighbours, ANN: Artificial Neural Network, PPV: Positive Predictive Value, MAE: Mean Absolute Error, CNN: Convolutional Neural Network

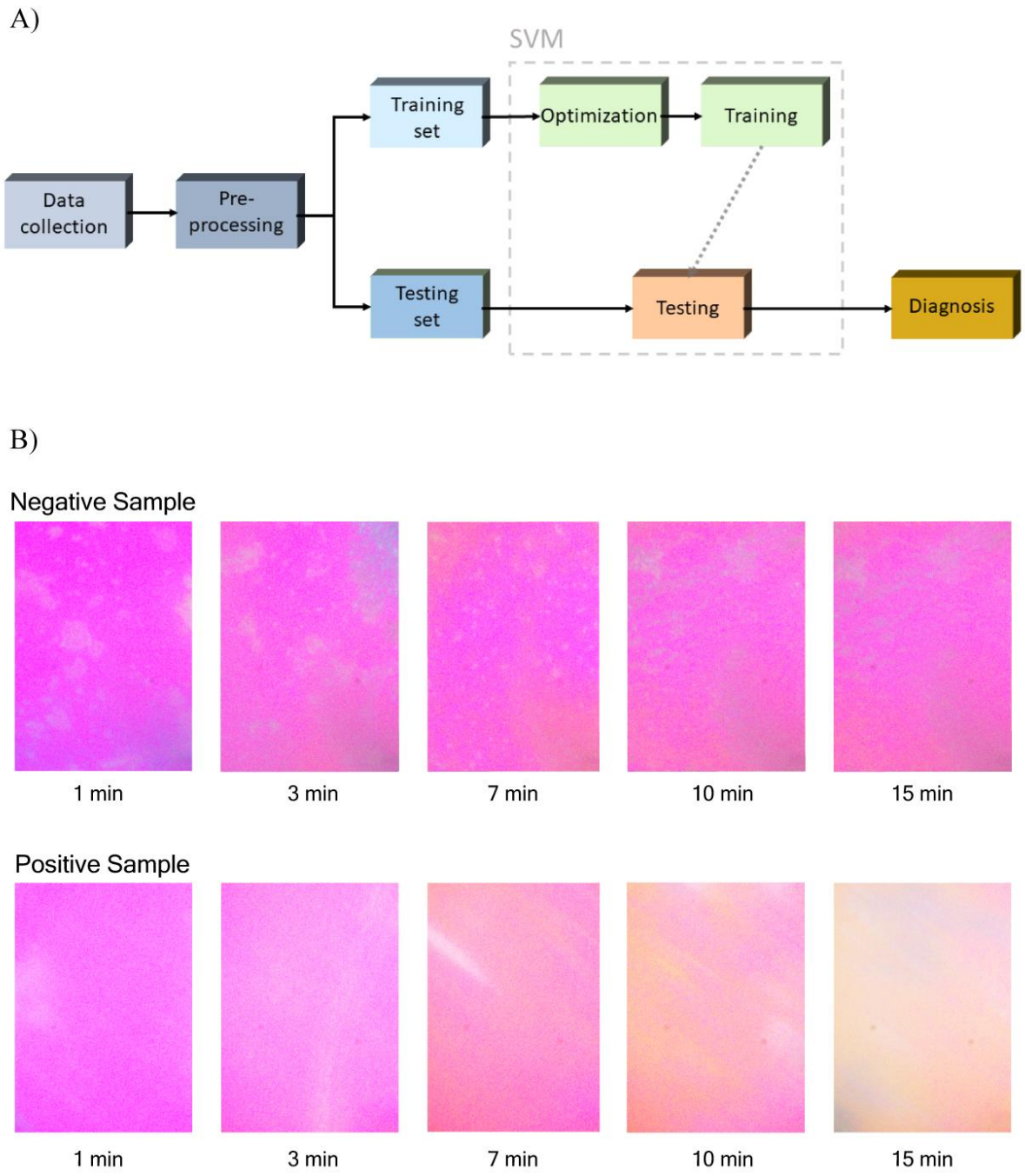

Figure 4.2: Flowchart and color matrix of the samples. A) Flowchart describing the analysis process, from image collection to final diagnosis. B) Color comparison between the colorimetric platforms of a negative control and a positive saliva sample. The time shown is the incubation time after virus lysis. The initial color for both platforms is pink, and only the positive sample presents a color change towards orange.

SVM is a supervised machine learning technique that classifies data based on its position with regards to a separating hyperplane[28,29]. The optimal hyperplane is determined during the training phase so that it best separates the data[29]. To start the analysis through SVM, the collected dataset of patients samples was split between training and testing sets (see Methods section). This is done to be able to test the SVM on images it has not seen before, therefore giving a representative indication of its performance under realistic conditions. The status of the patient samples as Covid-19-positive was verified by performing PCR on the samples. In order to determine what incubation time yields the best results of analysis, we compared three parameters, accuracy, sensitivity, and specificity of the different timepoints collected. These three parameters are defined as:

$$Accuracy = \frac{TP + TN}{n} \quad (eq. 1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (eq. 2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (eq. 3)$$

Where true positive (TP) and true negatives (TN) are the patients and healthy controls correctly classified as positive or negative to SARS-CoV-2, respectively. False positive (FP) and False negative (FN) were incorrectly classified, and *n* is the total number of samples used.

In addition to this, we have compared two color spaces, to determine which gives the best results. We have used the RGB color space, with additional features, and the CIE L*a*b* color space, with the same additional features, computed using the L*, a*, b* channels. In Table 2 we show the results of both color spaces for the different time points. L*a*b* and RGB perform similarly for the different timepoints, with L*a*b* yielding better results on the 10 and 15-minute timepoints.

Table 4.2: Accuracy, sensitivity, and specificity both in RGB and L*a*b* color space. L*a*b* performs better than the RGB color space in the 10 and 15 min timepoints.

		Timepoints				
		1 min	**3 min**	**7 min**	**10 min**	**15 min**
Accuracy	**RGB**	58,64	68,73	75,16	85,07	80,34
	L*a*b*	64,73	61,88	69,75	86,06	88,66
Sensitivity	**RGB**	57,33	67,60	74,87	85,66	75,78
	L*a*b*	67,19	58,51	66,86	86,64	86,14
Specificity	**RGB**	63,69	72,76	76,23	82,97	97,46
	L*a*b*	55,48	73,94	80,33	84,08	97,21

Accuracy, specificity and sensitivity on the test set each steadily increase as time passes, with the best accuracy and specificity reached at 10 min for the RGB color space (Figure 4.3a) and 15 min for the L*a*b* color space (Figure 4.3b).

As time is a crucial factor in point-of-care diagnosis applications, where high throughput is also essential, we chose to use the 10-minute timepoint as being the best compromise for a rapid test, while still obtaining robust effective predictions. Indeed, accuracy and sensitivity barely increase between 10 and 15 minutes in the L*a*b* space and decrease in RGB space. The ROC (Receiver Operating Characteristic) curves and corresponding AUCs (Area Under the Curve) also support this decision, with the 10 and 15-minute timepoints having the highest AUC. (Figure 4.3c, d). As a result, the 10-minute timepoint is used for patient experiments.

