

UNIVERSITY OF BIRMINGHAM

Development of Novel Nano Platforms and Machine Learning Approaches for Raman Spectroscopy

By

PAULO ALEXANDRE DE CARVALHO GOMES

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Advanced Nano-Materials Structures and Applications Group
School of Chemical Engineering
College of Engineering and Physical Sciences
University of Birmingham
July 2022

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

In Raman spectroscopy, data analysis occupies a large amount of time and effort; thus, it is paramount to have the proper tools to extract the most meaning from the Raman analysis. This thesis explores improved ways to analyse Raman data mostly by using machine learning techniques available in Python. The substrate used throughout this thesis has been patterned through an electrohydrodynamic process that patterns micrometric pillars onto the substrate, which, after being gold coated, can generate surface-enhanced Raman scattering. An initial theoretical background was laid for the electrohydrodynamic process and additional observations regarding the fluid mechanics. Furthermore, when the structures are fabricated, and Raman measurements are taken, we show that it is possible to create an effective convolutional neural networks that systematically evaluate these patterns' surface morphology and extracts features responsible for the surface-enhanced Raman scattering phenomenon. Being able to predict 90% of the time from optical microscope images and 99% of the time with atomic force microscopy images

Additionally, a thorough machine learning analysis of the Raman literature was done. The best machine learning algorithms were put together into a script combined with a graphical user Interface that can run multiple commands such as principal component analysis and self-organizing maps, all in a centralised way. This way, we managed to consistently extract information from Raman and surface-enhanced Raman scattering spectra to open possibilities for precise peak analysis methods using a multi-Lorentzian fit algorithm.

Acknowledgements

The work presented in this thesis was carried out under the supervision of **Professor Pola Goldberg Oppenheimer**, head of the Advanced Nano-Materials Structures and Applications (ANMSA) group in the school of Chemical Engineering at the University of Birmingham.

I would like to express my gratitude to all of those who have been part of this journey with me.

First, I am very grateful to **Professor Pola Oppenheimer** for letting me be part of this group and project, which made me grow as an individual and a professional. Thank you for all the meetings we had, which were always productive and always gave me the energy to continue and most of all, your crucial feedback. It has always been much appreciated.

Second, I would also want to thank all new and old members of the ANMSA group that have been working around me since my arrival and until my farewell. Specially **Dr. JJ Rickard**, **Dr. Carl Banbury**, **Dr. Mike Hardy**, **Mike MacGregor**, **Dr. Paolo Passaretti**, **Dr. Emma McCarthy**, **Dr. Michael Clancy**, **Dr. Liam Kelleher**, **Dr. Martin Chu**, **Georgia Harris** and **Clarissa Stickland**. You have always been there to talk about science and random life questions. We had good times, and I will always cherish them when looking back.

I would also want to give a heart-warming thanks to the very special Portuguese community at the University of Birmingham and Birmingham. Specially, **Inês Barroso**, **Ana Évora**, **António Costa**, **Patrícia Gírio**, **Ana Gomes**, **João Santos**, **Bárbara Simões**, **Carlos da Silva**. For some of them, I shared the working

office, lab equipment and even my Instituto Superior Técnico (IST) background. You made me feel at home.

Also, thank all the international PhD students who have made part in the lunch companionship, pub nights, and coffees with me. I will remember you!

Finally, to all my Portuguese mates in Portugal, especially **António Cardoso**, **Tiago Cristovão**, **Pedro Cristovão** and **Eduardo Costa**, for the multiple online meetings in discord, chilled CS online gaming and adventures while I was in Portugal for the holidays.

An unnamed thank you to all people I interacted with throughout my PhD period. I will remember you if I see you.

Finally, I want to thank all my family for always supporting me in my decisions and always being there for me when I am most lost! I can always count on them.

And last, a *special* thanks to **Dr. Rosana Villares** for all the help and patience she provided me and for all the talks, trips and moments we shared throughout these years.

Muito Obrigado a todos!

Table of Contents

Table of Contents

Abstract.....i

Acknowledgements.....ii

Table of Contents.....iv

List of Figures.....xiii

List of Tables.....xxii

List of Equations.....xxiii

Abbreviations.....xxvi

List of Publications.....xxvii

Introduction.....1

1.1 Thesis Outline.....4

1.2 State of the art.....5

1.2.1 Raman Scattering.....5

1.2.2 Surface-enhanced Raman Scattering.....8

1.2.3 Electrohydrodynamic Lithography (EHD).....13

1.2.4 Convolutional Neural Network (CNN).....14

1.2.5 Raman Data Preprocessing.....17

1.2.6 Multivariate Data Analysis.....20

1.2.7 Raman Data Visualisation.....22

1.2.8 Raman Peak Analysis.....23

1.2.9 Graphical User Interfaces.....23

1.2.10 Summary.....24

1.3 References.....24

Experimental Tools and Analytical Techniques.....31

2.1 Electrohydrodynamic Patterning.....31

2.1.1 Experimental Setup.....31

2.1.2 Substrate Preparation.....32

2.1.3 Patterned Materials.....34

2.1.4 Spin Coating Thin Films.....35

2.2 Thin Layer Gold Sputter Coating.....36

2.3 Characterisation Techniques.....37

2.3.1 Optical Microscopy.....37

2.3.2 Atomic Force Microscopy.....40

2.3.3 (Surface-Enhanced) Raman Spectroscopy.....43

2.4 Python.....45

2.4.1 Principal Component Analysis (PCA).....45

2.4.2 Self-Organizing Maps (SOM).....46

2.4.3 Convolutional Neural Network (CNN).....47

2.5 Other Software....49

2.6 References....49

Electrohydrodynamic Lithography: Theoretical Background, Simulations and Height estimations....51

3.1 Abstract....52

3.2 Electrohydrodynamic Theoretical Background....53

3.2.1 Parallel Capacitor: Electrohydrodynamic Dispersion Relation....53

3.2.2 Tilted Capacitor: Electrohydrodynamic Dispersion Relation....63

3.2.3 EHD simulations....66

3.3 Alternative methods to determine Electrohydrodynamic pattern heights....69

3.3.1 Colour to Height conversion....69

3.3.2 Electrohydrodynamic Mass conservation....73

3.4 Conclusion....75

3.5 References....76

Electrofluidynamic Patterning of Tailorable Nanostructured Substrates for Surface-Enhanced Raman Scattering....77

4.1 Abstract....78

4.2 Introduction....79

4.3 Results and Discussion....83

4.4 Conclusions....97

4.5 Methods....98
4.5.1 Materials....98
4.5.2 Electrohydrodynamic Surface Enhanced Raman Scattering Substrates Fabrication....99
4.5.3 Fabrication of Top Electrode....100
4.5.4 Raman Spectroscopy Measurements....101
4.5.5 Calculation of the Surface Enhanced Raman Scattering Enhancing Factor 101
4.5.6 Atomic Force Microscopy....102
4.5.7 Optical microscopy....103
4.5.8 Scanning Electron Microscopy....103
4.5.9 Deep Neural Network Image Analysis....104
4.6 References....105
Real-time determination of Surface-Enhanced Raman Scattering substrates via a Convolutional Neural Network approach....108
5.1 Abstract....109
5.2 Introduction....110
5.3 Methodology....113
5.3.1 Materials....113
5.3.2 Electrohydrodynamic Patterning Surface Enhanced Raman Scattering Substrates Fabrication....114

5.3.3 Optical Image Acquisition.....115

5.3.4 Surface Enhanced Raman Scattering Acquisition.....115

5.3.5 Surface-Enhanced Raman Signal Factor Calculation.....116

5.3.6 Surface enhanced Raman Scattering / non-Surface Enhanced Raman Scattering labelling and organization.....116

5.3.7 Data Augmentation.....118

5.3.8 Convolutional Neural Network.....121

5.3.9 Output from the Convolutional Neural Network model.....123

5.3.10 Video input onto the Convolutional Neural Network.....124

5.3.11 Screen-grab Optical Microscope Image.....125

5.3.12 Pillar/non-Pillar labelling, Data Augmentation and Convolutional Neural Network building.....126

5.3.13 Applying the pillar/non-pillar Convolutional Neural Network.....126

5.3.14 Applying the Surface Enhanced Raman Scattering / non-Surface Enhanced Raman Scattering Convolutional Neural Network.....127

5.3.15 Real-time run event proof-of-concept.....129

5.3.16 Real-time Surface Enhanced Raman Scattering pillar tracking summary 129

5.3.17 Scripts.....130

5.4 Results and Discussion.....130

5.4.1 Further work/Improvements.....139

5.5 Conclusions....142

5.6 References....143

Saliva data and peak analysis review....146

6.1 Abstract....148

6.2 Methodology....148

6.3 Sample size influence over Machine Learning analysis....149

6.4 Different data processing methods....154

6.5 Saliva peak analysis....158

6.6 Conclusion....164

6.6 References....165

Open-source software for Raman data plotting, Multivariate analysis and spectral Peak fitting....168

7.1 Abstract....169

7.2 Introduction....170

7.3 Methodology....173

7.3.1 Saliva Preparation....173

7.3.2 Raman Data Acquisition....173

7.3.3 Raman Data Collection....174

7.3.4 Python Code....176

7.3.5 Graphical User Interface....176

7.3.6 Data Analysis Methods.....176
7.3.6.1 Asymmetric Least Square Smoothing (ALS).....177
7.3.6.2 Smoothing.....177
7.3.6.3 Normalisation – Relative Intensity.....178
7.3.6.4 Standard Deviation.....178
7.3.6.5 Z-Score.....178
7.3.6.6 Principal Component Analysis (PCA).....179
7.3.6.7 Loadings.....180
7.3.6.8 Gaussian Mixtures.....181
7.3.6.9 Euclidean Distance.....182
7.3.6.10 K-Nearest Neighbours.....182
7.3.6.11 ROC Curve.....183
7.3.6.12 Confusion Matrix.....183
7.3.6.13 Self-Organizing Maps (SOM).....183
7.3.6 Peak Finding Methods.....185
7.3.6.1 Locating Local Maximum.....185
7.3.6.2 Multi-Lorentzian Curve Fit.....185
7.3.6.3 Residual Lorentzian Calculation.....186
7.3.6.4 Standard ‘Key’ and ‘Lock’ Peak Finding Methods.....187
7.3.6.5 Principal Components Analysis Loadings Peak Finding Method….189

7.3.6.6 Self Organizing Maps Activation Peak Finding Method.....190

7.3.7 Scripts.....191

7.4 Results and Discussion.....192

7.4.1 Plots.....192

7.4.2 Principal Component Analysis.....194

7.4.3 Peak Finder.....196

7.4.4 Principal Component Analysis Peak Finder.....199

7.4.5 Self Organizing Maps Peak Finder.....202

7.5 Conclusions.....205

7.6 References.....207

Summary and Future Outlook.....210

8.1 Summary.....210

8.2 Future Outlook.....211

8.3 References.....214

APPENDIX A.....216

SUPPORTING INFORMATION FOR CHAPTER 4.....216

S1 Representative Average Surface Roughness in the Centre of EHD Fabricated Pillar Topology.....216

S2 Calculation of the Enhancement Factor and Reproducibility Coefficient…217

S3 Reproducibility of EHD-SERS Substrates - Uniform Platforms for Reliable SERS.....218

S4 Stability Analysis.....219

S5 Spike-and-Recovery (S&R) Proof-of-Concept Assay of the EHD SERS Substrate TBI Biomarker.....220

S6 Characteristic Fingerprint SERS Spectra and Barcode of Peaks of the N-acetylaspartate Marker.....221

S7 Roughness of the Au-PS film.....222

S8 Quantitative Detection of Representative Biomarker via EHD-SERS.....223

S9 Detectable Biochemical Probes.....224

APPENDIX B.....225

PYTHON CODES.....225

B.1 Raman peaks from papers.....226

B.2 Saliva Peaks Whisker Plot.....229

B.3 EHD Colour to Thickness.....231

B.4 EHD Electrostatic Pressure Calculation for Parallel Configuration.....234

B.5 EHD Electrostatic Pressure Calculation for Tilted Configuration.....236

B.6 EHD 3D Simulations.....238

List of Figures

Figure 1.1: *Light scattering on a phonon. Three different scattering outcomes result from this interaction, namely: Rayleigh scattering, Δω (Raman shift) = 0, Stokes scattering Δω<0 and anti-Stokes scattering Δω>0 with their respective energy band diagram* *Page 8*

***Figure 1.2**:* *(A) Charge transfer mechanism between the adsorbate and the surface. This charge will slightly increase the enhancement of the SERS. (B) SERS due to nanometric roughness sites on the surface, makes highly localised enhancement of the electromagnetic field from the laser light, which will scatter with increasing intensity. (C) Interaction of electromagnetic radiation with the surface plasmons over the Nobel metal surfaces. (D) Electromagnetic field enhancement profile of two nanoparticles – where the enhancement is majorly centred around the space between the nanoparticles. And the enhancement profile between two roughness sites – roughness sites are simulated as hills and holes profile – where the enhancement is majorly centred in the hole part of the surface.* *Page 12*

***Figure 1.3**:* *(A) (i) Electrohydrodynamic lithography setup. Two parallel silicon substrates form a capacitor, with the bottom one coated with a micrometric thin polystyrene layer, after setting up a potential difference with a temperature above the glass transition temperature of the polymer (ii). The generated heat with the potential difference creates an electrostatic pressure that provokes surface instabilities onto the polystyrene, which begins to raise towards the top electrode. After reaching the top electrode and cooldown, the process is terminated. After disassembling, a thin gold coating layer is sputter-coated, making the patterned structures SERS active (iii). (B) EHD images from different image acquisition tools (i) scanning electron microscope (SEM), (ii) Renishaw optical microscope, (iii) AFM 3D, (iv) AFM 2D.* *Page 13*

***Figure 2.1**:* *EHD setup formation. (A) Bottom and top electrode assembled in a capacitor geometry with a polymer layer and an air gap. Applying a potential difference exerts electrostatic pressure that forces the polymer to rise at a high temperature. (B) Generation of pillars due to the electrostatic pressure and temperature.* *Page 32*

Figure 2.2: *(A) Schematic representation of the snow jet and respective (B) heating plate. (C) Schematic representation of the plasma cleaning process and respective (D) plasma cleaner tool. (E) Resulting clean silicon piece (top) ready to be sputtered coated (bottom).* *Page 34*

Figure 2.3: *(A) i) A solution of a solvent is pipetted onto the substrate. ii) The spin coater is then spun up, and the solution is forced to spread over the substrate. iii) The solvent evaporates as time passes leaving a thin homogenous layer. (B) Spin* *Page 35*

coater used for this work.

Figure 2.4: *(A) Automatic sputter coater schematic representation of the sputter coater. The figure shows a magnetron sputtering deposition. (B) The instrument used to sputter coat our EHD surfaces* *Page 36*

Figure 2.5: *A) Leica DM 2700M, (B) EHD patterns from the Leica DM 2700M microscope (scale bar: 20 μm). (C) Leica DM 2500 present in the InVia Qontor system, (D) EHD patterns before gold coating from the Leica DM 2500 (scale bar: 20 μm). Both EHD patterns (B) and (D) are from the same spot, the highlighted dash circle to help identify the common patterns and to inspect the difference in quality.* *Page 39*

Figure 2.6. *(A) NanoWizard II by JPK microscope. (B) i) Schematic representation of the tapping mode function while going though a gold structure while changing the laser reflection. ii) Laser alignment seen in the AFM software, maximise the laser sum and minimise the lateral and vertical deflection for optimal measurement. (C) Optical microscope image of the sample and cantilever, with a proper laser alignment (white spot). Measurement examples: (D) 1D line of a forming EHD structure, (E) 2D representation of a rough gold surface and (F) 3D of an E. coli bacteria (image obtained for [12]).* *Page 42*

Figure 2.7: *(A) Back part of the InVia Qontor Raman Spectrometer where the lasers are located. From these the 785 nm and 633 nm were the ones used. (B) Schematic representation of the Raman spectrometer showing the main components of a Raman spectrometer. The beam splitter splits the beam the objective aiming at the sample. Subsequently, the scattered radiation goes through a set of mirrors, lenses, filters and gratings that cleans the scattered radiation into the Raman scattering which is resolved in the CCD and inspected through the WiRE software.* *Page 44*

Figure 2.8: *(A) VGG16 architecture adapted for the new output layer SERS/non-SERS (B) VGG16 architecture for the fine-tuning of the bottom layers. The complete training of our CNN goes first through (A) and then (B).* *Page 49*

Figure 3.1: *Schematic diagram of a thin polystyrene film with a thickness h on a silicon substrate with an air gap and a top silicon electrode. Outlining the undulations, with a wavelength λ, caused by the polymer Brownian motion.* *Page 53*

Figure 3.2: *An applied electric field leads to the polarization of the dielectric polymer layer, yielding an attractive interaction between the charges at the polymer-air interface with the charges at the upper electrode. The layered system is equivalent to a sum of two capacitors in series.* *Page 58*

Figure 3.3: *(A) Contour plot of the 1/τ, in relation to q and h. To note regions of 1/τ>0 and 1/τ<0 and undetermined regime in blank. (B)i) 1/τ as a function of q, the stable* *Page 62*

and unstable regime present. ii) ∂pel/∂h variation with different polymer heights, showing breaking point at around 1.65×10-6m. iii) qc and qmax variation with polymer height, reaching a point where it breaks ∂pel/∂h<0 when. iv) pel variation with film thickness and representation of the breaking point.

Figure 3.4: *Tilting provoked by the different size spacers. With d being the incremental height resultant from the tilting angle.* *Page 63*

Figure 3.5: *Optical microscope image of an EHD pattern using different height spacers. Polystyrene colour to height conversion from [2].* *Page 64*

Figure 3.6: *1/τ as a function of q for different tilting δ values, the stable and unstable regime present. The parallel capacitance geometry is verified for δ=0 μm, by the similar dispersion regime identified in Figure 3.3Bi. When increasing the tilting values the amplified regime becomes extremely amplified, red line curvature from 0<δ<0.7μm (i - viii). For highest tilting δ=0.8μm (ix) the growth regime is completely damped.* *Page 66*

Figure 3.7: *(A) Cell of the initial EHD condition with small surface fluctuations given by a sine wave as a function of x and y distance. (B) Last iteration of the EHD process. Multiple nucleation site separated by a λ with one spot reaching the top of the electrode, stopping the iteration. (C) Check the spot of high growth in a 2D representation (slice of the 3D Figure 3.7B) at the start of the run and at the end. (D) 1D representation (point of the 3D Figure 3.7B) of the highest growth spot per iteration and observing the exponential growth trend.* *Page 69*

Figure 3.8: *(A) RGB values variation from the colour spectra of polystyrene, which are used to convert into a single value between 0 and 1, represented by the polystyrene calibration colours. (B) Polystyrene RGB values ranged between 0 and 1 (represented by the polystyrene calibration colours), and respective curve fit. (C) Top view of some EHD pillars examples and their colour to thickness conversion (i-iv), and respective matching in the calibration curve in (B).* *Page 71*

Figure 3.9: *(i-iv) 3D pillars image after applying the colour to thickness converter on the pillar optical images from Figure 3.8Ci-iv.* *Page 72*

Figure 3.10: *(A) i) Polystyrene colour variation with thickness, with added height points for 1-5 related to the different regions in ii) showing that these fit nicely onto the best fit line, and a cross from initial film thickness in iv). ii) Optical image of an EHD patterned surface, for proof-of-concept test of [Equation 41] over 5 different EHD region. iii) Table classification for the different pillar colours and respective parameters, number of pillars, n, radius, r, and the calculated thickness, t. iv) Initial polystyrene film optical image before EHD formation.* *Page 75*

Figure 4.1. *EHD process for SERS fabrication. To commence the pattering process, (a) a* *Page 85*

thin polymer film with an initial film thickness h is deposited between two electrodes spaced apart by a distance d. Applying an external voltage, V and liquefying at T > Tg leads to the generation of a high electric field, Ef (~108V/m) within the microcapacitor and (b) formation of an instability throughout the film with a characteristic wavelength λ, which determines the pitch between the final structures. The increasing amplitude of the film variations guides the film's redistribution, which eventually spans the gap between the two electrodes. (c) Top-view optical microscopy and (inset) scanning electron microscopy images of the EHD generated polymer structures, which are subsequently coated with a plasmon active gold nanolayer leading to high SERS signal enhancement. (inset) Representative SERS spectra of BT on top of various topologies of the formed pillars exemplifying that the roughened centre surrounded by a nanorim (top view, red) yields more than 50 times higher signal enhancement relative to the flat pillar topology.

Figure 4.2. *Representative EHD-generated topological overview. Cross-section AFM profiles, top-view OM images, SEM micrographs, and CNN images of the EHD generated structures. Corresponding SERS enhancements are shown in Figure 4.3d. (a) Pillar with Hrim = 250 nm and a nanoroughened centre, EF = 9.0 × 10 6. (b) Pillar with an equivalent nanorim of Hrim = 300 nm and a roughened centre, EF = 1.1 × 107. (c) Pillar with a one-sided nanorim of Hrim = 270 nm and a roughened centre, EF = 8.8 × 10 6. (d) Pillar with nanoroughness throughout and no rim, EF = 6.4 × 10 6. (e) Smooth pillar with one-sided nanorim of Hrim = 200 nm, EF = 6.1 × 10 6. (f) Pillar with a nanorim of Hrim = 150 nm, EF = 6.7 × 106. (g) Smooth pillar top with no rim, EF = 5.8 × 106. Scale bar SEM: 1 μm.* *Page 87*

Figure 4.3. *Tunability of structural dimensions of the metallo-dielectric EHD fabricated SERS substrates for controllable and optimal plasmon excitations. (a) Relative SERS enhancement versus the AR identifies an optimal AR for signal enhancement in the range between 0.6 and 0.8. Wedge-geometry induced AR variation generates bands of colour, due to the film thickness redistribution on top of the underlying silicon wafer, corresponding to the various aspect ratios (inset). The arrows correspond to the variation in AR and the corresponding colour change across the sample. (b) Relative SERS enhancement of the benzenthiol's 1070 cm–1 peak with varying nanorim heights indicate an optimal range of Hrim~300 ± 37 nm. (c) Spectrally resolved reflectance spectra relative to the flat gold, R0 as a function of the ARs. An increasing AR results in an increase in the extinction (R/R0) due to the enhanced coupling of light into LPSRs, which are tuned by the pillar geometry. (inset) The variation of the enhancement as a function of the gap between the pillars at the excitation wavelength of 785 nm reveals another important parameter, tunability of which and its effect on strength of localized surface plasmon resonances and surface Raman enhancement allows the optimization of the* *Page 89*

fabricated substrates. (d) Characteristic Raman peaks of benzenethiol at 1000, 1070, and 1575 cm^{-1} readily detected in the spectra acquired from the EHD SERS substrates showing the signal enhancement variation with the pillar's topological features shown in Figure 4.2.

Figure 4.4. *Engineering the EHD apparatus and reproducibly controllable structures. (a–c) Optical microscopy images and the overlaid Raman maps with the corresponding 2D near-field simulations of the Ef distribution (insets) of the SERS enhancement occurring (a) exclusively on top of each pillar, (b) at the perimeter of the pillar, and (c) between the pillars. Schematics (d) and AFM 3D images of the fabricated (e) top electrode with carefully designed parameters including the interelectrode spacing, AR, pitch, and the interstructural spacing of the protrusions, Λ m. (f) 3D schematic representation of the controllable EHD rig. (g) Photograph of the EHD prototype apparatus in the heating chamber.* *Page 93*

Figure 4.5 *(a) OM image of pillars formed under the reproducible EHD microlithography using planar top electrode, generating a homogeneous electric field throughout the capacitor, and (inset) pillars forming using lithographically predefined top electrode, introducing a heterogeneous electric field. AFM phase (b), with the corresponding 3D profile, and height (c) images of the high-fidelity EHD-replicated patterns. (d) Representative SERS spectrum of BT on a uniform substrate for the highlighted 1070 cm –1 peak, at an excitation of 785 nm laser, demonstrating reliable signal and substrate reproducibility (inset) of the SERS intensities on the EHD generated arrays from 10 randomly selected positions on the gold-coated EHD fabricated pillars. (e) Spike-and-recovery bioassay differentiates the characteristic peaks of biomarker in an artificial plasma with known concentrations (50pg/mL). SERS spectra of the spiked (gray) versus normal, nonspiked (black); analyte was spiked into the test sample matrix, and its response was recovered in the assay by comparison to an identical spike in the standard diluent with 95.5 ± 3.4% recovery of the biomarker.* *Page 96*

Figure 5.1: *(A)(i): Optical image of the EHD structures from the Leica optical microscope. (ii) The augmentation process generates rotated/mirrored/inverted images of the original Leica microscope image. (B)(i) Same EHD structures as in (A) (i) viewed through the Renishaw inVia optical microscope. The individual pillars are cropped using GIMP and are processed using Python for further image augmentation. (ii) The augmentation process generates rotated/mirrored/inverted images of the original Raman optical microscope image. (C) Raman spectra of a selection of 885 pillars show the separation between SERS and non-SERS spectra, with the red spectrum as a separation spectrum. Top red arrow pointing at the MB 1610 cm^{-1} most intense peak. (D) Histogram plot from the EF for each pillar shows the separation value between both classes in the red bar. (E) Whisker plot with dashed red marking the 3rd* *Page 118*

IQR separating SERS from non-SERS pillars.

Figure 5.2: *Example of a screen-grab from the Renishaw computer when acquiring a Raman measurement. (i) Widgets displayed from the WiRE interface. (ii) Optical microscope image from the sample. (iii) Raman spectrum from the acquired spot delimited by the cross-hair intersection in (ii). (iv) Measured sample list.* *Page 125*

Figure 5.3: *Applying an algorithm to break the WiRE screen-grab (A) into tiles, and then passing the tiles into the pillar/non-pillar CNN. (B) Output of the tiles being processed with the CNN, heat mapped by grad-CAM algorithm. (C) Padded final region of interest – optical microscope image- (D) final cropped optical image.* *Page 127*

Figure 5.4: *Application of the SERS/non-SERS CNN. (A) CrOMI from the WiRE software screen-grab. (B) Break the CrOMI into multiple tiles and pass them onto the SERS/non-SERS CNN outputting the grad-CAM results heatmap. (C) Reconstruct the CrOMI overlaid with the heatmap tiles. (D) Overlaid WiRE screen-grab with the highlighted heatmap SERS pillars on the optical image* *Page 128*

Figure 5.5: *SERS tracking method overview. (A) Continuous image acquisition using an inVia Renishaw Raman Spectrometer outputs SERS spectra corresponding to individual EHD pillars in the image. (B) The real-time screen-grab is analysed by the pillar/non-pillar CNN, revealing the heatmap activation for the optical microscope region (C). (D) Later padding it and cropping it (E). (F) The cropped image is passed onto the SERS/non-SERS CNN model and, the pillars labelled as SERS, are highlighted using the grad-CAM algorithm (G). (H) The output overlays the WiRE interface with the SERS/non-SERS activation regions over at the optical microscope screen section.* *Page 130*

Figure 5.6: *(A) Validation and Training accuracy change compared with the number of initial images. Marked as points of interest 1-5 for different data augmentation iterations related to Table 5.1. Orange circle related to (B). (B) Validation and training accuracy values changing with the number of augmented images from 5-7 related to Table 5.1. Orange circle related to (D). (C) Relation between the relative difference (%) between the training accuracy and validation accuracy and the ratio of the number of augmented images to the number of initial images. (D) Validation and training accuracy changes with different CNN model architecture, * notation related to models breaking down after some epochs. Changes between subsequent CNN architectures highlighted by a thicker square boarder.* *Page 136*

Figure 5.7: *(A) (i) CrOMI with outline SERS activation pillars. The red dashed square defines a test-run region. (ii) Zoom in on the test-run region with highlighted SERS activations. Table with the accuracy score test, the relative intensity of the pillar (I/I0), the real pillar label and the activation classification label. (B)* *Page 137*

SERS measurement tests. (i) SERS measurement with the appropriate activation over the pillar A pointed with an arrow. (ii) Non-SERS measurement without any activation over the pillar B pointed with an arrow.

Figure 6.1: *Stacked bar plot of the variation in sample size between different research papers. In orange, the number of unhealthy samples and the number of healthy samples in blue.* *Page 152*

Figure 6.2: *Specificity, sensitivity, and accuracy variation with a different number of patients. Increase the statistical values with an increasing number of patients (healthy + unhealthy). Sensitivity describes the portion of true positives correctly identified by a diagnostic test. Specificity is the proportion of the true negatives correctly identified in a diagnostic test. Accuracy is the proportion of accurate results, in a population.* *Page 153*

Figure 6.3: *Most relevant procedures for Raman data processing and analysis. In red, showing the most common data processing path among researchers. A: Data collection, B: baseline followed by normalisation, C: normalisation followed by baseline, D: data preprocessing, E: machine learning multivariable analysis, F: classification of the machine learning model, G: results and H: loadings.* *Page 155*

Figure 6.4: *Raman peak assignment on cancerous saliva samples. Their biomolecular group displays a doughnut pie chart of the most prominent peaks in the literature.* *Page 159*

Figure 6.5: *List of biomolecules present in saliva from cancerous patients with respective whisker plots for different Raman peak regions. Red is related to protein molecules, green nucleic acid, magenta sugars and blue lipids. When the peak assignment is singular, a stripe represents the symbol—complemented with Table 6.1* *Page 160*

Figure 7.1: *Raman data structure from the Renishaw system. A) Automated data formatting from WiRE 5.1 software. B) Processed data for quicker access and less memory storage.* *Page 175*

Figure 7.2: *First part of the Graphical Interface – Plotting. (A) The GUI shows multiple commands options that can be toggled and pushed to make different kinds of plots. Additionally, multiple display windows outline the files imported, a text area to display results and additional function tabs. When the options are toggled, the process is performed over the Raman data before plotting. For the data processes, we have the standard plot (Bi), toggled options for baseline correction (Bii), normalization (Biii) and standard deviation (Biv). Besides the standard plot, there are three other different plot options, stack plot (Ci), heatmap plot (Cii) and waterfall plot (Ciii).* *Page 193*

Figure 7.3: *Second tab of the Graphical Interface – PCA. A) PCA section, multiple* *Page 195*

commands can be computed in this panel by toggling different options and pressing the PCA 2D or 3D to run. B) Showcase a 3D PCA plot for 15 different datasets. C) 3D PCA with the same data points with Gaussian Mixtures applied. D) 3D PCA with the same data points with a Z-score applied to the data.

Figure 7.4: *A) Second part of the PCA section, exploring the window range, PCA loadings and distance plot in 2D. B)i) Euclidean distance whisker bar plot in the binary mode (healthy/unhealthy) obtaining for the distance between cluster ζ and cluster Ψ from PCA plot in Biii). ii) Confusion matrix for the classification test of ζ and Ψ samples, with accuracy (A), Specificity (Sp), Selectivity (Se), negative predictive value (NPV) and precision (P). iii) PCA plot with k-NN separation and separation percentage (52%).iv) ROC curve for the false positive rate (FPR) and true positive rate (TPR) relation. C)i) 2D PCA plot (same data as Figure 3), showing the respective loading components C)ii). D)i) 2D PCA and ii) loadings with windows setting on with a range from 950 cm -1 to 1500 cm -1 , highlighting a slightly different separation and loadings than D)i).* *Page 196*

Figure 7.5: *A)i) Peak Finder GUI tab highlighting the standard Raman peak analysis. Load a Raman TXT file (A ii) into 'key' and generate the Raman fit (B) with the peak information stored in a CSV file (C). Load the generated Raman (C) list into 'key' and load a Raman text file into 'lock' with the option 'Peak matching' ticked to generate the peak matching between the 'key' peaks and the 'lock' file, output results in (Di, ii).* *Page 198*

Figure 7.6: *A)i) GUI – PCA Peak Finder section. Highlighted in yellow and cyan are the two different operation methods of the script. Yellow generated the multi-Lorentzian fit CSV file of the loadings, and cyan compared the peaks from the loadings CSV file with the peak from the standard peak CSV file. B) PCA of three different samples, highlighted by four different regions in the PC1 and PC2 space (Q: quadrants, Q1 blue, Q2 orange, Q3 green and Q4 red). C) CSV output file of executing the highlighted yellow routine in A. Showing the multi-Lorentzian fitting parameters. D) Loading plot from PC1 and PC2 separated by two regions positive loadings (dark grey) and negative loadings (light grey). E) Multi- Lorentzian fit of the loadings for each quadrant. F) Peak matching between the multi-Lorentzian fit summed and the imported 'key' CSV file.* *Page 201*

Figure 7.7: *A)i) GUI - SOM Peak Finder section. Highlighted in yellow and cyan are the two different operation methods of the script. Yellow generates the multi-Lorentzian fit CSV file of the SOM neuron activations, and cyan compares the peaks from the activation CSV file with the peaks from the standard peak CSV file. B) Normalized average spectrum of the different samples loaded into 'key'. C) Self-Organizing maps in a three by three hexagonal configuration, separating the different samples into different hexagons highlighted with separation boundaries and edge-colouring related to the highest hexagon* *Page 204*

sample population. D) Activation Raman' spectrum' related to the different hexagons positions. E) Peak fitting parameter table of the different Lorentzian fits to the SOM activation 'spectrum' (D). F) Heatmap representing each SOM hexagon of the important peak fit centre values from the activation 'spectrum'. (D). G) Multi-Lorentzian sum for the different SOM activations per hexagon, with peak matching result between the SOM activation CSV and the imported 'key' CSV file.

List of Tables

Table 5.1:	*Data Augmentation trials.*	*Page 119*
Table 5.2:	*CNN trials.*	*Page 122*
Table 6.1:	*Raman biomolecule information from cancerous patient saliva with respective Raman region.*	*Page 161*
Table 6.2:	*Saliva biomolecular content change for the unhealthy patient. The change delta is in comparison with the healthy spectra. Arrow up: Condition spectra increases in that wavenumber. Arrow down: Condition spectra decrease in that wavenumber.*	*Page 163*

List of Equations

Equation 3.1	*Polymer height sinusoidal variation*	*Page 54*
Equation 3.2	*Navier-Stokes equation for an incompressible Newtonian fluid*	*Page 54*
Equation 3.3	*Navier-Stokes equation for an incompressible Newtonian fluid expanded in x*	*Page 54*
Equation 3.4.1	*Navier-Stokes equation in the z coordinate*	*Page 55*
Equation 3.4.2	*Navier-Stokes equation in x coordinate*	*Page 55*
Equation 3.5	*Fluid velocity as a function of z*	*Page 55*
Equation 3.6	*Solved Fluid velocity equation using the boundary conditions*	*Page 56*
Equation 3.7	*Mass conservation equation*	*Page 56*
Equation 3.8	*Average velocity equation*	*Page 56*
Equation 3.9	*Equation of motion (polymer height variation with time)*	*Page 56*
Equation 3.10	*Pressure acting on the polymer*	*Page 56*
Equation 3.11	*Simplified pressure acting on the polymer*	*Page 57*
Equation 3.12	*Thermodynamic Free Energy equation*	*Page 57*
Equation 3.13	*Electrostatic pressure*	*Page 57*
Equation 3.14	*Equivalent parallel capacitance equation*	*Page 57*
Equation 3.15.1	*Capacitance equation*	*Page 58*
Equation 3.15.2	*Capacitance variation with the polymer height*	*Page 58*
Equation 3.16	*Electrostatic pressure*	*Page 58*
Equation 3.17	*Laplacian pressure equation*	*Page 58*
Equation 3.18	*Complete pressure equation*	*Page 59*
Equation 3.19	*Equation of motion expanded*	*Page 59*
Equation 3.20	*Equation of motion arranged*	*Page 59*
Equation 3.21	*Dispersion relation*	*Page 59*
Equation 3.22	*Solving the dispersion relation*	*Page 59*

Equation 3.23	*Dispersion relation for the fastest growing mode*	*Page 60*
Equation 3.24	*Electrostatic pressure derivative to the polymer height*	*Page 60*
Equation 3.25	*Dispersion relation for maximum wavelength*	*Page 60*
Equation 3.26	*Maximum dispersion relation*	*Page 60*
Equation 3.27	*General dispersion relation*	*Page 61*
Equation 3.28	*Equivalent tilted capacitance equation*	*Page 63*
Equation 3.29	*Tilted capacitance equation (air)*	*Page 64*
Equation 3.30	*Tilted capacitance equation (air)*	*Page 64*
Equation 3.31	*Tilted capacitance equation (air)*	*Page 64*
Equation 3.32	*Equivalent tilted capacitance equation*	*Page 65*
Equation 3.33	*Equivalent tilted capacitance equation comparison with parallel capacitance*	*Page 65*
Equation 3.34	*Tilted capacitance equation*	*Page 65*
Equation 3.35	*2D Equation of motion*	*Page 66*
Equation 3.36	*2D Equation of motion expanded*	*Page 67*
Equation 3.37	*Complete pressure equation*	*Page 67*
Equation 3.38	*2D Equation of motion expanded*	*Page 67*
Equation 3.39	*2D Equation of motion height variation*	*Page 68*
Equation 3.40	*Electrohydrodynamic mass conservation*	*Page 73*
Equation 3.41	*Pillar height calculation*	*Page 73*
Equation 4.1	*Electrostatic Pressure*	*Page 84*
Equation 5.1	*Enhancing Factor calculation*	*Page 116*
Equation 5.2	*Relative accuracy difference (RAD) regarding the ratio of the number of augmented images to the number of original images*	*Page 133*
Equation 7.1	*Relative Raman Intensity calculations*	*Page 178*
Equation 7.2	*Raman Intensity standard deviation calculations*	*Page 178*

Equation 7.3	*Z-score calculation*	*Page 179*
Equation 7.4	*Principal Component Loadings calculation*	*Page 180*
Equation 7.5	*Loadings Lorentzian fitting per Quadrant*	*Page 181*
Equation 7.6	*Euclidean 2D/3D distance calculation*	*Page 182*
Equation 7.7	*Mathematical definition of a spectral peak*	*Page 185*
Equation 7.8	*Lorentzian Fitting Equations*	*Page 186*
Equation 7.9	*Residual from Lorentzian fit minus Raman signal*	*Page 186*
Equation 7.10	*2^{nd} order Residual calculations*	*Page 187*
Equation 7.11	*Complete Lorentzian fit equation for a compound ζ*	*Page 188*
Equation 7.12	*Complete Lorentzian fit equation for the whole Raman signal*	*Page 188*

Abbreviations

ANMSA Advanced NanoMaterials Structures and Applications
ANN Artificial Neural Network
AR Aspect ratio
AUC Area Under the Curve
BT Benzenethiol
CCD Charge Coupled Device
CNN Convolutional Neural Network
CrOMI Cropped Optical Microscope Image
CSV Comma-separated values
CV Cross validation
DC Direct current
DW Deionized Water
EF Enhancement Factor
EHD Electrohydrodynamic
FIB Focused Ion Beam
FN False negative
FP False positive
ft Fine-tune
GI Graphical interface
Grad-CAM Gradient-weighted Class Activation Mapping
GUI Graphical user interfaced
IDE Integrated development environment
IQR Interquartile range
k-NN k-nearest neighbours
LDA Linear Discriminate Analysis
LSP Localised surface plasmons
LSPR Localised surface plasmons resonance
MB Methylene Blue
ML Machine learning
NA Numerical Aperture
NP Nanoparticles
NPV Negative Predictive Value
OCR Optical Character Recognition
PEEK Polyether ether ketone
PC Principal components
PCA Principal component analysis
PDE Partial differential equation
PEEK Polyether ether ketone
PoC Point of Care
PPV Positive Prediction Value
PS Polystyrene
pt Pre-train
RAD Relative accuracy difference
ROC Receiver Operating Characteristic
RoI Region of Interest
SCD Spray Carbon Dioxide
SERS Surface-Enhanced Raman Scattering
SNV Standard Normal Variate
SOM Self-Organising Map
SP Surface Plasmons
SPP Surface Plasmon Polaritons
TA Training Accuracy
TN True Negative
TNR True Negative Rate
TP True Positive
TPR True Positive Rate
TXT Text file format
VA Validation Accuracy
VGG16 Visual Geometry Group

List of Publications

The following papers form the core of this thesis:

- **P. D. C. Gomes**, J. J. S. Rickard, and P. Goldberg Oppenheimer, “Electrofluidynamic Patterning of Tailorable Nanostructured Substrates for Surface-Enhanced Raman Scattering,” ACS Appl. Nano Mater., vol. 3, no. 7, pp. 6774–6784, 2020, doi: 10.1021/acsanm.0c01190.
- M. Hardy, L. Kelleher, **P. de Carvalho Gomes**, E. Buchan, H. O. M. Chu, and P. Goldberg Oppenheimer, “Methods in Raman spectroscopy for saliva studies–a review,” Appl. Spectrosc. Rev., vol. 57, no. 3, pp. 177–233, 2022, doi: 10.1080/05704928.2021.1969944.
- **P. de Carvalho Gomes**, A. Crossman, E. Massey, J. J. Stanley Rickard and P. G. Oppenheimer, “Real-time validation of Surface-Enhanced Raman Scattering substrates via convolutional neural network algorithm”. Informatics in Medicine Unlocked. Elsevier BV, Vol. 33, p. 101076, 2022, doi: 10.1016/j.imu.2022.101076.
- E. Buchan, M. Hardy, **P. de C. Gomes**, L. Kelleher, H. O. M. Chu and P. G. Oppenheimer, “Emerging Raman spectroscopy and saliva-based diagnostics: from challenges to applications”. Applied Spectroscopy Reviews. Informa UK Limited; p. 1–38, 2022, doi: 10.1080/05704928.2022.2130351.

Additional paper publications that are not part of the main thesis work:

- N. Chourmouziadi Laleni, **P. D. C. Gomes**, K. Gkatzionis, and F. Spyropoulos, “Propolis particles incorporated in aqueous formulations with enhanced antibacterial performance,” Food Hydrocoll. Heal., vol. 1, no. November, p. 100040, 2021, doi: 10.1016/j.fhfh.2021.100040.
- **P. De Carvalho Gomes**, M. Hardy, Y. Tagger, J. J. S. Rickard, P. Mendes and P. G. Oppenheimer. “Optimization of Nanosubstrates toward Molecularly Surface-Functionalized Raman Spectroscopy”. The Journal of Physical Chemistry C. American Chemical Society (ACS), Vol. 126, p. 13774–84, 2022, doi: 10.1021/acs.jpcc.2c03524.

Papers to be submitted:

- **P. D. C. Gomes** and P. Goldberg Oppenheimer, “Open-source software for Raman data Plotting, Multivariate analysis and spectral Peak Fitting”
- Michael Macgregor-Fairlie, **Paulo De Gomes**, Daniel Weston and Pola Goldberg Oppenheimer, “Hybrid Raman Spectroscopy and Artificial Neural Network Algorithm Discriminating Mycobacterium Bovis BCG and Members of the Order Mycobacteriales”.

Ongoing work:

- **P. D. C. Gomes** and P. Goldberg Oppenheimer, “Surface functionalisation over EHD pillars”.
- **P. D. C. Gomes** and P. Goldberg Oppenheimer, “Molecular impriting over EHD pillars”.

Conference posters presented:

- **P. de Carvalho Gomes**, P. Goldberg Oppenheimer [2018], “Study of the Optical Field Scattering Enhancement on a 2D Rough Surface Using COMSOL Multiphysics®”, COMSOL Conference, SwissTech Convention Center, Lausanne, Switzerland (Poster)
- **P. de Carvalho Gomes**, P. Goldberg Oppenheimer [2019], “”, Point-of-Care Biosensors & Mobile Diagnostics Europe, Rotterdam, The Netherlands (Poster)

CHAPTER 1

Introduction

The world we live in is made out of patterns.

In nature, spiral patterns are observed in the shells of snails and leaf formations in some plants. Other plants have fractal structures, and bees make tessellated honeycombs. We can identify a relationship among many natural patterns by studying these patterns. When analysing the number of patterns appearing in nature, they seem to be present in the Fibonacci sequence[1,2]. By applying the ratio of consecutive numbers, it converges to the unique number called the golden ratio, φ.

In Space, there are other patterns. A year on Earth refers to the Earth planet doing one lap around the sun. Earth shows four repeatable patterns, what we call seasons: spring, summer, autumn and winter per year. The moon has four different phases per month, and the sea tides have two low and two high tidal patterns per day.

Patterns can also appear in the scientific field. In mathematics, for instance, if you have a circle and divide its perimeter by its diameter, you will always find the same constant called pi (π) and many other examples in trigonometry. Even remote fields of science like the chaos theory have patterns, as shown by Feigenbaum, who discovered two constant relations from periodic patterns in chaos, named the delta (δ) and alpha (α) constants[3,4].

Interestingly, humans have become experts at analysing patterns and extracting meaning due to their inherently pleasing nature. In fact, we could even say that our ability to identify patterns is innate. Our brains have been wired in such a way that we have

learned to interpret patterns as an adaptive trait for human survival[5]. The thought process behind pattern recognition learning in humans has been extensively studied and curiously has opened a new field of research in computer science called machine learning (ML)[6]. Researchers have tried to mimic the thought process happening inside a human brain into a computer circuit to create artificial intelligence (AI). One of the first attempts in using AI was to detect different patterns in images[7]. With time, new ML methods have been developed, like convolutional neural networks (CNN)[8], which are very powerful for image analysis and will be a core component in this thesis.

Nowadays, pattern identification has stretched past the pure sciences and has been applied in other fields. For example, in the medical field, one way of identifying if someone has cancer is by observing optical imaging of cells[9,10]. There are subtle differences between healthy and cancerous cells, so the job of multiple doctors is to access and classify the extracted sample, which is an extremely time-consuming task. However, with the use of CNN, this complex task can be performed by a single computer in no time[11]. This use of a CNN is one of many examples that can highlight the potential of systematically translating specialised pattern recognition into computational language. In this thesis, the same logic has been extended to optimise the identification of substrates capable of generating a high optical signal enhancement. Since the specific features that generate the signal enhancement are not trivial to find, it is possible to have a deeper insight into these features using CNN. This aspect is covered in the first part of the thesis.

Another field where patterns can give insight into nature is quantum mechanics. Patterns are used at the nanoscale level to identify specific symmetry groups in molecules. For example, molecular symmetries provide information on allowed and forbidden energy band transitions for scattering events. In particular, it is of special interest for Raman

spectroscopy, and based on the probe molecule selection rules, only certain vibrational energy states can be achieved[12].

Raman scattering describes the interaction between a quantum of light, photon and a lattice vibration quantum, phonon[13]. The photons come from a monochromatic light source, usually from a laser, which are scattered over the material to analyse. When the photon frequency matches the vibration mode frequency of the material bonds, a resonance is measured as a stronger peak in the Raman spectra[14]. Thus, analysing these peaks gives important information about the molecules present in the samples since each molecular bond has a unique vibration mode fingerprint. The matching between peaks and molecules is easy for single molecules; however, when dealing with compounds, identifying the different components of each compound becomes a complicated task. In order to solve this issue, ML models are applied, specifically, those using advanced statistical analysis like the principal components analysis (PCA)[15]. Methods such as PCA are required to separate multi-peak signals into their main components, enabling insight into the various molecular vibrations through the multi-peak locations. The method of breaking the multi-peak Raman signal and the resulting insight are the main topics of discussion in this thesis's second section.

In summary, this work focuses first on using CNN to optimise a substrate from Raman spectroscopy data and images and second on data analysis methods for Raman spectral data, specifically towards Raman peak analysis.

1.1 Thesis Outline

The structure of this thesis starts with an introduction to Raman scattering and surface-enhanced Raman scattering (SERS) and the surface capable of generating SERS, the electrohydrodynamic lithography (EHD) patterns. The core Python functions are then introduced, as well as the convolutional neural networks and their architecture, and also data visualisation, preprocessing, processing, and multivariate analysis tools.

Following the Introduction chapter, the experimental tools and analytical techniques used to complete the research within this thesis are described in detail in **Chapter 2**. The materials used to achieve the nano and micro EHD structures, the gold sputter coating tool to make EHD patterns SERS active, and the characterisation techniques used to analyse the fabricated substrates and devices, including Raman spectroscopy and optical microscopy and atomic force microscopy (AFM).

In **Chapter 3**, different configuration types are introduced and explored through the theory of EHD fabrication, which is at the heart of all the fabricated device technologies in this thesis, with growth simulations and unique pillar height calculations performed solely on the optical imaging of the EHD patterns.

In **Chapter 4**, it was demonstrated that a CNN model could be successfully trained to identify AFM images of EHD pillars based on their SERS activity and to correctly label the AFM images as SERS or non-SERS pillars based merely on their AFM images.

In **Chapter 5**, the issues from Chapter 4 were once more addressed, but this time, optical microscope images were used instead of AFM images from the EHD-patterned surfaces. This led to the development of a CNN model that could distinguish SERS from non-SERS EHD patterns from optical microscope images in real-time.

In **Chapter 6**, the best technique for visualising, pre-processing, processing, and analysing Raman data was identified by carefully studying the various Raman data analysis literature for saliva samples.

In **Chapter 7**, many saliva samples were evaluated on a Raman spectrometer and subsequent data analysis in accordance with the procedure described in Chapter 6. The main subject of the chapter is the examination of these saliva samples using a customised Python programme. A graphical user interface with numerous buttons and widgets that manipulate the data is part of the Python script. It can analyse and track any Raman spectral shift by using machine learning and a cutting-edge peak finder technique to discover peaks in the Raman spectral data.

Initially, the thesis outline comprised of chapters 3, 4, 5, and instead of the current chapter 6 and 7 topics, it would have a molecular imprinting project part. However, due to the Covid-19 pandemic and the restrictions, some of the lab work parts were delayed or even scrapped, and thus the topic shifted to a more literature review and computation work component, which explains the current chapter 6 and 7 that seem slightly off-topic, but were a consequence of adapting to the Covid-19 restrictions imposed on my lab work. Despite this, some preliminary results from the molecular imprinting project were published, and others are still ongoing work, as highlighted in the List of Publications section.

1.2 State of the art

1.2.1 Raman Scattering

Raman scattering has extensive applications for both quantitative and qualitative sample analysis. It can provide information quickly and is very popular for determining biological

molecules [16,17], chemical processes [18,19] and material properties [20,21]. Raman spectroscopy is an excellent tool to analyse any samples, giving a specific fingerprint related to the molecular vibration modes of the material in a non-destructive way. A common use is process control in bioprocessing, as it can offer monitoring of key variables in the production process without much sample preparation [22]. Raman spectroscopy is used in all experiments throughout this work. When operating the Raman spectrometer, multiple components need to be considered, mainly the wavelength choice of the monochromatic laser light and the material composition to analyse. A scattering event involving a monochromatic laser light and a material can produce two distinct scattering results (Figure 1.1). The first is Rayleigh scattering, which occurs when the resulting scattered wave has the same energy as the incoming wave (elastic scattering)[23], and the second is Raman scattering, which occurs when the frequency of the dispersed light differs from the frequency of the incident radiation in these interactions (inelastic scattering)[24].

The difference between the Rayleigh and Raman scattering processes is related to the molecule's initial and final energy states. The Rayleigh scattering occurs when a molecule is initially in its ground or vibrationally excited energy state, is excited by a photon to a virtual excited energy state, and then is de-excited, re-emitting the photon with the same frequency as the excited one. In Raman scattering, the molecule can be in the ground state, be excited to a higher virtual energy state, and then re-emit the photon with slightly less energy, resulting in the difference in energy being converted into a molecular vibrational state (Stokes scattering), or the molecule can be already in a vibrational state, be excited to a higher virtual energy state, and then re-emit the photon with slightly more energy, resulting in the molecule returning to (anti-Stokes scattering). In both Stokes and anti-Stokes cases, the energy leftover and excess energy after the scattering has

information on the vibrational mode of the molecule. Thus Raman scattering is highly specific at probing chemical vibrational bonds because the vibrational state involved in the energy transfer is unique and specific to each molecule. Therefore, the Raman spectrum can give an accurate fingerprint of the compounds that are being analysed. However, the Raman scattering probability is very small, where only 1 in 10^8 of the scattering events are Raman scattered[25]. The other scattering events are all Rayleigh scattering. Hence, to characterise a sample with Raman spectroscopy, a Rayleigh filter is needed to filter out the Rayleigh scattering, to only let through the Raman scattering photons. Therefore, one of the major drawbacks of Raman scattering is the high sample concentration required. Because of the low number of scattered Raman photons, Raman scattering does not always detect low concentrations from a compound. Thus, Surface-enhanced Raman spectroscopy (SERS) is used to overcome this limitation.

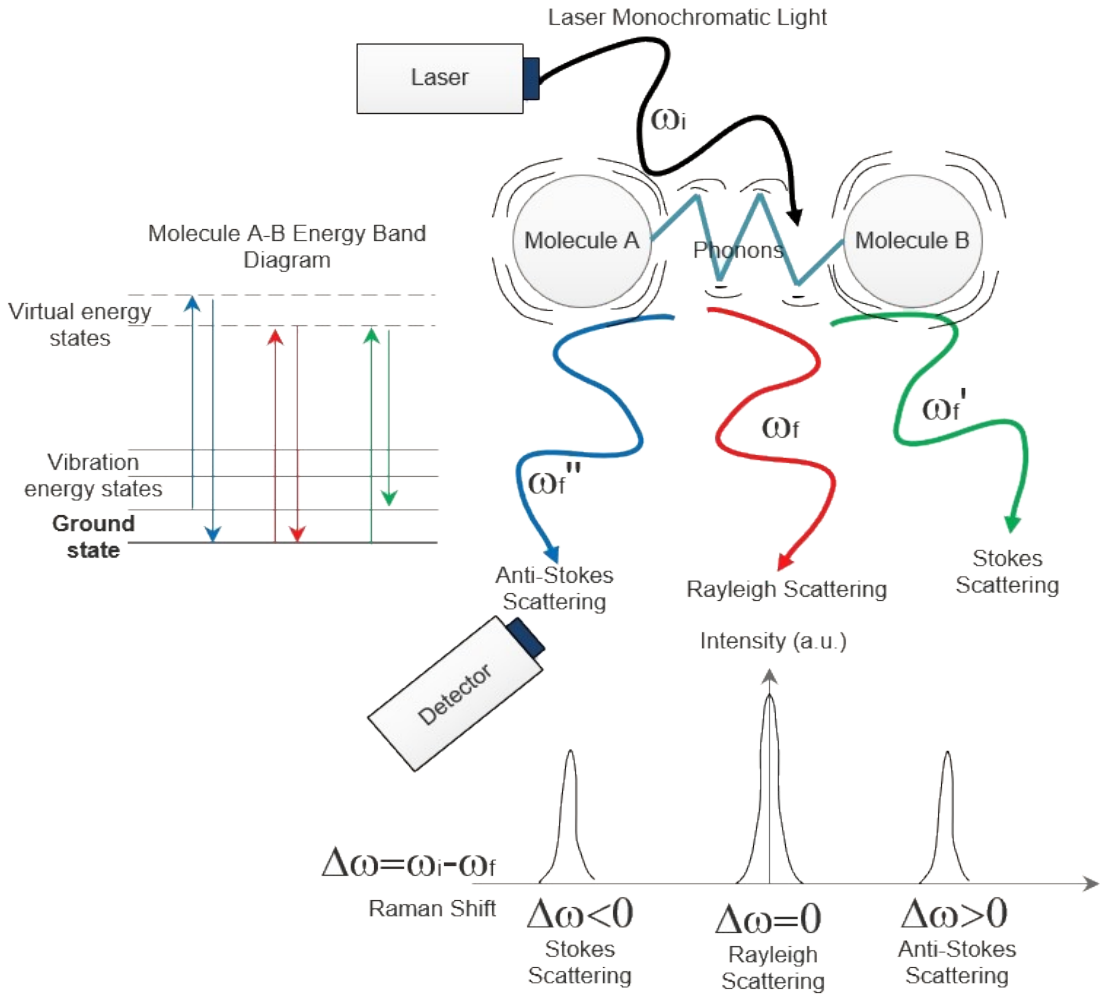


***Figure 1.1**: Light scattering on a phonon. Three different scattering outcomes result from this interaction, namely: Rayleigh scattering, $\Delta\omega$ (Raman shift) = 0, Stokes scattering $\Delta\omega<0$ and anti-Stokes scattering $\Delta\omega>0$ with their respective energy band diagram.*

1.2.2 Surface-enhanced Raman Scattering

Surface-enhanced Raman scattering (SERS) is a method used to amplify the Raman scattering signal. SERS was first discovered in 1974 by Fleischmann *et al.*[26], and after its discovery, many theories have been formulated to explain this phenomenon. The main theory explains that the enhancement comes in two forms, an electromagnetic and a chemical enhancement [26-35]. The electromagnetic enhancement comes from the surface plasmons from mainly transition metals such as gold, copper, and silver. Chemical enhancement, is derived from certain surface chemistries that improve the charge transfer

to target molecules and enhance the Raman signal. However, chemical enhancement is weaker than the electromagnetic enhancement. The maximum enhancement ascribed to chemical and electromagnetic enhancement is 10^2 [36] and 10^8 [37] respectively.

Instead of breaking down SERS into electromagnetic and chemical enhancement, we will split our SERS analysis into the entities that are present for SERS to occur, i.e., monochromatic light, a nanostructured rough surface, and an adsorbate. Moreover, the interaction among these entities is considered.

Laser light is a monochromatic electromagnetic source with power, energy, wavelength and frequency associated with the beam. The electromagnetic radiation from the laser has perpendicular, electric and magnetic field components, which are perpendicular to the wave propagation. Photons emitted from the laser are quantum in nature and have special properties (wave-particle duality), statistics (Bose-Einstein statistics) and polarisation inherent to their nature. A quantum mechanics treatment must detail the light interactions with a surface.

When a monochromatic light interacts with a surface, SERS is only verified for specific surfaces. Hence surfaces with the ability for SERS are called “SERS-active” surfaces. These are usually made from noble metals like silver, copper and gold. The appearance of SERS is a combination of roughness sites on the surface, close proximity for nano particles, and a plasmonic coating with noble metals. In the case of a silver coating, it shows better overall SERS along the nanometre to micrometre wavelength range [38]. However, it quickly oxidises in contact with air, and thus it is not entirely biocompatible without further functionalisation [39]. Here biocompatible is defined as not to cause toxic or harmful effects on a biological system[40]. Alternatively, gold can be used since it is mainly inert to air and biocompatible [41]. However, the enhancement is slightly

lower in wavelengths below 600nm because of interband transitions in gold that hinder the generation of plasmon-polaritons [38].

Finally, an adsorbate can be considered any molecule/entity/biomolecule adsorbed to the surface. There can be considerable differences when comparing the Raman and SERS spectra of the same adsorbate on different surfaces. Raman peaks that are not active can become active in SERS, and peaks located in a certain wavenumber can be shifted in SERS. The fact is that SERS measurements are more sensitive than Raman. Thus these subtle shifts and weak activations in Raman can be measured in SERS mostly because SERS is sensitive enough to probe how an adsorbate is orientated on the surface, which influences the spectral signal [42].

These are the three most important causes of SERS. However, because they all interact with one another, there are three types of relationships between these entities: laser-absorbate, adsorbate-surface, and laser-surface. The laser-adsorbate interaction gives the Raman spectra information of any adsorbate molecule on the surface. The information obtained from the Raman scattering is a fingerprint of the adsorbate, and it is unique for each adsorbate molecule. The surface and the adsorbate interaction can be considered an “adsorbate-surface” complex (**Figure 1.2.A**). The Raman signal from the adsorbate is enhanced from the surface through a charge transfer mechanism (chemical enhancement), where the adsorbate donates its charge to the surface, and these charges will oscillate as the free electrons inside the noble metal (electromagnetic enhancement), slightly increasing the enhancement (**Figure 1.2C**). Usually, when dealing with SERS, the focus is on the interaction between the laser light and surface-adsorbate complex. However, detailed information between separated entities is hard to predict.

Finally, the laser-surface interaction between the laser and the surface is responsible for the major enhancement. When dealing with the SERS surface, the

dielectric permittivity is the most important parameter to consider [43]. Dielectric permittivity determines how the material interacts with electromagnetic radiation and can be calculated by performing reflectance measurements over the surface. Finding how the dielectric permittivity changes in the micro-nano scale is hard to calculate precisely. The dielectric permittivity is usually given for bulk materials, but it can be roughness-dependent and frequency-dependent (**Figure 1.2B**). The dielectric permittivity of a noble metal SERS surface, such as gold, is negative at the excitation wavelengths (visible to near-infrared range) used for Raman scattering. When an electromagnetic field is dispersed across an interface between two materials where the dielectric function changes sign (e.g., the gold(negative)/air(positive) interface), generations of coherent delocalized electron oscillations known as surface plasmons (SP) are produced. SP on a metal surface generates an electromagnetic field outside the metal, resulting in the formation of a surface plasmon polariton (SPP). When SPP occurs at closed surfaces of small particles or roughness sites, it is referred to as localised surface plasmons (LSP). When different LSP modes are coupled, resonance can occur, which can boost the scattering field and is one of the mechanisms for electromagnetic enhancement in SERS. Similar enhancements are also observed over noble metal-coated nanoparticles when they are close to each other (**Figure 1.2D**). Overall, SERS has a fast detection time, with little or no sample preparation required, delivering highly sensitive results, making it one of the leading analytical techniques with multiple applications for sensors [44].

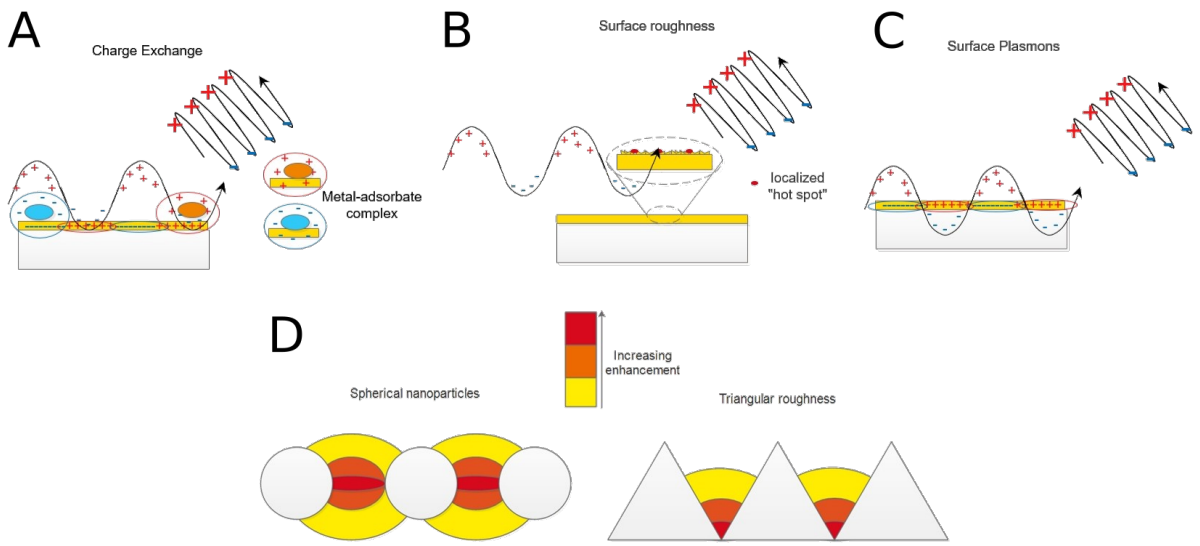


***Figure 1.2**: **(A)** Charge transfer mechanism between the adsorbate and the surface. This charge will slightly increase the enhancement of the SERS. **(B)** SERS due to nanometric roughness sites on the surface, makes highly localized enhancement of the electromagnetic field from the laser light, which will scatter with increasing intensity. **(C)** Interaction of electromagnetic radiation with the surface plasmons over the Nobel metal surfaces. **(D)** Electromagnetic field enhancement profile of two nanoparticles – where the enhancement is majorly centred around the space between the nanoparticles. And the enhancement profile between two roughness sites – roughness sites are simulated as hills and holes profile – where the enhancement is majorly centred in the hole part of the surface.*

Electrohydrodynamic (EHD) lithographed patterned structures[27] are the main surface used in this work for the SERS applications. Following the fabrication of the structures, they are gold-coated, which, when combined with the EHD patterned roughness, gives the patterns SERS properties. SERS is not located everywhere on the patterned structures, thus the aim is to develop a reliable method of finding SERS patterns using CNN yielding consistent high SERS signal pattern identification over the substrate.

1.2.3 Electrohydrodynamic Lithography (EHD)

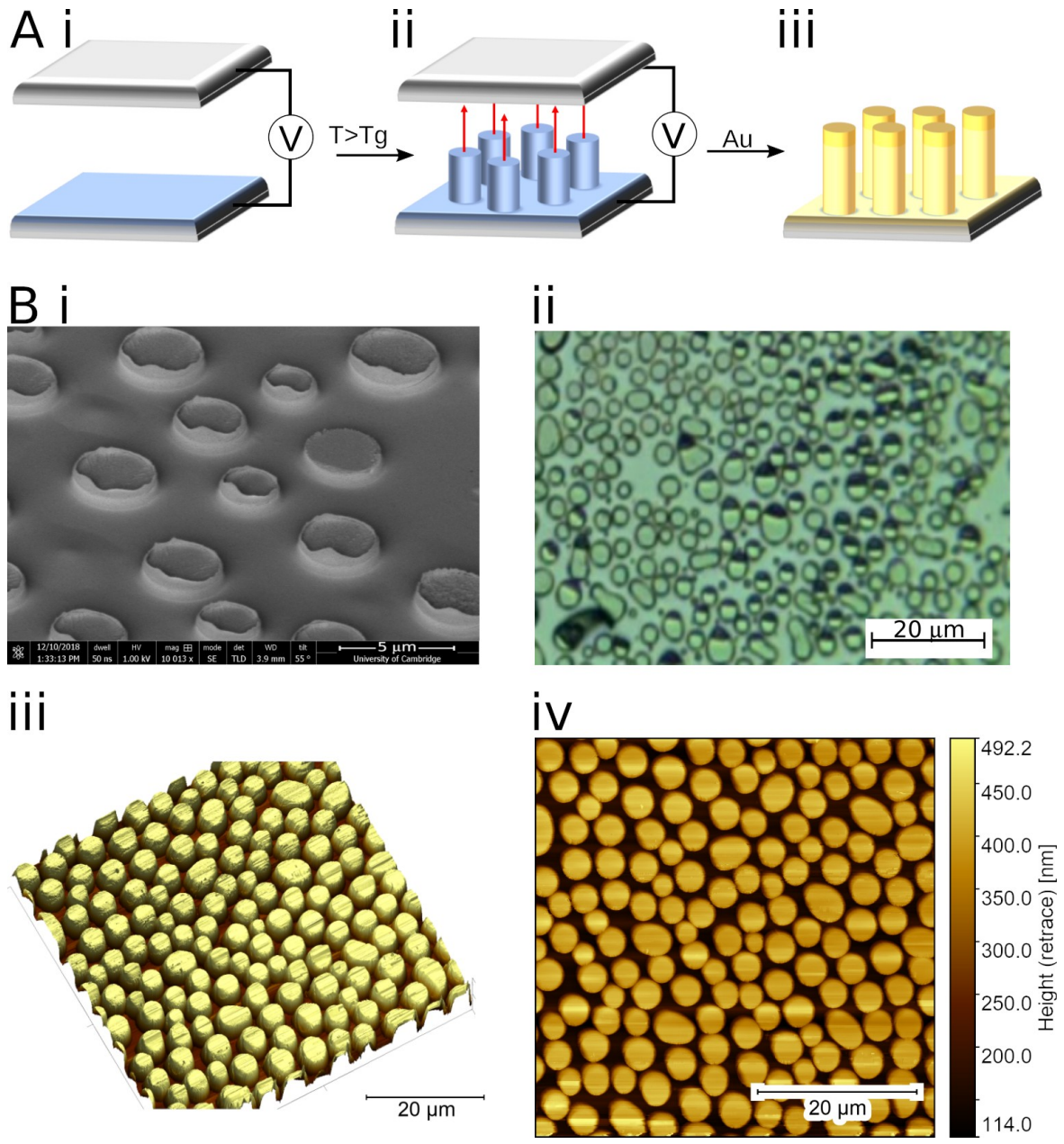


Figure 1.3: (A) (i) *Electrohydrodynamic lithography setup. Two parallel silicon substrates form a capacitor, with the bottom one coated with a micrometric thin polystyrene layer, after setting up a potential difference with a temperature above the glass transition temperature of the polymer* ***(ii)****. The generated heat with the potential difference creates an electrostatic pressure that provokes surface instabilities onto the polystyrene, which begins to raise towards the top electrode. After reaching the top electrode and cooldown, the process is terminated. After disassembling, a thin gold coating layer is sputter-coated, making the patterned structures SERS active* ***(iii)****.* ***(B)*** *EHD images from different image acquisition tools* ***(i)*** *scanning electron microscope (SEM),* ***(ii)*** *Renishaw optical microscope,* ***(iii)*** *AFM 3D,* ***(iv)*** *AFM 2D.*

To produce our SERS substrates, we used an EHD technique. The EHD process starts with the spin-coating of a thin polymer film onto a silicon substrate. A capacitor is formed when a conducting electrode is placed above the substrate and is separated by a micrometric spacer. The setup is heated above the polymer's glass transition temperature, and when an electrical potential difference is applied between the substrate and the top electrode, the film destabilises. The destabilisation causes disruptions of the capillary waves, which, as time passes, increase in amplitude, eventually reaching the top electrode that produces multiple arrays of pillars on the bottom electrode. Then, the temperature is reduced below the polymer's glass transition temperature, the voltage is then turned off, and the device is separated, leaving the formed pillars on the bottom substrate. Lastly, the patterns are sputter-coated with a thin gold layer, turning the structures SERS active. The experimental setup is shown in **Figure 1.3A**. If the pillars have a rough top, they can have up to 50 times more enhancement than a flat-topped pillar. Therefore, most SERS pillars will have rough tops [27,45]. However, identifying rough tops is not trivial and requires high-resolution tools to analyse them post-fabrication, which is time-consuming. As a result, ML algorithms are used to determine which structures are SERS and which are not. ML algorithms, such as CNN, can detect subtle variations in images and detect features that distinguish between SERS and non-SERS [27,46] that humans cannot. Another benefit is that the CNN model only needs to be trained once and can then be used whenever the model weights and biases are loaded.

1.2.4 Convolutional Neural Network (CNN)

A CNN is a form of ML popular for image classification problems. The use of CNN has dramatically increased with the improvement in computational power and has become

today's standard for computer vision applications[47]. The goal of any ML algorithm is to train on given data and correctly predict unseen data. Its applications have already been demonstrated to differentiate SERS from non-SERS EHD patterns AFM images [27]. Thus, CNN was chosen to analyse our optical images of pillars for real-time analysis on the Raman microscope.

Despite the architectural variations of a CNN, usually, it consists of 2D convolution layers [48], Max pooling layers[49], a Dropout layer[50], Flatten layer[51] and Dense layers[52]. These layers compute the input image in various ways so that at the end of the CNN, a categorical classification can be made.

