**Title**
Exploring the Challenges of Advancing Surface Enhanced Raman Spectroscopy-Based Biosensing From Research Laboratory to Clinics



**Author**
Liang, Owen Suyuan

**Publication Date**
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

# Exploring the Challenges of Advancing Surface Enhanced Raman Spectroscopy-Based Biosensing From Research Laboratory to Clinics

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Materials Science and Engineering

by

Owen Suyuan Liang

2020

ABSTRACT OF THE DISSERTATION

Exploring the Challenges of Advancing

Surface Enhanced Raman Spectroscopy-Based Biosensing

From Research Laboratory to Clinics

by

Owen Suyuan Liang

Doctor of Philosophy in Materials Science and Engineering

University of California, Los Angeles, 2020

Professor Ya-Hong Xie, Chair

Healthcare has seen a big boon in technological advancement with the last few years with next-generation genome sequencing, robotics, artificial intelligence, and bioinformatics. While the origins of these breakthroughs are deeply rooted in biology, life sciences, and medical science, the new age of technological breakthrough require an interdisciplinary approach. Materials science itself is an interdisciplinary field, with a huge contribution of physics and chemistry, and within the past few decades, advances in nanotechnology has allowed integration into healthcare giving materials science a new emerging role. Such is the case with a vibrational spectroscopy technique called Surface Enhanced Raman Spectroscopy (SERS).

SERS has the ability to measure the molecular bonds of the target analyte at the single molecular level. Yet, SERS is not a recognized spectrometry tool in healthcare. This dissertation examines the fundamental challenges that require a technique like SERS to transition from the academic research laboratories to clinical applications. While there are broad potential clinical applications, this dissertation focuses on the SERS measurement of diseases originating from cellular morphologies. The two main focal topics are cancer and microbiology research, where the proteome identification can provide new diagnostic and therapeutic treatment options.

The first major section is about testing the capabilities of our group's unique graphene gold nanopyramid hybrid platform in controlled research laboratory environments. Varies cases of cancers are examined: lung, skin, and breast, through their various size, morphology, and proteomic differences. Afterwards, the field of microbiology is explored with bacteria and fungus species.

The second major section involves the SERS's platform performance with clinical patient data. The first case study looks at the diagnosing meningitis from patients in China. The second case study looks at multidrug resistance detection of tuberculosis in rural Pakistan.

The last major section tries to consolidate the major lessons in this journey of bringing SERS to clinical applications. In particular, this section tackles the main question of why a technique like SERS has not been fully recognized in clinics yet and brings the case to why our SERS hybrid platform is very close to it.

The dissertation of Owen Suyuan Liang is approved.

Mark Goorsky

Rajan Kulkarni

Jaime Marian

Ya-Hong Xie, Committee Chair

University of California, Los Angeles

2020

*Dedicated to my wife, family, and close friends*

*that supported me through this long PhD journey.*

*Without everyone's constant support,*

*I would not have been able to cross this finish line.*

# Table of Contents

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

There are far too many people that I am forever indebted to that helped me get through this long PhD journey and without their never-ending support, I doubt I would have finished it; I'll try my best to acknowledge all of you.

On the academic research front, I would first and foremost like to give my sincerest gratitude to my advisor, Professor Ya-Hong Xie for giving me the opportunity to purse my PhD in your lab. I really appreciate your willingness to take me on as your undergraduate researcher all the way back in 2010, and look how far I have come now. I hope to continue to instill your passion of continuously learning, seeking the fundamental truths, and developing a "physical picture" in all my future endeavors; only then can one innovate and discover something new. To my group members, whether you were my graduate mentor, upper classmen, colleague, undergraduate mentee, foreign exchange student, thank you for helping me develop my academic sharpness.

On the familial front, I deeply want to thank my mom and my brother. Your constant emotional support, financial support, and never-ending family support, always gave me a sense of home that I can always return to. That last part always seems secondary until you hit those PhD walls and only by returning to that home was I able to find that motivation to breakthrough. Every time, it has undoubtedly reminded me how cohesive we are as unit to tackle any situation.

On the friends' front, you know who you are. If I could spend the time thanking each and every one of you on the countless efforts in getting me through this, I could probably write another hundred pages. Some notable examples are my close UCLA undergraduate friends, thanks for letting me be me with all my particularities and allowing me to be as snarky as possible. Our countless friend gatherings were definitely needed to help me wind down from the long grueling research hours. To my UCLA

graduate school friends, thanks for hearing me complain about my research. Those conversational stories talking science and going for midnight taco runs are some of the highlights in my 20s; what I would give to do that over again. To my high school friends, why are you reading this? Joking aside, the occasional visit to the hometown and us all meeting up were big stress relievers. Although it was always hard to go full science on you guys, I never diminished the opportunity to try and explain what I do. It's for you hard-working folks that us "eggheads" try to innovate for and it's always humbling to step down from that ivory tower and to break down the science to an understandable level; that I take as my life-long skill.

Finally, the most important person of them all, to my lovely wife, I dedicate all of this to you. Your unwavering supporting, countless encouragement, never-ending patience in expressing your differing viewpoint, the list goes on and on, all of which were essential in me finishing this PhD. Besides all the normal stuff from PhD, the number one thing I'm extremely grateful and proud of is that we were able to meet through it. What an exciting future that we will embark upon together.

# VITA

## EDUCATION

**University of California, Los Angeles** 2015-Present
Ph.D. Candidate in Materials Science and Engineering
**University of California, Los Angeles** 2012-2015
M.S. in Materials Science and Engineering
**University of California, Los Angeles** 2007-2011
B.S. in Materials Science and Engineering

## PUBLICATIONS

**Journal of Raman Spectroscopy: Amyloid β-Protein Quantification** **Co-Author**
DOI: https://doi.org/10.1002/jrs.5785 *May, 2019*
**Protein Science: SERS Distinguishes Amyloid β-Protein Isoforms** **Co-Author**
DOI: *https://doi.org/10.1002/pro.3434* *February, 2018*
**Biosensors & Bioelectronics: SERS of Colorectal Cancer** **1st-Author**
DOI: https://doi.org/10.1016/j.bios.2018.07.038 *July, 2018*
**Small: Sculpting Extreme EM Enhancement for Molecular Sensing** **Co-Author**
DOI: https://doi.org/10.1002/smll.201801146 *July, 2018*
**Cell Stem Cell: Cardiac Fibroblasts To Attenuate Heart Calcification** **Co-Author**
DOI: http://dx.doi.org/10.1016/j.stem.2016.10.005 *February, 2017*
**IEEE Transactions on Electron Devices: Graphene ESD Protection** **Co-Author**
DOI: https://doi.org/10.1109/TED.2016.2582140 *July, 2016*
**Journal of Physical Chemistry C: Single Molecule Manipulation for SERS** **Co-Author**
DOI: https://doi.org/10.1021/acs.jpcc.6b04761 *May, 2016*
**Analytical Chemistry: SERS Detection of Dopamine and Serotonin** **Co-Author**
DOI: https://doi.org/10.1021/acs.analchem.5b01560 *September, 2015*
**Adv Mat: Graphene-Plasmonic Hybrid Platform** **Co-Author**
DOI: http://dx.doi.org/10.1002/adma.201300635 *August, 2013*
**ACS Nano: Plasmonic Response - Graphene** **Co-Author**
DOI: http://dx.doi.org/10.1021/nn301694m *June, 2012*
**Book Chapter on Graphene** **Co-Author**
DOI: 10.5772/15543 *April, 2011*

## Manuscripts

**Owen Liang**, Jing Huang, Shan Huang, Luyue Jiang, Yangyang Liu, Wenjin Yu, Gang Niu, Gang Zhao, Wei Ren, Ya-Hong Xie; "*Neural Network for Aiding Rapid and Accurate Diagnoses of Infectious Central Nervous System Diseases*"

**Owen Liang**, Ya-Hong Xie; "*A Review of SERS on Potential Clinical Applications towards Diagnosing Colorectal Cancer*"

**Owen Liang**, Adnan Ali, Abdul Mateen, Tieyi Li, Shan Huang, Ashraf Fauzia, Zirui Liu, Kahlid Mahmood, Ya-Hong Xie; "*Clinical Applicability of SERS Platform for Effective Antimicrobial Stewardship of Tuberculosis in Rural Pakistan*"

## AWARDS

MSE Departmental Best TA Award *Winter 2013 & Spring 2013*

# 1. Motivation: Need of Alternative Biomedical Technologies

## 1.1.Introduction

Materials Science has pushed the frontiers of technology greatly in recent years with the advent of nanotechnology and Engineering has been able to make some these materials a reality. While the recent decade has seen a major leap in computer science with artificial intelligence, the biomedical field has also seen major breakthroughs as well. This has all been possible with the interdisciplinary combination of technologies that has allowed for novel applications into difficult clinical problems previously thought too difficult to solve. Materials Science has always been a big mix of physics and chemistry and the recent decade has seen more mix of biology applications breathing new possibilities previously unimaginable. One forefront aspect of Materials Science is characterization where various the underlining understandings or disciplines, e.g. surface science, crystallography, elemental analysis, can be applied to analyze any subject. One of the key foundational pillars of my dissertation is the application of materials science characterization to two relevant clinical applications: invasive cancer and drug-resistance bacteria.

## 1.2. Cancer Diagnosis

Cancer is a well-known disease affecting many lives in the US and in the world. In the US, it is the 2$^{nd}$ most prevalent mortality disease behind cardiovascular diseases, and accounts for 21.3% of deaths or 600,000 people in 2017 [1]. Currently, the causes of cancer are not completely understood but many risk factors can increase the likelihood of disease occurrence and/or severity. It is currently understood that a portion of cancer can be prevented with 42% of diagnosed cancers, 740,000 cases in 2019, as potentially avoidable; examples of prevention usually are related to behavioral changes like smoking, excess body weight, physical inactivity, etc [1]. In terms of societal costs, the burden of cancer is

estimated to be $80.2 billion (2015 USD), or adjusted to be $86.5 billion (2020 USD) [2]; this calculation takes into account direct medical costs and the total of all healthcare expenditures [3].

From the various cancer types, the top 5 new incidence cases are breast, lung, prostate, colorectal, and melanoma. Of these, my dissertation looks into breast, lung, and melanoma, where the selection was made because of the cancer prevalence and on the sample availability from our collaborator. While cancer prevention has come a long way and with cancer mortality dropping with better screening, the fundamental technology has not changed that much [3]. To this day, the early screening and diagnosis tools for breast, lung, and melanoma is still image based with mammograms for breast cancer, computational tomography for lung cancer, and by the naked eye from a physician for skin cancer [3]. Breast cancer can be thought as one of the "best" cancers to get as the 5-year survival rate for invasive breast cancer is 90% [3]. However, the early screening of breast cancer is depended mostly only on the mammogram, some high risk patients require an accompanying magnetic resonance imaging (MRI), and the mammogram has a large range of sensitivity from 54% to 81% with an addition risk of extra radiation to promote tumor growth [4], [5]. As such, there is an unmet medical need to improve cancer diagnosis by supplementing current imaging techniques with modern biomarker diagnosis techniques that can be done with genomics and proteomics.

## 1.3.Microbiology Drug Resistance Strains

For infectious diseases, the pathogens that can harm a human body are clinically fit under the umbrella of microbiology. In this field, the four main areas generally focused on in healthcare are bacteria, fungi, viruses, and parasites. Of these, the two fields focused on in this dissertation are bacteria and fungi. The main reason for selecting these two is because these two categories are cell-based and relatively easy to culture. The reason for exploring the field of microbiology is the well-known fear of "superbugs", which scientifically come from antibiotic resistance. This so called drug

resistance becomes a “superbug” when the single drug resistance becomes resistant to multiple antibiotic mechanistic pathways, typically antibiotics work through 5 typical mechanisms [6]. A bacteria’s resistance to one of these mechanisms can be thought of as a strain of the bacteria species. The current gold standard in identifying drug resistant strains is by doing a drug susceptibility test (DST), where a bacteria species is cultured along with the drug tablet to see if its growth is inhibited. However, such a process takes at least a day and typical clinical lab result turnarounds are usually recommended with a timeframe of 3 working days; as such, it is typically in a clinician’s interest to just use a cocktail of antibiotics to treat a patient. Overtime, from such rampant use of these cocktails, mutations develop and multiple drug resistance (MDR) strains become more readily prevalent. Such a scenario happened previously when penicillin was overly used and is now an obsolete drug because of widespread bacterial resistance [7].

These concerns of overly used antibiotics is what the World Health Organization (WHO) calls antimicrobial stewardship [8]. Antimicrobial stewardship is the call for targeted usage of antibiotics to control the rapid spread of MDR strains. The basic tenant is to only use one of the five types of antibiotics in curing a patient and not resort to the typical cocktail blanket usage of antibiotics. This is theoretically possible with DST cultures but is rarely done in most countries, because of the long culture time. As such, there is an unmet medical need for rapid identification of microbial species and possible drug-resistance strains. For current technology in bacteria species identification, there is a gene called the 16s rRNA gene which can be broadly used to identify >90% of bacteria genus and 65-83% of bacteria species [9][10]. However, this genomic technique is not typically used in clinical microbiology laboratories because of its weak antibiotic resistance strain detection and its high reagent cost. Thus current methods are not enough for sufficient antimicrobial stewardship and this is why there are still “superbugs” to this day [11], [12].

## 1.4.Clinical Significance

The motivation of our research is simply in the clinical significance. We believe that by using SERS for cell proteomic identification is a new alternative characterization method to help in the monitoring of a cell's homeostatic and metabolic state. To our knowledge, there are not that many techniques that can measure cells *in vitro* and especially not another technique that can measure the cell's proteome *in vitro*. The clinical significance comes from the application of our substrate to cancer cell differentiation; although at its core we can apply it to any cell. In cancer therapy, cancer cell heterogeneity is a difficult issue to tackle as the changing cellular environment can cause unforeseeable mutations and can lead to unsuccessful treatments. We believe our platform's ability to monitor an individual cell's state in that dynamic environment will shed some light into how a cell transforms itself into being untreatable. The key to all of this is the transmembrane proteins on the surface that maintains the cell's homeostatic state.

## 1.5. Cell Proteome Significance

The human genome project was an ambitious endeavor sought to answer the many mysteries from cancer mutations to human evolution. However, not every question could be answered as genotyping every sequence did not give insight to how certain cells behave. A large number of genes are annotated as hypothetical proteins or proteins with unknown function. This is because the role of each gene may be determined by phenotypic analyses followed by targeted gene disruption, which requires the construction of a deletion vector and transformation. Such a laborious role could not determine a protein's function, since transcriptome analysis does not reveal information of protein localization, quantification, and post-translational modifications; that is where proteomics comes in. [13]

A proteome analysis is a powerful technology for understanding global regulatory networks. Proteome reference maps can be developed for cytoplasmic proteins, membrane proteins,

phosphoproteins, and secreted proteins. These proteome networks will have to response dynamically to various environmental conditions such as nitrogen limitation, heat shock, pH changes etc.[13] Thus, developing a platform to help the study of the cell membrane's proteome is a much needed endeavor.

## 1.6. Focus of Dissertation

The focus of my thesis will be to explore the feasibility of using a SERS proteomic fingerprint for cell identification. The key term is to explore the feasibility and not to prove the feasibility. This is important because proving the feasibility will be an on-going process and stretches beyond the work that I can accomplish at UCLA. The idea of proving the feasibility will require a reasonably statistical average of cell proteome collected and to show that each one can be uniquely identified. My work is the beginning of that step and to handle the difficulties in setting up a protocol for this massive data collection.

# 2. Background: Principle of Current Disease Diagnostics and SERS

## 2.1.Current gold standard in disease diagnostics

For disease diagnostics, there is a multitude of modalities which are beyond the scope of this thesis. For now, the focus is on diagnostics that are relevant to SERS and for those techniques; they can be separated into proteomics and genomics.

### 2.1.1. Proteomics

Proteomics is the comprehensive study of all proteins in a cell, tissue, body fluid or organism. Proteins have many functions ranging from catalyzing the biochemical reaction to monitoring the internal/external environments of a cell. Proteins can differ amongst each cell and base themselves on the cell's main function; in other words, it is based upon how the cell will respond to its developmental and environmental signals. With that knowledge, analysis of the proteins of a cell under a controlled biological environment can provide insight into the activities of the cell or tissue. Proteomics can explain the many different functions of a cell including structure, expression profiling, protein function, and protein-protein interactions. [14]

The next logical question to ask is then: what is the advantage of proteomics compared to genomics? There are several advantages to studying the protein complement. The first is that although analysis of the genomic sequence can often identify the protein coding sequences and pathways, in many cases the function of the protein and the post-translational modifications that influence protein activity and cellular localization cannot be directly predicted from the genome. Thus, it is more straightforward to infer a protein's function by determining the conditions under which the cell can express the protein and when they will activate. In more biological terms, genomics can identify the expression profiles of the protein coding sequences with transcriptomics, however, mRNA levels do not

always correlate with protein levels and thus do not indicate the presence of active proteins. Transcriptomics measure the actively transcribed genes from the mRNA and these processes turn over rapidly thus information is lost. Proteomics on the other hand is to monitor the relatively more stable proteins, giving us unparalleled knowledge. [14]

#### *2.1.1.1. Cell Proteome*

The cell's proteomics can be generally thought of as the whole composition or the whole picture of a cell's profile. This picture will include the cell membrane proteome along with a portion of the inner cell membrane's protein.

#### *2.1.1.2. Western Blot*

Western Blot is another technique typically seen in proteomic analysis too. However, there needs to be a technique done beforehand to separate the protein and typically SDS-PAGE is used in conjuncture with Western Blot. After the proteins are separated, a specific antibody of the target protein is used so that the target protein can be seen in the photograph. The difference between Western Blot and SDS-PAGE is that 1) Western Blot requires a separation of proteins onto a nitrocellulose membrane and then identified with a target antibody; 2) SDS-PAGE is a method based on charge and molecular weight separation and then afterwards put into Western Blot so without a denaturing/separation process, Western Blot cannot be done. [14]

#### *2.1.1.3. 2D Gel Electrophoresis*

One of the current proteomic analysis methods is called Two Dimensional Gel Electrophoresis (2D-GE). This is a technique that uses a special gel to separate proteins by putting them through an immobilized pH gradient in one dimension. An applied bias will separate by the protein's net charge; this is because amino acids have ionizable groups that will contribute to the net charge of the protein. The degree of ionization is influenced by the pH of the solution. Afterwards, electric current is applied

through the gel and then the proteins separate based on their specific pH and this is called the isoelectric point, or when the overall charge of the protein is zero and unable to move. The second dimension of 2D-GE is by separating proteins by their molecular mass. This can be done at right angles to the first dimension through a Sodium Dodecyl Sulfate (SDS) – Polyacrylamide gel. Another name for this technique is called SDS-PAGE. Typically a dye is used with this technique to stain the separated proteins into a photograph; there are typically two kinds of dyes: Coomassie Brilliant Blue and a silver colloid stain. With SDS-PAGE, you can approximately separate 2,000 different proteins and it is typically used in conjuncture with a database of peptides. [14]

#### *2.1.1.4. Mass Spectrometry*

Mass Spectrometry (MS) is a technique used in conjecture with SDS-PAGE. First the SDS-PAGE is used to separate the proteins based on charge and molecular weight. Afterwards, the proteins are denatured into ionized fragments of the molecules; this is done by a digesting the proteins into peptides using a protease, such as trypsin. Next the MS will ionize the peptides and separate them according to their mass-to-charge ratio (m/z). Once the m/z ratios are collected, they are run through a peptide database and compared with for protein identification; this is a process of analysis called peptide mass fingerprinting. [14]

### 2.1.2. Genomics

While the focus of this dissertation does not involve genomics, it is typically the field that proteomics gets compared to and consequentially, SERS as well.