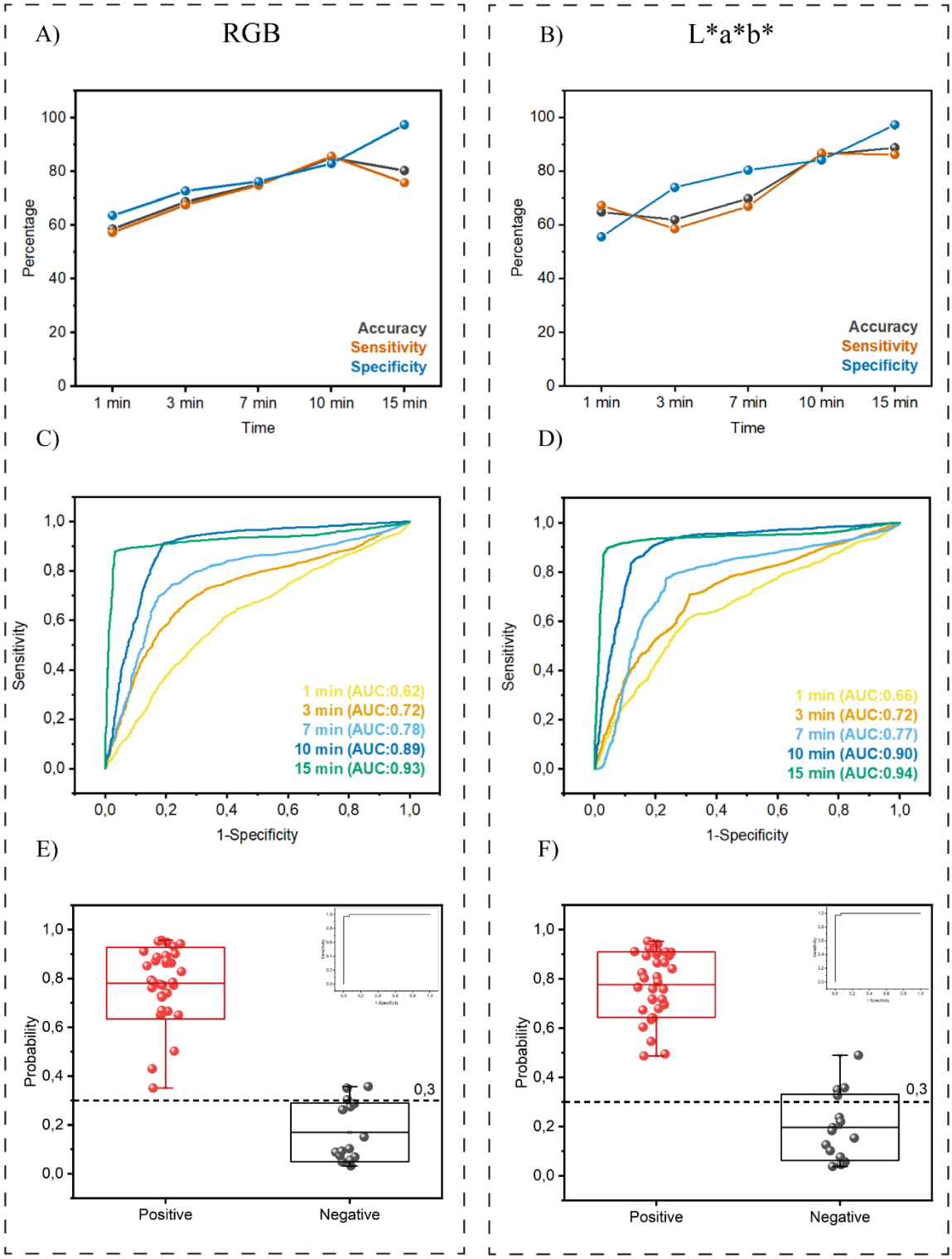

Figure 4.3: Diagnosis of Covid-19 in patients using the colorimetric point-of-care platform using SVM. A) Accuracy, Sensitivity, and Specificity of the SVM as the time increases until 10 minutes in RGB and B) in L*a*b* color spaces. C) ROCs for the different time points, with the AUCs specified in RGB and D) in L*a*b* color spaces. E) For each clinical sample obtained, the average probability of being predicted as Covid-19 positive and the threshold

established for the best specificity and sensitivity using the RGB color space. Inset: the ROC obtained from the probabilities and used to calculate the optimal threshold (AUC: 0,998). F) the average probability of being predicted as Covid-19 positive using the L*a*b* color space. Inset: the ROC obtained from the probabilities and used to calculate the optimal threshold (AUC: 0,998). The boxes represents 1 SD, the whiskers show the outliers.

The test samples were analyzed after an incubation time of 10 minutes, as this timepoint was selected for offering the best compromise between speed and successful prediction. Each image taken is, as previously described, subdivided into 20 sub-images and each sub-image is classified by the SVM. The probability of being a Covid-19 positive sample is obtained, and the average for all the sub-images for each sample is reported (Figure 4.3e, f). The results for the 7 and 15-minute timepoints are also determined for comparison (Figure S1), with again 10 minutes being the best compromise. For the final Covid-19 diagnosis for each sample, a cut-off value needs to be determined. A natural candidate would be 50%: if, on average, more than 50% of the image is predicted as positive, then the sample is classified as positive. However, it is possible to determine an optimal threshold, that performs better by using the ROC for each patient (Figure 4.3e and f insets) and finding the threshold that minimizes the distance to the point (0;1). The threshold is therefore determined to be 0.30, allowing to reach 92% accuracy, 100% sensitivity, and 73% specificity in the RGB color space, and 94% accuracy, 100% sensitivity and 80% specificity in the L*a*b* color space (Table 4.3). Interestingly, the threshold is the same for both color spaces, demonstrating good reproducibility regardless of the one chosen. L*a*b* has a slightly higher accuracy than RGB when classifying patients at 10 minutes, however, converting an image from RGB to L*a*b* color space adds an extra step to the image analysis. Therefore, it is possible to tailor the image analysis to the context, or to the resources available. Indeed, for point-of-care testing, especially if using computer or mobile devices with limited computational power, RGB might be preferred. If more advanced devices

are available, then the extra step of converting from RGB to L*a*b* can be done easily, and the user can benefit from improved accuracy.

Table 4.3: Results of the SVM for the patient diagnosis, using the 0.3 decision threshold.

	Sensitivity	Specificity	Accuracy	TP	TN	FP	FN
RGB	100%	73.33%	91.84%	34	0	4	11
L*a*b*	100%	80%	93.87%	34	0	3	12

4.5 Conclusion

Our studies using SVM have proven successful in plasmonic-assisted rapid detection of SARS-CoV2 viral RNA using phenol red-based colorimetric assays. Multiple timepoints were studied to find the point where the SVM model performs at its best while keeping detection time low, thus, enabling high throughput. RGB and L*a*b* color spaces were investigated, with L*a*b* yielding slightly better results than RGB. The 10-minute timepoint was chosen as the best compromise between speed and accuracy. Individual real patient samples were tested at this timepoint, yielding up to 94% accuracy, 100% sensitivity, and 80% specificity. The SVM demonstrated its capability to detect infection in clinical samples, making it a highly versatile testing method. Rapid and high throughput tests are critical for the monitoring of the spread of a disease, especially in the context of a pandemic, and this platform, coupled to ML analysis, could be a robust candidate to mitigate the impact of present and future pandemics.