A CNN starts with a 2D convolution (Conv2D) layer with two important parameters: filters and a kernel. The filters' parameter sets the number of convolution filters that a layer will learn. The second parameter, the kernel, defines a matrix size that convolves across the different channels from the input image, outputting a channel feature map. The concatenation of these feature maps results in the filters of the layer. The filters are initialised to small, random values which, during the CNN training, learn and are updated to distinguish the input image features. When the CNN model is optimised, the filters provide information about the classification mechanics of the CNN model. After the output feature map is created, it passes onto the Max pooling (MaxPooling) layer, which operates over the outputted feature map by downsampling the feature map and reassembling the maximums into a new feature map, which is then passed onto the next layer. Conv2D and MaxPooling layers can be stacked together into a multi-layered CNN, and these layers make up the main structure of a CNN. To conclude the CNN structure, additional layers are added for the appropriate classification outputs. The Dropout layer can be added, which applies to the last outputted feature map, which randomly sets to zero a percentage of output features of the layer. Usually, Dropout layers are used to prevent model

overfitting. Finally, the output feature map from the Dropout layer needs to be converted into a vector and to do this, a Flatten layer is used, which reduces the dimensionality of a feature map into a vector. The Flatten layer is always applied at the end of the CNN model to serve as input to a Dense layer, which can be connected to another Dense layer that together forms an output layer to the CNN. The last Dense layer is responsible for the classification of the input image.

A mathematical definition of an image is required to see how these different CNN layers affect the interpretation of an image. An image is represented by a 3D tensor, and it has a width *w*, a height *h* and a channel depth *c*. The width and height are related to the pixel size of the image, and the depth is related to the image channels. If it is a coloured image, it will have red, green, and blue channels (RGB), and if it is a greyscale image, it will have one channel value. Thus images can be described as a 3D tensor with shape (*w, h, 3*) if they are coloured and (*w, h, 1*) or (*w, h*) if they are in greyscale. After defining the mathematical representation of an image, we can analyse how the different layers used to build up a CNN operate over an image.

When put the CNN model into practice, large amount of data is required to optimise the different layers in order to be able to accurately predict the appropriate class for the input image [53]. The CNN optimisation starts with appropriate image data organisation and then separation into the three categories: training, validating, and testing. Training is used to train the model to adjust parameters and define important features within each image. While the model trains, it validates its training with the validation data, which is used to determine the model's accuracy and fine-tune different layer parameters. Finally, the testing data is used to evaluate the true accuracy of the model with no bias since these images were never seen before by the CNN, thus can give a real judgement on the optimised CNN model.

The CNN used is the VGG-16, a pre-trained CNN developed by Simonyan and Zisserman at the University of Oxford for the ImageNet image classification competition in 2014 [54]. In this competition, the goal was to create a program able to classify 1.2 million images into 1000 categories. VGG-16 came second in this competition with a test error of 7.3%, later improved to 6.8%. Since its excellent performance in the ImageNet competition, VGG-16 has been used as a base for countless image recognition tasks via transfer learning. Transfer learning uses pre-trained models as a base for other models with more generic features already understood by the network (for example, shapes and objects), so only model-specific features must be learnt, commonly by fine-tuning the final layers of the CNN [54–56].

An excellent example of VGG-16 application is within a model for pathological brain image classification. Magnetic resonance (MR) brain images were classified using Kaur and Gandhi's VGG-16 containing model [57], resulting in near perfect accuracy, specificity and sensitivity results, considerably better than all current gold standard classification methods. Another example of using the VGG16 neural network was to be able to distinguish between SERS-active EHD pillars and non-SERS-active EHD pillars and relate them to their AFM images. The model generated obtained an accuracy of 99% [27].

Besides the CNN, other ML techniques have been used to preprocess, process and analyse the Raman spectral data.

1.2.5 Raman Data Preprocessing

When measuring a sample through a Raman spectrometer, the scattering signal is acquired through a charge coupled device (CCD) located at the end of the instrument’s optical path and its signal is processed with the appropriate analysis software on a computer. When handling the raw data obtained from the measurement, it is required to be

preprocessed to remove undesirable features to make all data comparable for a proper data analysis.

Following the Raman preprocessing operation sequence proposed by Hardy et al. [10] any change to the raw data prior to the multivariable data analysis procedure is referred to as "data preprocessing." The most commonly used preprocessing methods for Raman spectroscopy data are: cosmic ray removal [58-64], baseline subtraction [60,65], smoothing [60,66-69], and principal component analysis (PCA) [62,70-73]. However, there is no standard preprocessing methodology, and each author uses their preferred preprocessing method. Here, the most commonly used methods are highlighted, along with their advantages and disadvantages.

After acquiring the raw Raman data, the first preprocessing is always cosmic ray removal [74]. Cosmic rays are everyday radiation from space unrelated to the sample being analysed. They are occasionally captured by the Raman spectroscope, which is unavoidable. Thus, the solution is to remove them using algorithms, and all authors consider this approach valid.

Following the removal of cosmic rays, baseline subtraction is performed. Baseline is the underline profile in a Raman spectrum that is sometimes removed or not depending on the author. The majority of authors approach baseline subtraction in the following manner:

If it is removed, the Raman spectra will lose relevant information. If it is not removed, the baseline becomes sample dependent. When looking for a relevant Raman fingerprint, overestimate its contributions due to baseline contributions, which is not desirable in data preprocessing. Finally, regardless of whether baseline subtraction is used or not, it should be used consistently across all data samples.

After the baseline subtraction is performed, smoothing operations can be used. Smoothing is a process of de-noising the signal to turn it smoother. Most authors avoid smoothing because it might remove important information related to specific spectra fingerprints, although some authors use smoothing just for aesthetic reasons for plotting. However, smoothing as a preprocessing tool is usually not performed.

Normalisation is the next step in data preprocessing. There are three ways to achieve normalisation. First, by resizing the spectra to a fixed range between zero and one, called the min-max normalisation. Second, resizing the spectra based on their average and the standard deviation is called the standard normal variate (SNV) normalisation. Third, the normalisation can be accomplished by using the area of a common peak in the spectra and resizing the spectra by this area. Like the baseline subtraction option, the normalisation also divides some authors. The argument is that since Raman spectra intensity values are related to a quantitative measurement (i.e. the amount of a chemical compound in the sample), if it is normalised, it will lose the quantitative aspect, and it is a mere qualitative measurement. Because of this, some authors usually reserve using normalisation for the multivariable data analysis part of the Raman analysis and leave it not normalised for all the quantitative analysis.

Finally, PCA can be used as preprocessing algorithm since PCA is used as a dimensional reduction algorithm, reducing the data into the most important components, which maximises the variance between different sample spectra. This process can reduce the initial data sample size into the most significant PCA components and thus compact the data into its most important components. When PCA is used as a preprocessing algorithm, its output components are subsequently analysed by another multivariable analysis algorithm like the linear discriminate analysis (LDA)[75,76] or support vector

machine (SVM) [77,78]. Thus, some authors use PCA as a preprocessing algorithm and others solely as a multivariable analysis tool.

1.2.6 Multivariate Data Analysis

After performing all the necessary preprocessing algorithms to the raw Raman data, the data is fed through the multivariate data analysis. The multivariate data analysis procedure takes care of analysing complex multi-class Raman spectra. When dealing with multi-class Raman spectra, we analyse a large amount of data, which quickly turns into a ‘big data’ problem [79]. Here computational performances in analysing large amounts of data start to matter, so it is required to have the proper hardware to tackle multivariate data analysis. Luckily with today’s advances in technology, most personal computers are powerful enough to perform some of these complex tasks. Usually, these procedures involve using Artificial Neural Networks [59,80-83] processes and complex statistical algorithms, which take a load on standard computers. However, these techniques are required to analyse multivariate data, so we the most up-to-date tools is needed to extract meaningful information and effectively separate it from insignificant information.

One of the most standard multivariate analyses is the principal component analysis (PCA) [62,70-73]. PCA is an unsupervised learning tool that separates the multi-class spectra by redefining the data into a new coordinate system that maximizes the variance between data points, thus sorting out the similar spectra into unique clusters. There is no need to classify the individual spectra to use PCA since PCA is an unsupervised tool. Thus, the algorithm can sort out the data without any label input by only inputting the preprocessed data, which shows the power of multivariate data analysis tools. Another important feature of the PCA analysis is the PCA loadings, representing the new coordinate systems with

the highest variance values defined by the PCA. The loadings provide further insight into how the different spectra are clustered, conveying each PC's most important Raman shift values. Thus, identifying all the pieces of information needed to analyse multi-class Raman spectra.

Besides PCA, there are also supervised learning tools like linear discriminant analysis (LDA)[75,76], support vector machining (SVM)[77,78] and partial least squares (PLS) [84]. These perform the same dimensionality reduction task as the PCA while considering the correlation between dependent (spectra class) and independent (Raman intensity) variables when performing the class separation. Here, the spectra data requires manual identification of which class the spectra belong to by the user. Usually, the output is more accurate using the supervised methods since they improve the variance between spectra and the variance between the different labelled classes; thus, they improve the classification slightly with this added consideration. Although researchers do not know the nature of their samples in some cases, these techniques cannot be applied since the samples cannot be correctly labelled. Besides these statistical multivariate analysis tools, there are also artificial neural networks (ANN) to analyse spectral data.

ANN algorithms function as the neuron network in the brain where multiple layers have interconnected nodes that are activated or not depending on the supplied data. Each node connection is assigned a weight and a bias trained to learn the differences among Raman spectra. The weights and biases are tuned *via* a backpropagation algorithm to minimize the predicted and desired output error. The ANN can be effectively trained to distinguish multi-class Raman samples by running many spectra data combinations. The output weights and biases can be inspected when the correct distinction of different classes occurs, which provides deep insight into the ANN classification of the different classes and are very important. These output weights and biases are called neuron

activation. They can be used in the data post-process to identify different Raman labels based on the activation of neurons signals.

One custom ANN that has worked well for Raman data is self-organizing maps (SOM) [80]. SOM is an unsupervised technique that reduces the dimensionality of the Raman data. The map organizes the same class spectra into the same region as clusters. Depending on the proximity of other clusters, the more the spectra are similar or different from each other. In the SOM map, each spot is represented as a neuron that only accepts similar spectra as the ones the neuron was trained on, thus providing information related to the different class fingerprints and similar classes based on the proximity.

1.2.7 Raman Data Visualisation

Visualizing the spectra to inspect the data is a fundamental tool for Raman spectra analysis. The visual inspections of the data can be performed before and after preprocessing to check if the preprocessing is correctly applied and after performing the multivariable analysis to interpret the output results. Different authors use different visual representations for the Raman spectra; most choose to average out the spectra of a sample, some authors normalize the spectrum, and others leave it with the appropriate intensity units. Some perform a baseline subtraction, others leave it as normal, and some show the average plot with a standard deviation outline to show the dispersion of the data. However, the most common procedure is to plot the average Raman spectra with baseline subtraction and without normalization. For the multivariable analysis using PCA, the choice is to plot the two most relevant principal components (two-dimensional plot) that show a proper separation of the multi-class Raman data and then fit a confidence ellipsoid[85] highlighting the different PCA clusters to help to visualise the cluster separation.

Occasionality 3D plots are also used, but they are not as visually explanatory as the 2D plots, so they are usually avoided.

1.2.8 Raman Peak Analysis

Finally, in quantitative Raman analysis, most authors focus on the intensity difference of common peaks or the appearance and disappearance of specific peaks when comparing multi-class Raman spectra. This comparison is usually a single comparison between individual spectra from different classes, and it fits the single analysis described previously. Single Raman spectrum analysis is fairly straightforward, usually involving peak Area/Intensity comparisons, as seen in multiple graphene Raman analysis studies [86]. For graphene, it is enough to calculate the intensity ratio between the 2D and G peaks to know if we have high or low-quality graphene. Although it is simple to perform, the information from this analysis can be very insightful in both quantitative and qualitative analysis. Usually, authors present a list of peaks that are being changed (positively or negatively) when comparing the Raman spectrum for different classes.

1.2.9 Graphical User Interfaces

All current available Raman analysis software have a graphical user interface. Some examples are WiRE software (Windows-based Raman Environment – from Renishaw), LabSpec (from Horiba), Spectragryph [87], Raman Tool Set [88] and spectools [89]. Raman analysis software are mostly closed-source software capable of acquiring Raman data and performing the basic process of the Raman data like baseline subtraction, normalization, cosmic ray removal, average calculation, and de-noising. However, more sophisticated multi-class analysis tools are not easy to use nor flexible, and in the end, they are not used

by most scientists, whom often prefer more flexible and easily tuneable data analysis procedures that can be achieved using home-built code in R-code [90] or Python. These can offer more dedicated modules to perform a more personalized multivariate data analysis.

1.2.10 Summary

This chapter sets the foundation for understanding the project by providing an overview of the physical interpretation of Raman scattering and introducing a specific surface fabrication technique using EHD that enhances the Raman signal (SERS) when coated with a plasmonic material. The chapter also introduces computational tools used for SERS spectral analysis and classification, including the use of CNN and Python-based algorithms for data preprocessing, analysis, and representation. These tools and techniques serve as the foundation for the main topics of the thesis.

1.3 References

[1] Newton, L. D. “Fibonacci and Nature: Mathematics Investigations for Schools”. *Mathematics in School*, Vol. *16*, no. 5, **1987**, p. 2–8. Available from: http://www.jstor.org/stable/30214387.

[2] Sinha, S. "The Fibonacci numbers and its amazing applications." *International Journal of Engineering Science Invention* 6, no. 9, **2017**, p. 7-14. Available from: https://www.researchgate.net/publication/330740074_The_Fibonacci_Numbers_and_Its_Amazing_Applications

[3] Feigenbaum, M. J. "Universal behavior in nonlinear systems." *Physica. D, Nonlinear phenomena* 7, no. 1-3, **1983**, p. 16-39.

[4] Briggs, K. "A precise calculation of the Feigenbaum constants." *Mathematics of computation* 57, no. 195, **1991**, p. 435-439.

[5] Mattson M. P. “Superior pattern processing is the essence of the evolved human brain.” *Frontiers in neuroscience*, *8*, **2014**, p. 265. Available from: https://doi.org/10.3389/fnins.2014.00265.

[6] Arbib, M.A. “The Handbook of Brain Theory and Neural Networks” second edition, Cambridge, Massachusetts, MIT Press, November 22, **2002**.

[7] Rosenfeld, A., and Wechsler, H. "Pattern recognition: Historical perspective and future directions." *International Journal of Imaging Systems and Technology* 11, no. 2, **2000**, p. 101-116.

[8] Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. "A survey of convolutional neural networks: analysis, applications, and prospects." *IEEE transactions on neural networks and learning systems*, Jun. **2021**.

[9] Hadjipanayis, C. G., Jiang, H., Roberts, D. W., and Yang, L. “Current and future clinical applications for optical imaging of cancer: from intraoperative surgical guidance to cancer screening.” *Seminars in oncology*, *38*(1), **2011**, p. 109–118. Available from: https://doi.org/10.1053/j.seminoncol.2010.11.008

[10] Hunter, B., Hindocha, S., and Lee, R. W. “The Role of Artificial Intelligence in Early Cancer Diagnosis.” *Cancers*, *14*(6), **2022**, p. 1524. Available from: https://doi.org/10.3390/cancers14061524.

[11] Kumar, Y., Koul, A., Singla, R., and Ijaz, M. F. "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda". Journal of ambient intelligence and humanized computing, 2022, p.1–28. Available from: https://doi.org/10.1007/s12652-021-03612-z

[12] Jorio, A., Mueller, N.S. and Reich S. “Symmetry-derived selection rules for plasmon-enhanced Raman scattering”. Vol. 95, Physical Review B. American Physical Society (APS); **2017**, 155409. Available from: http://dx.doi.org/10.1103/PhysRevB.95.155409.

[13] Böer, K.W. and Pohl, U.W. “Photon-Phonon Interaction”. Semiconductor Physics. Springer International Publishing; **2015**. p. 1–36. Available from: http://dx.doi.org/10.1007/978-3-319-06540-3_11-1.

[14] Clark, R.J.H. and Dines, T.J. “Resonance Raman Spectroscopy, and Its Application to Inorganic Chemistry. New Analytical Methods (27)”. Vol. 25, Angewandte Chemie International Edition in English. Wiley; **1986**. p. 131–58. Available from: http://dx.doi.org/10.1002/anie.198601311.

[15] Abdi, H. and Williams, L.J. “Principal component analysis”. Vol. 2, Wiley Interdisciplinary Reviews: Computational Statistics. Wiley; **2010**. p. 433–59. Available from: http://dx.doi.org/10.1002/wics.101.

[16] Allakhverdiev ES, Khabatova VV, Kossalbayev BD, Zadneprovskaya EV, Rodnenkov OV, Martynyuk TV, et al. Raman Spectroscopy and Its Modifications Applied to Biological and Medical Research. Vol. 11, Cells. MDPI AG; **2022**. p. 386. Available from: http://dx.doi.org/10.3390/cells11030386.

[17] Caserta G, Zuccarello L, Barbosa C, Silveira CM, Moe E, Katz S, et al. Unusual structures and unknown roles of FeS clusters in metalloenzymes seen from a resonance Raman spectroscopic perspective. Vol. 452, Coordination Chemistry Reviews. Elsevier BV; **2022**. p. 214287. Available from: http://dx.doi.org/10.1016/j.ccr.2021.214287.

[18] Kneipp K, Kneipp H, Itzkan I, Dasari RR, Feld MS. Ultrasensitive Chemical Analysis by Raman Spectroscopy. Vol. 99, Chemical Reviews. American Chemical Society (ACS); **1999**. p. 2957–76. Available from: http://dx.doi.org/10.1021/cr980133r.

[19] Schlücker S. Surface-Enhanced Raman Spectroscopy: Concepts and Chemical Applications. Vol. 53, Angewandte Chemie International Edition. Wiley; **2014**. p. 4756–95. Available from: http://dx.doi.org/10.1002/anie.201205748.

[20] Childres I., Jauregui L. A., Park W., Caoa H., and Chena Y. P., Raman spectroscopy of graphene and related materials, New Dev. Phot. Mater. Res., pp. 403–418, **2013**.

[21] Gouadec G, Colomban P. Raman Spectroscopy of nanomaterials: How spectra relate to disorder, particle size and mechanical properties. Vol. 53, Progress in Crystal Growth and Characterization of Materials. Elsevier BV; **2007**. p. 1–56. Available from: http://dx.doi.org/10.1016/j.pcrysgrow.2007.01.001.

[22] Goldrick S, Lovett D, Montague G, Lennox B. Influence of Incident Wavelength and Detector Material Selection on Fluorescence in the Application of Raman Spectroscopy to a Fungal Fermentation Process. Vol. 5, Bioengineering. MDPI AG; **2018**. p. 79. Available from: http://dx.doi.org/10.3390/bioengineering5040079.

[23] Miles, R.B., Lempertm W.R. and Forkeym J.N.. “Laser Rayleigh scattering”. Vol. 12, Measurement Science and Technology. IOP Publishing; **2001**. p. 33–51. Available from: http://dx.doi.org/10.1088/0957-0233/12/5/201.

[24] Hahn, D. W. "Raman scattering theory." *Department of Mechanical and Aerospace Engineering, University of Florida* (2007).

[25] Jones, R. R., Hooper, D. C., Zhang, L., Wolverson, D., and Valev, V. K. “Raman Techniques: Fundamentals and Frontiers”. *Nanoscale research letters*, *14*(1), **2019**, 231. Available from: https://doi.org/10.1186/s11671-019-3039-2.

[26] Fleischmann M, Hendra PJ, McQuillan AJ. Raman spectra of pyridine adsorbed at a silver electrode. Vol. 26, Chemical Physics Letters. Elsevier BV; **1974**. p. 163–6. Available from: http://dx.doi.org/10.1016/0009-2614(74)85388-1.

[27] Gomes PDC, Rickard JJS, Goldberg Oppenheimer P. Electrofluidynamic Patterning of Tailorable Nanostructured Substrates for Surface-Enhanced Raman Scattering. Vol. 3, ACS Applied Nano Materials. American Chemical Society (ACS); **2020**. p. 6774–84. Available from: http://dx.doi.org/10.1021/acsanm.0c01190.

[28] Hardy M, Kelleher L, de Carvalho Gomes P, Buchan E, Chu HOM, Goldberg Oppenheimer P. Methods in Raman spectroscopy for saliva studies – a review. Vol. 57, Applied Spectroscopy Reviews. Informa UK Limited; **2021**. p. 177–233. Available from: http://dx.doi.org/10.1080/05704928.2021.1969944.

[29] Goldberg-Oppenheimer P, Mahajan S, Steiner U. Hierarchical Electrohydrodynamic Structures for Surface-Enhanced Raman Scattering. Vol. 24, Advanced Materials. Wiley; **2012**. p. OP175–80. Available from: http://dx.doi.org/10.1002/adma.201104159.

[30] Mahajan S, Cole RM, Soares BF, Pelfrey SH, Russell AE, Baumberg JJ, et al. Relating SERS Intensity to Specific Plasmon Modes on Sphere Segment Void Surfaces. Vol. 113, The Journal of Physical Chemistry C. American Chemical Society (ACS); **2009**. p. 9284–9. Available from: http://dx.doi.org/10.1021/jp900661u.

[31] Mahajan S, Cole RM, Speed JD, Pelfrey SH, Russell AE, Bartlett PN, et al. Understanding the Surface-Enhanced Raman Spectroscopy “Background”. Vol. 114, The Journal of Physical Chemistry C. American Chemical Society (ACS); **2009**. p. 7242–50. Available from: http://dx.doi.org/10.1021/jp907197b.

[32] Du J, Xu J, Sun Z, Jing C. Au nanoparticles grafted on Fe3O4 as effective SERS substrates for label-free detection of the 16 EPA priority polycyclic aromatic hydrocarbons. Vol. 915, Analytica Chimica Acta. Elsevier BV; **2016**. p. 81–9. Available from: http://dx.doi.org/10.1016/j.aca.2016.02.009.

[33] Huang Y, Zhang X, Ringe E, Ma L, Zhai X, Wang L, et al. Detailed correlations between SERS enhancement and plasmon resonances in subwavelength closely spaced Au nanorod arrays. Vol. 10, Nanoscale. Royal Society of Chemistry (RSC); **2018**. p. 4267–75. Available from: http://dx.doi.org/10.1039/c7nr08959g.

[34] Ding SY, You EM, Tian ZQ, Moskovits M. Electromagnetic theories of surface-enhanced Raman spectroscopy. Vol. 46, Chemical Society Reviews. Royal Society of Chemistry (RSC); **2017**. p. 4042–76. Available from: http://dx.doi.org/10.1039/c7cs00238f.

[35] Li Y, Zhou L, Tang L, Li M, He JJ. Improved Surface Enhanced Raman Scattering Based on Hybrid Au Nanostructures for Biomolecule Detection. Vol. 8, IEEE Photonics Journal. Institute of Electrical and Electronics Engineers (IEEE); **2016**. p. 1–7. Available from: http://dx.doi.org/10.1109/JPHOT.2016.2619063.

[36] Su Y, Shi Y, Wang P, Du J, Raschke MB, Pang L. Quantification and coupling of the electromagnetic and chemical contributions in surface-enhanced Raman scattering. Vol. 10, Beilstein Journal of Nanotechnology. Beilstein Institut; **2019**. p. 549–56. Available from: http://dx.doi.org/10.3762/bjnano.10.56.

[37] Kaneko S, Watanabe S, Fujii S, Nishino T, Kiguchi M. The practical electromagnetic effect in surface-enhanced Raman scattering observed by the lithographically fabricated gold nanosquare dimers. Vol. 10, AIP Advances. AIP Publishing; **2020**. p. 025301. Available from: http://dx.doi.org/10.1063/1.5126981

[38] Lee KS, El-Sayed MA. Gold and Silver Nanoparticles in Sensing and Imaging: Sensitivity of Plasmon Response to Size, Shape, and Metal Composition. Vol. 110, The Journal of Physical Chemistry B. American Chemical Society (ACS); **2006**. p. 19220–5. Available from: http://dx.doi.org/10.1021/jp062536y

[39] Garg S, Rong H, Miller CJ, Waite TD. Oxidative Dissolution of Silver Nanoparticles by Chlorine: Implications to Silver Nanoparticle Fate and Toxicity. Vol. 50, Environmental Science & Technology. American Chemical Society (ACS); **2016**. p. 3890–6. Available from: http://dx.doi.org/10.1021/acs.est.6b00037.

[40] Dorland, W.A.N. "*Dorland's Illustrated Medical Dictionary*". 31st ed. Philadelphia PA: Saunders; **2007**.

[41] Kus-Liśkiewicz M, Fickers P, Ben Tahar I. Biocompatibility and Cytotoxicity of Gold Nanoparticles: Recent Advances in Methodologies and Regulations. Vol. 22, International Journal of Molecular Sciences. MDPI AG; **2021**. p. 10952. Available from: http://dx.doi.org/10.3390/ijms222010952.

[42] Park M, Choi JS, Yang L, Lee H. Raman Spectra Shift of Few-Layer IV-VI 2D Materials. Vol. 9, Scientific Reports. Springer Science and Business Media LLC; **2019**. Available from: http://dx.doi.org/10.1038/s41598-019-55577-x.

[43] Polubotko AM, Chelibanov VP. The theory of SERS on semiconductor and dielectric substrates. Vol. 122, Optics and Spectroscopy. Pleiades Publishing Ltd; **2017**. p. 937–43. Available from: http://dx.doi.org/10.1134/S0030400X17060169.

[44] Liu W, Man Z, Hua L, Chen A, Wang Y, Qian K, et al. Data mining methods of lung cancer diagnosis by saliva tests using surface enhanced Raman spectroscopy. 2014 7th International Conference on Biomedical Engineering and Informatics. IEEE; **2014**. Available from: http://dx.doi.org/10.1109/BMEI.2014.7002849.

[45] Mahajan S, Hutter T, Steiner U, Goldberg Oppenheimer P. Tunable Microstructured Surface-Enhanced Raman Scattering Substrates via Electrohydrodynamic Lithography. Vol. 4, The Journal of Physical Chemistry Letters. American Chemical Society (ACS); **2013**. p. 4153–9. Available from: http://dx.doi.org/10.1021/jz4018688.

[46] de Carvalho Gomes, P., Crossman, A., Massey, E., Stanley Rickard, J.J., Oppenheimer, P.G. "Real-time validation of Surface-Enhanced Raman Scattering substrates via convolutional neural network algorithm". Vol. 33, Informatics in Medicine Unlocked. Elsevier BV; **2022**. p. 101076. Available from: http://dx.doi.org/10.1016/j.imu.2022.101076.

[47] Sharma N, Jain V, Mishra A. An Analysis Of Convolutional Neural Networks For Image Classification. Vol. 132, Procedia Computer Science. Elsevier BV; **2018**. p. 377–84. Available from: http://dx.doi.org/10.1016/j.procs.2018.05.198.

[48] Convolutional 2D layer URL="https://keras.io/api/layers/convolution_layers/convolution2d/", accessed: 17/05/2022.

[49] 2D Max Pooling layer URL="https://keras.io/api/layers/pooling_layers/max_pooling2d/", accessed: 17/05/2022.

[50] Dropout layer URL="https://keras.io/api/layers/regularization_layers/dropout/", accessed: 17/05/2022.

[51] Flatten layer URL="https://keras.io/api/layers/reshaping_layers/flatten/", accessed: 17/05/2022.

[52] Dense layer URL="https://keras.io/api/layers/core_layers/dense/", accessed: 17/05/2022.

[53] Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. Vol. 35, IEEE Transactions on Medical Imaging. Institute of Electrical and Electronics Engineers (IEEE); **2016**. p. 1285–98. Available from: http://dx.doi.org/10.1109/TMI.2016.2528162.

[54] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv; **2014**. Available from: https://arxiv.org/abs/1409.1556

[55] Krishnaswamy Rangarajan A, Purushothaman R. Disease Classification in Eggplant Using Pre-trained VGG16 and MSVM. Vol. 10, Scientific Reports. Springer Science and Business Media LLC; **2020**. Available from: http://dx.doi.org/10.1038/s41598-020-59108-x.

[56] Tammina S. Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. Vol. 9, International Journal of Scientific and Research Publications (IJSRP). International Journal of Scientific and Research Publications (IJSRP); **2019**. p. p9420. Available from: http://dx.doi.org/10.29322/IJSRP.9.10.2019.p9420.

[57] Kaur T, Gandhi TK. Automated Brain Image Classification Based on VGG-16 and Transfer Learning. **2019** International Conference on Information Technology (ICIT). IEEE; 2019. Available from: http://dx.doi.org/10.1109/ICIT48102.2019.00023.

[58] Maitra I, Morais CLM, Lima KMG, Ashton KM, Date RS, Martin FL. Raman spectral discrimination in human liquid biopsies of oesophageal transformation to adenocarcinoma. Vol. 13, Journal of Biophotonics. Wiley; **2019**. Available from: http://dx.doi.org/10.1002/jbio.201960132.

[59] Ralbovsky NM, Halámková L, Wall K, Anderson-Hanley C, Lednev IK. Screening for Alzheimer's Disease Using Saliva: A New Approach Based on Machine Learning and Raman Hyperspectroscopy. Vol. 71, Journal of Alzheimer's Disease. IOS Press; **2019**. p. 1351–9. Available from: http://dx.doi.org/10.3233/JAD-190675.

[60] Muro CK, de Souza Fernandes L, Lednev IK. Sex Determination Based on Raman Spectroscopy of Saliva Traces for Forensic Purposes. Vol. 88, Analytical Chemistry. American Chemical Society (ACS); **2016**. p. 12489–93. Available from: http://dx.doi.org/10.1021/acs.analchem.6b03988.

[61] Al-Hetlani E, Halámková L, Amin MO, Lednev IK. Differentiating smokers and nonsmokers based on Raman spectroscopy of oral fluid and advanced statistics for forensic applications. Vol. 13, Journal of Biophotonics. Wiley; **2019**. Available from: http://dx.doi.org/10.1002/jbio.201960123.

[62] Feng S, Huang S, Lin D, Chen G, Xu Y, Li Y, et al. Surface-enhanced Raman spectroscopy of saliva proteins for the noninvasive differentiation of benign and malignant breast tumors. International Journal of Nanomedicine. Informa UK Limited; **2015**. p. 537. Available from: http://dx.doi.org/10.2147/IJN.S71811.

[63] Qian K, Wang Y, Hua L, Chen A, Zhang Y. New method of lung cancer detection by saliva test using surface-enhanced Raman spectroscopy. Vol. 9, Thoracic Cancer. Wiley; **2018**. p. 1556–61. Available from: http://dx.doi.org/10.1111/1759-7714.12837.

[64] Dies H, Raveendran J, Escobedo C, Docoslis A. Rapid identification and quantification of illicit drugs on nanodendritic surface-enhanced Raman scattering substrates. Vol. 257, Sensors and Actuators B: Chemical. Elsevier BV; **2018**. p. 382–8. Available from: http://dx.doi.org/10.1016/j.snb.2017.10.181.

[65] Rekha P, Aruna P, Brindha E, Koteeswaran D, Baludavid M, Ganesan S. Near-infrared Raman spectroscopic characterization of salivary metabolites in the discrimination of normal from oral premalignant and malignant conditions. Vol. 47, Journal of Raman Spectroscopy. Wiley; **2016**. p. 763–72. Available from: http://dx.doi.org/10.1002/jrs.4897.

[66] Cao G, Chen M, Chen Y, Huang Z, Lin J, Lin J, et al. A potential method for non-invasive acute myocardial infarction detection based on saliva Raman spectroscopy and multivariate analysis. Vol. 12, Laser Physics Letters. IOP Publishing; **2015**. p. 125702. Available from: http://dx.doi.org/10.1088/1612-2011/12/12/125702.

[67] Zamora-Mendoza BN, Espinosa-Tanguma R, Ramírez-Elías MG, Cabrera-Alonso R, Montero-Moran G, Portales-Pérez D, et al. Surface-enhanced raman spectroscopy: A non invasive alternative procedure for early detection in childhood asthma biomarkers in saliva. Vol. 27, Photodiagnosis and Photodynamic Therapy. Elsevier BV; **2019**. p. 85–91. Available from: http://dx.doi.org/10.1016/j.pdpdt.2019.05.009.

[68] Ralbovsky NM, Halámková L, Wall K, Anderson-Hanley C, Lednev IK. Screening for Alzheimer's Disease Using Saliva: A New Approach Based on Machine Learning and Raman Hyperspectroscopy. Vol. 71, Journal of Alzheimer's Disease. IOS Press; **2019**. p. 1351–9. Available from: http://dx.doi.org/10.3233/JAD-190675.

[69] QIU S, XU Y, HUANG L, ZHENG W, HUANG C, HUANG S, et al. Non-invasive detection of nasopharyngeal carcinoma using saliva surface-enhanced Raman spectroscopy. Vol. 11, Oncology Letters. Spandidos Publications; **2015**. p. 884–90. Available from: http://dx.doi.org/10.3892/ol.2015.3969.

[70] Li X. Spectral analysis of human saliva for detection of lung cancer using surface-enhanced Raman spectroscopy. Vol. 17, Journal of Biomedical Optics. SPIE-Intl Soc Optical Eng; **2012**. p. 037003. Available from: http://dx.doi.org/10.1117/1.jbo.17.3.037003.

[71] Brindha E, Rajasekaran R, Aruna P, Koteeswaran D, Ganesan S. Raman spectroscopy of bio fluids: an exploratory study for oral cancer detection. Alfano RR, Demos SG, editors. SPIE Proceedings. SPIE; **2016**. Available from: http://dx.doi.org/10.1117/12.2212684.

[72] Chen M, Chen Y, Wu S, Huang W, Lin J, Weng GX, et al. Raman spectroscopy of human saliva for acute myocardial infarction detection. Luo Q, Wang LV, Tuchin VV, editors. SPIE Proceedings. SPIE; **2014**. Available from: http://dx.doi.org/10.1117/12.2068876.

[73] Connolly JM, Davies K, Kazakeviciute A, Wheatley AM, Dockery P, Keogh I, et al. Non-invasive and label-free detection of oral squamous cell carcinoma using saliva surface-enhanced Raman spectroscopy and multivariate analysis. Vol. 12, Nanomedicine: Nanotechnology, Biology and Medicine. Elsevier BV; **2016**. p. 1593–601. Available from: http://dx.doi.org/10.1016/j.nano.2016.02.021.

[74] Schulze HG, Turner RFB. A Two-Dimensionally Coincident Second Difference Cosmic Ray Spike Removal Method for the Fully Automated Processing of Raman Spectra. Vol. 68, Applied Spectroscopy. SAGE Publications; **2014**. p. 185–91. Available from: http://dx.doi.org/10.1366/13-07216.

[75] Liu W, Sun Z, Chen J, Jing C. Raman Spectroscopy in Colorectal Cancer Diagnostics: Comparison of PCA-LDA and PLS-DA Models. Vol. 2016, Journal of Spectroscopy. Hindawi Limited; **2016**. p. 1–6. Available from: http://dx.doi.org/10.1155/2016/1603609.

[76] LIN J, LIN J, HUANG Z, LU P, WANG J, WANG X, et al. RAMAN SPECTROSCOPY OF HUMAN HEMOGLOBIN FOR DIABETES DETECTION. Vol. 07, Journal of Innovative Optical Health Sciences. World Scientific Pub Co Pte Lt; **2014**. p. 1350051. Available from: http://dx.doi.org/10.1142/S179354581350051X.

[77] Villa-Manríquez JF, Castro-Ramos J, Gutiérrez-Delgado F, Lopéz-Pacheco MA, Villanueva-Luna AE. Raman spectroscopy and PCA-SVM as a non-invasive diagnostic tool to identify and classify qualitatively glycated hemoglobin levels in vivo. Vol. 10, Journal of Biophotonics. Wiley; **2016**. p. 1074–9. Available from: http://dx.doi.org/10.1002/jbio.201600169.

[78] Li SX, Zeng QY, Li LF, Zhang YJ, Wan MM, Liu ZM, et al. Study of support vector machine and serum surface-enhanced Raman spectroscopy for noninvasive esophageal cancer detection. Vol. 18, Journal of Biomedical Optics. SPIE-Intl Soc Optical Eng; **2013**. p. 027008. Available from: http://dx.doi.org/10.1117/1.jbo.18.2.027008.

[79] Davenport TH, Barth P and Bean R. "How 'Big Data' Is Different." MIT Sloan Management Review 54, no. 1 (Fall **2012**).

[80] Banbury C, Mason R, Styles I, Eisenstein N, Clancy M, Belli A, et al. Development of the Self Optimising Kohonen Index Network (SKiNET) for Raman Spectroscopy Based Detection of Anatomical Eye Tissue. Vol. 9, Scientific Reports. Springer Science and Business Media LLC; **2019**. Available from: http://dx.doi.org/10.1038/s41598-019-47205-5.

[81] Acquarelli J, van Laarhoven T, Gerretzen J, Tran TN, Buydens LMC, Marchiori E. Convolutional neural networks for vibrational spectroscopic data analysis. Vol. 954, Analytica Chimica Acta. Elsevier BV; **2017**. p. 22–31. Available from: http://dx.doi.org/10.1016/j.aca.2016.12.010.

[82] Yang J, Xu J, Zhang X, Wu C, Lin T, Ying Y. Deep learning for vibrational spectral analysis: Recent progress and a practical guide. Vol. 1081, Analytica Chimica Acta. Elsevier BV; **2019**. p. 6–17. Available from: http://dx.doi.org/10.1016/j.aca.2019.06.012.

[83] Nguyen NT, Tojo S, Nguyen LM, Trawiński B, editors. Intelligent Information and Database Systems. Lecture Notes in Computer Science. Springer International Publishing; **2017**. Available from: http://dx.doi.org/10.1007/978-3-319-54430-4.

[84] Guo S, Rösch P, Popp J, Bocklitz T. Modified PCA and PLS: Towards a better classification in Raman spectroscopy–based biological applications. Vol. 34, Journal of Chemometrics. Wiley; **2020**. Available from: http://dx.doi.org/10.1002/cem.3202.

[85] Zonta ŽJ, Flotats X, Magrí A. Estimation of parameter uncertainty for an activated sludge model using Bayesian inference: a comparison with the frequentist method. Vol. 35, Environmental Technology. Informa UK Limited; **2014**. p. 1618–29. Available from: http://dx.doi.org/10.1080/09593330.2013.876450.

[86] Beams R, Gustavo Cançado L, Novotny L. Raman characterization of defects and dopants in graphene. Vol. 27, Journal of Physics: Condensed Matter. IOP Publishing; **2015**. p. 083002. Available from: http://dx.doi.org/10.1088/0953-8984/27/8/083002.

[87] Spectragryph, URL="https://www.effemm2.de/spectragryph/", accessed: 29/04/2022.

[88] Raman Tool Set, URL="https://sourceforge.net/projects/ramantoolset/", accessed: 29/04/2022.

[89] Spectra processing and analysis tools, URL="https://sourceforge.net/projects/spectools/", accessed: 29/04/2022.

[90] R project, URL="https://www.r-project.org", accessed: 29/04/2022.

CHAPTER 2

Experimental Tools and Analytical Techniques

This chapter describes the materials, and experimental and analytical techniques, used to make, prepare and characterise electrohydrodynamic surfaces for surface-enhanced Raman scattering (SERS). Additionally, computational tools to analyse, classify and simulate the data are also discussed, such as multivariate analysis and convolutional neural networks.

2.1 Electrohydrodynamic Patterning

2.1.1 Experimental Setup

The electrohydrodynamic (EHD) patterning method forms micro/nanostructures capable of generating surface-enhanced Raman structures when coated with gold, and thus it was an integral part of this work. EHD pattern formation has previously been achieved and here the same protocol is followed [1-3]. The EHD setup begins with the assembly of a capacitor, which is made up of a bottom conductive electrode made of commercially available silicon coated with a thin polymer film (100 nm) material to be patterned. Above is another conductive silicon electrode with a 1μm controllable air gap. A bias voltage (40V) is applied

between the two conductive electrodes to create a strong electric field (10^8V/m) due to the micrometric air gap. The whole setup is placed in an oven and heated above the material's glass transition temperature (180°C) softening the material until it is a viscous fluid. As time advances, the viscous fluid moves and redistributes to minimise the electrostatic forces between the electrodes (**Figure 2.1**). The process can be controlled by inspecting the current across the capacitor; when the current plateaus with time, the process is complete, and the entire device is cooled to room temperature and disassembled. After disassembly, the EHD generated structures are present on the bottom substrate.

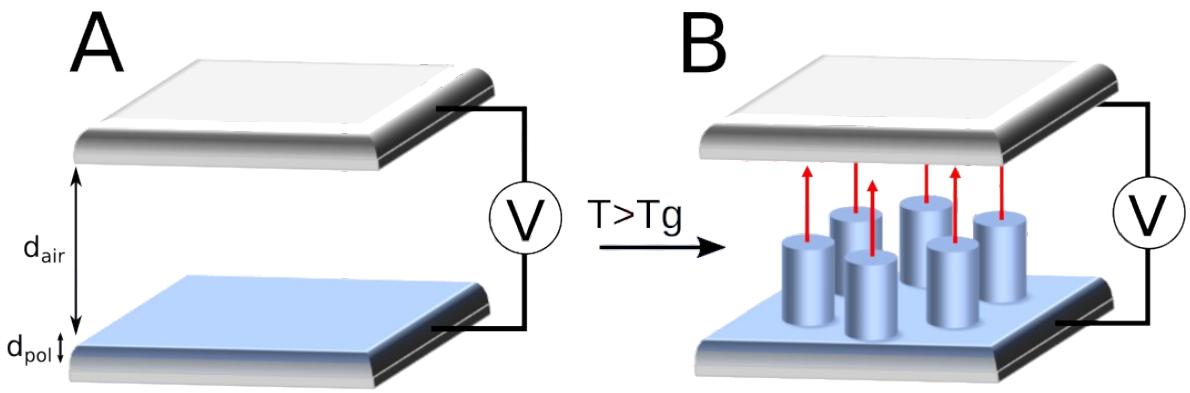


Figure 2.1: *EHD setup formation.* ***(A)*** *Bottom and top electrode assembled in a capacitor geometry with a polymer layer and an air gap. Applying a potential difference exerts electrostatic pressure that forces the polymer to rise at a high temperature.* ***(B)*** *Generation of pillars due to the electrostatic pressure and temperature.*

Below we give a detailed description of the materials, solvents and techniques used to produce and characterise the EHD structures.

2.1.2 Substrate Preparation

Single side polished silicon wafers were purchased from Si-Mat Silicon Materials with a diameter of 100mm and 525±25µm thickness. The wafers are n-type phosphorus-doped

with a 1-5Ωcm resistivity and surface roughness on the order of 1nm. These wafers were used as substrates for the EHD patterning. The silicon wafers were cleaved, using a diamond scribe and a metallic ruler, into 15×15mm^2 squares for the bottom electrode and 10×10mm^2 for the top electrode (**Figure 2.2E**). It is crucial to clean the silicon to give it an initially smooth, homogenous layer of the intended patterning material. A spray carbon dioxide (SCD) system was used for advanced cleaning, which was built in-house to remove the dust particles and other contaminations present on the silicon substrates (**Figure 2.2A**). The SCD set-up combined with an additional technique of heating the substrate at 240°C (**Figure 2.2B**) helps remove volatile hydrocarbon contamination and reduce humidity. This method of cleaning is called 'snow-jet cleaning' [1,2,4]. Following the snow-jet clean, an additional plasma cleaning procedure is carried over the silicon substrate (**Figure 2.2C**). The plasma tool used was a Harrick Plasma – PDC-32G-2 (**Figure 2.2D**) that function with a radio frequency oscillating electric field under a quasi-vacuum state. Combining these factors allows the air gas to ionise and thus generate a plasma. The plasma is accelerated onto the substrate thus removing adsorbed contaminations over the surface.[5]

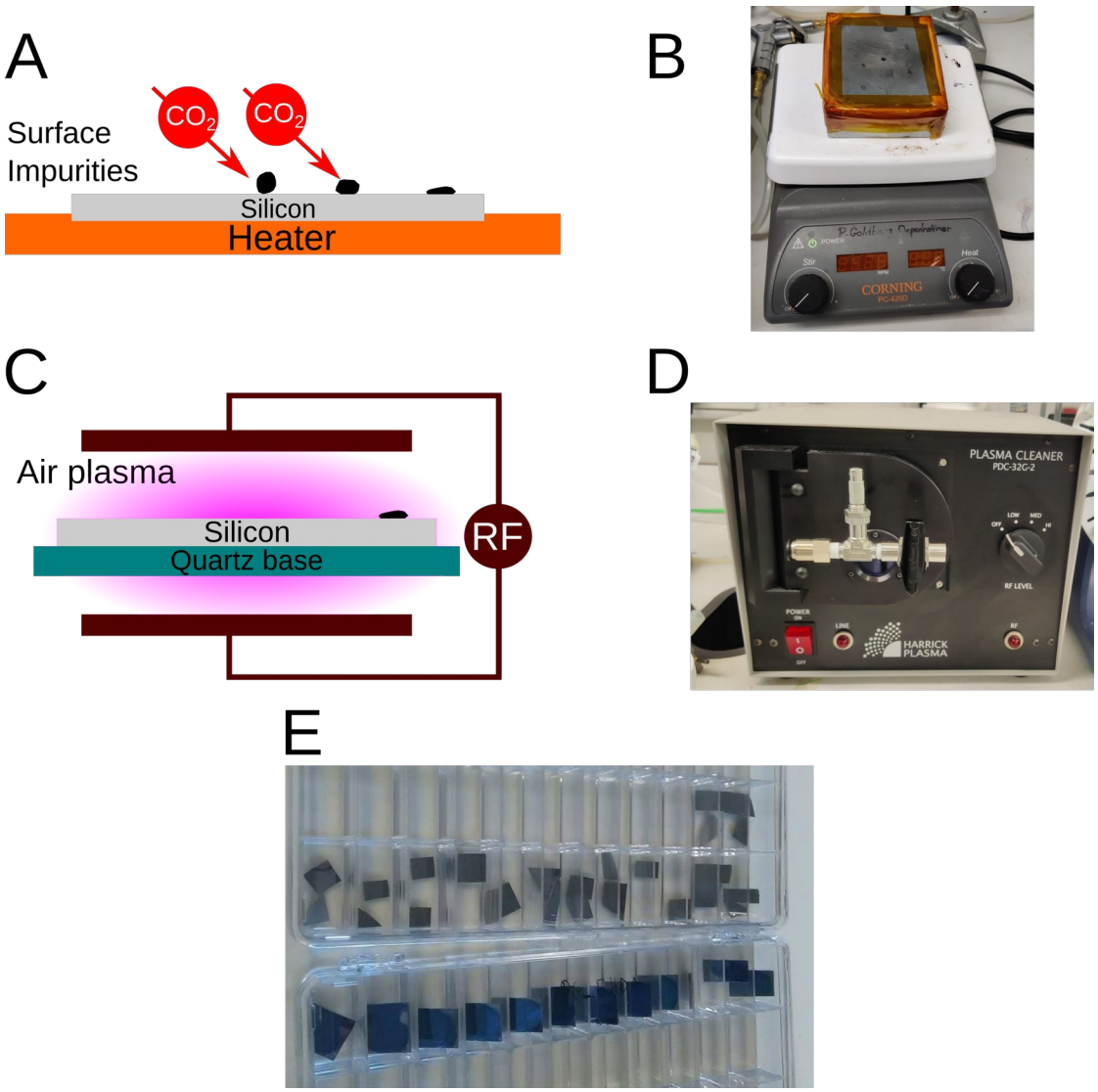


Figure 2.2: ***(A)*** *Schematic representation of the snow jet and respective* ***(B)*** *heating plate.* ***(C)*** *Schematic representation of the plasma cleaning process and respective* ***(D)*** *plasma cleaner tool.* ***(E)*** *Resulting clean silicon piece* ***(top)*** *ready to be sputtered coated* ***(bottom)****.*

2.1.3 Patterned Materials

To fabricate the EHD pillar structures, polystyrene with a molecular weight of 100,000g/mol, purchased from Sigma Aldrich, was used after being dissolved in toluene, also purchased from Sigma Aldrich.

2.1.4 Spin Coating Thin Films

Spin coating is a method of creating thin, homogeneous layers. For EHD lithography experiments, an ultra-thin, homogeneous polymer film with a maximum thickness of 200nm was required. The Laurel WS-650-23 (**Figure 2.3B**) spin coater instrument was used to create the required thin homogeneous films. A 75-mL solution of 3% polystyrene in toluene was deposited on the substrate and spun into a thin, homogeneous film, beginning at 3000 rpm and increasing the rate linearly until 6000 rpm, until the process was completed after 1 minute, yielding a thin polymer film substrate (**Figure 2.3A i-iii**).[6]

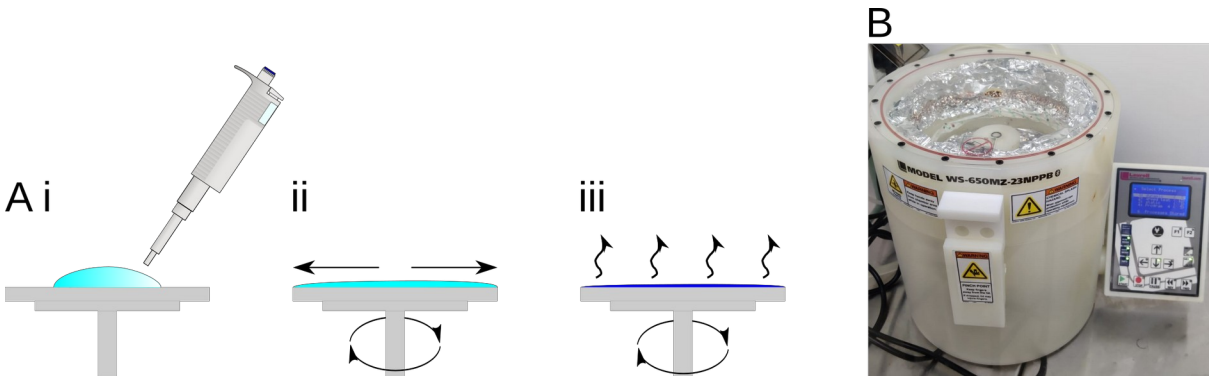


Figure 2.3: (A) i) *A solution of a solvent is pipetted onto the substrate.* ***ii)*** *The spin coater is then spun up, and the solution is forced to spread over the substrate.* ***iii)*** *The solvent evaporates as time passes leaving a thin homogenous layer.* ***(B)*** *Spin coater used for this work.*

The qualitative thickness of the spin-coated film was determined using an optical microscope (OM) and an atomic force microscope (AFM). Different colours visualised on an OM are related to different film heights in the thin film. The thin film height can be measured by manually performing a small indentation on the film and inspect the region on an AFM. The spin-coated film over the silicon substrate can now be patterned by using it as the bottom electrode on the EHD setup, allowing an EHD lithography process to be performed to generate the EHD patterns.

2.2 Thin Layer Gold Sputter Coating

It is not only essential to have micro/nano EHD structures but also a thin noble metal layer is required to turn the structures into SERS active structures. The noble metal choice was gold because of its compatible plasmonic enhancement for our pillar size, Raman laser wavelength requirements and its overall biocompatibility. Sputter coating used in this work functions by creating a plasma from input argon gas, and accelerating and confining the generated plasma ions into the gold target with an electric field and a magnetic field respectively. The impact of the ions allows gold atoms to be released from the target and accelerated through momentum transfer, and be deposited as a thin film on the substrate surface (**Figure 2.4A**).

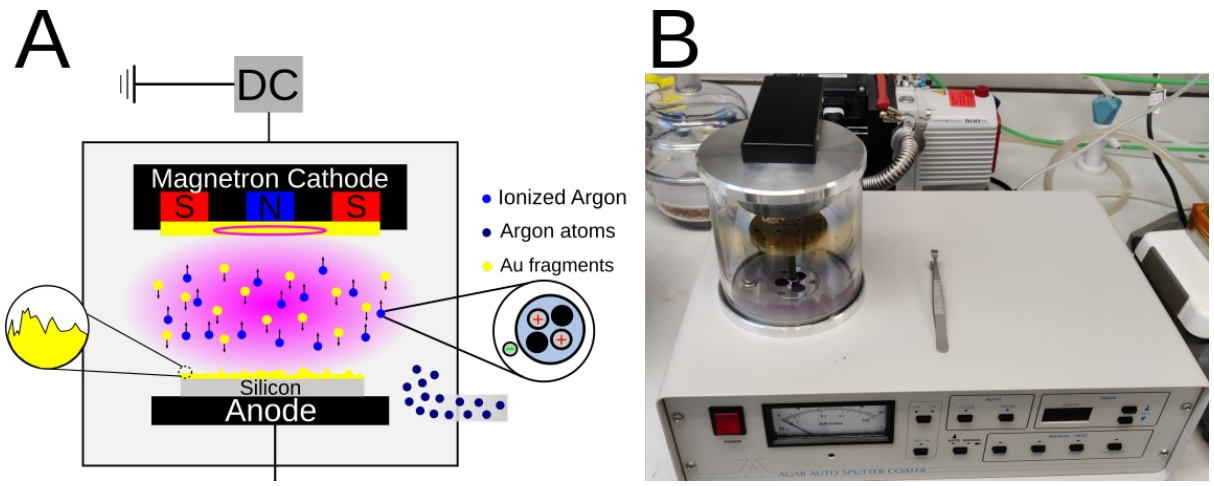


Figure 2.4: (A) *Automatic sputter coater schematic representation of the sputter coater. The figure shows a magnetron sputtering deposition.* ***(B)*** *The instrument used to sputter coat our EHD surfaces.*

The sputter coater AGB7341 by Agar Scientific[7] (**Figure 2.4B**) was employed to coat the EHD for SERS measurements. A range of gold thickness and roughness was obtained by varying the sputtering times and currents. The inner chamber argon gas pressure was kept constant at 0.08 mbar throughout all of the depositions. Testing multiple combinations with deposition times ranging from 20, 40, and 80 s and currents ranging

from 10, 20, 30, and 40 mA yielded a fit line between different current and time values in relation to gold thickness. The gold film surface was fine-tuned further by adjusting the sample radial distance on the sputter coater platform. By combining these parameters, it was possible to achieve a range of multiple grain sizes and roughness for the same thickness, which was then used to obtain the best SERS enhancement.

2.3 Characterisation Techniques

A range of imaging techniques were used to characterise the produced devices and substrates. Initially an optical microscopy was used however it does not resolve the smallest features produced in our experiments. Thus additionally, atomic force microscopy was used for a higher resolution and improved characterisation. For the analyte detection, Raman spectroscopy was used and when performed over a plasmon active surface, surface-enhanced Raman scattering was measured.

2.3.1 Optical Microscopy

Optical microscope images were an essential part of this work, first to assess the quality of the fabricated structures, and then as an essential part for acquiring images for the training of a convolutional neural network (CNN). Two different instruments were used for the image acquisitions, the Renishaw optical system (Leica DM 2500 – **Figure 2.5C**) and a research-grade microscope (Leica DM 2700M – **Figure 2.5A**). On both microscopes a set of 5x, 10x, 20x and 50x objectives were used to check the EHD sample but only the 50x was used for image and measurement acquisition. The Leica DM 2700M was equipped

with bright-field reflection and transmission capabilities. The Leica DM 2500M was equipped with bright-field reflection capabilities. The images were viewed and captured using an eye piece mounted CCD camera connected to a laptop. This setup allowed to observe the small sub-micron pillars produced and observe their colours, which indicate the thickness of the pillar structures as well as being able to image the pillar diameters.

A set of optical measurements were taken with both microscopes after EHD patterns were formed and after they were gold coated (**Figure 2.5B** and **Figure 2.5D**). All the pillar images were acquired as a top view image and saved for quantitative and qualitative analysis and for training a neural network model. Optical microscopy images were collected using the Leica Application Suite X (LAS X) [8] for the Leica DM 2700M and WiRE 5.3[9] for the Leica DM 2500.

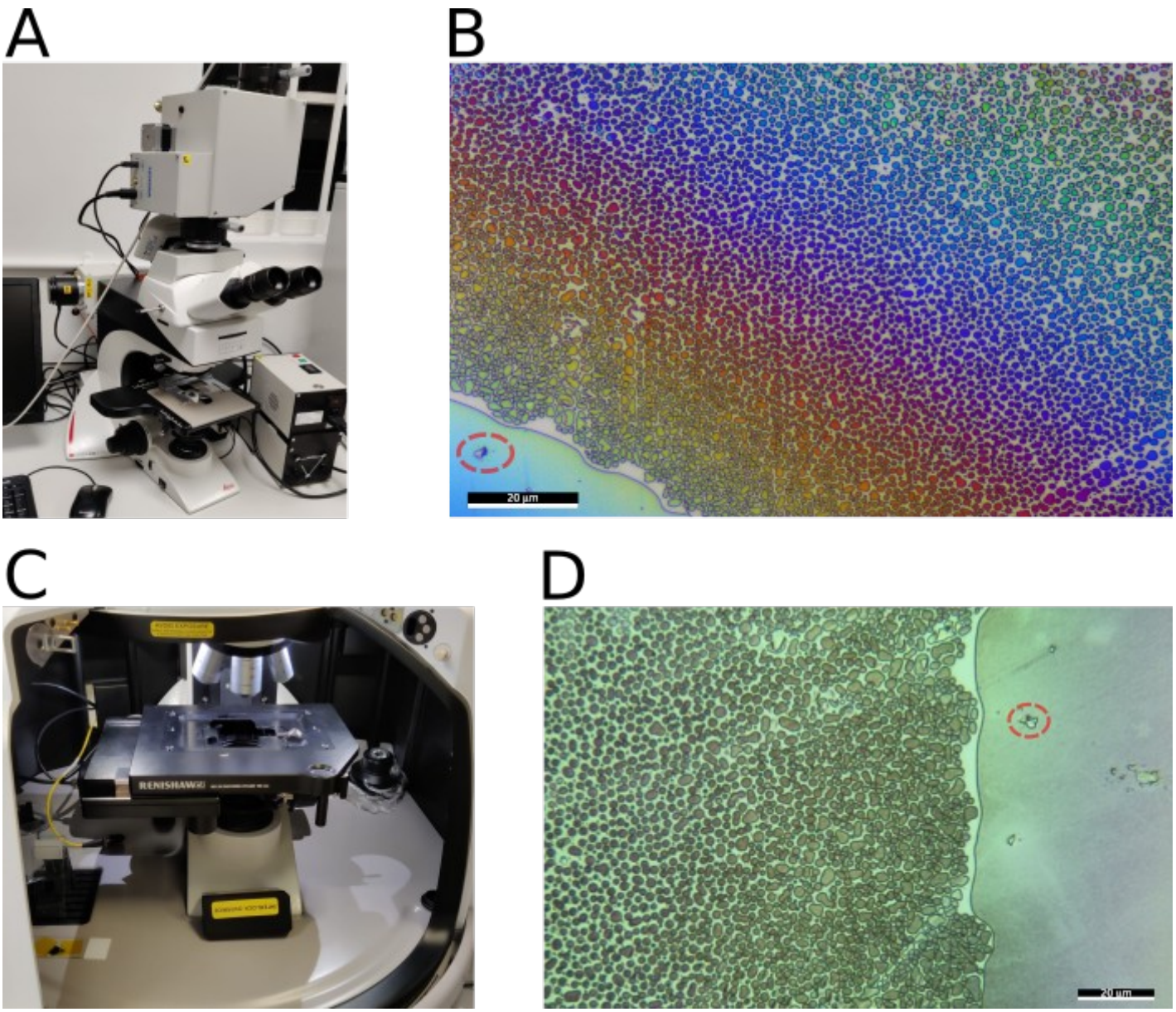


Figure 2.5: (A) *Leica DM 2700M,* ***(B)*** *EHD patterns from the Leica DM 2700M microscope (scale bar: 20 µm).* ***(C)*** *Leica DM 2500 present in the InVia Qontor system,* ***(D)*** *EHD patterns before gold coating from the Leica DM 2500 (scale bar: 20 µm). Both EHD patterns* ***(B)*** *and* ***(D)*** *are from the same spot, the highlighted dash circle to help identify the common patterns and to inspect the difference in quality.*

2.3.2 Atomic Force Microscopy

Atomic force microscopy (AFM) is an ideal tool to obtain precise information regarding surface topography and mechanical properties. Here we focus on the surface topography measurements. To operate the AFM there are two common operation modes, the contact mode and the tapping mode, which can be conducted in air or in liquid environments. Additionally, a cantilever needs to be installed onto the AFM system to be able to measure the fine surface details. When performing a scan across the surface in contact mode, the tip accompanies the sample's topography while maintaining a balance between the force experienced by the different surface profiles and the distance between the surface and the tip. Different topographical features can be inspected by monitoring the deflection of a laser beam off the back of the cantilever onto the photodiode detector sensing position. The laser reflection spot position provides the diode with a voltage input which is maintained constant by a feedback loop that updates the cantilever height if the balance is not verified. These height updates will recreate the surface profile, thus measuring the surface topography. In tapping mode, the probe only slightly contacts the surface thus this mode is particularly useful to prevent damage in soft, fragile or adhesive surfaces during scanning. Instead of controlling the cantilever's height for the feedback loop, here the amplitude of the cantilever's oscillation is used. A constant oscillation amplitude is maintained while scanning the surface, otherwise the amplitude is changed to accommodate for the surface topography and thus recording these movements relate to the topographical image (**Figure 2.6Bi**).

AFM measurements were performed using a NanoWizard II (**Figure 2.6A**) by JPK [10] microscope operated in the tapping mode and equipped with a NCHV-A probe (Thickness: 3.4 µm; length: 125µm; width: 40µm; Resonant frequency: 320kHz; spring

constant: 42Nm^{-1}). Image processing and analysis was carried out with the free software Gwyddion [11] (version 2.55). AFM measurements were used to acquire important experimental parameters including the initial film thickness, the lateral distances between the generated EHD pillar structures, their heights and diameters (i.e., their aspect ratio) and topographical imaging for the neural network.

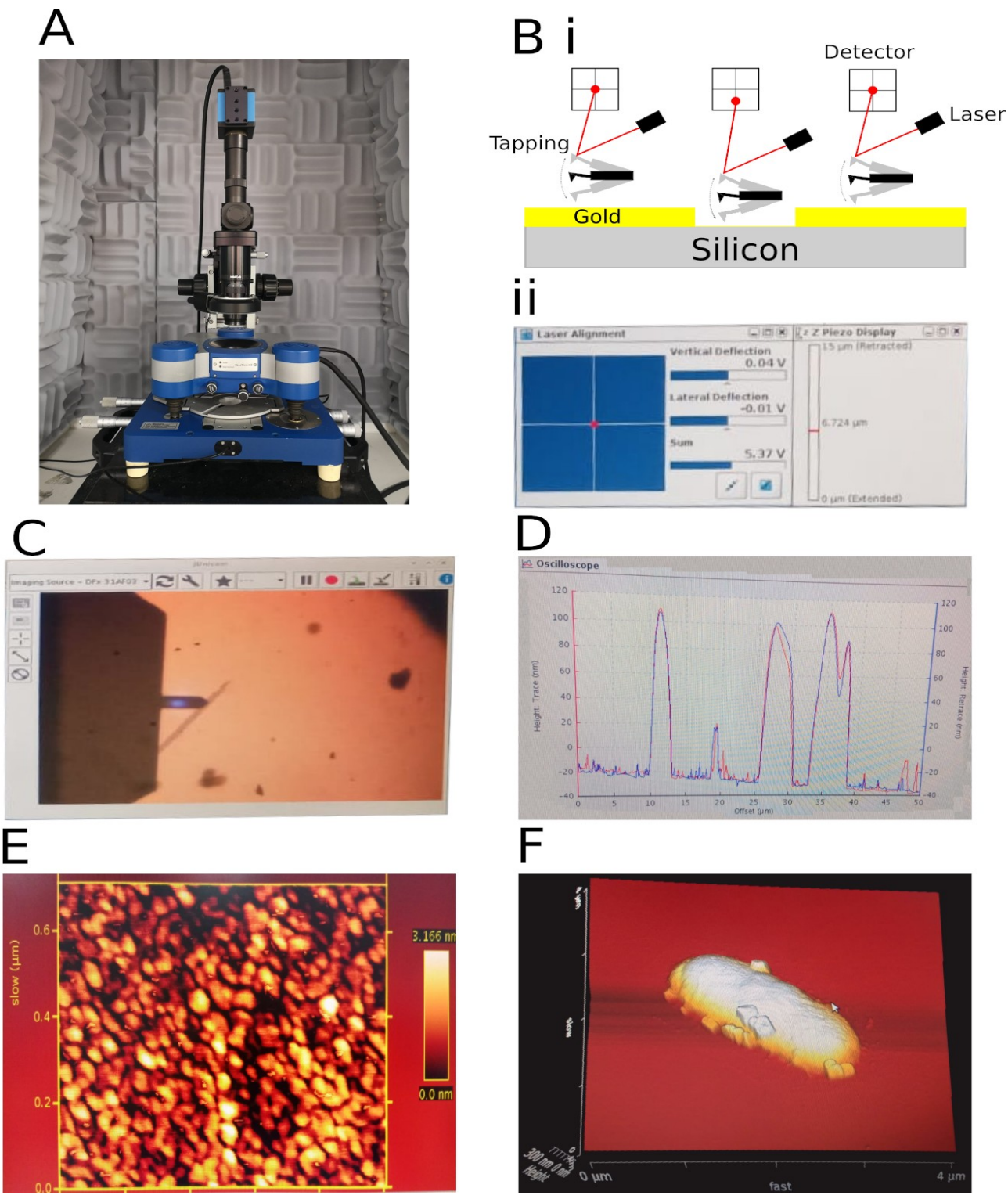


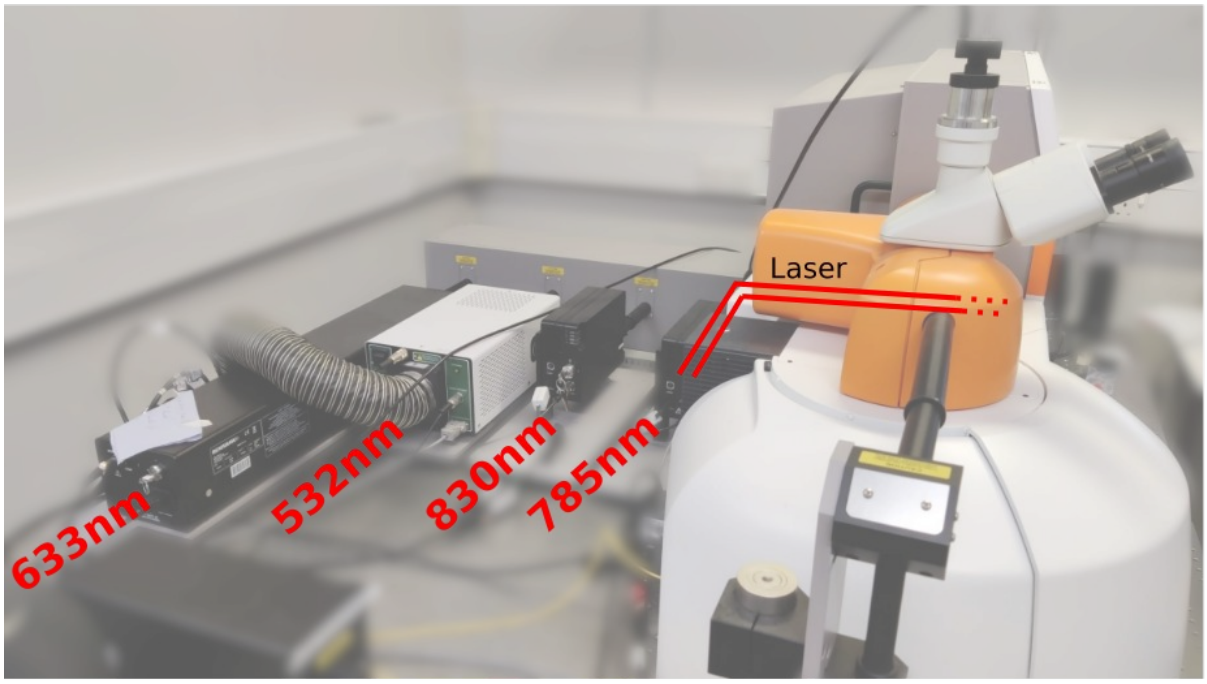
***Figure 2.6**. **(A)** NanoWizard II by JPK microscope. **(B) i)** Schematic representation of the tapping mode function while going though a gold structure while changing the laser reflection. **ii)** Laser alignment seen in the AFM software, maximise the laser sum and minimise the lateral and vertical deflection for optimal measurement. **(C)** Optical microscope image of the sample and cantilever, with a proper laser alignment (white spot). Measurement examples: **(D)** 1D line of a forming EHD structure, **(E)** 2D representation of a rough gold surface and **(F)** 3D of an E. coli bacteria (image obtained for [12]).*

2.3.3 (Surface-Enhanced) Raman Spectroscopy

Raman spectroscopy is a technique used to observe inelastic scattering interactions between monochromatic photons and chemical vibration modes. Hence, Raman spectroscopy is generally used in chemistry, material science and biology to identify molecules, compounds and pathogens by means of molecular vibrational fingerprints.[13] The Raman spectrometer (**Figure 2.7B**) consists of a monochromatic laser beam that is magnified by a microscope objective, scans a sample, and is back-scattered. The incident and back-scattered laser light are separated by a dichroic beam-splitter. Following the optical path, the remaining laser light is split: the out-of-focus light is removed and the in-focus-light passes through. Since Raman scattering events are very rare, the focused scattered light needs to go through a Rayleigh filter to suppress the Rayleigh scattering. Finally, a diffraction grating spatially disperses the different Raman bands before being measured on a CCD. The acquired signal is elaborated with a computer, which produces the final spectrum.

An *InVia* Qontor Raman spectrometer system by Renishaw equipped with a Leica confocal microscope DM2500 (50× objective with a numerical aperture of 0.75), 633nm RL633 Class 3B HeNe and 785nm lasers were used to acquire the Raman spectra of EHD, gold, EHD coated with gold, functionalised gold surfaces (thiophenol and methylene blue) and saliva. Raman analyses and data collection were performed using the dedicated software WiRE 5.1 [9].