### 2.1.3. Relation to our platform

When comparing these gold standard techniques to our platform, we must be clear that these techniques are complementary to ours, but we will not use their techniques. This is because what we do is fundamentally different than their technique and thus will only confuse the analysis. An example

of this is comparing our technique with the Mass Spectrometry (MS). MS is useful in that it is also a label-free technique, however it will give the peptides of all proteins and thus will give too much information; this will naturally convolute the information making analysis very difficult. Instead, our platform will just look at the transmembrane proteins and in a 1 μm diameter spot size. This small area is all we need to get a proteomic fingerprint which will give the same amount of information as the MS.

Another benefit with our technique is that time and sample preparation is saved compared to the gold standards. Our SERS substrate will only require a couple of droplets of the cell analyte (1-5 μL) and then can be scanned within seconds for a result. There is no need to run the samples through multiple platforms or through other mechanisms for further refinement. The proteins on our cancer surface are still intact and thus our method is a more straightforward choice; this will reduce the amount of error and loss of information, similar to a derivative requiring one dimension loss of information.

While there are many advantages with our platform, it is not without flaws. One large drawback of our technique is the established database for comparison. Since our technique has just started measuring pure standards for comparison, obviously the MS peptide database and the Western Blot's catalog of antibody-to-protein are much more reliable. Of course such a flaw will be offset with more time and data collection. Another flaw with our technique is that currently, the sample analyte needs to be dried onto the SERS substrate. This is because intimate contact is required for SERS to be effective; compared with the current standard which has established denaturing protocols, the current standard has more reliable procedures while ours require more research and time.

## 2.2.Principles of SERS Hybrid Platform

### 2.2.1. Raman Spectroscopy

As we can see from the previous gold standard of proteome techniques, these techniques give too much information and can convolute the analysis. The argument can thus be made that measuring the proteomic fingerprint can give the necessarily information with many benefits and very little negatives. To further understand our technique, one must first explain the Raman effect.

#### *2.2.1.1. Raman Effect*

Raman scattering is an effect named after the scientist Chandrasekhara Venkata Raman in 1921 in India. What he discovered, was an interesting phenomena where scattered photons have a higher energy than the incoming photon after interaction with a sample. Typical optical scattering phenomena, such as florescence, is a down-conversion of energy so when Raman discovered that his photons had actually increased in energy, he knew he had discovered a new optical scattering phenomenon [15].

Light can interact with matter in several ways: absorption, reflection, transmission, and scattering. Both the material and color (or wavelength) of the light can affect this type of interaction and the study of these interactions is called light spectroscopy. By using a Raman spectrometer, we can see that a tiny fraction of the scattered light has a different color because due to the scattering process, the photon has lost or gained energy from the vibrating atoms in the material. In studying the vibrations of the atoms in the material, we can discover useful information, such as the chemical and structural composition, using this technique called Raman spectroscopy. However, the Raman effect is actually inherently weak where only about 1 part in 10 million scattered photons will shift in wavelength. Thus, although Chandrasekhara Raman discovered the phenomena in the 1920s, Raman spectroscopy didn't see an explosion of usage until the invention of lasers and powerful detectors like the Charged Coupled Device (CCD). In general, all materials produce a Raman spectrum, with the exception of pure metals,

thus making Raman spectroscopy a very versatile technique. It is also non-contacting and thus non-destructive, requires little to no sample preparation since it only needs an optical microscope focusing lens, a laser spatial resolution of 1 µm, and can be analyzed in water unlike infrared absorption spectroscopy [15].

A more widely known technique that occurs when light interacts with atoms in molecules and causes them to vibrate is infrared absorption spectroscopy. However different selection rules apply and this is because for the Raman effect to occur, a change in the molecular polarisability is required during the vibration. An example of this difference can be seen in the study of carbon atoms for the diamond structure where Raman spectroscopy is superb at doing while infrared absorption spectroscopy is not [15]. To further understand the Raman effect, the energy levels of a molecule must be explored.

#### *2.2.1.2. Energy Levels in a Molecule*

For dealing with the interaction of light and molecules, the energy levels of the degrees of freedom of the molecule must be determined. These energy levels are either associated with the movement of the elections, the so called electronic energy levels and further differentiated by their spin, or by the movement of the atoms in the molecules called the motional energy states, differentiated by the vibrational, rotational, or translational energy levels. A great way to visualize these energy levels and their various transitions among them is through the Jablonski diagram as seen in Figure 1 [15].

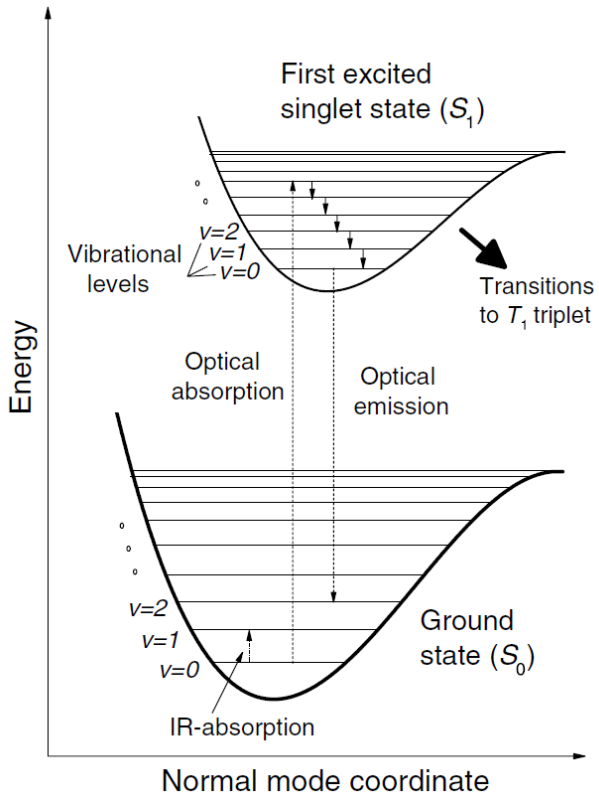



**Figure 1 – The Jablonski diagram that is typically seen in molecular spectroscopy. The dark bold curves represent the electronic energy levels and the light thin flat lines represent the vibrational levels. The possible radiative (dipole-allowed) transitions between states are indicated by the dotted lines, while the possible non-radiative transitions are represented by the solid lines.** [15]

From Figure 1, we can see that molecular electronic states are represented as the dark bold curves where the X-axis is the nuclear atomic coordinates, represented as a single variable called the normal mode coordinate, and the Y-axis is in increasing energy. As such, the minima of each curve represent the equilibrium position of the atoms in the molecule and the vibrational states are represented as the light thin flat lines. The electronic ground state of a molecule ($S_0$) is when the electrons occupy their lowest energy state as allowed by the Pauli exclusion principle. When an electron in a pair is transferred to an excited state, the electrons can have the same spin because they are in different states; thus the electronic excited state is usually four-times degenerate, representing the four

possible spin states. The singlet state ($S_1$) represents the only excited state configuration with a total spin of zero, and the triplet state ($T_1$) represents the triply-degenerate excited state with total spin of one. A radiative transition will involve an interaction with a photon where absorption is a transition to a higher energy level and emission is a transition to a lower energy level. A non-radiative transition will involve an interaction with the environment (solvent, or other molecules) or of internal interactions such as the intra-molecular vibrational redistribution (IVR). IVR is the redistribution of energy into the lower-energy vibrational or rotational states with a typical timescale of ~$10^{-12}$ seconds and can occur between the singlet and triplet states (inter-system crossing), unlike a radiative transition [15].

A simplified Jablonski diagram is typically used to represent Raman scattering as seen in Figure 2. For a scattered photon, there are two main groups this can be classified: elastic scattering and inelastic scattering. Elastic scattering is when the incident photon has exactly the same energy as the emitted photon, but typically of a different direction and/or polarization. In molecules, this process is often referred as the Rayleigh scattering where the molecule is in the same energy level after the scattering. This means that there is no transfer of energy between the molecule and the photon and therefore does not reveal much information about the internal structure. Inelastic scattering is when the incident photon ($E_L$) has a different energy as the emitted photon ($E_S$) and this energy difference correlates to a transition between two states in the molecule. The Raman scattering process falls into the category of the inelastic scatting. The important difference between Raman scattering and fluorescence is that the processes are instantaneous for Raman scattering while fluorescence is essentially a two step process, excitation and emission with a finite lifetime. This is unique for Raman because this means that the emitted Raman photon can happen without a direct absorption of the photon and thus there does not need to be an electronic transition within the molecule at that incident wavelength. Of course, such a

process is intrinsically a weak phenomenon compared to normal optical processes like absorption or fluorescence and thus requires the special equipment mentioned earlier [15].

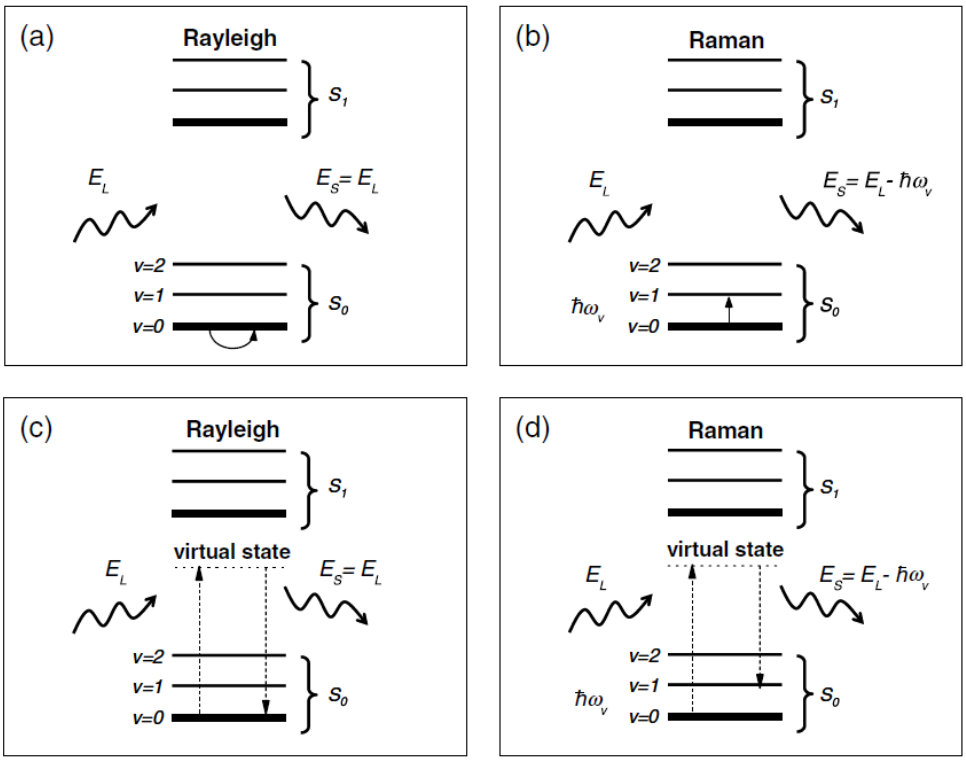



**Figure 2 – Simplified Jablonski diagram illustrating the Rayleigh (a) & (c) and Raman (b) & (d) scattering processes. A normal absorption to the electronic state and then emission can be seen in (a) and (b). However, the incident photon energy, $E_L$, does not need to be tuned to a specific transition in the electronic structure of the molecule and can be seen in (c) and (d). This type of scattering is viewed as two simultaneous processes where an absorption of a photon to transition to a virtual state and then a following emission to the ground state ($S_0$).** [15]

If the scattered photon has less energy than the incident photon ($E_S < E_L$), then this is called the Stokes process and represents the molecule excited to the first vibrational state with energy ($E_L$-$E_S$). However, if the scattered photon has energy larger than the incident photon ($E_S > E_L$), then this is called the anti-Stokes process and represents the molecule relaxing from an already excited vibrational state to its ground state. These two processes are typically plotted in an axis called the Raman shift (units cm$^{-}$

[1]) and is defined as $\Delta E_R = E_L - E_S$ which represents the energy lost by the photons during the scattering event. The graph will be plotted positive for a Stokes process and negative for an anti-Stokes process with the Y-axis labeled intensity, representing the number of photons hitting the detector and thus is in arbitrary units. An example of a Raman spectrum can be seen in Figure 3. For my research thesis regarding cell proteome however, the Raman signal is too weak and cannot be done with just normal Raman; thus Surface-Enhanced Raman Spectroscopy (SERS) is used to enhance the sensitivity of the platform and will be explain briefly in the next section [15].

### 2.2.2. Surface-Enhanced Raman Spectroscopy

Surface-Enhanced Raman Spectroscopy (SERS) uses the same spectrometer as regular Raman but involves some sort of a nano-feature metallic substrate to boost the signal of the Raman scattering from the molecules close to the substrate. This type of boosting can give enhancement factors as large as ~$10^9$ and allows for very low concentrations of materials, even single molecules as soon in Figure 3. Of course since SERS is still an active research field, the complexity of it cannot be fully described in this prospectus; thus, a basic background of the plasmon resonance and the enhancement factors are presented to give the reader a better understanding of our platform [15].

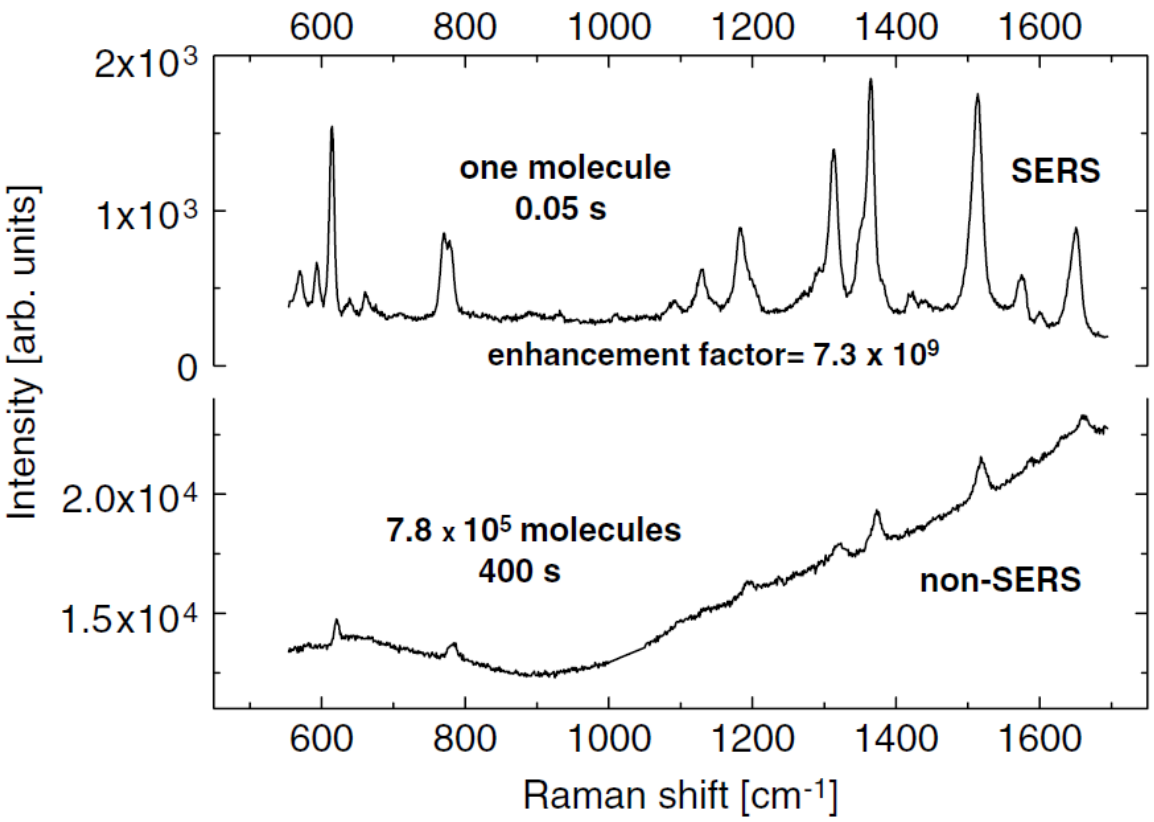



**Figure 3 – Raman spectrum and SERS spectrum of rhodamine 6G molecules, a commonly used dye. We can see the normal Raman spectrum on the bottom done with a long exposure time and with a large concentration of molecules. This is compared with the SERS spectrum on top done with only one molecule and for less than a second.** [15]

#### *2.2.2.1. Plasmonics*

Plasmonics is a relatively new term encompassing all areas of research and technology concerning the study, fabrication, and application of plasmon-supporting structures. Plasmonics and SERS are two areas of research that strongly overlap with each other since plasmons are at the core of the SERS electromagnetic effects and/or enhancements. To define a plasmon, we must first define plasma: the free electrons of a metal that move in the background of the fixed positive ions to ensure overall neutrality. Thus a plasmon is a quantum quasi-particle representing the elementary excitations, or modes, of the charge density oscillations in a plasma. A simple analogy can be described as this: a

plasmon is to the plasma charge density as photons are to the electromagnetic field. However, a plasmon is a quasi-particle because it is always lossy and highly interacting. Or in other words, the charge density oscillations must be maintained by an external source of energy, otherwise it will always decay due to various loss mechanisms, such as collisions. [15]

To relate the above definition of plasmons to our SERS platform, we must also introduce the term polaritons. Polaritons are modes where the electromagnetic wave excites the internal degrees of freedom inside a dielectric medium. In other words, when a photon couples with the internal degrees of freedom in a medium, it is no longer a quantum particle and its energy is shared between itself and the medium. Thus for a metal with a free-electron plasma and with an outside interaction of light, we get a plasmon-polariton. Along with this, one must also know that the optical response of a metal and its plasmonic effects from SERS can be described by its relative dielectric function, $\epsilon(\omega)$. Coupling all these terms into something more qualitative and practical are the electromagnetic and chemical enhancement factors. [15]

#### *2.2.2.2. Enhancement Factors*

By default, normal Raman spectroscopy has the average Raman intensity of a molecule directly proportional to the laser power density and to the Raman cross-section of the molecule (the portion of the molecule illuminated by the laser). This can also be generalized to SERS in that a given vibrational mode is also directly proportional to the laser power density and its normal Raman cross-section, but now also affected by an enhancement factor. The two main enhancement factors are the electromagnetic and chemical enhancement factor. [15]

##### 2.2.2.2.1. Electromagnetic Enhancement

The electromagnetic enhancement factor is due to the coupling of the incident and Raman electromagnetic fields with the SERS substrate. The electromagnetic enhancement relies on the

localized surface plasmons that give large local field enhancements extremely close to the metallic surface. The typical proximity distance from the SERS surface to the analyte is around 10 nm. In most cases to profit from this, the target molecule must actually be directly adsorbed onto the surface through physisorption or chemisorptions. [15] The concentration of the electromagnetic fields is typically coined the hotspot; for our SERS gold nano-pyramid platform, our range can go up to 100 nm along the faces of the pyramid, with the tip being the standard 10 nm (see Figure 4).