SVM is a type of algorithm that does not require high numbers of training data, nor high performance computers for its use[8]. While it is true that deep learning approaches can achieve higher accuracies, they remain dependent on training data in sufficient amounts and high-performance computers, or cloud platforms to operate. The diffusion of such computers and the

creation of massive open clinical databases allow clinicians and researchers to use the newest deep learning approaches. However, a lighter machine learning algorithm such as SVM is more suited to low-resource and point-of-care settings, while still being able to predict diseases in a clinically relevant manner, with 94% accuracy in the case of Covid-19. Indeed, the data for this work was acquired using a brightfield microscope with a color camera. Coupling SVM with truly portable optical approaches is the next ideal step in this project. Translation to the point-of-care is challenging, however, apparatus such as mobile imaging boxes with fully portable imaging systems, like the one developed in Mahshid Lab (AbdElFatah et al, manuscript under preparation) have the potential to solve this issue. Overall, coupling SVM with plasmonic color sensitive devices would prove to have great diagnostic potential in point-of-care and low resource settings, with high accuracy and throughput, critical to preventing contagion in populations.

4.6 Acknowledgements

The authors thank the Faculty of Engineering at McGill University, Natural Science and Engineering Research Council of Canada (NSERC, G247765), Canadian Institutes of Health Research (CIHR, 257352), Canada Foundation for Innovation (CFI, G248924) for financial support. CdRM thanks McGill Engineering Award (MEDA and Fonds du Recherche du Quebec (FRQnet).

4.7 References

1. Archenaa, J. & Anita, E. A. M. A Survey of Big Data Analytics in Healthcare and Government. *Procedia Computer Science* **50**, 408–413 (2015).
2. Fang, R., Pouyanfar, S., Yang, Y., Chen, S.-C. & Iyengar, S. S. Computational Health Informatics in the Big Data Age: A Survey. *ACM Comput. Surv.* **49**, 1–36 (2016).
3. Zitnik, M. *et al.* Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion* **50**, 71–91 (2019).

4. Rajkomar, A., Lingam, S., Taylor, A. G., Blum, M. & Mongan, J. High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. *J Digit Imaging* **30**, 95–101 (2017).
5. Cui, F., Yue, Y., Zhang, Y., Zhang, Z. & Zhou, H. S. Advancing Biosensors with Machine Learning. *ACS Sens.* **5**, 3346–3364 (2020).
6. Cooper, M. A. Optical biosensors in drug discovery. *Nat Rev Drug Discov* **1**, 515–528 (2002).
7. Damborský, P., Švitel, J. & Katrlík, J. Optical biosensors. *Essays in Biochemistry* **60**, 91–100 (2016).
8. Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: supervised methods. *Nat Methods* **15**, 5–6 (2018).
9. Piriya V.S, A. *et al.* Colorimetric sensors for rapid detection of various analytes. *Materials Science and Engineering: C* **78**, 1231–1245 (2017).
10. Zhao, V. X. T., Wong, T. I., Zheng, X. T., Tan, Y. N. & Zhou, X. Colorimetric biosensors for point-of-care virus detections. *Materials Science for Energy Technologies* **3**, 237–249 (2020).
11. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* **26**, 672–675 (2020).
12. Santarpia, J. L. *et al.* Aerosol and surface contamination of SARS-CoV-2 observed in quarantine and isolation care. *Sci Rep* **10**, 12732 (2020).
13. Carter, L. J. *et al.* Assay Techniques and Test Development for COVID-19 Diagnosis. *ACS Cent. Sci.* **6**, 591–605 (2020).
14. Mahshid, S. S., Flynn, S. E. & Mahshid, S. The potential application of electrochemical biosensors in the COVID-19 pandemic: A perspective on the rapid diagnostics of SARS-CoV-2. *Biosensors and Bioelectronics* **176**, 112905 (2021).
15. Tanner, N. A., Zhang, Y. & Evans, T. C. Visual detection of isothermal nucleic acid amplification using pH-sensitive dyes. *BioTechniques* **58**, 59–68 (2015).
16. Sergyan, S. Color histogram features based image classification in content-based image retrieval systems. in *2008 6th International Symposium on Applied Machine Intelligence and Informatics* 221–224 (IEEE, 2008). doi:10.1109/SAMI.2008.4469170.
17. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE* **104**, 148–175 (2016).
18. Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classification* **10**, (2000).
19. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 6 (2011).
20. Silva-Neto, H. A., Sousa, L. R. & Coltro, W. K. T. Colorimetric paper-based analytical devices. in *Paper-based Analytical Devices for Chemical Analysis and Diagnostics* 59–79 (Elsevier, 2022). doi:10.1016/B978-0-12-820534-1.00009-8.
21. Li, Z. *et al.* Plasmonic-based platforms for diagnosis of infectious diseases at the point-of-care. *Biotechnology Advances* **37**, 107440 (2019).
22. Mercan, Ö. B., Kılıç, V. & Şen, M. Machine learning-based colorimetric determination of glucose in artificial saliva with different reagents using a smartphone coupled µ PAD. *Sensors and Actuators B: Chemical* **329**, 129037 (2021).
23. Edachana, R. P. Paper-based device for the colorimetric assay of bilirubin based on in-situ formation of gold nanoparticles. *Microchim Acta* 9 (2020).
24. Tania, M. H., Shabut, A. M., Lwin, K. T. & Hossain, M. A. Clustering and Classification of a Qualitative Colorimetric Test. in 7–11 (2018). doi:10.1109/iCCECOME.2018.8658480.

25. Kim, H., Awofeso, O., Choi, S., Jung, Y. & Bae, E. Colorimetric analysis of saliva–alcohol test strips by smartphone-based instruments using machine-learning algorithms. *Appl. Opt.* **56**, 84 (2017).
26. Baik, S. *et al.* Diving beetle–like miniaturized plungers with reversible, rapid biofluid capturing for machine learning–based care of skin disease. *Sci. Adv.* **7**, eabf5695 (2021).
27. Joung, H.-A. *et al.* Point-of-Care Serodiagnostic Test for Early-Stage Lyme Disease Using a Multiplexed Paper-Based Immunoassay and Machine Learning. *ACS Nano* **14**, 229–240 (2020).
28. Guo, X. *et al.* Smartphone-based DNA diagnostics for malaria detection using deep learning for local decision support and blockchain technology for security. *Nat Electron* **4**, 615–624 (2021).
29. Noble, W. S. What is a support vector machine? *Nat Biotechnol* **24**, 1565–1567 (2006).
30. Xu, Y., Zomer, S. & Brereton, R. G. Support Vector Machines: A Recent Method for Classification in Chemometrics. *Critical Reviews in Analytical Chemistry* **36**, 177–188 (2006).