A



B

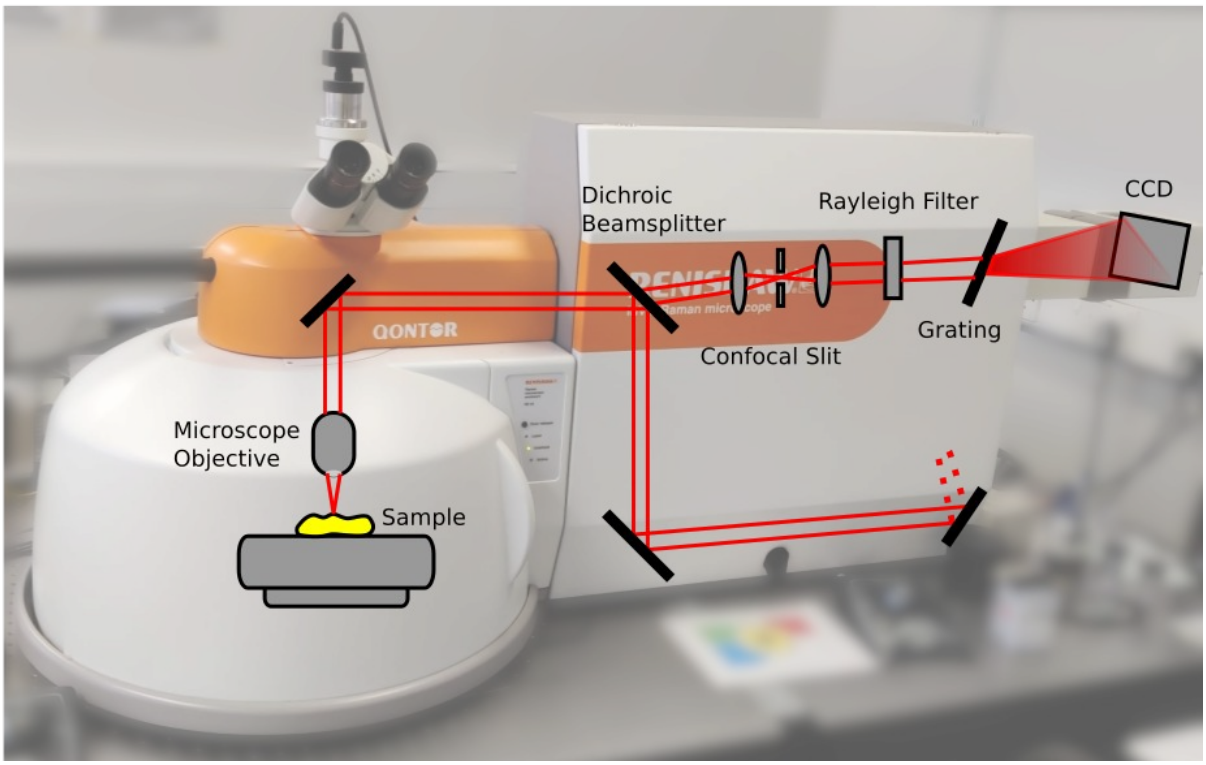


Figure 2.7: (A) *Back part of the InVia Qontor Raman Spectrometer where the lasers are located. From these the 785 nm and 633 nm were the ones used.* ***(B)*** *Schematic representation of the Raman spectrometer showing the main components of a Raman spectrometer. The beam splitter splits the beam the objective aiming at the sample. Subsequently, the scattered radiation goes through a set of mirrors, lenses, filters and gratings that cleans the scattered radiation into the Raman scattering which is resolved in the CCD and inspected through the WiRE software.*

2.4 Python

Throughout this work, Python has been the main tool used for all the data processing and machine learning (ML) work. Python [14] has gained a huge following in recent years due to its easy-to-learn language and large community online with multiple open-source code available that helps new users. Python is used for a variety of purposes, one of which being that it is the finest choice for practical neural networks. Analyzing the major elements of the multivariate data analysis techniques used in this work, including principal component analysis (PCA), self-organizing maps (SOM), and neural networks like convolutional neural networks (CNN). As an extra data analysis tool, especially for data visualization, Matplotlib package [15] was employed throughout the entire project.

2.4.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) [16] is a dimensionality reduction operation that projects the data into a lower-dimensional space. PCA is the main multivariable analysis tool to pre-process and analyse Raman data. PCA can be employed using the scikit-learn library, specifically sklearn.decomposition. PCA uses the normalised Raman spectra map for inputs and the number of PCA components to analyse. The number of components is related to how the Raman data will be transformed into the PCA representation. Hence, the number of components is related to the dimensions of the new representation. We used only two components for the 2D plot and three for the 3D plot in the script. After the component selection, the input normalised Raman spectra are 'fit_transform' onto the PCA by applying the PCA.fit_transform() function, which transforms the normalised spectra into the PCA space with the principal components (PC) PC 1 and PC 2 or PC1, PC2 and PC3

for the 3D case. A scattered plot with the PC1 and PC2 components is plotted to inspect the different Raman spectra in the PC space. Information regarding the different PC variances can be obtained by the PCA.explained_variace_ratio_, which outlines the percentage variance contribution in spectra separation for the different PC. This measure highlights how well the PC separates the different Raman spectra data.

2.4.2 Self-Organizing Maps (SOM)

The self-organizing maps (SOM) implementation is set up using the miniSOM[17] module, which handle the entire SOM implementation and mathematics. In context o this work, the SOM algorithm has been implemented to have only two input parameters, the Raman data to analyse and the size of the SOM grid map. A hexagon represents each spot on the grid map with inherent neural network weights, and biases randomly initialized at the start of the SOM calculation. Each hexagon's weight and bias represent a spectrum-like signature called the neuron activations or simply activations and is crucial to the SOM analysis. The different Raman spectra are represented by dots placed on the SOM grid map. The placement of the dots is determined by the spectrum-like weights and biases of the hexagon. The som.win_map() function can access each neuron activation. Additional information can be observed regarding the highest population sample per hexagon by the colouring of the hexagon edge. Finally, different greyscale intensities inside each hexagon give the boundary separations between adjacent hexagons; the darker the hexagon, the stronger the separation, whereas the lighter the hexagon, the weaker the separation.

2.4.3 Convolutional Neural Network (CNN)

For this work, the Python libraries used were Keras[18] and Tensorflow[19]. Keras is a user-friendly library that provides units to produce deep learning algorithms. This program runs with a Tensorflow backend, developed by Google. Both provide layers that can be used within CNNs and other deep learning tasks. These transform the data to allow the computer to extract and understand information from them.

From Keras, we imported a pre-trained neural network called the VGG-16 to train it on our EHD pillar images by following a similar procedure described in [20]. A pre-trained neural network is a saved network that has already been trained on previous data, thus the weights and biases of each interconnected neuron are already fixed. We write, "from keras.applications import VGG16" [21] to import the VGG-16 like VGG(weights=" ", include_top="""). For the "weights", we select the already trained neural network, in this case, the "imagenet" [22], which is an image database organized in hierarchies (only nouns) in which hundreds and thousands of images were used to train the VGG16 neural network. These weights are very robust and can be used for multiple applications. After importing the weights, we need to modify the output layer to adapt it to our SERS/non-SERS case. Hence, we need to exclude the "include_top" variable by adding "False" to the VGG function. Overall, the pre-trained neural network has 14,714,688 trained parameters that are all imported with the calling of the VGG function. By removing the VGG16 classes output, we need to manually define our densely connected classifier, which can be done by adding a Dense layer with two output nodes (SERS, non-SERS) "model.add(layers.Dense(2, activation='sigmoid'))" (**Figure 2.8BA**). Sigmoid here is an activation function that, depending on the input, the output will be activated (1) or turned

off (0), similar to how transistors pass current or not depending on the input gate voltage, thus predicting to which class the input image belongs to.

Additionally, when training our data with the pretrained neural networks, the parameters imported will stay fixed, hence we need to add an extra function to train only some parameters of our pre-trained model. This process is called fine-tuning. Here, only the bottom layers of the VGG16 will have their parameters ready to be trained for the new data. The process of tuning the layers to be able to learn is called "unfreezing" the layers. The overall fine-tune process gets its name because only the bottom layers will be trained, and the bottom layers are responsible for the abstract representation of the images; thus, training these will make the VGG16 "fine-tune" to the new dataset (**Figure 2.8B**).

On the other hand, earlier layers do not need to be trained since they are more responsible for encoding more generic, reusable features, which are mainly common in all objects, for instance, the edges of any shapes like circles, triangles, or squares. The fine-tune will be performed with a very low learning rate for the bottom layers to make the neuron weights more resistant to change, thus fine-tuning them.

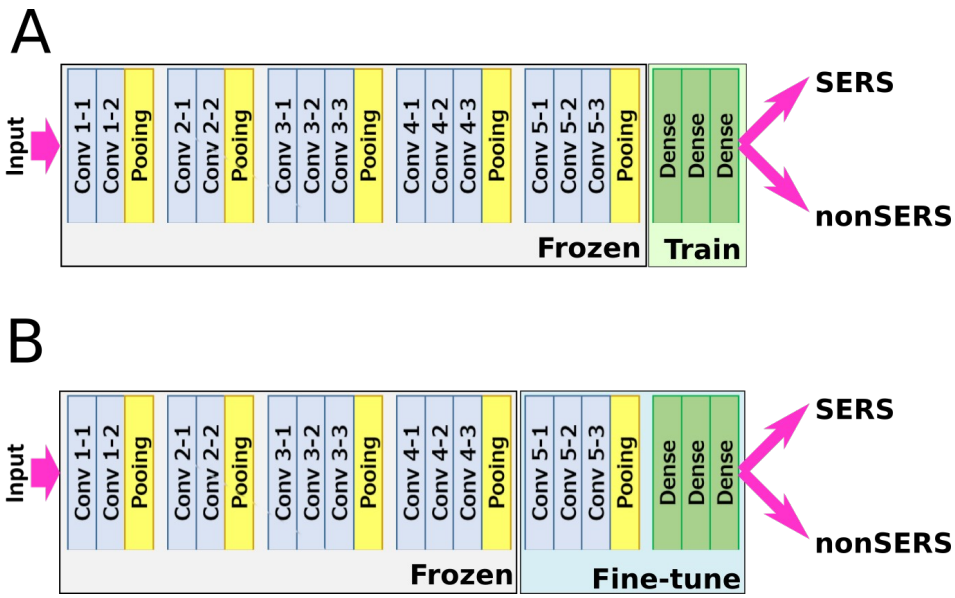


Figure 2.8: (A) *VGG16 architecture adapted for the new output layer SERS/non-SERS* ***(B)*** *VGG16 architecture for the fine-tuning of the bottom layers. The complete training of our CNN goes first through* ***(A)*** *and then* ***(B)****.*

Our dataset passes through an initial frozen VGG16 model to train the added Dense layers for our specific SERS/non-SERS dataset and then train the same VGG16 model with the bottom layers unfrozen, hence generating a very specific model for our dataset.

2.5 Other Software

All figures were composed using Inkscape 1.1.2 [23] in combination with GNU Image Manipulation Program (GIMP) 2.10.28 [24]. References were organised with Mendeley [25]. Document was done in Libre Office [26].

2.6 References

[1] Oppenheimer P. G., “Electrohydrodynamic Patterning of Functional Materials,” no. July, pp. 31–47, **2011**. PhD thesis, University of Cambridge, Cambridge, UK. Available: https://www.repository.cam.ac.uk/bitstream/handle/1810/245301/OppenheimerThesis.pdf

[2] Rickard J. J. S., “Advanced Micro- Engineered Platforms for Novel Device Technologies,” no. December, pp. 32–61, **2017**. PhD thesis, University of Birmingham, Birmingham, UK. Available: https://etheses.bham.ac.uk/id/eprint/8303/.

[3] Schaffer E., “Instabilities in Thin Polymer Films : Structure Formation and Pattern Transfer Instabilities in Thin Polymer Films: Structure Formation and Pattern Transfer,” no. September, pp. 23-30, **2001**. PhD thesis, University of Konstanz, Baden-Wurttemberg, DE. Available: https://d-nb.info/964088665/34.

[4] Sherman R, Hirt D, Vane R. Surface cleaning with the carbon dioxide snow jet. Vol. 12, Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films. American Vacuum Society; **1994**. p. 1876–81. Available from: http://dx.doi.org/10.1116/1.579021.

[5] Habib SB, Gonzalez E II, Hicks RF. Atmospheric oxygen plasma activation of silicon (100) surfaces. Vol. 28, Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films. American Vacuum Society; **2010**. p. 476–85. Available from: http://dx.doi.org/10.1116/1.3374738.

[6] Hall, D. B.; Underhill, P.; Torkelson, J. M., “Spin coating of thin and ultrathin polymer films.” Polymer Engineering & Science, 38, **1998**, p. 2039–2045.

[7] Sputter coating machine manual URL=”https://www.agarscientific.com/media/import/AGB7341-AGB7341SE-tech-data-july21.pdf”, accessed: 29/04/2022.

[8] Leica suite URL=”https://www.leica-microsystems.com/products/microscope-software/p/leica-las-x-ls/”, accessed: 29/04/2022.

[9] Renishaw WiRE URL=”https://www.renishaw.com/en/raman-software—9450”, access: 29/04/2022.

[10] JPK website URL=”https://www.jpk.com/app-technotes/products_atomic-force-microscopy”, access: 29/04/2022.

[11] Gwyddion URL=”http://gwyddion.net/”, accessed: 29/04/2022.

[12] Chourmouziadi Laleni N, Gomes PDC, Gkatzionis K, Spyropoulos F. “Propolis particles incorporated in aqueous formulations with enhanced antibacterial performance”. Food Hydrocolloids for Health. Elsevier BV; 1, **2021**, p. 100040. Available from: http://dx.doi.org/10.1016/j.fhfh.2021.100040.

[13] Renishaw SERS URL=”https://www.renishaw.com/en/sers-ters—25811”, access: 29/04/2022.

[14] Python URL=”https://www.python.org/”, access: 29/04/2022.

[15] Matplotlib URL=”https://matplotlib.org/”, access: 29/04/2022.

[16] Scikit-learn, PCA URL = ”https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html”, access: 29/04/2022.

[17] V. Giuseppe, “MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map,” **2018**, URL=”https://github.com/JustGlowing/minisom/”

[18] Convolutional Neural Network – Keras, URL=”https://keras.io/”, access: 29/04/2022.

[19] Convolutional Neural Network – Tensorflow, URL=”https://www.tensorflow.org/”, access: 29/04/2022.

[20] Chollet F., “Deep Learning with Python. Manning,” 2017. Available: https://books.google.es/books?id=wzozEAAAQBAJ

[21] VGG 16 keras, URL=“https://keras.io/api/applications/vgg/”, access: 29/04/2022.

[22] ImageNet, URL=”https://www.image-net.org/”, access: 29/04/2022.

[23] Inkscape, URL=”https://inkscape.org/”, access: 29/04/2022.

[24] GIMP, URL=”https://www.gimp.org/”, access: 29/04/2022.

[25] Mendeley, URL=”https://www.mendeley.com/”, access: 29/04/2022.

[26] LibreOffice, URL=”https://www.libreoffice.org/”, access: 29/04/2022.

CHAPTER 3

Electrohydrodynamic Lithography: Theoretical Background, Simulations and Height estimations

3.1 Abstract

The theoretical background of the electrohydrodynamic lithography physics using the Navier-Stokes equation is discussed in this chapter and expanded into five different small sections that function to understand better specific experimental conditions identified with the EHD lithography. The initial mathematical procedures to solve the Navier-Stokes equation and derive the dispersion relation have been made previously in [1-6], here these same equations were revisited and expanded for three situations:

1) Simulation of the dispersion relation for the parallel EHD configuration,

2) Solve the partial derivate equation of the simplified Navier-Stokes equation to simulate the EHD growth,

3) Develop the EHD framework for the tilted geometry capacitor with the respective simulation of the dispersion relation,

Additionally, two formulations related to how the EHD is analysed using an optical microscope were created to facilitate the estimate of the EHD pillar structure:

1) Formulate a colour-to-thickness converted for EHD optical image,

2) Identify an EHD mass conservation law that allows estimating the pillar heights from an optical microscope image.

These extensions to the EHD theoretical background and EHD applications have been the main addition to previous works and thus are part of this manuscript.

3.2 Electrohydrodynamic Theoretical Background

3.2.1 Parallel Capacitor: Electrohydrodynamic Dispersion Relation

A typical experimental setup is shown schematically in **Figure 3.1**. The z–coordinate of the film's interface is given by the height $h=h(x,t)$ where *x* is the lateral coordinate and *t* is the time variable.

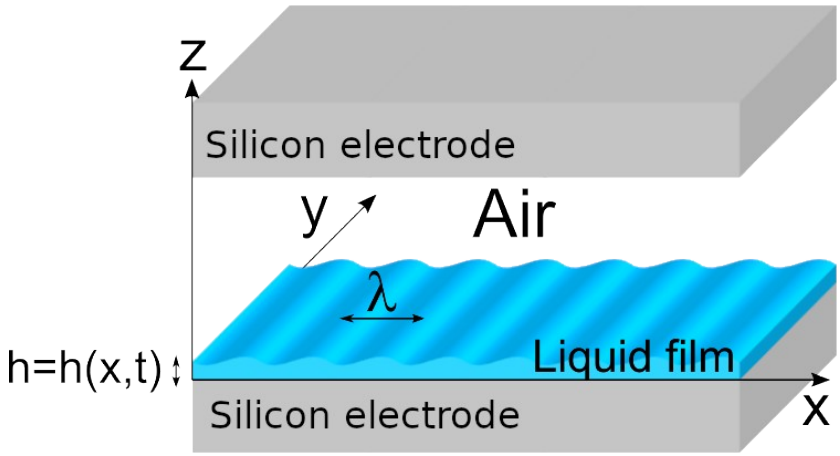


Figure 3.1: *Schematic diagram of a thin polystyrene film with a thickness h on a silicon substrate with an air gap and a top silicon electrode. Outlining the undulations, with a wavelength λ, caused by the polymer Brownian motion.*

The EHD analysis assumes multiple initial conditions.

- A thin polymer layer is spin-coated on to a silicon substrate and then placed in a capacitor setup, by placing another silicon substrate above the first coated substrate separated by an air gap.
- A Brownian motion exists on the surface polymer molecules, which have a low amplitude resonant wave at the surface.
- Surface tension, γ, is the dominant force opposing the surface fluctuations, attempting to reduce the overall surface area.

- an electric field E_P is applied across the two electrodes generating a high electrostatic pressure

Our goal is to simplify the Navier-Stokes equation to formulate the dispersion relation. For that, we can simplify the system as follows:

1) The height of the polymer is only dependent on one coordinate since the initial undulations are very small.

2) Assume the height variation is a constant plus a sinusoidal undulation with a wave number *q,* amplitude ζ and frequency $1/\tau$ in **Equation 3.1**

$$h(x,t)=h_0+\zeta e^{iqx+t/\tau} \tag{3.1}$$

3) Considering the Navier-Stokes equation for an incompressible Newtonian fluid[7]:

$$\rho\frac{\partial v}{\partial t}+\rho(v\cdot\nabla)v=-\nabla p+\eta\nabla^2 v+\rho g \tag{3.2}$$

With ρ the fluid density, v the fluid velocity, p the total pressure of the system and η the viscosity

Expanding the terms in **Equation 3.2** with respect to *x*.

$$\rho\left(\frac{\partial v}{\partial t}\right)+\rho\left[u\left(\frac{\partial v}{\partial x}\right)+v\left(\frac{\partial v}{\partial y}\right)+w\left(\frac{\partial v}{\partial z}\right)\right]=-\frac{\partial p}{\partial x}+\eta\left(\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}+\frac{\partial^2 v}{\partial z^2}\right)+\rho g \tag{3.3}$$

4) High viscosity results in a low flow velocity, thus the quadratic term in the Navier-Stokes equation can be neglected: $u\left(\frac{\partial v}{\partial x}\right)+v\left(\frac{\partial v}{\partial y}\right)+w\left(\frac{\partial v}{\partial z}\right)\rightarrow 0$

5) With slow dynamics, the velocity profile can be considered as a quasi-steady-state: $\rho\left(\frac{\partial v}{\partial t}\right)\rightarrow 0$

6) For thin-films with dimensions smaller than the capillary constant, gravity does not have an influence on the shape of the interface: $\rho g\rightarrow 0$

7) Considering that the film oscillations λ (**Figure 3.1**) are much larger than the film thickness h_0 and that the film thickness is much larger than the film fluctuations amplitude ζ (**Equation 3.1**)

Simplifying the Navier-Stokes equation into the following two-dimensional equation in the *z* and *x* coordinate:

$$\frac{\partial p}{\partial z}=0 \tag{3.4.1}$$

$$\eta\left(\frac{\partial^2 v}{\partial z^2}\right)=\frac{1}{\eta}\frac{\partial p}{\partial x} \tag{3.4.2}$$

From **Equation 3.4.1**, the pressure is constant along the z axis and it is only a function of x. Integrating **Equation 3.4.2** with respect to z,

$$v(z)=\frac{1}{2\eta}\left(\frac{\partial p}{\partial x}\right)z^2+c_1 z+c_2 \tag{3.5}$$

8) The velocity profile can be calculated by solving the boundary conditions using the *lubrication approximation*[8]: $v(z=0)=0\rightarrow c_2=0$

9) and no lateral stresses at the film surface: $v(z=h)\sigma_{xz}=\eta\frac{\partial v}{\partial z}=0\rightarrow c_1=2\frac{\partial p}{\partial x}h.$

Obtaining

$$v(z)=\frac{1}{2\eta}\left(\frac{\partial p}{\partial x}\right)z(z-2h) \tag{3.6}$$

10) Using mass conservation, any change in film height yields a flow in the lateral direction:

$$\frac{\partial V}{\partial t}=\frac{\partial A}{\partial t}dx=-\frac{\partial(A\bar{v})}{\partial x}dx \rightarrow \frac{\partial A}{\partial t}=-\frac{\partial j}{\partial x} \rightarrow \frac{\partial h}{\partial t}=-\frac{\partial(h\bar{v})}{\partial x} \tag{3.7}$$

The minus sign is related to the direction of flow. The flow is written as $j=A\bar{v}$ and $A=h$ in a one-dimensional case.

The average velocity in the film is the integration in z though the film height.

$$\bar{v}=\frac{1}{h}\int_0^h v(z)dz=\frac{h^2}{3\eta}\left(-\frac{\partial p}{\partial x}\right) \tag{3.8}$$

Finally, substituting **Equation 3.8** in **Equation 3.7**, yields the equation of motion for the film surface.

$$\frac{\partial h}{\partial t}=\frac{\partial}{\partial x}\left[\frac{h^3}{3\eta}\frac{\partial p}{\partial x}\right] \tag{3.9}$$

There are multiple pressure forces acting over the polymer film in the rest state.

$$p(h)=p_0+p_{vdW}(h)+p_L(h)+p_{ex}(h)+p_{dis}(h) \tag{3.10}$$

The ambient air pressure p_0, the van der Waals (vdW) intermolecular forces p_{vdW}, the Laplace pressure p_L, any excess pressures p_{ex}, which in this case is the electrostatic

pressure p_{el} and finally the disjoining pressure p_{dis} coming from the attractive force between two surfaces[9].

However, for the electrohydrodynamic setup, the dominant pressure is the electrostatic pressure caused by the differential potential applied between the two capacitors and the Laplacian pressure. Thus, the previous equation can be simplified into

$$p(h) = p_L(h) + p_{el}(h) \tag{3.11}$$

The electrostatic pressure can be calculated using the free energy stored in the capacitor equation,

$$F = F_0 - \frac{1}{2} C U^2 \tag{3.12}$$

With *C* being the capacitance and *U* the potential difference applied. The electrostatic pressure is given by:

$$p_{el} = \frac{1}{A}\frac{\partial F}{\partial h} = \frac{-U^2}{2A}\frac{\partial C}{\partial h} \tag{3.13}$$

The polymer-air double layer can be visualized as a series capacitor systems $\varepsilon_{air} \approx 1$

From **Figure 3.2**, the parallel capacitance setup is equivalent to having two capacitors in parallel. Thus,

$$\frac{1}{C} = \frac{1}{C_{dielectric}} + \frac{1}{C_{air}} = \frac{h}{\varepsilon_0 \varepsilon_p A} + \frac{d-h}{\varepsilon_0 \varepsilon_{air} A} = \frac{\varepsilon_p d - (\varepsilon_p - 1)h}{\varepsilon_0 \varepsilon_p A} \tag{3.14}$$

Reformulating with the inverse, and perform the derivative with respect to *h.*

$$C = \frac{\varepsilon_0 \varepsilon_p A}{\varepsilon_p d - (\varepsilon_p - 1) h} \tag{3.15.1}$$

$$\frac{\partial C}{\partial h} = - \frac{A \varepsilon_0 \varepsilon_p (\varepsilon_p - 1)}{\left[\varepsilon_p d - (\varepsilon_p - 1) h\right]^2} \tag{3.15.2}$$

Substituting **Equation 3.15.2** into **Equation 3.13**, the destabilizing electrostatic pressure equation is obtained[10] following the leaky dielectric model[11].

$$p_{el} = - \frac{U^2}{2A} \frac{\partial C}{\partial h} = - \frac{1}{2} \frac{\varepsilon_0 \varepsilon_p (\varepsilon_p - 1) U^2}{\left[\varepsilon_p d - (\varepsilon_p - 1) h\right]^2} \tag{3.16}$$

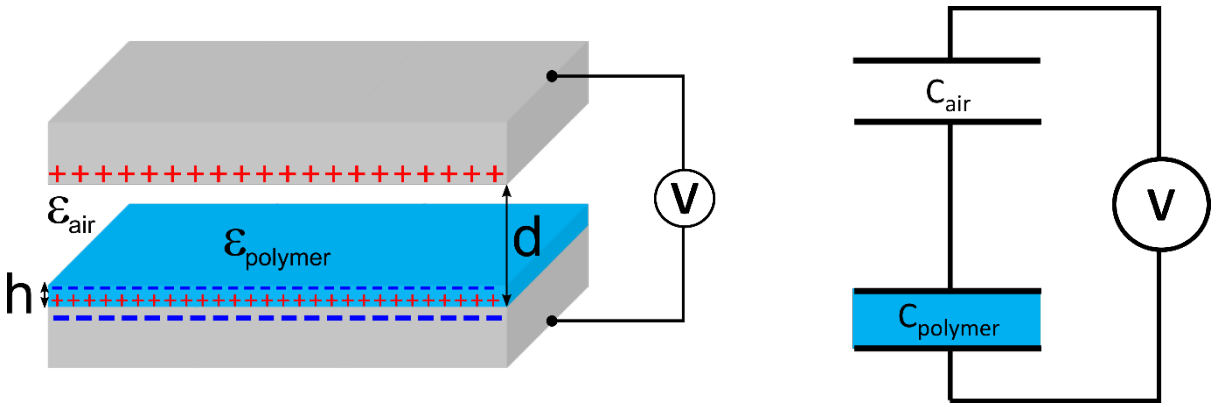


***Figure 3.2**: An applied electric field leads to the polarization of the dielectric polymer layer, yielding an attractive interaction between the charges at the polymer-air interface with the charges at the upper electrode. The layered system is equivalent to a sum of two capacitors in series.*

The Laplace pressure is given by the surface tension and the Laplacian to the position

$$p_L(h) = -\gamma \frac{\partial^2 h}{\partial x^2} \tag{3.17}$$

Combining **Equation 3.17**, **Equation 3.13** and **Equation 3.11**, leads to the total interfacial pressure combining the destabilizing electrostatic pressure and the stabilizing Laplace pressure.

$$p(h) = -\gamma \frac{\partial^2 h}{\partial x^2} - \frac{1}{2} \frac{\varepsilon_0 \varepsilon_p (\varepsilon_p - 1) U^2}{\left[\varepsilon_p d - (\varepsilon_p - 1) h\right]^2} \tag{3.18}$$

Which is described by the *linear stability analysis* of a free liquid film[2]. Substituting **Equation 3.18** in **Equation 3.9** and applying the chain rule.

$$\frac{\partial h}{\partial t} = \frac{h^2}{\eta} \frac{\partial h}{\partial x} \frac{\partial p}{\partial x} + \frac{h^3}{3\eta} \frac{\partial^2 p}{\partial x^2} = \frac{h^2}{\eta} \frac{\partial h}{\partial x} \left[-\gamma \frac{\partial^3 h}{\partial x^3} + \frac{\partial p_{el}}{\partial h} \frac{\partial h}{\partial x} \right] + \frac{h^3}{3\eta} \left[-\gamma \frac{\partial^4 h}{\partial x^4} + \frac{\partial^2 p_{el}}{\partial h^2} + \frac{\partial p_{el}}{\partial h} \frac{\partial^2 h}{\partial x^2} \right] \tag{3.19}$$

Finally, obtaining the dispersion relation by removing terms above $O(\zeta^2)$

$$\frac{\partial h}{\partial t} = \frac{h^3}{3\eta} \left[-\gamma \frac{\partial^4 h}{\partial x^4} + \frac{\partial p_{el}}{\partial h} \frac{\partial^2 h}{\partial x^2} \right] + O(\zeta^2) \tag{3.20}$$

Obtaining the dispersion relation by substituting **Equation 3.1** into **Equation 3.20**.

$$\frac{1}{\tau} = \frac{-h^3}{3\eta} \left[\gamma q^4 + \frac{\partial p_{el}}{\partial h} q^2 \right] \tag{3.21}$$

When $\tau \to \infty$, the right hand side of **Equation 3.21** can be rewritten as:

$$\left[\gamma q^4 + \frac{\partial p_{el}}{\partial h} q^2 \right] = 0 \Rightarrow \text{with } q^2 = x \Rightarrow \left[\gamma x^2 + \frac{\partial p_{el}}{\partial h} x \right] = 0 \tag{3.22}$$

Solving **Equation 3.22** for *x* and ignoring the trivial solution *x=0* results in $x = -\frac{1}{\gamma} \frac{\partial p_{el}}{\partial h}$.

Substituting back into *q* **Equation 3.22**, yields two possible solutions $0 < q = \sqrt{-\frac{1}{\gamma} \frac{\partial p_{el}}{\partial h}}$. *q* is a

real number if $\frac{\partial p_{el}}{\partial h}<0$ which would translate to an amplification of the film growth rate to an unstable regime. Otherwise, *q* is a complex number if $\frac{\partial p_{el}}{\partial h}>0$ which would equate to a damped solution thus making the film growth stable.

For the amplified case, the fastest growing mode is given by

$$\frac{d}{dx}\left[\gamma x^2+\frac{\partial p_{el}}{\partial h}x\right]=0\Rightarrow x_{max}=\frac{1}{2\gamma}\left|\frac{\partial p_{el}}{\partial h}\right|\Rightarrow q_{max}=\sqrt{\frac{1}{2\gamma}\left|\frac{\partial p_{el}}{\partial h}\right|} \tag{3.23}$$

Performing the derivative of the electrostatic pressure **Equation 3.16** with respect to *h*,

$$\frac{\partial p_{el}}{\partial h}=-\frac{\varepsilon_0\varepsilon_p\left(\varepsilon_p-1\right)U^2}{2}\frac{\partial}{\partial h}\left[\frac{1}{\left(\varepsilon_p d-\left(\varepsilon_p-1\right)h\right)^2}\right]=-\frac{\varepsilon_0\varepsilon_p\left(\varepsilon_p-1\right)^2U^2}{\left(\varepsilon_p d-\left(\varepsilon_p-1\right)h\right)^3} \tag{3.24}$$

In terms of wavelength, $\lambda=2\pi/q$ combined with **Equation 3.24**, the maximum wavelength is written as

$$\lambda_{max}=2\pi\sqrt{\frac{2\gamma}{\left|\partial p_{el}/\partial h\right|}}=2\pi\sqrt{\frac{2\gamma\left(\varepsilon_p d-\left(\varepsilon_p-1\right)h\right)^3}{\varepsilon_0\varepsilon_p\left(\varepsilon_p-1\right)^2U^2}} \tag{3.25}$$

Finally, the associated maximal growth scales as proportional to the fourth power of the dominant wave vector, and is proportional to the ratio of surface tension to viscosity for a given λ:

$$\frac{1}{\tau_{max}}=\frac{\gamma h^3}{3\eta}q_{max}^4 \tag{3.26}$$

Combining **Equation 3.21** and **Equation 3.24**, results in the general dispersion relations

$$\frac{1}{\tau} = -\frac{h^3}{3\eta}\left[\gamma q^4 - \frac{\varepsilon_0 \varepsilon_p (\varepsilon_p - 1)^2 U^2}{(\varepsilon_p d - (\varepsilon_p - 1)h)^3} q^2\right] \quad (3.27)$$

Let us give some values to all the variables in **Equation 3.27** $\gamma = 30\,mN/m$, $\eta = 1.5\times10^4\,Pa.s$, $\varepsilon_0 = 8.85\times10^{-12}\,F/m$, $\varepsilon_p = 2.5$, $d = 10^{-6}\,m$, $U = 40\,V$.

The independent variable in this case is the initial polymer height $h = 10^{-6}\,m$ - $10^{-9}\,m$, and the wavenumber $q = 1.000$ - $2.000.000\,m^{-1}$ from the dispersion relationship.

The Python code to execute **Figure 3.3** is given in the Annex section **B.3**. In the dispersion relation we highlight the variation of the frequency $1/\tau$ as a function of both polymer height and wavenumber *(h, q)* in (**Figure 3.3 A**). Highlighting the region when damped (violet and dark blue) to amplified (green, yellow and red). Some values of *(h, q)* result in a non-physical frequency highlighted by the blank region. In **Figure 3.3B**, we highlight the different parameters relation to each other, **Figure 3.3Bi** related the dispersion relation frequency $1/\tau$ as a function of *q*, the stable/damped in violet and blue and unstable/amplified regime in cyan, green, yellow and red are presented. **Figure 3.3Bii** highlight the electrostatic pressure derivative with polymer height for different polymer heights, showing the polymer breaking point at around $1.65\times10^{-6}m$ height. The breaking point is also replicated in the electrostatic pressure plot (**Figure 3.3Biv**). The polymer breaking point represents the change between the polymers' growth regimes, from stable to unstable. Finally in **Figure 3.3Biii** we show the relation of q_c and q_{max} variation with polymer height, reaching a point where it breaks when $\partial p_{el}/\partial h < 0$.

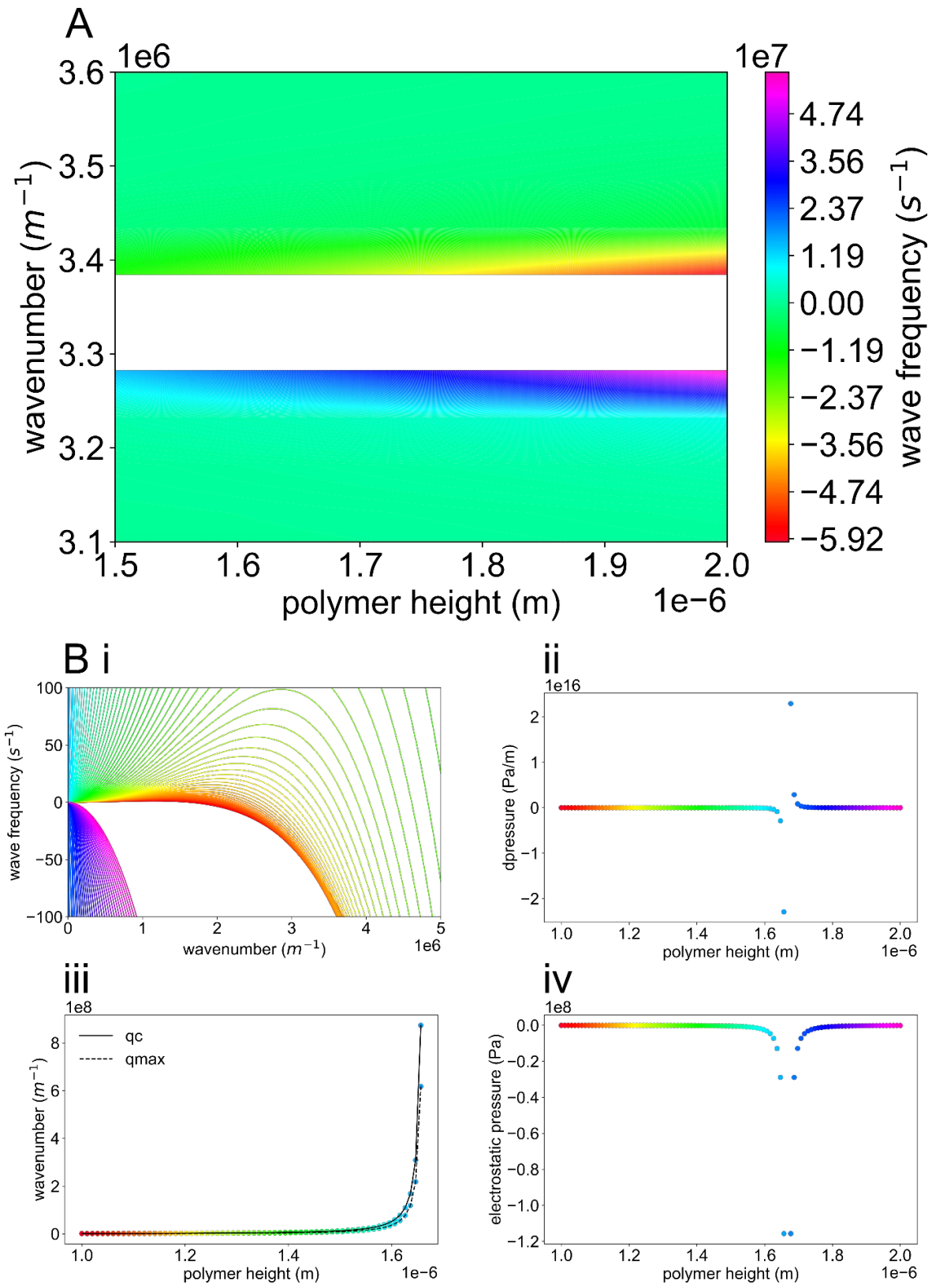


Figure 3.3: ***(A)*** *Contour plot of the 1/τ, in relation to q and h. To note regions of 1/τ>0 and 1/τ<0 and undetermined regime in blank.* ***(B)i)*** *1/τ as a function of q, the stable and unstable regime present.* ***ii)*** $\partial p_{el}/\partial h$ *variation with different polymer heights, showing breaking point at around* 1.65×10^{-6}*m.* ***iii)*** q_c *and* q_{max} *variation with polymer height, reaching a point where it breaks* $\partial p_{el}/\partial h<0$ *when.* ***iv)*** p_{el} *variation with film thickness and representation of the breaking point.*

3.2.2 Tilted Capacitor: Electrohydrodynamic Dispersion Relation

In practise, the EHD setup is not perfectly parallel, and there is usually a slight tilt between the substrates. A new mathematical formulation for the dispersion relation can be obtained using this new geometry.

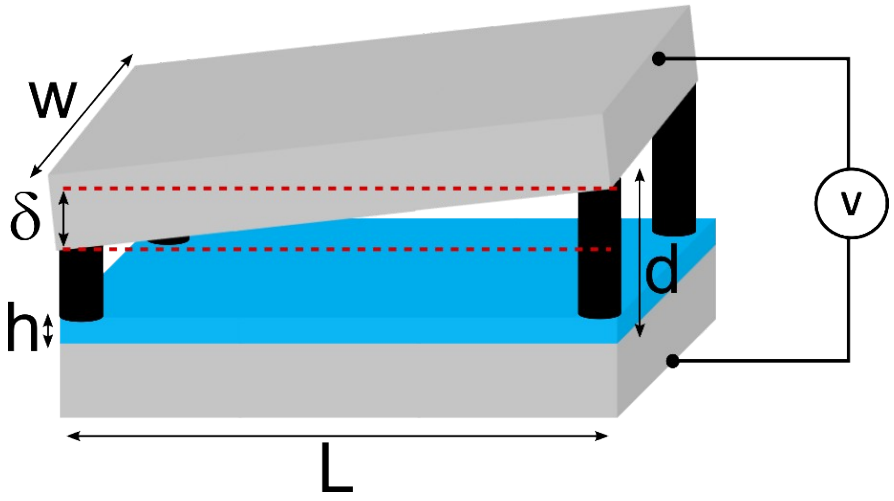


***Figure 3.4**:Tilting provoked by the different size spacers. With δ being the incremental height resultant from the tilting angle.*

The small tilt shown in (**Figure 3.4**) results in a slightly different air capacitor term to the parallel capacitance equation in **Equation 3.14**.

$$\frac{1}{C}=\frac{1}{C_{dielectric}}+\frac{1}{C_{air}'} \tag{3.28}$$

By considering the tilt geometry, the additional contribution from the tilt to the air capacitance can be calculated by integrating over the region. Assuming $\varepsilon_{air}\approx 1$

$$C_{air}^{'}=\int_0^L \frac{\varepsilon_0(wdx)}{d-\delta-h+\frac{\delta x}{L}}=\frac{\varepsilon_0 w}{d-\delta-h}\int_0^L \frac{dx}{\left(1+\frac{\delta x}{L(d-\delta-h)}\right)}=\frac{\varepsilon_0 w}{d-\delta-h}\int_0^L \left(1+\frac{\delta x}{L(d-\delta-h)}\right)^{-1}dx \tag{3.29}$$

Simplifying by using the following relation, $\delta<d \rightarrow (1+x)^n \approx 1+nx$ and $A=wL$

$$C_{air}^{'}=\frac{\varepsilon_0 w}{d-\delta-h}\int_0^L \left(1-\frac{\delta x}{L(d-\delta-h)}\right)dx \tag{3.30}$$

Let $\Delta=d-\delta-h$ and integrating with respect of *x* results in the new air capacitance formula

$$C_{air}^{'}=\frac{\varepsilon_0 A}{\Delta}\left[1-\frac{\delta}{2\Delta}\right] \tag{3.31}$$

The tilt is indeed something that can be observed through an optical image acquisition of an EHD pattern. A colour gradient can be observed among different EHD patterns (**Figure 3.5**). Each colour represents a different height.

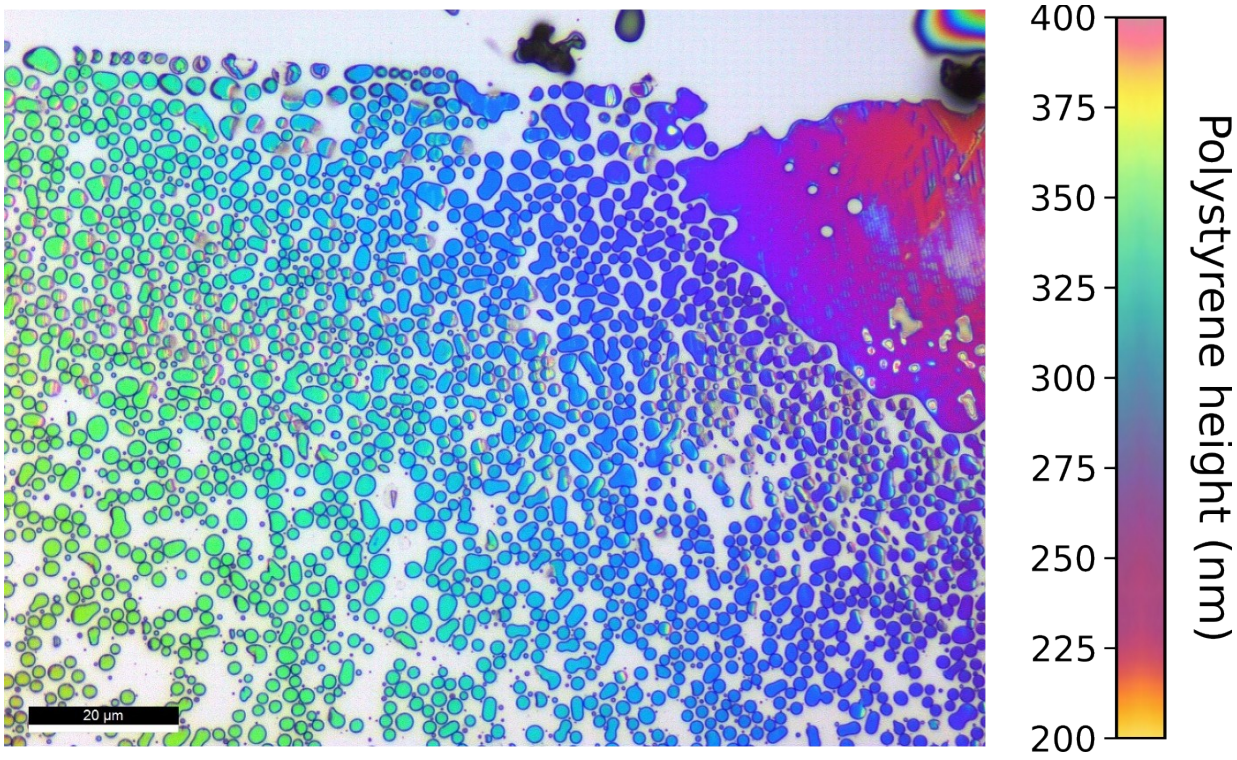


***Figure 3.5**: Optical microscope image of an EHD pattern using different height spacers. Polystyrene colour to height conversion from [2].*

The equivalent capacitance in this tilted geometry is then given by

$$\frac{1}{C'}=\frac{1}{C_{dielectric}}+\frac{1}{C'_{air}}=\frac{h}{\varepsilon_0\varepsilon_p A}+\frac{\Delta}{\varepsilon_0 A}\left[\frac{1}{1-\frac{\delta}{2\Delta}}\right]$$
$$\approx\frac{h}{\varepsilon_0\varepsilon_p A}+\frac{\Delta}{\varepsilon_0 A}\left[1+\frac{\delta}{2\Delta}\right]=\frac{h}{\varepsilon_0\varepsilon_p A}+\frac{d-h-\delta}{\varepsilon_0 A}+\frac{\delta}{2\varepsilon_0 A} \quad (3.32)$$

Which can be written as **Equation 3.14** plus the tilting contribution.

$$\frac{1}{C'}=\frac{h}{\varepsilon_0\varepsilon_p A}+\frac{d-h}{\varepsilon_0 A}-\frac{\delta/2}{\varepsilon_0 A}=\frac{\varepsilon_p d-(\varepsilon_p-1)h}{\varepsilon_0\varepsilon_p A}-\frac{\delta/2}{\varepsilon_0 A}=\frac{1}{C_{||}}+\frac{1}{C_{//}} \quad (3.33)$$

Let $d'=d-\delta/2$, substituting in **Equation 3.33** and get an equivalent form to **Equation 3.15.1**.

$$C'=\frac{\varepsilon_0\varepsilon_p A}{\varepsilon_p d'-(\varepsilon_p-1)h} \quad (3.34)$$

The dispersion relation calculation follows as in the parallel case, now with an added parameter δ which by itself can influence which regime the polymer growth is in. The same formulation as previously described was tested, with $0<\delta<0.8\,\mu m$. When $\delta=0$, the parallel plate condition is verified (**Figure 3.6i**). The polymer height was changed from $0.1\,\mu m<h<0.2\,\mu m$ the height air spacing was fixed $d=1\,\mu m$. The air gap at the highest tilt point can reach approximately *0* when $d=1\,\mu m$, $\delta=0.8\,\mu m$ and $h=0.2\,\mu m$ (**Figure 3.6ix**). At the lowest tilt point (i.e. parallel capacitor set up) the air gap reaches $0.9\,\mu m$ when $d=1\,\mu m$, $h=0.1\,\mu m$ and $\delta=0\,\mu m$ (**Figure 3.6i**). The air gap height difference influences the growth regime in the EHD growth.

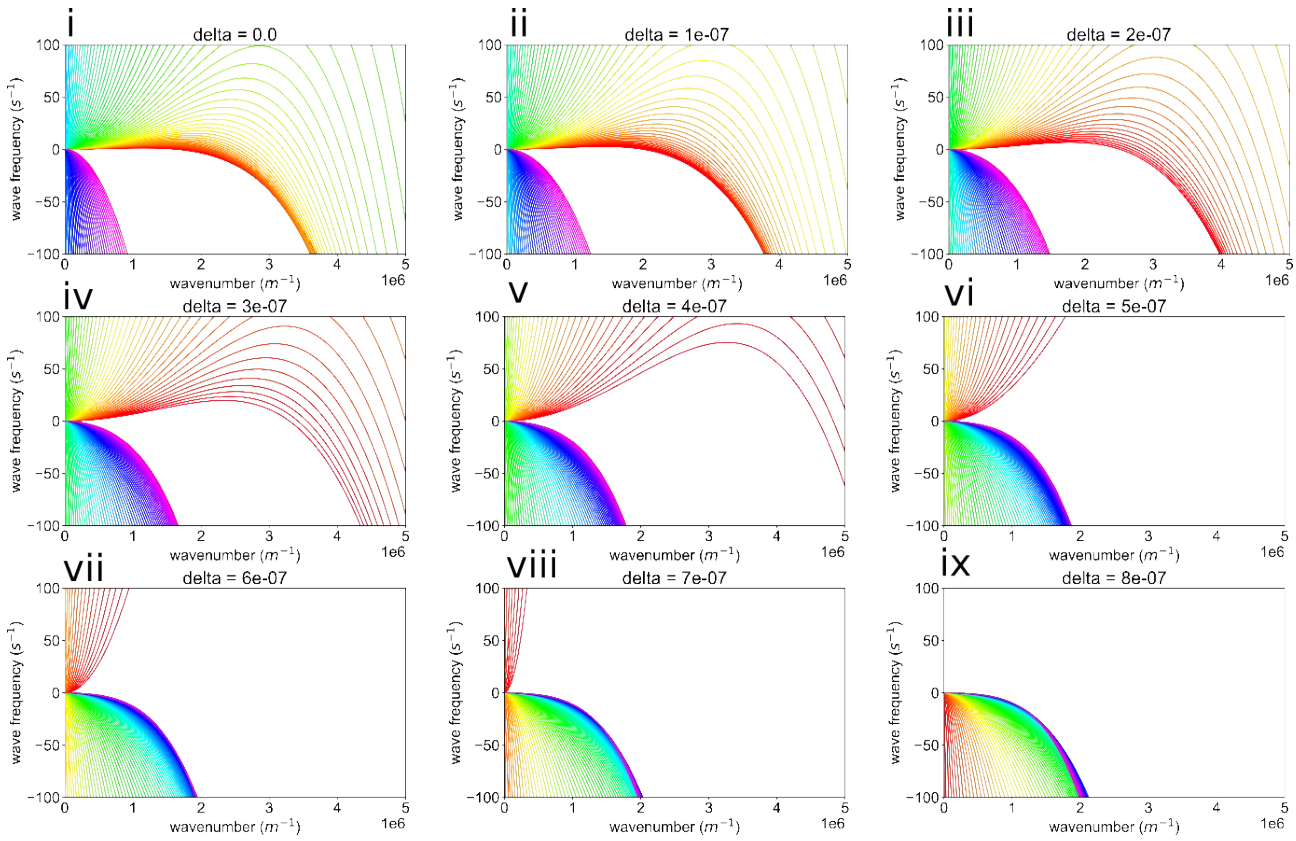


***Figure 3.6**: 1/τ as a function of q for different tilting δ values, the stable and unstable regime present. The parallel capacitance geometry is verified for δ=0 μm, by the similar dispersion regime identified in **Figure 3.3Bi**. When increasing the tilting values the amplified regime becomes extremely amplified, red line curvature from 0<δ<0.7μm (**i** - **viii**). For highest tilting δ=0.8μm (**ix**) the growth regime is completely damped.*

Hence explains the difficulty in precise control for the EHD growth without patterning the top electrode[4]. The full Python script used for the plots can be inspected in annex section **B.5**.

3.2.3 EHD simulations

To finalize the EHD discussion a simulation work of the differential equation given in **Equation 3.19** for the parallel EHD configuration, but now in 2D

$$\frac{\partial h}{\partial t}=\frac{h^2}{\eta}\frac{\partial h}{\partial x}\frac{\partial p}{\partial x}+\frac{h^2}{\eta}\frac{\partial h}{\partial y}\frac{\partial p}{\partial y}+\frac{h^3}{3\eta}\frac{\partial^2 p}{\partial x^2}+\frac{h^3}{3\eta}\frac{\partial^2 p}{\partial y^2} \tag{3.35}$$

In partial differential equations (PDE) format the previous equation can be written as

$$h_t = \frac{h^2}{\eta} h_x \left[-\gamma \left(h_{xxx} + h_{yyx} \right) + h_x p_h \right] + \frac{h^2}{\eta} h_y \left[-\gamma \left(h_{xxy} + h_{yyy} \right) + h_y p_h \right] + \frac{h^3}{3\eta} \left[-\gamma \left(h_{xxxx} + h_{yyxx} \right) + p_{hh} + p_h h_{xx} \right] + \frac{h^3}{3\eta} \left[-\gamma \left(h_{xxyy} + h_{yyyy} \right) + p_{hh} + p_h h_{yy} \right] \tag{3.36}$$

In this equation, subscripts *x* and *y* denote the spatial derivatives and subscript *t* denotes the time derivative. Here *p* is the electrostatic pressure.

Rewriting as well the electrostatic pressure from **Equation 3.24** as,

$$p = -\frac{1}{2} \frac{\varepsilon_0 \left(\varepsilon_p - 1 \right) U^2}{\varepsilon_p d^2 \left[1 - \frac{\varepsilon_p - 1}{\varepsilon_p d} h \right]^2} \tag{3.37}$$

And simplify with $A = -\frac{\varepsilon_0 U^2}{2d}$ and $B = \frac{\varepsilon_p - 1}{\varepsilon_p d}$ so that $p = \frac{AB}{\left(1 - Bh \right)^2}, p_h = -\frac{2AB^2}{\left(1 - Bh \right)^3}$ and

$p_{hh} = \frac{6AB^3}{\left(1 - Bh \right)^4}$.

thus **Equation 3.37** turns into

$$\begin{aligned} h_t = {} & \frac{h^2}{\eta} h_x \left[-\gamma \left(h_{xxx} + h_{yyx} \right) - \frac{2AB^2}{\left(1 - Bh \right)^3} h_x \right] \\ & + \frac{h^2}{\eta} h_y \left[-\gamma \left(h_{xxy} + h_{yyy} \right) - \frac{2AB^2}{\left(1 - Bh \right)^3} h_y \right] \\ & + \frac{h^3}{3\eta} \left[-\gamma \left(h_{xxxx} + h_{yyxx} \right) + \frac{6AB^3}{\left(1 - Bh \right)^4} - \frac{2AB^2}{\left(1 - Bh \right)^3} h_{xx} \right] \\ & + \frac{h^3}{3\eta} \left[-\gamma \left(h_{xxyy} + h_{yyyy} \right) + \frac{6AB^3}{\left(1 - Bh \right)^4} - \frac{2AB^2}{\left(1 - Bh \right)^3} h_{yy} \right] \end{aligned} \tag{3.38}$$

Equation 3.38 was written into a Python script and is in Annexe **B.6**. To solve the PDE, a loop was run over the small height increment to the film thickness, until the sum of the

initial film thickness and the increment passed the air gap. The height increment was calculated as:

$$dh = dt \times \left\{ \begin{array}{l} \frac{h^2}{\eta} h_x \left[-\gamma \left(h_{xxx} + h_{yyx} \right) - \frac{2AB^2}{\left(1-Bh\right)^3} h_x \right] \\ + \frac{h^2}{\eta} h_y \left[-\gamma \left(h_{xxy} + h_{yyy} \right) - \frac{2AB^2}{\left(1-Bh\right)^3} h_y \right] \\ \frac{+h^3}{3\eta} \left[-\gamma \left(h_{xxxx} + h_{yyxx} \right) + \frac{6AB^3}{\left(1-Bh\right)^4} - \frac{2AB^2}{\left(1-Bh\right)^3} h_{xx} \right] \\ \frac{+h^3}{3\eta} \left[-\gamma \left(h_{xxyy} + h_{yyyy} \right) + \frac{6AB^3}{\left(1-Bh\right)^4} - \frac{2AB^2}{\left(1-Bh\right)^3} h_{yy} \right] \end{array} \right\} \quad (3.39)$$

Iteratively, the new h is calculated as $h_i = h_{i-1} + dh_{i-1}$ with i the iteration value, with the stopping condition when the new h_i reaches the top of the electrode. The results of the EHD growth in 2D are shown below (**Figure 3.7**). Highlighting the difference between the initial stage and the final EHD growth stage (**Figure 3.7A** and **Figure 3.7B**), with the stop condition achieved when the film thickness passes the capacitance air gap (**Figure 3.7C** and **Figure 3.7D**) on the 40th iteration.

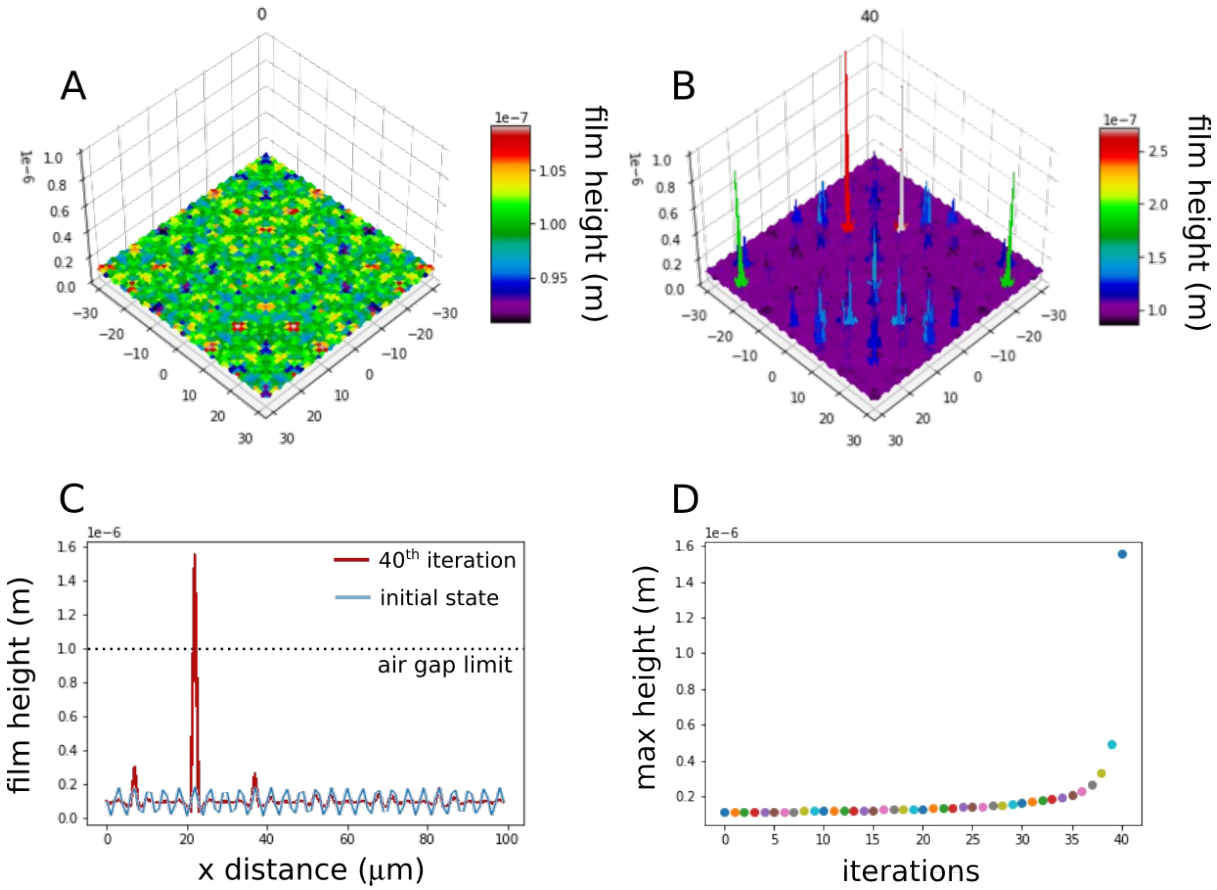


Figure 3.7: (A) *Cell of the initial EHD condition with small surface fluctuations given by a sine wave as a function of x and y distance.* ***(B)*** *Last iteration of the EHD process. Multiple nucleation site separated by a λ with one spot reaching the top of the electrode, stopping the iteration.* ***(C)*** *Check the spot of high growth in a 2D representation (slice of the 3D* ***Figure 3.7B****) at the start of the run and at the end.* ***(D)*** *1D representation (point of the 3D* ***Figure 3.7B****) of the highest growth spot per iteration and observing the exponential growth trend.*

3.3 Alternative methods to determine Electrohydrodynamic pattern heights

3.3.1 Colour to Height conversion

As previously mentioned, the tilting growth of the EHD patterns makes the appearance of different coloured structures. Colours in the different structures result from the interference of the two reflected beams at the polymer-air and the polymer-substrate interfaces of the incident white light of the microscope. These two beams can either interfere constructively or destructively. Therefore, interference from a structure with a given height enhances or

cancels particular wavelengths, and the film will appear differently coloured depending on the structure height. From [2], EHD colours have a distinct colour formation for different film thicknesses. From 20nm (light brown), 70nm (dark brown), 100nm (dark blue), 140nm (light blue) 200nm (yellow), 250–280nm (purple), 290nm (blue), 310nm (turquoise), 330nm (green), 350nm (yellow), 400nm (light purple), 420nm (green), 460nm (yellow), 520nm (pink) and alternating light green and pink up to approximately 1.5μm until it changes to a transparent grey.

Thus, using this matching between colour (RGB) and thickness (nm) values, thickness information can be extracted from an optical image. To manage the conversions from optical colour image to height image, a Python script was written. The code is available in Annex Section B.3. The relationship to convert from RGB to polystyrene film thickness was obtained from Figure 3.8B by fitting a linear curve $thickness(RGB) = a \times RGB \in [0,1] + b$ to the previous data points. An additional auxiliary function (Figure 3.8A) was used to track the three individual RGB values and convert them into a scalar number between 0 and 1, which, when combined with the fit line, converts the optical image into the appropriate height.

From the best-fit line the parameters obtained were $a = 456.6$ and $b = 1.1$. Thus, any optical EHD pillar image can be loaded, and processed by using these parameters and thus convert the image into a thickness EHD pillar image. Here are some examples to illustrate how this method works.

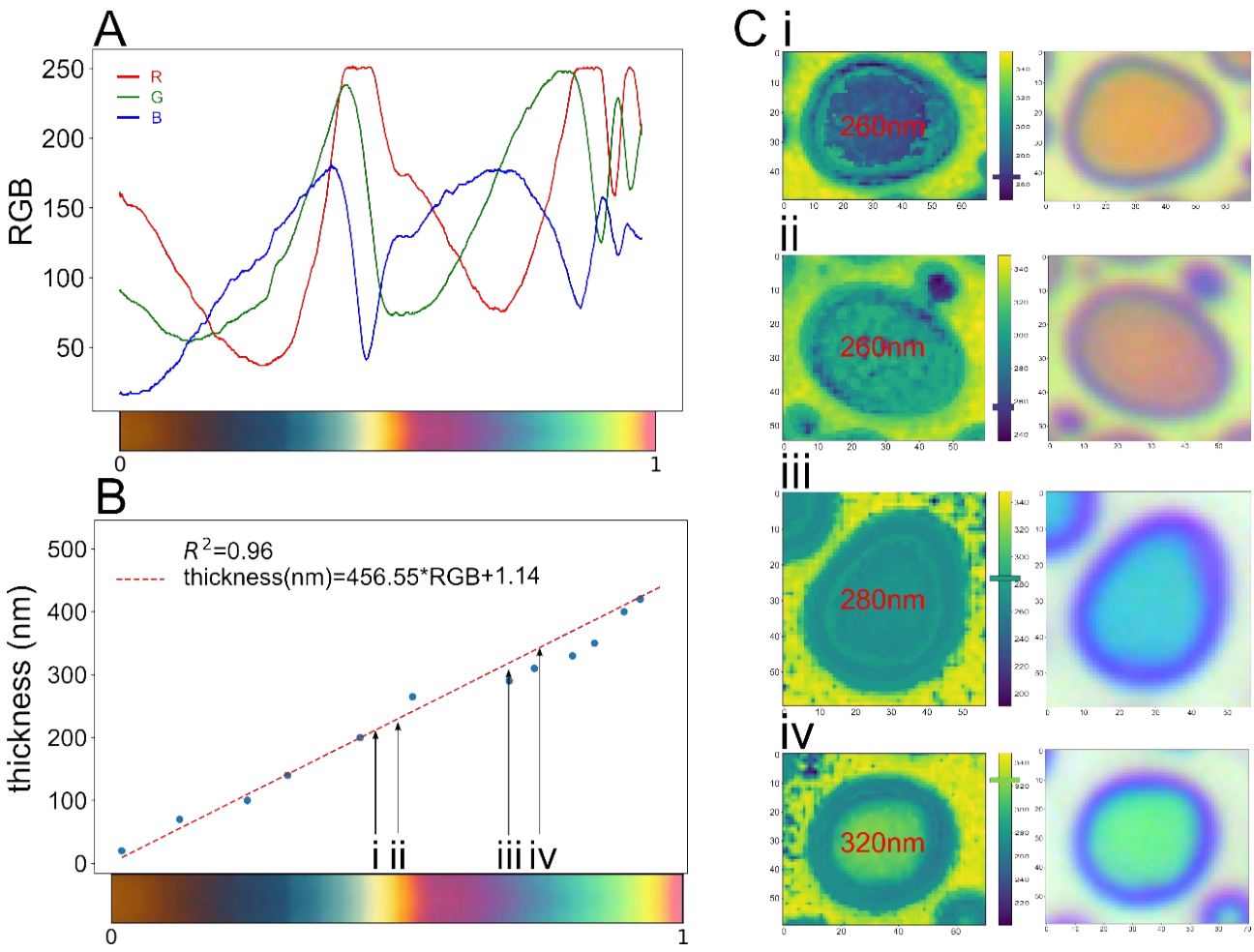


***Figure 3.8**: **(A)** RGB values variation from the colour spectra of polystyrene, which are used to convert into a single value between 0 and 1, represented by the polystyrene calibration colours. **(B)** Polystyrene RGB values ranged between 0 and 1 (represented by the polystyrene calibration colours), and respective curve fit. **(C)** Top view of some EHD pillars examples and their colour to thickness conversion **(i-iv)**, and respective matching in the calibration curve in **(B)**.*

The colour to thickness conversion method has its shortcomings. Colours are not clearly defined when zooming in on the individual pillars. The colour becomes distorted due to the quality of the image file formats. There are also colour repetitions like yellow, blue and pink/purple that appear more than once, which makes the script inaccurate for these colours. Furthermore, the background is not completely white, so a script needs to be developed to mask the substrate when computing the colour-to-height converter for the pillars. Moreover, the EHD pillars have a border that affects the colour and height. The optical microscope acquisition brightness setup can also influence the colour of the pillar and thus its thickness conversion.

Despite these drawbacks, this method opens many possibilities for representing the pillars. A pseudo-3D AFM image can be acquisition by plotting the converted height to a 3D plot. The *x* and *y* values are by the image scale bar and the *z* by the colour-height conversion script. Here are some example (**Figure 3.8Ci-iv**) of the conversion into full pseudo-3D pillars (**Figure 3.9i-iv**).

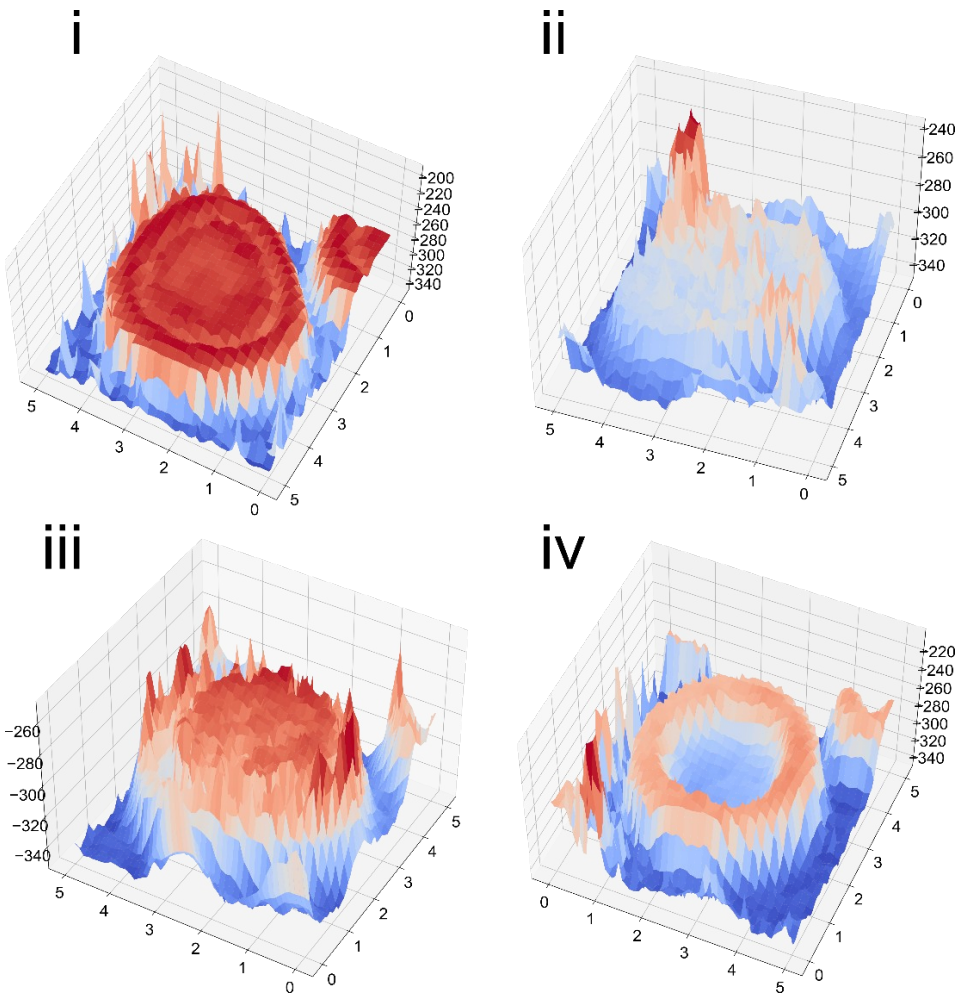


***Figure 3.9**: **(i-iv)** 3D pillars image after applying the colour to thickness converter on the pillar optical images from **Figure 3.8Ci-iv**.*

The 3D pillar representation from the colour to thickness conversion script outlines the lack of definition from the optical image. The individual pillar images have around 50-70 pixels per size, while an AFM image would obtain around 256-1024 pixels per side. The information gained from the AFM makes the image quality higher, although the time to acquire these images is also higher.

Despite all the drawbacks of this method, it opens up the possibility of getting height information related to optical images by having a calibration curve to translate from colour (RGB values) to height (nm), which can potentially be used in any polymer that behaves the same way. Furthermore, this method will improve with better optical setups.