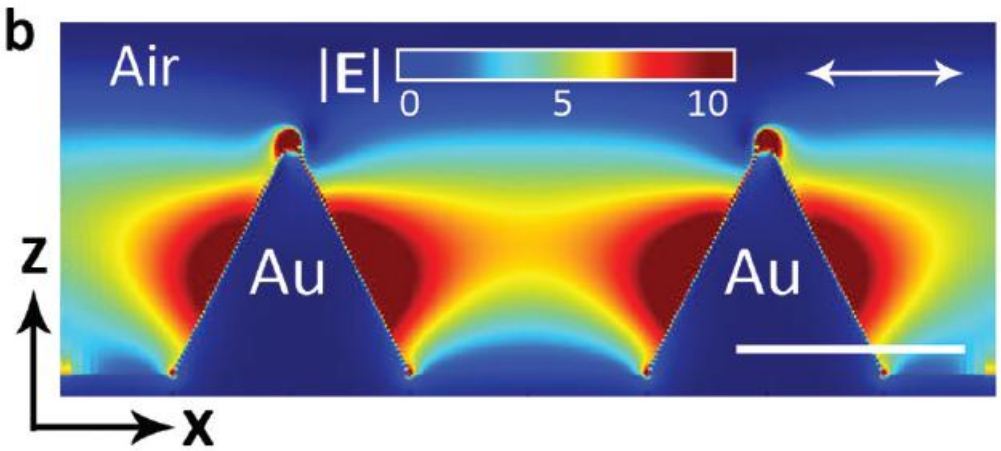



**Figure 4 – Our group's SERS gold nano-pyramid platform under FDTD simulation of the electric field distribution. The scale bar is 200 µm, the white arrows on the top right are the polarization direction of the incident light, and the incident light wavelength is 633 nm. As we can see, the large electric field concentration is on the pyramid faces and on the tip of the pyramid with 50-100 nm and 10 nm respectively.**[16]

#### 2.2.2.2.2. Chemical Enhancement

The other main enhancement for SERS is the chemical enhancement and is actually still a subject of debate to its definition. The technical definition is any modification of the Raman polarizability tensor upon adsorption of the molecule onto the metallic surface, meaning this definition covers quenching and enhancement of the signal. However, most researchers bring up the chemical enhancement when the situation gives rise to a more resonant modified Raman polarizability. The most

studied mechanism at the moment is called the charge-transfer mechanism. This effect occurs when the adsorbate does not bind covalently to the metal and the metal causes a perturbation in the electronic structure of the analyte, which thus causes a mild change in its electronic distribution. This mild change can be thought of as a corresponding change of the polarizability and thus ultimately the change in the Raman efficiency of the mode. In more lament terms, the SERS surface material can have an affinity with the target analyte affecting the Raman signal [15]. For our platform, we have previously proved [16] that the addition of graphene has improved the Raman yield by one order of magnitude and we attribute this to the chemical enhancement as have others [17].

## 2.3.Hybrid Platform

#### *2.3.1.1. Graphene*

Our SERS platform will not be complete without mentioning one of the key components: graphene. Graphene is a single atomic layer of carbon arranged in a hexagonal honeycomb shape pattern. It was discovered in the late 2000s as the first experimental proof of a stable 2D material and with its unique properties, became a hotbed for research topics. Some of these properties include its material tensile strength, chemical inertness, and excellent thermal and electrical properties. In our hybrid platform, graphene rests on top of the gold nanofeatures.

##### 2.3.1.1.1. Purpose

There are five main reasons for the usage of graphene. 1) Graphene is used as our calibrator for the plasmon resonance. What this means is that we use the graphene's Raman G-peak as a marker to determine the intensity of the SERS signal at the hotspot. This can be continuously checked among all SERS substrates of the same type and this gives us repeatable results. 2) A monolayer thick of graphene also allows us to do some kind of quantitative SERS measurement. This is because other groups use SERS markers with dye molecules and their concentration is unknown on the surface; compared to our

graphene which at its worst quality can be 2 layers, this will only introduce a known error of 2x the counts in the G-peak. This is a quantifiable number compared to the uncertainty in dye molecule SERS measurements. 3) Graphene is chemically inert and allows for our substrate to be biocompatible. The is because graphene is a natural diffusion barrier and thus can isolate the analyte from the metal. This will open the possibility of using other metals beside gold, such as silver which has a better quality factor for better SERS resonance in visible light. 4) Since graphene can be used as a built-in gauge, we can also determine the spatial location of each hotspot within the error factor of the Raman spectrometer's stage. The current Renishaw spectrometer has a stage that can move in 100 nm increments and thus will give us error to half the pitch size of the current gold nano-pyramid setup (of course pitch size can be altered during the fabrication process). 5) Lastly, graphene has a chemical enhancement and this was proven by our group. We did a SERS measurement with and without graphene and found that biomolecules are attracted to graphene and give an order of magnitude in enhancement. [18]

### 2.3.2. SERS substrate fabrication

The SERS substrate fabrication was headed by a previous group member, Ming Xia, and he did a lot to finalize a repeatable procedure. This prospectus will not go into the details of the SERS substrate fabrication, since my thesis is on the application of the SERS substrate, and will only highlight the main parts. More details on the fabrication process can be found in our published work. [19]

The SERS substrate fabrication currently consists of the following feature shapes: nano-pyramids, nano-triangles, and nano-pillars; these features can also be made out of silver and gold. Figure 5 displays a schematic diagram of the hybrid graphene-gold nano-pyramid substrate. For this prospectus, only the gold nano-pyramid fabrication will be covered.

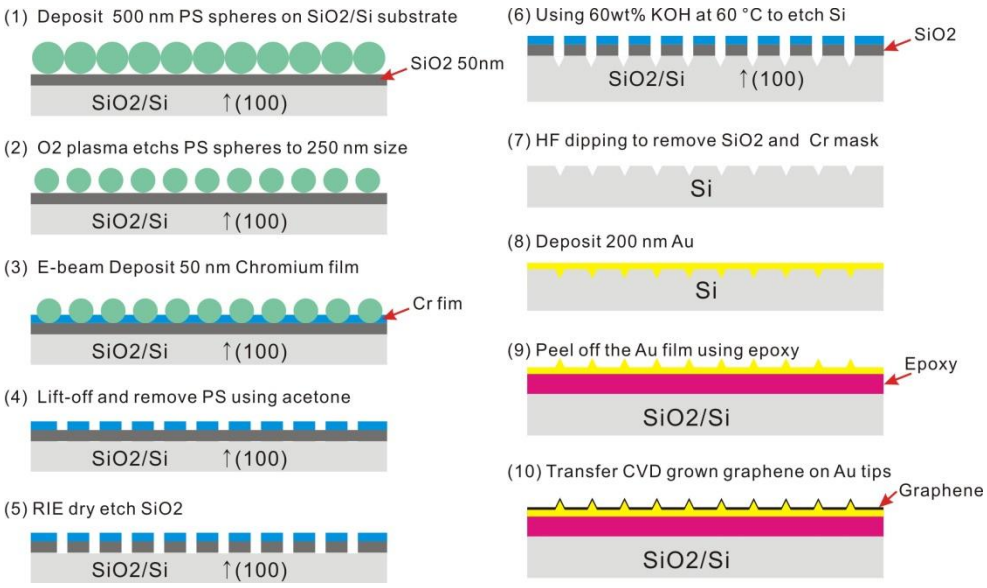



**Figure 5 – Fabrication Diagram of the graphene/gold nano-pyramid**

First, the target $SiO_2$/Si wafer is pre-cleaned in a Piranaha and RCA solution to ensure very good hydrophilicity. After the pre-clean step, polystyrene (PS) nano-spheres are deposited onto the surface via a drip jig. The PS nano-spheres will provide a mask for a dummy metal deposition and this metallic mask will be defining the nano-pattern feature for KOH etching. The KOH is the key to forming the Si mold pits. Once those pits are made (see Figure 6), gold is deposited on top and a simple glue like epoxy can be used to detach the gold film from the mold. Copper grown Chemical Vapor Deposition (CVD) graphene is transferred onto this substrate and this is referred to as our hybrid SERS platform.

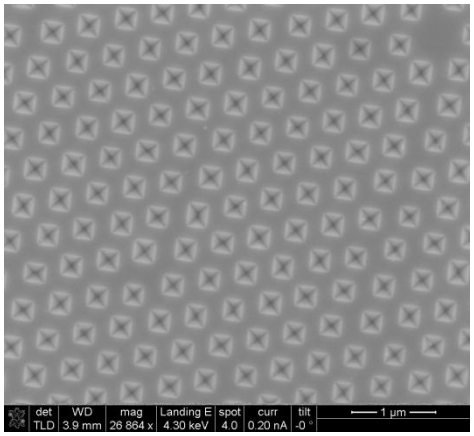


Figure 6 – SEM image of the Si pyramidal mold before gold sputtering. We can see the periodicity of the nano-features due to the self-assembly of the PS nano-sphere mask with very little defects. The scale bar is 1 µm.

#### *2.3.2.1. Raman Measurement Procedure*

Raman is typically done on the Materials Science & Engineering Department's Renishaw Raman spectrometer located on the first floor of Engineering V. Typically the 785 nm laser is used along with a 50x Leica LWD (long working distance) lens. The machine is usually calibrated to 10% of its max power (20 mW) and usually exposures run from 1-10 seconds. Mappings are now normally done in 10 µm x 10µm with 1 µm spacing, but there were variances in previous data collections; typically mappings are structured to the purpose of the experiment. A silicon reference wafer is used to calibrate the machine before every use.

#### *2.3.2.2. Raman Data Post-Processing*

Raman post-data processing is done with the Wire 4.2 software provided by Renishaw. This is a powerful software that allows us to subtract the baseline spectrum, eliminate cosmic ray peaks, run a noise filter software, and curve fit peaks if necessary. After the data is post-processed, it is converted to individual text files and then run through our in-lab's PCA software. This software is comprised of a custom python code for vector input and then ran through R-studio, where the language R is more typically used by statistical students. The R-studio program will run the converted python text files and form a PCA graph.

## 2.4. Multivariate Analysis

### 2.4.1. Principle Component Analysis

With hundreds of SERS spectra being taken per Raman mapping, and multiple Raman mapping done on each cell, the amount of Raman spectra can easily run over the 1000s. As such a computer aided tool is necessarily for data processing. The reason we use Principle Component Analysis (PCA) for

our data analysis is because it is typically used as a standard in data processing in biology. Although this is our initial reason, we are currently working on developing more advanced machine learning programs to further aid in the data analysis; but that is beyond the scope of this dissertation.

PCA works by being a dimension reduction tool. This is essentially a useful technique to help humans interpret data in a more visual plane. In other words, instead of taking into account 70+ variables from the dataset, PCA can reduce that number of variables down to 2 and this will make it very quick to interpret the result. PCA works in a similar fashion as a change of coordinates; i.e. from Cartesian to spherical. PCA looks for patterns within a dataset and separates them by their variance, or relationship to one another. After finding the most variant data points, they are plotted on the PC1 axis and the second most variant axis is plotted orthogonal to PC1. Thus, the processed data will cluster themselves into groups and this will allow us for easy identification; i.e. cleaning the left cluster as cancer and right cluster as normal healthy cells.

## 2.5. Machine Learning

### 2.5.1. Supervised Learning vs Unsupervised Learning Algorithms

Two machine learning data analysis tools are used in this study: DNN and PCA, and both are chosen for their particular purposes. The two important terms to understand are supervised learning and unsupervised learning. Supervised learning is a class of tools that require a key or answer to be fed into the program so that the algorithms can provide feedback and mold the data to these labels. In DNN, the program requires a large identified sample dataset to form its neural model. Using this neural model, the program will then sample a test dataset and return the True Positives (TPs), False Positives (FPs), True Negative (TNs), and False Negatives (FNs). These values can then be used to calculate the sensitivity and specificity of separating monocyte and lymphocyte. The other class of tools is unsupervised learning and as the name suggests, it does not require an identified sample dataset to

form its conclusion. Instead, PCA is purely a mathematical transformation of coordinates systems, similar to the analogy of Cartesian to Polar, but transforms the data based on the value of variance. The end result is that the original data is viewed from another point of view for easier interpretation of results. Both techniques will not be explained further as well as the benefits and disadvantages of each class.

While there are many examples of supervised learning, such as Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), the method used in this study is DNN. The main reason for selecting DNN over the other methods is the lack of a clear understanding in the relationship between a Raman spectrum and the cell species identification. In other words, it is not known if a linear separator, hyperplane classifier, or majority voting of nearest entities is the best method for clearly identifying the Raman fingerprint. As such, DNN is chosen to factor in these possibilities and form what it thinks is the best classifier model on its own, without inherent bias from the user. The DNN model is formed on the basis of minimizing the loss function, the metric used to characterize the effectiveness of the model. A back propagation function is included which updates the neural model for each correct and incorrect input, along with a hyperparameter, effectively the learning rate of the program and modifies the percentage change of the neural model for each input.

For unsupervised machine learning, the method selected in this study was PCA. While there are also other unsupervised methods, such as hierarchal clustering or k-means clustering, PCA was chosen because it belongs to the blind signal separation category and this means that PCA requires the least amount of information for data separation when compared with other methods. It is based on this that the assumption can be made that PCA is the check on the data, without any inherent data bias, to check the result of the DNN. One of the biggest disadvantages of DNN is the requirement of an already labeled standard that the program can build its neural model and also a very large dataset to confirm

the model. The obvious flaw in this approach is when an imperfect standard is given to the DNN and the neural model detracts from the true answer. The other flaw of a large dataset is that while it is necessary for building a good model, it may introduce the issue of over-modeling; in this case, an over-modeled neural model cannot handle slight differences in the data and immediately skews its answers as all right or wrong. In this study, over-modeling occurred for 4% of evaluation rounds and to combat this issue, averaging is done.

For PCA, over-modeling does not occur as it can be thought up as a simple orthogonal transformation. The metric used for the separation of data in PCA space is variance, or the square of the standard deviation. This allows the user to compare data based upon differences and similarities. Applying this to the Raman spectrum, the key Raman peaks selected are used for the whole dataset, and the intensity values are compared and variance calculated. The eigenvalues and eigenvectors are calculated from a matrix of these variances and these two values form the Principal Components (PCs) of the original data. Thus, a PC represents a linear combination of weighted values based on the important Raman peaks selected. For ease of data interpretation, the two highest eigenvalue PCs, or the two PCs with the highest variance, are plotted into a 2D graph with a 68% Normal Probability oval for each dataset. The sensitivities can be calculated from the amount of data nodes that fall into the correct confidence oval.

### 2.5.2. Linear Discriminant Analysis

In regards to clinical applicability, disease diagnosis is heavily dependent on the patient's biovariability. This variability is based on the inherent patient's characteristics and there is always a question to whether the uniqueness of a SERS fingerprint is greater than the patient's biovariability. The SERS nature of the biovariability can be attributed to the various intricacies of proteins, e.g. size, molecular composition, SERS polarizability, etc. In the SERS and PCA-LDA disase diagnosis scheme,

hundreds of SERS spectra are measured per patient or many patients per disease category. Therefore the PCA representation can be depicted by ovals encompassing each dataset and the natural spread of the oval area can be thought as the inherent biological variability, as seen in Figure 7c. To the best of our knowledge, biovariability is not an issue thus we believe the variations within SERS is greater than the biovariability found in patients, making clinical diagnosis possible [20]. Such a phenomenon will also mean this biovariability saturates in a line graph as seen in Figure 7d. Doing such a study will give the necessary information of the minimum number of measurements needed per patient to make an effective diagnosis.