4.8 Supporting information: Nano-plasmonically Boosted Nucleic Acid Amplification Coupled with Support Vector Machine for Minute Colorimetric Classification of Viral RNA

Olivia Jeanne[1], Carolina del Real Mata[1], Tamer AbdElFatah[1], Mahsa Jalali[1], Haleema Khan[1], Sara Mahshid[1]*

[1] *Department of Bioengineering, McGill University Montreal, QC, Canada*

sara.mahshid@mcgill.ca

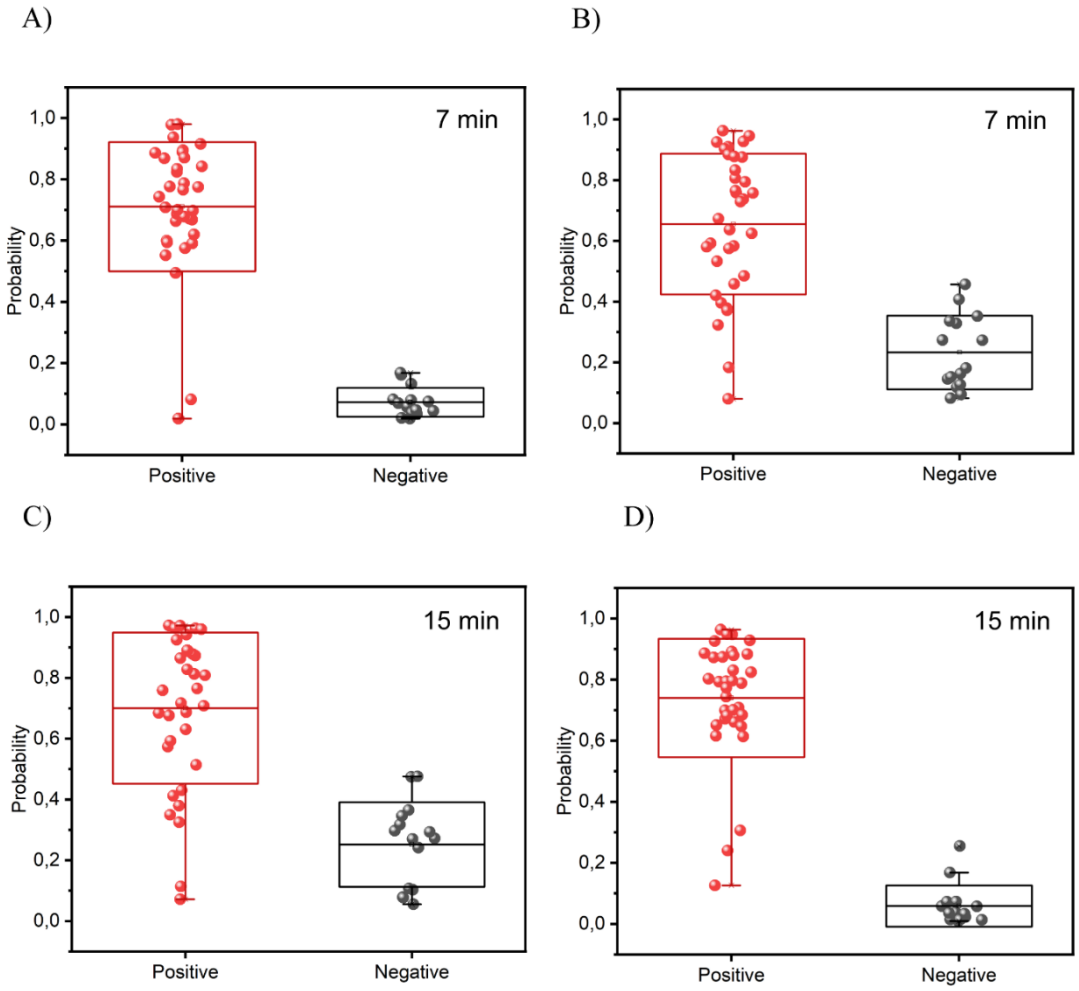

Figure 4.S1: Average probabilities for the 7 and 15 min timepoints. A) For each clinical sample obtained, the average probability of being predicted as Covid-19 positive at 7 min in RGB color space and B) average probability of being predicted as Covid-19 positive at 7 min in L*a*b* color space. C) For each clinical sample obtained, the

average probability of being predicted as Covid-19 positive in RGB color space and D) in the L*a*b* color space. The boxes represent 1 SD, the whiskers show the outliers.

5 Comprehensive discussion of findings

5.1 Comparison with literature

In this work, two analyses are presented, using SVM for the classification of optical data: SERS spectra and colorimetric images. The results on the binary classifications are excellent, with over 94% accuracy both when using SERS spectra, for the GBM task, and colorimetric images, for the Covid-19 task. This is in line with previous studies using SERS spectra and colorimetric images. Indeed, glioma detection using SERS and SVM achieved 97.9% accuracy [85], Covid-19 detection through Raman spectroscopy achieved 90% accuracy [73], and previous colorimetry analyses are also in this range [20,96,118]. Some rapid Covid-19 testing devices approved by the U.S. Food and Drug Administration using RT-LAMP, such as the Sherlock CRISPR SARS-CoV-2 kit reach 100% sensitivity and specificity, for a detection time of 60 min[119]. This, compared to our colorimetric device couple with ML analysis, which can achieve 100% sensitivity and 80% specificity in 10 min, demonstrate that our platform is very promising, both in terms of time-to-result, and sensitivity and specificity, although some optimization still needs to be performed.

SVM performs a bit less successfully on the classification task involving multiple mutated cell lines with a high number of classes, with an accuracy of 70.04%, and it is possible that both the high number of classes and the limited data available, although efforts were made to have the most extended dataset possible, reduced its performances. However, this result is fairly similar to the 74.9% accuracy obtained by Ho et al. on a 30-class classification[74]. In the future, additional data that is acquired could prove to increase the performance of the SVM on this specific task.

5.2 Machine Learning model

The work presented here classifies different datasets into discrete classes: cell line mutations, positive or negative to GBM and Covid-19. Automated analysis of data can take many forms and one of those is machine learning. However, multiple machine learning techniques exist. As a result, it is necessary to be familiar with the different existing machine learning techniques, and their different uses to select the one best suited for the task that needs to be performed. We discuss this choice here. Since classification is what is studied in this work, machine learning approaches that perform regression are not suitable, be they approaches developed for regression, or models designed initially for classifications which are then modified to perform regression. Similarly, unsupervised approaches are not suitable for the studies we are performing, as we already know the classes we want to classify the data into. Therefore, we need to choose a supervised approach that enables classification. There are still many candidates that can be chosen, and one must be selected. SVM, despite the advent of deep learning techniques that have become more popular in the past decade, still holds much potential in the field of medicine, for diagnostic purposes. It can be implemented using readily available languages and toolkits, such as sci-kit learn. Optimization and validation of the model can also be done through these modules, which allow for efficient integration in a data processing and analysis pipeline. Hyperparameter optimization can be performed using a Bayesian search. This search is less time consuming than a grid search, as it does not systematically evaluate all the combination of possible hyperparameters, while yielding good results[120]. It is therefore a good compromise between exhaustivity and prohibitive computational time and power. Care must also be taken when training ML models, as they can be prone to overfitting, and SVM is no exception. We have used the cross-validation technique to ensure that this does not occur, and that the models will perform properly. This is evidenced by the results we obtain, which show

that the SVM models that we have used generalize well, and can be applied on new data that was never seen before by the algorithms while maintaining good performance.

5.3 Biological samples

Real human samples are highly complex biological material, thus testing a sensor using clinical samples is a crucial step when determining its efficacy. Very different samples were collected and successfully analyzed: EVs collected from the plasma of patients, genetic material present inside clinical samples, namely naso-pharyngeal swabs and saliva samples. They were further analyzed using two very different optical readouts: SERS, and colorimetry respectively. This further demonstrates the versatility of machine learning in general, and SVM in particular, and its ability to find and use information present in any type of sample, even when they are analyzed through very different readout methods. It can be used on SERS spectra, which contains a lot of information on the chemical structures and components present, but also on colorimetric platforms, where the quantity of information is less, but needs to be detected consistently, with high sensitivity.