3.3.2 Electrohydrodynamic Mass conservation

There is an interesting property when forming EHD surfaces. The initial polymer mass is conserved after the EHD patterns generation. Imagine that we initially have a polystyrene film with thickness *t* and area coverage *A*. This occupies a volume and a mass, that after the EHD process, is converted into an *n* number of pillars with height *h* and radius *r*. The mass before the EHD formation and after needs to be conserved and thus:

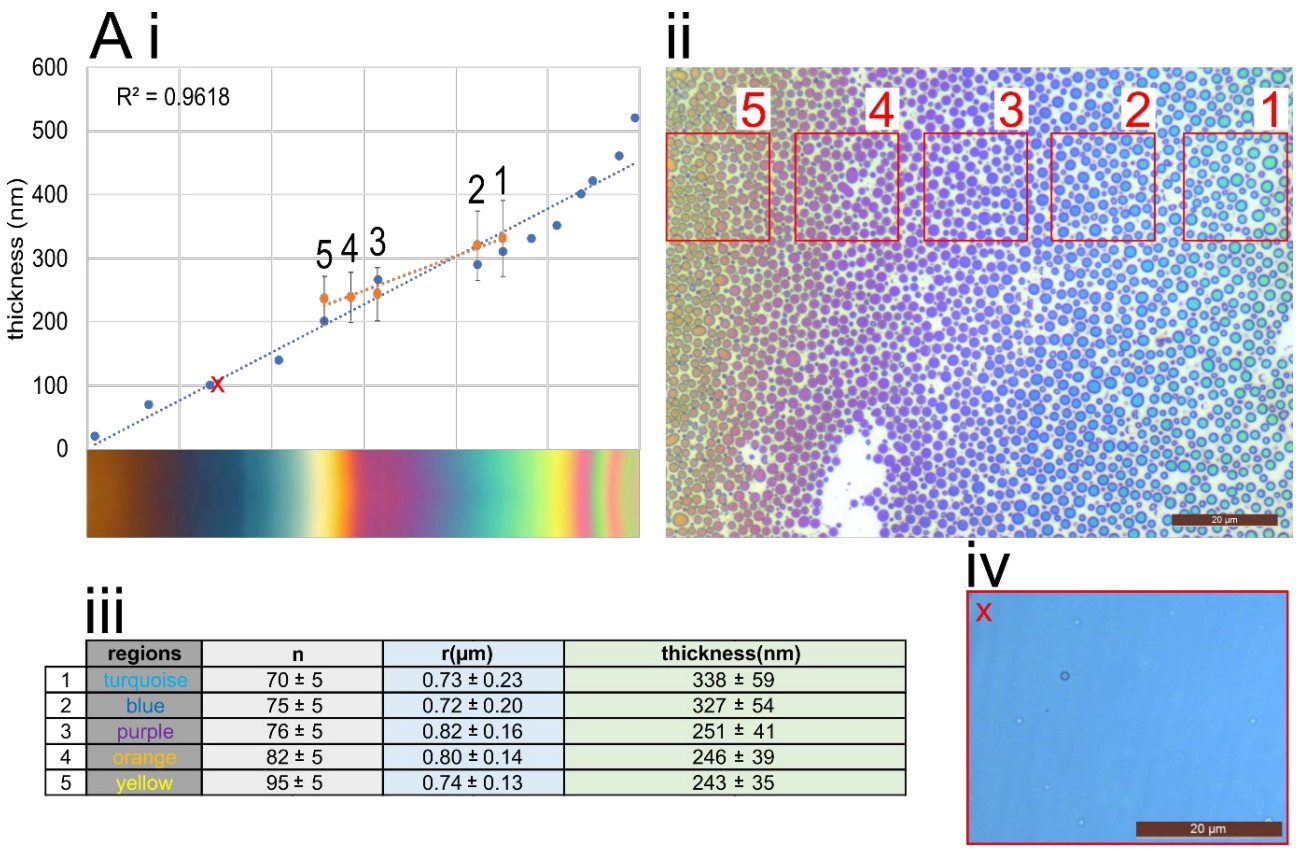
$$\Delta m = 0 \rightarrow \rho_{polymer} A_{film} h_{film} = \rho_{polymer} n_{pillar} h_{pillar} \pi r_{pillar}^2 \quad (3.40)$$

By rearranging **Equation 3.40**, we can formalize an equation to approximate the height of the pillars in a region.

$$h_{pillar} = \frac{A h_{film}}{n \pi r_{pillar}^2} \quad (3.41)$$

To test this formula, we calculated the pillar heights of 5 different spots (**Figure 3.10Aii** red squares) in an optical image as a proof-of-concept of our method. The film is initially coated with a $h_{film} = 100\,nm$ layer, represented in **Figure 3.10Aiv**, and respective the thickness point is illustrated in **Figure 3.10Ai** with a red cross. The number of pillars inside each red square, *n*, was recorded, and since the red square inevitably makes some pillars

appear in the boundaries, the error for the number of pillars was given as an average value for the number of pillars cut from this selection. The radius of each pillar, *r*, was calculated using ImageJ software, first by selecting the scale bar dimension and then calibrating it to select all the different pillar diameters, then halved for the radius value. However, there is some irregularity in the pillar radius, and the pillar dimensions have a large variance, as can be observed in **Figure 3.10Aiii**, having an error between 17-32% of the value. Finally, the area of the selected red regions *A* was chosen to be equal to $400\,\mu m^2$, the area of each red area. With these considerations, we can estimate the height of the pillars, which we have calculated in **Figure 3.10Aiii**. These points were highlighted in **Figure 3.10Ai** next to the previous colour *versus* thickness diagram. These new calculated points match the overall trend fitted in the previous chapter. With these calculations, we show a simple way of estimating the pillar height based on the number of pillars, pillar radius, initial height, and areal region, all parameters that can be obtained from an optical microscope image.



	regions	n	r(µm)	thickness(nm)
1	turquoise	70 ± 5	0.73 ± 0.23	338 ± 59
2	blue	75 ± 5	0.72 ± 0.20	327 ± 54
3	purple	76 ± 5	0.82 ± 0.16	251 ± 41
4	orange	82 ± 5	0.80 ± 0.14	246 ± 39
5	yellow	95 ± 5	0.74 ± 0.13	243 ± 35

***Figure 3.10**: **(A) i)** Polystyrene colour variation with thickness, with added height points for 1-5 related to the different regions in **ii)** showing that these fit nicely onto the best fit line, and a cross from initial film thickness in **iv). ii)** Optical image of an EHD patterned surface, for proof-of-concept test of [**Equation 41**] over 5 different EHD region. **iii)** Table classification for the different pillar colours and respective parameters, number of pillars, n, radius, r, and the calculated thickness, t. **iv)** Initial polystyrene film optical image before EHD formation.*

3.4 Conclusion

This chapter examined the physics underlying the production of EHD. Different experimental geometries were examined, including the common parallel capacitor configuration and its extension to the tilted configuration, where differences in the dispersion relation between the two cases were examined. Additionally, the parallel configuration dispersion relation was extended to 2D, and the PDE was simulated with similar results to reality. Moreover, two different analyses were done on the EHD optical imaging. One is related to the conversion between EHD pattern colours and heights by following a linear relation between EHD colours and thickness, and the second is a relation

between mass conversion between the initial film deposition state and the after EHD pattern formation. Both are alternative ways of estimating the height of EHD patterns with only an optical image without performing additional measurements.

3.5 References

[1] Oppenheimer P. G., “Electrohydrodynamic Patterning of Functional Materials,” no. July, pp. 11–30, **2011**. PhD thesis, University of Cambridge, Cambridge, UK. Available: https://www.repository.cam.ac.uk/bitstream/handle/1810/245301/OppenheimerThesis.pdf

[2] Schaffer E., “Instabilities in Thin Polymer Films : Structure Formation and Pattern Transfer Instabilities in Thin Polymer Films: Structure Formation and Pattern Transfer,” no. September, pp. 31-49, **2001**. PhD thesis, University of Konstanz, Baden-Wurttemberg, DE. Available: https://d-nb.info/964088665/34.

[3] Rickard J. J. S., “Advanced Micro- Engineered Platforms for Novel Device Technologies,” no. December, pp. 62–74, **2017**. PhD thesis, University of Birmingham, Birmingham, UK. Available: https://etheses.bham.ac.uk/id/eprint/8303/.

[4] Rickard JJS, Farrer I, Goldberg Oppenheimer P. Tunable Nanopatterning of Conductive Polymers via Electrohydrodynamic Lithography. Vol. 10, ACS Nano. American Chemical Society (ACS); **2016**. p. 3865–70. Available from: http://dx.doi.org/10.1021/acsnano.6b01246.

[5] Nazaripoor, H., Koch, C.R., Sadrzadeh, M. and Bhattacharjee, S. “Thermo-Electrohydrodynamic Patterning in Nanofilms”. Vol. 32, Langmuir. American Chemical Society (ACS); **2016**. p. 5776–86. Available from: http://dx.doi.org/10.1021/acs.langmuir.6b01810.

[6] Ruckenstein, E., Jain, R.K.. “Spontaneous rupture of thin liquid films". Vol. 70, Journal of the Chemical Society, Faraday Transactions 2. Royal Society of Chemistry (RSC); **1974**. p. 132. Available from: http://dx.doi.org/10.1039/f29747000132

[7] L.D Landau and E.M. Lifshitz. Fluid mechanics. London: Pergamon Press, **1959**.

[8] Oron, A, Davis, S.H. and Bankoff, S.G. “Long-scale evolution of thin liquid films”. Vol. 69, Reviews of Modern Physics. American Physical Society (APS); **1997**. p. 931–80. Available from: http://dx.doi.org/10.1103/RevModPhys.69.931

[9] Teletzke, G.F., Davis, H.T. and Scriven, L.E. Wetting hydrodynamics. Vol. 23, Revue de Physique Appliquée. EDP Sciences; **1988**. p. 989–1007. Available from: http://dx.doi.org/10.1051/rphysap:01988002306098900

[10] Voicu, N.E., Harkema, S. and Steiner, U. “Electric-Field-Induced Pattern Morphologies in Thin Liquid Films”. Vol. 16, Advanced Functional Materials. Wiley; **2006**. p. 926–34. Available from: http://dx.doi.org/10.1002/adfm.200500470.

[11] Pease L.F. and Russel, W.B. “Electrostatically induced submicron patterning of thin perfect and leaky dielectric films: A generalized linear stability analysis”. Vol. 118, The Journal of Chemical Physics. AIP Publishing; **2003**. p. 3790–803. Available from: http://dx.doi.org/10.1063/1.1529686

CHAPTER 4

Electrofluidynamic Patterning of Tailorable Nanostructured Substrates for Surface-Enhanced Raman Scattering

This work was published has "***Gomes, P.D.C.**, Rickard, J.J.S., Goldberg Oppenheimer, P. "Electrofluidynamic Patterning of Tailorable Nanostructured Substrates for Surface-Enhanced Raman Scattering". Vol. 3, ACS Applied Nano Materials. American Chemical Society (ACS); **2020**. p. 6774–84. Available from: http://dx.doi.org/10.1021/acsanm.0c01190 "*, and my contribution as the first author was in the EHD fabrication, SERS data acquisition, AFM data acquisition, analysis of the relationship between the aspect ratio, nano-rims, nano-cavities, and nano-roughness to the SERS signal enhancement, and development of a CNN for the identification of 2D AFM profiles of EHD pillars regarding their SERS enhancement (IF SERS active or not).

Other authors' contributions to the paper included the creation of the top-electrode patterning with the ideal EHD shape for SERS enhancement, as shown in **Figure 4.4**. The mechanical setup for fine-tuning the height adjustment, signal enhancement simulations for these EHD patterns, and finally the spike-and-recovery assay with spiked artificial blood plasma deposited on the fabricated substrates.

4.1 Abstract

The unique attributes of surface-enhanced Raman spectroscopy (SERS) make it well suited to address the challenges associated with portable diagnostics. However, despite the remarkable progress in this field, in which the instrumentation has made great strides forward providing a route to the miniaturization of sensing devices, to date producing three-dimensional low-cost SERS substrates which simultaneously fulfill the multitude of criteria of high sensitivity, reproducibility, tunability, multiplexity, and integratability for rapid sensing has not yet been accomplished. Successful implementation of SERS requires readily fine-tuned nanostructures, which create a high enhancement.

Here, an advanced electrofluidynamic patterning (EHD) technique enables rapid fabrication of SERS active topographic morphologies with high throughput and at a nanoresolution via the spatial and lateral modulation of the dielectric discontinuity due to the high electric field generated across the polymer nanofilm and air gap. The subsequent formation of displacement charges within the nanofilm by coupling to the electric field yield a destabilizing electrostatic pressure and an amplification of EHD instabilities enabling the controllable pattern formation. The top of each gold coated EHD fabricated pillar generates controllable, high SERS enhancement by coupling surface plasmon modes on top of the pillar, with each nanostructure acting as an individual sensing unit. The absolute enhancement factor depends on the topology as well as the tuneable dimensions of the nanostructured units, and these are optimized in the design and engineering of the dedicated EHD apparatus for reproducible, low-cost fabrication of the three-dimensional nanoarchitectures on macrosurfaces, rendering them easy for integration in further lab-on-a-chip devices. This unique combination of nanomaterials and nanospectroscopic systems lay the platform for a variety of applications in chemical and biological sensing.

4.2 Introduction

Surface-enhanced Raman spectroscopy (SERS) has been undergoing a renaissance in the past decade, on a trajectory to become the best-positioned analytical technique to posture significant impact on portable sensing. SERS, due to its distinctive attributes of rapid detection, high-sensitivity heightened with the capability of detection down to single molecule levels [1–7], as well as multiplexing, all while requiring no complex sample preparation, can be easily deployed for a breadth of applications ranging from homeland security through to medical diagnostics, biochemical sensors and onto food pathogenesis. Nevertheless, despite the large amount of SERS-related research activity being pursued,

the field still poses many challenges to be addressed due to the numerous conditions with the majority of which must be fulfilled simultaneously for SERS to become a well established sensing tool. Any imperfections in substrates will have a significant effect on the definitive response, alongside this, SERS-active platforms are often plagued by irreproducibility, instability, and a lack of tunability toward specific molecules. While high-enhancement is possible with SERS-active nanoparticles (NPs), only a miniscule fraction exhibits SERS activity, which also have a high-sample variability and poor batch-to-batch reproducibility, all substantially affecting the reproducibility of the measured signal[8–12]. This is exacerbated by both the use of surfactants to stabilize the NPs and the control of their shapes and aspect ratios, which in turn tend to reduce the subsequent attachment of the target analytes as along with the lack of established methods to further assemble the fabricated SERS-active NPs into distinctive “hot-spots” on macroscale ranges. Furthermore, SERS substrates with tuneable localized surface plasmon resonances (LSPR) which can be easily interfaced externally (e.g., in electrochemical devices) or integrated internally (e.g., inside microfluidic chips) are needed to gain the required, consistent, tailored high enhancement, a challenging task for NPs-based SERS systems. Therefore, lithographically structured substrates exhibit a great potential to intrinsically overcome these problems for enabling consistently reproducible SERS measurements. However, despite the significant advances of patterning methods and enhanced microfabrication techniques, the vast majority of synthetic “bottom-up” and “top-down” routes to generate SERS nanostructures are still based on conventional lithographic approaches. These include for instance photolithography, electron beam lithography, and focused ion beam (FIB), which are not only time-consuming and cumbersome but also require precise integration of multistep processes, [8,10,12–15] limiting the scalability of the resulting platforms, resulting in substrates that are far too expensive for practical

applications. Inadequate resolution of the mould, two-dimensionality, the inevitable defects, poor mechanical stability, and pattern distortion collectively constitute the further limiting factors of micronanolithographies and other imprinting techniques. The fluctuating SERS enhancement measured across a single, as well as seemingly identical substrates, further exacerbates the downsides of the current fabrication processes. These limitations have driven extensive efforts to explore novel processes for the fabrication of highly ordered 3D structures with nanometer periodicities and underline the continuous, unmet need for high-throughput, reliable, and cost-effective methods to develop tuneable, three-dimensional micronanoarchitectured surfaces to deliver highly reproducible, laterally and spatially consistent SERS enhancement. [16–18]

Here, an advanced technique of fabricating highly reproducible structures *via* controllable electrofluidynamic patterning (EHD) combined with a novel nanolithographic device was developed to determine the needed level of control required to generate tailor-made, structured submicrometer SERS platforms (**Figure 4.1**) with fine-tuned nanoarchitectures (in contrast to NPs) wherein uniquely, each of the individual structural units yields a high enhancement. Importantly, this contrasts typical SERS active substrates, which act as collective surfaces for signal augmentation and which often encounter numerous difficulties in controlling the fabricated morphologies ranging from nano to micrometer scale onto the overall macrosurface. It has been previously demonstrated that the combined use electric field and hydrodynamics for lithography by patterning dielectric materials, conductive and crystalline materials, superapolar lotus-to-rose hierarchical nanosurfaces, with controllable alignment of the internal nanomorphologies. [19–23] Building upon these principles, in this study a systematically exploration of the localized plasmons in different nanogeometries, delivering an intuitive guidance for appropriate nanostructure design and introducing intelligent engineering of a

novel nanolithographic device, which is further exploited to fabricate and control the metallic nanomorphologies with an inherent nanoroughness for tuneable surface-enhanced Raman spectroscopy. Optimized EHD further enables the controllable fabrication of the desired platforms for multiplexed SERS detection. Building upon the vibrational spectroscopy principles, intimately dependent on the interplay between the laser excitation wavelength and the structural unit shapes and dimensions, an in-depth characterization and optimization of the EHD generated architectures have further been carried out to enable the fabrication of tailor-made SERS substrates (**Figure 4.2**). Tunability of the structural dimensions was possible via the independent control of various experimental parameters. For instance, the strength of the applied voltage (and thus, the generated electric field) allowed control of the speed and the consistency of the patterning as well as dictate the characteristic wavelength of the instability which, in turn, determined the spacing/pitch between the fabricated pillars. The filling ratio enabled the regulation of the structures' diameters. The small gap between the electrodes was carefully adjusted via the use of (nano) spacers with various dimensions which, controlled the achievable heights of the structures. Tuning any of these parameters allowed the fabrication of controllable diameter, height, and pitch, all dictating the optical responses of the structures. This tunability yielded the subsequent design of substrates that generate strong localized electromagnetic fields at optical wavelengths that are required for the optimum SERS excitation by different laser sources. Interestingly, depending on the EHD generated topographic morphologies of the top structural individual unit surfaces, which varied between a nanorim, planar, or roughened top, certain topologies exhibited considerably more enhanced SERS signal, which further guided the design of the dedicated electrode for reproducible fabrication of SERS substrates with optimum performance. Subsequently, a novel apparatus for fabricating high-precision, controllable EHD SERS-active platforms

with optimal structural parameters for reproducible SERS was designed and engineered. Validation bioassay comprised of the gold coated predesigned pillar structures are further utilized for rapid detection of traumatic brain injury indicative biomarkers, demonstrating that patterned EHD arrays can act as straightforward and cost-effective substrates for high-throughput SERS detection. Overall, EHD-based patterning postures salient advantages over other lithographic methods, including the inherent tunability of the fabricated nanoarchitectures (via directly adjusting the various experimental parameters), straightforward single-step patterning, the speed which can be easily controlled via the chosen material to be patterned (with the instability being directly proportional to the viscosity [21,24]), cost-effectiveness of the fabricated substrates (made from low-cost polymers coated with a gold nanolayer), scalability, and the ease of integration within a lab-on-a-chip. These characteristics lay the basis for the design of highly reproducible SERS-active arrays with controllable localized plasmonic fields for a variety of lab-on-chip sensing devices.

4.3 Results and Discussion

The EHD method of fabricating a platform for SERS comprises of assembling a microcapacitor device with bottom and top surfaces, facing each other, used as two opposing electrodes. A thin polymer film with an initial thickness, h (h = 80–100 ± 21 nm) is deposited on a bottom electrode, and a second top planar electrode is placed at a controllable distance, d (d = 150–800 ± 53 nm), leaving a certain air gap, which can be adjusted to micronanometer levels of precision (**Figure 4.1a**). The layered system is therefore equivalent to a sum of two capacitors in series filled-in with polymer and air layers. Subsequently, the film is fluidified, by heating the whole microcapacitor to an

elevated temperature (i.e., above the glass transition temperature, T_g, of the polymer film) and in parallel applying an external voltage, V, which in turn generates a high electric field, E_f (~10^8 V/m) within the small capacitor's nanogap, d' across the dielectric material. The EHD induces and successively exploits the evolving surface instabilities to generate tunable lithographically defined structural morphologies on the bottom electrode. For the sufficiently strong E_f generated inside the microcapacitor, the destabilizing electrostatic pressure, p_{el} overcomes the stabilizing effects of the Laplace pressure, p_L acting at the liquid–air interface (**Equation 4.1**), which to minimize the free electrostatic energy of the system acts to align the interface, separating the two media, in parallel to the E_f lines:

$$p_{el}\left(=-1/2\,\varepsilon_0\,\varepsilon_p\left(\epsilon_p-1\right)E_f^2\right) > p_L\left(=-\gamma\frac{\partial^2 h}{\partial x^2}\right) \tag{4.1}$$

where ε_0 is the dielectric permittivity of the vacuum, ε_p is the dielectric constant of the polymer, and γ is the surface tension. Consequently, the thin polymer film adopts a profile of pillars forming on the bottom substrate that correspond to the locations of the initially induced perturbations (**Figure 4.1b**).

Once the EHD is complete, the electric field is maintained while the temperature is allowed to drop to below the fluidifying transition temperature, i.e., the T_g of the polymer, and the bottom substrate is released. This preserves the final structures and avoids distortions or fluctuations in the polymer. Subsequently, coating the EHD generated submicron architectures with a plasmon-active gold nanolayer (25.0 ± 2.7 nm) yields effective enhancement of the local electromagnetic field (**Figure 4.1c**). EHD thus enables a single-step fabrication of submicron SERS active substrates, with each of the individual pillars yielding a considerable signal enhancement (**Figure 4.1c** and **Figure 4.3**). [20-22,24]

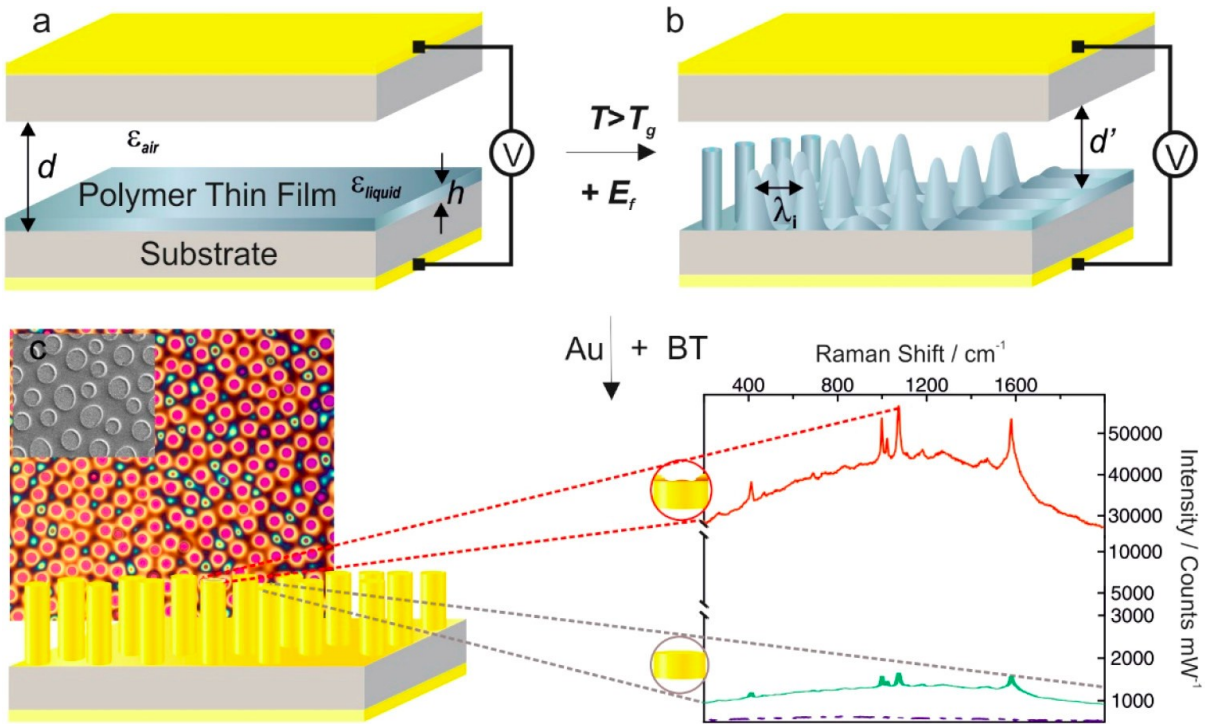


Figure 4.1. *EHD process for SERS fabrication. To commence the pattering process,* ***(a)*** *a thin polymer film with an initial film thickness h is deposited between two electrodes spaced apart by a distance d. Applying an external voltage, V and liquefying at T > T_g leads to the generation of a high electric field, E_f (~10^8V/m) within the microcapacitor and* ***(b)*** *formation of an instability throughout the film with a characteristic wavelength λ, which determines the pitch between the final structures. The increasing amplitude of the film variations guides the film's redistribution, which eventually spans the gap between the two electrodes.* ***(c)*** *Top-view optical microscopy and (inset) scanning electron microscopy images of the EHD generated polymer structures, which are subsequently coated with a plasmon active gold nanolayer leading to high SERS signal enhancement.* ***(inset)*** *Representative SERS spectra of BT on top of various topologies of the formed pillars exemplifying that the roughened centre surrounded by a nanorim (top view, red) yields more than 50 times higher signal enhancement relative to the flat pillar topology.*

The EHD technique allows the patterning of micron and nanostructures with a broad range of feature sizes and topological morphologies of the final structural units (**Figure 4.1c, inset** and **Figure 4.2**). These can be accomplished by alternating the termination time of the patterning process, which can lead to either conelike structures or sharp pyramids with a round base or alternatively, due to the disassembly process of the top electrode can lead to the variation in shape and roughness of the tops of the formed structures. Salinising the top electrode prior to the onset of the patterning process renders it apolar to surfaces grafting a monolayer of octadecyltrichlorosilane between the upper and lower substrates and facilitates the smooth release of the substrates. This enables

easy disassembly due to the nonstickly self-assembled nanolayer. Alternatively, it is possible to directly disassemble the microcapacitor setup without any additional salinization steps. These enable the production of a range of structures where the top surface of each pillar is either smooth, concave, roughened, or a combination of those (**Figure 4.2**). Out of these, certain structural topologies were found to provide a further degree of signal enhancement on top of each pillar (**Figure 4.1c, inset**). Representative EHD micronanoarchitectures' cross-section atomic force microscopy (AFM) profiles, bright field optical microscopy (OM) top-view images, scanning electron microscopy (SEM) images, and convolutional neural network (CNN) images are shown in (**Figure 4.2**). The CNN algorithm for image classification of the EHD patterned structures for SERS labelling includes several stages (see **4.5 Methods** for details): (i) data gathering, (ii) data preprocessing, (iii) data augmentation, (iv) feature extraction using a pre-trained convolutional neural network, followed by (v) fine-tuning and the visualization of the activated neurons for the image classification. CNN images allow the classification of the formed pillars as a function of the enhancement and, thus, the correlation of the images with the degree of enhancement of the scattered radiation. The results from our data show that the centre of the smooth pillar tops has exhibited weak activation regions for "SERS"-like features while the nanorim like edges exhibited highly activated regions indicating those being an important factor for "SERS" active CNN neurons. The various generated geometries and topologies of the pillars included the flattop with no rim, equivalent nanorim surrounding roughened or smooth centres, with various rim heights, H_{rim} ($0 \leq H_{rim} \leq 500$ nm) and one-sided nanorim. A standard analyte of benzenethiol (BT) was subsequently used to evaluate the SERS enhancement factor (EF) from each of these different patterns, acquired under identical experimental conditions (**Figure 4.3d** and **Appendix AS1** and **S2**). Establishing optimal top feature morphologies is used in the

subsequent design of the dedicated top electrode in conjunction with the controllable microlithographic EHD apparatus.

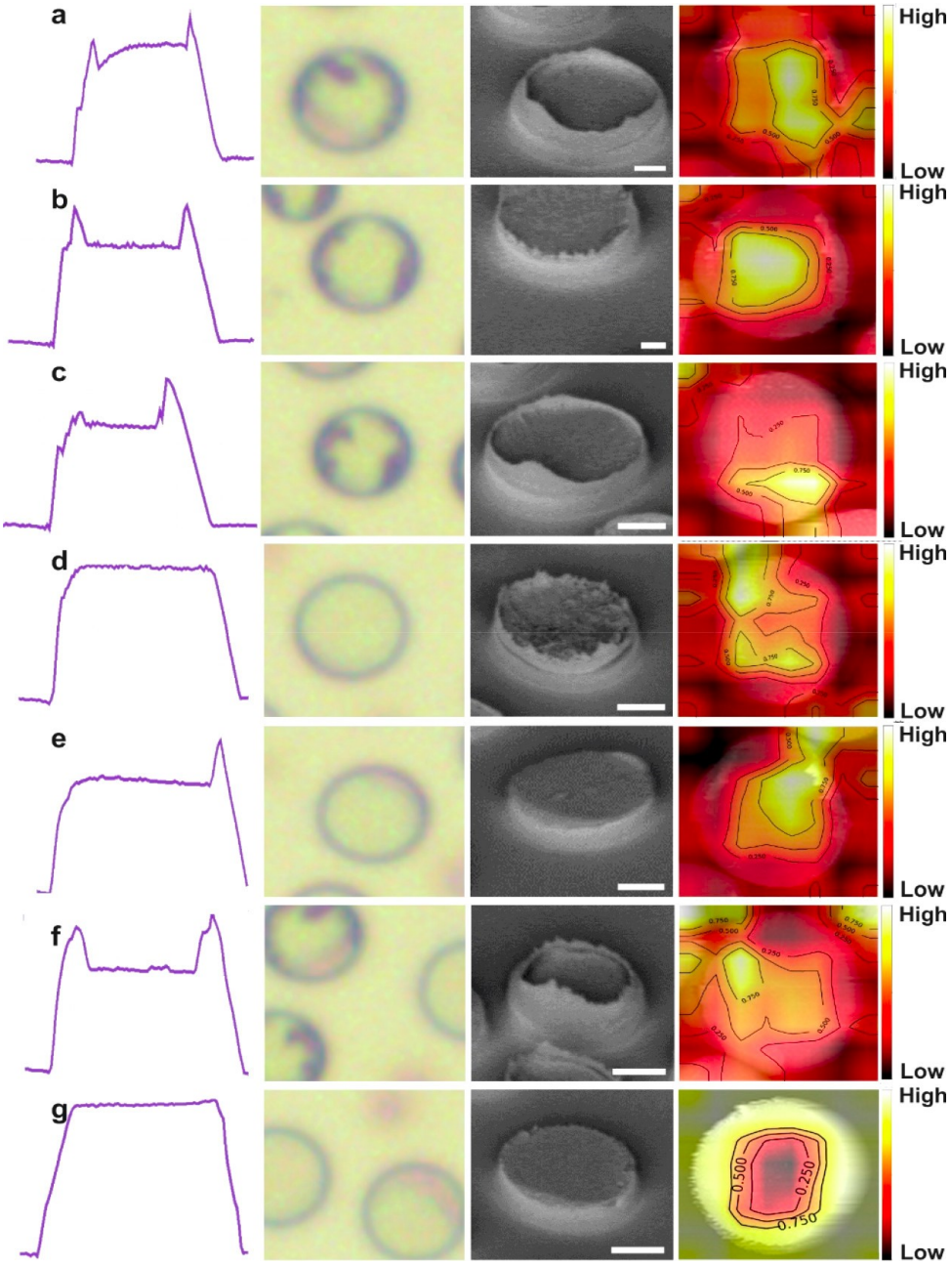


Figure 4.2. *Representative EHD-generated topological overview. Cross-section AFM profiles, top-view OM images, SEM micrographs, and CNN images of the EHD generated structures. Corresponding SERS enhancements are shown in* ***Figure 4.3d****.* ***(a)*** *Pillar with* H_{rim} *= 250 nm and a nanoroughened centre, EF = 9.0 × 10*^{6}*.* ***(b)*** *Pillar with an equivalent nanorim of* H_{rim} *= 300 nm and a roughened centre, EF = 1.1 × 10*^{7}*.* ***(c)*** *Pillar with a one-sided nanorim of* H_{rim} *= 270 nm and a roughened centre, EF = 8.8 × 10*^{6}*.* ***(d)*** *Pillar with nanoroughness throughout and no rim, EF = 6.4 × 10*^{6}*.* ***(e)*** *Smooth pillar with one-sided nanorim of* H_{rim} *= 200 nm, EF = 6.1 × 10*^{6}*.* ***(f)*** *Pillar with a nanorim of* H_{rim} *= 150 nm, EF = 6.7 × 10*^{6}*.* ***(g)*** *Smooth pillar top with no rim, EF = 5.8 × 10*^{6}*. Scale bar* $_{SEM}$*: 1 µm.*

Subsequent to identifying optimal candidate topologies for achieving increased EFs, we have investigated the impact of other structural parameters, all of which can be tuned during the EHD process, on the resulting signal enhancement. A systematic analysis reveals that the relative signal enhancement strongly depends on the aspect ratio (AR) of the pillars, highlighting a key factor for the design of optimized EHD-generated SERS-active substrates (**Figure 4.3a**). Furthermore, due to the slight misalignment of the capacitor plates in the EHD capacitor, there is a “wedge” geometry across the substrate with a variation ranging from 100nm to 1μm across for a 1cm wide sample, which subsequently leads to the lateral variation in the resultant pillar’s aspect ratio. This has been exploited to our advantage to generate a breadth of pillars with varying heights and diameters. In the fabricated surfaces it was found that, the ratio of the thickness of one film coating on the surface of the substrate to the diameter of the formed structure should be at least 0.6 and no more than 0.8 for the optimal SERS enhancement factor (**Figure 4.3a**). The height of the rim-like topology in the formed pillar was found to have an optimal range with $0.30 \leq H_{rim} \leq 0.37$μm exhibiting the highest achievable signal enhancement (**Figure 4.3b**).

The generated pillars of the bottom substrate, once adopted, are spaced from one another at a certain pitch distance, λ. During the EHD, the step of adjusting the gap between the top and bottom electrodes to a micrometer level of precision and further adjusting the gap to a nanometer level of precision can be used to control the spacing between adjacent pillars. This pitch, λ, can additionally be controlled by alternating the strength of the applied electric field. The influence of the pitch on the resulting SERS signal shown in the inset of **Figure 4.3c** demonstrates that the electric field excited on the

gold coated structures has a local minimum at λ = 1μm and otherwise increases considerably at λ > 1μm.

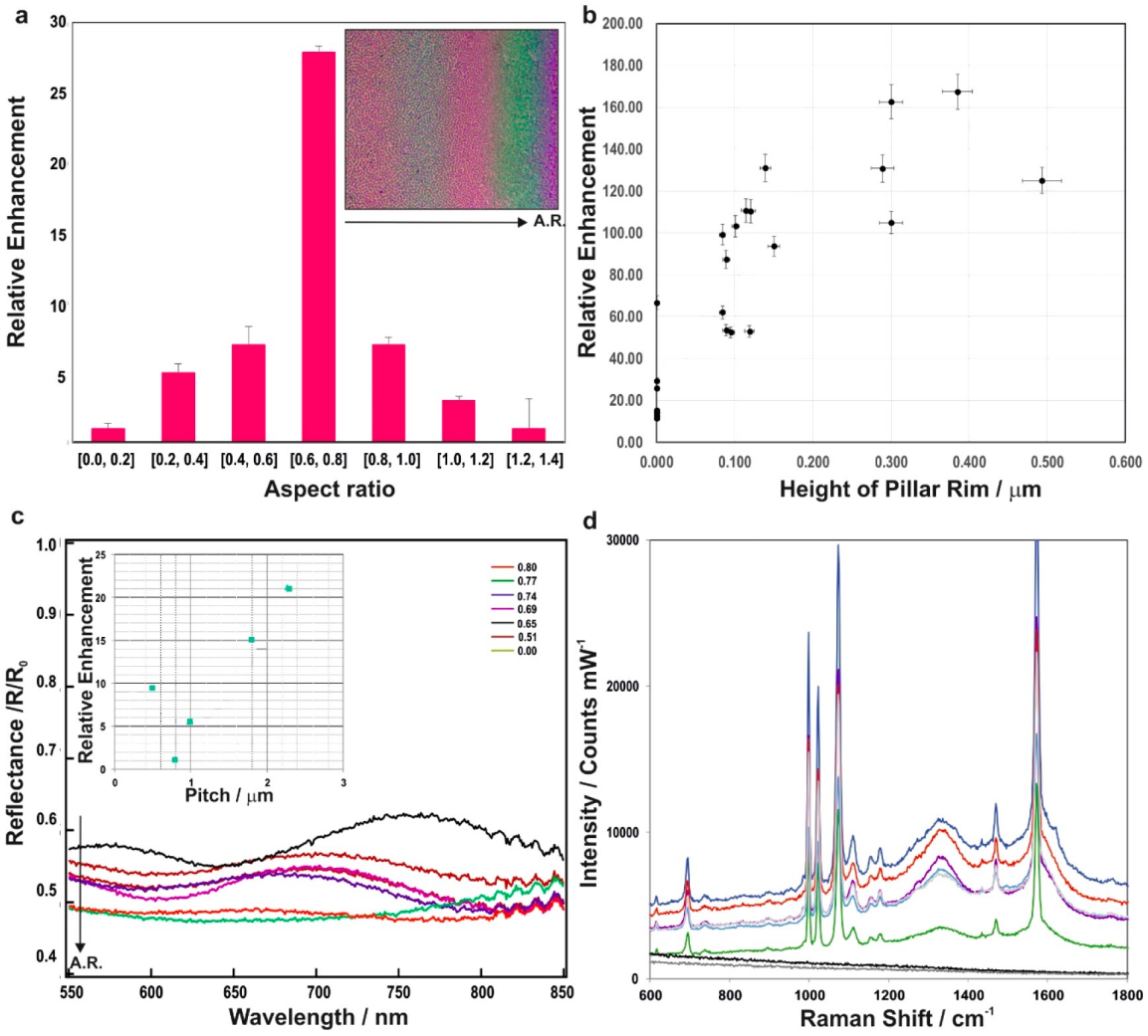


Figure 4.3. *Tunability of structural dimensions of the metallo-dielectric EHD fabricated SERS substrates for controllable and optimal plasmon excitations.* ***(a)*** *Relative SERS enhancement versus the AR identifies an optimal AR for signal enhancement in the range between 0.6 and 0.8. Wedge-geometry induced AR variation generates bands of colour, due to the film thickness redistribution on top of the underlying silicon wafer, corresponding to the various aspect ratios (inset). The arrows correspond to the variation in AR and the corresponding colour change across the sample.* ***(b)*** *Relative SERS enhancement of the benzenthiol's 1070 cm^{-1} peak with varying nanorim heights indicate an optimal range of H_{rim}~300 ± 37 nm.* ***(c)*** *Spectrally resolved reflectance spectra relative to the flat gold, R_0 as a function of the ARs. An increasing AR results in an increase in the extinction (R/R_0) due to the enhanced coupling of light into LPSRs, which are tuned by the pillar geometry. (inset) The variation of the enhancement as a function of the gap between the pillars at the excitation wavelength of 785 nm reveals another important parameter, tunability of which and its effect on*

strength of localized surface plasmon resonances and surface Raman enhancement allows the optimization of the fabricated substrates. ***(d)*** *Characteristic Raman peaks of benzenethiol at 1000, 1070, and 1575 cm^{-1} readily detected in the spectra acquired from the EHD SERS substrates showing the signal enhancement variation with the pillar's topological features shown in* ***Figure 4.2****.*

Furthermore, depending on the pillar's interstructural pitch, patterned arrays can exhibit SERS enhancement either on top of each individual structure (**Figure 4.4a**), at their perimeter (**Figure 4.4b**) or in the area between the structures (**Figure 4.4c**). The spacing of the pillars is thus also of an importance to the localization of the enhancement and the "correct" pitch ($1.2 \leq \lambda \leq 2.5\mu m$) can assist in enabling each pillar to act as an individual detection centre.

Dependence of reflectance with the wavelength reveals an increase in intensity with an increasing aspect ratio up to optimal dimensions at AR = 0.80 ± 0.12, variation that mainly arises from coupling of light into plasmon resonances, which are tuned by the pillar dimension, and therefore, the relative SERS enhancement is correlated with the increased strength of plasmonic coupling into the structure. Experimentally measured reflectance across pillars, R/R_0 with laterally varying ARs, exhibits increased extinction with an increasing aspect ratio, with a nearly 30 percent increase that occurs in extinction when the AR varies from 0.5 to 0.8 (**Figure 4.3c**).

Evaluation of the SERS signal evolving from each of the representative topological morphologies in **Figure 4.2** demonstrates variation in the enhancement factor of up to 40% between the highest and the lowest enhancing pillars. The absolute EF for the optimal pillar with an equivalent nanorim of H_{rim} = 300 nm and roughened centre was found to be $1.1 \times 10^7 \pm 1.7 \times 10^6$ in contrast to the smallest enhancement achieved from the smooth pillar with no rim with an EF of $5.8 \times 10^6 \pm 1.2 \times 10^6$ using a 785 nm excitation

laser. It is plausible that the EHD structures with optimal enhancing AR and nanorim like topology with a nanorough centre create a coupling effect combining the plasmon and whispering-gallery resonance for trapping and enhancing the signal. Such optical “nanocavities” could be confining light to smaller volumes by resonant photon circulation and thus yielding an improved SERS enhancement. For the subsequent reproducible pattern design, a pillar with an equivalent nanorim and roughened centre was found to be a favourable topology with the optimal feature sizes of these pillars identified to preferably have an AR ~0.7–0.8, rim height of 300nm, and $1.2 \leq \lambda \leq 2.5\mu m$.

Establishment of the EHD structural factors for optimal SERS enhancement has subsequently guided the fabrication of reproducible and controllable three-dimensional SERS substrates via designing and engineering a novel high-precision lithographic rig in conjunction with the dedicated top electrode. The developed microlithographic setup combines micromanipulator and piezoactuators for the finest adjustments to enable parallel capacitor-like EHD patterning for consistent fabrication of submicron pillars (**Figure 4.4f**). It is designed to allow for several degrees of freedom of movement and well-aligned parallel positioning of the top and bottom electrodes, to yield the integrity and accuracy of the intercapacitor distance down to the submicron scale, for the fabrication of highly reproducible SERS-active substrates with the ideal structural topological morphologies, aspect ratios, and pitches. The controllable microlithographic EHD apparatus comprises a support for a bottom substrate that has a film coating on the top surface thereof and a second one facing the bottom, supporting a second top substrate with a controllable gap between the surfaces. The base of it is made of aluminium. Extending upward from the base is a pillar comprised of a rectangular aluminium pillar screwed to the base. Attached through an opening in the base is a micrometer, which has a nonrotating head, with a copper block mounted onto its spindle. A bottom substrate to be

patterned is mounted on the copper block. A spring clamp (a thin beryllium copper spring clamp) is used to hold the bottom substrate in place on the copper block, also ensuring electrical connection of the bottom substrate to ground. A spar that extends from the pillar supports a top substrate assembly suspended over the bottom substrate. At the end of the spar a block of insulating polyether ether ketone (PEEK) is attached and then attached to that is a copper block with a spring clamp to which the top electrode/substrate is placed. The assembly includes a support arm underneath which is mounted a piezo-actuator, an upper block, a second piezo-actuator, a lower block, and the top substrate. The top substrate is clamped to the assembly using a beryllium copper spring clamp. The beryllium copper spring clamps are thinner than the height of the silicon substrates and hold the substrates at their edges to allow a bias voltage to be applied to the top and bottom substrates. A first gap modification enables adjustment to a micron scale precision, *via* adjusting a micrometer screw gauge, with the second gap adjustment provided to allow adjusting the gap to the nanometer level (via piezo actuators) of precision, yielding a significant improvement in the reproducibility and thus the enhancement factor from the surface. The gap during the patterning is monitored by measuring the current flowing between the two electrodes using a microammeter. Subsequently, a connection is provided to a source of electrical power for applying an electric field across the gap between the facing surfaces of the first and second substrates (**Figure 4.4g**). A sensor is concurrently used to sense a *finite* current flowing through the first and second substrates, and the whole apparatus is homogeneously heated to an elevated temperature to enable the thin film's fluidity. This further ensures that there is no temperature gradient around the substrates as well as allowing for more accurate control of the gap between the substrates.

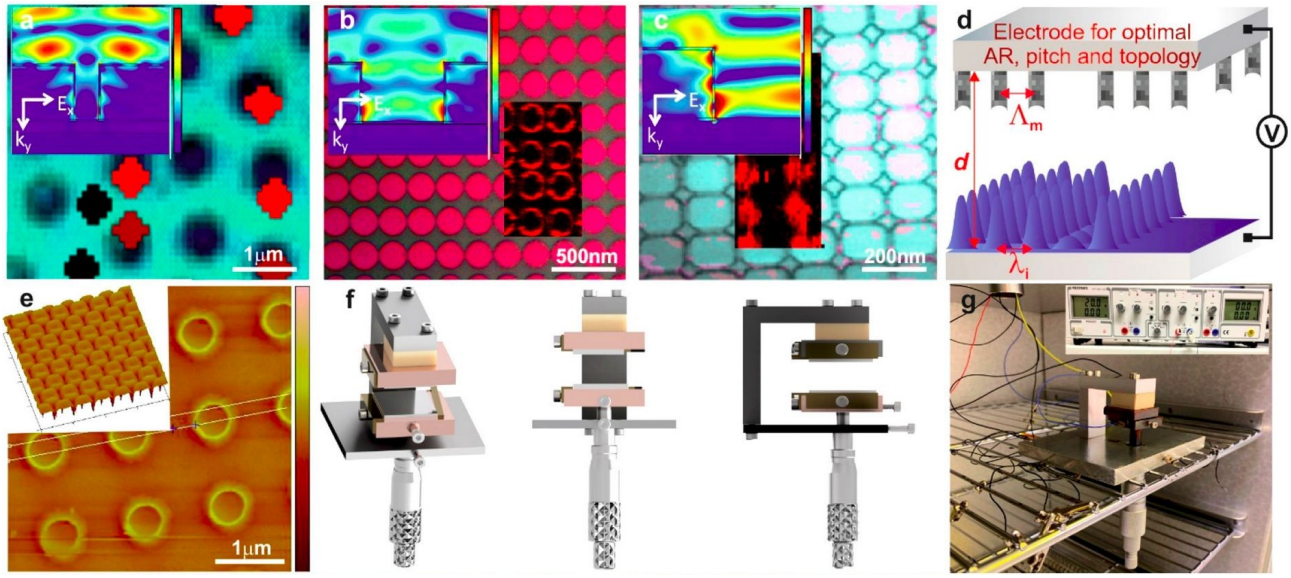


Figure 4.4. *Engineering the EHD apparatus and reproducibly controllable structures. (a–c) Optical microscopy images and the overlaid Raman maps with the corresponding 2D near-field simulations of the E_f distribution (insets) of the SERS enhancement occurring* ***(a)*** *exclusively on top of each pillar,* ***(b)*** *at the perimeter of the pillar, and* ***(c)*** *between the pillars. Schematics* ***(d)*** *and AFM 3D images of the fabricated* ***(e)*** *top electrode with carefully designed parameters including the interelectrode spacing, AR, pitch, and the interstructural spacing of the protrusions, Λ_m.* ***(f)*** *3D schematic representation of the controllable EHD rig.* ***(g)*** *Photograph of the EHD prototype apparatus in the heating chamber.*

While the surface of the upper electrode for the study and optimization of the EHD structural parameters for SERS was a flat (planar) surface, in the reproducible microlithographic setup, the surface of the top electrode comprises a dedicated, carefully designed templated pattern of structures that protrude toward the surface of the bottom substrate. We have established that the highest EF of gold covered SERS active pillars is achieved when the colour of the pillars is blue or green, corresponding to a pillar aspect ratio of 0.7–0.8, as shown in **Figure 4.2a, inset**. The design and fabrication of the dedicated top electrode's structures was thus dictated by the established optimal parameters of the EHD yielding the formation of high-fidelity positive replicas of the imposed morphology and fabrication of the structures in a more controllable and repeatable manner (**Figure 4.4d, e**). The master electrode is of a high structural integrity and durability and can be used to consistently produce structures of interest from low-cost materials using the EHD to guide the pattern formation and, therefore, enabling high-

reproducibility, cost-efficiency, and scalability of the generated nanosurfaces, rendering this technique even more technologically appealing. The uniquely designed top electrode provides the ultimate dimensions for high enhancement but also enables the creation of rimmed nanocavities and is comprised of downward protruding pillars with AR of 0.8, pitch of 1.2 μm, and H_{rim} = 300 nm nanorim on their perimeter (**Figure 4.4d**). Assembling this top electrode within the microlithographic EHD apparatus (**Figure 4.4g**) and applying a voltage across the two plates yields the generation of a heterogeneous electric field. In this case, the liquefied polymer film is drawn toward the protrusions of the top electrode, where the electrostatic pressure is the highest and the destabilization process is the fastest, faithfully reproducing the imposed patterns as shown in **Figure 4.5a-c**. Guiding the uniform structures fabrication, the setup ensures that the protrusions match each other and the height of a given pillar substantially matches the height of another pillar. The AFM images show the reproducibly fabricated morphologies comprised of pillars with the optimized dimensions and topology spanning SERS substrate areas of 200 × 200 μm^2 with high fidelity patterns. Upon examining 10 arbitrary capacities across each substrate with the monolayer of BT as the SERS probe, the EHD fabricated SERS surfaces exhibited reliable signal intensities and substrate reproducibility. A repeatable SERS response was obtained from the pillars with average standard deviation of less than 3.0% in the framework of one sample (**Figure 4.5d, inset** and **Appendix AS3**) and between the different samples (height of the bars in **Figure 4.5d, inset** and **Appendix AS3**) for the 1070 cm^{-1} Raman peak comparable to the state-of-the-art SERS substrates. [25,26] Uniform SERS signal was also obtained across several substrates with the EF values being narrowly distributed around the average of $1.1 \times 10^7 \pm 4.3 \times 10^5$ (**Appendix AS2b**). For the 0.95 confidence interval data it is therefore, expected that the difference in EF as measured between the different areas to lie between (9.4×10^6, 1.1×10^7) for 95% of the future measurements.

The experimentally calculated reproducibility coefficient indicates that the absolute difference between any three subsequent EF obtained on a particular substrate are estimated to be no greater than 8.55% on 95% of occasions. The wavelengths of the LSPR and the strength of the electric field are highly sensitive to the dimensions (*e.g.*, AR, diameter, pitch), the shape (the inherent nanoroughness in the central parts of the structures enclosed by the surrounding nanorim) and coupling modes of the nanostructures. The controlled pitch combined with the consistent nanorim topology (**Figure 4.2** and **Appendix AS1**) renders the adjacent pillars not plasmonically coupled to each other, generating a strong localization on top of the structures (with the optical enhancement on the surface mainly due to red-shifted apex plasmons) and consequently, high fields and SERS enhancements, which arise almost exclusively from a single structure. The EF and sensitivity achieved with these nanomorphologies is of a similar magnitude to both the commercially available as well as the various SERS substrates demonstrated in the literature while exhibiting an improved signal reproducibility and stability. [16,17,27-40] EHD-SERS substrates retain their structural integrity over the period of several months with no agglomeration or pattern distortion discovered. The substrates show excellent stability throughout the spectral collection period of 3 months with an average relative standard deviation of 3.9% (**Appendix AS4**).

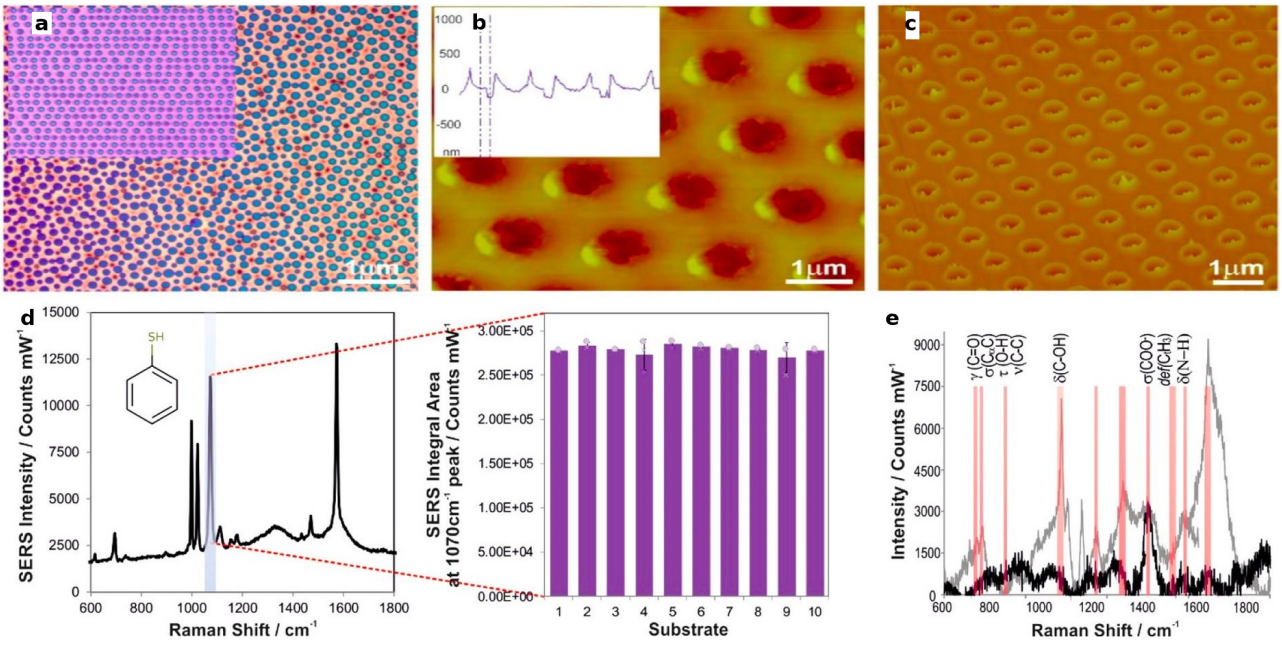


Figure 4.5. (a) *OM image of pillars formed under the reproducible EHD microlithography using planar top electrode, generating a homogeneous electric field throughout the capacitor, and (inset) pillars forming using lithographically predefined top electrode, introducing a heterogeneous electric field. AFM phase* ***(b)****, with the corresponding 3D profile, and height* ***(c)*** *images of the high-fidelity EHD-replicated patterns.* ***(d)*** *Representative SERS spectrum of BT on a uniform substrate for the highlighted 1070 cm^{-1} peak, at an excitation of 785 nm laser, demonstrating reliable signal and substrate reproducibility* ***(inset)*** *of the SERS intensities on the EHD generated arrays from 10 randomly selected positions on the gold-coated EHD fabricated pillars.* ***(e)*** *Spike-and-recovery bioassay differentiates the characteristic peaks of biomarker in an artificial plasma with known concentrations (50pg/mL). SERS spectra of the spiked (gray) versus normal, nonspiked (black); analyte was spiked into the test sample matrix, and its response was recovered in the assay by comparison to an identical spike in the standard diluent with 95.5 ± 3.4% recovery of the biomarker.*

To demonstrate the diagnostic potential of the developed EHD SERS active substrates, we have assembled a proof-of-concept spike-and-recovery assay comprised of *N*-acetylaspartate spiked artificial blood plasma deposited on the fabricated substrates at a concentration of 50 picograms per milliliter. *N*-acetylaspartate is a biomarker indicative of acute, early-stage traumatic brain injury and is exclusively localized in neurons, reflecting their functional status (as well as of axons) in the brain. [41-43] It is considered a marker for neuronal health and viability, highly specific to the central nervous system, not released by other organs and thus, is a very sensitive measure of neuronal compromise. This

biomarker is thus of importance for development of early stage triaging point-of-care diagnostic technologies for traumatic brain injuries. The EHD-SERS substrates were used to establish the specific fingerprint spectra (**Figure 4.5e** and **Appendix AS5** and **S6**) and to quantitatively determine the lowest detectable concentration of the target biomarker. The characteristic peaks of the analytes were monitored and a linear trend was observed with a good correlation (R^2 = 0.9910) between biomarker concentrations (100nM to 1pM) and SERS intensity (**Appendix AS8**). No signal enhancement was observed on the flat parts on the substrate between the pillars (**Figure 4.4a, inset** and **Appendix AS7**). SERS spectra for detection of the *N*-acetylaspartate show successful differentiation of the biomarker's fingerprints,[42,44] spiked at known concentrations (**Figure 4.5e**) in the artificial plasma (top) versus the control sample (bottom), indicating the potential for detection and differentiation from clinical blood and plasma samples for future development of biodiagnostic SERS assays.

4.4 Conclusions

A tailored approach was demonstrated for the manufacture of surface-enhanced Raman scattering active nanoarchitectures via the EHD lithography. The plasmon dependent phenomena of the fabricated substrates can be straightforwardly adjusted via the lithographic key-factors, allowing the development of reliably engineered and cost-effective substrates with reproducible and high SERS enhancement. SERS signal enhancement of these substrates is governed by not only the pillar's nanotopology but also by the dimensions of the fabricated structures, parametrized by the ratio of the height to diameter as well as the interstructural pitch. Thus, the controllable aspect ratio via the EHD and its effect on strength of the enhancement factor further enables the possibility of fine-tuning

the substrates responses for different laser excitation wavelengths. The ability to carefully design and fabricate highly reproducible optimized SERS active nanoarrays enables the control of localized plasmonic fields, laying a platform for developing portable lab-on-chip devices and sensors to detect and analyse a variety of biochemicals (**Figure S9**) pertinent for medical, environmental, and forensic applications. Intriguingly, due to the individually enhancing optimized structural units, each topological structure in the array could further be functionalised with specific receptors for a range of target molecules for rapid, multiplexed, highly specific and sensitive detection. These will be, in turn, useful for the development of rapid-response, addressable miniaturized sensors, which are selective for targeted chemical and biological molecules.

4.5 Methods

4.5.1 Materials

Polystyrene (PS, M_w 100 kg/mol, $M_w/M_n < 1.07$) was used as a polymer for the patterning (purchased from Polymer Standards Service GmbH) and toluene was used as a main solvent in this process (Fisher Scientific). Highly polished *p*-doped silicon (Si) wafers with ⟨100⟩ crystal orientation (Wafernet, GMBH, Eching, Germany) were used during the EHD as bottom electrodes (*X-lith eXtreme* Lithography). 3 wt/wt % PS solutions were filtered through a Teflon filter with a pore diameter of 100 nm. BT (Analytical Standard, Sigma-Aldrich) was used without further purification. Silver paint (Electrodag 1415M) used to contact the electrodes. Standard biomarkers (S one-hundred calcium-binding protein B and *N*-acetyl aspartate) for individual fingerprint signature studies were acquired from the

Human ELISA Kit (Abnova) and Sigma-Aldrich (Saint Louis, MO). The 99.999% purity gold target for sputtering was purchased from Kurt J. Lesker.

4.5.2 Electrohydrodynamic Surface Enhanced Raman Scattering Substrates Fabrication

PS was used as the polymer for thin film deposition and patterning during the EHD, and highly polished p-doped silicon wafers were used as bottom and top electrodes in the capacitor setup. Silicon wafers were thoroughly cleaned in a “piranha” solution consisting of sulphuric acid and hydrogen peroxide (3:1, accordingly) followed by rinsing with deionized water and drying under nitrogen flow. An initial solution of PS in toluene 3% w/w was subsequently spin-coated onto the freshly cleaned 1.0 × 1.0cm^2 silicon electrodes and these were then electrically connected to the external voltage supply and homogeneously heated to 175°C. The whole EHD patterning rig was enclosed inside a high-precision performance convection oven ensuring optimal temperature uniformity throughout the chamber. The temperature in the oven was thermostatically controlled (a thermometer was placed in an oven which was set at the desired operating temperature. The temperature reading of the thermometer was compared to the oven thermostat reading). Furthermore, a sensor was concurrently used to sense a *finite* current flowing through the first and second substrates and the whole apparatus was homogeneously heated to an elevated temperature to enable the thin film’s fluidity. This ensured that there was no temperature gradient around the substrates. Once the EHD process has been completed, the capacitor setup was cooled down to room temperature to quench the formed structures and the external voltage supply was disconnected. To facilitate the smooth release of the patterned surfaces, the substrate was grafted with an octadecyltrichlorosilane self-assembled monolayer. The patterned substrates were finally

coated with a gold nanolayer with an average thickness of 25.0 ± 2.7nm (Emitech sputter-coater) using a DC argon plasma and a gold target. The gold deposition was performed onto all the EHD-generated substrates as well as the polymer film as reference. Subsequently, benzenethiol monolayer was assembled on the gold-coated surfaces from ethanolic solution at varying concentrations.

4.5.3 Fabrication of Top Electrode

Focused ion beam (FIB) process was used to fabricate the dedicated EHD electrode. For this purpose, initially silicon wafers were cleaved and then cleaned using a carbon dioxide spray system. Subsequently, the electrode was produced using an FEI Helios Nanolab 650 dual beam scanning microscope, equipped with an electron beam for imaging and a separate beam of gallium ions at 52° to the electron beam. The gallium ions are accelerated up to 30kV to provide them with a high energy and the generated beam of ions is focused to a spot by several electrostatic lenses and then scanned over the object by scan plates. The gallium ion beam was used to sputter away material to create the required patterns. The associated patterning system of the Helious (Nanobuilder) was further employed by using a bitmap image to control the ion beam to recreate the pattern from the designed image. Nanobuilder was subsequently used to shift the stage and repeat the process to create large patterned areas.

4.5.4 Raman Spectroscopy Measurements

SERS measurements were carried out using InVia Qontor spectrometer confocal Raman microscope (Renishaw) equipped with 785nm excitation laser, which was adjusted for optimal throughput, fluorescence control, and sensitivity. The spectra were typically acquired at 10s exposure time and a laser power of 1–10mW with a 785nm laser to avoid photochemical effects in the SERS spectra, sample damage, or degradation. The equation of the linear fit obtained from the dilution studies quantitatively calculating the limit of detection as 3 times the standard deviation of the blank, derived from the gradient of the straight line, where σ is the standard deviation of the response and *S* is the slope of the calibration curve. SERS maps were generated in a Stream-line mode scan with an exposure time of 10s and power of 5mW at 785nm. A 50× objective with a numerical aperture of 0.75 was used for SERS measurements over a range of 500 to 2500cm^{-1} relative to the excitation Raman shift. Optical measurements were carried out with a specially adapted research grade microscope (Leica DM 2700M) equipped with an incoherent white light source, allowing confocal measurements with 2.0μm depth resolution. The spectra were normalized with respect to those recorded on flat gold or gold-covered flat polymer film surfaces. An intelligent-fitting filter was applied for baseline subtraction.

4.5.5 Calculation of the Surface Enhanced Raman Scattering Enhancing Factor

Numerical modelling was carried out using a finite element solver COMSOL v4.3. The enhancement factor calculation is reproduced from ref [42]. In brief, SERS enhancement

factor was calculated by comparing the intensities of the non-enhanced Raman scattering peak (I_{Raman}) at 1 070cm^{-1} of pure BT obtained by focusing the laser into a quartz cell and the corresponding SERS signals (I_{SERS}) obtained from the SERS substrates. The detection volume of the solution-phase benzenethiol sample, V_f, was calculated using the following relation: V_f = (depth of focus) × (focus area) = $(1.4n\lambda/NA^2) \times \pi(0.4\lambda/2NA)^2$. The surface density of the adsorbed BT molecules on the structured surface was taken as ρ_s = 3.3molecules/nm^2, and the enhanced area, A, was defined as the diffraction limited spot size ($\pi(0.4\lambda/2NA)^2$), with NA, numerical aperture; n, refractive index; λ, wavelength; ρ_v, surface density of molecules in the detection volume. The EF was calculated using the relation: $EF = [I_{SERS}/(\rho_s A)]/[I_{Raman}/(\rho_v V_f)]$.

4.5.6 Atomic Force Microscopy

A JPK NanoWizard II atomic force microscope was used to thoroughly characterize the surfaces' topography. The AFM measurements were performed using tapping mode via an intermittent contact of the tip with the sample in ambient conditions. NCHV-A cantilevers with a resonance frequency of 320kHz and stiffness of 42Nm^{-1} were used. Height and phase images were analysed with Gwyddion's software (version 2.55). The AFM measurements were used to acquire important parameters including the film thickness h, the interelectrode spacing d, the characteristic pitch λ, and the AR of the generated structures. A total of 13 areas of 50 × 50μm^2 were scanned throughout each sample. The area was chosen according to the different coloured regions of the EHD samples. The heights ranging from 200 to 700nm spanned a lateral distance of approximately 700μm. Each scan was performed at 1024 × 1024 pixel quality at approximately 3s/line raster rate,

for a total of 1h/region. The set point was adjusted for each region to compensate for the pillar height changes while the rest of the structural parameters were kept constant. Ten pillars were chosen per scanning region, totalling in 130 pillars characterized. The pillars were chosen as representative examples of the different topologies obtained during the EHD lithography.

4.5.7 Optical microscopy

For the concurrent tracking of the AFM images, a corresponding optical microscopic image of the pillars was also acquired. These optical images guided the navigation of the AFM measurements between different regions (Leica DM2700 M). Pattern formation in thin polymer films was qualitatively analysed using bright-field reflection microscopy (Olympus Optical Microscopes GX61, GX2, and BX40). The reflection of white light from the sample enabled resolving the submicron-scale features. The evolving patterns were also observed by *in situ* monitoring of the EHD pattern formation and replication in thin films using a digital camera and imaging software (Carl-Zeiss VisioCam, AxioVision).

4.5.8 Scanning Electron Microscopy

Samples for SEM imaging were prepared by placing a post-experiment, disassembled substrate with the generated patterns on an inclined or cross-sectional holder to enable the imaging of the top and cross-sectional views. Scanning electron micrographs were acquired using a thermally assisted field emission source scanning electron microscope (FEI helios dual beam) with a lateral resolution of 2–5nm with a typical acceleration voltage of 1–10kV.

4.5.9 Deep Neural Network Image Analysis

For the data gathering stage, the studied individual pillars were in the range of heights varying between 200 and 700nm and diameters of 0.5–5μm. All the images were acquired as a top view and height differences are given by a different intensity level on the greyscale. The ability of the pillar to enhance the signal was either labelled as “No enhancement”, meaning there is no or negligible Raman signal enhancement on the pillar and as “SERS” where there is a noticeable signal enhancement from the pillar. Each of the chosen pillars was then characterized via the AFM and the corresponding surface profiles were also obtained. These constitute the initial data set for the CNN. The subsequent step of the data pre-processing included ensuring that all the images are on the identical greyscale and same pixel size. For the data augmentation, the data set was increased by conducting rotations and inversions of the acquired images while keeping the labels fixed. This is possible because the Raman measurements are the same for the pillar independent of the rotation. Following the data augmentation, a complete training data set is generated. In the next step, a VGG16 pre-trained convnet was used, which was later modified with a fully connected classifier by changing the output layer to “SERS” and “No enhancement” and subsequently, the last block of neuron layers prior to the classification was fine-tuned. After training the modified model with the data set, the model was tested with the data set, where the testing data included pillar images that were never “seen” by the model. This information identifies the features that are most important for the SERS-active pillar via the class activation map (CAM) by producing a heat map image of the neuron activation over the input test image for a specific label, which results in how intensely the input image activates a given label. The overlaid heat map has consequently a high activity on features of the input image that are “SERS”-like features.

4.6 References

[1] Han, X. X.; Ozaki, Y.; Zhao, B. Label-Free Detection in Biological Applications of Surface-Enhanced Raman Scattering. TrAC, Trends Anal. Chem. **2012**, 38, 67−78.

[2] Han, X. X.; Zhao, B.; Ozaki, Y. Surface-enhanced Raman Scattering for Protein Detection. Anal. Bioanal. Chem. **2009**, 394 (7), 1719−1727.

[3] Kahraman, M.; Mullen, E. R.; Korkmaz, A.; Wachsmann-Hogiu, S. Fundamentals and Applications of SERS-based Bioanalytical Sensing. Nanophotonics **2017**, 6, 831−852.

[4] McQueenie, R.; Stevenson, R.; Benson, R.; MacRitchie, N.; McInnes, I.; Maffia, P.; Faulds, K.; Graham, D.; Brewer, J.; Garside, P. Detection of Inflammation in Vivo by Surface-Enhanced Raman Scattering Provides Higher Sensitivity Than Conventional Fluorescence Imaging. Anal. Chem. **2012**, 84 (14), 5968−5975.

[5] Moore, T.; Moody, A.; Payne, T.; Sarabia, G.; Daniel, A.; Sharma, B. In Vitro and In Vivo SERS Biosensing for Disease Diagnosis. Biosensors **2018**, 8 (2), 46−71.

[6] Wang, R.; Chon, H.; Lee, S.; Ko, J.; Hwang, J.; Choi, N.; Cheng, Z.; Wang, X.; Choo, J. Biomedical Applications of Surface-Enhanced Raman Scattering Spectroscopy. In Frontiers and Advances in Molecular Spectroscopy; Laane, J., Ed.; Elsevier, **2018**; Chapter 9, pp 307−326.

[7] Wang, Y.; Kang, S.; Khan, A.; Ruttner, G.; Leigh, S. Y.; Murray, M.; Abeytunge, S.; Peterson, G.; Rajadhyaksha, M.; Dintzis, S.; Javid, S.; Liu, J. T. C. Quantitative Molecular Phenotyping with Topically Applied SERS Nanoparticles for Intraoperative Guidance of Breast Cancer Lumpectomy. Sci. Rep. **2016**, 6, 21242−21254.

[8] Ackermann, L.; Althammer, A.; Born, R. Catalytic Arylation Reactions by C−H Bond Activation with Aryl Tosylates. Angew. Chem., Int. Ed. **2006**, 45 (16), 2619−2622.

[9] Kho, K. W.; Qing, K. Z. M.; Shen, Z. X.; Ahmad, I. B.; Lim, S. S. C.; Mhaisalkar, S.; White, T. J.; Watt, F.; Soo, K. C.; Olivo, M. Polymer-Based Microfluidics with Surface-Enhanced Raman-Spectroscopy-Active Periodic Metal Nanostructures for Biofluid Analysis. J. Biomed. Opt. **2008**, 13 (5), 054026.

[10] Lim, C.; Hong, J.; Chung, B. G.; deMello, A. J.; Choo, J., Optofluidic platforms based on surface-enhanced Raman scattering. *Analyst* **2010,** *135* (5), 837-844.

[11] Pennathur, S.; Fygenson, D., Improving fluorescence detection in lab on chip devices. *Lab on a Chip* **2008,** *8* (5), 649-652.

[12] Piorek, B. D.; Lee, S. J.; Santiago, J. G.; Moskovits, M.; Banerjee, S.; Meinhart, C. D., Free-surface microfluidic control of surface-enhanced Raman spectroscopy for the optimized detection of airborne molecules. *Proceedings of the National Academy of Sciences* **2007,** *104* (48), 18898-18901.

[13] Kahl, M.; Voges, E.; Kostrewa, S.; Viets, C.; Hill, W., Periodically structured metallic substrates for SERS. *Sensors and Actuators B: Chemical* **1998,** *51* (1), 285-291.

[14] Qian, X.; Zhou, X.; Nie, S., Surface-Enhanced Raman Nanoparticle Beacons Based on Bioconjugated Gold Nanocrystals and Long Range Plasmonic Coupling. *Journal of the American Chemical Society* **2008,** *130* (45), 14934-14935.

[15] Tantra, R.; Brown, R. J. C.; Milton, M. J. T., Strategy to improve the reproducibility of colloidal SERS. *Journal of Raman Spectroscopy* **2007,** *38* (11), 1469-1479.

[16] Ngo, Y. H.; Li, D.; Simon, G. P.; Garnier, G. Gold Nanoparticle−Paper as a Three-Dimensional Surface Enhanced Raman Scattering Substrate. Langmuir **2012**, 28 (23), 8782−8790.

[17] Santoro, G.; Yu, S.; Schwartzkopf, M.; Zhang, P.; Koyiloth Vayalil, S.; Risch, J. F. H.; Rubhausen, M. A.; Hernandez, M.; Domingo, C.; Roth, S. V. Silver Substrates for Surface Enhanced Raman Scattering: Correlation Between Nanostructure and Raman Scattering Enhancement. Appl. Phys. Lett. **2014**, 104 (24), 243107− 243112.

[18] Wang, A. X.; Kong, X. Review of Recent Progress of Plasmonic Materials and Nano-Structures for Surface-Enhanced Raman Scattering. Materials **2015**, 8 (6), 3024−3052.