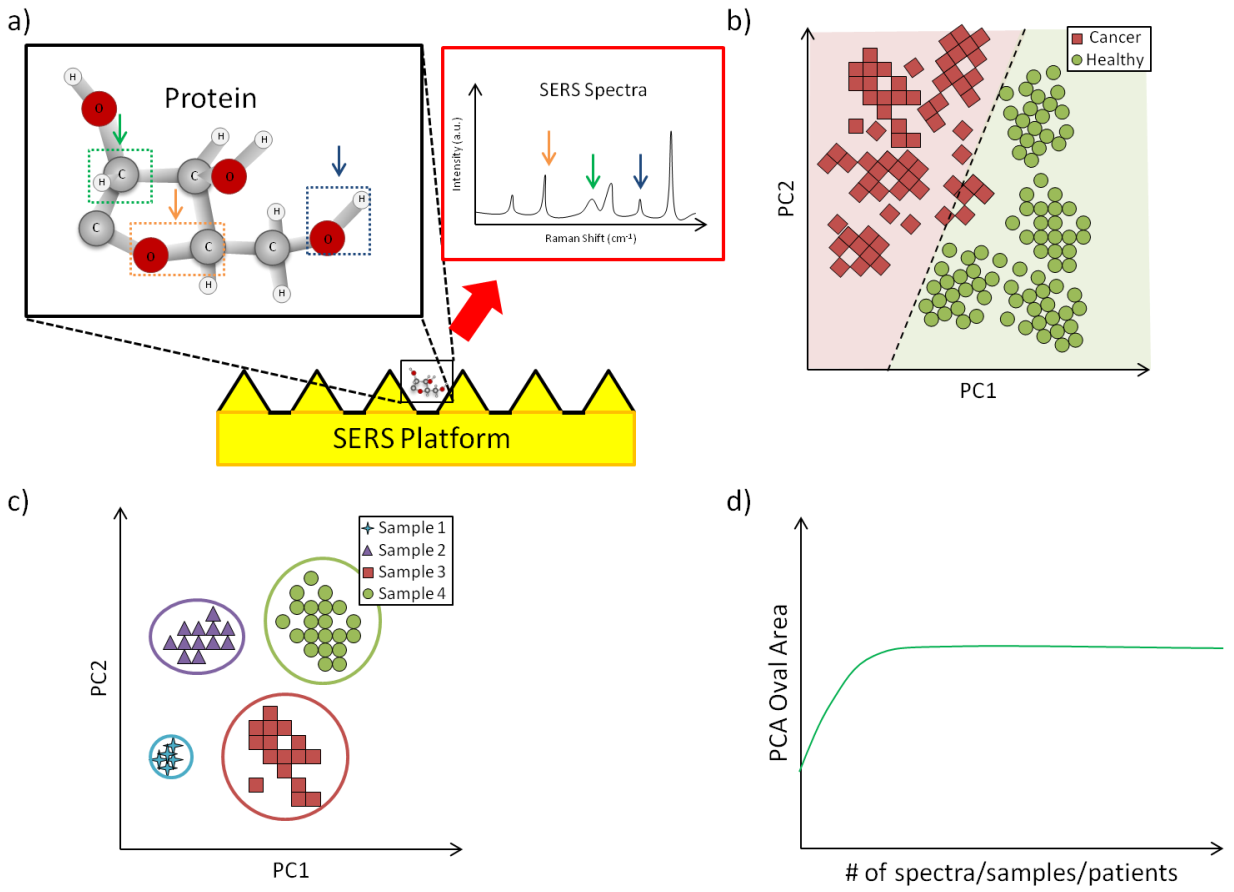



**Figure 7 – Schematic of the SERS measurement and analysis. a) SERS measurement is done on a protein molecule with specific bonds of its amino acids, as shown in the insert black box. Examples of those bonds are represented by their respective colors: green C-H, orange C-O, and blue O-H, and each of these bonds will have a unique SERS peak, as shown on the insert red box representing a possible SERS spectra. b) will show the subsequent analysis done through PCA-LDA. The**

**complex SERS spectra are re-represented into the PCA space by their respective cancer and healthy datasets. A classifier algorithm such as LDA is performed on the PCA to give their respective classifying labels, in this case red is cancer and green is healthy. Sensitivity and specificity can be calculated from the data points that fall into the correct/incorrect categories; in this example, the cancerous red square is misclassified into the healthy green region, termed a false negative, therefore the sensitivity is less than 100% as opposed to the 100% specificity. c) PCA shows the inherent variance amongst its data and if an oval encompasses each dataset, the area of that oval can represent the biovariability. This inherent biovariability will saturate if enough spectra are taken per sample, or enough patients are measured per disease. d) a line graph can be used to represent such a phenomena where saturation is the biovariability of a target disease and also explains the minimum amount of patients necessary for a study.**

### 2.5.3. Deep Neural Network

The DNN model employed for this study is based on an object detection image-based neural network built on Tensorflow and pre-trained on the COCO dataset [21]. The basics of a neural network can be considered as a repeating algorithm that classifies the importance of an input based on an activation function. An activation function is similar to the action potential of a human neuron cell, where a necessary stimulus causes the firing of the neuron and this is an all or nothing process. This is analogous to artificial neural networks where the activation function is a mathematical threshold value and once that is met, the result is like the firing of a human neuron. There are additional nuances to this mathematical equation with coupling of weights and bias values, and the resultant firing is not a step function, but a specialized mathematical function containing in-between 0 and 1 activation values; an example is the sigmoid function. However, the main concepts translate to the idea that only the important characteristics of an image will be filtered through this activation function with each of these characteristics being represented as a neuron in one layer of the neural network. The addition of multiple layers gives rise to the non-linearly of a DNN and these features allow a DNN to recognize an image, similar to mimicking the image processing of a human brain. Coupled with the introduction of convolution neural networks (CNN), the processing requirement for image-based neural networks dropped significantly, paving the way for large advancements in the field [22]. However, the detailed description and workings of each of these improvements are beyond the scope of this study and a sample of this literature can be found in Ref. [23]–[25].

### 2.5.4. Decision Trees

Another form of supervised machine learning (ML) is called decision tree and in combination with boosting is our group's current best answer for SERS ML-based data analysis. In the decision tree model, one can get a classification result similar to the DNN model shown above, but the process is fundamentally different. Where the DNN model builds upon the forward bias and each neuron can be thought up as an important spectral feature, the black box nature of a DNN does not elude to inform the data scientist a clear picture of what the ML model is doing. It is from that high-level thinking that decision tree can help provide that ML insight were each tree can be back-tracked and provide a clear indication of each SERS fingerprint's uniqueness in each spectra.

The basic format of a decision tree is shown in Figure 8 where a decision tree is composed of three parts: 1) the main node, 2) the branches, and 3) the sub nodes. The simplest way to understand the decision tree format is to image each main node as a set of if-then statements typically found in computer programming. As one can therefore imagine, a large set of if-then statements will form a case-structure and the conditions of that is the high level framing of a decision tree. For each main node, the if-then statement is trying to reach a goal, which is a target threshold value and that is set by the user. These threshold values are arbitrary but the goal is to represent key features in SERS spectra, therefore in a sense, each sub node is a SERS peak of great importance. The branches in this tree represent the topological structure and unlike DNN, does not have a weight value. The program follows along each main node, the if-then statements, satisfying each branch's condition and repeats itself until arriving at the target threshold values represented in each sub node. Thus as one can imagine, in a perfect decision tree classifier, enough sub nodes are selected with enough key main nodes to present a full SERS spectrum [26].

The following is an example of a SERS dataset broken down to fit an example clinical application. For a clinical report, the main output is the binary: patient or healthy. The very first node is the first if-then statement splitting this decision. If we follow the patient branching line, each subsequent main node will represent important threshold requirements based upon the SERS intensity. The number of layers of a decision tree is part of the art and will finally represent the "model" to classify the label patient. The very end sub nodes are the SERS features that fit this classifier label and can be thought as of as the unique key biomarker SERS peaks.

The obvious follow-up question of how does one construct the "perfect" decision tree is the foundation of the machine learning algorithm. If the perfect decision tree is known, then there is no need for the ML portion, but in real-world scenarios, this is part of the unknown and the ML portion defines a regressive learning portion to optimize the best decision tree structure. To determine the structure of a decision tree, the ML algorithm follows a similar "greedy strategy" of minimizing a partition criterion called a Gini coefficient. Details of this is beyond the scope of the thesis but can be found by literature [26] but the essential part is a minimization of the mean square error. Therefore, the ML algorithm repeats itself to minimize this error, based upon the user set number of categories and number of samples.

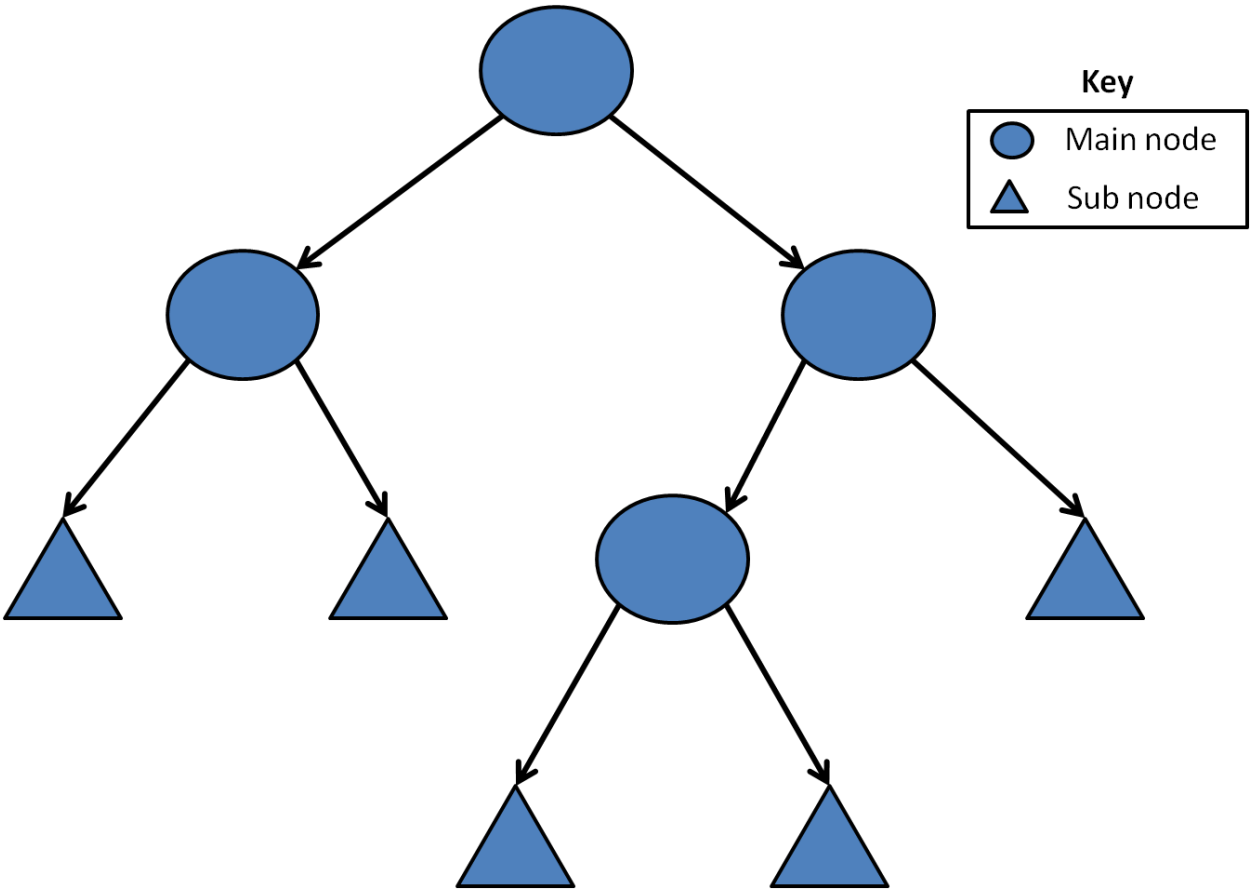



Figure 8 – Decision Tree layout with circles representing the main nodes and the triangles representing the sub nodes.

### 2.5.5. Boosting

The second part in optimizing the decision tree ML is combining it with boosting. Boosting is a common ensemble-based method which is a fancier way of saying it combines various models to determine a more accurate classifier. For our case, this is achieved by boosting the various decision trees of each SERS spectra to give an overall database evaluation of labeling each dataset a patient or healthy control. The general theory of boosting is similar to the DNN model where a learner determines which decision tree is the best based on their performance and through iterations, weights are assigned to each learner to give greater influence for the final ensemble's final prediction. The most popular boosting algorithm is called AdaBoost, short for adaptive boosting, and this is the final form our ML

platform uses. The general idea of Adaboost is to optimizing the boosting method by focusing on the weakest classifiers or the outliers that are the most difficult to classify by the algorithm. In the mathematical sense, normal boosting applies weights to learners giving a preference in the final prediction but this will only cover a majority of the dataset and outlier data gets lost in the average weighting. Adaboost solves this focusing on these misclassified datasets and giving them more weight in subsequent iterations. Thus, and adaboost function will try to lower its overall error rate with this additional constraint and give an overall better accurate classifier [27].

Coupling the adaboost with decision trees is our research group's current solution at supervised ML. Following the same logic as before, the ensemble of decision trees are optimized by boosting where a learner is determining the importance of each decision tree for the final prediction of patient or healthy control. As is the case with most clinical applications, outliers are the usual focus and thusly the adaboost philosophy fits rather well to this situation. Where misclassified results may typically lower the weight value of other algorithms, adaboost inverts this and adds these weak classifiers to its overall error loss function. The goal of the adaboost algorithm is to minimize this loss function therefore helping to select the best decision trees that represent both the majority features and the features that can handle the outlier datasets[26].

# 3. Initial Evidence in Controlled Environments: Laboratory Setting

## 3.1.SERS in Differentiating Biological Entities

The decision to select the cancer cell lines as our first cell type to identify is as follows. The first reason is the biomedical importance, where in 2015 it is estimated that there were 1.6 million new cancer cases and a reported 590,000 deaths.[28] From this we can see why cancer research is being heavily researched upon and also why there are many established cancer cell lines. Human cancer cell lines are essentially tumor cells that have been removed from patients and then genetically modified to be immortal so they can be mass-produced. These immortal or continuous cell lines can divide indefinitely if given the correct cell culture conditions[29]; this make makes these continuous cell lines extremely useful in research.

The systematic approach in understanding SERS application to clinics is to test the platform in the best case scenario of laboratory settings. Human cancer cell lines represent these controlled environments by providing the key proteome signature for each cancer disease without the introduction of patient biovariability. As noted in the background section, the SERS protocol stays the same with the only differentiating factor being the different cell lines measured. The general protocol is to grow each cancer cell line, re-suspend them into liquid format via PBS, perform the SERS measurement, and then analyze the data with machine learning. The main cancer diseases examined in this thesis are lung, skin, and breast cancer and for microbiology, E. *coli* is the main bacteria cell line along with cultured fungus species isolated from patient CSF.

### 3.1.1. Human Cancer Cell Lines

#### *3.1.1.1. Lung Cancer*

Lung cancer is the most prevalent cancer in terms of incidence and the largest cause of cancer deaths for both men and women today. The current 5-year survival rate for late-stage diagnosis is 5% as opposed to 50% if the patient is screened early [3]. The current diagnostic tool is low-dose CT scan and while it has lowered mortality by 20% in the past decade it is not perfect.

The overall theme of looking into lung cancer is to see if SERS proteome identification can lead to new diagnosis not previously known and our first attempt at a second cancer disease following the success of the previous study involving colon cancer [30]. One major difference from that study to this one is the free flowing of cancer cells in the liquid solution and allowing the cells to dry onto the substrate. The previous study also involved colon cancer cells from a cell line, but those cells were cultured onto the substrate and then measured. For these particular experiements, cells were released from their cultured dish with trypsin, an enzyme that attacks the adhesion proteins of a cultured cell to release it into the media, and the free float cells are resuspended into PBS via a regular centrifuge. The purpose is to see the differentiability of the remaining cell membrane proteome when measured by SERS and if that is enough to identify the different between healthy and lung cancer cells. To simulate a healthy control, a T-cell culture line is selected, and two different lung cancer cell lines were used: A549 and H157. One unique feature of these various cell lines is that with an optical microscope (OM), a technician can separate these two cell lines if needed.

##### 3.1.1.1.1. Preliminary Data

For this dataset, the goal is to check if we can even differentiate a cancer cell from a healthy cell. For the lung cancer cell of this variant (A549), a T-cell is compared to as the healthy cell since the lung cancer disease is known to circulate in the blood stream. So our comparison is to simulate a blood vial

sample with every component removed except for the lung cancer cell and the T-cell; this is of course not practically realistic but a good starting point.

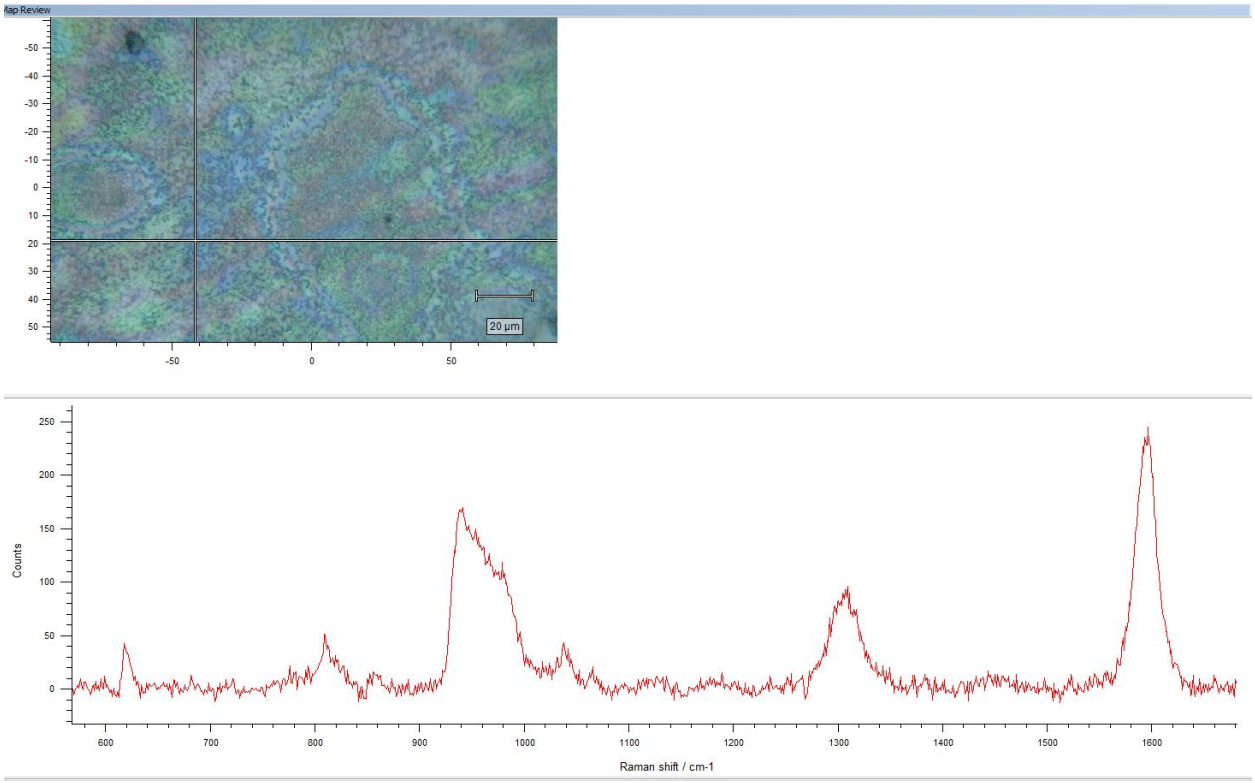



**Figure 9 – Raman mapping of the substrate before adding the T-cell or Lung cancer cell. The same SERS substrate was used for all the experiments. The top image is the optical microscope image with a scale bar of 20 µm and depicts the area that was mapped out with 10 µm step size. At the crosshair, we get an idea of what a typical Raman spectrum will look like in that mapping**