The biological samples used for these studies also hold promise for developing less invasive diagnostic techniques. Indeed, a liquid biopsy was sufficient to diagnose a patient as healthy or suffering from GBM, making it a less invasive procedure than a traditional biopsy of the brain tissue. This is significant especially in the context of cancer, as it suffers from late and invasive diagnosis procedures, which usually consist of solid biopsies[103]. These have additional limitations as they cannot always allow the clinicians to appreciate the full heterogeneity of the tumor, nor monitor its evolution. GBM also presents an additional challenge because a solid biopsy in the brain, a high-risk area is required for definitive diagnosis. In comparison, the analysis of EVs through liquid biopsies offer the possibility to have a comprehensive description of tumor heterogeneity, and to follow the evolution of the tumor through time, as

liquid biopsies can be repeated much more easily than solid biopsies[103]. Machine learning combined with liquid biopsy can therefore be a useful tool to diagnose, but also to personalize the care and treatment for each patient, as it has successfully demonstrated its ability to analyze the data obtained from such procedures.

Similarly, in the Covid-19 study, we have proven that saliva or naso-pharyngeal swabs could be used and successfully classified by the SVM. Saliva samples are less painful to acquire than naso-pharyngeal swabs and could be favored as a sample in the future, especially if repeated testing over multiple days is necessary.

5.4 Future directions

Some future directions involve the use of other ML algorithms. Indeed, it might be of interest in the future to compare SVM with other ML techniques, such as deep learning techniques, to determine whether or not SVM is the optimal solution, or if better results can be obtained, simply by changing the type of ML. This has started to be explored in Jalali et al. (manuscript under review), by using a residual CNN on the same spectral data used in this thesis. When dealing with the SERS data, it performs better than the SVM, both on the mutated cell line classification and the patient classification. It has also been successful for additional patient investigation. Indeed, it can be used to directly analyze patient samples to look for specific markers of mutations. The results obtained from the CNN are consistent with the patients' clinical annotations and could be used as a first diagnosis for tumor composition, or as a complement of traditional clinical annotation methods. This CNN could also be tested on the colorimetric dataset, to determine if it yields higher accuracy or a faster detection time than the SVM that has been presented here. It would also be of interest to determine whether deeper and more complex architectures, such as AlexNet or Inception[32,121], or models using similar architecture, could outperform both the SVM used in this work, and the CNN.

Conversely, it is also possible that due to constraints in computing power or speed, SVM might be preferred, especially in contexts where rapid results are of high importance, such as for the Covid-19 diagnostic test. In addition to this, a lighter architecture such as SVM can be easily implemented in a local device, without risking running out of memory space, thus eliminating the need for a good internet connection for remote cloud access, or the use of high-performance computers. Indeed, it has been shown that lighter models, that do not use deep learning can perform at least as well as deep learning models without needing as much computational power[97]. Therefore, a systematic comparison of the accuracy compared to the computational expenditure and the time necessary for training and testing the data could be a good way to find the best compromise between accuracy and efficient resource management. It would also allow to tailor the ML model to the application: a lighter, faster algorithm could be most beneficial when operating in environments with heavy constraints on resources, while slower but more accurate algorithms could be preferred in less constrained settings. One could envision that in applications such as the one described for Covid-19 detection, a light, fast and high-throughput approach could be preferred, while in an application such as mutation detection, a slower but more powerful model could be favored.

Furthermore, given the very promising results obtained on one cancer, GBM, and one infectious disease, Covid-19, these analysis pipelines could be generalized to other cancers and infectious diseases. Paediatric cancers, such as astrocytoma or paediatric GBM could also be a good candidate, given their similarly late diagnosis, heterogeneity and difficulty to operate. In addition to this, paediatric GBM has, unsurprisingly, some similarities to GBM[122], and would be a good place to start further studies, as we have shown we can successfully detect GBM and GBM-related mutations within cell lines. Other cancers that are challenging to treat, such as triple negative breast cancer, which suffers from poor prognosis, high risk of relapse after

surgery and resistance to molecular therapy, could also be diagnosed and investigated in patients using this platform coupled with ML, and EVs extracted from liquid biopsies[123].

Generalizing the diagnostic method presented here to other infectious diseases could also have very promising applications in global health monitoring and be used on already existing infections such as MERS, or different flu strains. This would enable to monitor multiple diseases with high rates of infection, and help isolate those tested positive, helping to curb the spread of the infection. Fully automating the entire process, from saliva collection to the final diagnosis given by the ML model is also a future step envisioned for this work, that would allow maximal ease of use for patients and healthcare professionals administering the test. Finally, having a multiplexed platform, adapted to detect multiple diseases or strains of the same disease, coupled with SVM for detection, could be very efficient for surveillance purposes and could be further investigated in the future.

6 Summary and Conclusion

SVM has demonstrated here its versatility, with high success rates in tasks employing two very different types of data: one dimensional spectra, with a high number of features associated with molecular change, and two dimensional images, with color change associated with a positive diagnosis. It has proven able to recognize the subtle changes in SERS spectra, enabling to classify cellular mutations, and also to diagnose patients suffering from GBM with 95% accuracy. In the context of a colorimetric test, SVM has successfully detected patients suffering from Covid-19 using an on-chip RT-LAMP assay after a 10-minute incubation period. SVM is effective both when performing in-depth analysis of data, and also when trying to be as fast as possible when dealing with time-constrained detection. As a ML technique that's does not demand high computing resources nor extensive training data, it has high potential in diagnostics, especially in point-of-care settings, where it can be implemented as a user-friendly, highly accurate, and quick analysis method. In addition to this, both studies were label-free and offered less invasive collection of samples than the traditional gold standard methods. Machine learning could in the future be implemented to analyze data generated by new applications of such devices, helping to construct an integrated analysis pipeline and bringing them from the lab to the real world.

7 Master bibliography

Introduction, literature review, discussion

1. Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J.* 21 (1959).
2. El Naqa, I. & Murphy, M. J. What Is Machine Learning? in *Machine Learning in Radiation Oncology* (eds. El Naqa, I., Li, R. & Murphy, M. J.) 3–11 (Springer International Publishing, 2015). doi:10.1007/978-3-319-18305-3_1.
3. Lussier, F., Thibault, V., Charron, B., Wallace, G. Q. & Masson, J.-F. Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering. *TrAC Trends Anal. Chem.* **124**, 115796 (2020).
4. Alloghani, M., Al-Jumeily D, A., Mustafina, J., Hussain, A. & Aljaaf, A. J. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. in *Supervised and Unsupervised Learning for Data Science* (eds. Berry, M. W., Mohamed, A. & Yap, B. W.) (Springer, Cham, 2020).
5. Devos, O., Ruckebusch, C., Durand, A., Duponchel, L. & Huvenne, J.-P. Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. *Chemom. Intell. Lab. Syst.* **96**, 27–33 (2009).
6. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
7. Paliwal, M. & Kumar, U. A. Neural networks and statistical techniques: A review of applications. *Expert Syst. Appl.* **36**, 2–17 (2009).
8. Mor-Yosef, S. Ranking the Risk Factors for Cesarean: Logistic Regression Analysis of a Nationwide Study. *Obstet. Gynecol.* 944–947 (1990).
9. Weng, S. *et al.* Deep learning networks for the recognition and quantitation of surface-enhanced Raman spectroscopy. *The Analyst* **145**, 4827–4835 (2020).