[19] Busà, C.; Rickard, J. J. S.; Chun, E.; Chong, Y.; Navaratnam, V.; Goldberg Oppenheimer, P. Tunable Superapolar Lotus-to-Rose hierarchical Nanosurfaces via Vertical Carbon Nanotubes Driven Electrohydrodynamic Lithography. Nanoscale **2017**, 9 (4), 1625− 1636.

[20] Goldberg Oppenheimer, P.; Eder, D.; Steiner, U. Carbon Nanotube Alignment via Electrohydrodynamic Patterning of Nano-composites. Adv. Funct. Mater. **2011**, 21 (10), 1895−1901.

[21] Goldberg Oppenheimer, P.; Hutter, T.; Chen, B.; Robertson, J.; Hofmann, S.; Mahajan, S., Optimized Vertical Carbon Nanotube Forests for Multiplex Surface-Enhanced Raman Scattering Detection. *The Journal of Physical Chemistry Letters* **2012,** *3* (23), 3486-3492.

[22] Goldberg Oppenheimer, P.; Steiner, U. Rapid Electro- hydrodynamic Lithography Using Low-Viscosity Polymers. Small **2010**, 6 (11), 1248−1254.

[23] Rickard, J. J. S.; Farrer, I.; Goldberg Oppenheimer, P. Tunable Nanopatterning of Conductive Polymers via Electrohydrodynamic Lithography. ACS Nano **2016**, 10 (3), 3865−3870.

[24] Goldberg Oppenheimer, P.; Mahajan, S.; Steiner, U. Hierarchical Electrohydrodynamic Structures for Surface-Enhanced Raman Scattering. Adv. Mater. **2012**, 24 (23), OP175−OP180.

[25] Huang, J.-A.; Zhao, Y.-Q.; Zhang, X.-J.; He, L.-F.; Wong, T.-L.; Chui, Y.-S.; Zhang, W.-J.; Lee, S.-T., Ordered Ag/Si Nanowires Array: Wide-Range Surface-Enhanced Raman Spectroscopy for Reproducible Biomolecule Detection. *Nano Letters* **2013,** *13* (11), 5039-5045.

[26] Liu, X.; Shao, Y.; Tang, Y.; Yao, K.-F. Highly Uniform and Reproducible Surface Enhanced Raman Scattering on Air-stable Metallic Glassy Nanowire Array. Sci. Rep. **2015**, 4 (1), 5835.

[27] Srichan, C.; Ekpanyapong, M.; Horprathum, M.; Eiamchai, P.; Nuntawong, N.; Phokharatkul, D.; Danvirutai, P.; Bohez, E.; Wisitsoraat, A.; Tuantranont, A. Highly-Sensitive Surface-Enhanced Raman Spectroscopy (SERS)-based Chemical Sensor Using 3D Graphene Foam Decorated with Silver Nanoparticles as SERS Substrate. Sci. Rep. **2016**, 6, 23733.

[28] Schmidt, M. S.; Hü bner, J.; Boisen, A. Large Area Fabrication of Leaning Silicon Nanopillars for Surface Enhanced Raman Spectroscopy. Adv. Mater. **2012**, 24 (10), OP11−OP18.

[29] Japan, S. URL="http://www.stjapan.de/en/shop/accessories-for-spectroscopy-microanalysis/raman-sers-substrate/", accessed: 29/05/2020.

[30] StellarNet Inc. URL="https://www.stellarnet.us/spectrometers-accessories/sers-substrates/", accessed: 27/05/2020.

[31] Analysers, R.-T. URL="http://sersitive.eu/shop/silver-substrate-5pcs-bundle/", accessed: 29/05/2020.

[32] Mesophotonics URL="http://www.mesophotonics.com/products/klarite.html", accessed: 29/05/2020.

[33] Jiao, Y.; Ryckman, J. D.; Ciesielski, P. N.; Escobar, C. A.; Jennings, G. K.; Weiss, S. M. Patterned Nanoporous Gold as an Effective SAB. Nanotechnology **2011**, 22 (29), 295302.

[34] Baltog, I.; Primeau, N.; Reinisch, R.; Coutaz, J. L. Surface Enhanced Raman Scattering on Silver Grating: Optimized Antenna-like Gain of the Stokes Signal of 10 4 . Appl. Phys. Lett. **1995**, 66 (10), 1187−1189.

[35] Ward, D. R.; Grady, N. K.; Levin, C. S.; Halas, N. J.; Wu, Y.; Nordlander, P.; Natelson, D. Electromigrated Nanoscale Gaps for Surface-Enhanced Raman Spectroscopy. Nano Lett. **2007**, 7 (5), 1396−1400.

[36] Taminiau, T. H.; Stefani, F. D.; Segerink, F. B.; van Hulst, N. F. Optical Antennas Direct Single-molecule Emission. Nat. Photonics **2008**, 2 (4), 234−237.

[37] Xu, X.; Kim, K.; Li, H.; Fan, D. L. Ordered arrays of Raman nanosensors for Ultrasensitive and Location Predictable Biochemical Detection. Adv. Mater. **2012**, 24 (40), 5457−5463.

[38] Zhu, W.; Wen, B.-Y.; Jie, L.-J.; Tian, X.-D.; Yang, Z.-L.; Radjenovic, P. M.; Luo, S.-Y.; Tian, Z.-Q.; Li, J.-F. Rapid and Low-Cost Quantitative Detection of Creatinine in Human Urine with a Portable Raman Spectrometer. Biosens. Bioelectron. **2020**, 154, 112067−112075.

[39] Cheng, H.-W.; Huan, S.-Y.; Wu, H.-L.; Shen, G.-L.; Yu, R.-Q. Surface-Enhanced Raman Spectroscopic Detection of a Bacteria Biomarker Using Gold Nanoparticle Immobilized Substrates. Anal. Chem. **2009**, 81 (24), 9902−9912.

[40] Szekeres, G. P.; Kneipp, J. SERS Probing of Proteins in Gold Nanoparticle Agglomerates. Front. Chem. **2019**, 7, 30.

[41] Patel, T. B.; Clark, J. B. Synthesis of N-acetyl-L-aspartate by Rat Brain Mitochondria and its Involvement in Mitochondrial/cytosolic Carbon Transport. Biochem. J. **1979**, 184 (3), 539−546.

[42] Rickard, J. J. S.; Di-Pietro, V.; Smith, D. J.; Davies, D. J.; Belli, A.; Oppenheimer, P. G. Rapid Optofluidic Detection of Biomarkers for Traumatic Brain Injury via Surface-enhanced Raman Spectroscopy. Nat. Biomed. Eng. **2020**, 4, 610.

[43] Shannon, R. J.; van der Heide, S.; Carter, E. L.; Jalloh, I.; Menon, D. K.; Hutchinson, P. J.; Carpenter, K. L. H. Extracellular N-Acetylaspartate in Human Traumatic Brain Injury. Journal of neurotrauma **2016**, 33 (4), 319−329.

[44] Socrates, G. Infrared and Raman Characteristic Group Frequencies: Tables and Charts, 3rd ed.; Wiley, **2004**; pp 1−366.

CHAPTER 5

Real-time determination of Surface-Enhanced Raman Scattering substrates via a Convolutional Neural Network approach

This work has been recently published as "***de Carvalho Gomes P.****, Crossman A., Massey E., Stanley Rickard J.J., Oppenheimer P.G. Real-time validation of Surface-Enhanced Raman Scattering substrates via convolutional neural network algorithm. Vol. 33, Informatics in Medicine Unlocked. Elsevier BV;* ***2022****. p. 101076. Available from: http://dx.doi.org/10.1016/j.imu.2022.101076*", and my contribution as the first author was the majority of the work. The whole project was designed by me and sectioned into two components for the master's thesis programmes at the University of Birmingham. Laboratory work was carried out by myself, who then supervised and solved the computer vision algorithm and real-time workflow.

5.1 Abstract

Many future sensors are likely to use surface-enhanced Raman scattering (SERS) substrates due to their high sensitivity on low concentrated analytes. The problem with using SERS substrates, for example as a medical biomarker qualitative test, is that the enhancement from different parts of the substrate can be inconsistent. We have thus implemented convolutional neural networks (CNN) to analyse SERS-active substrates produced by electrohydrodynamic lithography (EHD). For doing so, a computer programme was developed to identify regions of high and consistent SERS activity on optical microscope images of EHD patterns in real-time. The implemented script has two main components. The first part of the script optimises the CNN model, and the second part processes the input optical microscope image and outputs the predicted SERS-active regions overlaid onto the input optical microscope image using the model developed in the first part of the script. The results show that the best CNN for the first part of the script is pt-VGG16-ft, a pre-trained CNN with a fine-tuned last layer. The pt-VGG16-ft achieved

90% accuracy in classifying high-SERS active EHD pillars via an optical microscope image, allowing for real-time predictive tools for SERS substrates. The second part of the script generates a heatmap that provides feedback on the regions of high SERS activity and is superimposed on the optical image to aid in SERS spectral acquisition measurement. Finally, a proof-of-concept run of the script obtained an accuracy of 87.5% for a real-time SERS pillar measurement acquisition, showing that it is feasible to identify SERS structures by using a real-time CNN processing of an optical image.

5.2 Introduction

Surface-enhanced Raman scattering (SERS) is a phenomenon discovered by Martin Fleischmann and Patrick J. Hendra in 1974 [1], who identified enhanced Raman scattering on a rough metal surface. Since then, SERS technology has shown a multitude of applications in biology [2–4] and chemistry [5–7] in the sensing field and SERS-active substrates are critical for recording accurate and reproducible spectral fingerprints of a plethora of molecules. To get these high SERS signals, substrates need to be carefully crafted to have a high level of chemical stability, biocompatibility [8], and reproducible fabrication methods [9]. The most used SERS active materials include gold, silver, and copper; but gold is usually the choice for biocompatibility [8, 10], and silver is preferred for higher SERS signals [2]. The optimisation in the SERS field has led to many advances, particularly within the field of biosensors. Very promising applications can be achieved when complemented with microfluidics for a lab-on-a-chip device, where multiple analyses can be performed over the same substrate which expands the multiplexing possibilities of SERS [11,12]. SERS outputs very specific fingerprints that can uniquely identify biomolecules. SERS substrates can also be integrated with a portable Raman system,

thus becoming a portable sensor system [5,13,14]. Additionally, the acquisition time is almost instantaneous, and with new machine learning techniques [15–18], the Raman signal analysis can give diagnostic results immediately, enabling earlier diagnosis for many health conditions.

However, some obstacles must be overcome before SERS sensors fulfil their potential in the medical, biological, and chemical sectors. Scientists within the SERS field have identified the need for reproducible signal enhancement from SERS substrates to be consistent before SERS sensors can become widely used [19, 20]. These inconsistencies cause measurements on the same substrate to differ, consequently hampering the duplication of previous studies. A feasible solution to tackle these obstacles could be to use machine learning (ML) because it can complement SERS in many ways.

In this work, we apply the ML technique called convolutional neural networks (CNN). CNN have improved image processing mechanisms compared to other neural networks. Convolution layers learn local patterns from the input image and these learned patterns have translation invariant properties i.e., they can be recognised anywhere in an image, whereas in other neural networks like a densely connected network, would need to learn the patterns anew [21]. For this reason, CNN are more efficient at image processing than other neural networks, needing fewer training samples to learn general trends in the images. Moreover, CNN can learn spatial hierarchies of the patterns, which are mapped over to the different layers of the CNN. In the CNN hierarchy, the first layer learns small local patterns like edges, and the next layers will learn from the edges and so on, allowing a CNN to learn increasingly complex and abstract visual concepts automatically. CNN have been previously used to identify SERS pillars from atomic force microscope (AFM) images [11]. To do so, a pre-trained CNN was used and later fine-tuned to our differently classified (SERS/non-SERS) EHD structures. Furthermore, in the work by Gomes *et al.*[11]

the AFM images have better resolution than the optical images, thus, in the present work we aimed to find out if we could achieve similar results with these lower resolution optical images. Since optical images are easier and much quicker to acquire and do not need special training compared to the AFM, it will drastically reduce the experiment and data acquisition time. Thus, optimizing not only the SERS reproducibility but also its throughput. All features of the CNN used by Gomes *et al.* [11] are extended for optical images with further implementation for real-time SERS detection, thus automating the selection of SERS structures for a fast and consistent measurement.

The structures we analysed were optical microscope images of gold-coated polymer pillars produced by electrohydrodynamic (EHD) lithography [22, 23] and the goal was to determine whether the pillars are SERS active or non-SERS active. To test the hypothesis, five substrate samples were made using EHD lithography. These were analysed and used to optimise a CNN model that can distinguish SERS-active pillars from non-SERS active pillars and then predict the SERS-active ones. The project was split into two parts. First, to create an accurate CNN to predict SERS from non-SERS pillars and second, using this CNN to predict the SERS-active structures in real-time while acquiring Raman measurements.

CNN capabilities lie in understanding local feature patterns in an image and identifying them elsewhere. Thus making it possible to use the CNN developed here to test other forms of SERS substrates. Since the CNN previously learned what a SERS substrate is, the same features can be identified in other SERS substrates, making the SERS identification process generalised. Therefore, the main objective of this work is to improve the reproducibility of the SERS-active substrates signal enhancement for their implementation in lab-on-a-chip devices and point-of-care testing. Using the power of ML for the analysis, we envisioned an improvement in the efficiency of the production of EHD

SERS substrates leading to potential improvements of production techniques in the biomedical field, and a deeper understanding of the SERS mechanisms.

The end goal of this work is to apply this research to point-of-care sensors by creating consistent reproducible results of the SERS signal, the manufacturing process, and the understanding of SERS-active substrate production to help develop SERS sensors in lab-on-a-chip devices. A recent publication by Rickard *et al.* [24] uses SERS to detect blood-based biomarkers in traumatic brain injuries from blood samples, demonstrating that SERS technology can be used for innovative sensors. The lab-on-a-chip market is of particular interest, with an estimated market size of £7.4 billion by 2027 [25]. Its potential profitability means there will be significant interest in current research on SERS-active substrates.

5.3 Methodology

5.3.1 Materials

Polystyrene (PS) Mw 100 kg/mol, Mw/Mn < 1.07 was used as the polymer for the patterning (purchased from Polymer Standards Service GmbH), and toluene was used as the main solvent in this process (Fisher Scientific Ltd). Highly polished p-doped silicon (Si) wafers with <100> crystal orientation (Wafernet, GmbH, Eching, Germany) were used during the EHD as electrodes (X-lith eXtreme Lithography). Methylene blue (MB) (Analytical Standard, Sigma-Aldrich Ltd) was used without further purification. Silver paint (Electrodag 1415M) was used to contact the electrodes. The 99.999% purity gold target for sputtering was purchased from Kurt J. Lesker Ltd.

5.3.2 Electrohydrodynamic Patterning Surface Enhanced Raman Scattering Substrates Fabrication

PS was used as the polymer for the thin film deposition and patterning during the EHD, and highly polished p-doped silicon wafers were used as bottom and top electrodes in the capacitor setup. Silicon wafers were thoroughly cleaned in a "piranha" solution consisting of sulphuric acid and hydrogen peroxide (3:1, accordingly), then rinsed with deionized water and dried under nitrogen flow. An initial solution of PS in toluene 3% w/w was subsequently spin-coated onto the freshly cleaned 1.0 × 1.0cm^2 bottom silicon electrodes, and these were then electrically connected to the external voltage supply and homogeneously heated to 175°C. The whole EHD patterning rig was enclosed in a convection oven, ensuring optimal temperature uniformity throughout the chamber. The temperature in the oven was thermostatically controlled.

Furthermore, a current meter was concurrently used to sense a finite current flowing between the first and second substrates and the whole apparatus was homogeneously heated to the elevated temperature above the glass transition temperature of the PS to enable the thin film's fluidity. Once the EHD process was completed, the capacitor setup was cooled down to room temperature to quench the formed structures, and the external voltage supply was disconnected. The patterned substrates were finally sputter coated with a gold nano-layer with an average thickness of 40.3 ± 2.7nm (Emitech sputter-coater) using a DC argon plasma and a gold target. Subsequently, an analyte to be detected in an ethanolic solution with a 10μM concentration of MB was dry-cast over the substrate and left to dry in air.

5.3.3 Optical Image Acquisition

After the EHD patterning was formed and disassembled, but before gold coating, a set of optical measurements were taken using the Renishaw optical system, and an improved set of optical measurements were taken with a specially adapted research-grade microscope (Leica DM 2700M) equipped with an incoherent white light source. Following the gold coating step, the same EHD patterns were collected again on the two optical systems. Following the SERS measurement of each pillar, a connection of the SERS signal to the pillar image from both microscopes was made to model the CNN. All the pillar images were acquired as a top view image. The choice for using images from two microscopes was related to improving the generalization of our CNN. Obtaining pillar images from both microscopes resulted in more data for developing the CNN model for the two different classes (SERS/non-SERS). Mixing the low and high-resolution images enhanced the model's training and the results when testing the low-resolution images from the Renishaw microscope in the real-time acquisition measurements. The high-resolution images from the Leica microscope are only needed for the model training.

5.3.4 Surface Enhanced Raman Scattering Acquisition

After applying the gold coating of the patterns, and dry casting the analyte (MB) for detection, the sample is taken to the Raman system for further analysis. Using an *InVia* Qontor confocal Raman (Renishaw) spectrometer, SERS measurements were carried out using a 633nm excitation laser. SERS maps were generated using the point mapping mode scan with an exposure time of 1 second per point in the map, with 3 accumulations using a power of 5mW. A 50× objective lens with a numerical aperture of 0.75 was used

for the SERS measurements and optical image collection, with the SERS collected over a range of 500 to 2500cm^{-1} relative to the excitation wavelength. The spectra were later processed by a home-built Python script [26], which performs a custom baseline subtraction followed by a Lorentzian fit to the highest intensity peaks. Afterwards, the enhancing factor (EF) was computed.

5.3.5 Surface-Enhanced Raman Signal Factor Calculation

The SERS enhancement factor was calculated by comparing the non-enhanced Raman scattering peak (I_{raman}) intensities at 1,610 cm^{-1} of pure MB and the corresponding SERS signals (I_{SERS}) obtained from the SERS substrates. The detection volume of the solution-phase MB sample, V_f, was calculated using the following relation:

$$V_f = (\text{depth of focus}) \times (\text{focus area}) = (1.4\, n\, \lambda / NA^2) \times \pi (0.4\, n\, \lambda / 2\, NA)^2 \quad (5.1)$$

The surface density of the adsorbed MB molecules on the structured surface was taken as ρ_s = 3.3 molecules/nm^2, and the enhanced area, A, was defined as the diffraction-limited spot size ($\pi(0.4\lambda/2NA)^2$), with: NA, numerical aperture; n, refractive index; λ, wavelength; ρ_v, the surface density of molecules in the detection volume, estimated from [27]. Following these considerations, the EF was estimated to be EF~1000 × I_{SERS}.

5.3.6 Surface enhanced Raman Scattering / non-Surface Enhanced Raman Scattering labelling and organization

The most important step in building a CNN is having well-defined classes and appropriate images to represent these classes. The images were split into two different classes

depending on their Raman signal (SERS/non-SERS). This definition allows the CNN to learn and differentiate the two classes based on the pillar images. Thus, gaining the ability of making future predictions using the same CNN model to differentiate SERS from non-SERS pillars from an unknown EHD sample.

For the Raman signal acquisition, a total of 2800 Raman map points were acquired over five different EHD substrates. Each map point was measured over a unique pillar. These unique pillars were cropped individually from the Raman microscope image (**Figure 5.1B**) and Leica microscope image (**Figure 5.1A**) and sequentially organized regarding their Raman signal strength (**Figure 5.1C**). All images were processed using GNU Image Manipulation Program (GIMP – https://www.gimp.org/). The pillar images were processed by cropping, making sure only the pillars of interest were present and labelled according to their Raman signal, and later exported into a tif format. The labelling was defined based on the third interquartile range (3^{rd} IQR) intensity of the 1610 cm^{-1} MB peak in **Figure 5.1C** (red arrow), whose value is represented as the pillar EF value in histogram form **Figure 5.1D** (red bar) and whisker plot form **Figure 5.1E** (red dashed line). The EF was calculated as described in the previous section. Each spectrum was categorized as SERS or non-SERS depending on whether the Intensity or EF fell above or below the red line **Figure 5.1C**, **Figure 5.1D** and **Figure 5.1E**. With these instructions, each cropped image could be labelled into their respective SERS/non-SERS class according to the intensity/EF values, and then be used to model the CNN.

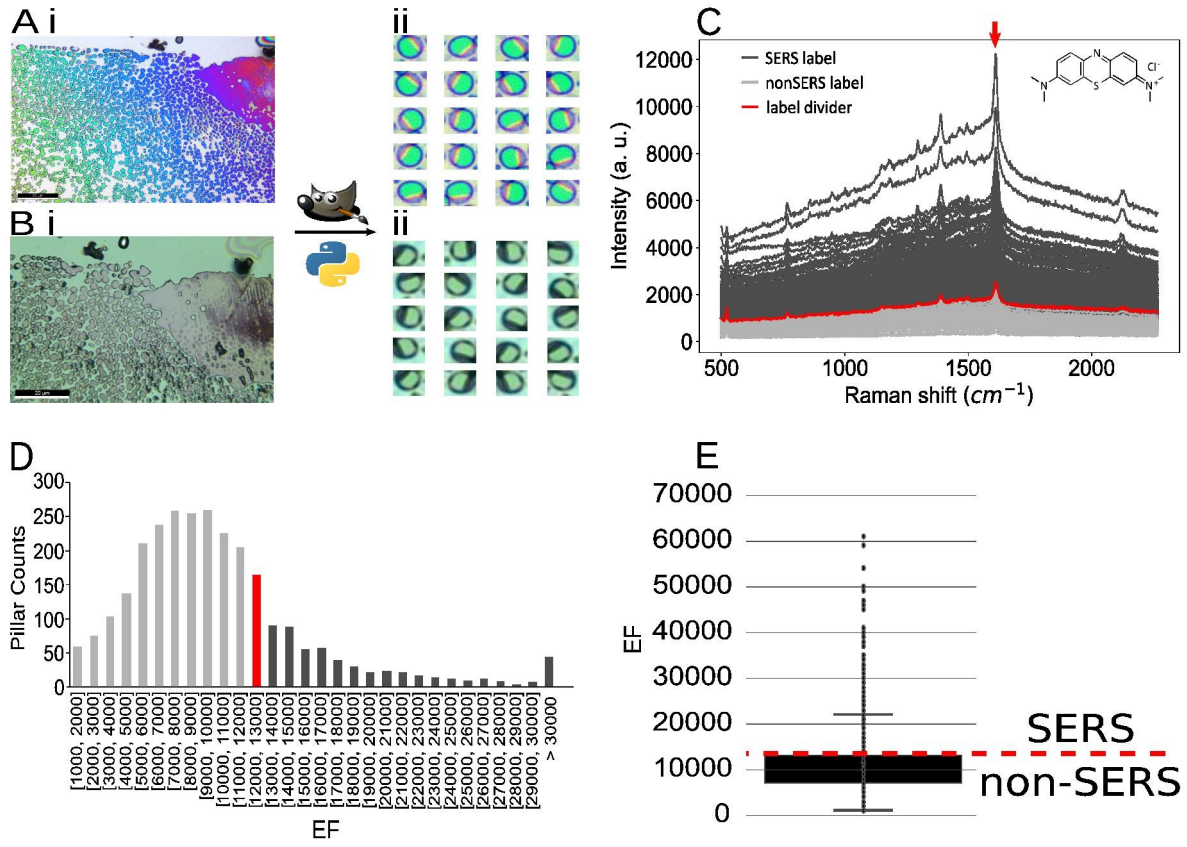


Figure 5.1: (A)(i): *Optical image of the EHD structures from the Leica optical microscope.* ***(ii)*** *The augmentation process generates rotated/mirrored/inverted images of the original Leica microscope image.* ***(B)(i)*** *Same EHD structures as in* ***(A)(i)*** *viewed through the Renishaw inVia optical microscope. The individual pillars are cropped using GIMP and are processed using Python for further image augmentation.* ***(ii)*** *The augmentation process generates rotated/mirrored/inverted images of the original Raman optical microscope image.* ***(C)*** *Raman spectra of a selection of 885 pillars show the separation between SERS and non-SERS spectra, with the red spectrum as a separation spectrum. Top red arrow pointing at the MB 1610 cm^{-1} most intense peak.* ***(D)*** *Histogram plot from the EF for each pillar shows the separation value between both classes in the red bar.* ***(E)*** *Whisker plot with dashed red marking the 3rd IQR separating SERS from non-SERS pillars.*

5.3.7 Data Augmentation

Data augmentation is an image manipulation technique to artificially increase the data sample size by changing the original images with rotations, inversions and mirroring operations while maintaining the EF labels for each pillar. This is possible because the Raman measurements are the same for the pillar, independent of the image rotation. An example is shown in **Figure 5.1Aii** and **Figure 5.1Bii** with both microscopes having the

same pillar subjected to a data augmentation process. Data augmentation ultimately improves the model generalization, reducing the overfitting of the model.

Different data augmentation setups were tried for the CNN optimization and a summary of them are displayed in **Table 5.1**. The number of pillars per class follows the 3rd interquartile range (IQR) of the EF rule described in the previous section.

***Table 5.1**: Data Augmentation trials.*

Iteration	Initial Image dataset		Sum	Image Augmented dataset				Sum
				Train		Validation		
	SERS	non SERS		SERS	non SERS	SERS	non SERS	
1	221	664	885	32,000		8,000		40,000
2	525	1,575	2,100					
3	750	2,250	3,000					
4	950	2,850	3,800					
5	1,050	3,150	4,200					
6				80,000		20,000		100,000
7				160,000		40,000		200,000

An initial sample size of 885 pillar images was **first** divided into 221 SERS labelled pillars and 664 non-SERS labelled pillars. From the 885 pillar images, a total of 40,000 augmented images were created. From these 40,000, 80% were split into training (32,000) and 20% for validation (8,000). Since there were more non-SERS pillars (664) than SERS pillars (221), the SERS pillars were subjected to more cycles of data augmentation due to the need of a balanced number of different classes to make a proper generalization of the CNN [28]. Thus, from the 32,000 training dataset, 16,000 augmented SERS pillar images were created from an initial 221 pillars dataset, and 16,000 non-SERS pillar images were data augmented from an initial number of 664 pillars. The same logic applies for the validation dataset, therefore, from a total of 8,000 pillars, 4,000 pillars were data augmented. From 221 SERS pillar images, 4,000 data augmented were created 664 from non-SERS pillars. This consideration was applied for all the remaining cases.

We continued with a **second** test where 2,100 original images were fed into the CNN model. From these, 1,680 were non-SERS and 420 were SERS. From the initial 2,100 image dataset, a total of 40,000 augmented images were created. 32,000 (16,000 SERS / 16,000 non-SERS) images were used for training and 8,000 (4,000 SERS / 4,000 non-SERS) for validation.

In the **third** test, additional images from the second optical microscope were introduced into the previous dataset. For this test, a total of 3,000 pillar images were used with 2,800 pillars from the Renishaw microscope and 200 pillars from the Leica DM 2700M microscope. From the 3,000 image dataset, a total of 40,000 augmented images were created. 32,000 (16,000 SERS / 16,000 non-SERS) images were used for training and 8,000 (4,000 SERS / 4,000 non-SERS) for validation.

During the **fourth** test, the number of optical images from the Leica DM 2700M microscope were increased to obtain 1,000 images and combined these with the Renishaw images to give 3,800 images. From the 3,800 image dataset, a total of 40,000 augmented images were created. 32,000 (16,000 SERS / 16,000 non-SERS) images were used for training and 8,000 (4,000 SERS / 4,000 non-SERS) for validation.

For the **fifth** test, the number of optical images from the Leica DM 2700M microscope were increased to obtain 1,400 images combined with the 2,800 pillars from the Renishaw microscope to give 4,200 images. From the 4,200 image dataset, a total of 40,000 augmented images were created. 32,000 (16,000 SERS / 16,000 non-SERS) images were used for training and 8,000 (4,000 SERS / 4,000 non-SERS) for validation.

For the **sixth** test, the total number of pillar images remained at 4,200 and we changed the data augmentation number. From the 4,200 image dataset, a total of 100,000

augmented images were created. 80,000 (40,000 SERS / 40,000 non-SERS) images were used for training and 20,000 (10,000 SERS / 10,000 non-SERS) for validation.

Finally, in the **seventh** test the total number of pillar images remained at 4,200 but the data augmentation number was changed. From the 4,200 image dataset, a total of 200,000 augmented images were created. 160,000 (80,000 SERS / 80,000 non-SERS) images were used for training and 40,000 (20,000 SERS / 20,000 non-SERS) for validation.

5.3.8 Convolutional Neural Network

After obtaining the images, separating them into their respective labels and data augmenting them, we proceeded into the CNN generation and optimization to decide which CNN structure is adequate for our dataset.

For the first test, we designed a CNN with one Conv2D(64,(3,3)) layer. The 3 by 3 coefficient is related to the kernel size, which represents the pixel window height and pixel window width. The filter number related to the output feature map depth is 64. In **Table 5.2** the window size has been omitted since it is the same in every Conv2D. After the initial Conv2D, we have a MaxPooling((2,2)) layer, which consisted of extracting a 2 by 2 pixel window from the input feature maps and outputting the maximum value of each channel. In **Table 5.2** the window size has been omitted since it is the same in every MaxPooling operator.

***Table 5.2**: CNN trials.*

Model Number	1	2	3	4	5	6	7	8	9	10	Pre-Trained VGG-16 Frozen	Pre-Trained VGG-16 Fine-Tuned
Learning rate	0.001											
Optimizer	RMSProp											
Loss	Categorical Cross-Entropy											
Input shape	(50, 50)											
L1	64	64	64	64	64	64	64	64	64	64	64	64
L2		64	64	64	64	64	64	64	64	64	64	64
L3	128											
L4		128	128	128	128	128	128	128	128	128	128	128
L5	256		128	128	128	128	128	128	128	128	128	128
L6		256										
L7	512		256	256	256	256	256	256	256	256	256	256
L8		512		256	256	256	256	256	256	256	256	256
L9	512		512		256	256	256	256	256	256	256	256
L10		512		512								
L11			512		512	512	512	512	512	512	512	512
L12	256			512		512	512	512	512	512	512	512
L13	2	256			512		512	512	512	512	512	512
L14		2	256			512						
L15			2	256			512	512	512	512	512	512
L16				2	256			512	512	512	512	512
L17					2	256			512	512	512	512
L18						2	256					
L19							2	256		0,3	0,3	0,3
L20								2	256			
L21									2	256	256	256
L22										2	2	2

Conv2D
MaxPooling
Dropout
Flatten
Dense

All CNN end with a Flatten layer followed by a Dense layer. These layers operate over the final feature map tensor and flattens the tensor (3D) into a vector (1D). The Flatten layer is necessary to make the connection between the feature map tensor output from the series of Conv2D and MaxPooling layers to the classification output. The classification output is the binary classification for SERS and non-SERS pillar images. After flattening the tensor into a vector, this vector is then passed into two Dense layers that converge the Flatten layer into the desired binary classification.

Additionally, when compiling the CNN to learn with the training and validation datasets, three important parameters are required: the learning rate, the optimizer function, and a loss function. The learning rate is a measure of how much the neurons will change per iteration. In our case, we fixed it at 0.001, a standard value for a CNN [28]. The optimizer function was the ‘RMSprop’ and the loss function 'categorical_crossentropy'. These two functions are used to optimize the overall accuracy of the model to distinguish

SERS from non-SERS pillar images. The different CNN's used in this work are presented in **Table 5.2**.

The CNN layer optimization tests start with a basic CNN format (**Table 5.2 model number: 1**) and becomes incrementally more complex with the addition of new Conv2D layers and MaxPooling layers. After incrementing the number of layers for the first nine tests, for the tenth, we added a Dropout layer with a 0.3 factor. The Dropout layer with the 0.3 factor randomly sets 30% of the neuron weight entries to zero. Turning some entries to 0 guarantees a reduction in the overall model overfit. For the eleventh test (pt-VGG16), we added more complexity into the system by importing a pre-trained CNN, the VGG16 [29]. To adapt it to our CNN we changed the terminations of the VGG16 for our SERS/non-SERS labels using the integral VGG16 CNN and adding the Dropout, Flatten and two Dense layers. Additionally, the VGG16 was imported into our CNN model with its weights and biases. For this test, these weights and biases were frozen, making them unmodified when training the model with the SERS/non-SERS dataset. The only trained layers were the added layers, i.e. the Dropout, Flatten and the two Dense layers. Weights and biases are parameters associated to each neuron in our model. For the VGG16, 138 million neuron parameters were loaded and used for our adapted CNN. Finally, for the twelfth test (pt-VGG16-ft), we improved the eleventh CNN, and altered the last set of Conv2D and MaxPooling of the VGG16 model to be unfrozen, so their parameters could be fine-tuned to our SERS/non-SERS data.

5.3.9 Output from the Convolutional Neural Network model

After the model is trained and validated, we passed test image data though the model to check the true accuracy of our model in distinguishing between SERS and non-SERS

pillars. Here we implemented not only a method to distinguish pillars according to their classification but also to highlight, using an overlaid heatmap, regions from the image that were specifically activated on the CNN in the presence of a SERS pillar. Activation can be low or high depending on the neuron activity towards a specific patch of the image. To apply these activations, we used a gradient class activation map (grad-CAM). Grad-CAM employs a mathematical operation to the CNN prediction of the test image outputting the activation neurons from the last Conv2D layer. Since the last layer of the CNN is the interface between the last layer and the classification layer of a CNN, it highlights the most prominent features that make the classification decision.

5.3.10 Video input onto the Convolutional Neural Network

After setting up the CNN model, we apply it to real-time Raman data acquisition. To that effect, we incorporated the CNN over the Renishaw computer and pass it screen-grab images, which included the optical microscope display part (**Figure 5.2A**). There is also some unnecessary information from the screen-grab (**Figure 5.2Ai, iii** and **iv**), so we applied a new image processing CNN onto the screen-grab that crops the optical microscope image region from the whole screen grab and feeds it to the CNN automatically (**Figure 5.2Aii**). The reason to get the screen-grab is that the process can happen in real-time while the user is scanning the surface without manually exporting the image.

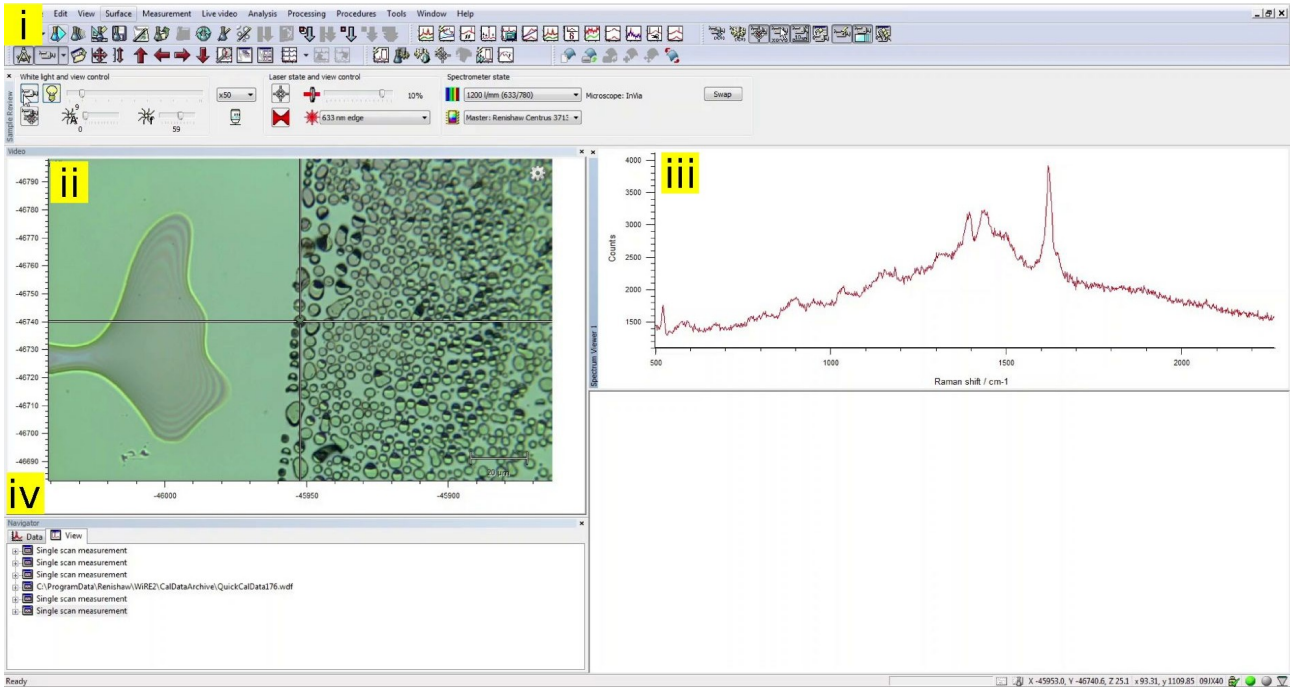


Figure 5.2*: Example of a screen-grab from the Renishaw computer when acquiring a Raman measurement.* ***(i)*** *Widgets displayed from the WiRE interface.* ***(ii)*** *Optical microscope image from the sample.* ***(iii)*** *Raman spectrum from the acquired spot delimited by the cross-hair intersection in* ***(ii)****.* ***(iv)*** *Measured sample list.*

5.3.11 Screen-grab Optical Microscope Image

Since we have accumulated many pillar images from the previous SERS/non-SERS CNN, we can now use them to build another CNN for screen mapping the optical microscope. This new CNN distinguishes between pillar and non-pillar datasets. The ‘pillar’ label corresponds to the whole pillar selections that were captured from the previous SERS/non-SERS CNN, and ‘non-pillar’ refers to scraped images from the WiRE-feed that do not involve the optical microscope part (**Figure 5.2Aii**). Scraped images are randomly selected sections from the WiRE interface (**Figure 5.2Ai, iii** and **iv**), which will be data augmented to compensate for the sample size difference.

5.3.12 Pillar/non-Pillar labelling, Data Augmentation and Convolutional Neural Network building

From the SERS/non-SERS pillars, we used 4,200 images with a new label ‘pillar’. For the non-pillar label, around 20 scraped images from the WiRE interface were used. The scraped image cuts varied in shape and were taken from diverse spots in the WiRE screen-grab image (**Figure 5.2Ai**, **iii** and **iv**).

These images were data augmented with the following conditions: a width shift range of 0.2, a height shift range of 0.2, a shear range of 0.2, a zoom range of 0.2 and horizontal flip and vertical flip. These new configurations generated distorted images from pillars and from the scraped images that helped to generalize the CNN. A total of 40,000 augmented images were created with 20,000 being used for the pillar label and 20,000 for the non-pillar label. From these, 16,000 were assigned for training and 4,000 for validation. The CNN used was the pt-VGG16-ft, described previously. The fitting of the CNN did not require much time (approximately 1 minute) to run since both labels were very distinguishable and the model could reach a validation and training accuracy of 100%.

5.3.13 Applying the pillar/non-pillar Convolutional Neural Network

The pillar/non-pillar CNN was trained on overall smaller images than the actual whole WiRE screen-grab (1920 pixels by 1280 pixels) (**Figure 5.3i**), we used a script to break the screen-grab image into smaller tiles (100 by 100 pixels) and pass these tiles onto the fitted pillar/non-pillar CNN model. The model used the grad-CAM method to assign the tiles with the pillar (heatmap) or non-pillar (black) label (**Figure 5.3ii**). Since there were some

elements that the pillar/non-pillar CNN did not understand, for instance the cross-hair, the scale bar, or the optical microscope image contours, these elements were interpreted as the non-pillar class as a result of the lack of proper training for these images. To mitigate this, we wrote a second script that preformed a padding operation towards accepting these areas and thus completed the enclosed region of the microscope image (**Figure 5.3iii**). After the padding and obtaining the final defined optical microscope region, we could manipulate it by passing it onto the SERS/non-SERS CNN.

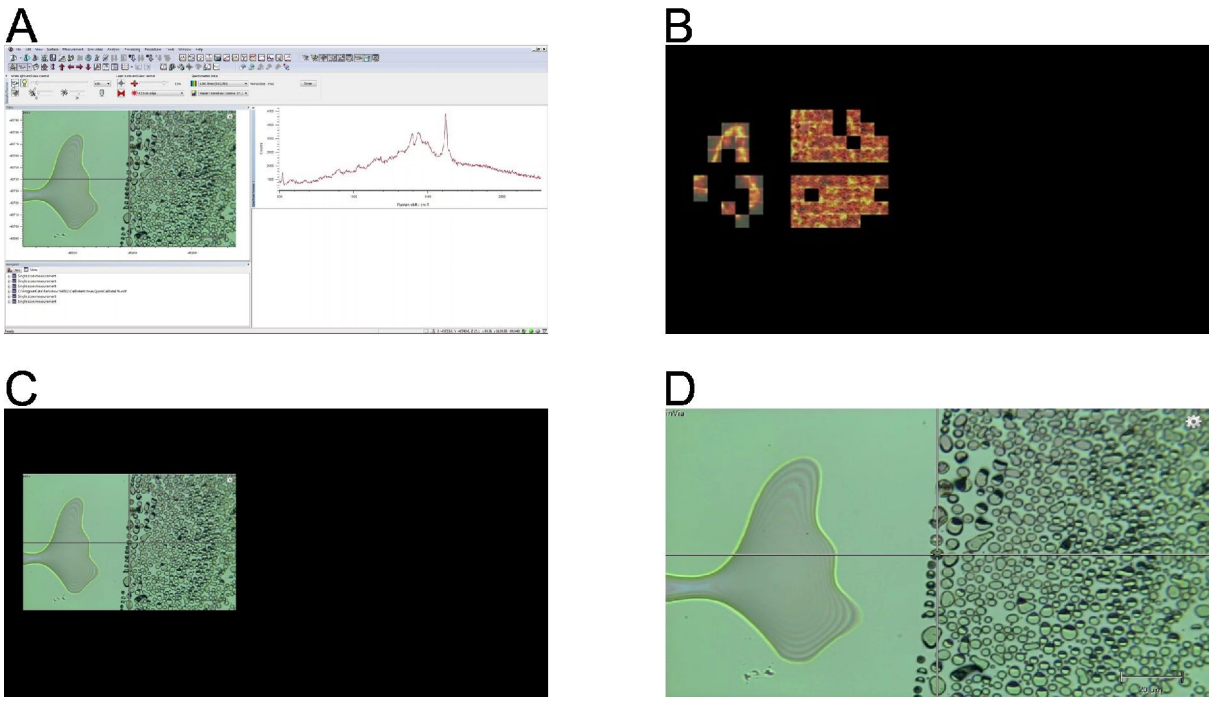


Figure 5.3*: Applying an algorithm to break the WiRE screen-grab* ***(A)*** *into tiles, and then passing the tiles into the pillar/non-pillar CNN.* ***(B)*** *Output of the tiles being processed with the CNN, heat mapped by grad-CAM algorithm.* ***(C)*** *Padded final region of interest – optical microscope image-* ***(D)*** *final cropped optical image.*

5.3.14 Applying the Surface Enhanced Raman Scattering / non-Surface Enhanced Raman Scattering Convolutional Neural Network

Once we obtained the cropped optical microscope image (CrOMI) (**Figure 5.4A**), we fed the image onto the SERS/non-SERS CNN. As we faced again the same image size issue,

we decided to break the CrOMI into tiles and fed the tiles into another script. This auxiliary script cropped the pillars inside the tiles into single pillar images to be tested on the SERS/non-SERS CNN. This script uses Pythons' cv2 module (https://opencv.org/), specifically the 'cv2.findContours function' that locates circle shapes in an image. Afterwards, the SERS/non-SERS CNN classified the pillars and outputs the pillar image with the grad-CAM heatmap overlaid to show if the pillar was SERS. Then there is a reconstruction process where the pillar with the overlaid grad-CAM heatmap was reconstructed onto the tile image (**Figure 5.4B**). The tile images with the grad-CAM heatmap overlaid pillars were then reconstructed onto the CrOMI (**Figure 5.4C**), which was in turn reconstructed onto the WiRE screen-grab (**Figure 5.4D**). The grad-CAM heatmap locations give a visual cue to the Raman microscope user to direct the measurement over the pillars that are SERS overlaid with the heatmap.

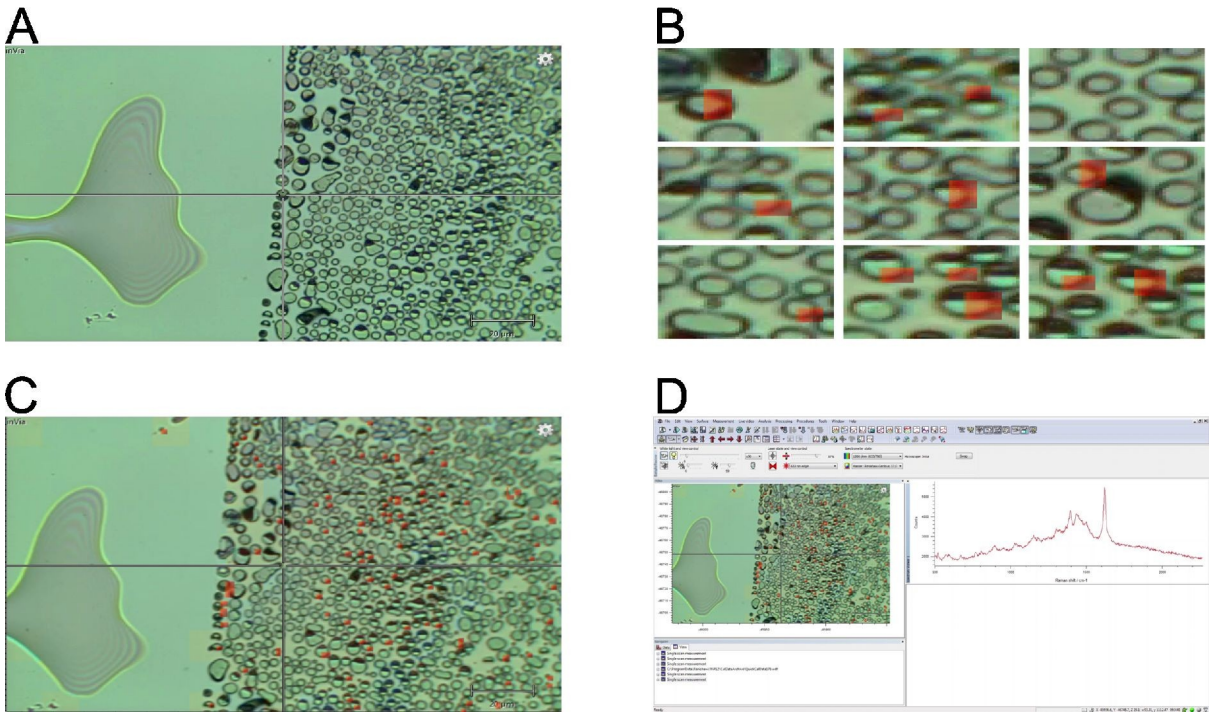


***Figure 5.4**: Application of the SERS/non-SERS CNN. **(A)** CrOMI from the WiRE software screen-grab. **(B)** Break the CrOMI into multiple tiles and pass them onto the SERS/non-SERS CNN outputting the grad-CAM results heatmap. **(C)** Reconstruct the CrOMI overlaid with the heatmap tiles. **(D)** Overlaid WiRE screen-grab with the highlighted heatmap SERS pillars on the optical image.*

5.3.15 Real-time run event proof-of-concept

Unfortunately, it is not feasible to pass a real-time feed from the WiRE software itself into our Python script due to Renishaw WiRE software limitations. To solve this problem, we screen grabbed the whole computer screen and fed the recording into the script. Resulting from this, we can provide a proof-of-concept test with a video feed from a real-time 8 pillar measurement run; and subsequently, the WiRE feed was analysed with the full functionalities of the CNN. Firstly, with the pillar/non-pillar CNN to obtain the CrOMI, and secondly with the SERS/non-SERS CNN to obtain the grad-CAM of each SERS pillars. The screen grab video was recorded using OBS software (https://obsproject.com/).

5.3.16 Real-time Surface Enhanced Raman Scattering pillar tracking summary

Firstly, initialize the Raman system and set it up for Raman acquisition. While searching for the EHD SERS pillars, the script screen-grabs the WiRE software in real-time (**Figure 5.5A**). From the screen-grab, a script that uses the pillar/non-pillar CNN model (**Figure 5.5B**) processes the screen-grab and identified the regions where there were pillars. The identified region (**Figure 5.5C**, **D**) was then cropped into the CrOMI (**Figure 5.5E**) and the CrOMI was analysed using the SERS/non-SERS CNN model (**Figure 5.5F**). This analysis highlighted SERS pillars using the grad-CAM algorithm, and the heatmap activation regions were overlaid on the CrOMI (**Figure 5.5G**). Then the CrOMI was overlaid onto the WiRE software, facilitating the researchers' task of measuring using the highlighted pillars which were SERS active.

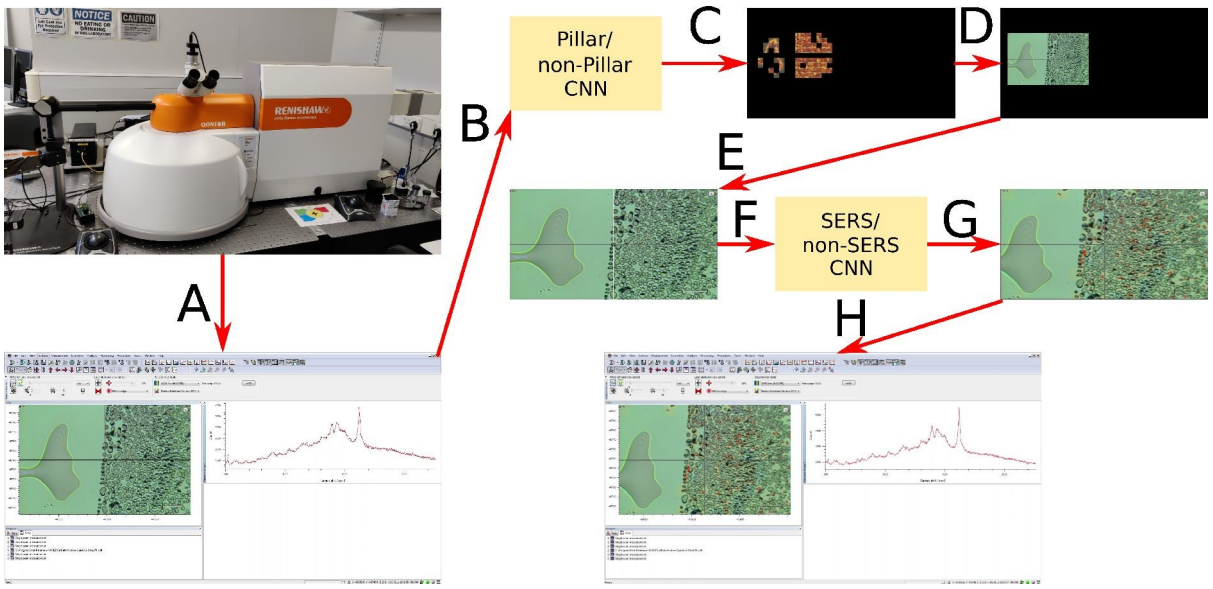


Figure 5.5: *SERS tracking method overview.* ***(A)*** *Continuous image acquisition using an inVia Renishaw Raman Spectrometer outputs SERS spectra corresponding to individual EHD pillars in the image.* ***(B)*** *The real-time screen-grab is analysed by the pillar/non-pillar CNN, revealing the heatmap activation for the optical microscope region* ***(C)****.* ***(D)*** *Later padding it and cropping it* ***(E)****.* ***(F)*** *The cropped image is passed onto the SERS/non-SERS CNN model and, the pillars labelled as SERS, are highlighted using the grad-CAM algorithm* ***(G). (H)*** *The output overlays the WiRE interface with the SERS/non-SERS activation regions over at the optical microscope screen section.*

5.3.17 Scripts

All the scripts written for this work were done in Python (https://www.python.org/) and can be found in [30]. The available scrips are the scripts to create the CNN (SERS/non-SERS and pillar/non-pillar), the script to fit multiple peaks from Raman maps, the data augmentation script, and the script that integrates all the scripts together and outputs the screen-grab with the appropriate SERS activations layer.

5.4 Results and Discussion

Starting with the optimization of the SERS/non-SERS CNN, **Figure 5.6A** and **Figure 5.6B** show different combinations of initial data and data augmentation, respectively. In **Figure**

5.6A there are multiple regions labelled from **i** to **v**. In **i**, a total of 885 SERS/non-SERS pillars were used and split into 221 SERS pillars and 664 non-SERS and later augmented to a total of 40,000 images. Since the initial number of SERS pillar images are three times less than the non-SERS, the augmented images from the SERS class had three times more data augmentation cycles than the non-SERS. Because of this, the SERS augmented data were more likely to be similar to each other. This fact is noticed by the slight difference between the validation (91%) and training (94%) accuracy. The 3% difference is related to a slight overfit of the model from the fewer SERS images compared to the amount of initial non-SERS images.

For **ii**, a total of 2,100 original images were used and augmented to the same 40,000 images and then fed into the CNN model. Compared to **i**, the accuracy of the results obtained in **ii** were lower, and the discrepancy between validation (84%) and training (89%) accuracy increased (5%). This can be explained by the addition of pillar images from other EHD samples. Since EHD samples can vary, the pillars grown for each sample can also vary, thus by adding different pillars to the image pool, it increases the variability of pillar images which affects the model fitting accuracy. Furthermore, in **i** only one EHD sample was used thus the accuracy obtained was better compared to **ii**, because learning from similar pillar images outputs better accuracy. However, the model is then not general for different EHD samples which explains the slight decrease in the model accuracy.

For **iii**, a total of five EHD samples were used, as well as additional pillar images from another optical microscope with better resolution and added colouring to help the generalization of the CNN model. Overall, 3,000 pillars total (2,800 pillar samples from the Renishaw microscope, 200 pillars from the Leica optical microscope) were augmented to 40,000 and analysed by the CNN. With these changes, we observe a decrease in the

validation accuracy to its lowest value (81%). This decrease in accuracy can be explained by the addition of a few new images from another microscope and the addition of more EHD images from different samples, which negatively impacts the overall accuracy of the model as a consequence of the increased variance with the additional pillar images. However, these added images help the overall generalization of the model, in spite of lowering its accuracy.

More images from the Leica optical microscope can be added to help the generalization of the CNN by increasing the variability of the pillar images. In **iv**, the number of optical images was increased to 3,800 pillars, resulting in a slight improvement of the model's validation (83%) and training (91%) accuracy that remained unchanged after increasing the total number of optical images to 4,200 pillars **(v)**. After this point no more pillar images were added to the dataset to prioritise the improvement of the CNN model by optimizing the data augmentation cycles.

Moving on to **Figure 5.6B**, the number of original images constant (4,200) while improving the total number of augmented image data. From the previous tests, a total of 40,000 images were data augmented (20,000 SERS and 20,000 non-SERS) which resulted in an 83% validation accuracy and 91% training accuracy. By increasing the augmented images to 100,000 (50,000 SERS 50,000 non-SERS), the values obtained for the validation accuracy were improved to 88%, shortening the distance between validation, and training accuracy (91%), hence turning into a more general model with less overfitting. Finally, with 200,000 augmented images we obtained similar validation accuracy (90%) and training accuracy (90%) value, showing a reliable model generalization.

Another interesting finding in this work refers to the relation between the relative accuracy difference (RAD) of the validation accuracy (VA) and the training accuracy (TA)

regarding the ratio of the number of augmented images to the number of original images, as shown in **Figure 5.6C**.

$$RAD = \frac{TA - VA}{TA} \times 100 \tag{5.2}$$

There is a clear decrease in the RAD (**Equation 5.2**) values when the ratio between augmented and original images increases, reaching a minimum for the ratio between 4,200 original images to the final 200,000 augmented images. With these settings, we managed to train a model with very low overfit since both training and validation accuracy are high and the same, 90%. Moreover, this model is capable of distinguishing SERS from non-SERS pillars and is the final setting used for the proof-of-concept run.

To develop a generalised CNN model it is important to work with the highest number of images that can describe an EHD sample. By comparing the difference between the validation and training accuracy values obtained for the initial 885 images, which all come from one sample, (90.5%/94%) with the final 4,200 image dataset, coming from 5 samples and 2 different microscopes, (90%/90%). The higher accuracy obtained for the 885 images could give the impression that the model is well trained and generalised. However, in reality the model would be overfitting to a specific set of pillar images from only one substrate and not generalizing to the whole possibility of EHD pillars from many substrates, which would lead to mistakes when running a test sample. Even though the training of different pillars images from different samples would results in a loss of accuracy, it is worth it for a better overall generalization of the EHD sample.

To check if there were simpler alternatives matching the pt-VGG16-ft CNN accuracy, some additional tests were performed over simpler CNN. More complexity to the CNN was progressively added and the training and validation accuracy was tracked until

reaching the pt-VGG16-ft architecture described in **Figure 5.6D**. For this analysis, initial layers were added and progressively added intermediate and terminal layers. When we added the top layers to the initial models (model **1-4**), the performance of the CNN stayed around 75% for both training and validation accuracy. When adding the intermediate layers (after model **4**), both training and validation accuracy decreased, especially for the model **6** and **7** (highlighted with a *****). The model broke down and reverted to 50% training and validation accuracy, most likely due to overfitting issues. This issue only occurred when changing the intermediate layers.

Lastly, when adding terminal layers (models **8** and **9**), the model improved slightly (74%), but the accuracy never got back to the level of the simpler models (models **1**,**2** and **3**). The potential reason for this is that the pillar images used in this work were small (approximately 50 by 50 pixels in size). For this image size, a simpler CNN can obtain better accuracy than a complex CNN since the more complex CNN will have more redundant information when trained on its own, which can lower the model accuracy. However, with the incorporation of the pre-trained layers (pt-VGG16) the model accuracy (model **11**) improved and approached the 80% accuracy, proving the power of incorporating the pre-trained VGG16 into a model. Finally, when fine-tuning the terminal layers of the pt-VGG16 – model **12**, the CNN reached the maximum level of 90% accuracy for both training and validation datasets. With these sets of tests, we showed how powerful the combination of the pre-trained VGG16 and the fine-tuning are for the accuracy improvement of a CNN, and that is the reason why we used it for our proof-of-concept.

Comparing the pt-VGG16-ft results with those from [11], the validation and training accuracy found was around 99% with the same model, similar EHD samples but a different image quality. In [11], images were obtained from AFM. Since these images have better resolution and the colouring of the AFM image is related to pillar height changes,

they are more comparable between each other. Thus, the AFM images provide more information to the CNN. Moreover, compared with the optical microscopic images, which have the advantage of including colour information which helps with the CNN modelling however, the lack of height profile information means it is not as defined to the model as the AFM images. This is the reason why we incorporated images from the Leica microscope to give the model additional colour information to train on. Nevertheless, they are not comparable to the height information obtained from the AFM, so we are mostly working with the pillar shape and colour variation as the only aspects for the CNN to consider from the optical images, and not any information related to pillar height difference.

The downside to using an AFM for image collection is the long acquisition times required to obtain high quality images. Because of this, having large amounts of AFM data is hard and so only a total of 130 images were acquired in [11]. The low amount of original data can affect the overall accuracy of the model and give it a better accuracy value than it should be, as it was seen in **Figure 5.6Ai**. Hence, here in this work an optical microscope system was used to acquire more images and be able to do this in real-time rather than several hours for AFM, although at the cost of image quality. With these considerations, the present work generalizes better for SERS pillars than in [11] with a faster method and with only a 9% decrease in accuracy.

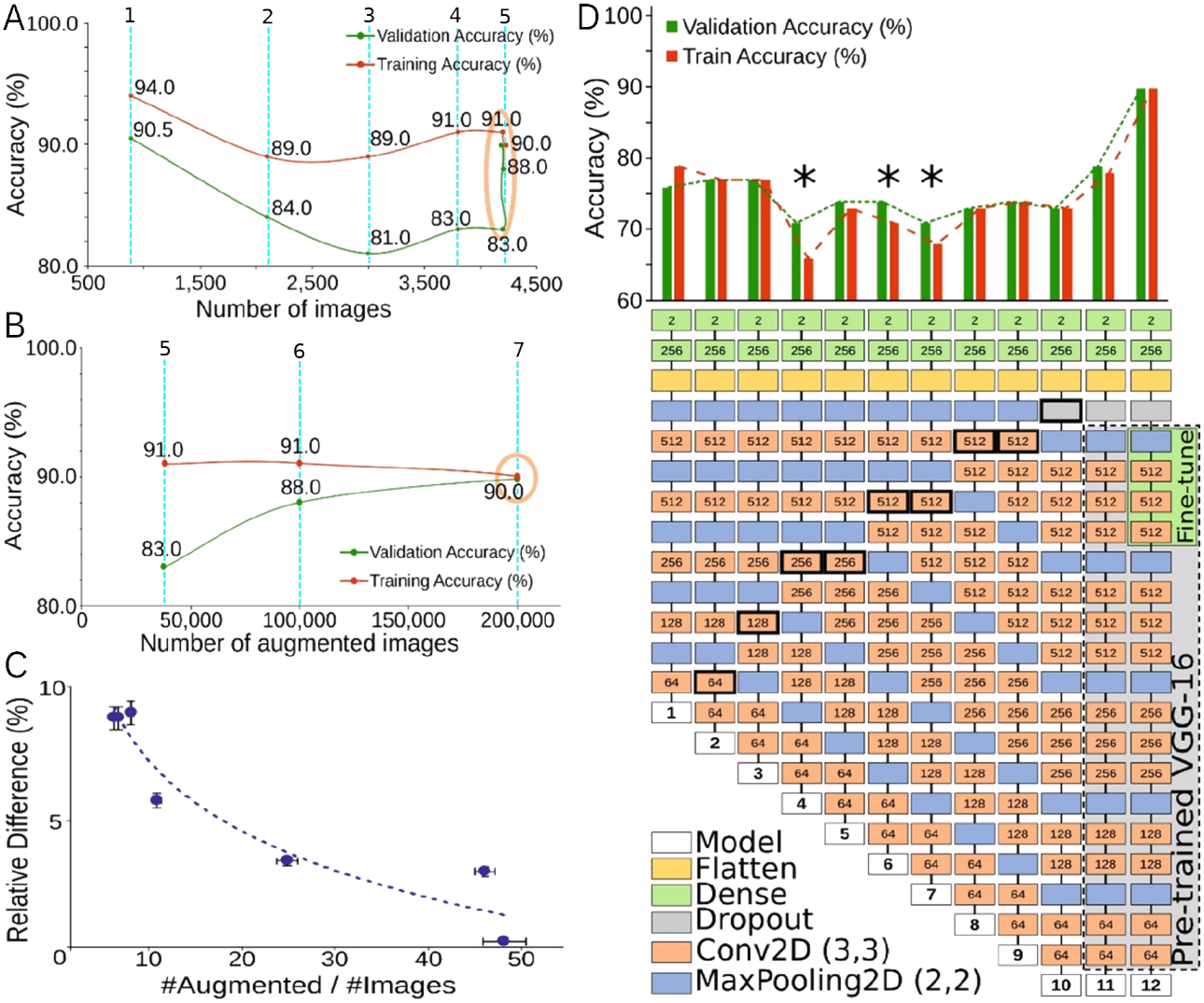


Figure 5.6: (A) *Validation and Training accuracy change compared with the number of initial images. Marked as points of interest* ***1-5*** *for different data augmentation iterations related to* ***Table 5.1****. Orange circle related to* ***(B)****.* ***(B)*** *Validation and training accuracy values changing with the number of augmented images from* ***5-7*** *related to* ***Table 5.1****. Orange circle related to* ***(D)****.* ***(C)*** *Relation between the relative difference (%) between the training accuracy and validation accuracy and the ratio of the number of augmented images to the number of initial images.* ***(D)*** *Validation and training accuracy changes with different CNN model architecture,* ******* *notation related to models breaking down after some epochs. Changes between subsequent CNN architectures highlighted by a thicker square boarder.*

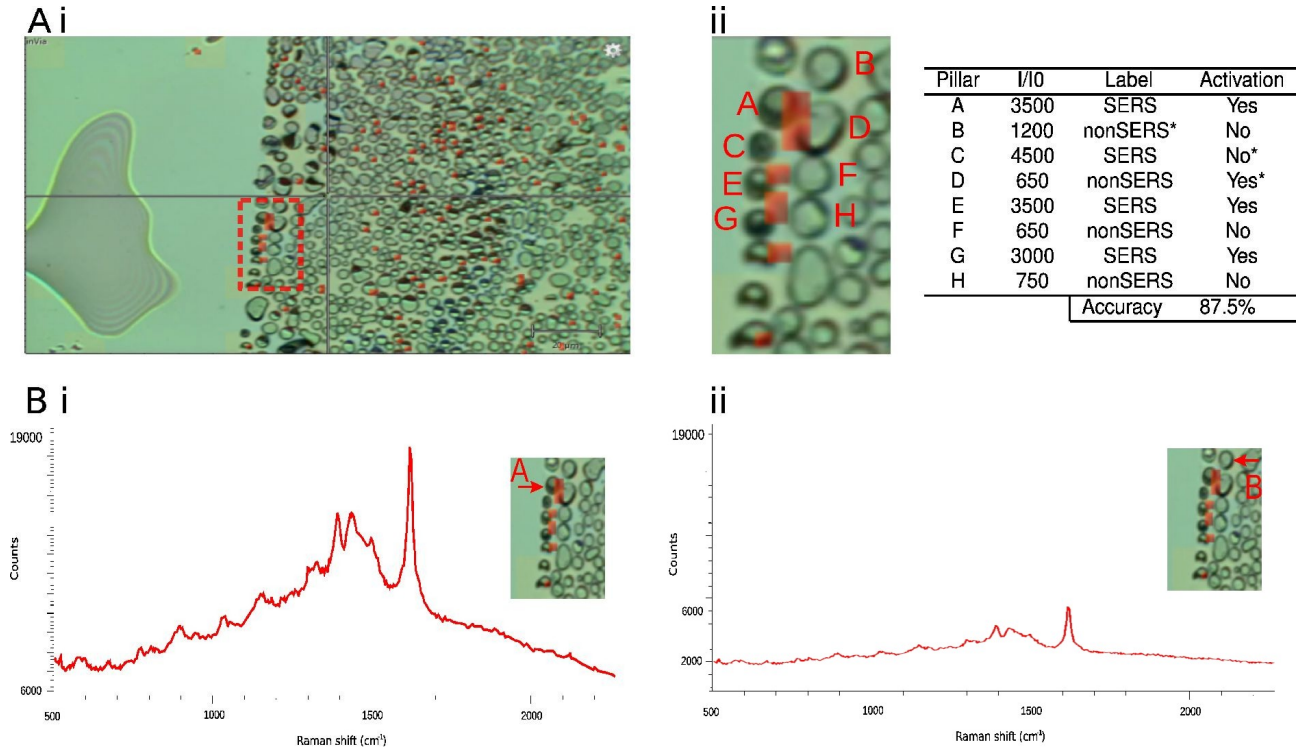
After the development of our SERS/non-SERS CNN model, we wanted to automate the process of identifying SERS pillars while using the Renishaw Raman system. To that effect, we load the Renishaw system over the sample, use the pillar/non-pillar CNN model to first get the CrOMI from the Renishaw video feed and then passed the CrOMI into the SERS/non-SERS pt-VGG16-ft model, as described in **Figure 5.5**. The identified pillars were labelled according to their SERS/non-SERS CNN classification. If the pillars were

labelled as SERS, the grad-CAM algorithm would overlay the activation spot over the pillar image. The activation spot was designed to only highlight activations above 0.8 (value between 0 and 1). In other words, weak SERS activations between 0.5 and 0.8 were not highlighted. These considerations positively affect the algorithm by decreasing the false positives rate since, during our testing of the model, some pillars were labelled as SERS but had a weak activation value of around 0.5-0.55. These activation values mean that the model classifies these pillars almost as a coin flip. A non-SERS label was applied to prevent false positives in these circumstances since the model was unsure of which class the pillar belonged to.

Whenever the activation value was above 0.8, the WiRE image was reassembled with the grad-CAM regions showing the SERS pillars, **Figure 5.7Ai**. As a test-run, a selected patterned area in **Figure 5.7Aii** was selected and ran a total of eight different pillars for the SERS/non-SERS proof-of-concept **Figure 5.7Aii (A→H)**. The different intensity ratio, real label, and activation are shown in **Figure 5.7Aii** table. In **Figure 5.7Bi** and **Figure 5.7Bii** we highlighted two different Raman acquisitions showing a clear distinction for the signal of the different pillars and their respective activation. Overall, the classification labels have been correct six times while the remaining two classes were unclear to which pillar they belonged to, thus for these cases, a ***** was added in **Figure 5.7Aii** table and half a point was given to the classification. With these considerations, the accuracy of the test-run was 87.5% which is a similar accuracy value obtained for the pt-VGG16-ft model of 90.0%.

A couple of subtle considerations in this approach should be considered. Firstly, the original image had some unknown features to the SERS/non-SERS model such as the cross-hair and the scale-bar, which are part of the WiRE interface by default, **Figure 5.7Ai**. These regions were never considered when training the SERS/non-SERS model. Thus,

pillars next to these details were always inaccurately classified. Despite this, there is still room for many pillars to be analysed. Another consideration appears during the process of breaking the CrOMI into tiles, **Figure 5.4B**. The method selects a fixed pixel size to section off the CrOMI. Each tile is a square, and in some cases, the crop happens over a pillar which makes them unusable for the SERS/non-SERS classification, so the model does not consider these cases. Also, only the pillars enclosed by the circle contour script are considered, so some pillars are inevitably left out if they are not considered a circle by the script.



Pillar	I/I0	Label	Activation
A	3500	SERS	Yes
B	1200	nonSERS*	No
C	4500	SERS	No*
D	650	nonSERS	Yes*
E	3500	SERS	Yes
F	650	nonSERS	No
G	3000	SERS	Yes
H	750	nonSERS	No
		Accuracy	87.5%

Figure 5.7: (A) (i) *CrOMI with outline SERS activation pillars. The red dashed square defines a test-run region.* ***(ii)*** *Zoom in on the test-run region with highlighted SERS activations. Table with the accuracy score test, the relative intensity of the pillar (I/I0), the real pillar label and the activation classification label.* ***(B)*** *SERS measurement tests.* ***(i)*** *SERS measurement with the appropriate activation over the pillar A pointed with an arrow.* ***(ii)*** *Non-SERS measurement without any activation over the pillar B pointed with an arrow.*

With all these considerations, the number of activated pillars in the final image tend to be less than the real SERS pillars. However, the goal is to improve the SERS pillar

prediction process to facilitate making the decision to measure SERS pillars, which improves the SERS reproducibility, and with the aid of the CNN we have improved greatly the SERS selection in comparison to our first approach of random guessing, thus making the spectra collected more consistent and data collection can be achieved in a shorter time.