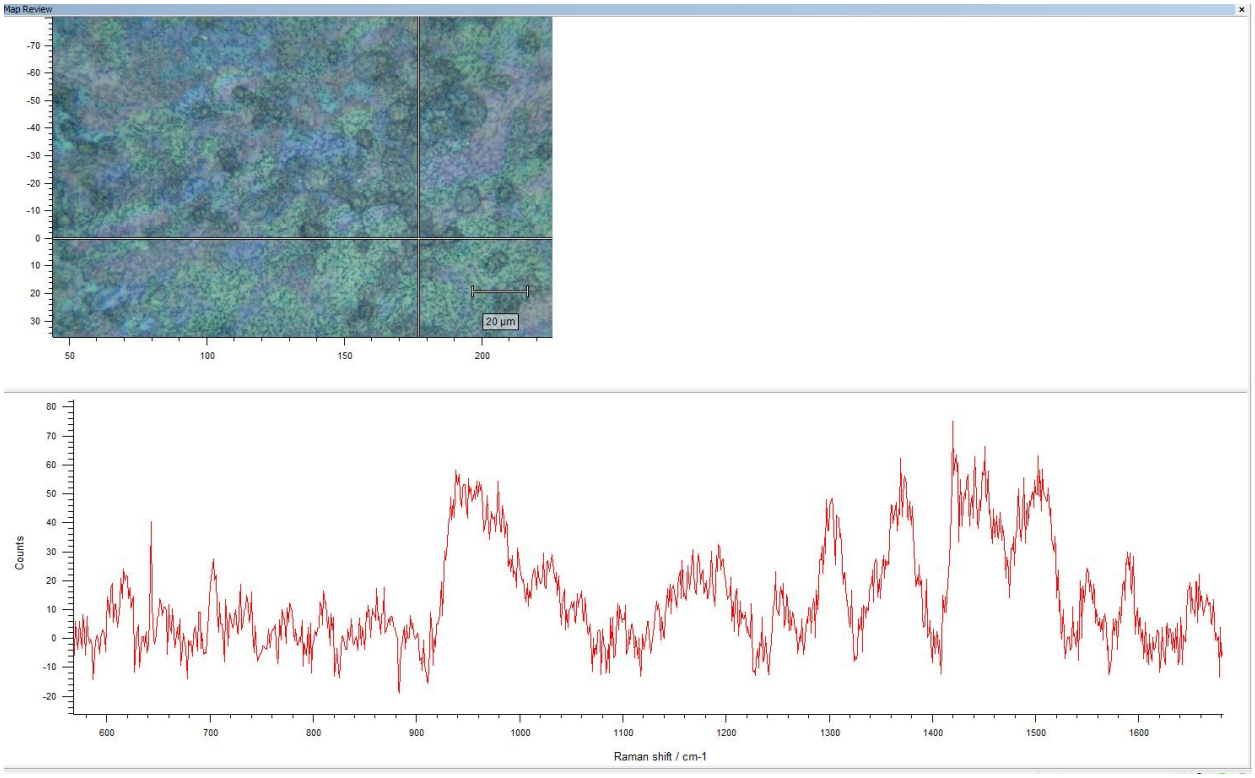



**Figure 10 – Raman mapping of the SERS substrate with the addition of T-cells. The top image is the optical microscope with a scale bar of 20 µm and a Raman mapping step size of 10 µm. The bottom Raman spectrum is an example of the typical Raman spectra in the mapping.**

From Figure 9, we can see an idea of what the SERS substrate spectra looks like before the addition of the T-cells and the Lung cancer cells. A large mapping dataset was taken and the bottom insert in Figure 9 shows a typical Raman spectrum of the mapping file. Comparing Figure 9 with Figure 10, we can already see peak differences in the Raman spectra, but we will need to use PCA to process the 500 datapoints.

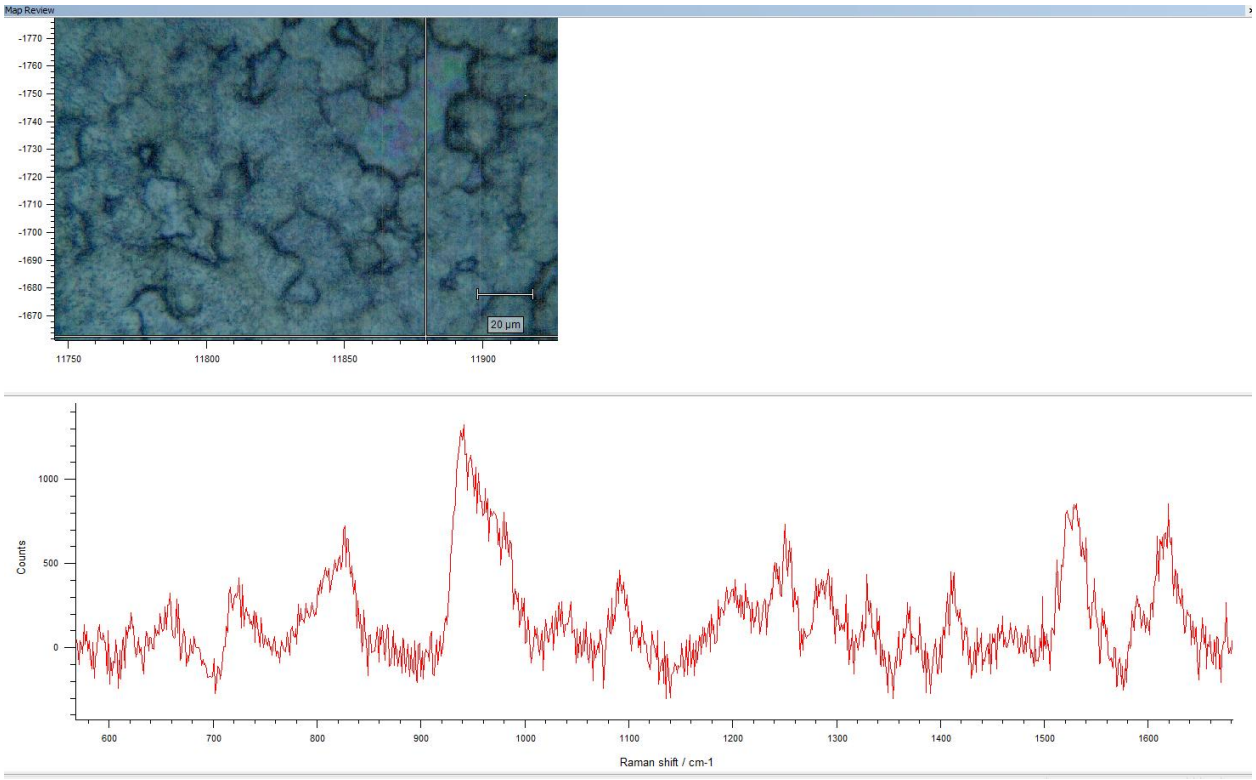



**Figure 11 - Raman mapping of the SERS substrate with Lung cancer cells. The top image is the optical microscope with a scale bar of 20 µm and a Raman mapping step size of 10 µm. The bottom Raman spectrum is an example of the typical Raman spectra in the mapping.**

If we now compare Figure 10 with Figure 11, then we also see there are major differences in the SERS peaks. After processing all those datasets, they were put through the PCA program and this is the result.

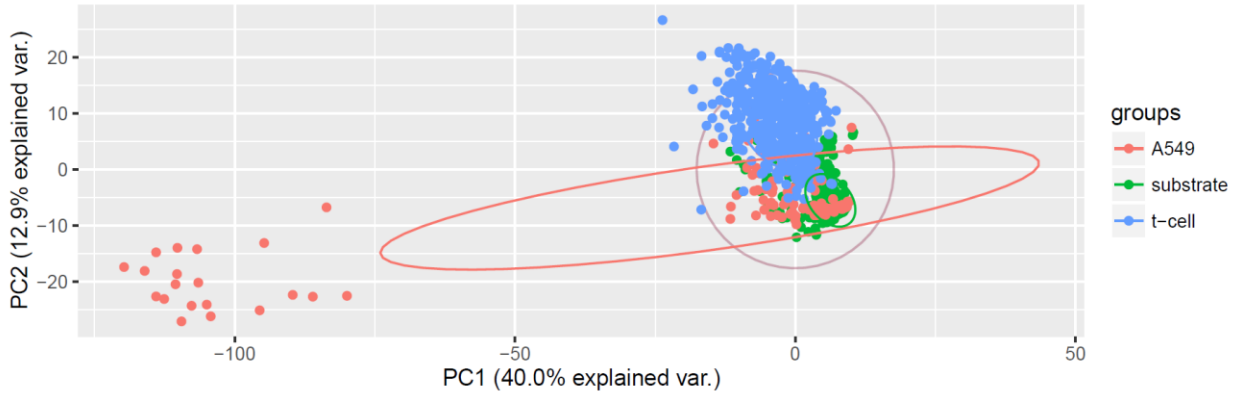


**Figure 12 – PCA graph of the lung cancer (A549), the healthy cells (t-cell), and of the substrate. Each dot represents a complete Raman spectrum; unfortunately the PCA program is overlapping some data points and will be updated to shrink the dots in the future (for clearer viewing). The colored ovals represent 67% of the dataset so the large red oval shows that there are two majority datasets and hence the large longitudinal split of the oval. However, there is still clear differentiation among the datasets.**

From Figure 12, we see three major colors where red represents the lung cancer dataset, green represents the SERS substrate, and blue represents the healthy cell dataset. Each dot in Figure 12 represents a complete Raman spectrum so we can clearly see the amount of data collected for this analysis to be statistically relevant. Each oval in the PCA graph represents 67% of all their respective colored dataset; in particular the red oval stands out because we can see a split in the collected dataset where the 20ish red dots on the left are completely different from the substrate or t-cells. We also see in Figure 12, that there is some overlap of the lung cancer dataset with the t-cell; this is most likely from the similar SERS peaks that represent a human cell. We also see that there are some lung cancer datasets inside the green oval representing the substrate; this is most likely because the Raman mapping was done over a large area and thus some regions contained no lung cancer and only the substrate was measured.

#### 3.1.1.1.2. Combing all the Data Together

Following the positive results of the preliminary data collection, another cell line was also measured, H157. This is the same cell morphology as A549 but has a different shape thus if we just focus on size differences, the lung cancer cells are similar in size and the T-cell is smaller in size, when viewed by OM. SERS mapping was done on each cell line and the resulting datasets were aggregated into the ML algorithm. At the time of measurement for the preliminary dataset, PCA was our best data analysis tool. As noted in the background section, PCA is inherently not a classifier but can be used to differentiate species if the datasets are distinctly different. The differentiation process becomes difficult

when there are enough outliers that skew the ellipse shape and also the PCA is typically combined with LDA for binary differentiation. In lieu of PCA+LDA, adaboost is used as the advanced form of the classifying tool. The additional benefit of using adaboost is the ability of multi-classification so therefore the data of the two lung cancer cell line and the control data can be compared altogether. Figure 13 represent the confusion matrix produced from the adaboost algorithm that compares the predictions from the model against the actual labeled truth. As can be seen in Figure 13, the diagonal represents the matched or correct predictions from the adaboost and any numbers in the squares beside the diagonal represent incorrect predictions. Another way to think about this is if the focus is to differentiate lung cancer vs healthy, then the diagonal corners are the true positives values, the center square is the true negative values, and the remaining squares represent the false positive and false negative.

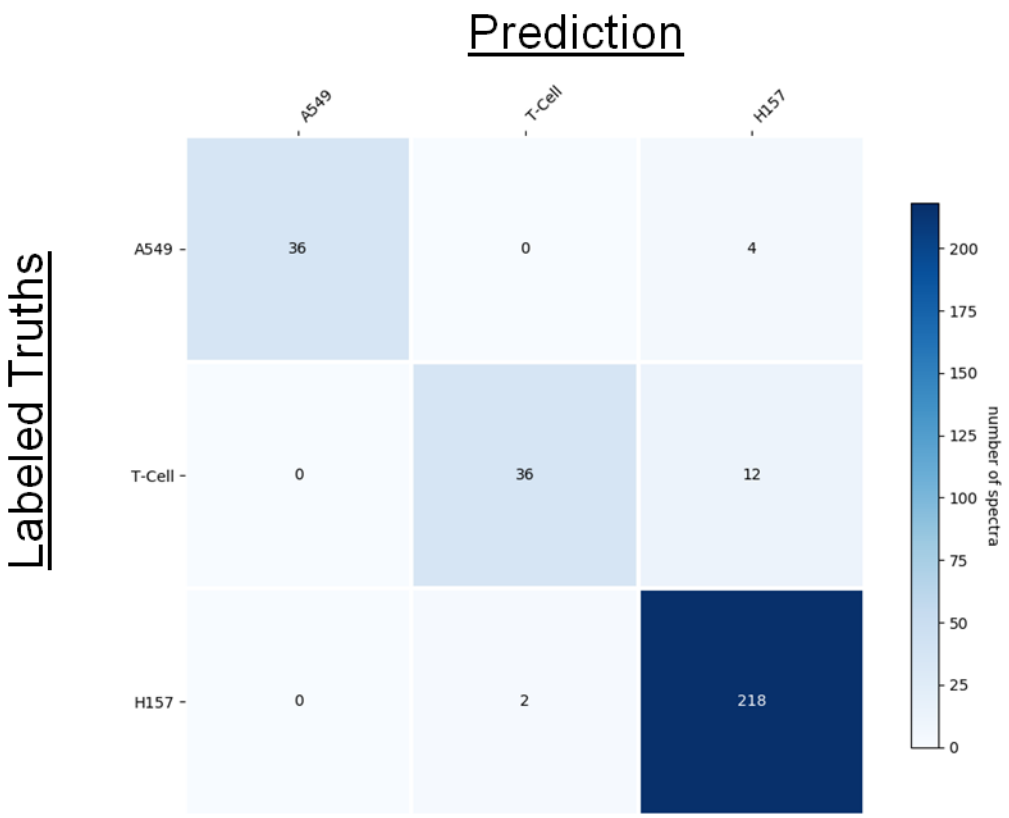


**Figure 13 – This is a multi-classification of the lung cancer vs a control for cultured cells measured with SERS. Adaboost is used to train a dataset and the trained model is used to test a separate dataset to produce the following confusion matrix. The left side represent the actual label for each classifier in each row and the top side represents the adaboost model's prediction for each classifier in each column. All matched up values, therefore along the diagonal, means successful match while any values outside the diagonal represent mismatched predictions from the ML model.**

**Table 1 – The confusion matrix values and the calculated sensitivity and specificity. To compare the effectiveness of the platform in differentiating lung cancer vs healthy patients, the last row and the highlighted red values show this. In this case, the SERS platform was 99% sensitive in detecting lung cancer and 75% specific in detecting healthy control.**

| Name | TP | FP | TN | FN | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| A549 | 36 | 0 | 268 | 4 | 90% | 100% |
| H157 | 218 | 16 | 72 | 2 | 99% | 82% |
| T-Cell | 36 | 2 | 258 | 12 | 75% | 99% |

#### *3.1.1.2. Skin Cancer*

Skin cancer is the most commonly diagnosed cancer in the US with a most recent study in 2019 estimating 96,480 new cases in the US alone [3]. Signs of skin cancer typically require personal monitoring of skin lesions monthly and consistently evaluated by a health care provider. An alternative proteomic technique can provide an alternative to this visual monitoring.

For the purpose of proving out the SERS platform, the particular skin cancer cell lines chosen look similar in shape and size under a typical OM. Thusly, clear differentiation needs to rely on alternative technique such as gene sequencing. We believe that the SERS platform can provide a cheaper alternative with no reagents needed to detect the cell proteome differences. Following the same protocol establish from our preliminary lung cancer study, a skin cancer cell line and the control cell line, a keratinocyte, are used for comparison.

#### 3.1.1.2.1. Measurement and Data Analysis

For this dataset, we wanted to investigate the proteomic fingerprint for a cancer disease compared with another cancer disease. The idea behind this is to check if the abnormal transmembrane proteins will be present in both cells and if these SERS peaks will show up and convolute the PCA. Below are two example Raman spectra of the skin cancer control (HaCaT) in Figure 14 and the skin cancer (A431) in Figure 15.

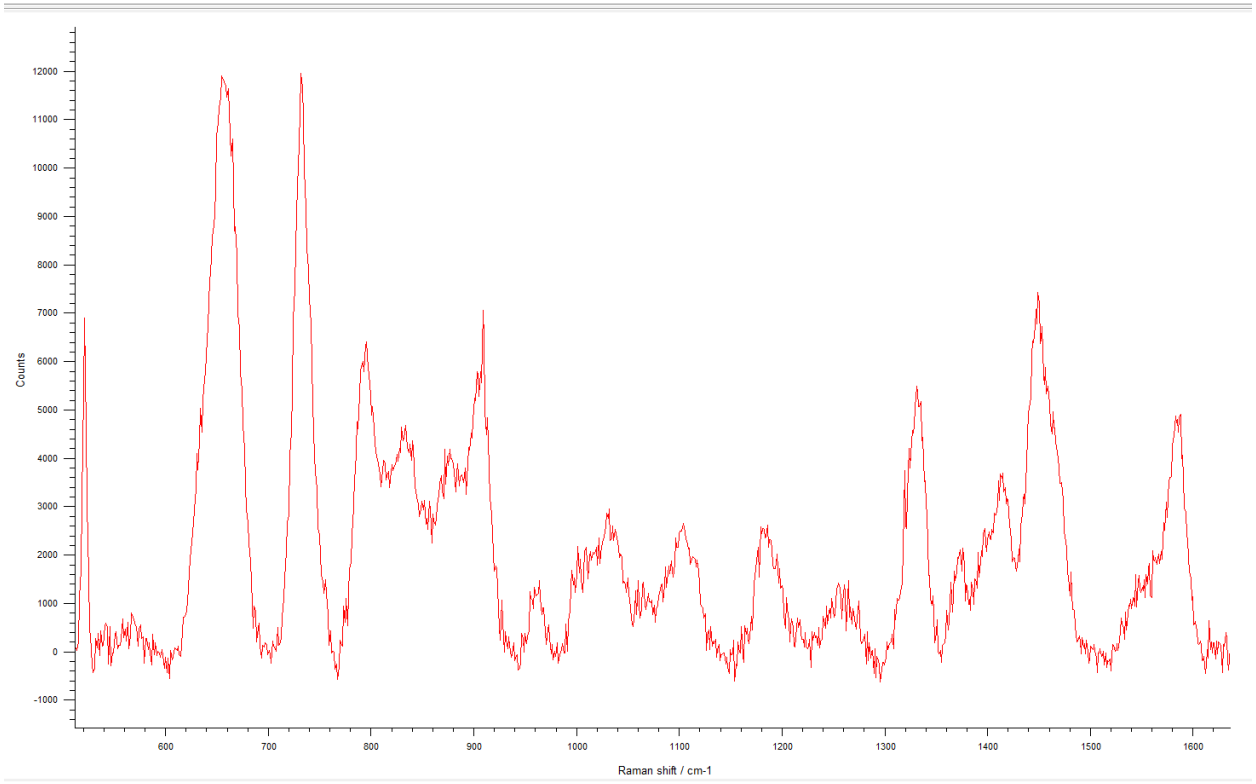



**Figure 14 – Raman spectrum of the skin cancer control (healthy cell) taken with the 785 nm laser. This can be compared to the Raman spectrum of the skin cancer cells (Figure 15) for further analysis.**

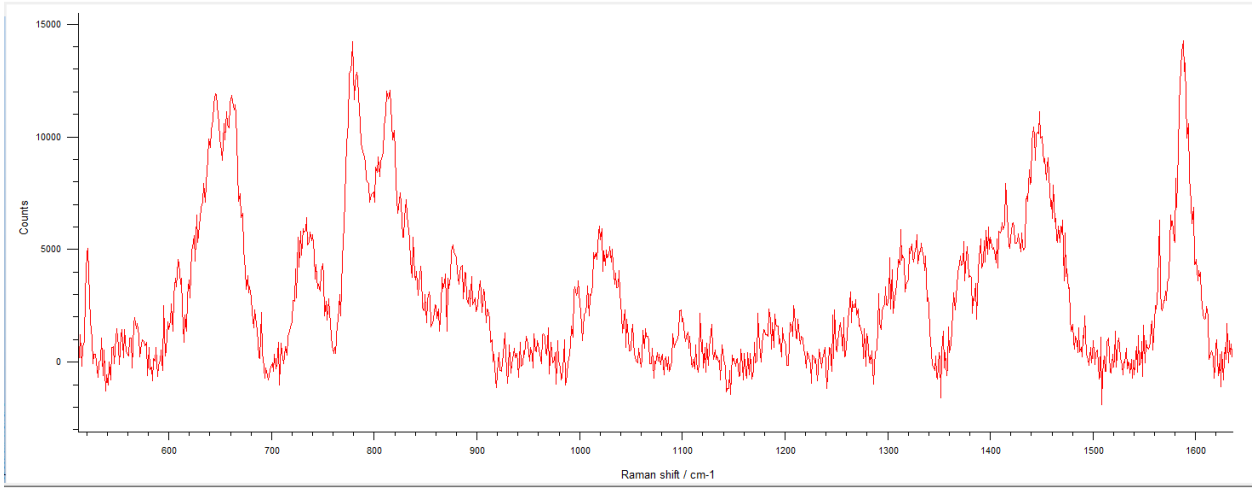



**Figure 15 – Raman spectrum of the skin cancer cell (A431) taken with the 785 nm laser. This can be compared with Figure 14 and we see some similarities which make sense for these similar cell types, but still have SERS peak differences as seen in the double peak at 650 $cm^{-1}$ vs the single peak in Figure 14.**

Doing a PCA analysis of the SERS dataset gives the following PCA graph as seen in Figure 16. Since this data analysis is just two different species, a PCA graph can clearly that most of the data is not overlapping and therefore mostly differentiable. An easy way to visualize this separation is to trace the white vertical line when PC1 = 0 and therefore most of the cancer data points are on the left and the control data points are on the right. Running the adaboost ML algorithm on this dataset shows an accuracy of 81.4% with 80% sensitivity and 82.5% specificity. Our current consensus is if the ML program is >80%, then we deem the SERS platform a success.