10. Carlomagno, C. *et al.* COVID-19 salivary Raman fingerprint: innovative approach for the detection of current and past SARS-CoV-2 infections. *Sci. Rep.* **11**, 4943 (2021).

11. Zitnik, M. *et al.* Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Inf. Fusion* **50**, 71–91 (2019).

12. Titano, J. J. *et al.* Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat. Med.* **24**, 1337–1341 (2018).

13. Rajkomar, A., Lingam, S., Taylor, A. G., Blum, M. & Mongan, J. High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. *J. Digit. Imaging* **30**, 95–101 (2017).

14. Lindsey, R. *et al.* Deep neural network improves fracture detection by clinicians. *Proc. Natl. Acad. Sci.* **115**, 11591–11596 (2018).

15. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).

16. Hofmann, T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Mach. Learn.* **42**, 177–196 (2001).

17. Gentleman, R. & Carey, V. J. Unsupervised Machine Learning. in *Bioconductor Case Studies* 137–157 (Springer New York, 2008). doi:10.1007/978-0-387-77240-0_10.

18. Rebrošová, K. *et al.* Rapid identification of staphylococci by Raman spectroscopy. *Sci. Rep.* **7**, 14846 (2017).

19. Gewers, F. L. *et al.* Principal Component Analysis: A Natural Approach to Data Exploration. *ArXiv180402502 Cs Stat* (2018).

20. Tania, M. H., Shabut, A. M., Lwin, K. T. & Hossain, M. A. Clustering and Classification of a Qualitative Colorimetric Test. in 7–11 (2018). doi:10.1109/iCCECOME.2018.8658480.

21. Cunningham, P. & Delany, S. J. k-Nearest Neighbour Classifiers - A Tutorial. *ACM Comput. Surv.* **54**, 1–25 (2022).

22. Sandhu, T. H. Machine Learning and Natural Language Processing - A Review. *Int. J. Adv. Res. Comput. Sci.* **9**, 582–584 (2018).

23. Kruber, F., Wurst, J., Morales, E. S., Chakraborty, S. & Botsch, M. Unsupervised and Supervised Learning with the Random Forest Algorithm for Traffic Scenario Clustering and Classification. in *2019 IEEE Intelligent Vehicles Symposium (IV)* 2463–2470 (IEEE, 2019). doi:10.1109/IVS.2019.8813994.

24. Kasera, S., Herrmann, L. O., Barrio, J. del, Baumberg, J. J. & Scherman, O. A. Quantitative multiplexing with nano-self-assemblies in SERS. *Sci. Rep.* **4**, 6785 (2015).

25. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58**, 109–130 (2001).

26. Stoltzfus, J. C. Logistic Regression: A Brief Primer. *Acad. Emerg. Med.* **18**, 1099–1104 (2011).

27. Clemmensen, L., Hastie, T., Witten, D. & Ersbøll, B. Sparse Discriminant Analysis. *Technometrics* **53**, 406–413 (2011).

28. Hastie, T., Tibshirani, R. & Friedman, J. H. Linear methods for classification. in *The elements of statistical learning: data mining, inference, and prediction* (2009).

29. Song, Y.-Y. & Lu, Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch. Psychiatry* **27**, 130–135 (2015).

30. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

31. Brownlee, J. *Deep Learning With Python*. (Machine Learning Mastery, 2016).

32. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).

33. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016). doi:10.1109/CVPR.2016.90.

34. Dong, R., Weng, S., Yang, L. & Liu, J. Detection and Direct Readout of Drugs in Human Urine Using Dynamic Surface-Enhanced Raman Spectroscopy and Support Vector Machines. *Anal. Chem.* **87**, 2937–2944 (2015).

35. Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: supervised methods. *Nat. Methods* **15**, 5–6 (2018).

36. Vapnik, V. N. *The nature of statistical learning theory*. (Springer, 2010).

37. Xu, Y., Zomer, S. & Brereton, R. G. Support Vector Machines: A Recent Method for Classification in Chemometrics. *Crit. Rev. Anal. Chem.* **36**, 177–188 (2006).

38. Radzol, A. R. M., Lee, K. Y. & Mansor, W. Classification of salivary based NS1 from Raman Spectroscopy with support vector machine. in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 1835–1838 (IEEE, 2014). doi:10.1109/EMBC.2014.6943966.

39. Noble, W. S. What is a support vector machine? *Nat. Biotechnol.* **24**, 1565–1567 (2006).

40. Suzuki, T. & Amano, Y. NLOS Multipath Classification of GNSS Signal Correlation Output Using Machine Learning. *Sensors* **21**, 2503 (2021).

41. Kelis Cardoso, V. G. & Poppi, R. J. Cleaner and faster method to detect adulteration in cassava starch using Raman spectroscopy and one-class support vector machine. *Food Control* **125**, 107917 (2021).

42. Pierna, J. A. F., Abbas, O., Dardenne, P. & Baeten, V. Discrimination of Corsican honey by FT-Raman spectroscopy and chemometrics. *Biotechnol Agron Soc Env.* 10 (2011).

43. Pinkham, D. W., Bonick, J. R. & Woodka, M. D. Feature optimization in chemometric algorithms for explosives detection. in (eds. Broach, J. T. & Holloway, J. H.) 83571K (2012). doi:10.1117/12.923387.

44. Barakat, N., Bradley, A. P. & H. Barakat, M. N. Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. *IEEE Trans. Inf. Technol. Biomed.* **14**, 1114–1120 (2010).

45. Hamza, A. An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique. *Int. J. Adv. Comput. Sci. Appl.* **8**, (2017).

46. Ashraf, S. F. *et al.* Predicting benign, preinvasive, and invasive lung nodules on computed tomography scans using machine learning. *J. Thorac. Cardiovasc. Surg.* S0022522321002580 (2021) doi:10.1016/j.jtcvs.2021.02.010.

47. Khan, S. *et al.* Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM). *Biomed. Opt. Express* **7**, 2249 (2016).

48. Çamlica, Z., Tizhoosh, H. R. & Khalvati, F. Medical Image Classification via SVM Using LBP Features from Saliency-Based Folded Data. in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)* 128–132 (2015). doi:10.1109/ICMLA.2015.131.

49. Schulz, M.-A. *et al.* Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nat. Commun.* **11**, 4238 (2020).

50. Guo, X. *et al.* Smartphone-based DNA diagnostics for malaria detection using deep learning for local decision support and blockchain technology for security. *Nat. Electron.* **4**, 615–624 (2021).

51. Shin, H. *et al.* Early-Stage Lung Cancer Diagnosis by Deep Learning-Based Spectroscopic Analysis of Circulating Exosomes. *ACS Nano* **14**, 5435–5444 (2020).

52. Erzina, M. *et al.* Precise cancer detection via the combination of functionalized SERS surfaces and convolutional neural network with independent inputs. *Sens. Actuators B Chem.* **308**, 127660 (2020).

53. Colthup, N. B., Daly, L. H. & Wiberley, S. E. Vibrational and Rotational Spectra. in *Introduction to infrared and Raman spectroscopy* (Academic Press, 1975).