5.4.1 Further work/Improvements

Some improvements can be made to the CNN to obtain a better score than 90% accuracy. The first one refers to the design of our CNN because it was made to have an input image of 50 by 50 pixels. The pixel size was defined to optimize the computational time, however, defining this pixel size creates a problem when using the grad-CAM algorithm. Multiple 2DMaxPooling layers are used, and after each passing of this layer the image is shrunk to half the pixel size, resulting in a feature map of 3 by 3 pixels at the end of the CNN. Thus, the grad-CAM algorithm outputs the heatmap with these same dimensions, which are too small to identify any detailed features in the original image (50 by 50 pixel). Although there is a SERS activation region, it is mapped as a square of 50/3 by 50/3 pixels on the larger pillar image of 50 by 50 pixels. These pixel differences make the grad-CAM heatmap overlay less precise to give information about the exact spot where SERS activation will be detected on the pillar. To solve this issue, we need to increase the input size to 150 by 150 pixels or more. This change comes with an increase in the run time of the script but, in the end, the grad-CAM algorithm would output a more defined window to plot a more detailed grad-CAM SERS activation. This fact was considered when dealing with pillar **C** and **D** in **Figure 5.7Aii**, where both pillars shared the same region when fed into the SERS/non-SERS model, which was later SERS activated. The activation was highlighted in the same

region for both images, thus the reason why we considered adding half a point for this imprecise classification.

Another area for improving the system is concerned with adding additional pillar images from a more precise tool such as the AFM. AFM images have shown great results when used for SERS/non-SERS classification [11] although the AFM acquisition time is higher than the optical microscope acquisition time. Perhaps this work could have been improved if some AFM images from the same pillars were incorporated into the dataset, as well as added new pillar images using a higher magnification objective such as the x100 objective. Potentially, with this new dataset, a better accuracy for the SERS/non-SERS model could have been obtained.

Regarding the script used to obtain the CrOMI, an addition to the script section can be further implemented regarding the image processing. Such as an improved tile cropping method (**Figure 5.4B**) and a script to perform padding over the cross-hair and scale-bar sections (**Figure 5.4A)**. Here, two potential implementations were describe on how to correct these issues. First, there is an image processing function in Python called cv2.laplacian that can remove straight lines and could be a viable candidate to remove the cross-hair and scale-bar lines. The algorithm automatically fills the missing lines by padding the region with average colours from the pixels surrounding cross-hair lines. However, the padding can affect the pillar images if they are located near the lines. Another alternative is to overlay two CrOMI images by making a small adjustment to the sample's position between each image acquisition, then combining them. The combining process would automatically remove the cross-hair and the scale-bar since the images are taken from different spots, so the script could automatically stitch a new image without them. This method also helps with the tiling of the CrOMI process, which occasionally cuts pillars when cropping the tiles, making them unavailable for classification. With this option,

the two CrOMI images could be tiled together and whenever a pillar was cut in one CrOMI, the other CrOMI would not have it cut, so a selection of all the uncut pillars could be done and later used to feed into the SERS/non-SERS model. With these considerations, the resulting grad-CAM image would have more SERS pillar activations since more pillars would be processed, resulting in more SERS pillars being found and, potentially, more SERS pillars available to select.

Finally, the decision of the SERS/non-SERS intensity can also be improved. Since the intensity value distribution is not a normal distribution, a risk was taken when assigning SERS labels to 'weak' intensity values. This approach, however, underestimated SERS values and considered SERS samples EF~10^4, which might not be the best decision when dealing with SERS. After some deliberation, a better approach would be to choose EF~10^6 to ensure that the only pillars used for measurements of the analyte were very good SERS pillars, even though this approach comes with a drawback, as the number of pillars with the SERS label would drastically decrease. This is the reason behind the use of the 3rd IQR of the intensity as a more moderate divider between classes. Another different division system could be tried in the future that would output better accuracy results. For instance, a range for SERS pillars could be added instead, considering SERS pillars with the EF between, 10^4<EF<2×10^4. If this approach was followed, the new dataset would have been built around the two labels: one for average-SERS with 10^4<EF<2×10^4 and another not-average-SERS with EF<10^4 and EF>2×10^4. This approach would make a more refined SERS reproducibility case because it would force the classification of SERS pillars to be between a certain range of EF. This would give a consistent enhancement to the analyte, making it very useful for clinical diagnosis of biomarkers in a clinical environment.

Overall, the CNN developed in this work has reached a 90% accuracy in real-time detection of SERS pillars from an optical image. These results are very promising and extendable to other SERS substrates. Hopefully, a contact with the Renishaw company could be made to implement this software onto their products so that every user could have a specialized SERS detection function in their machines.

5.5 Conclusions

In the present work, a CNN was developed that can accurately predict 90% of the SERS pillars in an EHD SERS substrate using optical microscope images. This high accuracy value for optical images was possible with a large dataset of 4,200 raw images and subsequent large data augmentation of 200,000 samples, 100,000/100,000 for each SERS/non-SERS class. Combining the whole image dataset with a pre-trained CNN with fine-tuning of the last layers, pt-VGG16-ft, made the high levels of accuracy of the CNN model possible.

Pre-trained models used like the pt-VGG16-ft have shown to be suitable to produce high accuracies within models. However, it is very dependent on the computational power available to the user running the script. Potentially the script can also be executed through a cloud service. With these additional computational resources, more raw data images could be added past the 4,200 and further extension of the data augmentation process. Additionally, adding better quality images for the training from different sources, for example the AFM and better resolution objectives, would further reduce overfitting and improve the overall accuracy of the model, perfecting the generalization to the EHD patterns.

A final script of this work was designed with the function to analyse WiRE-feed real-time images of EHD pillars through the optical microscope image and give accurate predictions over the CrOMI, highlighting SERS pillars to be used for data acquisition of the required analyte. There are some areas for improvement in future iterations of this script, such as the way pillars are cropped into tiles and further developing the pillar contour function to improve pillar selection. These changes would increase the number of identified SERS pillars shown over the CrOMI, which would increase the selection of SERS pillars to the user. Some additional changes could be done in terms of the classification rule imposed for labelling SERS and non-SERS pillars that could potentially give better accuracies to the model and more consistent SERS reproducibility.

Overall, the ML has been hugely beneficial in this analysis to find SERS-active substrates produced from EHD lithography. This work has demonstrated the power of ML and how it can be applied to many areas of analysis in the engineering field with positive consequences. In this way, the CNN has proven to be a valuable tool for analysis of SERS-active substrates in real-time.

5.6 References

[1] Fleischmann M, Hendra PJ, McQuillan AJ. Raman spectra of pyridine adsorbed at a silver electrode. Vol. 26, Chemical Physics Letters. Elsevier BV; **1974**. p. 163–6. Available from: http://dx.doi.org/10.1016/0009-2614(74)85388-1.

[2] Vo-Dinh T, Wang HN, Scaffidi J. Plasmonic nanoprobes for SERS biosensing and bioimaging. Vol. 3, Journal of Biophotonics. Wiley; **2009**. p. 89–102. Available from: http://dx.doi.org/10.1002/jbio.200910015.

[3] Tripp RA, Dluhy RA, Zhao Y. Novel nanostructures for SERS biosensing. Vol. 3, Nano Today. Elsevier BV; **2008**. p. 31–7. Available from: http://dx.doi.org/10.1016/S1748-0132(08)70042-2.

[4] Bantz KC, Meyer AF, Wittenberg NJ, Im H, Kurtuluş Ö, Lee SH, et al. Recent progress in SERS biosensing. Vol. 13, Physical Chemistry Chemical Physics. Royal Society of Chemistry (RSC); **2011**. p. 11551. Available from: http://dx.doi.org/10.1039/c0cp01841d.

[5] Tran V, Walkenfort B, König M, Salehi M, Schlücker S. Rapid, Quantitative, and Ultrasensitive Point-of-Care Testing: A Portable SERS Reader for Lateral Flow Assays in Clinical Chemistry. Vol. 58, Angewandte Chemie International Edition. Wiley; **2018**. p. 442–6. Available from: http://dx.doi.org/10.1002/anie.201810917.

[6] Fan M, Andrade GFS, Brolo AG. A review on recent advances in the applications of surface-enhanced Raman scattering in analytical chemistry. Vol. 1097, Analytica Chimica Acta. Elsevier BV; **2020**. p. 1–29. Available from: http://dx.doi.org/10.1016/j.aca.2019.11.049.

[7] Kleinman SL, Frontiera RR, Henry AI, Dieringer JA, Van Duyne RP. Creating, characterizing, and controlling chemistry with SERS hot spots. Vol. 15, Phys. Chem. Chem. Phys. Royal Society of Chemistry (RSC); **2013**. p. 21–36. Available from: http://dx.doi.org/10.1039/c2cp42598j.

[8] Potara M, Maniu D, Astilean S. The synthesis of biocompatible and SERS-active gold nanoparticles using chitosan. Vol. 20, Nanotechnology. IOP Publishing; **2009**. p. 315602. Available from: http://dx.doi.org/10.1088/0957-4484/20/31/315602.

[9] Miljanić S, Frkanec L, Biljan T, Žinić ZMM, "Recent Advances in linear and nonlinear Raman spectroscopy I," J. Raman Spectrosc. 38(April), 1538–1553 (**2007**). Available from: https://doi.org/10.1002/(ISSN)1097-4555.

[10] Liu H, Zhang L, Lang X, Yamaguchi Y, Iwasaki H, Inouye Y, et al. Single molecule detection from a large-scale SERS-active Au79Ag21 substrate. Vol. 1, Scientific Reports. Springer Science and Business Media LLC; **2011**. Available from: http://dx.doi.org/10.1038/srep00112.

[11] Gomes PDC, Rickard JJS, Goldberg Oppenheimer P. Electrofluidynamic Patterning of Tailorable Nanostructured Substrates for Surface-Enhanced Raman Scattering. Vol. 3, ACS Applied Nano Materials. American Chemical Society (ACS); **2020**. p. 6774–84. Available from: http://dx.doi.org/10.1021/acsanm.0c01190.

[12] Huang JA, Zhang YL, Ding H, Sun HB. SERS-Enabled Lab-on-a-Chip Systems. Vol. 3, Advanced Optical Materials. Wiley; **2015**. p. 618–33. Available from: http://dx.doi.org/10.1002/adom.201400534.

[13] Quang LX, Lim C, Seong GH, Choo J, Do KJ, Yoo SK. A portable surface-enhanced Raman scattering sensor integrated with a lab-on-a-chip for field analysis. Vol. 8, Lab on a Chip. Royal Society of Chemistry (RSC); **2008**. p. 2214. Available from: http://dx.doi.org/10.1039/b808835g.

[14] LUO BS, LIN M. A PORTABLE RAMAN SYSTEM FOR THE IDENTIFICATION OF FOODBORNE PATHOGENIC BACTERIA. Vol. 16, Journal of Rapid Methods & Automation in Microbiology. Wiley; **2008**. p. 238–55. Available from: http://dx.doi.org/10.1111/j.1745-4581.2008.00131.x.

[15] Pelt DM, Sethian JA. A mixed-scale dense convolutional neural network for image analysis. Vol. 115, Proceedings of the National Academy of Sciences. Proceedings of the National Academy of Sciences; **2017**. p. 254–9. Available from: http://dx.doi.org/10.1073/pnas.1715832114.

[16] Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. 2017 International Conference on Engineering and Technology (ICET). IEEE; **2017**. Available from: http://dx.doi.org/10.1109/ICEngTechnol.2017.8308186.

[17] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Vol. 349, Science. American Association for the Advancement of Science (AAAS); **2015**. p. 255–60. Available from: http://dx.doi.org/10.1126/science.aaa8415.

[18] Khan S, Ullah R, Shahzad S, Anbreen N, Bilal M, Khan A. Analysis of tuberculosis disease through Raman spectroscopy and machine learning. Vol. 24, Photodiagnosis and Photodynamic Therapy. Elsevier BV; **2018**. p. 286–91. Available from: http://dx.doi.org/10.1016/j.pdpdt.2018.10.014.

[19] Cialla D, März A, Böhme R, Theil F, Weber K, Schmitt M, et al. Surface-enhanced Raman spectroscopy (SERS): progress and trends. Vol. 403, Analytical and Bioanalytical Chemistry. Springer Science and Business Media LLC; **2011**. p. 27–54. Available from: http://dx.doi.org/10.1007/s00216-011-5631-x.

[20] Sharma B, Fernanda Cardinal M, Kleinman SL, Greeneltch NG, Frontiera RR, Blaber MG, et al. High-performance SERS substrates: Advances and challenges. Vol. 38, MRS Bulletin. Springer Science and Business Media LLC; **2013**. p. 615–24. Available from: http://dx.doi.org/10.1557/mrs.2013.161.

[21] Chollet F, Deep Learning with Python, 1st ed. USA: Manning Publications Co. **2017**. p 123.

[22] Goldberg-Oppenheimer P, Steiner U. Rapid Electrohydrodynamic Lithography Using Low-Viscosity Polymers. Vol. 6, Small. Wiley; **2010**. p. 1248–54. Available from: http://dx.doi.org/10.1002/smll.201000060.

[23] Rickard JJS, Farrer I, Goldberg Oppenheimer P. Tunable Nanopatterning of Conductive Polymers via Electrohydrodynamic Lithography. Vol. 10, ACS Nano. American Chemical Society (ACS); **2016**. p. 3865–70. Available from: http://dx.doi.org/10.1021/acsnano.6b01246.

[24] Rickard JJS, Di-Pietro V, Smith DJ, Davies DJ, Belli A, Oppenheimer PG. Rapid optofluidic detection of biomarkers for traumatic brain injury via surface-enhanced Raman spectroscopy. Vol. 4, Nature Biomedical Engineering. Springer Science and Business Media LLC; **2020**. p. 610–23. Available from: http://dx.doi.org/10.1038/s41551-019-0510-4.

[25] Lab On a Chip Market Size, Trends & Growth: 2021 to 2026, https://www.marketdataforecast.com/market-reports/lab-on-a-chip-market, Market Data Forecast. Date access: November 17, **2019**.

[26] Peak finder script available at URL="https://github.com/PauloAxcel/Raman-Peak-Finder".

[27] Hinterwirth H, Kappel S, Waitz T, Prohaska T, Lindner W, Lämmerhofer M. "Quantifying Thiol Ligand Density of Self-Assembled Monolayers on Gold Nanoparticles by Inductively Coupled Plasma–Mass Spectrometry". Vol. 7, ACS Nano. American Chemical Society (ACS); **2013**. p. 1129–36. Available from: http://dx.doi.org/10.1021/nn306024a

[28] Kandel I, Castelli M. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset [Internet]. Vol. 6, ICT Express. Elsevier BV; **2020**. p. 312–5. Available from: http://dx.doi.org/10.1016/j.icte.2020.04.010.

[29] Zhang X, Zou J, He K, Sun J. Accelerating Very Deep Convolutional Networks for Classification and Detection. Vol. 38, IEEE Transactions on Pattern Analysis and Machine Intelligence. Institute of Electrical and Electronics Engineers (IEEE); **2016**. p. 1943–55. Available from: http://dx.doi.org/10.1109/TPAMI.2015.2502579.

[30] CNN script available at URL="https://github.com/PauloAxcel/SERS-EHD-pillars/tree/main/article".

CHAPTER 6

Saliva data and peak analysis review

This chapter serves as an introduction to the following one, "Open-source software for Raman data plotting, multivariate analysis, and spectral peak fitting." The goal of developing software capable of analysing Raman data stems from the need to identify the various ways Raman data is preprocessed and processed and, as a result, to create a unique code to manipulate Raman data. As a result, this chapter discusses various Raman data preprocessing and processing procedures for saliva sample analysis. Because there is a significant amount of interest in figuring out how saliva can be used as a bioanalyte due to how easy and non-invasive it is to collect samples, there have been a plethora of studies on identifying specific biomarkers in saliva that can translate into health problems. As a result, the Raman data analysis software was expanded to include a peak analysis component, and saliva literature was reviewed for significant peaks and relevant biomarkers for the presence of any health-related disease. As a result, the research in this chapter focuses on the findings in the literature concerning the preprocessing and processing of saliva Raman data. Parts of this chapter have been incorporated into two literature reviews related to saliva and Raman spectroscopy. The papers are the following: “Hardy, M., Kelleher, L., **de Carvalho Gomes, P.**, Buchan, E., Chu, H.O.M. and Goldberg Oppenheimer P. “Methods in Raman spectroscopy for saliva studies – a review”. Vol. 57, Applied Spectroscopy Reviews. Informa UK Limited; **2021**. p. 177–233. Available from: http://dx.doi.org/10.1080/05704928.2021.1969944 “ mainly section “9 Data analysis”. The other review paper published was: “Buchan, E., Hardy, M., **Gomes, P. de C.**, Kelleher, L., Chu, H.O.M. and Oppenheimer, P.G. “Emerging Raman spectroscopy and saliva-based diagnostics: from challenges to applications”. Applied Spectroscopy Reviews. Informa UK Limited; 2022. p. 1–38. Available from: http://dx.doi.org/10.1080/05704928.2022.2130351 “ mainly parts in section “2.3 Healthcare – Cancer types”.

6.1 Abstract

This chapter discusses the Raman data and peak analysis related to multiple saliva studies regarding cancer. One of the most important issues with Raman spectra is their analysis. Not only individual analysis but a comparative analysis between similar Raman spectra. Here a guide for different approaches to tackling Raman spectral analysis is provided with a special focusing on saliva. Saliva has been used as a biofluid since there has been a special interest in analysing its Raman spectral components and relating them to specific diseases.

Additionally, the saliva collection method is simple and non-invasive. Thus, the patient is more likely to donate saliva, and it is easier to pass ethical approvals. However, Raman intensity information from saliva is usually weaker than other forms of biofluids (blood, plasma, urine), hence the need for additional computational methods to differentiate biomolecular fingerprints of the weak saliva signals.

6.2 Methodology

A python script was developed to inspect downloaded papers in pdf file format and then detect Raman peak information in order to create the saliva Raman peak libraries. Furthermore, not only the peak values but also the sentence in which the peak value appears were saved for context about the nature of the peak. These scripts can quickly check many papers, with all peaks and their context written in a csv file. This can then be processed by the user. To interpret text written in a pdf file, the pdf file must be converted into an image using the pdf2image package[32]. Because pdf files are vectorized[33], converting a pdf to an image allows it to be interpreted by the PyTesseract module[34], a

python version of the Tesseract-OCR (Optical Character Recognition) developed by Google[35]. Using these tools, the Python script can load and interpret various articles. Because Raman peak locations are the information of interest in this study, and Raman peak information is always written with the cm^{-1} termination in the literature, a template location method[36] can be used to identify cm^{-1} sentences. By searching for this specific keyword, the code extracts the peak and its context. A table of Raman peaks, their context (where the peak came from, what the peak is related to), and the paper to which they belong, can be generated, quickly read through, and processed as has been done in this work. The script for the complete code can be accessed through [37].

6.3 Sample size influence over Machine Learning analysis

This section covers the data analysis used throughout many publications related to the Raman spectra of healthy/unhealthy saliva patients. Data analysis can be performed using different machine learning (ML) models involving advanced statistical analysis. These models help differentiate minute features between healthy and unhealthy patients that are difficult to distinguish through direct observation and manual comparison. The results coming from these models are interpreted, and relevant inferences can be made regarding specific features that classify the different health conditions. However, many considerations are required when dealing with ML. One of the most important is the sample size, i.e., the different number of patients. These patients can be separated according to their healthy versus unhealthy conditions, and often subsets of unhealthy patients arise, such as degrees of infection or different cancerous stages. Afterwards, the number of samples is split into training and testing samples. To train an efficient ML model it requires a sufficiently large number of samples to have meaningful results. There are

many difficulties in obtaining a large number of samples in the medical setting, mainly because the number of patients with a specific cancer type and without much in-class variation is often small.

In the end, only a small sample size is gathered and used. According to [1], the ideal sample size per class is around 75-100 to get meaningful results. The importance of having a large enough sample size is to reduce the overfitting and improve the differences between classes. Overfitting means the inability of the model to generalise, i.e. the model becomes biased towards the training samples. Thus, it is unable to interpret samples from outside of that training sample space. Overfitting is a significant challenge when dealing with small sample size and detecting when it happens is not trivial. Therefore, getting the largest possible sample size to train a ML model is essential for accurate predictions.

From the papers analysed[1-21], it is clear there is a great degree of variation between sample sizes', some from only 20 total samples (10 healthy/10 unhealthy) to a larger sample size of 128 (64 healthy/64 unhealthy). However, most of the large sample size data is gathered from Raman data banks online [2-5], which are easier to access than obtaining samples directly from patients (**Figure 6.1**). The number of samples used for data analysis can drastically change the overall results of the analysis. For instance, samples with only 20 samples (10 healthy/10 unhealthy) [6-8] will have overfitting complications in their machine learning models due to the lack of training and test data.

The following example shows how a small dataset affects model overfitting.

The effect of a small dataset on the overfitting of a model can be illustrated by the following example. To train/test a model, the initial sample needs to be split into train/test data, usually in the 80%/20% proportions. If there are only ten different patient samples,

the number of train data would be eight, and the test data would be two. The most common way to train the model is using the cross-validation (CV) algorithm[22].

The model will be trained by consecutively splitting the train data into training and validating data. CV is associated with an integer that determines how the dataset is split. For instance, from the eight training data, six will be used to train the model and two will be used to validate the model. This action will be repeated until the model has seen all the different validation data. Finally, after the CV, the model is tested on the two samples from the testing dataset (note that the model has never seen these samples). The accuracy of the model can be extracted from the testing samples.

When dealing with machine learning for diagnostic applications, some questions arise. If a model is well-trained, a train with only eight samples can be used to determine the impact of overfitting. What is the performance of a model when only two samples are available? How well do two samples from different patients represent the entire sample space of a patient? Is it possible to extrapolate conclusions from two samples to the entire population? Will future patients believe predictions made by a machine trained on only two samples? The answer to both of these questions is unequivocally no. However, the negative aspects should not be the primary consideration. They must be respected, but the vital point is to focus on the positives and the potential that these new ML techniques can have in assisting with disease diagnostics, therapeutics, and monitoring. The only drawback is the small sample size, which, if overcome, will increase patients' and clinicians' trust in ML diagnostics.

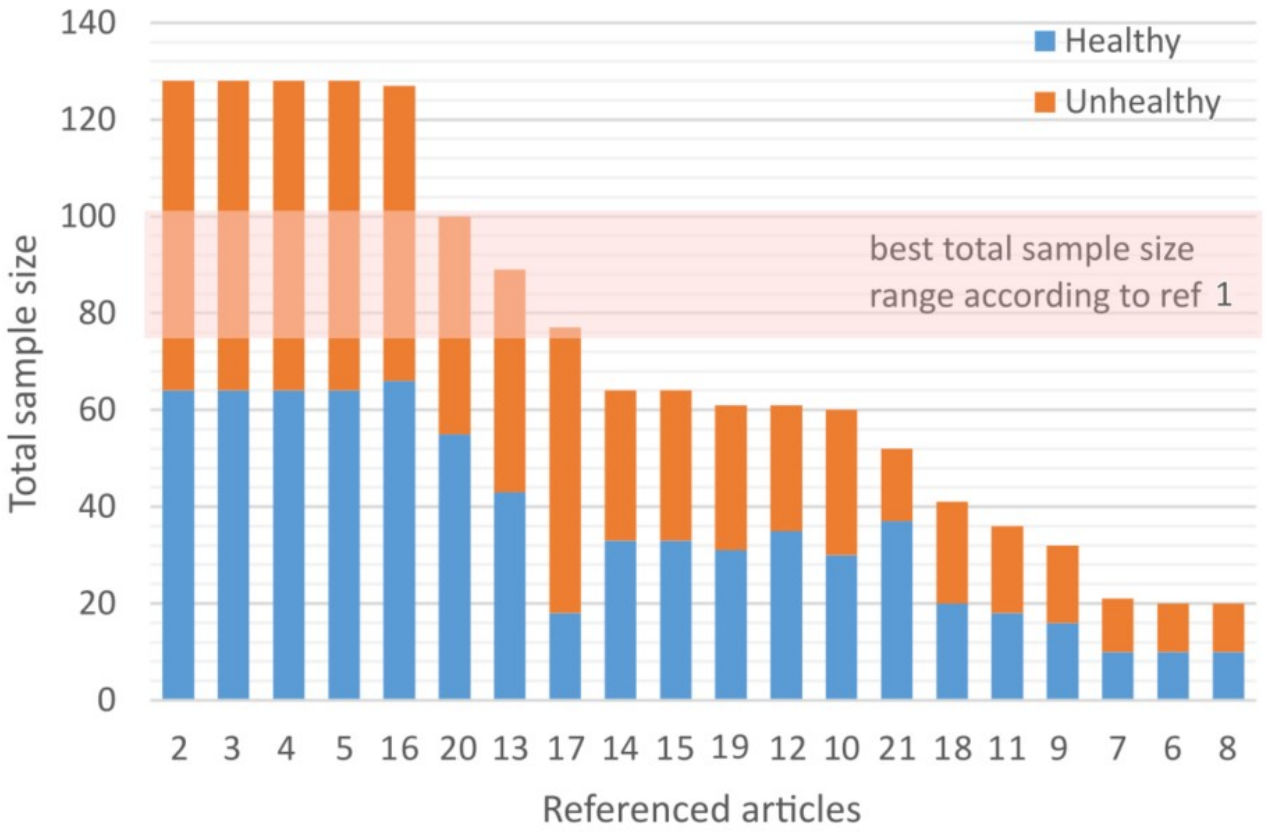


***Figure 6.1**: Stacked bar plot of the variation in sample size between different research papers. In orange, the number of unhealthy samples and the number of healthy samples in blue.*

Many other factors can affect the way that ML impacts the sample analysis. Such as the intrinsic differences between the same group of patients, i.e. the patient's gender, ethnicity, and age, are all crucial factors that can slightly change the same group data samples (increase within-class variance)[23]. These changes can be further amplified when small sample size is required to train. The outcomes/conclusions will be skewed, which is undesirable for a generalised model. Another nuance that can affect the ML model is that the preprocessing done on the data before an ML analysis can also impact ML diagnostics. Great care needs to be given when applying any aggressive changes to the raw data, smoothing/denoising/filters and thus, should be avoided.

Going into detail about the separation of samples into training and testing data and how the training was validated using a CV algorithm. After the ML model is trained, it

needs to be tested on data that the model has never seen before. A set of statistical quantities like sensitivity, specificity, and accuracy measures how well the model gets the correct identifications. The relation between the sample size and the statistical parameters is observed in **Figure 6.2**. The larger the dataset, the better the statistical parameters become. However, the trend is not perfectly linear since there are cases where a small dataset (30) reaches 100% accuracy, but a large dataset (130) can only reach values between 90-100% of accuracy. Although for the case with a small sample size, getting 100% per cent accuracy might indicate a model bias to the dataset, and the results should be rejected until more data is available. As a golden rule for ML results interpretation, if the sample size is small and the statistical parameters are in the 100% range, the ML model is most likely biased, and the results did not generalise. The solution to this issue is to acquire more data. Besides these methods, using cross-validation when training is always an option; however, the results will again be biased if the data size is small.

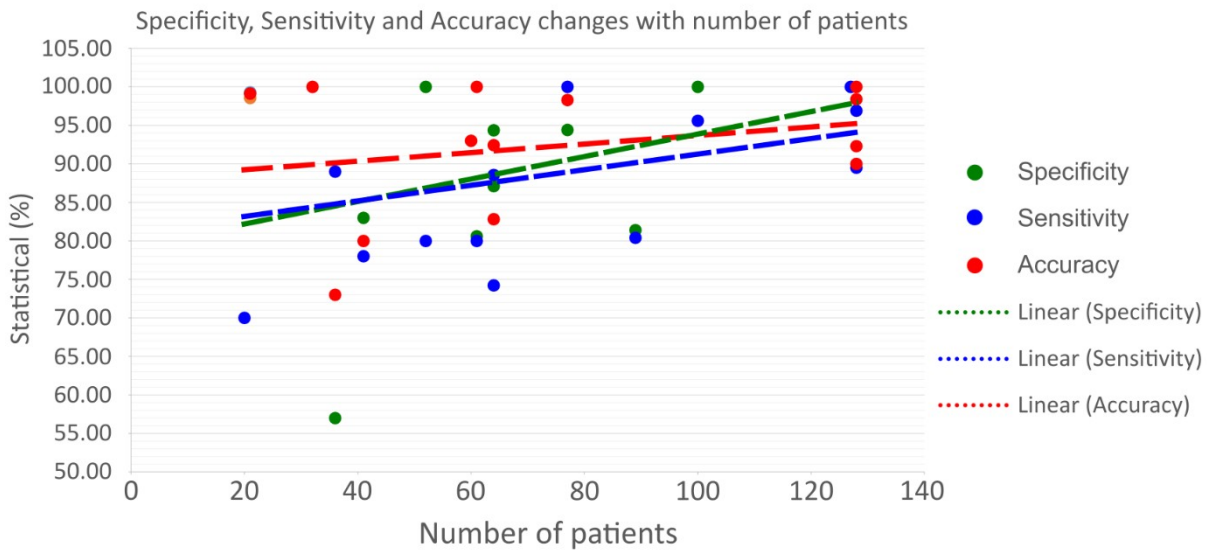


***Figure 6.2**: Specificity, sensitivity, and accuracy variation with a different number of patients. Increase the statistical values with an increasing number of patients (healthy + unhealthy). Sensitivity describes the portion of true positives correctly identified by a diagnostic test. Specificity is the proportion of the true negatives correctly identified in a diagnostic test. Accuracy is the proportion of accurate results, in a population.*

6.4 Different data processing methods

Finally, we summarise the data processing done after data acquisition until the ML diagnosis result (**Figure 6.3A-G**). The different data processing varies between authors, but most prefer to follow the red line in **Figure 6.3** schematics. The data processing starts with an initial baseline subtraction (**Figure 6.3B**), followed by a vector normalisation of the spectrum (**Figure 6.3C**). Next, a preprocessing (**Figure 6.3D**) of the Raman data, usually with a principal component analysis (PCA) treatment, rejecting the components with the least significance and keeping the most important ones for the multivariable analysis (**Figure 6.3E**). The ML model choice was split among authors from least discriminant analysis (LDA), artificial neural networks (ANN) to support vector machine (SVM), all three showing decent results (90%-100% accuracy), with the ANN being the best. The model (LDA/SVM/ANN) is then trained/tested by first performing a data split between 20% for testing data and 80% for training data and then scored using the leave-one-out cross validations (LOOCV) classification algorithms (**Figure 6.3F**). From the different classification results (**Figure 6.3G**), accuracy, sensitivity, and specificity can be extracted, accuracy being the best indicator of the quality of the generated model. Complementary to these results, most authors also present a receiver operating characteristic (ROC) curve.

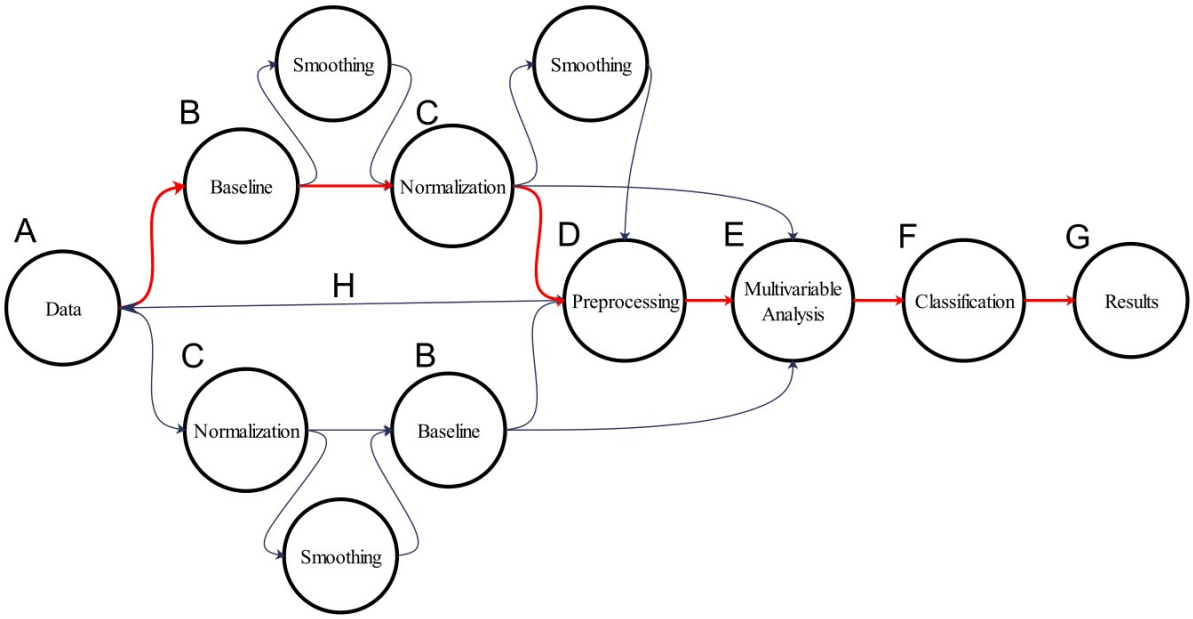


***Figure 6.3**: Most relevant procedures for Raman data processing and analysis. In red, showing the most common data processing path among researchers. A: Data collection, B: baseline followed by normalisation, C: normalisation followed by baseline, D: data preprocessing, E: machine learning multivariable analysis, F: classification of the machine learning model, G: results and H: loadings.*

Going in-depth for each data analysis procedure, for the baseline subtraction, most authors opt for an asymmetric least square smoothing (AsLS) [7,9-12] or a high order polynomial fit [2,13-16,24]. Next, one author performed a smoothing after baseline subtraction [2]. This step was avoided by the remaining authors mainly because any smoothing would affect the true nature of the Raman peaks. After the baseline subtraction, most groups perform an area normalisation [7,9,10,13,15], others normalise the data to the maximum intensity value [2,14,15,24]. Normalisation is essential for comparing Raman spectra from different samples on an equal footing. Normalisation closes the intensity gap between different spectra, allowing for a more accurate comparison of peak intensities. If no normalisation was performed, the ML model would distinguish primarily based on the difference in intensity (samples with more vs. less intensity signal) rather than the minute differences between samples (same peaks have lower vs. higher intensities). It is up to the user to decide whether to perform a normalisation based on maximum peak area or

maximum peak intensity, as the normalisation affects subsequent multivariate analysis procedures and thus the final results and conclusions. Also, a small minority of authors [17,18] prefer first to use a normalisation followed by baseline subtraction. Although it is not the norm, the results might be that the same values between different groups are still comparable. However, it is always preferred to normalise after the baseline correction since the intensity values are consistently between zero and one for a maximum intensity normalisation. Or the area under the curve between zero and one for the area normalisation. After the normalisation, two authors [10,24] perform smoothing of the baseline subtracted normalised spectra. They perform a Savitzky-Golay filter smoothing. For data analysis, smoothing should be avoided, there are occasionally cases where the spectra are very noisy, and if it is impossible to acquire more measurements from the sample, then a low factor smoothing can be performed as a last resort. Most authors, however, avoid smoothing altogether and proceed into preprocessing.

The preprocessing step serves to simplify the corrected Raman spectra data into a low dimensionality space. Most authors perform a principal component analysis (PCA) [4-6,11-13,15,18,19,24] over the Raman data. They also perform a complementary t-test over the principal components (PC) in order to select the best components. Other authors select the number of components until the variance sums reach 90%-95%. Besides the PCA, a group performed a genetic algorithm (GA) [7] which selected the most important features from different Raman spectra. GA is a machine learning algorithm based on the theory of evolution, where features from the Raman spectra that dominate over others will be stored and the others ignored. Only the important peaks will be used when feeding into the ML algorithm.

The preprocessing step is solely used to help the computational load and thus can be avoided. It is challenging to figure out whether decreasing the complexity of our data

can improve the diagnostic capabilities of our model, i.e. if less and condensed data is better than more but scattered data. As a result, even if the computation process takes longer, the improvement in model accuracy may be overall beneficial, so the ML algorithm should use the entire set of components. After the data preprocessing it follows the multivariable analysis. Here the PCA preprocessed data as input for the different ML algorithms. The input data needs to be split into training and testing data. The most common separation is 80%/20% [2,3,9,10,12], however there were some groups that did 90%/10% [20] and 60%/40% [17]. For the ML algorithm choice, some authors prefer to use LDA [5,6,11,13,18,19], others ANN [4,7,9] and others SVM [10,16,17,20,24]. ML models' function is to learn patterns in the feed-in data. Initially, the choice of LDA or SVM is perfectly fine. If the output results are not adequate, ANN can be used to solve more complex multivariate analysis problems; however, it needs developed intuition to interpret the results. To finalise, the model needs to be classified. If ANN is used, a choice for the optimiser, loss function and metrics are required.

The choice for the optimiser and loss function is very much dependent on the case study. Usually, for healthy/unhealthy patients, the binary cross-entropy for the loss function is optimal since it is optimised for two outcomes. For the metrics, accuracy is the most used parameter. The optimiser function adjusts the differences in internal weights according to two consecutive iterations scores using a Backpropagating algorithm. The goal of the optimiser is to lower the loss score. When the loss score plateaus, the optimisation is done, and the metrics can be calculated [4,7,9]. For the LDA and SVM, the classification method of choice is the LOOCV [7,11,13-17,20,24]. It is the most common method for classifying an ML model. For the sensitivity, specificity and accuracy results, considering the number of available patients is required when evaluating these statistical

results, as already mentioned in **Figure 6.2**. The best results were obtained for ANN, although both LDA and SVM were acceptable.

It is essential to have an accurate diagnosis from ML analysis and identify potential biomarkers specific to the patients' condition. The following section is dedicated to the most common healthy and unhealthy saliva biomarkers that can potentially identify changes between different patient groups. As described in **Figure 6.3 D→A**, when the preprocessed data is interpreted back into Raman data, this new Raman data will be slightly modified and different from standard Raman data. However, it can still be analysed and provide unique insight that can potentially open new diagnosis possibilities or treatments. Some authors have performed a GA to track the most important peaks [7,9,12]. Others convert the principal components from the PCA analysis into respective loadings. The loadings can easily be calculated by matrix multiplication of the eigenvectors with the square root of the eigenvalues. The specific biomarkers gather from this analysis are related to the saliva of unhealthy patients (acute myocardial infarction [13,19], cancer [6,11,12,14,15,16-18,21,25], viruses [2-5,20,26], Alzheimer [7]) or others state altering condition (smokers [21], drugs [16,25,27-30], radiation [8]) and healthy patients.

6.5 Saliva peak analysis

Below a statistical analysis of the biochemical content in saliva is presented related to the previous publications[6-13,15-18,31], complementing it with a detailed comparison of which peaks are effectively changing from healthy to unhealthy patients[6,8,11,13,15-17].

Most of the peaks present in saliva are related to protein/amino acid differences. Overall, 50% of the peaks assignment in the literature related to saliva from cancerous

patients are from proteins [6-13,15,16-18,31]. Almost 29% are peaks associated with other biomolecules, like antioxidants, amino sugars and other vibration bonds with no indication of the molecule related to [1-6,18,24,26], 9% to nucleic acid [6,7,9,13,16-18,31], 7% to lipids [6-8,11,15,17,18,31] and 5% to sugars [6-9,11,12,16-18]. From the protein biomolecule group, the most mentioned chemical vibration was assigned to phenylalanine with 33 mentions, followed by tyrosine (31), tryptophan (26), amide III (tertiary) and I (primary) (11,10) (**Figure 6.4 red**). For the nucleic acids, cytosine (6), adenine (5), guanine (5), thymine (5), guanosine and uracil (2,2). For lipids, phospholipids (6), cholesterol (2) and phosphatidylcholine (2) and lastly, for sugars, D-glucose (7), glycine (2), D-mannose (2). Other, less frequently mentioned biomolecules found in saliva (1) were not considered in the statistical analysis. Furthermore, when dealing with a specific group of biomolecules, occasional peak assignments were ambiguous and were left as 'unspecified' or put into the mixed group if they were related to ambiguous bond vibrations.

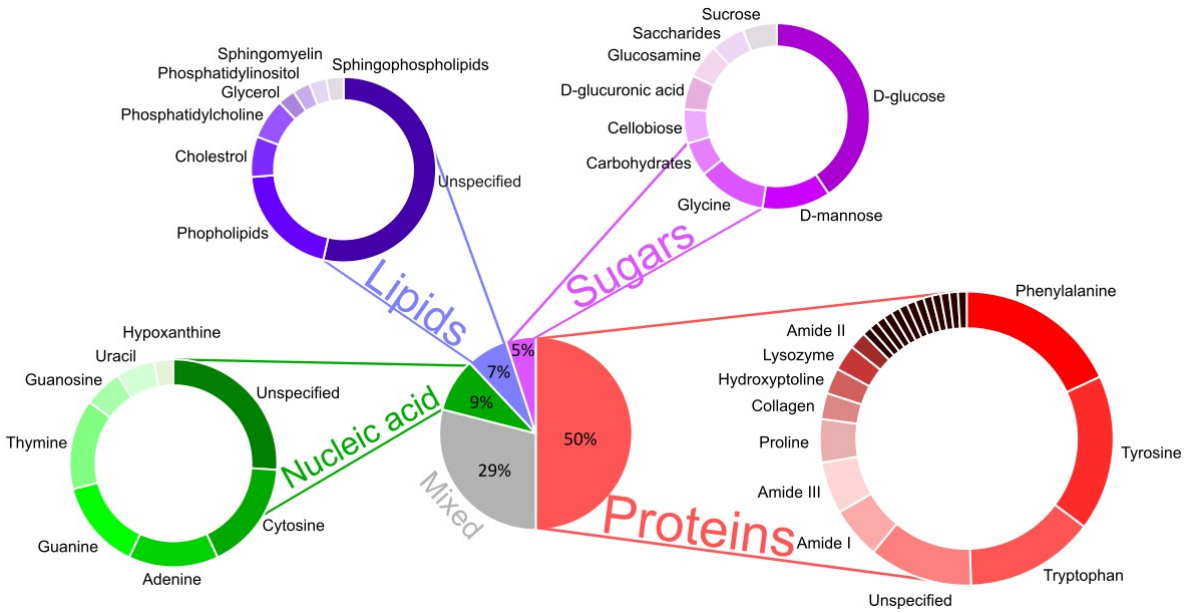


***Figure 6.4**: Raman peak assignment on cancerous saliva samples. Their biomolecular group displays a doughnut pie chart of the most prominent peaks in the literature.*

A whisker plot was constructed from all the Raman peaks found in the literature (**Figure 6.5**). The x-axis is the Raman shift locations of the different biomolecules with respective whisker plot statistics (quartiles, median, max and min). In the literature, some authors only mentioned the general biomolecule (e.g. protein) without giving the specific amino acid or the specific protein vibration peak. A generic protein label was incorporated into the whisker plot despite lacking specific information. Furthermore, in **Table 6.1**, the peaks can be traced back to their source.

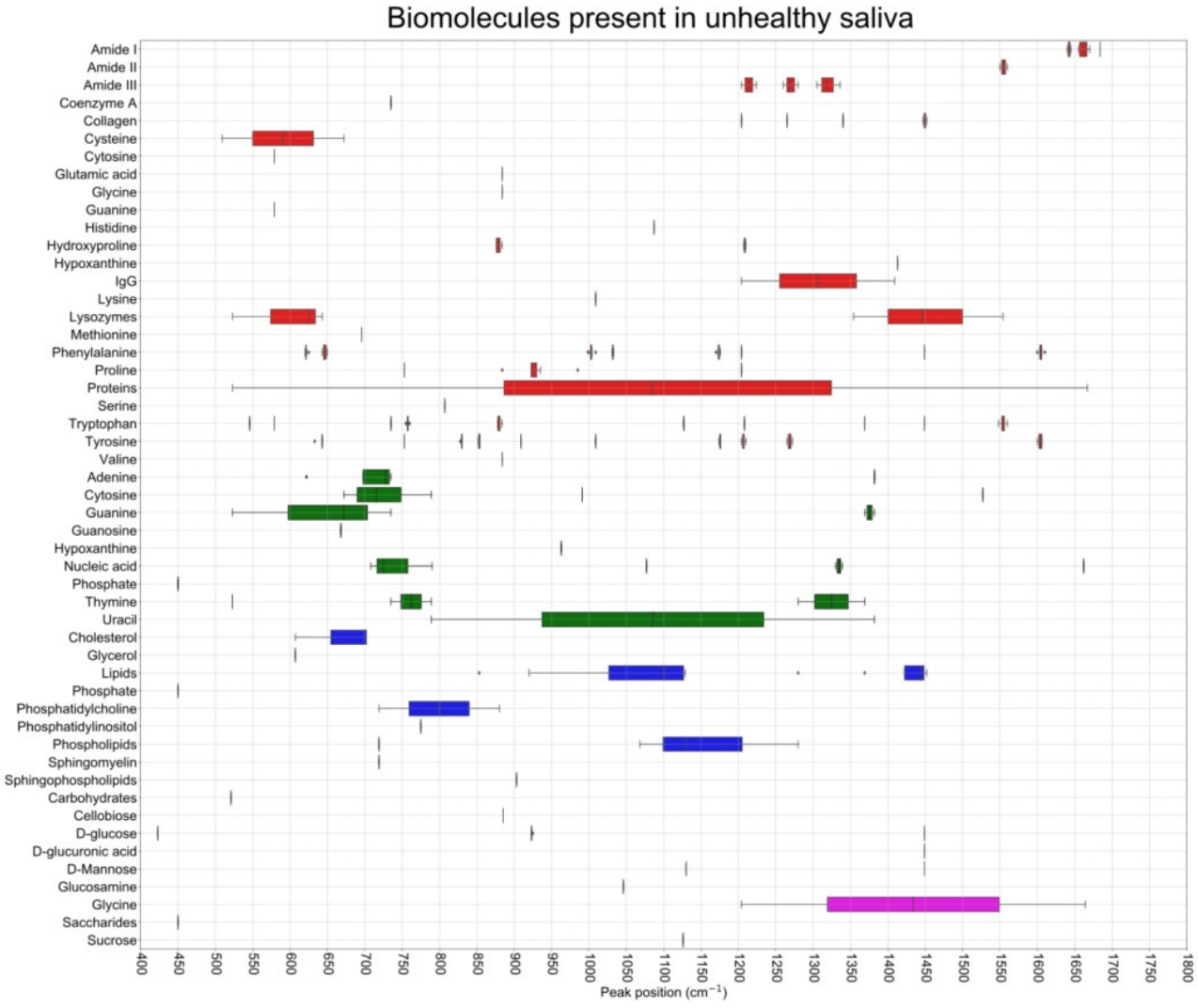


***Figure 6.5**: List of biomolecules present in saliva from cancerous patients with respective whisker plots for different Raman peak regions. Red is related to protein molecules, green nucleic acid, magenta sugars and blue lipids. When the peak assignment is singular, a stripe represents the symbol—complemented with **Table 6.1**.*

The data presented here span the 400-1800cm^{-1} region of Raman shifts. This region is the most common when analysing biological systems like saliva. Many different amino acids are present throughout separate Raman shift regions regarding the biomolecular content, and even some share the same wavenumber. Due to this peak degeneracy, finding a specific contribution to a peak can be complicated, so most authors present an attempt peak label, which might not be accurate. This situation not only serves to illustrate the complexity of peak labelling on a complex biological system like saliva but also seeks to provide a peak label roadmap for saliva.

***Table 6.1**: Raman biomolecule information from cancerous patient saliva with respective Raman region.*

Protein Tentative assignment	Peak position (cm^{-1})
Amide I	1640[6],1645[31],1655[10,7],1661[8],1664[12],1667[7],1670[13],1684[15]
Amide II	1550[9],1555[13],1560[17]
Amide III	1204[11],1224[11],1260[6],1265[15],1268[16],1275[11],1280[18],1305[31],1318[10],1336[12]
Coenzyme A	735[18]
Collagen	1204[11],1265[15],1340[15],1447[15],1452[8]
Cysteine	509[13],672[16]
Cytosine	579[17]
Glutamic acid	884[18]
Glycine	884[18]
Guanine	579[17]
Histidine	1087[9]
Hydroxyproline	876[13,15],880[8],884[18],1207[7],1210[13]
Hypoxanthine	1413[7]
IgG	1204[11],1409[11]
Lysine	1009[18]
Lysozyme	523[18],625[7],643[7],1354[9],1446[8],1554[7]
Methionine	696[18]
Phenylalanine	620[8],621[7,10,15],622[18],625[7],643[15],650[8],999[16,17],1002[6,8,13],1003[9,10],1004[15,31],1009[18],1031[13],1032[10],1033[15],1170[6],1174[9,10],1176[7],1204[11],1449[16],1600[16],1603[17],1605[7],1606[10],1610[6]
Proline	753[7],884[18],923[7,16,17],925[18],935[15],985[7],1204[11]
Proteins	523[18],622[18],723[17],732[16],853[31],920[31],1047[8],1049[15],1077[18],1084[6,15],1125[31],1280[18],1369[18],1440[31],1449[13,16],1667[7]
Serine	807[6]
Tryptophan	546[17],579[17],735[18],755[13],757[7],758[9,10],760[15],878[9,10],880[7],884[18],1126[7],1127[9,10],1208[8,15],1369[18],1449[9,10],1548[6],1552[15],1554[7],1556[9,10],1560[17]
Tyrosine	633[6],643[7,9,10,13],753[7],828[13],830[7,9,10],852[16,17],853[13],854[9,10,15],909[18],1009[18],1174[10],1176[7,15],1204[11],1206[10],1207[7],1210[13],1265[9],1272[7],1600[16],1603[17],1606[9,10]
Valine	884[18]

Nuclei Acid Tentative assignment	Peak position (cm^{-1})
Adenine	622[18],723[17],732[16],735[18],1382[17]
Cytosine	672[16],696[18],735[18],789[18],991[9],1527[17]
Guanine	523[18],672[16],735[18],1369[18],1382[17]
Guanosine	668[31]
Hypoxanthine	963[7]
Unspecified	450[31],708[6],725[6],790[6],1077[18],1330[13],1339[6],1662[6]
Thymine	523[18],735[18],789[18],1280[18],1369[18]
Uracil	789[18],1382[17]

Lipids Tentative assignment	Peak position (cm^{-1})
Cholestrol	607[7],702[7]
Glycerol	607[7]
Unspecified	853[31],920[31],1062[7],1077[18],1125[31],1126[11],1129[6],1280[18],1369[18],1440[31],1447[15],1452[8]
Phosphatidycholine	719[7],880[7]
Phosphatidylinositol	775[7]
Phospholipids	450[31],719[7],1068[7],1130[17],1280[18]
Sphingomyelin	719[7]
Sphingophospholipids	903[7]

Sugars Tentative assignment	Peak position (cm^{-1})
Carbohydrates	521[9]
Cellobiose	885[11]
D-glucose	423[16,17],923[7,16,17],925[18],1449[17]
D-glucuronic acid	1449[17]
D-mannose	1130[6],1449[17]
Glucosamine	1046[16]
Glycine	1204[11],1664[12]
Saccharides	450[8]
Sucrose	1126[11]

To conclude this section, we need to approach the specific peak that changes when dealing with unhealthy and healthy patients. The peak changes can be characterised by intensity changes, Raman shift changes or both. Here we outline the most prominent peak changes. Authors in [13] have shown that tryptophan peak at 755 cm^{-1}, phenylalanine at 1002 cm^{-1} and C-CH3 bond peak at 1047 cm^{-1} increase for unhealthy patients. Other peaks for tyrosine at 1210 cm^{-1} and 1670 cm^{-1} amide I decreased for unhealthy patients. Authors in [11] showed that the peak of 1224 cm^{-1} and 1275 cm^{-1} from amide III, 1409 cm^{-1} from IgG and 1417 cm^{-1} quinoid ring increased for unhealthy patients, whereas the peak of 885 cm^{-1} from cellobiose, 1126 cm^{-1} from lipids/proteins/sucrose and 1204 cm^{-1} from amide III, collagen, tyrosine, phenylalanine, IgG, proline and glycine decreased. Authors in [15] showed that 1004 cm^{-1} phenylalanine, 1208 cm^{-1} tryptophan, 1265 cm^{-1} amide III and collagen, 1340 cm^{-1} collagen, and 1684 cm^{-1} amide I increase for unhealthy patients, whereas 621 cm^{-1} phenylalanine and 1176 cm^{-1} tyrosine increases for healthy patients. In paper [17] the 423 cm^{-1} and 1450 cm^{-1} of D-glucose, 548 cm^{-1} tryptophan, 579 cm^{-1} tryptophan, cytosine and guanine and 717 cm^{-1} adenine increase for unhealthy patients. Whereas 853 cm^{-1} tyrosine, 999 cm^{-1} phenylalanine, 1130 cm^{-1} phospholipids, 1382 cm^{-1} adenine, guanine and uracil, 1527 cm^{-1} cytosine and 1603 cm^{-1} tyrosine and phenylalanine decreases. For the author, in [16] peak of 676 cm^{-1} from cysteine, cytosine and guanine, 854 cm^{-1} of proline and tyrosine, 1000 cm^{-1} of phenylalanine, 1046 cm^{-1} of glucosamine, 1263 cm^{-1} of amide III and 1604 cm^{-1} of phenylalanine increases for unhealthy patients, whereas peak 423 cm^{-1} of D-glucose and 727 cm^{-1} of adenine decreases. The author in [6] describes an increase in peak 1167 cm^{-1} glutathione, 1170 cm^{-1} phenylalanine, 1147 cm^{-1} in carotenoids, and 1339 cm^{-1} and 1662 cm^{-1} related to nucleic acids. Author [19] showed increase in peak 1002 cm^{-1}, 1330 cm^{-1}, 1449 cm^{-1} and 1670 cm^{-1} related to amino acids. The author in [8] noticed an increase in the 2107 cm^{-1} of thiocyanate for patients after

radiation treatment. Finally, authors in [12] showed an increase in peak 1336 cm^{-1} of amide III. A list of all these peaks is provided in **Table 6.2** below.

***Table 6.2**: Saliva biomolecular content change for the unhealthy patient. The change delta is in comparison with the healthy spectra. Arrow up: Condition spectra increases in that wavenumber. Arrow down: Condition spectra decrease in that wavenumber.*

Raman shift (cm^{-1})	Assignment	Δ	Condition
422[17]	D-glucose	↑	Lung cancer
423[16]	D-glucose	↑	Lung cancer
548[17]	Tryptophan	↑	Lung cancer
579[17]	Tryptophan, cytosine, guanine	↑	Lung cancer
621[15]	Phenylalanine	↓	Breast tumor
635[17]	C-S twist	↓	Lung cancer
636[16]	C-H torsion, COO-wag; O-C=O in plane deformation; C-C-C in phase deformation	↓	Lung cancer
676[16]	Cysteine, cytosine, guanine	↓	Lung cancer
717[17]	Protein, adenine	↑	Lung cancer
727[16]	Protein, adenine	↑	Lung cancer
755[13]	Tryptophan	↑	Acute myocardial infarction
853[17]	Tyrosine, proline ring	↓	Lung cancer
854[16]	Tyrosine, proline ring	↓	Lung cancer
885[11]	Cellobiose	↓	Oropharyngeal cancer
923[17]	Proline ring, glucose, lactic acid	↓	Lung cancer
999[17]	Phenylalanine	↓	Lung cancer
1000[16]	Phenylalanine	↓	Lung cancer
1002[13]	Phenylalanine	↑	Acute myocardial infarction
1004[15]	Phenylalanine	↑	Breast tumor
1034[17]	Stretching vibration of the ring, deformation in plane C-H	↓	Lung cancer
1046[16]	Glucosamine	↓	Lung cancer
1047[13]	C-CH3	↑	Acute myocardial infarction
1049[15]	Proteins	↓	Breast tumor
1062[6]	Lipids	↑	Oral cancer
1084[15]	Proteins	↓	Breast tumor
1126[11]	Lipid, protein, disaccharide, sucrose	↓	Oropharyngeal cancer
1129[6]	Lipids	↑	Oral cancer
1130[17]	Phospholipids	↓	Lung cancer
1147[6]	Carotenoids	↑	Oral cancer
1167[6]	Glutathione	↑	Oral cancer
1170[6]	Tyrosine	↑	Oral cancer
1176[15]	Tyrosine	↓	Breast tumors
1204[11]	Amide III, collagen, tyrosine, phenylalanine, IgG, proline, glycine	↓	Oropharyngeal cancer
1208[15]	Tryptophan	↑	Breast tumors
1210[13]	Tyrosine	↓	Acute myocardial infarction
1224[11]	Amide III	↑	Oropharyngeal cancer
1247[17]	Plane deformation of C-H	↓	Lung cancer
1260[6]	Amide III	↑	Oral cancer
1263[16]	Amide III	↓	Lung cancer
1265[15]	Amide III, collagen	↑	Breast tumors
1275[11]	Amide III	↑	Oropharyngeal cancer
1339[6]	Nucleic acids	↑	Oral cancer
1340[15]	Collagen	↑	Breast tumors
1382[17]	Adenine, Guanine, Uracil	↓	Lung cancer
1409[11]	COO-IgG	↑	Oropharyngeal cancer
1417[11]	Quinoid ring	↑	Oropharyngeal cancer
1450[17]	D-glucose, D-glucuronic acid, D-mannose	↑	Lung cancer
1527[17]	Carotenoids, cytosine	↓	Lung cancer
1559[17]	Tryptophan, amide II	↓	Lung cancer
1603[17]	Tyrosine, phenylalanine	↓	Lung cancer
1604[16]	Tyrosine, phenylalanine	↓	Lung cancer
1610[6]	Tryptophan	↑	Oral cancer
1640[6]	Amide I	↑	Oral cancer
1662[6]	Nucleic acids	↑	Oral cancer
1670[13]	Amide I	↓	Acute myocardial infarction
1684[15]	Amide I	↑	Breast tumors
2107[8]	Thiocyanate	↑	Radiation

Interestingly, the Raman peak changes from unhealthy to healthy are varied. There is no single indicator among different Raman peaks to predict whether a patient is healthy or unhealthy. Even for the same peak, for example, phenylalanine 1000-1004cm^{-1}, patients with breast cancer had an increase in its signal, but patients with lung cancer had a

decrease (**Table 6.2**). This shows there is no notable trend amongst cancer patients, and depending on the cancer location, it can lead to different biomolecules expression in saliva. Thus, further increasing the complexity of using saliva as a biofluid for Raman diagnostics.

6.6 Conclusion

In this chapter the various methodologies used for data analysis were examined, including not only data processing but also experimental planning. It is fundamental to have a good sample size that represents the population. It was shown that 75 different donor samples are already a good enough start to have statistical significance for Raman data analysis. However, weighing all the different in-class variations inside that population is complicated, and sometimes the only remedy is to mitigate and work with what is possible. Additionally, getting many people for these tests requires specific features for the test patients and additional ethical bureaucratic paperwork that might not be accessible to everyone, so it is a complicated task. However, using saliva as an analyte can remove some of those hindrances, thus why the choice of saliva.

Additionally, the fundamental Raman analysis provided in this chapter shows a whole spectrum of detailed analytes that can be tracked through saliva and multiple potential applications for diagnostics. A thorough Raman data processing analysis was registered when analysing saliva Raman data. A complete description was given of the different preprocessing, processing, classification and results, which are standard in multivariate analysis, especially for saliva as a biofluid. Furthermore, the tracking of specific important peaks that can reveal a variety of illnesses is easily accessible through the multiple multivariate analysis exposed here.

6.6 References

[1] Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. Vol. 760, Analytica Chimica Acta. Elsevier BV; **2013**. p. 25–33. Available from: http://dx.doi.org/10.1016/j.aca.2012.11.007.

[2] Othman NH, Lee KhuanY, Radzol ARM, Mansor W, Rashid URM. K-Nearest Neigbour: Detection of NS1 from SERS spectra of adulterated saliva. 2016 IEEE Region 10 Conference (TENCON). IEEE; **2016**. Available from: http://dx.doi.org/10.1109/TENCON.2016.7848314.

[3] Othman NH, Lee KhuanY, Radzol ARM, Mansor W, Rashid URM. Optimal PCA-EOC-KNN Model for Detection of NS1 from Salivary SERS Spectra. 2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS). IEEE; **2018**. Available from: http://dx.doi.org/10.1109/ICIIBMS.2018.8549984.

[4] Othman NH, Y. Lee K, Radzol ARM, Mansor W. TERMINATION CRITERION FOR PCA WITH ANN FOR DETECTION OF NS1 FROM ADULTERATED SALIVA. Vol. 78, Jurnal Teknologi. Penerbit UTM Press; **2016**. Available from: http://dx.doi.org/10.11113/jt.v78.9044.

[5] Othman NH, Lee KY, Radzol ARM, Mansor W, Ramlan NNM. Linear discriminant analysis for detection of salivary NS1 from SERS spectra. TENCON 2017 - 2017 IEEE Region 10 Conference. IEEE; **2017**. Available from: http://dx.doi.org/10.1109/TENCON.2017.8228352.

[6] Brindha E, Rajasekaran R, Aruna P, Koteeswaran D, Ganesan S. Raman spectroscopy of bio fluids: an exploratory study for oral cancer detection. Alfano RR, Demos SG, editors. SPIE Proceedings. SPIE; **2016**. Available from: http://dx.doi.org/10.1117/12.2212684.

[7] Ralbovsky NM, Halámková L, Wall K, Anderson-Hanley C, Lednev IK. Screening for Alzheimer's Disease Using Saliva: A New Approach Based on Machine Learning and Raman Hyperspectroscopy. Vol. 71, Journal of Alzheimer's Disease. IOS Press; **2019**. p. 1351–9. Available from: http://dx.doi.org/10.3233/JAD-190675.

[8] Colceriu-Șimon, Hedeșiu, Toma, Armencea, Moldovan, Știufiuc, et al. The Effects of Low-Dose Irradiation on Human Saliva: A Surface-Enhanced Raman Spectroscopy Study. Vol. 9, Diagnostics. MDPI AG; **2019**. p. 101. Available from: http://dx.doi.org/10.3390/diagnostics9030101.

[9] Al-Hetlani E, Halámková L, Amin MO, Lednev IK. Differentiating smokers and nonsmokers based on Raman spectroscopy of oral fluid and advanced statistics for forensic applications. Vol. 13, Journal of Biophotonics. Wiley; **2019**. Available from: http://dx.doi.org/10.1002/jbio.201960123.

[10] Muro CK, de Souza Fernandes L, Lednev IK. Sex Determination Based on Raman Spectroscopy of Saliva Traces for Forensic Purposes. Vol. 88, Analytical Chemistry. American Chemical Society (ACS); **2016**. p. 12489–93. Available from: http://dx.doi.org/10.1021/acs.analchem.6b03988.

[11] Connolly JM, Davies K, Kazakeviciute A, Wheatley AM, Dockery P, Keogh I, et al. Non-invasive and label-free detection of oral squamous cell carcinoma using saliva surface-enhanced Raman spectroscopy and multivariate analysis. Vol. 12, Nanomedicine: Nanotechnology, Biology and Medicine. Elsevier BV; **2016**. p. 1593–601. Available from: http://dx.doi.org/10.1016/j.nano.2016.02.021.

[12] Maitra I, Morais CLM, Lima KMG, Ashton KM, Date RS, Martin FL. Raman spectral discrimination in human liquid biopsies of oesophageal transformation to adenocarcinoma. Vol. 13, Journal of Biophotonics. Wiley; **2019**. Available from: http://dx.doi.org/10.1002/jbio.201960132.

[13] Cao G, Chen M, Chen Y, Huang Z, Lin J, Lin J, et al. A potential method for non-invasive acute myocardial infarction detection based on saliva Raman spectroscopy and multivariate analysis. Vol. 12, Laser Physics Letters. IOP Publishing; **2015**. p. 125702. Available from: http://dx.doi.org/10.1088/1612-2011/12/12/125702.

[14] Wu W, Gong H, Liu M, Chen G, Chen R. Noninvasive breast tumors detection based on saliva protein surface enhanced Raman spectroscopy and regularized multinomial regression. 2015 8th International Conference on Biomedical Engineering and Informatics (BMEI). IEEE; **2015**. Available from: http://dx.doi.org/10.1109/BMEI.2015.7401503.

[15] Feng S, Huang S, Lin D, Chen G, Xu Y, Li Y, et al. Surface-enhanced Raman spectroscopy of saliva proteins for the noninvasive differentiation of benign and malignant breast tumors. International Journal of Nanomedicine. Informa UK Limited; **2015**. p. 537. Available from: http://dx.doi.org/10.2147/IJN.S71811.

[16] Qian K, Wang Y, Hua L, Chen A, Zhang Y. New method of lung cancer detection by saliva test using surface-enhanced Raman spectroscopy. Vol. 9, Thoracic Cancer. Wiley; **2018**. p. 1556–61. Available from: http://dx.doi.org/10.1111/1759-7714.12837.

[17] Liu W, Man Z, Hua L, Chen A, Wang Y, Qian K, et al. Data mining methods of lung cancer diagnosis by saliva tests using surface enhanced Raman spectroscopy. 2014 7th International Conference on Biomedical Engineering and Informatics. IEEE; **2014**. Available from: http://dx.doi.org/10.1109/BMEI.2014.7002849.

[18] Li X. Spectral analysis of human saliva for detection of lung cancer using surface-enhanced Raman spectroscopy. Vol. 17, Journal of Biomedical Optics. SPIE-Intl Soc Optical Eng; **2012**. p. 037003. Available from: http://dx.doi.org/10.1117/1.JBO.17.3.037003.

[19] Chen M, Chen Y, Wu S, Huang W, Lin J, Weng GX, et al. Raman spectroscopy of human saliva for acute myocardial infarction detection. Luo Q, Wang LV, Tuchin VV, editors. SPIE Proceedings. SPIE; **2014**. Available from: http://dx.doi.org/10.1117/12.2068876.

[20] Wang Yan, Hua Lin, Liu Jinghua, Qu Dian, Chen Anyu, Jiao Yi, et al. Preliminary study on the quick detection of acquired immure deficiency syndrome by saliva analysis using surface enhanced Raman spectroscopic technique. 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE; **2009**. Available from: http://dx.doi.org/10.1109/IEMBS.2009.5333131.

[21] Zermeño-Nava J de J, Martínez-Martínez MU, Rámirez-de-Ávila AL, Hernández-Arteaga AC, García-Valdivieso MaG, Hernández-Cedillo A, et al. Determination of sialic acid in saliva by means of surface-enhanced Raman spectroscopy as a marker in adnexal mass patients: ovarian cancer vs benign cases. Vol. 11, Journal of Ovarian Research. Springer Science and Business Media LLC; **2018**. Available from: http://dx.doi.org/10.1186/s13048-018-0433-9.

[22] Berrar, D. "Cross-Validation." **2019**, p.542-545.

[23] George, N., Singh, H., Jotaniya, R. and Pandya, S.R. “Raman spectroscopy for the determination of forensically important bio-fluids”. Vol. 340, Forensic Science International. Elsevier BV; **2022**. p. 111441. Available from: http://dx.doi.org/10.1016/j.forsciint.2022.111441.

[24] Dies H, Raveendran J, Escobedo C, Docoslis A. Rapid identification and quantification of illicit drugs on nanodendritic surface-enhanced Raman scattering substrates. Vol. 257, Sensors and Actuators B: Chemical. Elsevier BV; **2018**. p. 382–8. Available from: http://dx.doi.org/10.1016/j.snb.2017.10.181.

[25] Andreou C, Hoonejani MR, Barmi MR, Moskovits M, Meinhart CD. Rapid Detection of Drugs of Abuse in Saliva Using Surface Enhanced Raman Spectroscopy and Microfluidics. Vol. 7, ACS Nano. American Chemical Society (ACS); **2013**. p. 7157–64. Available from: http://dx.doi.org/10.1021/nn402563f.

[26] Eom G, Hwang A, Kim H, Yang S, Lee DK, Song S, et al. Diagnosis of Tamiflu-Resistant Influenza Virus in Human Nasal Fluid and Saliva Using Surface-Enhanced Raman Scattering. Vol. 4, ACS Sensors. American Chemical Society (ACS); **2019**. p. 2282–7. Available from: http://dx.doi.org/10.1021/acssensors.9b00697.

[27] Shende C, Brouillette C, Farquharson S. Detection of codeine and fentanyl in saliva, blood plasma and whole blood in 5-minutes using a SERS flow-separation strip. Vol. 144, The Analyst. Royal Society of Chemistry (RSC); **2019**. p. 5449–54. Available from: http://dx.doi.org/10.1039/c9an01087d.

[28] D’Elia V, Montalvo G, Ruiz CG, Ermolenkov VV, Ahmed Y, Lednev IK. Ultraviolet resonance Raman spectroscopy for the detection of cocaine in oral fluid. Vol. 188, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. Elsevier BV; **2018**. p. 338–40. Available from: http://dx.doi.org/10.1016/j.saa.2017.07.010.

[29] Farquharson S, Shende C, Sengupta A, Huang H, Inscore F. Rapid Detection and Identification of Overdose Drugs in Saliva by Surface-Enhanced Raman Scattering Using Fused Gold Colloids. Vol. 3, Pharmaceutics. MDPI AG; **2011**. p. 425–39. Available from: http://dx.doi.org/10.3390/pharmaceutics3030425.

[30] Sivashanmugan K, Squire K, Tan A, Zhao Y, Kraai JA, Rorrer GL, et al. Trace Detection of Tetrahydrocannabinol in Body Fluid via Surface-Enhanced Raman Scattering and Principal Component Analysis. Vol. 4, ACS Sensors. American Chemical Society (ACS); **2019**. p. 1109–17. Available from: http://dx.doi.org/10.1021/acssensors.9b00476.

[31] Yuen C, Zheng W, Huang Z. Improving surface-enhanced Raman scattering effect using gold-coated hierarchical polystyrene bead substrates modified with postgrowth microwave treatment. Vol. 13, Journal of Biomedical Optics. SPIE-Intl Soc Optical Eng; **2008**. p. 064040. Available from: http://dx.doi.org/10.1117/1.3050447.

[32] PDF to image converter package website: https://pypi.org/project/pdf2image/ visited 16:41 06/01/2023

[33] PDF vectorized file : https://www.adobe.com/creativecloud/design/discover/vector-file.html visited 16:39 06/01/2023

[34] Pytesseract package website : https://pypi.org/project/pytesseract/ visited 16:14 06/01/2023

[35] Tesseract-OCR package website : https://github.com/tesseract-ocr/tesseract visited 16:35 06/01/2023

[36] Template locator script: https://docs.opencv.org/4.x/d4/dc6/tutorial_py_template_matching.html visited 16:44 06/01/2023

[37] Code for Raman peak fetching through pdf articles: https://github.com/PauloAxcel/Raman-peaks-from-papers visited 18:29 06/01/2023

CHAPTER 7

Open-source software for Raman data plotting, Multivariate analysis and spectral Peak fitting

7.1 Abstract

Proper Raman data analysis involves a high level of expertise often missing in most professional software. Moreover, they do not provide the flexibility to execute out-of-the-box solutions for the Raman data analysis. To address this issue, in this chapter, several scripts were developed that can be accessed through a costumed graphical user interface (GUI) to solve multiple Raman data analysis problems thoroughly and flexibly. The scripts have been optimized to run smoothly with few inputs from the user, only requiring the Raman data and one parameter. The scripts focus on three main components:

- creating a customized peak library for any chemical compound
- solving multivariate data analysis and classification problems
- generating multiple data visualization plots

Firstly, the scripts offer individual, stacked, waterfall and heatmap plotting, with multiple preprocessing options like asymmetric least squares baseline, normalization or standard deviation. Secondly, they perform multivariate data analysis using principal component analysis (PCA), with k-nearest neighbour, Euclidean distance, receiving operating characteristic (ROC) curve and confusion matrix classification. Additionally, the scripts offer a neural network, self-organizing maps (SOM), to solve multivariate analysis outputting activation plots, peak activation heatmaps and classification results. Lastly, a multi-Lorentzian peak fitting script fits imported Raman data, PCA loadings and SOM activations, storing the peak fitting results in a comma separated value (CSV) file, facilitating a peak comparison between different samples and providing a complete Raman data analysis. With the executable scripts developed in this paper, we believe it would potentially solve some of the most common Raman data analysis problems found in the literature, leading to publishable results and plots when executed.