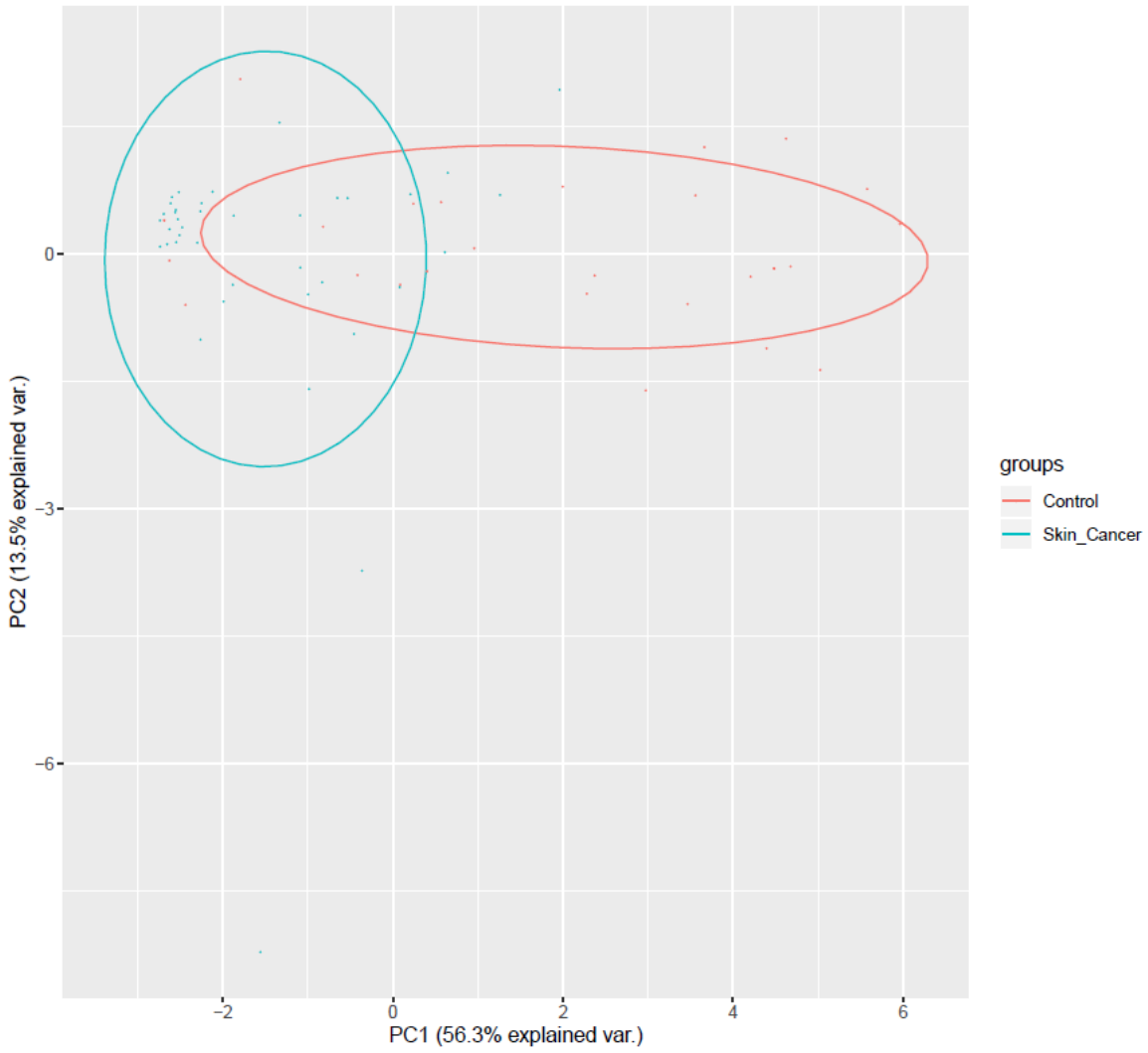



**Figure 16 – PCA of the SERS skin cancer cell line A431 vs the kertainocyte control cell line (HaCaT). If only two species are being compared, then the PCA is a good visual representation than the confusion matrix. The ovals represent 67% of all data points in each respective category.**

#### *3.1.1.3. Breast Cancer*

Breast cancer is one of the top five cancers in the US with an estimated 268,600 new cases every year [3]. However, the prognosis is relatively good compared to other cancers due to the early detection and treatment options. The current unmet medical needs for breast cancer is in improving the early detection by reducing the false positive rate and precision medicine for more targeted treatment options.

In regards to the SERS platform, doing a study on breast cancer cells gives the platform another data point in showing the adaptability of the platform and its ability to handle any type of cancer. In addition, the two cell lines being studied here: MCF-7 and MDA-MB-231, are specifically chosen to be from the same disease (adenocarcinoma), same cell morphology (epithelial), same cell type (adherent), and from the same tissue (mammary gland, derived from metastatic site: pleural effusion). The main differentiating factor between the two will be derived from the original host, or in other words, because they came from two different people. Thus, at the time in planning this experiment, it was our group's first attempt at seeing if SERS can differentiate two different "patient" samples and preliminary to the actual clinical setting.

#### 3.1.1.3.1. Data Analysis

The SERS measurement was done in the same format as previously mentioned. Only the SERS spectra of high quality were chosen and given to the ML algorithm. However, before the ML was done, a simple dimension reduction, or PCA, was done to visualize the two datasets. This can be seen in Figure 17, where the two different breast cancer cell lines are completely separated as seen by the red and teal ovals. Non-overlapping ovals mean 100% sensitivity and specificity if we rely purely on the PCA graph. To verify this, the adaboost algorithm was applied to the two datasets and resultant accuracy is 99.4%, with 99.6% sensitivity and 98.1% specificity.

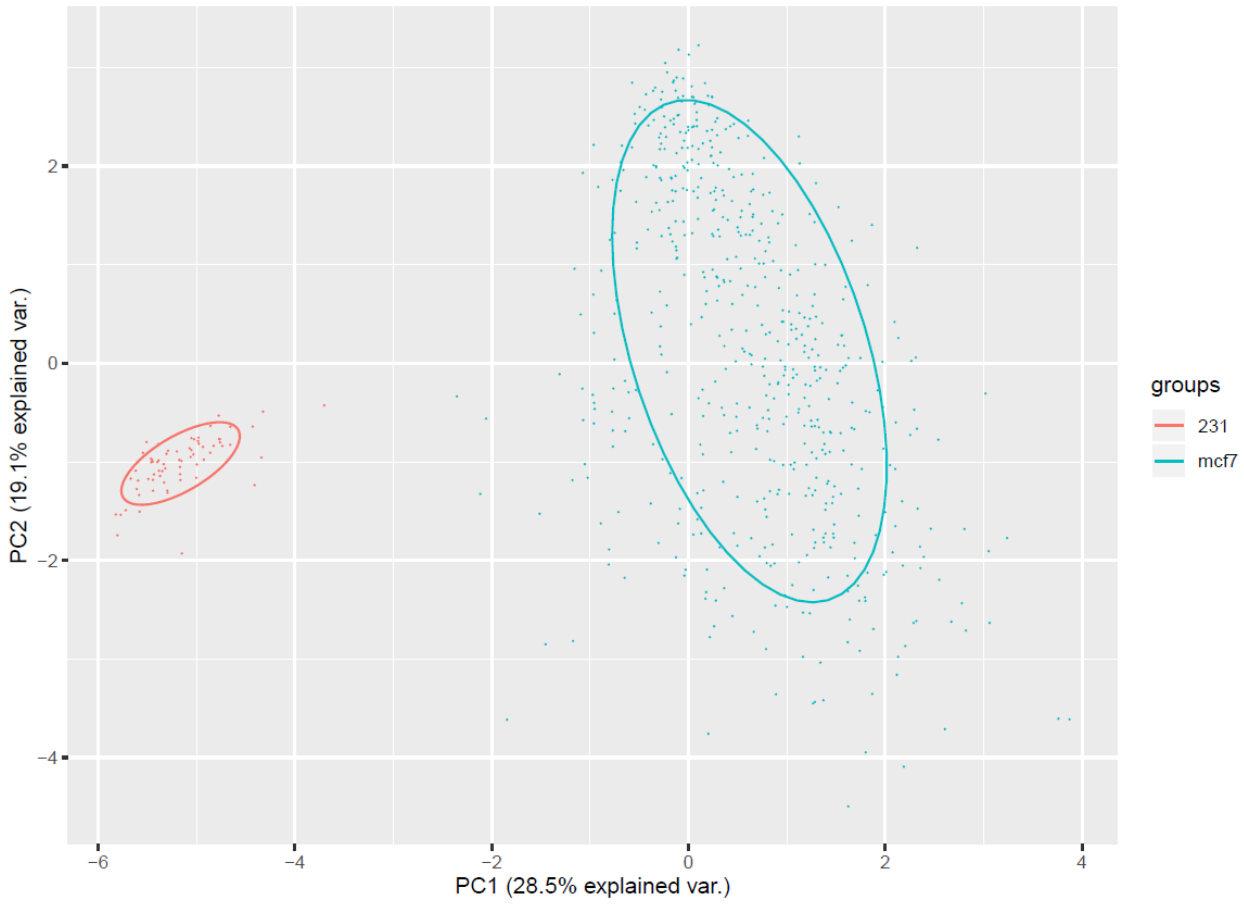



**Figure 17 – the PCA graph of the two breast cancer cell lines, MCF-7 and MDA-MB-231. As can be seen, the datasets were not overlapping and were separated by a high degree of variance; this means that the SERS spectra are distinctly different.**

#### *3.1.1.1. Combining All Cancer Data*

With the success of the breast cancer cell line study, we have preliminary shown that the SERS platform has the feasibility of differentiating cancer species down to the patient-level. Thus with this clinical possibility, a quick summary of all cancer species is presented. The next logical question for the SERS platform is to see if it has the ability to differentiate all cancer species if their data is mixed together. This could simulate clinical situations of patients with isolated cancer cells from their biological fluid and being able to uniquely identify the cancer species from just measuring that cancer cell membrane's proteome. All the SERS dataset were combined and normalized so they could be effectively compared and the multi-classification adaboost was applied to check their differentiability.

Figure 18 shows the confusion matrix of all cultured cancer cell lines for breast cancer (MDA-MB-231 and MCF7), skin cancer (A431), and lung cancer (A549 and H157). Reading it in similar fashion as the previous confusion matrices, the left column is the actual labeled truths and the top row is the model's predicted labels. Therefore any diagonal values are true positive for each respective cancer cell line. The most notable observation can be seen in the middle row, representing skin cancer, which the adaboost program had difficulty differentiating among the other two cancers. Upon closer examination of the skin cancer dataset, we determined that the amount of data was not sufficient to fully train the ML algorithm. This also lines up with the skin cancer vs control study of being the lowest sensitivity/specificity. However, even with this upset, breast and lung cancer are clearly differentiable, with skin cancer partially differentiable. For each particular cancer species differentiability, Table 2 shows these values, very clearly showing the weak differentiability of skin cancer.

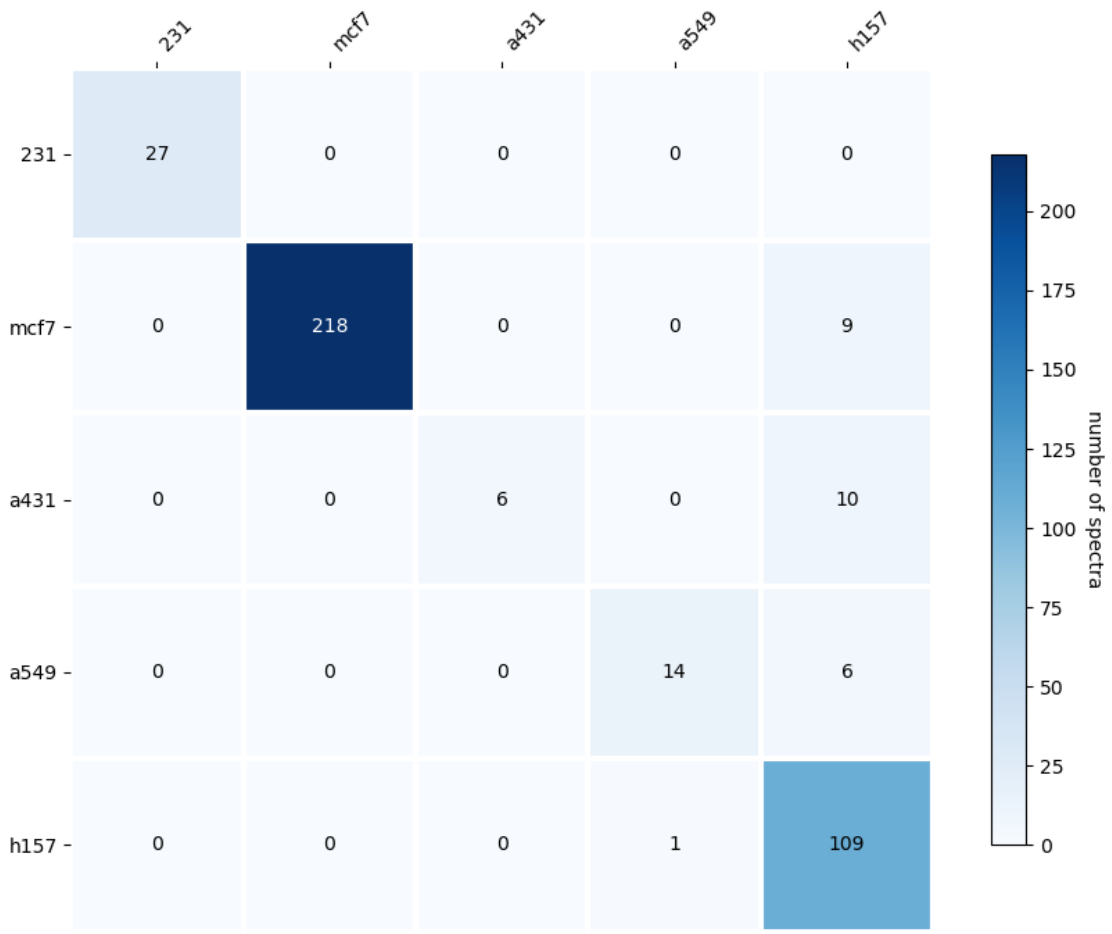



**Figure 18 – Confusion matrix of all cancer cell lines species normalized and combined together for adaboost data analysis. Similar to the other confusion matrices, the left column represents the actual labeled truths and the top row represent the model prediction labels. The first two rows are for breast cancer (MDA-MB-231 and MCF7), the middle row is for skin cancer (A431), and the last two rows are for lung cancer (A549 and H157). Right away, it is clear that most of the cancer species are correctly classified with skin cancer having the most difficulty.**

**Table 2 – the sensitivity and specificity calculations of the cultured cancer cell line via the adaboost algorithm. The focus of this table is to focus just on the cancer species disease differentiability.**

| Name | TP | FP | TN | FN | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Breast | 245 | 0 | 129 | 9 | 96% | 100% |
| Skin | 6 | 0 | 368 | 10 | 38% | 100% |
| Lung | 123 | 19 | 251 | 0 | 100% | 93% |

### 3.1.2. Microbiology

Microbiology is the study of bacteria, fungus (yeast), and viruses. While the current talk of the time is focused on viruses due to the 2019 coronavirus pandemic, before this pandemic, the most talked about problem in this category is typically about antibiotic resistance. In a report by the Center for Disease Control (CDC) published in 2019 [31], more than 2.8 million antibiotic-resistant infections occur in the US each year and more than 35,000 people die as a result. Another worrying fact is that while antibiotic development has slowed down due to low attention and funding, the speed of antibiotic resistance strain evolution has not. An evidence of this can be seen in a trendline in [31] where the first antibiotic (penicillin) was developed in 1943 and the first resistant strain was identified in 1967; nowadays a new antibiotic developed (e.g. ceftazidime-avibactam) can be developed in 2015 but the first antibiotic strain was also identified in the same year.