54. Ellis, D. I. & Goodacre, R. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. *The Analyst* **131**, 875 (2006).

55. Moskovits, M. Surface-enhanced spectroscopy. *Rev. Mod. Phys.* **57**, 783–826 (1985).

56. Stremersch, S. *et al.* Identification of Individual Exosome-Like Vesicles by Surface Enhanced Raman Spectroscopy. *Small* **12**, 3292–3301 (2016).

57. Li, L., Hutter, T., Steiner, U. & Mahajan, S. Single molecule SERS and detection of biomolecules with a single gold nanoparticle on a mirror junction. *The Analyst* **138**, 4574 (2013).

58. Vitol, E. A. *et al.* In Situ Intracellular Spectroscopy with Surface Enhanced Raman Spectroscopy (SERS)-Enabled Nanopipettes. *ACS Nano* **3**, 3529–3536 (2009).

59. Ngo, H. T., Wang, H.-N., Fales, A. M. & Vo-Dinh, T. Plasmonic SERS biosensing nanochips for DNA detection. *Anal. Bioanal. Chem.* **408**, 1773–1781 (2016).

60. Lee, J.-H., Kim, B.-C., Byeung-Keun, O. & Choi, J.-W. Rapid and Sensitive Determination of HIV-1 Virus Based on Surface Enhanced Raman Spectroscopy. *J. Biomed. Nanotechnol.* **11**, 2223–2230 (2015).

61. Zheng, X.-S., Jahn, I. J., Weber, K., Cialla-May, D. & Popp, J. Label-free SERS in biological and biomedical applications: Recent progress, current challenges and opportunities. *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* **197**, 56–77 (2018).

62. Yan, Z. *et al.* A Label-Free Platform for Identification of Exosomes from Different Sources. *ACS Sens.* **4**, 488–497 (2019).

63. Zhang, Y. *et al.* Plasmonic Colorimetric Biosensor for Sensitive Exosome Detection via Enzyme-Induced Etching of Gold Nanobipyramid@MnO2 Nanosheet Nanostructures. *Anal. Chem.* **92**, 15244–15252 (2020).

64. Piriya V.S, A. *et al.* Colorimetric sensors for rapid detection of various analytes. *Mater. Sci. Eng. C* **78**, 1231–1245 (2017).

65. Askim, J. R. & Suslick, K. S. Colorimetric and Fluorometric Sensor Arrays for Molecular Recognition. in *Comprehensive Supramolecular Chemistry II* 37–88 (Elsevier, 2017). doi:10.1016/B978-0-12-409547-2.12616-2.

66. Morbioli, G. G., Mazzu-Nascimento, T., Stockton, A. M. & Carrilho, E. Technical aspects and challenges of colorimetric detection with microfluidic paper-based analytical devices (μPADs) - A review. *Anal. Chim. Acta* **970**, 1–22 (2017).

67. Zhu, D., Liu, B. & Wei, G. Two-Dimensional Material-Based Colorimetric Biosensors: A Review. *Biosensors* **11**, 259 (2021).

68. Priyadarshini, E. & Pradhan, N. Gold nanoparticles as efficient sensors in colorimetric detection of toxic metal ions: A review. *Sens. Actuators B Chem.* **238**, 888–902 (2017).

69. Namera, A., Nakamoto, A., Saito, T. & Nagao, M. Colorimetric detection and chromatographic analyses of designer drugs in biological materials: a comprehensive review. *Forensic Toxicol.* **29**, 1–24 (2011).

70. Chen, Z. *et al.* Detection of exosomes by ZnO nanowires coated three-dimensional scaffold chip device. *Biosens. Bioelectron.* **122**, 211–216 (2018).

71. Xu, W., Xue, X., Li, T., Zeng, H. & Liu, X. Ultrasensitive and Selective Colorimetric DNA Detection by Nicking Endonuclease Assisted Nanoparticle Amplification. *Angew. Chem. Int. Ed.* **48**, 6849–6852 (2009).

72. Dao Thi, V. L. *et al.* A colorimetric RT-LAMP assay and LAMP-sequencing for detecting SARS-CoV-2 RNA in clinical samples. *Sci. Transl. Med.* **12**, eabc7075 (2020).

73. Yin, G. *et al.* An efficient primary screening of COVID-19 by serum Raman spectroscopy. *J. Raman Spectrosc.* **52**, 949–958 (2021).

74. Ho, C.-S. *et al.* Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nat. Commun.* **10**, 4927 (2019).

75. Banaei, N. *et al.* Machine learning algorithms enhance the specificity of cancer biomarker detection using SERS-based immunoassays in microfluidic chips. *RSC Adv.* **9**, 1859–1868 (2019).

76. Chen, S. *et al.* Raman Spectroscopy Reveals Abnormal Changes in the Urine Composition of Prostate Cancer: An Application of an Intelligent Diagnostic Model with a Deep Learning Algorithm. *Adv. Intell. Syst.* **3**, 2000090 (2021).

77. Zhang, Y. *et al.* Identification and distinction of non-small-cell lung cancer cells by intracellular SERS nanoprobes. *RSC Adv.* **6**, 5401–5407 (2016).

78. Dies, H., Raveendran, J., Escobedo, C. & Docoslis, A. Rapid identification and quantification of illicit drugs on nanodendritic surface-enhanced Raman scattering substrates. *Sens. Actuators B Chem.* **257**, 382–388 (2018).

79. Yogesha, M. *et al.* A micro-Raman and chemometric study of urinary tract infection-causing bacterial pathogens in mixed cultures. *Anal. Bioanal. Chem.* **411**, 3165–3177 (2019).

80. Wang, C., Madiyar, F., Yu, C. & Li, J. Detection of extremely low concentration waterborne pathogen using a multiplexing self-referencing SERS microfluidic biosensor. *J. Biol. Eng.* **11**, 9 (2017).

81. Walter, A., März, A., Schumacher, W., Rösch, P. & Popp, J. Towards a fast, high specific and reliable discrimination of bacteria on strain level by means of SERS in a microfluidic device. *Lab. Chip* **11**, 1013 (2011).

82. Dawuti, W. *et al.* Urine surface-enhanced Raman spectroscopy combined with SVM algorithm for rapid diagnosis of liver cirrhosis and hepatocellular carcinoma. *Photodiagnosis Photodyn. Ther.* **38**, 102811 (2022).

83. Li, X. *et al.* Different classification algorithms and serum surface enhanced Raman spectroscopy for noninvasive discrimination of gastric diseases: Algorithms and serum SERS for discriminating gastric diseases. *J. Raman Spectrosc.* **47**, 917–925 (2016).

84. Radzol, A. R. M., Lee, K. Y., Mansor, W., Wong, P. S. & Looi, I. PCA-MLP SVM distinction of salivary Raman spectra of dengue fever infection. in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 2875–2878 (IEEE, 2017). doi:10.1109/EMBC.2017.8037457.

85. Sun, J. *et al.* Detection of glioma by surface-enhanced Raman scattering spectra with optimized mathematical methods. *J. Raman Spectrosc.* **50**, 1130–1140 (2019).

86. Li, B. *et al.* Non-invasive diagnosis of Crohn's disease based on SERS combined with PCA-SVM. *Anal. Methods* **13**, 5264–5273 (2021).