7.2 Introduction

C.V. Raman discovered the Raman scattering phenomenon in 1928 [1], identifying that when photons interacted with matter, a small percentage of photons were getting inelastically scattered in the matter. An inelastic scattering results in a difference in energy between incoming and scattered radiation caused by the interactions between molecular vibrations and the incident photons. When probing the scattering signal, information regarding the vibration modes of the molecules is obtained, which translates to specific information regarding the kinds of molecules and bonds present in the sample. Many research groups discovered the utility of using Raman spectroscopy in performing physics [2,3], chemistry [4,5], biology [6,7], material [8,9], medicine [10,11], and forensics [12,13] samples analyses from these findings. Hence, more recently, not only are Raman spectrometers present in many top research laboratory facilities but also in the clinical environment [14,15] to perform diagnostic measurements [16,17].

However, the high availability of Raman systems in all these facilities creates a problem for users, who are mainly experts in their research field, to deal with the high complexity of Raman data analysis that comes with the sample data processing. In both medical fields and research laboratories, non-data scientists perform most data analysis. It is usually performed from a ‘black box’ approach, ignoring the complex multi-parameter options for data analysis, thus losing better insight into the data. Moreover, each Raman user has a unique way of analysing the Raman data, and thus there is no systematic way of analysing the Raman data among different research groups. The existence of multiple approaches and inconsistencies between analysis procedures used by different research groups and even members within those research groups damages the reproducibility [18] and validation of results.

Due to these circumstances, many Raman spectrometer companies help scientists in their data analysis tasks by offering a software product that comes with the Raman spectrometer and solves most data analysis procedures. For instance, WiRE [19] software from Renishaw and LabSpec [20] from Horiba. These products can perform the Raman data acquisition and the respective data analysis. However, the data analysis tools offered are simplified enough not to overwhelm scientists with intricate multi-parameter algorithms. Despite this, the simplified algorithms in the software can perform most data processing algorithms and are good enough for initial data analysis. However, the data process is very one dimensional by ignoring more in-depth complex data analysis, which scientists should not avoid. To address this problem, the different ways Raman spectra data are analysed were investigated and sort them into the best combination for optimal results from the Raman data, thus, proposing our best method to analyse Raman data. A Python script[21] including these techniques has been created and can be accessible to everyone. Moreover, the script developed requires a maximum of one parameter as input besides the Raman spectral data, so it was designed to be as straightforward as possible while using the best data analysis configurations from the literature to be simple to understand by any scientist.

To program our alternative procedure, Python was used since it is the current top programming language [22]. It provides the necessary flexibility to perform complex data science tasks. Furthermore, Python is free; it has a wide online community with multiple open-source libraries which are easy to use and modify to our needs. Also, the Python language is user friendly with many intuitive features, so the entry level is lower than other programming languages. In our work, we have developed a complete Raman data analysis code by combining, modifying and optimizing existing Raman data analysis modules with our own to create a unique Python script. Throughout this work, examples

were provided of the different components of the scripts being applied to our case study Raman data while highlighting their respective functions and utilities.

The script comes with a GUI ideal for interacting with Raman data. It simplifies the user-machine interaction while still offering its full range of complex data science analysis options. A GUI, integrated with different algorithms, was developed to systematically analyse the Raman data that both Raman and non-Raman scientists can easily use. The GUI developed here assembles multiple Raman pre-processing tools that allow quick data analysis with simple clicks and complicated multi-class Raman analysis operations simplified with easy-to-use parameters. Among its affordances, it is possible to create Raman plots, pre-processing data, multivariate data analysis of multi-class Raman data, use a Neural Net option, and create dedicated Raman peak libraries for comparison or storage. The tools created to assemble this Raman GUI are inspired by how multiple authors analysed their Raman data.

The ultimate goal of this work is to achieve a complete in-depth understanding of Raman spectrum data and multi-class Raman spectra data. To demonstrate the goal of the scripts, a hypothetical case of two Raman spectra ζ and Ψ is used, where Ψ has some components of ζ in it. Let ζ be a known substance, but Ψ is only partially unknown substance, although it is related to ζ. Hence, Ψ can be written as Ψ, akin to ζ, or simply Ψ(ζ). Typical examples in the literature would be, extracting from the same tissue a healthy sample, ζ, and an unhealthy sample, Ψ. ζ can be defined as the control state of the tissue, and Ψ is the altered state. Moreover, since the tissue is the same, in terms of the Raman spectra, Ψ will have most ζ components, which makes it a tissue, and other components inherent to Ψ that make it an unhealthy tissue. The objective is to completely discern the differences and similarities between Ψ(ζ) from ζ by performing a throughout Raman spectra analysis. Hence, obtaining a complete report regarding the healthy,

unhealthy and what makes healthy tissue unhealthy from the Raman data features—in summary, having the ability to discern between healthy and unhealthy samples ζ, Ψ(ζ) and only Ψ by processing the Raman data with multiple tools, such as peak finder algorithms, multivariate analysis and neural networks. Subsequently, post-processing results from the multivariate analysis and neural networks combined with the peak finder algorithm to have complete insight into the different samples' fingerprints.

7.3 Methodology

7.3.1 Saliva Preparation

Saliva samples were collected from a single patient. The patient did not eat for 1 hour before the sample was collected by spitting into a container. Afterwards, 10μL of saliva was pipetted from the container into an aluminium foil substrate, left to dry in ambient conditions, and then taken into the Raman spectrometer.

7.3.2 Raman Data Acquisition

Using an InVia Qontor confocal Raman microscope (Renishaw), Raman measurements were carried out using a 785nm excitation laser. Raman maps were generated using the map mode scan with an exposure time of one second per map point and three accumulations using a laser power at the surface of 5mW. A 50× objective lens with a numerical aperture of 0.75 was used for the Raman measurements. Finally, the Raman spectrum was collected over a range of 600–1700cm^{-1}.

7.3.3 Raman Data Collection

WiRE software version 5.1 was used for our Raman spectral acquisition, which possesses very up-to-date functionalities that acquire Raman data from the instrument and perform data analysis. Here WiRE is used for data acquisition and cosmic ray removal. All other data processing and analysis is done through our scripts. The exported Raman data from the WiRE software is formatted as in **Figure 7.1A**. It is composed of a header of seven tabs, '#X',' empty tab', '#Y', 'empty tab',' #Wave', 'empty tab' and' #Intensity', and then the body of the text file with four columns, the X coordinate of the map, the Y coordinate of the map, the Raman shift values in cm^{-1} and the intensity values in arbitrary units (a.u.). The data structure has multiple repetitive values, making the files excessively large to store (**Figure 7.1A**). Depending on the Raman setting, a measurement can be extended (go from 100-3200cm^{-1} wavenumber) or static: user-defined midpoint – *k* – and then from that midpoint, a range is defined depending on the grating/laser settings, which for our setup was 600cm^{-1} < k < 1700cm^{-1}. The Raman shift range used is stored for every point in the map, so each intensity will have the same Raman shift associated with it, thus occupying unnecessary space. Besides these values, the constant *x* and *y* coordinate of the measured spot are also repeatedly stored. For instance, a 100 static point map will have 100*x*and*y*combinations of coordinates, on average, 1015 Raman shift values per point (depending on the filter and the system) and a respective 1015 intensity values (to match the Raman shift points) (**Figure 7.1A**). For a 100 point map scan with the Renishaw WiRE settings, the number of data points, besides the header, is 1015 data points per map times four variables (*x, y*, wavenumber and intensity) times 100 points in a map; for a total of 406,000 data points. Depending on the sample, it can be required to acquire more map points, so the number of data points grows linearly with the number of map points.

To reduce the space occupied by WiRE's exported data files, the data is reorganized in the following manner: the Raman shift values as the header and two additional headers for the map point's *x* and *y* coordinates. (**Figure 7.1B**). The intensity is written horizontally for a single *x* and *y* coordinates along with the header Raman shift values. This method occupies 1017 (1015+2) columns and 101 (100 intensity values per point + 1 Raman shift values) rows, amounting to 102,717 data points. Writing the data points horizontally instead of vertically saves 75% in terms of memory space compared with the standard data output format. Furthermore, this formatting facilitates access to the data points since each row is now a spectrum on the map.

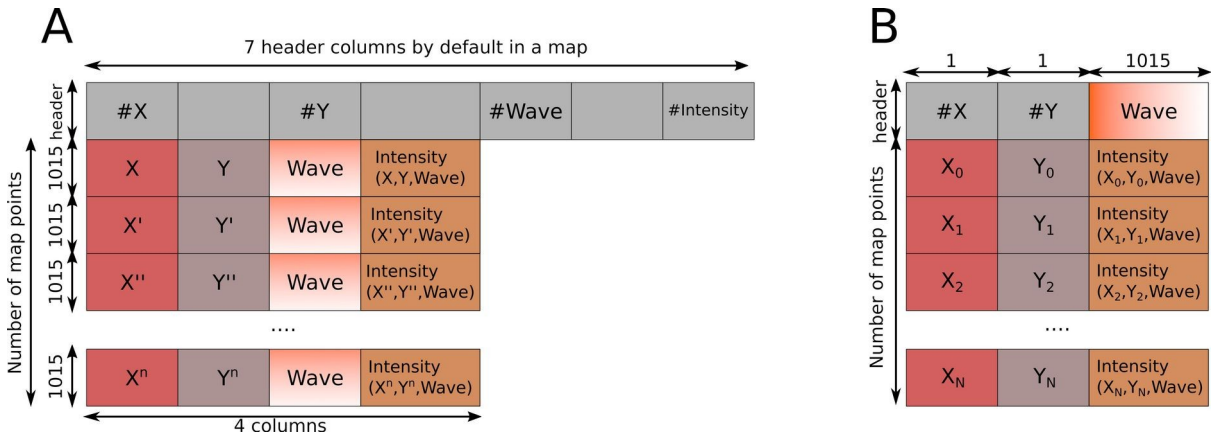


Figure 7.1*: Raman data structure from the Renishaw system. A**) Automated data formatting from WiRE 5.1 software. B**) Processed data for quicker access and less memory storage.*

For a single spot spectrum, the Raman data is expressed with a three-column header' #Wave', 'empty tab',' #Intensity' and then the data body as two columns with the Raman shift values in cm^{-1} and the intensity values in arbitrary units. Since standard Raman measurements are done in maps for statistical significance, the script was not optimized for individual Raman spectrum but only for Raman maps.

7.3.4 Python Code

All the scripts were written in Python version 3.7. If nothing is said otherwise, the conditions used in each different Python model/algorithm are the default options for each module. The code has been developed using Spyder[23] version 5.0.0 integrated development environment (IDE). The script is executed by selecting the operation to perform and then providing the raw Raman data in TXT format and an individual parameter option. The algorithm has been optimized to ask for minimal inputs from the user.

7.3.5 Graphical User Interface

The GUI was coded with Qtdesign using the PyQT5 library [24]. This library gives the user the freedom to make their own costumed GUI. With the help of Widgets, the user can click on the desired action on the GUI and receive as output the results. QpushButton, QcheckBox and Qlabel were the most commonly used widgets in the script. The different widgets are programmed to run specific functions assigned with an additional script written in Python. The functions programmed are highlighted below.

7.3.6 Data Analysis Methods

The various data analysis methods are discussed in this section, and examples demonstrating the effects of the various methods on a test Raman spectra are shown in the following section.

7.3.6.1 Asymmetric Least Square Smoothing (ALS)

The initial processing of the Raman data is to remove the background from the Raman spectra. The method used is an asymmetric least-square baseline (ALS) created by Paul Eilers [25,26] and written in Python from [27], which was further adapted for our script. The main package used in Python is the SciPy package [28], more specifically, the linalg, which analysis spectral signals with the "solveh_banded" function [29]. The main ALS function takes as inputs a Raman spectra and parameters related to the ALS configuration: the asymmetry, smoothness, iterations, and threshold, and outputs the baseline profile of the spectra. These options can be tuned from the 'Baseline' tab displayed on the right-hand side of the Graphical Interface (**Figure 7.2A**). By default, the values used are the standard ones: 0.05 'asymmetry', 1,000,000 'smoothness', 10 'iterations', and 0.000001 'threshold'.

7.3.6.2 Smoothing

Overall, smoothing is avoided when showing data and analysis data due to the potential loss of information, so it should be handled with care. Smoothing was used mainly for the peak fitting section of the code since we want to avoid the potential fitting of peaks to noise. Thus, a smoothing function was applied to the Raman spectra before running the peak finding to avoid this event. The smoothing function was adapted from the Savitzky-Golay filter [30].

7.3.6.3 Normalisation – Relative Intensity

A typical way to plot Raman data is to normalize different Raman signals to their respective maximum and minimum point. The relative spectra are ideal for comparing different Raman spectra by scaling them to the same range values. The normalization script converts the intensity into a relative intensity ranging between 0 and 1 (**Equation 7.1**).

$$I_{Rel} = \frac{I(x) - min(I(x))}{max(I(x)) - min(I(x))} \tag{7.1}$$

7.3.6.4 Standard Deviation

There is also the option of plotting the average spectra while highlighting the standard deviation to inspect the dispersion of the Raman data over a map.

$$std(I)_i = \sqrt{\frac{\sum_i^N (I_i - mean(I))}{N}} \tag{7.2}$$

7.3.6.5 Z-Score

Two ways to tackle outliers in the data were coded into the complete script. One was the Z-score [31] option. The Z-score is given by the difference of a variable (x_i) and its mean $\bar{x}$, and divided by its standard deviation σ as in (**Equation 7.3**):

$$z_i \equiv \frac{x_i - \overline{x}}{\sigma} \tag{7.3}$$

The Z-score is available from the SciPy.stats package. If the Z-score option is used, it will remove all data points with three standard deviations away from the mean spectra.

7.3.6.6 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a dimensionality reduction operation that projects the data into a lower-dimensional space. PCA is the main multivariable analysis tool to preprocess and analyse Raman data. PCA can be employed using the scikit-learn library [32], specifically sklearn.decomposition. PCA uses the normalised Raman spectra map for inputs and the number of PCA components to analyse. The number of components is related to how the Raman data will be transformed into the PCA representation. Hence the number of components is related to the dimensions of the new representation. We used only two components for the 2D plot and three for the 3D plot in the script. After the component selection, the input normalised Raman spectra are 'fit_transform' onto the PCA by applying the PCA.fit_transform() function, which transforms the normalised spectra into the PCA space with the principal components (PC) 1 and PC 2 or PC1, PC2 and PC3 for the 3D case. A scattered plot with the PC1 and PC2 components is plotted to inspect the different Raman spectra in the PC space. Information regarding the different PC variances can be obtained by the PCA.explained_variace_ratio_, which outlines the percentage variance contribution in spectra separation for the different PC. This measure highlights how well the PC separates the different input data.

7.3.6.7 Loadings

The principal components obtained from the PCA analysis can be subsequentially processed into the spectral loadings (**Equation 7.4**). The loadings indicate how the different PC separate the samples regarding their Raman spectra. The loadings mimic the Raman signal profile by computing the PC's eigenvalues and eigenvectors by applying the PCA to the Raman data. The loadings result from the coordinates with maximum variance and thus are very important to explain the different class separations obtained from the Raman spectra in the PC space.

The eigenvalues are obtained by the PCA.explained_variance_ representing the amount of variance by the different components. The eigenvectors are obtained from the PCA.components_, which are the principal axes in the PC feature space representing the direction of maximum variance in the data. By combining these two terms, we obtain the loadings of the PCA, which are relevant for the interpretation of the PCA plot.

$$Loadings = Eigenvectors \cdot \sqrt{Eigenvalues} \tag{7.4}$$

The loadings signal are thoroughly analysed by splitting them into positive and negative signals. For instance, when the Loadings from PC1 (LPC1) are positive, we are dealing with positive PC1. When they are negative, we are dealing with the negative part of PC1, which was generalized for all the different principal components. So in the 2D case, PC2 and LPC2 follow the same relation. Since in 2D we plot four different situations, i.e. PC1 > 0 and LPC1 > 0, PC1 < 0 and LPC1 < 0, combined with PC2 > 0 and LPC2 > 0, and PC2 < 0 and LPC2 < 0, which are related to the different quadrants in the PC space. Hence we have four quadrants (Q), Q1, Q2, Q3 and Q4, that respect the previous

inequalities Q1 (LPC1>0 and LPC2>0), Q2 (LPC1<0 and LPC2>0), Q3 (LPC1<0 and LPC2<0) and Q4 (LPC1>0 and LPC2<0) which can be written as **Equation 7.5**.

$$\begin{aligned} Q1 &= LPC1>0 + LPC2>0 \\ Q2 &= -LPC1>0 + LPC2>0 \\ Q3 &= -LPC1>0 - LPC2>0 \\ Q4 &= LPC1>0 - LPC2>0 \end{aligned} \tag{7.5}$$

Thus, depending on which PC quadrant a sample has populated, the loading 'spectra' can be identified based on the quadrant loading inequalities. For instance, if a sample is heavily populated in Q4, we would have to look into the LPC1>0 and LPC2<0 regions of the loading plots to see how the Q4 loading 'spectra' would look like, which is related to how the sample is differentiated compared to the others.

7.3.6.8 Gaussian Mixtures

Occasionally, there will be outliers in the Raman data, represented as points away from the sample cluster when processed into the PCA space. A Gaussian Mixtures [33] process is used to remove these outliers. This method fits an ellipsoid onto the PCA cluster location. It will exclude points outside the ellipsoid three standard deviations away from the cluster centre, similar to the Z-score option but now in the PC space. Thus, cleaning up the graphical representation of the Raman data PCA.

7.3.6.9 Euclidean Distance

Finalizing the PCA section, a statistical comparison between datasets is implemented by calculating the Euclidean distance (**Equation 7.6**) between different clusters in the PC space by employing.

$$\left(Euclidean\,distance\right)_{nD}=\sqrt{\sum_{n=1}^{2\vee 3}\left(PC_i-PC'_i\right)^2} \tag{7.6}$$

The Euclidean distance provides information on how different two clusters are from each other.

7.3.6.10 K-Nearest Neighbours

K-nearest neighbours (k-NN) classification from the sklearn.neighbors[34] library was used for classifying the PCA 2D data. The imported spectra data is split into training and validation data with the respective labels using the train_test_split from sklearn.model_selection [35]. The k-NN model fits the training data, and the k-NN score is calculated based on the cross-validation score [36] from sklearn.model_selection. The scoring output reflects the separability of the different data points. If the score is near 100%, the two classes are perfectly distinguishable, and if the score is approximately 50% (equal to random guessing, for the case of two classes), the score is poor. The k-NN classifier can predict classes using the ‘predict’ function with which we generate the confusion matrix and the receiver operating characteristic (ROC) curve from the prediction values.

7.3.6.11 ROC Curve

The Receiver Operating Characteristic (ROC) [37] is imported from the sklearn.metrics library using the roc_curve and auc package. The micro-average and macro-average ROC curve and respective ROC area (AUC) are computed and show the different sensitivity vs the false positive rate (FPR), FPR = FP/(FP+TN) per class used.

7.3.6.12 Confusion Matrix

The confusion matrix [38] is computed regarding the previously mentioned predicted classes using the k-NN classifier vs the actual classes imported from the model. Hence the confusion matrix evaluates the overall accuracy of the k-NN classifier model. From the confusion matrix, we get a complete predictive analysis of our model by computing the true positive (TP), the false negative (FN), the false positive (FP) and the true negative (TN). From there, we calculate the sensitivity (true positive rate) by the formula TPR = TP/(TP+FN), the specificity (true negative rate) TNR = TN/(TN+FP), the precision (positive predictive value) PPV = TP/(TP+FP), the negative predictive value NPV = TN/(TN+FN), and the overall accuracy = (TP+TN)/(TP+FP+FN+TN) of the model.

7.3.6.13 Self-Organizing Maps (SOM)

The SOM implementation is set up using the modules in miniSOM [39], which takes care of the whole SOM implementation and mathematics and is further improved in the script presented here. The Hexagonal Topology [40] work example on the MiniSOM GitHub page[39] served as a starting point for the script extension shown here. Where a hexagon

represents each spot on the grid map with inherent neural network weights, and biases randomly initialized at the start of the SOM calculation. Each hexagon's weight and bias represent a spectrum-like signature called the neuron activations or simply activations and is crucial to the SOM analysis. However, the implementation in [40] does not show the number of occurrences per hexagon; it only shows one occurrence, making it impossible to determine how many events or if different sample classes occur in the same hexagon. As a result, these features were added to the script shown here, and multiple occurrences of the various classes per hexagon can now be seen, as shown in **Figure 7.7C** of an example of a SOM map. However, another module, the Circlify[41, 42], was required to properly represent the multiple dots inside the hexagon. Where each different dot represents a different Raman spectrum on the SOM-space. The Circlify module handles rearranging multiple dots in a hexagon into an orderly geometry that minimises empty spots between dots (check **Figure 7.7C** dot arrangement). Aside from the dot arrangements, the hexagon edge colouration was changed to match the colour of the most populous class on that hexagon, and finally, the plot of neuron activations per hexagon, as shown in **Figure 7.7D**, which was obtained by checking the "win map"[43] variable and plotting it in relation to the Raman shift values. In summary, the modifications include a cleaner topographical arrangement discretising individual spectrum, an additional neuron activations plot with respect to the topographical arrangement of the spectra and addition of a topographical heatmap revealing the highest neuron activation bands. Moreover, SOM has been implemented here with only two input parameters, the Raman data to analyse and the size of the SOM grid map. The placement of the dots is determined by the spectrum-like weights and biases of the hexagon. Additional information can be observed regarding the highest population sample per hexagon by the colouring of the hexagon edge. Finally, different greyscale intensities inside each hexagon give the boundary

separations between adjacent hexagons, darker the hexagon, the stronger the separation, the lighter the hexagon weaker the separation.

7.3.6 Peak Finding Methods

7.3.6.1 Locating Local Maximum

The peak finding method starts by identifying the important peaks in a Raman spectrum and proceeds to perform multiple Lorentzian fits. In order to find the peaks in a Raman spectrum, a characterization of what is a peak is required. In mathematical terms, a peak happens when the difference between the Raman signal intensity in a spot *i* is smaller than in spot *i+1* and simultaneously, spot *i+1* is larger than spot *i+2* as described in **Equation 7.7**.

$$I_i < I_{i+1} \text{ and } I_{i+1} > I_{i+2} \tag{7.7}$$

In the case of noisy Raman data, noisy spectra can be identified as multiple peaks without being an actual peak. Hence, before checking for the local maxima, we apply a smoothing function to the Raman signal, previously described in the smoothing section.

7.3.6.2 Multi-Lorentzian Curve Fit

After identifying the local maxima, we obtain a list of different peaks that can be fitted by a Lorentzian curve (**Equation 7.8**)

$$L(x) = A\frac{\Gamma}{(x - x_0)^2 + (\Gamma)^2}, A, \Gamma, x_0 = constant \quad (7.8)$$

The work in [44] inspired the full peak finder script and then modified it to work for all identified peaks in a spectrum. The python package used to perform the Lorentzian fit is the SciPy.optimize.curve_fit module. We can identify the fitting parameters and calculate the complete Lorentzian curve fit, which fits the whole spectrum.

7.3.6.3 Residual Lorentzian Calculation

After obtaining the Lorentzian curve fit sum, we subtract it from the original Raman signal, and obtain the residual Lorentzian (**Equation 7.9**).

$$Residual = I(x) - \sum_{i=0}^{N} A_i \frac{\Gamma_i}{(x - x_i)^2 + (\Gamma_i)^2} N = list\ of\ peaks \quad (7.9)$$

For a perfect Lorentzian fit, the residual should be approximately zero. However, there are always discrepancies when comparing the Raman signal with the fitted Lorentzian sum, so we proceed by applying a second Lorentzian peak fitting to the residual.

There are two outcomes for the residual spectrum calculation. First, if the Lorentzian sum is higher than the Raman signal, the residual will be negative, and if the residual is positive, the Lorentzian sum is lower than the signal. Thus, the residual has a peak if the difference between the Raman signal and the Lorentzian sum is positive and a

trough if negative. When the residual has a peak, the initial Lorentzian fit did not cover enough of the peak in the original Raman signal. Thus we need to perform another Lorentzian fit to the residual peak.

On the other hand, the initial Lorentzian fit covers too much of the original Raman signal peak when it is a trough. Thus, it is required to subtract a Lorentzian fit to the trough of the residual to compensate for this. In summary, a de-convolution is performed of the initial Lorentzian into two positive Lorentzian if the residual is positive and a positive and a negative Lorentzian if the residual is negative, obtaining the final fitting equation (**Equation 7.10**). The prime notation for the residual peaks and the double prime for the residual trough. This double Lorentzian fit results in an improved second-order Lorentzian fit, resulting in a near-zero second-order residual.

$$2^{nd}\,order\,Residual = Residual + \sum_{i=0}^{N'} A_i' \frac{\Gamma_i'}{\left(x - x_i'\right)^2 + \left(\Gamma_i'\right)^2} - \sum_{i=0}^{N''} A_i'' \frac{\Gamma_i''}{\left(x - x_i''\right)^2 + \left(\Gamma_i''\right)^2} \qquad (7.10)$$

The identified peaks from this double Lorentzian fit procedure are stored in a CSV file and exported every time the code is run.

7.3.6.4 Standard ‘Key’ and ‘Lock’ Peak Finding Methods

The peak finder script will find all peaks present in the different imported files as previously described, with the multi-Lorentzian peak fitting. The peak finder goal is to automatically compare the peaks of different samples, yielding a comprehensive description of in common peaks and their relation and singular peaks among different samples. For this we introduced two nomenclatures created for this script, with the following description: a 'key', which implies a known sample that will function as a peak key (i.e. peak barcode) to be

used to check a 'lock' sample (i.e. unknown sample). Hence the other name is the 'lock' sample, which is unknown but has some information about the 'key' sample. When combining these nomenclatures with the general concept of peak identification described in the last part of the introduction, the 'key' spectra are the ζ spectra, and the 'lock' spectra are the Ψ(ζ). Furthermore, regarding the 'lock', i.e. Ψ(ζ) sample, the common peaks are the ζ peaks that come from the 'key', and the unknown peaks will come uniquely from Ψ ('lock'). So, we can have the complete analysis of ζ, Ψ and Ψ(ζ) in the end.

When performing the peak analysis, multiple input files can be provided to the script: the standard Raman spectra file, the PCA loadings, and the SOM activations. For the standard Raman data in TXT file format, we first load the ζ spectra into the 'key' section and run the code to generate the Lorentzian peak CSV file according to **Equation 7.11** Lorentzian fit equation.

$$\xi \simeq \sum_{i=0}^{N} A_i \frac{\Gamma_i}{\left(x - x_i\right)^2 + \left(\Gamma_i\right)^2} + \sum_{i=0}^{N'} A_i' \frac{\Gamma_i'}{\left(x - x_i'\right)^2 + \left(\Gamma_i'\right)^2} - \sum_{i=0}^{N''} A_i'' \frac{\Gamma_i''}{\left(x - x_i''\right)^2 + \left(\Gamma_i''\right)^2} \tag{7.11}$$

With the different xi, xi' and xi'' as the peaks representing the ζ spectra. Then the Ψ(ζ) spectra is loaded in the 'lock' section and the CSV file of the ζ spectra in the 'key' section and press run. Ψ(ζ) spectra has the form of **Equation 7.12**

$$\begin{aligned} \Psi(\xi) \simeq & \sum_{i=0}^{N} B_i \frac{H_i}{\left(y - y_i\right)^2 + \left(H_i\right)^2} + \sum_{i=0}^{N'} B_i' \frac{H_i'}{\left(y - y_i'\right)^2 + \left(H_i'\right)^2} - \sum_{i=0}^{N''} B_i'' \frac{H_i''}{\left(y - y_i''\right)^2 + \left(H_i''\right)^2} \\ & + \sum_{i=0}^{N} A_i \frac{\Gamma_i}{\left(x - x_i\right)^2 + \left(\Gamma_i\right)^2} + \sum_{i=0}^{N'} A_i' \frac{\Gamma_i'}{\left(x - x_i'\right)^2 + \left(\Gamma_i'\right)^2} - \sum_{i=0}^{N''} A_i'' \frac{\Gamma_i''}{\left(x - x_i''\right)^2 + \left(\Gamma_i''\right)^2} \end{aligned} \tag{7.12}$$

With the different xi, xi' and xi'' as the peaks representing the ζ spectra in Ψ(ζ) and the different yi,yi' and yi'' as the peaks representing the overall Ψ(ζ) spectra. This combination of files will output the 'lock' spectra with the ζ highlighted peaks that match the 'lock' file.

Multiple combinations can be made with this script. Another possibility is to get both the peak CSV list from both ζ and Ψ(ζ) spectra followed by loading both ζ and Ψ(ζ) CSV files into the 'key' section and the 'lock' only the Ψ(ζ) spectra. This combination will match the Ψ(ζ) Raman spectra with both ζ and Ψ(ζ) peaks. Then we can combine the peaks from the two samples highlighted and sort out the different contributions. If the Ψ(ζ) peaks only highlight the Ψ(ζ) spectra, then it is a specific peak to the Ψ(ζ) spectra. If the Ψ(ζ) spectra are highlighted by both the ζ and the Ψ(ζ) peaks, it is an ζ peak. Potentially it can be both contributions, and this information can be discerned by consulting the intensity values in the output CSV file. Hence, the overall peak matching method serves as a good indicator of the contribution of different peaks and to which spectra it belongs. Furthermore, there are multiple other combinations of 'key' and 'lock' files that can be done, resulting in a deeper understanding of each sample.

7.3.6.5 Principal Components Analysis Loadings Peak Finding Method

Since the PCA multivariable analysis method can distinguish differences between different classes, if we compare ζ and Ψ(ζ), PCA will separate the samples based on the differences between the two samples or mix based on the similarities. So we can distinguish the ζ and Ψ contribution from the Ψ(ζ). This information is identified by analysing the PCA loadings from the PCA analysis. The loadings from the PC1 and PC2 are Lorentzian fitted as the spectral peaks to write the Loadings as in **Equation 7.11**. Thus

each quadrant will also have its Lorentzian equation combining **Equation 7.5** with **Equation 7.11**.

A practical case would be to generate a PCA/loading peak finder from ζ and Ψ(ζ) spectra. These would output a CSV file of the loadings peaks and their correspondent quadrants. Then we would load this CSV file into the lock and load into the key the generated ζ and Ψ(ζ) CSV files from the standard peak finding method and perform the peak matching. The output plot will highlight the location of the ζ and Ψ(ζ) peaks in the different quadrant regions. If the Ψ(ζ) peaks only highlight the quadrant spectra, it is a specific peak to the Ψ spectra. If the quadrant spectra are highlighted by the ζ and the Ψ(ζ) peaks, it is an ζ peak. Hence, it serves as a good indicator of the contribution of different peaks and to which quadrant they belong.

To better understand the data, both the PCA/loadings peak fitting analysis and the standard peak fitting analysis need to be analysed together to achieve better conclusions about your data.

7.3.6.6 Self Organizing Maps Activation Peak Finding Method

Finally, the same procedure for the peak finding in the PCA loadings is applied to the SOM activations. Information obtained from the SOM activations is related to the difference between different classes, i.e. knowing ζ and Ψ(ζ) and distinguishing the different spectral features through the SOMs' activations by highlighting all ζ, Ψ(ζ) and Ψ features.

Each neuron and its main peaks can be inspected by applying a multi-Lorentzian fit to each activation. So that a Lorentzian fit equation would represent each neuron as in **Equation 7.11**. Besides this information, we can track the peaks identified per neuron as

an activation heatmap, getting a clear idea of how the different peaks change per neuron. As a result, all the different peaks in each neuron spot are quantified and written over a CSV file.

A practical case would be to generate a SOM activation peak finder from ζ and Ψ(ζ) spectra. These would output a CSV file of the activation peaks and their correspondent hexagon i and j index positions. Then we would load this CSV file into the lock and load into the key the generated ζ and Ψ(ζ) CSV files from the standard peak finding method and perform the peak matching. The output plot will highlight the location of the ζ and Ψ(ζ) peaks in the different hexagon index i, j index positions. If the Ψ(ζ) peaks only highlight the hexagon spectra, it is a specific peak to the Ψ spectra. If the hexagon spectra are highlighted by the ζ and the Ψ(ζ) peaks, it is an ζ peak. Hence, it serves as a good indicator of the contribution of different peaks and which hexagon they belong to.

To better understand the data, all SOM/activations peak fitting analysis, PCA/loadings peak fitting analysis, and the standard peak fitting analysis needs to be analysed together to achieve better conclusions about your data.

7.3.7 Scripts

The entirety of the scripts for this work can be found at https://github.com/PauloAxcel/Raman-Peak-Finder.git.

7.4 Results and Discussion

This section compiles the outputs from the different scripts developed in this work and the GUI instructions on how to execute them.

Overall, the section is divided into three parts, an initial plotting script that takes care of the different plotting options for the data, a second part dedicated to the PCA analysis and classification performance and a third part dedicated to the peak analysis.

7.4.1 Plots

The created GUI (**Figure 7.2A**) has a main screen dedicated to plotting and three tabs assigned to different data processing functions. It also has a 'Browse' button that opens a window for the user to select a Raman data file (must be in TXT format). The Raman files are directly produced from the Raman software (in our case, the Renishaw and its software WiRE v5.1), and the imported data do not require any pre-processing. All pre-processing is done internally by the script. After loading the Raman data files, the user can plot the data imported with different data manipulation like baseline correction, standard deviation and normalisation. As an illustration of the script's capabilities, different saliva samples were arranged to showcase the different plot options. For the baseline subtraction, the script takes as input **Figure 7.2Bi** and outputs **Figure 7.2Bii**. The normalisation option converts the intensity of **Figure 7.2Bii** into **Figure 7.2Biii**. Finally, the standard deviation option highlights the standard deviation to inspect the dispersion of the acquired Raman data changing the plot from **Figure 7.2Biii** to **Figure 7.2Biv**. These options are toggled commands from the main screen, and more than one option can be selected simultaneously. Only the baseline can be further fine-tuned from the available

command operations. Four parameters can be tuned for the asymmetric least squares smoothing baseline for the Raman plots, located on the ' Baseline' tab section.

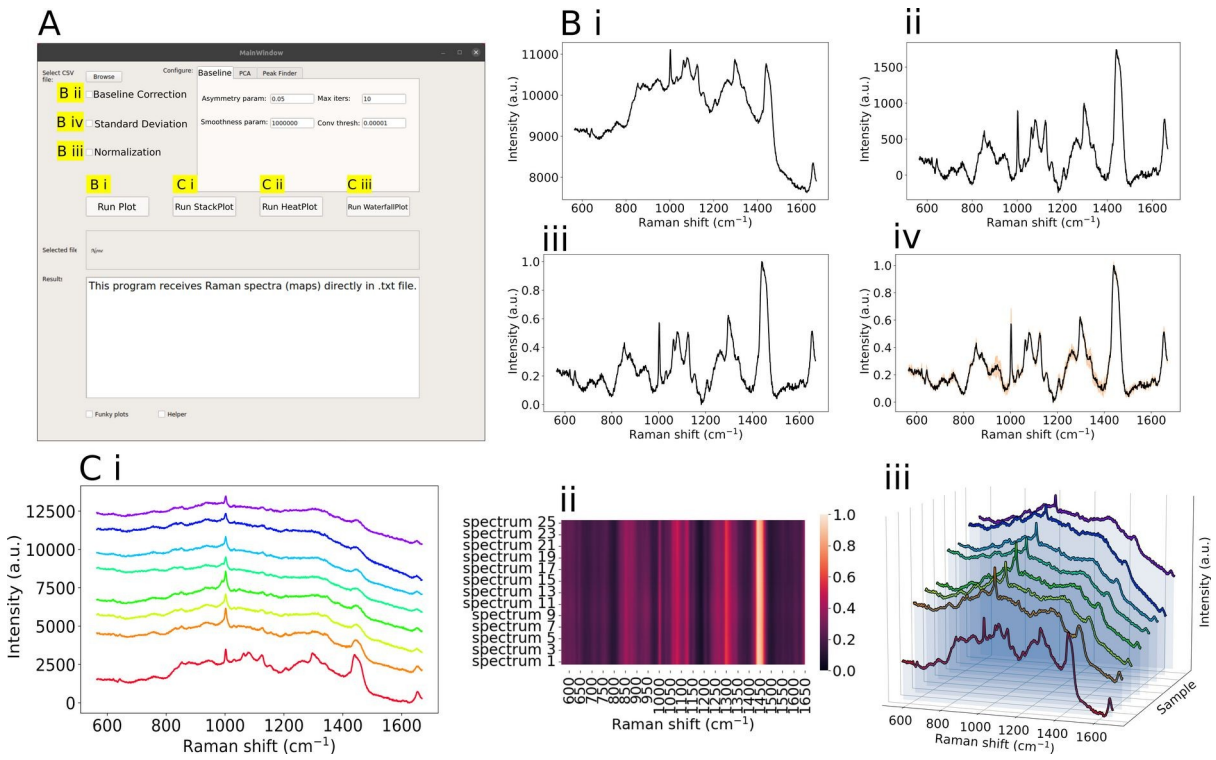


***Figure 7.2**: First part of the graphical interface – Plotting. (**A**) The GUI shows multiple commands options that can be toggled and pressed to make different kinds of plots. Additionally, multiple display windows outline the files imported, a text area to display results and additional function tabs. When the options are toggled, the process is performed over the Raman data before plotting. For the data processes, we have the standard plot (**Bi**), toggled options for baseline correction (**Bii**), normalization (**Biii**) and standard deviation (**Biv**). Besides the standard plot, there are three other different plot options, stack plot (**Ci**), heatmap plot (**Cii**) and waterfall plot (**Ciii**).*

Furthermore, multiple plot options are available to visualise the data. There are average map or individual plots (**Figure 7.2Bi**), stack plots (**Figure 7.2Ci**), heatmap plots (**Figure 7.2Cii**) and 3D plots (**Figure 7.2Ciii**) which can be pressed. The data is then processed according to the toggled operations and subsequentially plotted after the “run plot” button is pressed.

7.4.2 Principal Component Analysis

After inspecting the imported data with the different plot options, we can continue the data analysis procedure. In the second tab section, an option for multivariable analysis was coded to allow the user to analyse the imported Raman spectra using PCA (**Figure 7.3A**). The data loaded and processed with the PCA option are by default baseline subtracted (ALS method with default parameters) and normalisation (**Equation 7.1**) by the script.

The PCA script output was designed to plot the multiple outcomes from a typical PCA data analysis, such as 2D and 3D (**Figure 7.3B**) plot options. Besides the plotting option, multiple preprocessing options are available to manipulate the data before the PCA operation is performed. We have the Z-score (**Figure 7.3D**), a window setting (**Figure 7.4C** and **D**) and a Gaussian mixtures method (**Figure 7.3C**). The Z-score and window setting operations change the initial Raman data before the PCA processing, thus producing different results than the initial PCA analysis. The Z-score command will remove outlier spectra from the whole spectra collection, thus clustering the spectra more tightly together. For the window setting, by changing the Raman shift values, the Raman intensity values will be truncated to a specific region making the PCA accommodate this (**Figure 7.3D**). The window option can be run when tracking a specific spectra region resulting in a different and clearer PCA separation than the standard settings. The last parameter, the Gaussian Mixtures operation, is done after the PCA is performed. It functions similar to the Z-score. It fits an ellipsoid that accommodates a certain standard deviation range over the sample cluster in the PC space, excluding all spectra points outside that range (**Figure 7.3C**). Different combinations of parameters can improve the PCA results from the PCA analysis. The important observation is the different clustering of the classes and the subsequent loading analysis (**Figure 7.4C** and **D**).

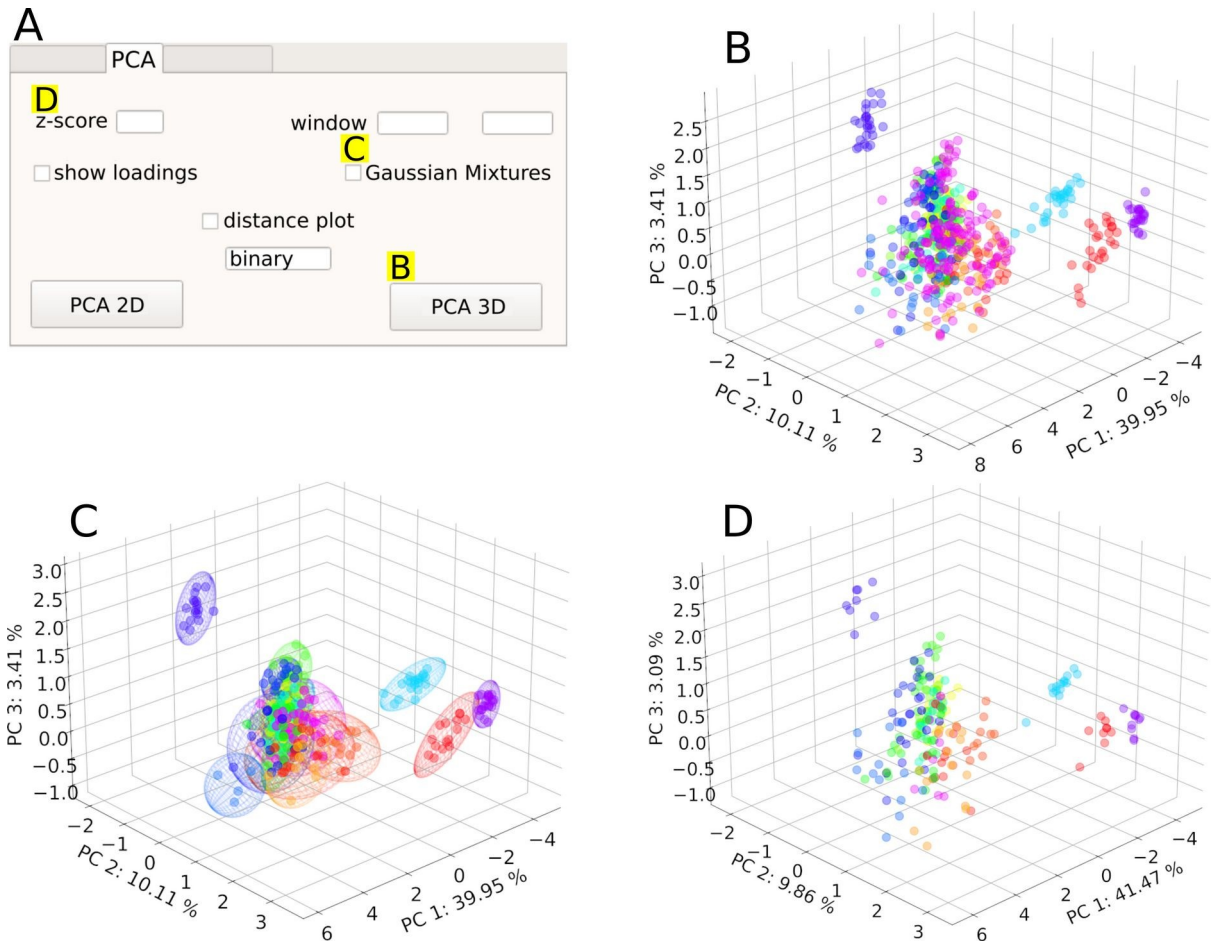


***Figure 7.3**: Second tab of the Graphical Interface – PCA. **A)** PCA section, multiple commands can be computed in this panel by toggling different options and pressing the PCA 2D or 3D to run. **B)** Showcase a 3D PCA plot for 15 different datasets. **C)** 3D PCA with the same data points with Gaussian Mixtures applied. **D)** 3D PCA with the same data points with a Z-score applied to the data.*

Finalising the PCA section, a statistical comparison between the principal components (PC) is implemented by calculating the Euclidean distance (**Equation 7.4**) (**Figure 7.4Bi**) between different clusters in the PC space (**Figure 7.4Biii red** vs **Figure 7.4Biii green**). Since PCA is used to compare multiple samples, the most common cases in the bioscience field are healthy vs unhealthy samples. Some authors also add an intermediate state, for instance, healthy vs cancer vs benign cancer. So, we have a binary comparison or a trinary comparison in the end. For these cases, an additional script does a classification test over the samples with a k-nearest neighbours algorithm (**Figure 7.4Biii**), highlighting the separation boundary between the two classes and outputting the

accuracy of the two classes' separation. Additionally, a ROC curve (**Figure 7.4Biv**) and a confusion matrix (**Figure 7.4Bii**) are plotted to give more information regarding the class analytical parameters like sensitivity and specificity.

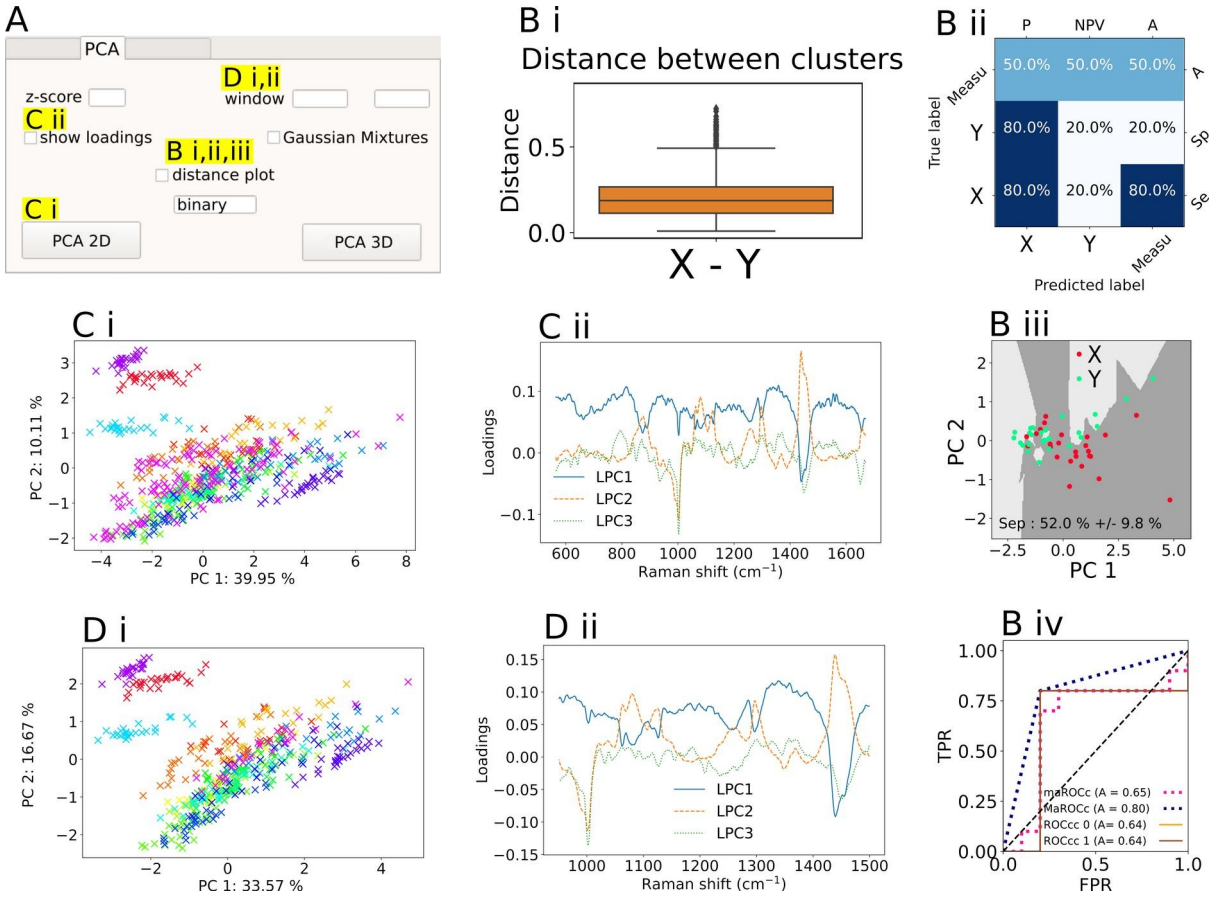


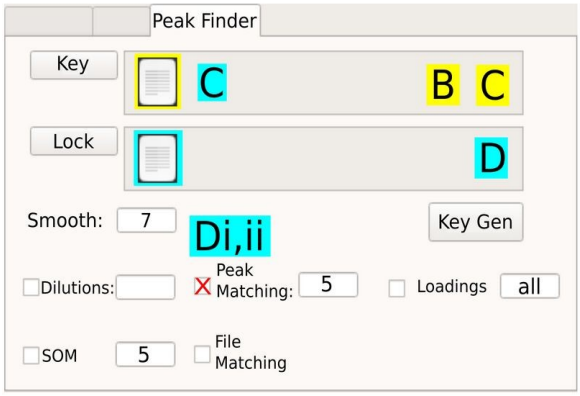
Figure 7.4*:*A) *Second part of the PCA section, exploring the window range, PCA loadings and distance plot in 2D.* ***B)i)*** *Euclidean distance whisker bar plot in the binary mode (healthy/unhealthy) obtaining for the distance between cluster ζ and cluster Ψ from PCA plot in* ***Biii)****.* ***ii)*** *Confusion matrix for the classification test of ζ and Ψ samples, with accuracy (A), Specificity (Sp), Selectivity (Se), negative predictive value (NPV) and precision (P).* ***iii)*** *PCA plot with k-NN separation and separation percentage (52%).* ***iv)*** *ROC curve for the false positive rate (FPR) and true positive rate (TPR) relation.* ***C)i)*** *2D PCA plot (same data as* ***Figure 3****), showing the respective loading components* ***C)ii)****.* ***D)i)*** *2D PCA and* ***ii)*** *loadings with windows setting on with a range from 950 cm^{-1} to 1500 cm^{-1}, highlighting a slightly different separation and loadings than* ***D)i)****.*

7.4.3 Peak Finder

For the last tab of the GUI, we present our Raman multi-peak fitting script. The script is designed to perform multi-Lorentzian fits to individual Raman peaks in a spectrum. Two

different procedures can be executed when importing the Raman spectra map (**Figure 7.5Aii**) onto the GUI (**Figure 7.5Ai**). The first is highlighted in yellow. The Raman TXT file is imported onto the key menu followed by pressing the keygen button resulting in the creation of the multiplot in **Figure 7.5B** and a CSV table with all the peaks found as in **Figure 7.5C**. **Figure 7.5B** outlines the multi-Lorentzian fit procedure. First, the peaks are identified by locating local maximum points in the Raman spectra (**Figure 7.5Bi** red crosses) and then fitted by the same number of Lorentzian curves (**Figure 7.5Bi** coloured Lorentzian fits). All Lorentzian fits are added and subtracted over the Raman spectra, resulting in the blue scatter points (**Figure 7.5Bii**). The outcome residual peaks are again fitted with multiple Lorentzian curves (**Figure 7.5Bii** coloured Lorentzian fits), obtaining the second-order residual. The second-residual plotted should approximate zero in case of a good Lorentzian fitting (**Figure 7.5Biii** red line). Finally, overlapping the multi-Loretzian fit sum (**Figure 7.Biv** blue) with the initial Raman spectra (**Figure 7.Biv** orange). The CSV file generated (**Figure 7.5C**) with all the fitted peaks has additional information related to the peak intensity, intensity error, width, and Raman shift. After generating this CSV file, it can be used as a 'key' option to open 'lock' spectra which leads to the second part of the peak finder code. Following the cyan labels in **Figure 7.5Ai**, the generated CSV file is inserted on the 'key' option, and an unknown Raman spectrum is inserted in the 'lock' option. Afterwards, pressing the keygen button with the peak matching box toggled (**Figure 7.5Ai** red cross) generates **Figure 7.5D** plots. The peak matching box associated value is related to the acceptance window between peaks, i.e., peaks found between the two loaded files match if their Raman shift difference is inferior to the value inserted in the text box.

A i



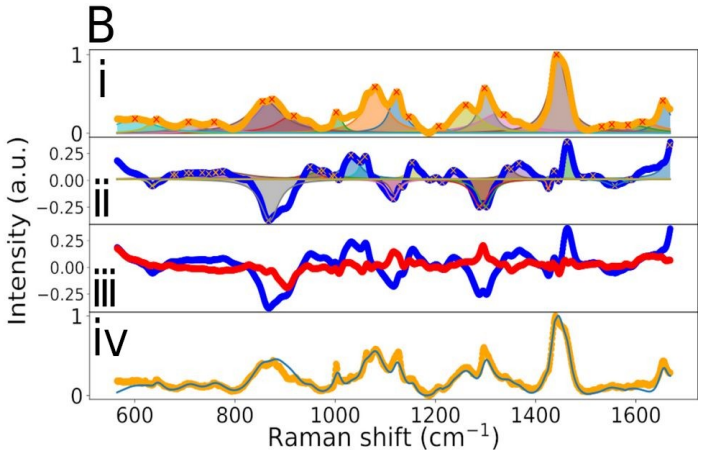
ii

#X	#Y	#Wave	#Intensity
29750.509761	5769.346500	1668.609375	6961.682129
29750.509761	5769.346500	1667.625000	7057.558105
29750.509761	5769.346500	1666.641602	7016.005859
29750.509761	5769.346500	1665.657227	7123.545410
29750.509761	5769.346500	1664.672852	7199.100098
29750.509761	5769.346500	1663.687500	7091.025391
29750.509761	5769.346500	1662.703125	7129.311035
29750.509761	5769.346500	1661.717773	7127.686523
29750.509761	5769.346500	1660.732422	7253.693848
29750.509761	5769.346500	1659.747070	7353.060547
29750.509761	5769.346500	1658.760742	7311.520996
29750.509761	5769.346500	1657.775391	7429.431152
29750.509761	5769.346500	1656.789063	7430.398926

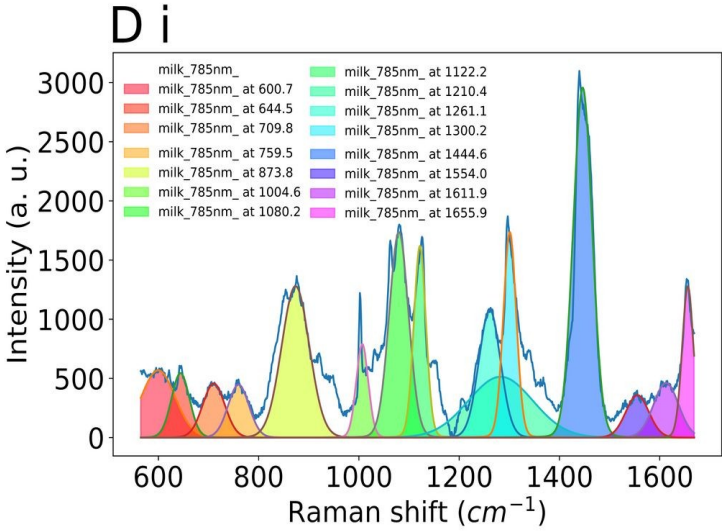
C

label	height	center	width	Intensity (a.u.)	err	importance
milk_785nm_	0.2	600.7	44.9	570.8	0.8	major
milk_785nm_	0.2	644.5	21.8	541.5	3.8	major
milk_785nm_	0.0	681.4	20.0	106.2	2.2	minor positive
milk_785nm_	0.0	707.4	31.3	141.5	0.7	minor positive
milk_785nm_	0.1	709.8	28.4	445.0	1.0	major
milk_785nm_	0.0	733.4	34.1	148.5	0.6	minor positive
milk_785nm_	0.1	759.5	24.8	438.9	1.2	major
milk_785nm_	0.1	775.0	22.9	173.6	1.2	minor positive
milk_785nm_	0.1	780.4	52.0	169.0	7.6	minor positive
milk_785nm_	0.4	860.7	41.1	1268.3	5.5	major
milk_785nm_	0.4	873.8	35.3	1337.6	4.1	major
milk_785nm_	0.2	913.6	34.6	686.4	4.1	major
milk_785nm_	0.1	952.7	13.6	286.4	4.8	minor positive
milk_785nm_	0.1	972.3	28.4	192.5	2.4	minor positive
milk_785nm_	0.0	988.3	11.4	103.2	21.3	minor positive
milk_785nm_	0.3	1004.6	10.8	830.4	15.9	major
milk_785nm_	0.2	1033.1	19.3	545.0	2.3	minor positive
milk_785nm_	0.2	1058.2	10.8	520.6	20.6	minor positive
milk_785nm_	0.6	1080.2	26.0	1812.8	7.4	major
milk_785nm_	0.5	1122.2	16.1	1632.0	15.1	major
milk_785nm_	0.1	1155.8	9.8	395.2	10.8	minor positive
milk_785nm_	0.0	1201.3	2.2	39.3	16.2	minor positive
milk_785nm_	0.1	1210.4	14.4	282.3	6.2	major
milk_785nm_	0.1	1237.5	12.0	200.3	3.0	minor positive
milk_785nm_	0.4	1261.1	29.2	1111.4	0.9	major
milk_785nm_	0.6	1300.2	15.9	1782.9	13.0	major
milk_785nm_	0.2	1325.4	46.9	764.7	14.3	major
milk_785nm_	0.1	1347.0	13.6	261.0	7.5	minor positive
milk_785nm_	0.1	1368.0	18.4	357.1	0.9	minor positive
milk_785nm_	1.0	1444.6	20.1	3136.5	28.8	major
milk_785nm_	0.3	1463.9	10.0	876.1	14.3	minor positive
milk_785nm_	0.0	1515.4	15.7	74.9	0.2	minor positive
milk_785nm_	0.1	1540.9	38.0	269.0	3.0	major
milk_785nm_	0.1	1554.0	23.2	348.1	1.1	major
milk_785nm_	0.1	1592.2	42.6	331.2	1.8	major
milk_785nm_	0.1	1611.9	42.3	449.2	0.6	major
milk_785nm_	0.4	1655.9	14.5	1294.0	5.9	major

B



D i



D ii

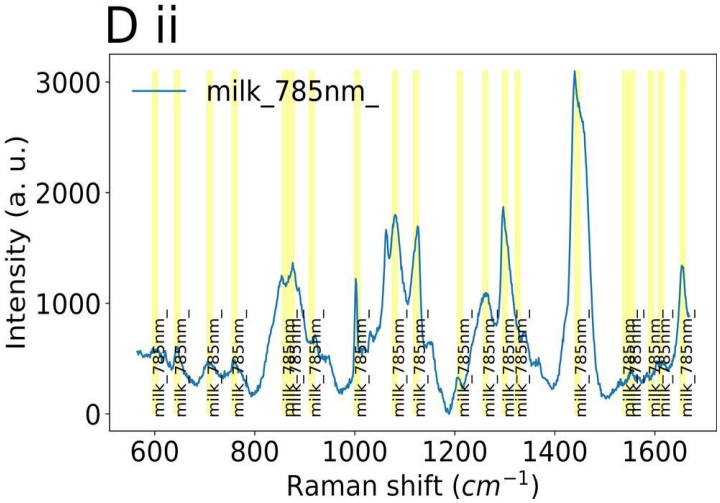


Figure 7.5*:*A)i) *Peak Finder GUI tab highlighting the standard Raman peak analysis. Load a Raman TXT file* ***(A ii)*** *into 'key' and generate the Raman fit* ***(B)*** *with the peak information stored in a CSV file* ***(C)****. Load the generated Raman* ***(C)*** *list into 'key' and load a Raman text file into 'lock' with the option 'Peak matching' ticked to generate the peak matching between the 'key' peaks and the 'lock' file, output results in* ***(Di, ii)****.*

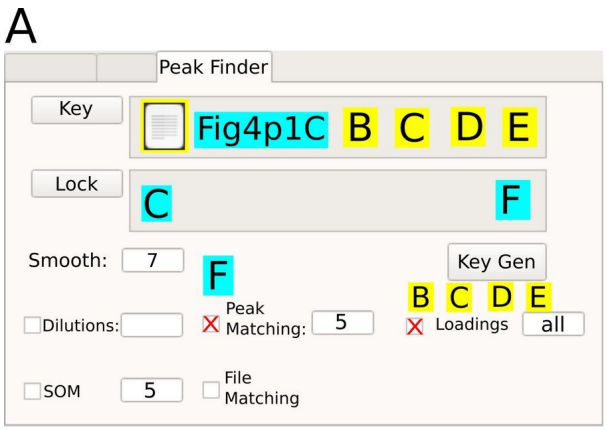
Usually, a number approximately five or lower is accepted. From the peak matching option, the generated plots (**Figure 7.5D**) match the 'key' peaks onto the 'lock' file and highlight them and fit them if there is a match (**Figure 7.5Dii, i**, respectively).

7.4.4 Principal Component Analysis Peak Finder

Besides the standard peak finder method described previously, the script can analyse the peaks from the two multivariate analysis methods of PCA and SOM. PCA peak finding method works over the loading values of the two initial principal components (PC1 and PC2). To run these script functionalities, there are two options highlighted in **Figure 7.6A**. Following the yellow highlighting, we load the different spectra into the key section, toggle the 'loadings' option and press the keygen button. The script runs and outputs the PCA plot with the two initial components, PC1 and PC2 (**Figure 7.6B**). Together with the respective loadings plot (**Figure 7.6D**). The PCA plot shows how the sample distributes among the PC space with additional highlighted regions marking the four different quadrants (blue Q1, yellow Q2, green Q3 and red Q4). The loadings plot also highlights regions that are related to the different PCA quadrants, separating the positive and negative values in dark grey and light grey. In the 2D PCA space, there are four different quadrants (Q), Q1 blue with the PC1 and PC2 values > 0, Q2 PC1<0 and PC2>0, Q3 PC1<0 and PC2<0 and Q4 PC1>0 and PC2<0. These inequalities are related to the contributions of the Loadings spectra. For instance, in Q4, which is heavily populated by the red samples in (**Figure 7.6B**), the Loadings give information that for this quadrant, we have to look for LPC1>0 and LPC2<0 in **Figure 7.6D**. Generalizing, the two PC loadings can be deconstructed into the four different inequality (**Equation 7.5**), and multi-Lorentzian peak fit them as outlined in **Figure 7.6E**. Thus, the peak analysis over the loadings is split

between the four different quadrants of the PCA space, and each quadrant is individually multi-Lorentzian fitted, and the values are recorded in CSV format (**Figure 7.6C**) with information regarding the peak fit height, centre, width, and the respective quadrant to which the loading inequality belonged.

The second option to run the PCA peak finder script is highlighted in cyan and requires the loadings multi-Lorentzian fit CSV file (**Figure 7.6C**) inserted in the 'lock' file option. Together with the standard Raman spectra multi-Lorentzian peak fit CSV file (**Figure 7.5C**) inserted in the 'key' section. To run this routine, toggle both the 'peak matching' option and the 'loadings' option and push the keygen button. This combination of options outputs the sum of the multi-Lorentzian fits (**Figure 7.6F**) with the respective quadrant location, highlighted by the colour matching between the loadings 'spectra' and the PCA's quadrant colour (**Figure 7.6B**). The resulting plot matches the peaks between the two CSV files loaded, i.e. the loadings 'spectra' and the peaks from the standard Raman spectra multi-Lorentzian fit. Each matching depends on which loadings quadrant the match occurs; thus, each highlighted yellow bar in (**Figure 7.6F**) only matches a specific peak per quadrant, which results in a better understanding of the different samples in terms of their multivariate analysis and peak distributions.



C

label	height	center	width	importance
saliva_785n	0.121	613.6	37.4	Q1
saliva_785n	0.121	613.6	37.4	Q4
saliva_785n	0.084	690.7	38.0	Q1
saliva_785n	0.084	690.7	38.0	Q4
saliva_785n	0.116	755.4	41.5	Q1
saliva_785n	0.116	755.4	41.5	Q4
saliva_785n	0.008	813.5	21.9	Q1
saliva_785n	0.008	813.5	21.9	Q2
saliva_785n	0.120	820.0	31.9	Q1
saliva_785n	0.120	820.0	31.9	Q4
saliva_785n	0.053	873.6	19.3	Q1
saliva_785n	0.053	873.6	19.3	Q2
saliva_785n	0.146	937.5	28.1	Q1
saliva_785n	0.146	937.5	28.1	Q4
saliva_785n	0.178	997.6	21.6	Q1
saliva_785n	0.178	997.6	21.6	Q4
saliva_785n	0.038	998.6	10.8	Q3
saliva_785n	0.038	998.6	10.8	Q4
saliva_785n	0.075	1080.2	26.3	Q1
saliva_785n	0.075	1080.2	26.3	Q2
saliva_785n	0.059	1081.0	12.1	Q2
saliva_785n	0.059	1081.0	12.1	Q3
saliva_785n	0.055	1119.5	20.8	Q1
saliva_785n	0.055	1119.5	20.8	Q2
saliva_785n	0.023	1119.9	4.0	Q2
saliva_785n	0.023	1119.9	4.0	Q3
saliva_785n	0.097	1180.6	39.1	Q1
saliva_785n	0.097	1180.6	39.1	Q4
saliva_785n	0.003	1186.6	5.4	Q3
saliva_785n	0.003	1186.6	5.4	Q4
saliva_785n	0.005	1202.2	7.7	Q3
saliva_785n	0.005	1202.2	7.7	Q4
saliva_785n	0.106	1210.2	41.8	Q1
saliva_785n	0.106	1210.2	41.8	Q4
saliva_785n	0.005	1231.9	7.8	Q1
saliva_785n	0.005	1231.9	7.8	Q2
saliva_785n	0.026	1256.6	15.3	Q1
saliva_785n	0.026	1256.6	15.3	Q2
saliva_785n	0.086	1280.1	14.2	Q1
saliva_785n	0.086	1280.1	14.2	Q4
saliva_785n	0.037	1299.4	11.9	Q1
saliva_785n	0.037	1299.4	11.9	Q2
saliva_785n	0.118	1336.8	41.3	Q1
saliva_785n	0.118	1336.8	41.3	Q4

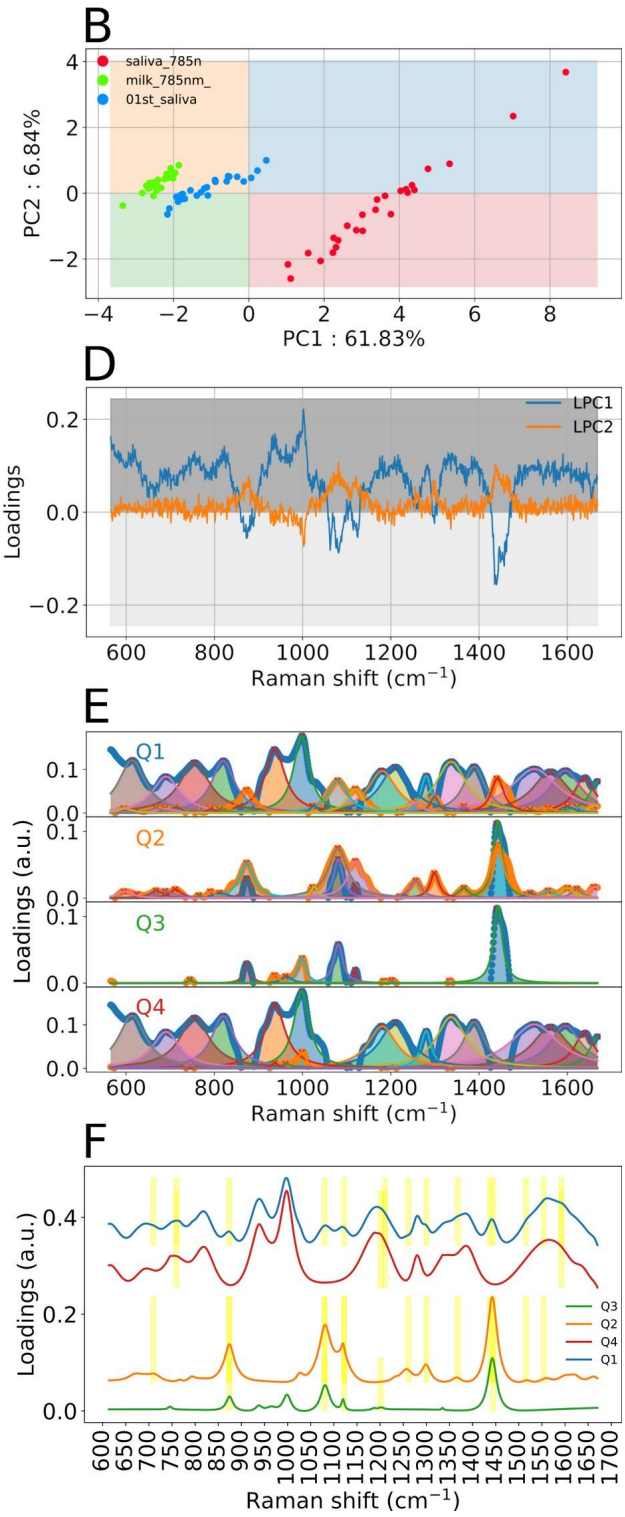


Figure 7.6*:*A)i) *GUI – PCA Peak Finder section. Highlighted in yellow and cyan are the two different operation methods of the script.* ***Yellow*** *generated the multi-Lorentzian fit CSV file of the loadings, and* ***cyan*** *compared the peaks from the loadings CSV file with the peak from the standard peak CSV file.* ***B)*** *PCA of three different samples, highlighted by four different regions in the PC1 and PC2 space (Q: quadrants, Q1 blue, Q2 orange, Q3 green and Q4 red).* ***C)*** *CSV output file of executing the highlighted yellow routine in* ***A****. Showing the multi-Lorentzian fitting parameters.* ***D)*** *Loading plot from PC1 and PC2 separated by two regions positive loadings (****dark grey****) and negative loadings (****light grey****).* ***E)*** *Multi-Lorentzian fit of the loadings for each quadrant.* ***F)*** *Peak matching between the multi-Lorentzian fit summed and the imported 'key' CSV file.*

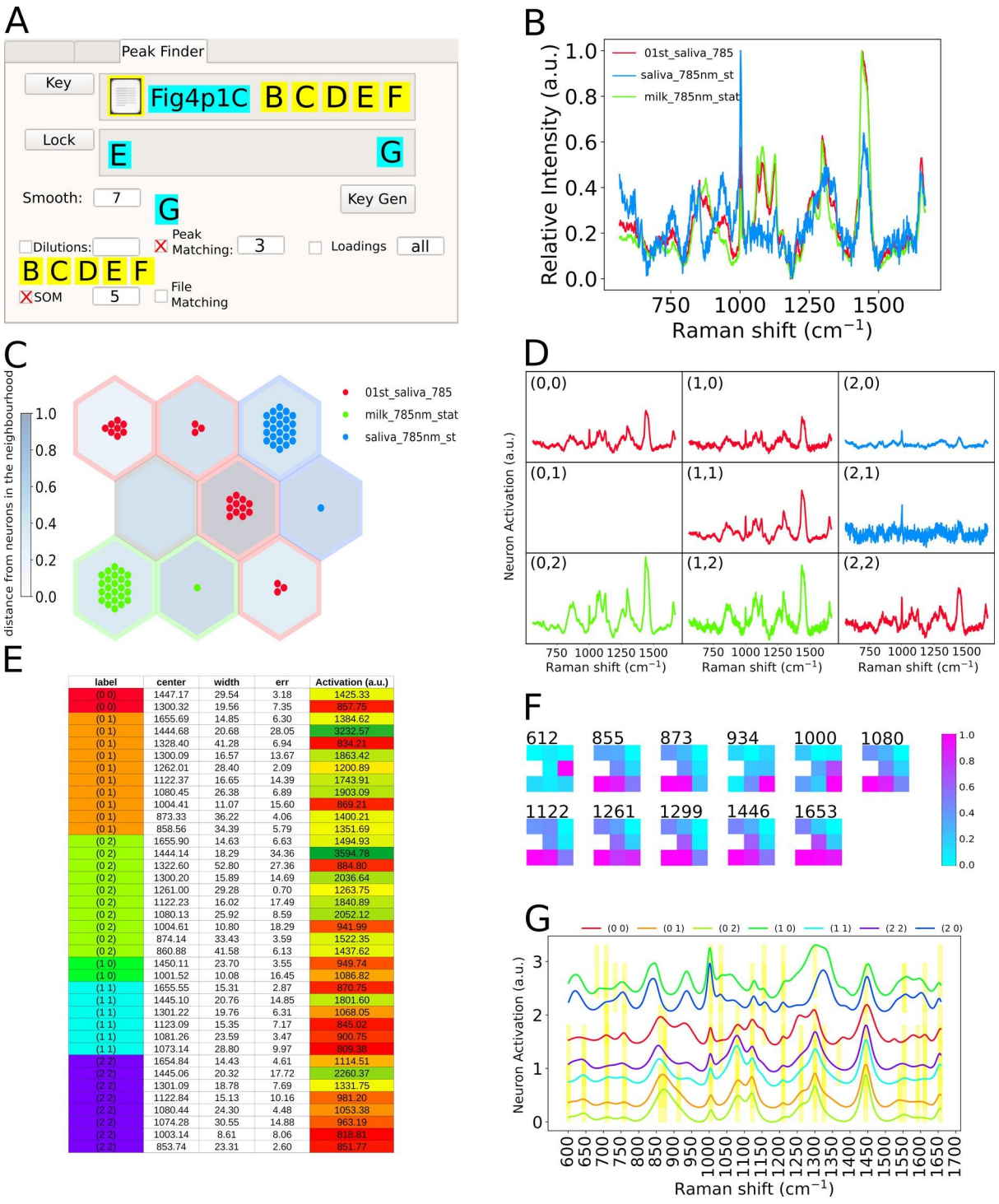
7.4.5 Self Organizing Maps Peak Finder

The SOM peak finding method works similarly to the PCA. However, we check the neuron activation spectra instead of the loadings. The SOM script can be executed following two procedures highlighted in yellow and cyan (**Figure 7.7A**). Starting with the one highlighted in yellow, we load a Raman map file onto the 'key' section as described in **Figure 5Aii**, toggle the SOM tick-box, and press the keygen button. The SOM size can be customized by adjusting the parameter next to the SOM toggle box. A good approximation of the SOM size is the fourth root of the total number of spectra present in all the sample maps. After running the script, multiple plots are output, such as the normalized intensity Raman plot of all the samples (**Figure 7.7B**). The SOM plot (**Figure 7.7C**) highlights three important features. The different sample placements within the map. Each dot present in the SOM represents a spectrum in the Raman spectra map. The hexagons that compose the SOM have an inner colour representing how specific the neuron activation is for the spectrum present in the hexagon. Finally, the hexagon edge colour is related to the highest sample population in each hexagon. Complementing the SOM plot, we have the neuron activations plot (**Figure 7.7D**), highlighting the different activation 'spectra' per neuron. Information regarding each hexagon's spectrum profile is obtained from this plot by matching the x and y coordinate position to the hexagon position, and the peak finding method is applied to these activation plots. The multi-Lorentzian peak fitting method is applied to each activation 'spectra' as described previously, and the fitting parameters are stored in a CSV file (**Figure 7.7E**), with additional information related to the SOM hexagonal position for the peaks to be tracked.