In regards to proving out the SERS platform, the overarching theme is to figure out the fundamental limits of the platform. Typically, a cell is referred to as an animal cell and that is the cell type we were measuring for cancer cells. The thinking to switch to bacteria and to yeast is because the cell membrane also has an outer layer called a cell wall and most of the cell membrane proteins we typically measure with SERS are tucked behind this. There was a concern that the differentiability in the cell wall proteins are not enough compared to the membrane and we wanted to test this limitation with the SERS platform. For a potential clinical application, we also focused on another advantage of SERS in its fast measurement time. To reiterate, SERS has an advantage over Raman in that with an enhancement of the signal, less time is needed for a usable SERS signal; this is at least a 2 order magnitude of time savings. What this means is that for time sensitive clinical applications, SERS can help solve these unmet medical needs.

The medical disease that our group focused on is meningitis. Meningitis is a pathogen infection in the central nervous system, typically found in the cerebral spinal fluid (CSF) in the spinal cord, and the most common pathogen is bacteria, followed by fungi. The important thing about meningitis is the need for fast diagnosis, where a patient requires treatment within 24-48 hrs. The typical treatment is to immediately start a cocktail of antibiotics when meningitis is suspected and that is where unnecessary usage of antibiotics introduces the evolution of these resistant strains. The typical gold standard for meningitis diagnosis is culturing but such processes require 2-3 days thus the clinician gives the treatment at any suspected inflammation and not even wait for the diagnosis. A targeted species treatment will help reduce the spread of antibiotic resistance and it's one of the guiding principles of antibiotic stewardship [32][33].

#### *3.1.2.1. First Cultured Bacteria – E.coli*

The first species to test our platform on is Escherichia coli or E. *coli*. E. *coli* was chosen because of it has very quick doubling time, the time it takes for one cell to double, therefore one can culture a complete petri within 18 hrs. Additionally, E. *coli* is one of the possible species that can be found for meningitis so it was a good first feasibility check for our platform.

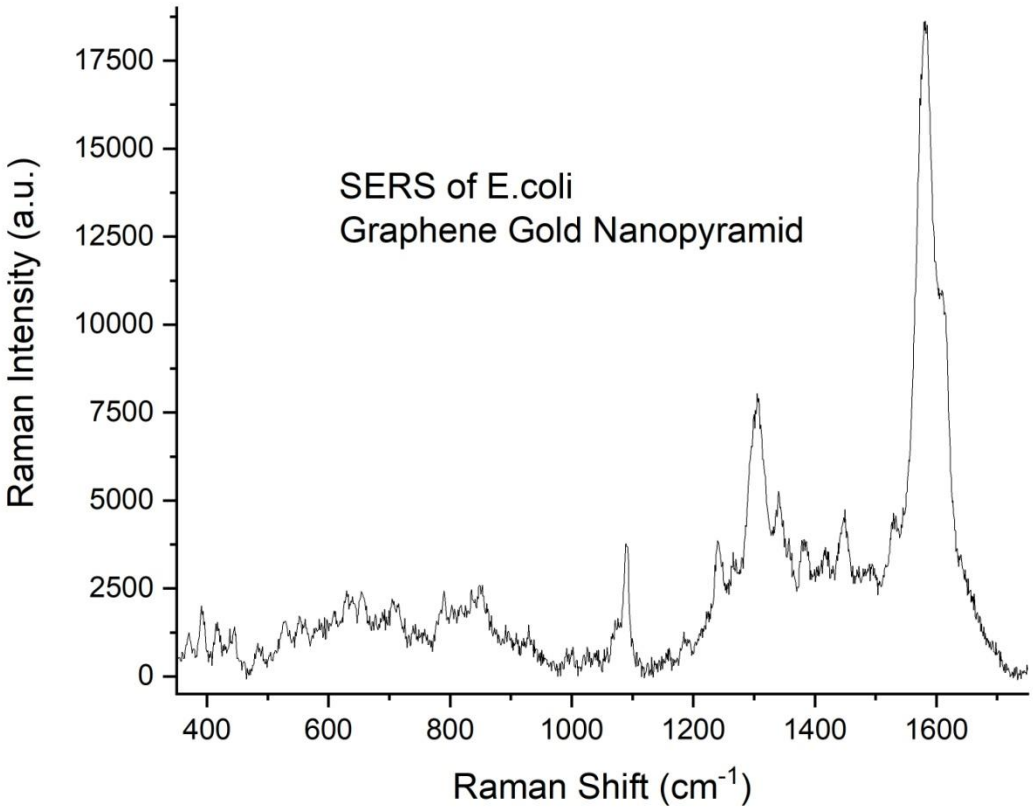



**Figure 19 – This is an example SERS spectra of E.coli measured with the SERS nanochip.**

#### *3.1.2.2. Fungus in CSF - Cryptococcus*

With the success of the E.coli SERS results, we were able to form a collaboration to measure meningitis patient samples. But before we could handle patient samples, our platform had to be first tested and the opportunity to test out fungus samples came to be. While fungal infections are much fewer in prevalence when compared with bacterial infections, the main fungal species related to meningitis is the Cryptococcus family. Within the Cryptococcus family, there are two species of interest: neoformans and gattii. Typically, to differentiate these two species, genome sequencing is needed and that is typically not done, except academically. We were able to set out to investigate if SERS is able to

differentiate these two fungi species. Although patient samples were used, these fungi were isolated and then grow in culture, thus only the dominant species is used for identification.

In regards to proving out the SERS platform, the theme for the study is to continue testing the limitations of the SERS platform with another cell type: the yeast cell structure. The yeast cell structure is an essentially a plant cell and thus for the SERS platform, it will be a mix of the previous bacteria and cancer cell studies. Once again, the question of whether there is enough cell membrane proteome to differentiate species and potentially subspecies drew us to investigate this study.

#### 3.1.2.2.1. Cryptococcus neoformas vs Crytococcus gattii

Following our established standard protocol of our SERS measurement, SERS data was collected for the two different fungus species. The datasets were then first analyzed with PCA to see if they could be differentiated and Figure 20 is the result. As one can see, while there is an overlap of the green and blue ovals, the majority of these two datasets can be separated as indicated by the horizontal black line. This shows that while imperfect, a simple PCA can differentiate these two species. In fact, a follow-up ML analysis shows that the adaboost algorithm show 99.7% accuracy in differentiating between these two species. The red oval is when we add in the E.coli data collected from the previous study. This is to mimic the possibility of another meningitis pathogen and to see if the SERS platform will confuse the three species. When we have shown this PCA graph to the doctors, they were quite impressed and gave us the go-ahead to measure their patient samples. The vertical black line in Figure 20 shows complete separation between the bacteria and fungi species and for completion, the adaboost program had 99% accuracy in separating the bacteria from fungus.

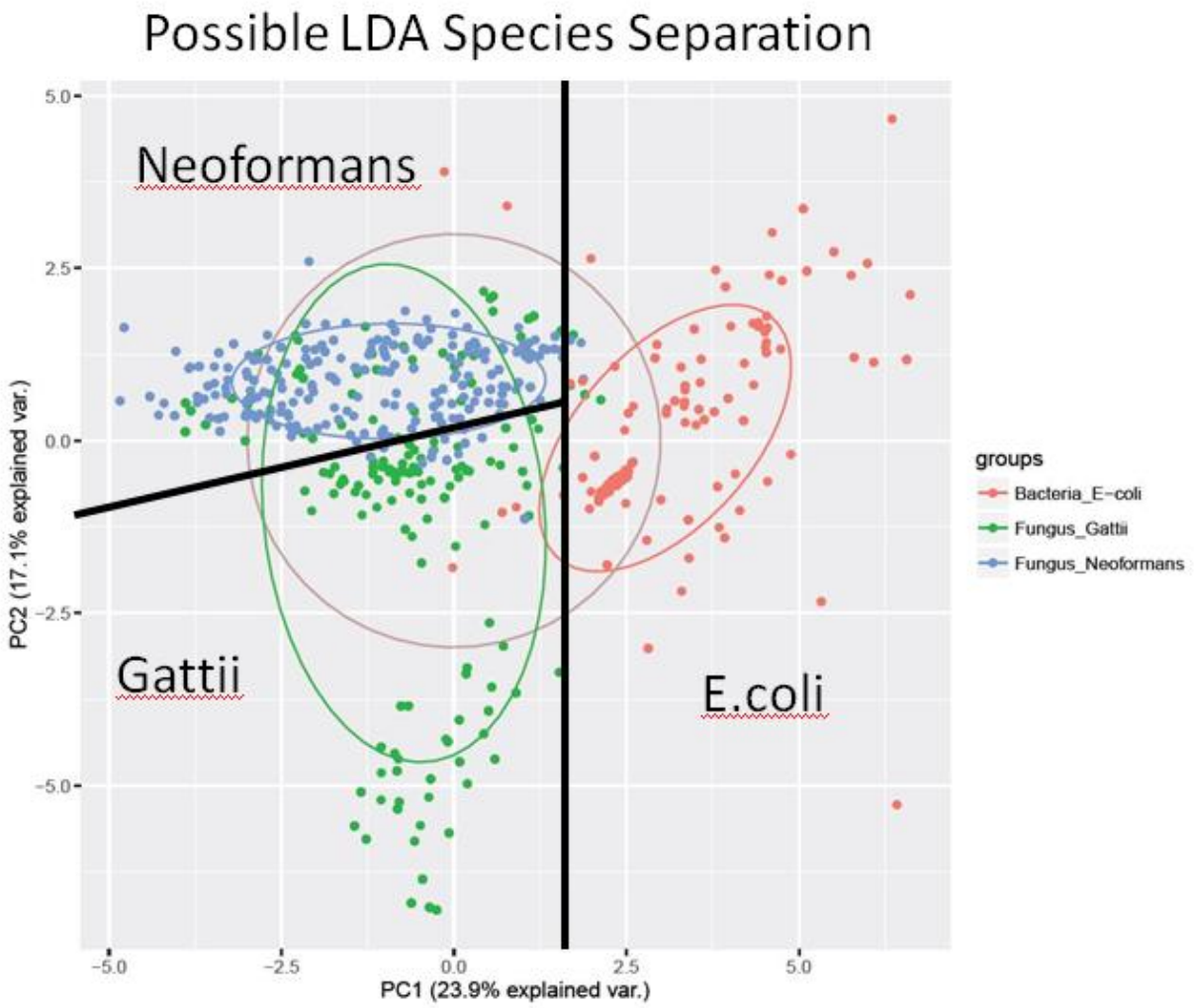



**Figure 20 – PCA graph of the SERS measurement of two Cryptococcus species (neoformans and gattii) when compared with E.coli. These species can be potentially found for patients diagnosed with meningitis. What is particular to note is the three different regions on the PCA graph for each species, as denoted by the black lines. This shows that even with just a multivariate analysis, the data is unique enough to separate the three species.**

Another interesting part of the study is that we were actually given three fungi samples. The third fungus species is another subset of the gattii species; however, their genetic sequence is a little different. These two gattii species were labeled ST7 and ST106, and we labeled as two different subspecies. To separate these two species, their complete genome had to be sequenced. We measured their SERS spectra and while the PCA was not able to differentiate these two subspecies, the adaboost showed great differentiability with 99.8% accuracy. This gave us great preliminary evidence that the SERS platform has great potential and that we were ready for the true clinical studies.

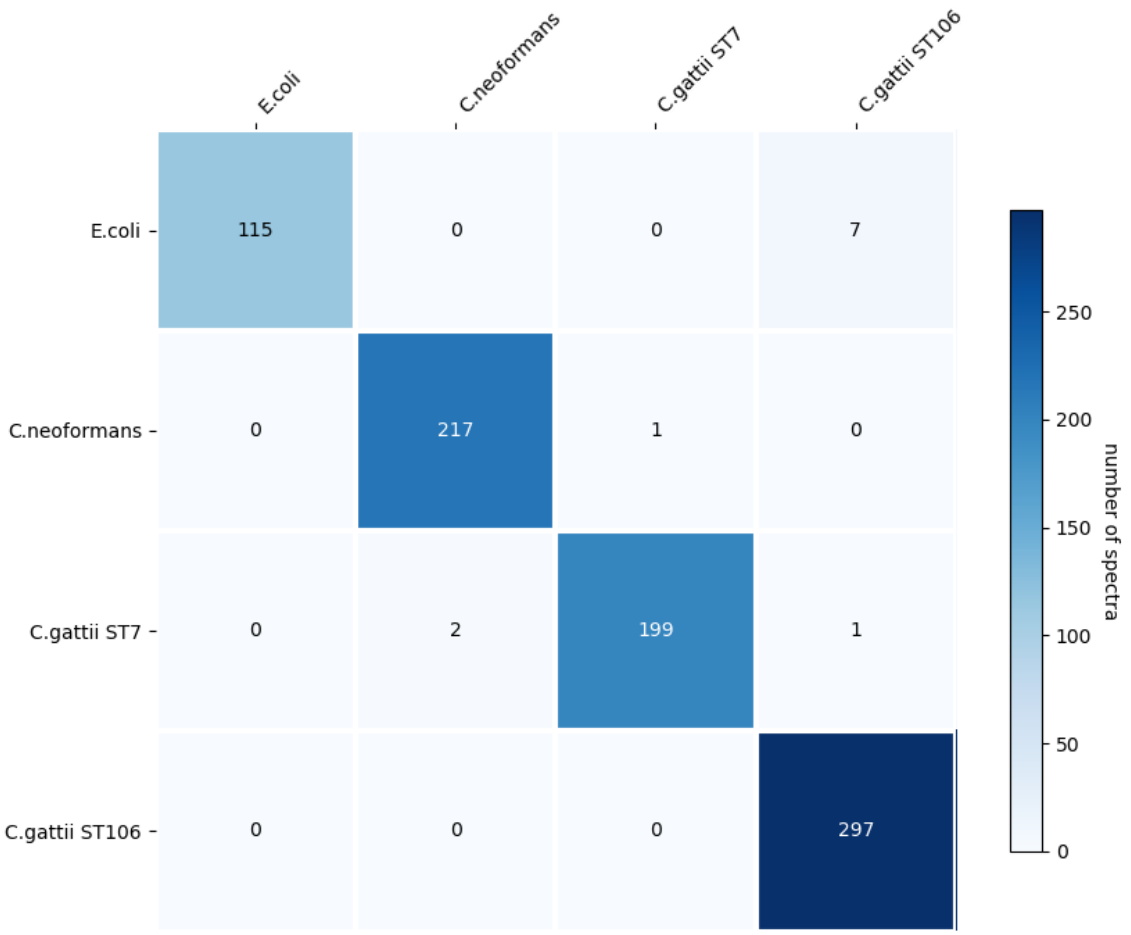



**Figure 21 – Multi-classification adaboost of the fungus and bacteria species for meningitis. What is interesting to note is the subspecies involved for the Cryptococcus gattii, ST7 and ST106. Where these two subspecies can only be separated by genetic sequencing, SERS + ML has also shown the same differentiability with 99.8% accuracy.**

# 4. Case Studies of Clinical Applicability

Building upon the successes of the cell line datasets, I was given a couple of opportunities to test the SERS platform on clinical patient samples. One of those opportunities occurred in China where I spent a year in the city of Suzhou and another year in the city of Xi'an working with our collaborator from Xi'an Jiaotong University. The focus of this Chinese collaboration was to apply the SERS platform to a major high-tier hospital in the city of Xi'an where its prestige drew patients from all five nearby provinces. Because of this, their patient influx is incomparable compared to USA hospitals and this gave us the opportunity to test the SERS platform on actual meningitis patients, which we were able to preliminary show good proof in Chapter 3. My second opportunity came through a collaboration effort with the Government College University of Faisalabad (GCUF) where the project focused on applying the SERS platform to help diagnose multi-drug resistance for tuberculosis bacteria strains. This was made possible by shipping the SERS platform to Pakistan and then working very closely with the collaborator at GCUF and then doing the data analysis back at UCLA. Below are more detailed information for each particular study.

## 4.1. Meningitis in China

With the increase of antibiotic resistant microorganisms in recent years, antimicrobial stewardship is a much debated and necessary topic. However, the implementation of a gold standard, such as microbial culturing, is still not widely-enforced where doctors can most of the time diagnosis patients empirical with a high degree of accuracy. This situation is further worsened in meningitis patients where patient survival is determined by the hours and the lengthy time of a culture for species identification may not change a doctor's antibiotic treatment. In addition, the scenario becomes even more complex in developing countries where patient influx is several orders of magnitude higher than developed nations. Such a triage scenario leaves hospitals with no choice but to follow the diagnosis of

experienced doctors and relying on basic initial observations, such as cell count, white blood cell (WBC) type, CSF turbidity, etc, to determine a patient's aliment.

A case study of the above-mentioned scenario can be seen in North-Western China, in the city of Xi'an, Shaanxi province. Xijing Hospital is a top Chinese military hospital that is nowadays open to the general public. Based on the country's health care system, hospital visits are fairly inexpensive and the same price among the various tiered hospitals. As such, generally the most prestigious of hospitals will receive the most influx of patients, as seen with Xijing Hospital which is currently ranked the best within its nearby five provinces. This large influx of patients gives an unique situation in the neurological ward, where patients that are suspected of cerebrealspinal infections are brought in. Typically, the hospital expects 5-30 patients per day that require a lumbar puncture; a comparative example with the Ronald Regan Hospital in Los Angeles, California, USA, is to expect 1 patient per month that requires a lumbar puncture. In addition, Meningitis patients are more common in developing countries and around rural areas and these factors help boosts Xijing Hospital's unique numbers. However, the downside of large patient numbers is the strain on the hospital's microbiology clinical testing laboratories, where it is more practical for doctor diagnosis based on initial white blood cell count differentials than the gold standard of culturing. As outlined by the World Health Organization (WHO), the gold standard for Meningitis patients is for Cerebral Spinal Fluid (CSF) culturing [34].

An alternative method is present that does not require culturing and with its low patient analyte requirement, can be used in parallel with current methods: Raman Spectroscopy. This spectroscopy technique is classified under vibrational spectroscopy, other examples include infrared spectroscopy, and it measures the vibrational bonds of molecules in a material. The intriguing application of this technique to clinical studies is that the laser-induced bond vibrations are energy based and can provide a unique spectroscopic spectrum, when coupled with the proteins, lipids, and DNA of a cell. This allows

Raman spectroscopy to produce unique identifying "fingerprints" that can be used to differentiate cells on a species taxonomy level. The collection of multiple relevant disease species in the form of a Raman database can allow Raman spectroscopy to be used in clinical application with advantages such as single cell analysis, fast scan and disease diagnosis (when compared to the gold standard of culturing), and less human-error prone, based on reliance to automated computational analyses.