87. Silva-Neto, H. A., Sousa, L. R. & Coltro, W. K. T. Colorimetric paper-based analytical devices. in *Paper-based Analytical Devices for Chemical Analysis and Diagnostics* 59–79 (Elsevier, 2022). doi:10.1016/B978-0-12-820534-1.00009-8.

88. Solmaz, M. E. *et al.* Quantifying colorimetric tests using a smartphone app based on machine learning classifiers. *Sens. Actuators B Chem.* **255**, 1967–1973 (2018).

89. Rahmat, R. F. *et al.* Automated color classification of urine dipstick image in urine examination. *J. Phys. Conf. Ser.* **978**, 012008 (2018).

90. Edachana, R. P. Paper-based device for the colorimetric assay of bilirubin based on in-situ formation of gold nanoparticles. *Microchim Acta* 9 (2020).

91. Sajed, S., Arefi, F., Kolahdouz, M. & Sadeghi, M. A. Improving sensitivity of mercury detection using learning based smartphone colorimetry. *Sens. Actuators B Chem.* **298**, 126942 (2019).

92. Sajed, S., Kolahdouz, M., Sadeghi, M. A. & Razavi, S. F. High-Performance Estimation of Lead Ion Concentration Using Smartphone-Based Colorimetric Analysis and a Machine Learning Approach. *ACS Omega* **5**, 27675–27684 (2020).

93. Luo, R. *et al.* Machine learning for total organic carbon analysis of environmental water samples using high-throughput colorimetric sensors. *The Analyst* 7 (2020).

94. Tania, M. H. *et al.* Assay Type Detection Using Advanced Machine Learning Algorithms. in *2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)* 1–8 (IEEE, 2019). doi:10.1109/SKIMA47702.2019.8982449.

95. Wang, J. High-precision recognition of wheat mildew degree based on colorimetric sensor technique combined with multivariate analysis. *Microchem. J.* 8 (2021).

96. Kim, H., Awofeso, O., Choi, S., Jung, Y. & Bae, E. Colorimetric analysis of saliva–alcohol test strips by smartphone-based instruments using machine-learning algorithms. *Appl. Opt.* **56**, 84 (2017).

97. Tania, M. H. *et al.* Intelligent image-based colourimetric tests using machine learning framework for lateral flow assays. *Expert Syst. Appl.* 22 (2020).

98. Mutlu, A. Y. *et al.* Smartphone-based colorimetric detection via machine learning. *The Analyst* **142**, 2434–2441 (2017).

99. Esposito, A., Criscitiello, C., Locatelli, M., Milano, M. & Curigliano, G. Liquid biopsies for solid tumors: Understanding tumor heterogeneity and real time monitoring of early resistance to targeted therapies. *Pharmacol. Ther.* **157**, 120–124 (2016).

100. Alix-Panabières, C. The future of liquid biopsy. *Nature* **579**, S9–S9 (2020).

101. Ignatiadis, M., Sledge, G. W. & Jeffrey, S. S. Liquid biopsy enters the clinic — implementation issues and future challenges. *Nat. Rev. Clin. Oncol.* **18**, 297–312 (2021).

102. Vlassov, A. V., Magdaleno, S., Setterquist, R. & Conrad, R. Exosomes: Current knowledge of their composition, biological functions, and diagnostic and therapeutic potentials. *Biochim. Biophys. Acta BBA - Gen. Subj.* **1820**, 940–948 (2012).

103. Zhou, B. *et al.* Application of exosomes as liquid biopsy in clinical diagnosis. *Signal Transduct. Target. Ther.* **5**, 144 (2020).

104. Jalali, M. *et al.* Plasmonic nanobowtiefluidic device for sensitive detection of glioma extracellular vesicles by Raman spectrometry. *Lab. Chip* **21**, 855–866 (2021).

105. Kazemzadeh, M. *et al.* Space curvature-inspired nanoplasmonic sensor for breast cancer extracellular vesicle fingerprinting and machine learning classification. *Biomed. Opt. Express* **12**, 3965 (2021).

106. Lee, W., Lenferink, A. T. M., Otto, C. & Offerhaus, H. L. Classifying Raman spectra of extracellular vesicles based on convolutional neural networks for prostate cancer detection. *J. Raman Spectrosc.* **51**, 293–300 (2020).

107. Ćulum, N. M. *et al.* Characterization of ovarian cancer-derived extracellular vesicles by surface-enhanced Raman spectroscopy. *The Analyst* **146**, 7194–7206 (2021).

108. Carmicheal, J. *et al.* Label-free characterization of exosome via surface enhanced Raman spectroscopy for the early detection of pancreatic cancer. *Nanomedicine Nanotechnol. Biol. Med.* **16**, 88–96 (2019).

109. Banaei, N., Moshfegh, J. & Kim, B. Surface enhanced Raman spectroscopy-based immunoassay detection of tumor-derived extracellular vesicles to differentiate pancreatic cancers from chronic pancreatitis. *J. Raman Spectrosc.* **52**, 1810–1819 (2021).

110. Uthamacumaran, A. *et al.* Machine learning characterization of cancer patients-derived extracellular vesicles using vibrational spectroscopies: results from a pilot study. *Appl. Intell.* (2022) doi:10.1007/s10489-022-03203-1.

111. Boba, M. *et al.* False-negative results of breast core needle biopsies – retrospective analysis of 988 biopsies. 5.

112. Jackman, R. J. & Marzoni, F. A. Needle-localized breast biopsy: why do we fail? *Radiology* **204**, 677–684 (1997).

113. Disease Outbreak News. https://www.who.int/emergencies/disease-outbreak-news.

114. Sun, H. *et al.* AI-aided on-chip nucleic acid assay for smart diagnosis of infectious disease. *Fundam. Res.* **2**, 476–486 (2022).

115. Shabut, A. M. *et al.* An intelligent mobile-enabled expert system for tuberculosis disease diagnosis in real time. *Expert Syst. Appl.* **114**, 65–77 (2018).

116. Joung, H.-A. *et al.* Point-of-Care Serodiagnostic Test for Early-Stage Lyme Disease Using a Multiplexed Paper-Based Immunoassay and Machine Learning. *ACS Nano* **14**, 229–240 (2020).

117. Rapid Diagnostic Testing for Influenza: Information for Clinical Laboratory Directors | CDC. https://www.cdc.gov/flu/professionals/diagnosis/rapidlab.htm (2019).

118. Mercan, Ö. B., Kılıç, V. & Şen, M. Machine learning-based colorimetric determination of glucose in artificial saliva with different reagents using a smartphone coupled µ PAD. *Sens. Actuators B Chem.* **329**, 129037 (2021).

119. Khan, W. A., Barney, R. E. & Tsongalis, G. J. CRISPR-cas13 enzymology rapidly detects SARS-CoV-2 fragments in a clinical setting. *J. Clin. Virol.* **145**, 105019 (2021).

120. Frazier, P. I. A Tutorial on Bayesian Optimization. *ArXiv180702811 Cs Math Stat* (2018).

121. Szegedy, C. *et al.* Going deeper with convolutions. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1–9 (IEEE, 2015). doi:10.1109/CVPR.2015.7298594.

122. Jones, C., Perryman, L. & Hargrave, D. Paediatric and adult malignant glioma: close relatives or distant cousins? *Nat. Rev. Clin. Oncol.* **9**, 400–413 (2012).

123. Yin, L., Duan, J.-J., Bian, X.-W. & Yu, S. Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Res.* **22**, 61 (2020).