Moreover, the peaks with the highest activation values are stored, and a heatmap plot is constructed regarding these peak position values per hexagon (**Figure 7.7F**).with

each heatmap position matches the SOM position. Thus, the peaks can be tracked to the SOM locations from the heatmap.

Finally, the second option to run the SOM peak finder script is highlighted in cyan and requires the activation multi-Lorentzian fit CSV file (**Figure 7.7E**) inserted in the 'lock' file option. Together with the standard Raman spectra multi-Lorentzian peak fit CSV file (**Figure 7.5C**) inserted in the 'key' section. To run this routine, toggle both the 'peak matching' option and the 'SOM' option and push the keygen button. This combination of options outputs the sum of the multi-Lorentzian fits (**Figure 7.6G**) with the respective SOM x and y location. The resulting plot matches the peaks between the two CSV files loaded, i.e. the SOM activation 'spectra' and the peaks from the standard Raman spectra multi-Lorentzian fit. Each matching depends on which hexagon coordinates the matching occurs; thus, each highlighted yellow bar in (**Figure 7.7G**) only matches a specific peak per hexagon, which helps to understand better the different samples in terms of their multivariate analysis and peak distributions.



label	center	width	err	Activation (a.u.)
(0 0)	1447.17	29.54	3.18	1425.33
(0 0)	1300.32	19.56	7.35	857.75
(0 1)	1655.69	14.85	6.30	1384.62
(0 1)	1444.68	20.68	28.05	3232.57
(0 1)	1328.40	41.28	6.94	834.21
(0 1)	1300.09	16.57	13.67	1863.42
(0 1)	1262.01	28.40	2.09	1200.89
(0 1)	1122.37	16.65	14.39	1743.91
(0 1)	1080.45	26.38	6.89	1903.09
(0 1)	1004.41	11.07	15.60	869.21
(0 1)	873.33	36.22	4.06	1400.21
(0 1)	858.56	34.39	5.79	1351.69
(0 2)	1655.90	14.63	6.63	1494.93
(0 2)	1444.14	18.29	34.36	3594.78
(0 2)	1322.60	52.80	27.36	884.80
(0 2)	1300.20	15.89	14.69	2036.64
(0 2)	1261.00	29.28	0.70	1263.75
(0 2)	1122.23	16.02	17.49	1840.89
(0 2)	1080.13	25.92	8.59	2052.12
(0 2)	1004.61	10.80	18.29	941.99
(0 2)	874.14	33.43	3.59	1522.35
(0 2)	860.88	41.58	6.13	1437.62
(1 0)	1450.11	23.70	3.55	949.74
(1 0)	1001.52	10.08	16.45	1086.82
(1 1)	1655.55	15.31	2.87	870.75
(1 1)	1445.10	20.76	14.85	1801.60
(1 1)	1301.22	19.76	6.31	1068.05
(1 1)	1123.09	15.35	7.17	845.02
(1 1)	1081.26	23.59	3.47	900.75
(1 1)	1073.14	28.80	9.97	809.38
(2 2)	1654.84	14.43	4.61	1114.51
(2 2)	1445.06	20.32	17.72	2260.37
(2 2)	1301.09	18.78	7.69	1331.75
(2 2)	1122.84	15.13	10.16	981.20
(2 2)	1080.44	24.30	4.48	1053.38
(2 2)	1074.28	30.55	14.88	963.19
(2 2)	1003.14	8.61	8.06	818.81
(2 2)	853.74	23.31	2.60	851.77

***Figure 7.7**: **A)i)** GUI - SOM Peak Finder section. Highlighted in yellow and cyan are the two different operation methods of the script. Yellow generates the multi-Lorentzian fit CSV file of the SOM neuron activations, and cyan compares the peaks from the activation CSV file with the peaks from the standard peak CSV file. **B)** Normalized average spectrum of the different samples loaded into 'key'. **C)** Self-Organizing maps in a three by three hexagonal configuration, separating the different samples into different hexagons highlighted with separation boundaries and edge-colouring related to the highest hexagon sample population. **D)** Activation Raman' spectrum' related to the different hexagons positions. **E)** Peak fitting parameter table of the different Lorentzian fits to the SOM activation 'spectrum' **(D)**. **F)** Heatmap representing each SOM hexagon of the important peak fit centre values from the activation 'spectrum'. **(D)**. **G)** Multi-Lorentzian sum for the different SOM activations per hexagon, with peak matching result between the SOM activation CSV and the imported 'key' CSV file.*

The multiple peak analysis described here was used to characterise the sample, which included multi-Lorentzian peak fitting to average spectra, PCA loadings, and SOM activations. It is now possible to gain complete insight into the various nuances among different samples by using multiple methods to cross-validate peak changes, resulting in greater certainty in the analysis and a dependable multivariate analysis methodology.

7.5 Conclusions

Proper Raman data analysis requires high levels of expertise in the data science field, which might not be common to all Raman spectrometer users. Additionally, Raman data analysis software included with the Raman spectrometer lack advanced multivariate analysis tools and thus have limited options in the data processing tools for the user. Finally, when performing Raman compound analysis, the libraries available are usually very specific to general chemical compounds and lack the flexibility for unknown substances. With the aid of an intuitive GUI, a script capable of handling various Raman data analysis choices was created. The GUI has three main tools, one for Raman data plotting with multiple options, individual, stacked, heatmap and waterfall plot. Another for multivariate analysis with a complete PCA analysis with the PCA plot, loadings and classification stores, k-nearest neighbours, confusion matrix and the ROC curve for a complete multivariate analysis and samples classification. Finally, a peak analysis script that uses a multi-Lorentzian fit approach to fit all the peaks present in a Raman map in the PCA loadings and the SOM activations. The multi-Lorentzian fit analysis parametrizes the fit into a CSV file, thus creating any customized peak library for any compound for a posteriori comparison. Moreover, the script has been optimized to run using the least

number of inputs, mainly the Raman map TXT file and one parameter. Thus, it is very convenient and easy to use.

Furthermore, each user can have their peak library and share it with other scientists. These libraries can also be compared with user-built CSV files from literature peaks. The matching feature also outputs an intensity ratio comparison for matched peaks making the script a quantitative and quantitative analysis tool for Raman spectra.

Overall, the multiple scripts created and combined through the GUI solve most Raman data analysis problems in standard Raman literature. Moreover, with the option of multiple Raman peak fitting methods, which generates each user their peak library, which can be used to compare with other Raman samples, PCA Loadings and SOM activation. Thus, making all output libraries compatible, and peak matching can be perform between them.

With this solution at hand, the problem posed in the introduction of finding the relation between sample ζ and sample Ψ, which depends on ζ, i.e. Ψ(ζ), can be completely understood with your peak finding solution, using the combination of Raman peak tracking and PCA/SOM peak tracking to get a complete understanding of the samples. Thus, discerning between sample ζ and Ψ and which parts of sample Ψ resemble ζ.

For future work, integrating a combination of Gaussian, Lorentzian and Voigt for the peak fitting method instead of only Lorentzian and let the user decide on which to use, thus, perfecting the 2nd residual calculation to approximate zero. Additionally, the Raman analysis could consider the x and y coordinate values of the Raman map and use these to generate a 2D intensity map highlighting the important peaks. Finally, incorporate a

supervised multivariate analysis method, the linear discriminant analysis (LDA), into the analysis methods.

7.6 References

[1] Raman CV. A new radiation. Vol. 37, Proceedings of the Indian Academy of Sciences - Section A. Springer Science and Business Media LLC; **1953**. p. 333–41. Available from: http://dx.doi.org/10.1007/BF03052651.

[2] Pezzotti G. Raman spectroscopy of piezoelectrics. Vol. 113, Journal of Applied Physics. AIP Publishing; **2013**. p. 211301. Available from: http://dx.doi.org/10.1063/1.4803740.

[3] Ferrari AC, Meyer JC, Scardaci V, Casiraghi C, Lazzeri M, Mauri F, et al. Raman Spectrum of Graphene and Graphene Layers. Vol. 97, Physical Review Letters. American Physical Society (APS); **2006**. Available from: http://dx.doi.org/10.1103/PhysRevLett.97.187401.

[4] Jehng JM, Wachs IE. Structural chemistry and Raman spectra of niobium oxides. Vol. 3, Chemistry of Materials. American Chemical Society (ACS); **1991**. p. 100–7. Available from: http://dx.doi.org/10.1021/cm00013a025.

[5] Sarycheva A, Gogotsi Y. Raman Spectroscopy Analysis of the Structure and Surface Chemistry of Ti3C2Tx MXene. Vol. 32, Chemistry of Materials. American Chemical Society (ACS); **2020**. p. 3480–8. Available from: http://dx.doi.org/10.1021/acs.chemmater.0c00359.

[6] Kuhar N, Sil S, Verma T, Umapathy S. Challenges in application of Raman spectroscopy to biology and materials. Vol. 8, RSC Advances. Royal Society of Chemistry (RSC); **2018**. p. 25888–908. Available from: http://dx.doi.org/10.1039/C8RA04491K.

[7] Zhang C, Zhang D, Cheng JX. Coherent Raman Scattering Microscopy in Biology and Medicine. Vol. 17, Annual Review of Biomedical Engineering. Annual Reviews; **2015**. p. 415–45. Available from: http://dx.doi.org/10.1146/annurev-bioeng-071114-040554.

[8] Zhang X, Qiao XF, Shi W, Wu JB, Jiang DS, Tan PH. Phonon and Raman scattering of two-dimensional transition metal dichalcogenides from monolayer, multilayer to bulk material. Vol. 44, Chemical Society Reviews. Royal Society of Chemistry (RSC); **2015**. p. 2757–85. Available from: http://dx.doi.org/10.1039/c4cs00282b.

[9] R. Beams, L. Gustavo Cançado, and L. Novotny, “Raman characterization of defects and dopants in graphene,” J. Phys. Condens. Matter, vol. 27, no. 8, 2015, doi: 10.1088/0953-8984/27/8/083002.

[10] Ember KJI, Hoeve MA, McAughtrie SL, Bergholt MS, Dwyer BJ, Stevens MM, et al. Raman spectroscopy and regenerative medicine: a review. Vol. 2, npj Regenerative Medicine. Springer Science and Business Media LLC; **2017**. Available from: http://dx.doi.org/10.1038/s41536-017-0014-3.

[11] Nabiev I, Chourpa I, Manfait M. Applications of Raman and surface-enhanced Raman scattering spectroscopy in medicine. Vol. 25, Journal of Raman Spectroscopy. Wiley; **1994**. p. 13–23. Available from: http://dx.doi.org/10.1002/jrs.1250250104.

[12] Izake EL. Forensic and homeland security applications of modern portable Raman spectroscopy. Vol. 202, Forensic Science International. Elsevier BV; **2010**. p. 1–8. Available from: http://dx.doi.org/10.1016/j.forsciint.2010.03.020.

[13] Muehlethaler C, Leona M, Lombardi JR. Review of Surface Enhanced Raman Scattering Applications in Forensic Science. Vol. 88, Analytical Chemistry. American Chemical Society (ACS); **2015**. p. 152–69. Available from: http://dx.doi.org/10.1021/acs.analchem.5b04131.

[14] Pence I, Mahadevan-Jansen A. Clinical instrumentation and applications of Raman spectroscopy. Vol. 45, Chemical Society Reviews. Royal Society of Chemistry (RSC); **2016**. p. 1958–79. Available from: http://dx.doi.org/10.1039/c5cs00581g.

[15] Choo-Smith LP, Edwards HGM, Endtz HP, Kros JM, Heule F, Barr H, et al. Medical applications of Raman spectroscopy: From proof of principle to clinical implementation. Vol. 67, Biopolymers. Wiley; 2002. p. 1–9. Available from: http://dx.doi.org/10.1002/bip.10064.

[16] Austin LA, Osseiran S, Evans CL. Raman technologies in cancer diagnostics. Vol. 141, The Analyst. Royal Society of Chemistry (RSC); **2016**. p. 476–503. Available from: http://dx.doi.org/10.1039/c5an01786f.

[17] Kallaway C, Almond LM, Barr H, Wood J, Hutchings J, Kendall C, et al. Advances in the clinical application of Raman spectroscopy for cancer diagnostics. Vol. 10, Photodiagnosis and Photodynamic Therapy. Elsevier BV; **2013**. p. 207–19. Available from: http://dx.doi.org/10.1016/j.pdpdt.2013.01.008.

[18] Xiong M, Ye J. Reproducibility in surface-enhanced Raman spectroscopy. Vol. 19, Journal of Shanghai Jiaotong University (Science). Springer Science and Business Media LLC; **2014**. p. 681–90. Available from: http://dx.doi.org/10.1007/s12204-014-1566-7.

[19] Renishaw WiRE, URL="https://www.renishaw.com/en/raman-software—9450".

[20] Horiba LabSpec 6 Spcetroscopy Suite Software, URL="https://www.horiba.com/int/products/detail/action/show/Product/labspec-6-spectroscopy-suite-software-1843/".

[21] Python, URL="https://www.python.org".

[22] Top 10 Programmin Languages to Learn in 2022 URL="https://www.geeksforgeeks.org/top-10-programming-languages-to-learn-in-2022/".

[23] Spyder , URL="https://www.spyder-ide.org/".

[24] PyQy5, URL="https://pypi.org/project/PyQt5/".

[25] Eilers PHC. A Perfect Smoother. Vol. 75, Analytical Chemistry. American Chemical Society (ACS); **2003**. p. 3631–6. Available from: http://dx.doi.org/10.1021/ac034173t.

[26] Eilers PHC and Boelens HFM, "Baseline correction with asymmetric least square smoothing [unpublished]," 2005. It may be available on the internet.

[27] ALS python, URL="https://github.com/perimosocordiae".

[28] SciPy, URL="https://scipy.org/".

[29] Solve bands, URL="https://docs.scipy.org/doc/scipy/reference/generated/scipy.linalg.solveh_banded.Html".

[30] Savgol filter, URL="https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.savgol_filter.html".

[31] Z-score, URL="https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.zscore.html".

[32] PCA, URL="https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html".

[33] Gaussian Mixtures, URL="https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html".

[34] k-Nearest Neighbours, URL="https://scikit-learn.org/stable/modules/neighbors.html".

[35] Train-test split, URL="https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html".

[36] Cross validation score, URL="https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html".

[37] Receiver Operating Characteristic (ROC), URL="https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html".

[38] Confusion Matrix, URL="https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html".

[39] V. Giuseppe, “MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map”,2018, URL= https://github.com/JustGlowing/minisom/

[40] Huang WQ, Li Y, Li CM, Xu RC. New heuristics for packing unequal circles into a circular container. Vol. 33, Computers & Operations Research. Elsevier BV; 2006. p. 2125–42. Available from: http://dx.doi.org/10.1016/j.cor.2005.01.003.

[41] SOM base hexagon script: https://github.com/JustGlowing/minisom/blob/master/examples/Hexagonal

Topology.ipynb visited 20:20 07/01/2023.

[42] SOM win_map script section: https://github.com/JustGlowing/minisom/blob/master/minisom.py visited 20:20 07/01/2023.

[43] Circlify, URL=”https://pypi.org/project/circlify/” visited 20:25 07/01/2023.

[44] Peak fitting, URL=”https://github.com/emilyripka/BlogRepo” visited 20:30 07/01/2023.

CHAPTER 8

Summary and Future Outlook

8.1 Summary

This doctoral thesis has demonstrated the capabilities of using convolutional neural networks to help classify different patterns regarding their SERS activation with an accuracy between 90-99%. Additionally, multiple approaches were elaborated for an ideal way of analysing Raman data, such as self-organising maps, principal component analysis and multi-peak fitting analysis resulting in an in-depth singleplex and multiplex Raman spectral data analysis.

In particular, in Chapter 4, convolutional neural networks were used to characterise atomic force microscopy images of patterned surfaces in order to gain a better understanding of their SERS activations. In Chapter 5 optical images, it was possible to identify SERS active EHD pillars and even to integrate the code in a real-time experimental data acquisition test, resulting in accurately identifying SERS pillars for a consistent and reproducible SERS measure.

Another aspect to highlight is the use of Python as the chosen language for writing and executing all coded scripts dues to its capabilities in executing multiple different tasks. From 3D simulation of the EHD dispersion equation in Chapter 3 to using different CNN

architectures in Chapters 4 and 5 and multiple ML algorithms like SOM and PCA in Chapter 7. Furthermore, in Chapter 7, multiple plot options were investigated (2D and 3D), and a GUI was developed for an interactive Raman data analysis experience, thus showing Python's flexibility for problem-solving in the scientific world.

Finally, in Chapter 6, a systematic analysis of Raman spectral data was carried out in the most systematic analysis methods performed by many authors with a special interest in the saliva Raman field. Since saliva extraction is a non-invasive procedure and is easy to acquire, it naturally developed a special interest with potential applications for the detection of different illnesses when used as an analyte. In order to do so, the different analysis methods were organised and coded into a functional GUI presented in Chapter 7 that can be accessed and run by just giving the Raman data as input. Scripts are executed by just clicking a button, and then the results are output and saved with high quality. For every script executed, an additional text file is generated that collects the multi-Lorentzian peak fitting parameters for each Raman special loaded, thus automatically generating a Raman peak library for each sample that can be shared and used by other Raman scientists.

8.2 Future Outlook

The work started here lays initial foundations that can be picked up and explored for the future. Starting from the first experimental chapter, chapter 3, the tilted geometry and its effects in the EHD setup was explored resulting in the generation of an assortment of different pillar structures. Thus a microfabricated top electrode surface with a controllable tilting angle can be employed to generate controllable EHD patterns structures with a specific height gradient that could open paths for new optical effects using EHD structures.

In chapters 4 and 5, the use of pretrained CNN for tracking SERS EHD structures was explored using two different approaches: AFM images and optical microscope images. Two different bottlenecks were identified for this task. Regarding the AFM images, the bottleneck is the acquisition time to arrange sufficient images to have a large representation of the different classes. It requires large amounts of time, and then to have all images with very high quality, even more, time is required. Thus, a solution to this problem was to use the optical microscope, which makes it easy to acquire large amounts of images. However, the bottleneck is now the resolution of said images. Since training CNN always requires large amounts of data to reduce the overfitting of the model, obtaining the images through the optical microscope will always be a better option. The methodology of using data augmentation for the images and the pre-trained VGG16 with fine-tuning were very good. The data augmentation method managed to reduce the overfitting, and the pre-trained VGG16 with fine-tuning generated models with very high accuracy. Thus these methods seem ideal for generating any other kind of image identification/classification solution. One aspect that can be improved is to automate the segmentation of individual images from larger images. We manually selected all images in our work, which is a very time-consuming task. Thus automated way of using the cv2.contour function[1] was proposed. That can create contours in images, individually select images of interest, and crop them for an automated image generator. This tool can be used to make the task of identifying images to classify quicker, requiring no manual work. Another improvement could be made by setting up a new base architecture for CNN. The CNN architecture used in this work was the VGG16. However, other architectures might work faster and potentially more accurately, such as the YOLOv3 [2], ResNet [3], and AlexNet [4], which can also be integrated while acquiring images in real-time. Besides applications for SERS/non-SERS classifications, the biomedical field has many different

problems requiring proper classification, usually done by highly qualified medical professionals. However, if properly trained, these CNN can replace the medical profession in the diagnoses and free the professional to perform more complex tasks besides the classification tasks. So we envisioned that CNN would be employed in multiple medical settings in the future.

In chapter 6, saliva as a bioanalyte was introduced, as was how multiple research groups analyse saliva in terms of its Raman fingerprint. It was demonstrated that a correlation could be obtained for differences in the saliva spectra with different illnesses. However, these changes are usually very subtle, and thus machine learning methods are employed to distinguish the differences properly. Here, saliva is a good analyte for diagnostics of different cancer types, so we can estimate that saliva will be more used for these cancer tracing due to its non-invasive properties. Moreover, machine learning techniques have shown to be very powerful in identifying small changes between spectra, and thus it is a requirement to use them in proper medical diagnostics. All the different techniques inspected here culminate with the development of the GUI in chapter 7. Where all the different tools can be employed and used on the spot.

Additionally, the most important tool is the multi-Lorentzian peak fitting, which was developed to generate the Raman signal of the sample and inspect all the relevant peaks from it. These peaks can be inspected through the normal Raman spectral analysis and the different machine learning tools, PCA and SOM. For this different analysis, a CSV file with all the spectra peaks is created. This file can be used to compare different spectra and research groups and share them among peers, thus creating a common experimental Raman peak library. This generated CSV can be used anywhere, and it will be a valuable tool for diagnostics.

Overall, the interdisciplinary relationship between computational tools, micro-fabricated patterns, and a variety of characterization techniques demonstrated in this thesis is groundbreaking. The Python-based framework used in this research is highly versatile and can be applied to a wide range of other pattern-generating techniques such as focused ion beam lithography, electron beam lithography, optical lithography, nanoimprinting, microfabrication, and nanoparticles, as well as measurement tools like SEM and Fourier-transform infrared spectroscopy (FTIR). The potential of this programming tool to empower users to explore their own creativity during testing and problem-solving processes is truly innovative and has the potential to revolutionize the field. The combination of traditional methods with novel computational tools during the analysis has been proven, throughout this thesis, to be an extremely effective approach to the analysis of micro-fabricated patterns and the processing of their data. The proposed innovations in this research are particularly impactful for the scientific community, as the micro-fabricated patterns studied have a wide range of potential applications in various fields, particularly in the field of health.

8.3 References

[1] Cv2 Contour, URL="https://docs.opencv.org/4.x/d4/d73/tutorial_py_contours_begin.html", access: 29/04/2022

[2] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. arXiv; **2018**. Available from: https://arxiv.org/abs/1804.02767.

[3] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv; **2015**. Available from: https://arxiv.org/abs/1512.03385.

[4] Alom MZ, Taha TM, Yakopcic C, Westberg S, Sidike P, Nasrin MS, et al. The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. arXiv; **2018**. Available from: https://arxiv.org/abs/1803.01164.

Appendices

APPENDIX A

SUPPORTING INFORMATION FOR CHAPTER 4

S1 Representative Average Surface Roughness in the Centre of EHD Fabricated Pillar Topology

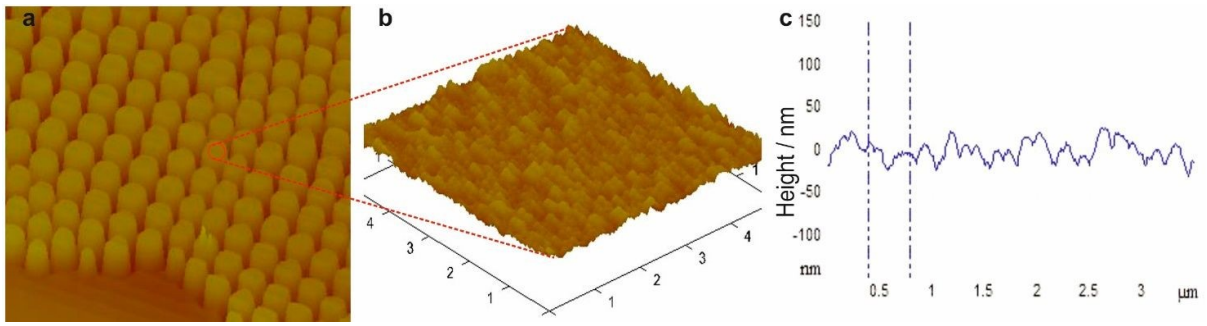


Figure S1. (a) *Representative three-dimensional AFM height image of the EHD fabricated pillar structures with the higher magnification* ***(b)*** *AFM image of the average surface roughness.* ***(c)*** *The calculated arithmetic average of the height deviations from the centre plane of the surface in the central pillar region was on average of 30.0±12.0 nm per 5 µm surface area.*

S2 Calculation of the Enhancement Factor and Reproducibility Coefficient

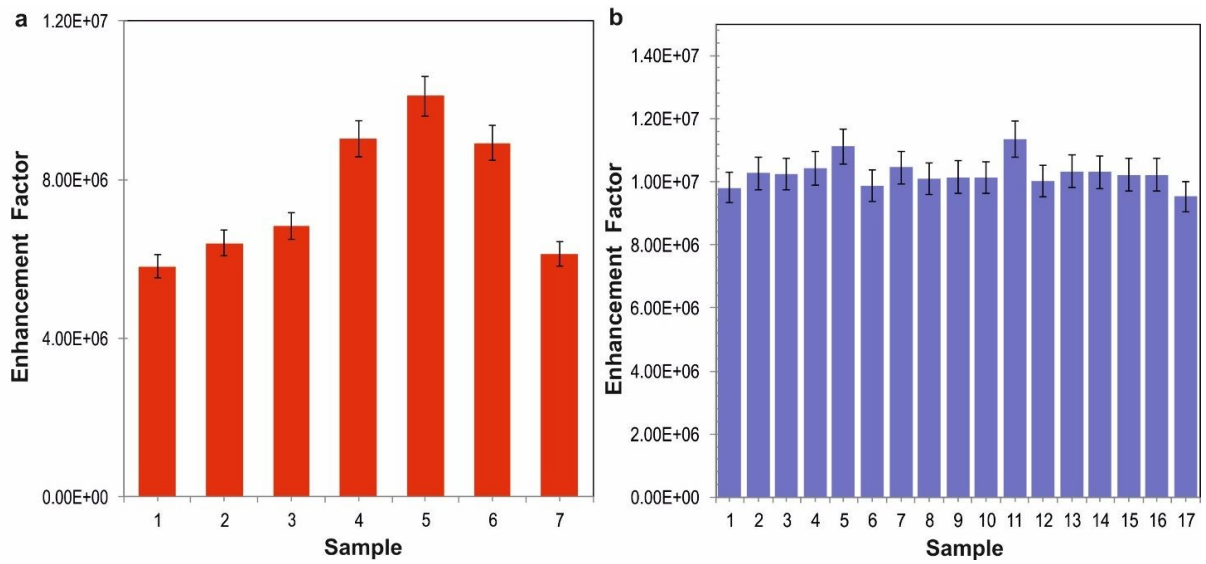


Figure S2. (a) *Absolute enhancement factor values of the seven representative pillars in* ***Figure 2****.* ***(b)*** *For the enhancement factors from the 17 reproducible substrates fabricated via the micronanolithographic EHD apparatus, the reproducibility coefficient and the confidence interval of the mean EF were found to be: Standard Deviation*$_{Relative\ EF}$*×2.8×100%=8.55%: (9.4x10*^{6}*, 1.1x10*^{7}*). All of the achieved EFs are on the order of magnitude of x10*$^{6-7}$ *with the small variation in the pre-factor values, with the largest difference in the SERS EFs of less than two fold for the between the highest and the lowest substrate, indicating a good signal reproducibility. This is in agreement with the conventional SERS platforms, considered highly reproducible even when showing nearly one order of magnitude difference.* [22] *The confidence interval limits for the enhancement factor were calculated based on the mean difference of -1.96 × standard deviation (differences) and mean difference of +1.96 × standard deviation (differences).*

S3 Reproducibility of EHD-SERS Substrates - Uniform Platforms for Reliable SERS

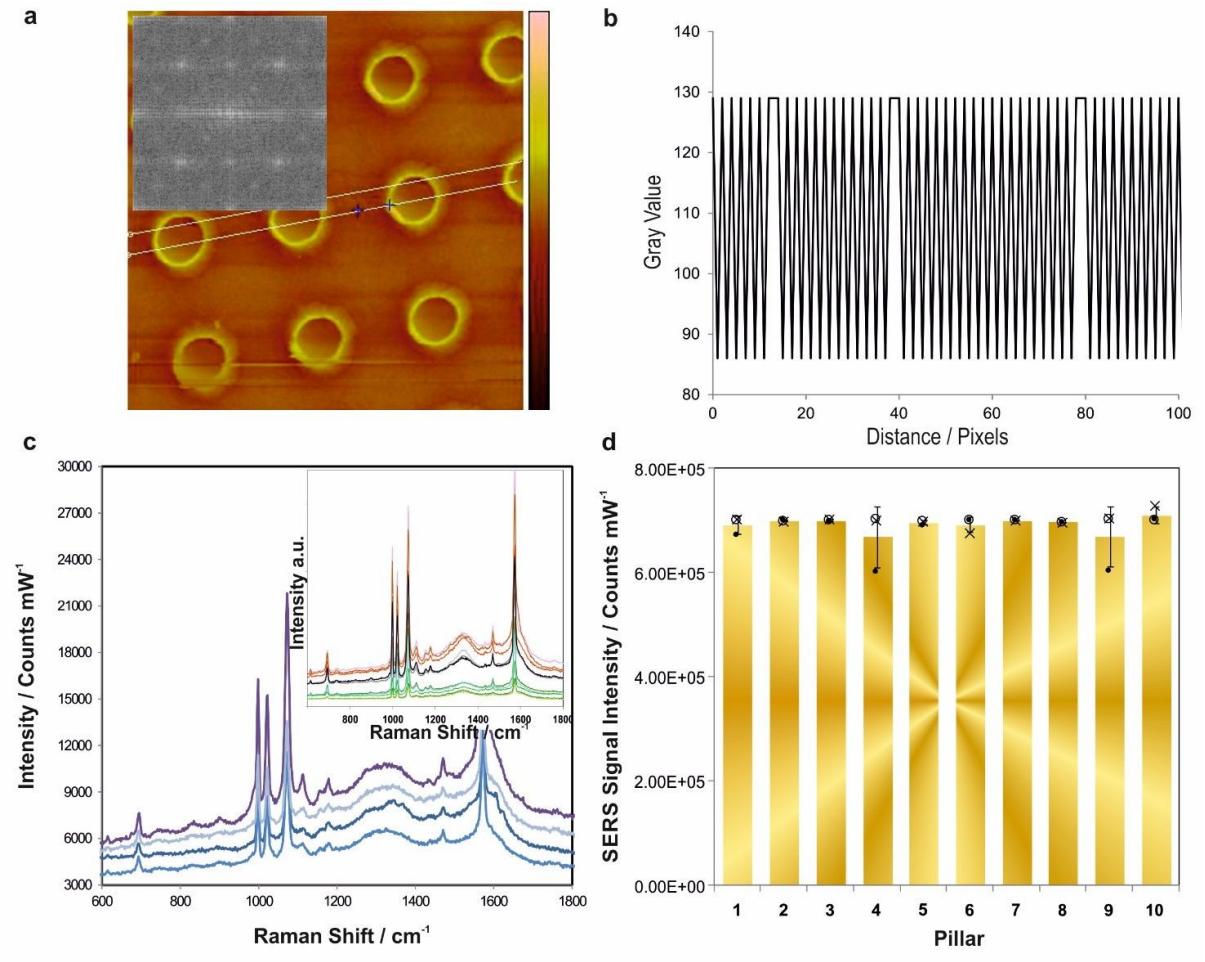


Figure S3. (a) *AFM image of the uniform periodic EHD-fabricated nanoplatforms and the corresponding fast Fourier transform (FFT) image* ***(inset)****.* ***(b)*** *Ingrained by the plot profile of the AFM image the grey value distribution as function of distance across the image, demonstrating highly-consistent structural units with low and high frequencies containing similar image information with two dominating directions in the FFT image passing through the centre, both originating from the uniform structures in the background of the original image.* ***(c)*** *These structural units provide a highly-consistent SERS signal over the entire substrate surface area (200x200μm^2) with pillars distribution at the surface in regards to the laser spot size (diameter: ~1.0μm) yielding uniform signal. SERS spectra acquired from four random locations across each substrate as well as from* ***(inset)*** *three different substrates at average of five locations (i.e., pillars) each.* ***(d)*** *Representative SERS spectra of BT on uniform substrates across ten random pillars on each substrate, for the 1070 cm^{-1} peak. Repeatable SERS response was obtained from the pillars with relative standard deviation values of less than 3.0% with an average variation of less than 7.5% in terms of the relative peak intensities in the framework of one substrate (error bars) and between the different samples (height of the bars).*

S4 Stability Analysis

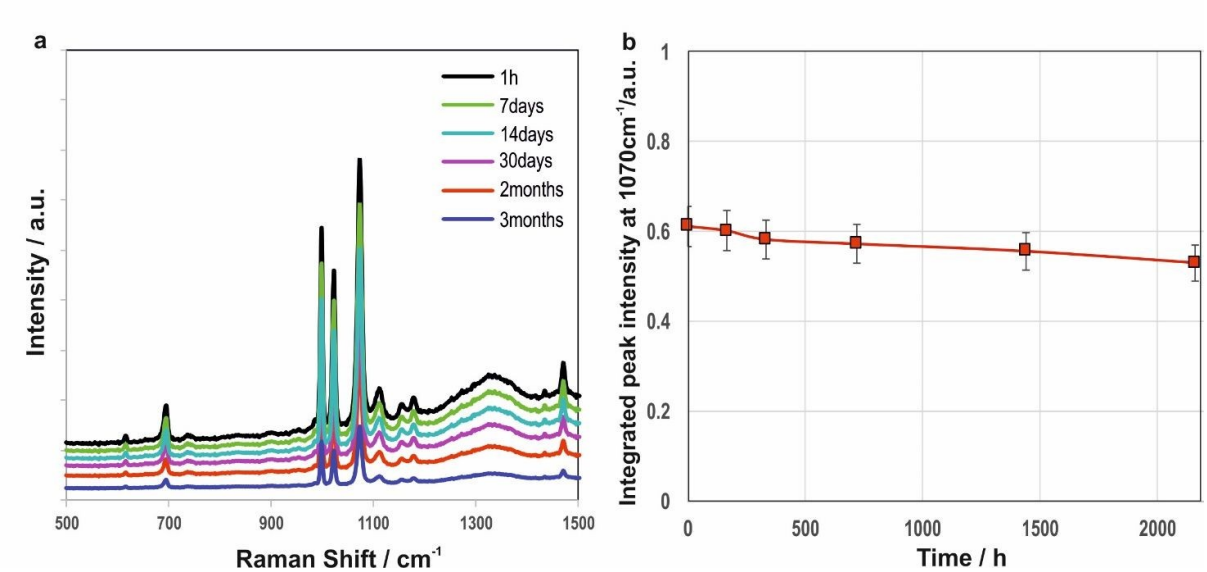


Figure S4. (a) *Characterization of SERS activity with the representative Raman spectra of BT on the EHD-SERS substrates.* ***(b)*** *Temporal evolution of the signal intensity of BT at 1070 cm^{-1} with a very minor variation of less than 1% within the first 2 weeks and overall of 5.7% over the period of 3 months. Within the experimental error, the substrates are stable over long periods of time.*

S5 Spike-and-Recovery (S&R) Proof-of-Concept Assay of the EHD SERS Substrate TBI Biomarker

a

Sample		No spike (0 pg.mL)	Spiked (~50 pg/mL)
Control (diluent)		0	53
1		0	51.5
2		0.7	48.9
3		0	51.7
4		0	47.1
5		0	50.9
6		0.7	49.3
7		0.5	52.8
8		2.3	51.5
9		0	49.8
10		0.3	53
Average			50.65
			0.044339623
Mean Recovery %			**95.56603774**
(+/- S.D.)			(+/-3.4%)

b

Raman Band/cm^{-1}	Peak Assignment
682m	γ (C=O) τ (O-H) in *N*-acetyl bending mode of C=O·O-
717m	$\sigma(C_{\alpha}C)$. *def*(O-H)
930s	ν(C-C) and N-C symmetric stretching
1010w	out of plane $\rho(CH_3)$, $\tau(CH_2)$ γ (C=O) amides
1168m	γ (C=O) and δ(C-OH)
1420m	$\sigma(COO^-)$, acetylamide in the CH_3-NH-C=O of the *N*-acetyl asymmetric *def*(C_tH_3) and $\delta_s(CH_3)$
1517s	δ(N–H) in (NH-C=O-CH_3) group

Figure S5. (a) *S&R validation assay of the EHD-SERS substrates establish a difference between the diluent and the analytical sample matrix. 10μL of analyte spiked in the 50μL test sample matrix showed a mean recovery (%) response ± standard deviation (S.D.) of 95.5 ± 3.4 in the assay relative to a duplicate spike in the standard diluent.* ***(b)*** *Representative SERS peak assignments of the N-acetylaspartate.*

S6 Characteristic Fingerprint SERS Spectra and Barcode of Peaks of the N-acetylaspartate Marker

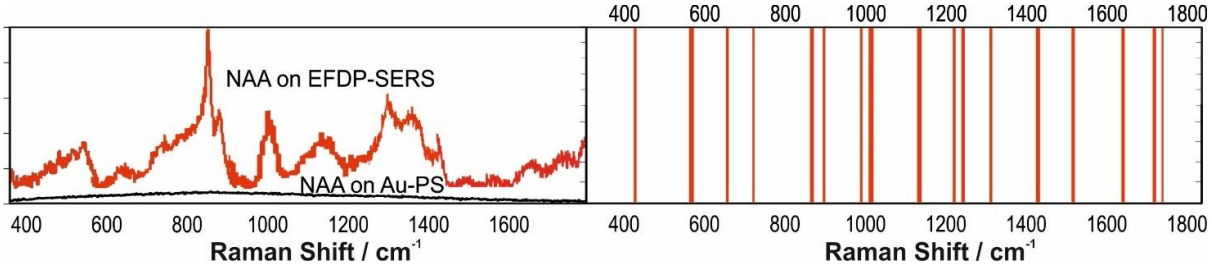


Figure S6. Left: *Average Raman spectra of the N-acetylaspartate using the EHD-SERS substrates (red) and using the reference gold-nanocoated PS film of the same thicknesses used for the EHD patterning but with no structures on the surfaces (black) recorded at 785 nm excitation wavelength, 3 mW laser power, with the EHD substrates, used as fingerprint references.* ***Right:*** *The corresponding barcode for the marker. The characteristic finger-prints include the peak at 682 cm^{-1}, attributed to the C=O wagging and twisting of –OH bonds in the N-acetyl, the 717 cm^{-1} peak originating from the stretching of $C_{\alpha}C$ bonds as well as the deformation of the O-H groups. The peaks at and between 1010 cm^{-1}, 1168 cm^{-1} are from the out-of-plane CH_3 rocking mode, twisting of the CH_2, the wagging of the C=O amides and the N-H vibrations in the NH-C=O-CH_3 group at 1517 cm^{-1} in the N-acetyl. The peak at 1420 cm^{-1} originates from the acetylamide in the CH_3-NH-C=O of the asymmetric vib CH_3 deformations.*

S7 Roughness of the Au-PS film

The average roughness (Ra) of the pristine as-spun polymer films and Au coated polymer substrates were analysed using AFM. The mean roughness value represents the arithmetic average of the surface height deviations measured from the centre plane of the sample calculated from five independent measurements.

The AFM analysis indicates that the roughness of the pristine polymer film is slightly lower compared to those after the deposition of the gold nano-layer. The overall roughness after gold deposition is only of the order of ~nm which is not enough to generate any SERS signal under our experimental conditions.

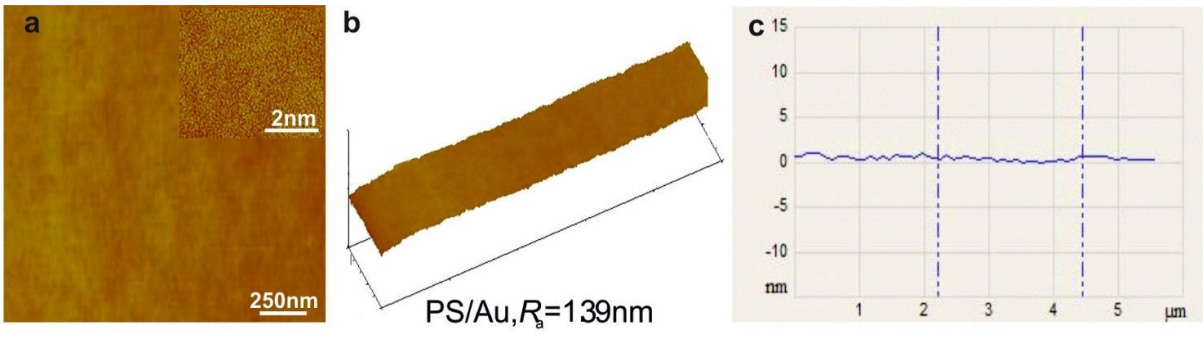


Figure S7*. AFM height* ***(left)*** *and three-dimensional morphology* ***(middle)*** *images of* ***(a)*** *as-spun polymer film and of gold layers* ***(b)*** *deposited after two cycles of 30 s long sputtering at 70 mA on the polymer coated substrate. The cross-sections of the corresponding AFM scans* ***(right hand side)*** *in (**a** and **b**) show an average roughness of pristine polymer film and the gold.*

S8 Quantitative Detection of Representative Biomarker via EHD-SERS

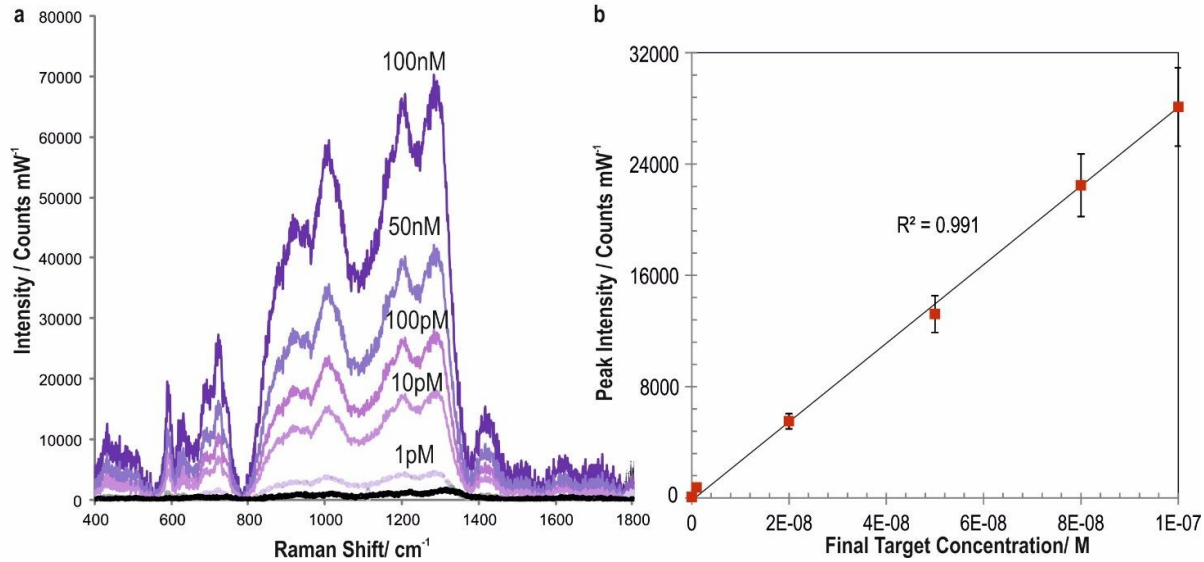


Figure S8. *SERS substrates were used as an assay to perform dilutions using the specified concentration range for the biomarker.* ***(a)*** *SERS spectra detected from dilutions at 100 nM to 1 fM concentrations and the* ***(b)*** *corresponding calibration curve used to calculate the limits of detection, which is three times the standard deviation of the blank divided by the gradient of the linear curve. Error bars denote the standard deviation.*

S9 Detectable Biochemical Probes

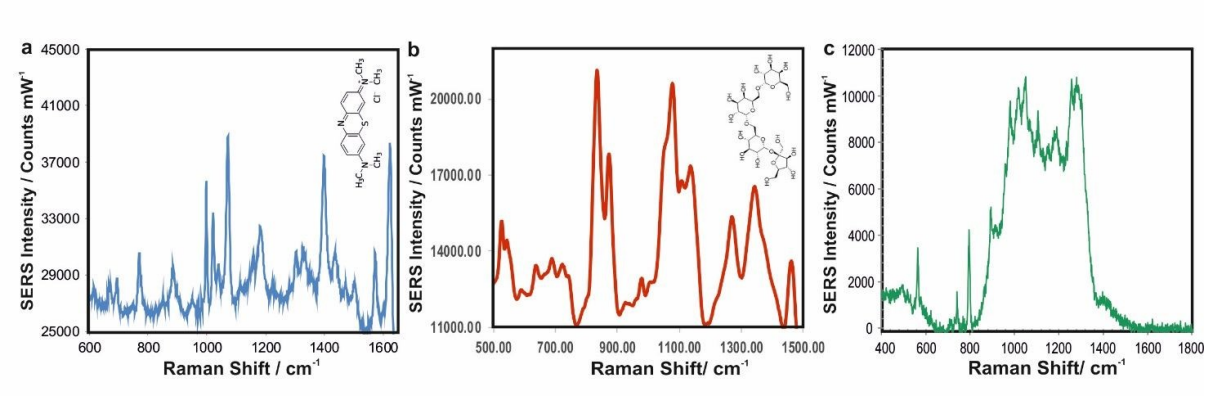


Figure S9. *Representative fingerprint SERS spectra for* ***(a)*** *Methylene Blue,* ***(b)*** *Stachyose and* ***(c)*** *S-one-hundred calcium-binding protein B on EHD-SERS substrates.*

APPENDIX B

PYTHON CODES

The main python scripts created in this document are the whole convolutional neural network code and RoI code discussed in **Chapter 4**, and **Chapter 5** can be inspected in:

https://github.com/PauloAxcel/SERS-EHD-pillars

And the Raman peak finder code with the graphical user interface code from **Chapter 10** can be inspected in:

https://github.com/PauloAxcel/Raman-Peak-Finder

Additionally to the codes already mentioned in the document other codes were implemented for smaller task. Here we provide a description to them plus their code.

B.1 Raman peaks from papers

*This code was designed to read pdfs and extract the Raman peaks from them with the sentence where they appear. This work helped to create the saliva tables in **Chapter 9**.*

***Description** : Easy python code to look for peaks in your papers. Uses pytesseract package (related to tesseract from google) and pdf2image package as the main pdf read/interpret text package. Followed by a search for a specific key work, in this case cm, (which is relate to the Raman peak units, found in all papers). The code extracts the peak, the sentece where that value came, and 3 other columns ('height','importance' and 'width') that are irrelevante in this context but are useful for other ongoing work.*

There are 3 important notes for this code to work you need to 1st install tesseract from https://github.com/UB-Mannheim/tesseract/wiki then pip install pytesseract 2nd install poppler from https://github.com/oschwartz10612/poppler-windows/releases/ 3rd download all your pdf files, that you want to extract the peak detailed information into, and put them into a folder

os.chdir() pytesseract.pytesseract.tesseract_cmd = <your tesseract.exe file dir (terminated in \tesseract.exe)> poppler_path =

***Code:** https://github.com/PauloAxcel/Raman-peaks-from-papers*

“

```
import slate3k as slate
import os
from tqdm import tqdm
from PyPDF2 import PdfFileReader
import re
import pandas as pd
import numpy as np
```

```
import matplotlib.pyplot as plt
from matplotlib import style
from PIL import Image
import pytesseract
import argparse
import cv2
import os
from pdf2image import convert_from_path

pdf_files = []

os.chdir(r"C:\Users\paulo\OneDrive - University of Birmingham\Desktop\birmingham_02\
saliva\GUI for raman analysis\reading raman values from papers\saliva")

for file in os.listdir():
    if file.split('.')[-1] == 'pdf':
        pdf_files.append(file)


pytesseract.pytesseract.tesseract_cmd =  r'C:\Program Files\Tesseract-OCR\tesseract.exe'
poppler_path   =   r'C:\Users\paulo\OneDrive   -   University   of   Birmingham\Desktop\
birmingham_02\articles\saliva\poppler-21.03.0\Library\bin'

all_string = []

new_string = []

num = []

for f in pdf_files:
    pages = convert_from_path(f,poppler_path=poppler_path)
    num.append(len(pages))
    for p in pages:
        text = pytesseract.image_to_string(p)
        string = list(filter(None,text.split('\n')))
        for s in string:
            new_string.append(list(filter(None,s.split('\x00'))))
        flat_s = [item for sublist in new_string for item in sublist]
        b_string = ' '.join(flat_s).split('. ')
        all_string.append(b_string)



all_string = pd.DataFrame(all_string).T
all_string.columns = list(range(sum(num)))

info_to_obtain = 'cm'
info = []
numbers = []

for m in range(sum(num)):
                all_string_clean    =    all_string.iloc[:,m].replace(to_replace='None',
value=np.nan).dropna()
    info.append(all_string_clean[all_string_clean.str.contains(info_to_obtain)].tolist())


flat = list(set([item for sublist in info for item in sublist]))

final_num = []

#range for a control check
low = 100
high = 3200

final_df = []

for t in flat:
    numbers = re.findall(r"[-+]?\d*\.\d+|\d+ "+info_to_obtain, t)
```

```
#         if detect_special_characer(t):
#             pass
#         else:
    for n in numbers:
        if len(n.split(' ')) !=1:
            n = n.split(' ')[0]
        if (float(n)>low and float(n)<high):
            t = ' '.join(t.split(','))
            t = re.sub('[^a-zA-Z0-9\n\.]', ' ', t)
            final_df.append([n,t])

final_df = pd.DataFrame(final_df,columns=['center','label'])
final_df['height'] = [1]*final_df.shape[0]
final_df['importance'] = ['major']*final_df.shape[0]
final_df['width'] = [1]*final_df.shape[0]
final_df.to_csv('peaks_'+f+'.csv',index=False,sep=';')

“
```

B.2 Saliva Peaks Whisker Plot

*This code was designed to plot the saliva peaks whisker plot in **Chapter 9**.*

Code: *https://github.com/PauloAxcel/saliva-peaks-whisker-plot*

“

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt


font = {'family' : 'arial',
        'size'   : 26}

plt.rc('font', **font)


rangex = np.arange(400,2200+1,25)

df = pd.read_csv('saliva data_v04.txt', sep='\t')
df = df.dropna().reset_index(drop=True)

labels = list(df['group'].drop_duplicates())
colors = ['blue','green','red','magenta']


limit = [0,0,0,5]
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score


i=-1
for label , color in zip(labels,colors):
    i = i+1

    index = df['group'] == label



    mols = list(df[' Assignments'][index].drop_duplicates())

    plt.figure(figsize=(36,len(mols)/1.618))

    for mol in mols:
        print(len(mols))

        index2 = df[' Assignments'] == mol
        X = df['Peak position (cm−1)'][index2]


        from sklearn.cluster import MeanShift, estimate_bandwidth
        from sklearn import preprocessing

        if X.shape[0]>1:
            if X.shape[0]>15:
                q = 0.1
            else:
                q = 0.5
```

```
            x1 = list(X.values)

            x = list(zip(x1,np.zeros(len(x1))))
            x = preprocessing.scale(x)
            bandwidth = estimate_bandwidth(x, quantile=q)
            if bandwidth == 0:
                sns.boxplot(y=df[' Assignments'][index],x=X,color = color)

            else:

                ms = MeanShift(bandwidth=bandwidth, bin_seeding=True)
                ms.fit(x)
                labels2 = ms.labels_
                cluster_centers = ms.cluster_centers_

                labels_unique = np.unique(labels2)
                n_clusters_ = len(labels_unique)

                for k in range(n_clusters_):
                    my_members = labels2 == k
                          sns.boxplot(y=df[' Assignments'][index],x=X[my_members],color =
color)
        else:
            sns.boxplot(y=df[' Assignments'][index],x=X,color = color)

    plt.xlabel('Peak position (cm$^{-1}$)')
    plt.title(label)
    plt.xticks(np.arange(400,1800+1,50),rotation=-90)
    plt.xlim(400,1800)
    plt.grid(True, linestyle='--')



“
```

B.3 EHD Colour to Thickness

This code was designed to convert EHD RGB values (from an optical image) into thickness (nm values) in ***Chapter 3****.*

Code: *https://github.com/PauloAxcel/EHD-colour-to-thickness*

“

```
from PIL import Image
import numpy as np
from collections import namedtuple, OrderedDict
import pandas as pd
import matplotlib.pyplot as plt
from scipy.optimize import curve_fit
import pylab
import scipy as sp
Polynomial = np.polynomial.Polynomial
from matplotlib.pyplot import figure
import pandas
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
import numpy as np
import cv2
import matplotlib.pyplot as plt


############## CALIBRATION TO GET RGB TO THICKNESS CONVERTION

font = {'family' : 'Arial',
        'size'   : 22}

plt.rc('font', **font)

img = cv2.imread('calibration.png')
cal =  cv2.cvtColor(img, cv2.COLOR_BGR2RGB)
#cal1d = cal[cal.shape[0]//2,:,:]
cal1d = np.mean(cal,axis=0,dtype=int)
#stdcal1d = np.max(np.std(cal,axis=0),axis=1)
stdcal1d = 5
#tabulated values from Erik Schaffer

thickness=[20,70,100,140,200,265,290,310,330,350,400,420,460,520]
Rval = [154,98,43,47,253,176,78,88,168,244,252,158,228,247]
Gval = [86,58,76,117,231,76,166,207,243,247,130,223,207,174]
Bval = [15,48,107,143,119,130,176,175,140,94,154,123,128,133]



def indexation(cal1d,stdcal1d,Rval,Gval,Bval):
    counter = 0
    x = []
    xdim = np.linspace(0,1,len(cal1d))

    for r,g,b in zip(Rval,Gval,Bval):
```

```
        Rmin = r - stdcal1d < cal1d[:,0]
        Rmax = cal1d[:,0] < r+ stdcal1d

        Gmin = g - stdcal1d < cal1d[:,1]
        Gmax = cal1d[:,1] < g + stdcal1d

        Bmin = b - stdcal1d < cal1d[:,2]
        Bmax = cal1d[:,2] < b+stdcal1d


        index = list(Rmax & Rmin & Gmax & Gmin & Bmin & Bmax)

         position = np.mean([index for index, element in enumerate(index) if element ==
True],dtype=int)

        if position<len(cal1d)+1 and position>0:
            x.append(xdim[position])
        else:
            x.append(0)
            counter +=1


    return x,counter



from scipy.optimize import curve_fit

def objective(x, a, b):
        return a * x + b


def rsquared(x, y):
    """ Return R^2 where x and y are array-like."""

    slope, intercept, r_value, p_value, std_err = sp.stats.linregress(x, y)
    return r_value**2


x,counter = indexation(cal1d,stdcal1d,Rval,Gval,Bval)

popt, _ = curve_fit(objective, x, thickness)

a, b = popt

y_line = [objective(xs, a, b) for xs in x]

plt.figure(figsize=(9,9/1.618))
plt.plot(x,thickness,'o')
plt.plot(x,  y_line,  '--',  color='red',label='$R^2$='+str(round(rsquared(x,thickness),2))
+'\n thickness(nm)='+str(round(a,2))+'*RGB'+str(round(b,2)))
plt.legend(loc='best')
plt.show()

#### CASE STUDY TO MAP RGB INTO HEIGHT

img2 = cv2.imread('test_image.png')
test =  cv2.cvtColor(img2, cv2.COLOR_BGR2RGB)
tnew = np.zeros((test.shape[0],test.shape[1]))

Rnew = test[:,:,0]
Gnew = test[:,:,1]
Bnew = test[:,:,2]

cnt = []
filt = 25
```

```
for i in range(Rnew.shape[0]):
    print(i)
    xnew,counter = indexation(cal1d,filt,list(Rnew[i,:]),list(Gnew[i,:]),list(Bnew[i,:]))
    tnew[i,:] = [a*xnew2+b for xnew2 in xnew]
    cnt.append(counter)


fig = plt.figure()
ax = fig.add_subplot(121)
im = ax.imshow(tnew)
fig.colorbar(im)

ax2 = fig.add_subplot(122)
ax2.imshow(test)

xx = np.linspace(0,5,tnew.shape[1])
yy = np.linspace(0,5,tnew.shape[0])

X,Y = np.meshgrid(xx,yy)

Z = tnew

from matplotlib import cm
from matplotlib.ticker import LinearLocator

fig = plt.figure()
ax = plt.axes(projection='3d')
surf = ax.plot_surface(X, Y, -Z, cmap=cm.coolwarm,  linewidth=0, antialiased=False)
ax.view_init(elev=50, azim=-180)
"
```

B.4 EHD Electrostatic Pressure Calculation for Parallel Configuration

This code was designed to calculate the dispersion relationship for the EHD physics in parallel capacitor condition in ***Chapter 3****.*

Code: *https://github.com/PauloAxcel/EHD-dispersion-relation*

“

```
import pandas as pd
import numpy as np
import math

import numpy as np
import pandas as pd
import scipy
import peakutils
import matplotlib.pyplot as plt
import numpy as np

from scipy.signal import find_peaks
from scipy.signal import find_peaks_cwt
import numpy as np
import peakutils
from scipy.signal import lfilter
from scipy.linalg import solveh_banded
from brokenaxes import brokenaxes

font = {'family' : 'Arial',
        'size'   : 22}

plt.rc('font', **font)


gamma=0.03
eta=15000
h=0.0000001
epsilon0=8.85E-12
epsilonp=2.5
A=0.0001
d=0.000001
U=40
delta = 0

space1 = 100
space2 = 100

h = np.linspace(1E-6,2E-6,space1)
q = np.linspace(0,5000000,space2)

freq = np.zeros((space1,space2))

qmax = np.zeros(space1)
qc = np.zeros(space1)
dp = np.zeros(space1)
p = np.zeros(space1)


NUM_COLORS = space1

cmaps = plt.get_cmap('gist_rainbow')

colours = [cmaps(1.*i/NUM_COLORS) for i in range(NUM_COLORS)]

plt.figure(figsize=(9,9/1.618))
for i in range(space1):
```

```
      p[i] =  -0.5*(epsilon0*epsilonp*(epsilonp-1)*U**2)/(epsilonp*(d-delta/2)-(epsilonp-
1)*h[i])**2
      dp[i] =  -(epsilon0*epsilonp*(epsilonp-1)**2*U**2)/(epsilonp*(d-delta/2)-(epsilonp-
1)*h[i])**3
    qmax[i] = np.sqrt(-dp[i]/(2*gamma))
    qc[i] = np.sqrt(-dp[i]/gamma)

    for j in range(space2):
        freq[i,j] = -(h[i]**3/(3*eta))*(gamma*q[j]**4+dp[i]*q[j]**2)


    plt.plot(q,freq[i,:],color=colours[i])
    plt.xlabel('wavenumber ($m^{-1}$)')
    plt.ylabel('wave frequency ($s^{-1}$)')
    plt.xlim(0,5000000)
    plt.ylim(-100,100)
    plt.show()

fig = plt.figure(figsize=(9,9/1.618))
ax = fig.add_subplot(111)

fig2 = plt.figure(figsize=(9,9/1.618))
ax2 = fig2.add_subplot(111)

fig3 = plt.figure(figsize=(9,9/1.618))
ax3 = fig3.add_subplot(111)

for ind, c in enumerate(colours):

    ax.plot(h[ind],p[ind],'o',label='pressure',color=c)
    ax.set_ylabel('electrostatic pressure (Pa)')
    ax.set_xlabel('polymer height (m)')
#    ax.legend(loc='best',frameon=False)


    ax2.plot(h[ind],dp[ind],'o',label='dpressure',color=c)
    ax2.set_ylabel('dpressure (Pa/m)')
    ax2.set_xlabel('polymer height (m)')


    ax3.plot(h[ind],qc[ind],'o',color=c)
    ax3.plot(h[ind],qmax[ind],'o',color=c)
    ax3.set_ylabel('wavenumber ($m^{-1}$)')
    ax3.set_xlabel('polymer height (m)')

ax3.plot(h,qc,label='qc',color='k')
ax3.plot(h,qmax,ls='--',label='qmax',color='k')
ax3.legend(loc='best',frameon=False)
plt.show()

plt.figure(figsize=(9,9/1.618))
plt.contourf(h, q, freq, 1000, cmap='gist_rainbow')
plt.xlabel('polymer height (m)')
plt.ylabel('wavenumber ($m^{-1}$)')
plt.xlim(1.5e-6,2e-6)
plt.ylim(3.1e6,3.6e6)
plt.colorbar(label='wave frequency ($s^{-1}$)')
"
```

B.5 EHD Electrostatic Pressure Calculation for Tilted Configuration

*This code was designed to calculate the dispersion relationship for the EHD physics in tilted capacitor condition in **Chapter 3**.*

Code: *https://github.com/PauloAxcel/EHD-dispersion-relation*

“

```
import pandas as pd
import numpy as np
import math

import numpy as np
import pandas as pd
import scipy
import peakutils
import matplotlib.pyplot as plt
import numpy as np

from scipy.signal import find_peaks
from scipy.signal import find_peaks_cwt
import numpy as np
import peakutils
from scipy.signal import lfilter
from scipy.linalg import solveh_banded
from brokenaxes import brokenaxes

font = {'family' : 'Arial',
        'size'   : 22}

plt.rc('font', **font)


gamma=0.03
eta=15000
h=0.0000001
epsilon0=8.85E-12
epsilonp=2.5
A=0.0001
d=0.000001
U=40
delta = 0.0000001

space1 = 100
space2 = 100
space3 = 9

delta = np.linspace(0,8E-7,space3)
h = np.linspace(1E-6,2E-6,space1)
q = np.linspace(0,5000000,space2)

freq = np.zeros((space1,space2))

qmax = np.zeros(space1)
qc = np.zeros(space1)
dp = np.zeros(space1)
p = np.zeros(space1)

NUM_COLORS = space1

cmaps = plt.get_cmap('gist_rainbow')

colours = [cmaps(1.*i/NUM_COLORS) for i in range(NUM_COLORS)]
```

```
for k in range(space3):
    plt.figure(figsize=(9,9/1.618))
    for i in range(space1):
             p[i]  =  -0.5*(epsilon0*epsilonp*(epsilonp-1)*U**2)/(epsilonp*(d-delta[k]/2)-
(epsilonp-1)*h[i])**2
             dp[i]  =  -(epsilon0*epsilonp*(epsilonp-1)**2*U**2)/(epsilonp*(d-delta[k]/2)-
(epsilonp-1)*h[i])**3
        qmax[i] = np.sqrt(-dp[i]/(2*gamma))
        qc[i] = np.sqrt(-dp[i]/gamma)

        for j in range(space2):
            freq[i,j] = -(h[i]**3/(3*eta))*(gamma*q[j]**4+dp[i]*q[j]**2)


        plt.plot(q,freq[i,:],color=colours[i])
        plt.xlabel('wavenumber ($m^{-1}$)')
        plt.ylabel('wave frequency ($s^{-1}$)')
        plt.title('delta = '+str(delta[k]))
        plt.xlim(0,5000000)
        plt.ylim(-100,100)
        plt.show()

“
```

B.6 EHD 3D Simulations

This code was designed to calculate the dispersion relationship for the EHD physics in tilted capacitor condition in ***Chapter 3****.*

Code: *https://github.com/PauloAxcel/EHD-3D-simlations*

“

```
import numpy as np
import matplotlib.pyplot as plt
import sys

eta = 1.5*10**4
gamma = 30*10**-3
mu = 1.0
A = -1.5e-21
eps_0 = 8.85e-12
eps_1 = 2.5
eps_air = 1.0
mol = 0.001
h_0 = 100e-9
d = 1.0e-6
l_0 = 50.0e-9
V = 40.0
e = 1.6e-19
NA = 6.02e23
temp = 200+273.15
kB = 1.38e-23
k = np.sqrt(2000*e**2*NA*mol/(eps_1*eps_0*kB*temp))

size = 100
dx = 1. / size
dy = 1. / size

T = 9.0
dt = .0001
n = int(T / dt)



import sympy as sym
from sympy import symbols ,diff
from sympy.utilities.lambdify import lambdify
import math
import scipy.integrate as integrate
import scipy.special as special
from scipy import interpolate



A = -eps_0*V**2/(2*d)
B = (eps_1-1)/(eps_1*d)


x, y = np.mgrid[-10*np.pi:10*np.pi:100j, -10*np.pi:10*np.pi:100j]
z = h_0 + h_0*0.1*np.sin(x*y)
z0 = z


from matplotlib.ticker import MaxNLocator
from mpl_toolkits.mplot3d import axes3d
from matplotlib import cm
from matplotlib.ticker import LinearLocator, FormatStrFormatter


step_plot = n//90
```

```
i = 0
#for i in range(n):
while z.max() <= d:


    hc = z

    p = A*B/(1-B*z)**2
    p_h = 2*A*B**2/(1-B*z)**3
    p_hh = 6*A*B**3/(1-B*z)**4

    dhdx = np.gradient(hc , dx)
    d2hdx2 = np.gradient(dhdx[0] , dx)

    d3hdx2dy = np.gradient(d2hdx2[0] , dy)
    d4hdx2dy2 =  np.gradient(d3hdx2dy[0] , dy)

    d3hdx3 = np.gradient(d2hdx2[0] , dx)
    d4hdx4 = np.gradient(d3hdx3[0] , dx)

    dhdy = np.gradient(hc , dy)
    d2hdx2 = np.gradient(dhdx[0] , dx)

    d3hdx2dy = np.gradient(d2hdx2[0] , dy)
    d4hdx2dy2 =  np.gradient(d3hdx2dy[0] , dy)

    d3hdx3 = np.gradient(d2hdx2[0] , dx)
    d4hdx4 = np.gradient(d3hdx3[0] , dx)

    dhdy = np.gradient(hc , dy)
    d2hdy2 = np.gradient(dhdy[0] , dy)

    d3hdy2dx = np.gradient(d2hdy2[0] , dx)
    d4hdy2dx2 =  np.gradient(d3hdy2dx[0] , dx)

    d3hdy3 = np.gradient(d2hdy2[0] , dy)
    d4hdy4 = np.gradient(d3hdy3[0] , dy)



     dh = 1e4*dt/(3*eta)*(3*hc**2*dhdx[1]*(-gamma*(d3hdx3[1] + d3hdy2dx[1])-p_h*dhdx[1])+
hc**3*(-gamma*(d4hdx4[1] + d4hdy2dx2[1]) + p_hh  - p_h * d2hdx2[1])+
         3*hc**2*dhdy[1]*(-gamma*(d3hdy3[1] + d3hdx2dy[1])-p_h * dhdy[1])+
         hc**3*(-gamma*(d4hdy4[1] + d4hdx2dy2[1]) + p_hh - p_h * d2hdy2[1]))
  z = hc - dh
    z[0:5,:] = z0[0:5,:]
    z[-5:,:] = z0[-5:,:]
    z[:,0:5] = z0[:,0:5]
    z[:,-5:] = z0[:,-5:]



    if i%1 == 0:
        levels = MaxNLocator(nbins=15).tick_values(z.min(), z.max())

        fig = plt.figure(figsize=(9,9/1.618))
        ax1 = fig.add_subplot(111, projection="3d")
        plt.title(i)
        # Plot the surface.
        surf = ax1.plot_surface(x, y, z, cmap='nipy_spectral',
                                linewidth=0, antialiased=False)

        # Customize the z axis.
        ax1.set_zlim(0,d )
        ax1.set_xlim(-10*np.pi,10*np.pi)
        ax1.set_ylim(-10*np.pi,10*np.pi)

        # Add a color bar which maps values to colors.
```

```
        fig.colorbar(surf, shrink=0.5, aspect=5)
        ax1.view_init(elev=45, azim=45)
        plt.show()

    i = i + 1

“
```