### 4.1.1. Current Gold Standard – Cytology and Culture

In this study, the identification and differentiation of two white blood cell species, Monocyte and Lymphocyte, are compared based on the Raman proteomic signature from hospital patient CSF samples. It is shown that the conventional Raman signature is enough to clearly differentiate the two cell species and show the advantages of using this alternative technique compared to the current gold standard. Figure 22 shows a schematic of the various protocols carried out by 1) Xijing Hospital's typical routine, 2) the WHO's gold standard for CSF patients involving cell culturing, and 3) the proposed Raman protocol for handling CSF patients. As shown in Figure 22a, an outline of the typical patient protocol for Xijing Hospital is presented and it is seen that doctor diagnosis typically just relies on the white blood cell count differential. There are some cases, based on the patient's history, where additional procedures are requested, such as cell culturing and/or serological testing. Figure 22b shows the WHO guideline which basically summarizes that all patient CSF needs to be cultured for microbe species identification before antibiotics can be given. Figure 22c is the proposed Raman alternative that adapts most of the current Xijing Hospital protocol, but introduces Raman spectroscopy for cell species identification. The obvious advantage of following the WHO guideline is to advance antimicrobial stewardship at the cost of time, while the protocol of Xijing Hospital shows the difficulty of implementing the culturing gold standard and the necessity of quick patient diagnosis. The Raman alternative is introduced to take advantage of both methods by giving the ability of species identification

without the complications of culturing. In addition, the process of Raman mapping each cell from a patient's CSF can be automated and the result analyzed with high sensitivity and specificity such that the chance of human error is routinely minimized.

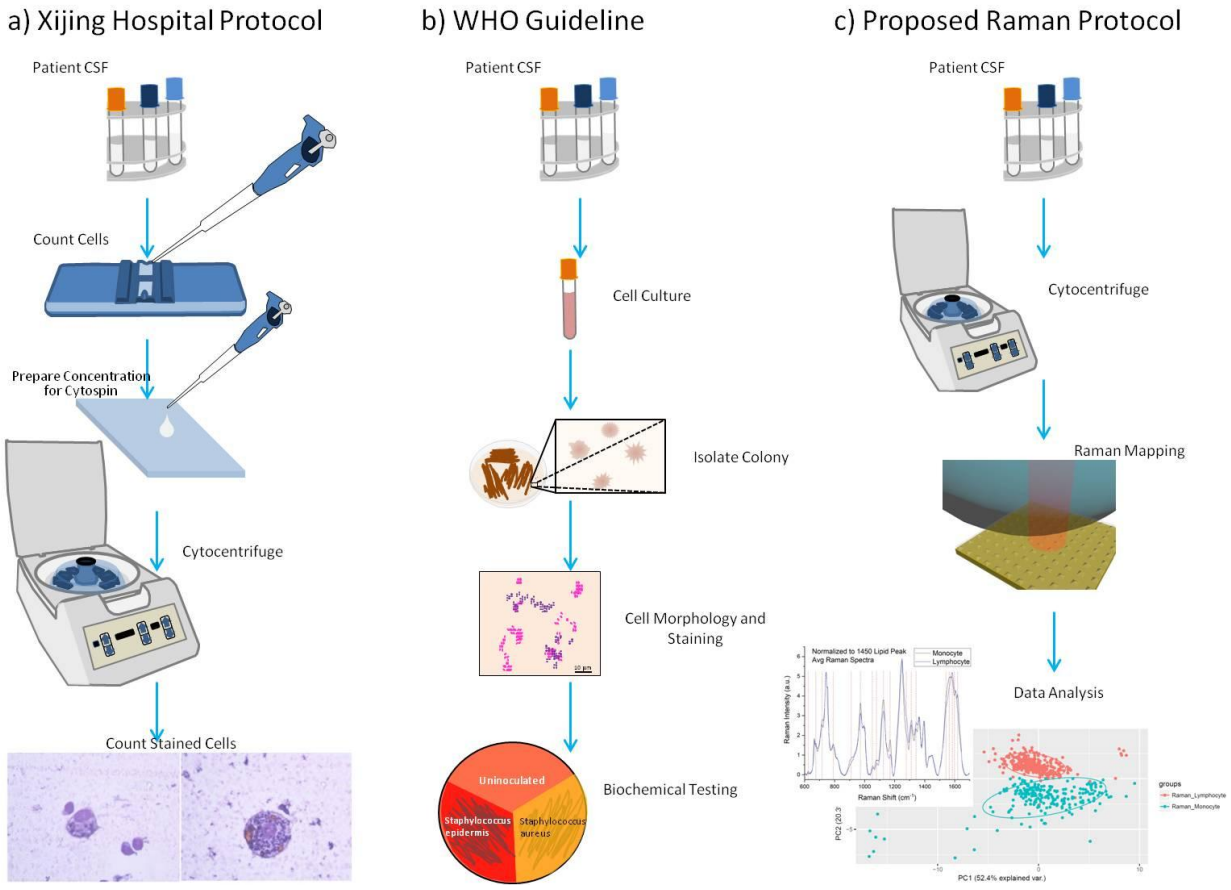



**Figure 22 – Schematic diagram of three different methods for handling CSF patients. (a) Xijing Hospital's staining and white blood cell count. The basics of the protocol is outlined here: 1) Patient CSF is extracted via lumbar puncture, 2) aliquot to a hemocytometer for initial cell count, 3) based on the cell count, a set concentration is aliquoted for cytocentrifuge, 4) cytocentrifuge, 5) fix and stain cells and count/identify each species. (b) WHO guideline for culturing patient CSF. The protocol basics: 1) Patient CSF is extracted via lumbar puncture, 2) sample is cultured in incubator, 3) if positive, microbes will form colonies and isolated based on shape, 4) further identification can be done with Gram staining and cell morphology determined by optical microscope, 5) species identification can be determined via biochemical reagent testing. (c) Proposed Raman protocol for handling patient CSF. The protocol basics: 1) Patient CSF is extracted via lumbar puncture, 2) a gold film is prepared and the CSF is cytospun to concentrate the cells, 3) Raman mapping is done for each cell, 4) data is feed into a computer for analysis that reports the sensitivity and specificity.**

#### *4.1.1.1. Typical Hospital Procedure*

All patients gone through this study receive normal hospital treatment and only leftover CSF collected from the patient are used to test the Raman spectroscopy portion. CSF is collected from the

patient via lumbar puncture and typically 1-10 mL of CSF is collected. 100 µL is set aside for the Raman protocol. Meanwhile, 10 µL of CSF is aliquot for initial cell counting with the hemocytometer to get a concentration per µL. A typical healthy patient contains 0 red blood cells (RBCs) and 0-5 WBCs per µL.

Even if the patient at this point is deemed healthy, cell staining and WBC species identification is done for confirmation. This is done by aliquoting a certain amount of CSF for cytocentrifuge from a conversion table based on the initial cell count. At 800 xg for 10 mins, the cytocentrifuge concentrates the cells onto a microscope slide and residual CSF are collected by the perimeter filter paper. Cells are then fixed with acetonformaldehyde for 1-2 mins, and then the May-Grünwald-Giemsa (MGG) staining reagent is applied for 1-2 mins. The slide is then washed with De-Ionized (DI) water several times. Cells are then viewed under an Optical Microscope (OM) at 400x and 1,000x. About 200 stained cells are counted and differentiated based on size, shape, granules, and number of lobes of nuclei according to the respective WBC type; RBCs are also stained but not counted.

### 4.1.2. SERS Feasibility of Differentiation WBC Type via Proteome

In this study, the focus will be on how to apply the proposed Raman protocol, as mentioned in Figure 22, to Xijing Hospital and evaluate its effectiveness. Figure 22a depicts the Xijing Hospital protocol with the crucial step being the cell staining, identification, and counting. An example of some stained cell images is depicted in Figure 23a showing some MGG-stained lymphocyte cells. The MGG staining clearly shows the nucleus of the lymphocyte as compared to an unfixed lymphocyte that is dried onto a gold film substrate as seen in Figure 23b. The identity or label of each cell is verified by a professional doctor from the hospital. The OM image of Figure 23b is the typical cell shape to search for before a lipid Raman spectrum is measured to verify that this image imprint is a cell. Figure 23c is an example of this lipid measurement and the vibration modes typically associated to these peaks are the C-H stretching and $CH_2$ asymmetric stretching bond Table 3. Once a cell is verified with this lipid

measurement, a mapping is done to extract the protein-rich Raman spectra of the cell. An example of a typical Raman mapping can be seen in Figure 23d, where a lipid mapping measurement is done over the area of two lymphocyte cells. Figure 23d is an OM image with the lipid mapping measurement overlaid on top; the faint blue is the substrate background noise with a dark blue color showing the outline of the cell, and finally, a bright green, yellow, or red representing the intense signal of the lipid of the cell.

Once a cell is verified, a square protein-rich Raman mapping is collected. The Raman mapping criteria are to ensure minimum overlapping of data via large step sizes and high signal/noise by restricting the map size to be within the cell. Although some information may be lost by not measuring the cell boundary, it was found that typically each Raman spectrum in a mapping is very similar in intensity and shape. Initially, it was expected that Raman spectra may vary due to dissimilar proteome lipid rafts on a cellular membrane surface; however, it is believed that the large laser spot size averaged out the differences. Although it is well-known that various CSF proteins are present and should be everywhere, it is observed that there are very little salt and protein crystallization as most of this is suctioned away via the filter paper in the cytocentrifuge. This reduced concentration of CSF proteins does not show up in the Raman spectra and it is believed that these residual proteins are coupled into our background noise.

One major advantage of this Raman technique is the automated process that can be carried out by the Raman computer. For the typical Xijing Hospital protocol, each patient will need at least 200 cells counted by an individual doctor and these cells will need to be sorted by each cell species. In contrast, a Raman mapping of a patient sample can be theoretically automated by first, a generic mapping with the lipid measurement, and then a protein-rich measurement for cell species fingerprint comparison. Such an automated method will be less error-prone and free up the necessary hospital personnel from such a tedious task.

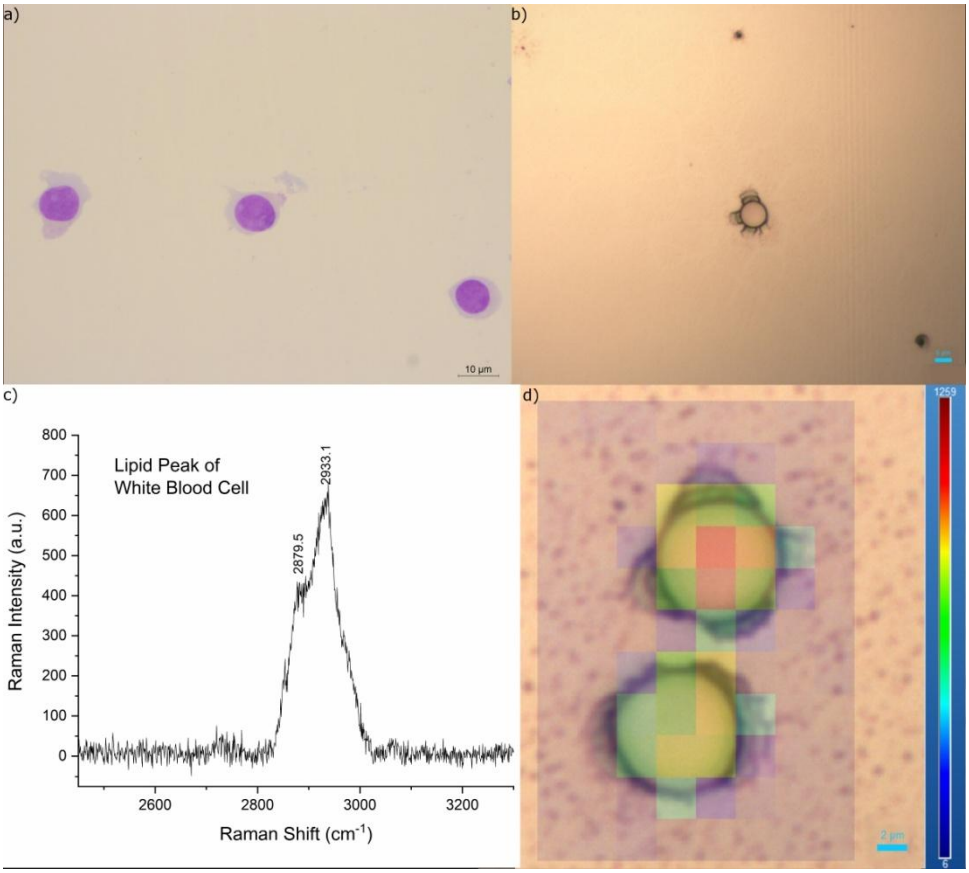



**Figure 23 – 4 insert panel showing the Raman evidence of the presence of WBC. (a) shows the MGG stasining of cytospun WBC cells onto a microscope slide. (b) shows the WBC viewed under the OM of the Raman spectrometer when dried normally and unfixed on our chip. (c) is the Raman spectrum of the WBC at the high wavenumber region showing the presence of lipid peaks. (d) is the Raman mapping cross-overlay ontop of an OM image similar to (b) where the red color represents high intensity of the lipid peak and the blue color represents the low intensity of the lipid peak.**

The averaged Raman spectrum of monocyte and lymphocyte is shown in Figure 24a. The various peaks can be seen for each cell species and their respective vibrational mode is listed in Table 3. Although typically, a Raman fingerprint will have different wavenumber peak locations, yet as seen in the figure, the intensity is the major difference between the two species. This is attributed to the similarity of the vibration modes in the various amino acids and the abundance of each mode, signifies the different functions of the two WBC species. When the spectra are normalized, as seen in Figure 24b, the spectra similarities are further highlighted. To differentiate between these species, a computerized analysis program is necessary and thus PCA and DNN were used. For PCA, important peaks that can

differentiate between the two species are selected and highlighted in Figure 24 by the red dotted lines. The important peaks were chosen based on the greatest distance, peak-to-valley, and peaks of similar intensities were assumed to be essential proteins typically found in all cells. These important peaks are the variables that form the linearly weighted vectors of a principal component. The graphical representation of the two most important principal components is displayed in a PCA graph, as seen in Figure 25.

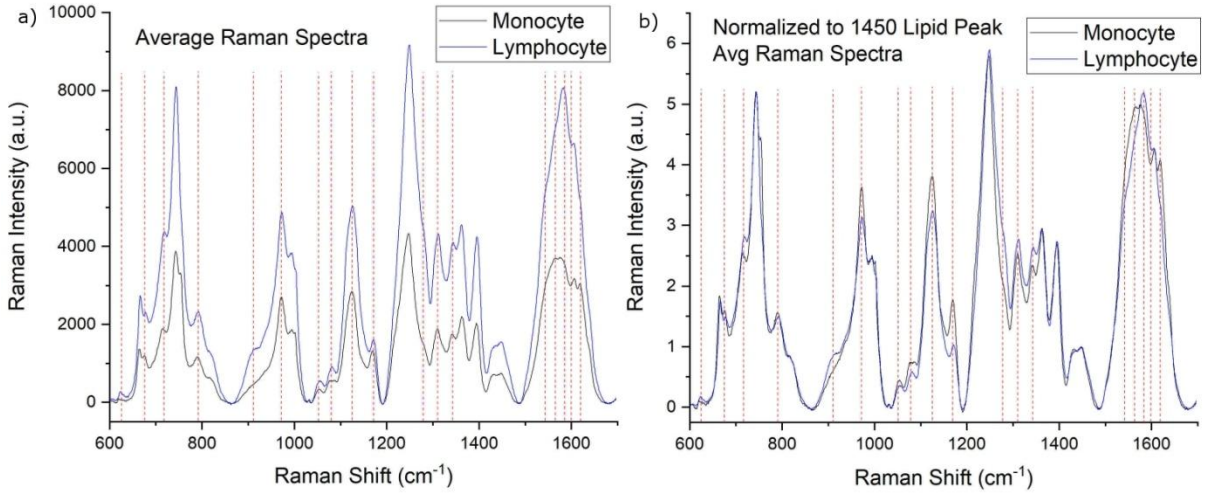



**Figure 24 – Raman spectra of the two types of WBCs typically found in the CSF of patients. (a) shows the averaged Raman spectra of monocyte and lymphocyte. (b) shows the averaged spectra but normalized to the lipid peak to accentuate the key differences between the two types. The dashed red lines are the key Raman peaks listed in Table 3.**

**Table 3 – Raman Assignment of the Biologically Rich Regime**

| Raman Peak Center | Bond Assignment |
|---|---|
| 624 | C-C twisting mode of phenylalanine (proteins) |
| 675 | Ring breathing modes in the DNA bases |
| 717 | C-N (membrane phospholipids head)/adenine $CN^{-}(CH3)_3$ (lipids) |
| 789 | O-P-O stretching in DNA |
| 911 | Glucose |
| 971 | v(C-C) wagging |
| 1050 | C-O stretching, C-N stretching (protein) |
| 1080 | Ttypical phospholipids, phosphate vibrations, collagen |
| 1124 | *v* (C-C) skeletal of acyl backbone in lipid |
| 1171 | Tyrosine, (CH) phenylalanine |
| 1279 | Amide III (α-helix) |
| 1314 | $CH_3CH_2$ twsting mode of collagen/lipid |
| 1343 | $CH_3$, $CH_2$ wagging mode of collagen, glucose |
| 1542 | Amide II |
| 1563 | Tryptophan |
| 1582 | δ(C=C), phenylalanine |
| 1597 | Amide I band of proteins (C=O stretching) |
| 1620 | *v* (C=C), porphyrin |

The data analysis results on the possible meningitis patients are listed in Table 4. All data was normalized before given to the PCA program and 70% of the data was used for the training, with the remaining 30% used for testing the model. Columns 2-5 in Table 4 list the TP, FN, TN, and FP values assuming that Monocyte is listed as a positive result and Lymphocyte listed as a negative result. However, since labeling a cell species as positive or negative is arbitrary, the same is true for defining a sensitivity or specificity. Thus, the sensitivities are calculated for each species respectively; recall that sensitivity is the ratio of TP over the sum of the TP and FN. Column 6-7 lists the sensitivity values of monocyte and lymphocyte respectively; these values are calculated for all the patient samples used in this study. To compare the effectiveness of the DNN program, the results are compared with the routine hospital report. However, since correctly identifying two white blood cell species does not

convey whether a patient is healthy or not, the ratio of the number of lymphocyte/monocyte cells are compared instead.

For PCA, over-modeling does not occur as it can be thought up as a simple orthogonal transformation. The metric used for the separation of data in PCA space is variance, or the square of the standard deviation. This allows the user to compare data based upon differences and similarities. Applying this to the Raman spectrum, the key Raman peaks selected are used for the whole dataset, and the intensity values are compared and variance calculated. The eigenvalues and eigenvectors are calculated from a matrix of these variances and these two values form the Principal Components (PCs) of the original data. Thus, a PC represents a linear combination of weighted values based on the important Raman peaks selected. For ease of data interpretation, the two highest eigenvalue PCs, or the two PCs with the highest variance, are plotted into a 2D graph with a 68% Normal Probability oval for each dataset. The sensitivities can be calculated from the amount of data nodes that fall into the correct confidence oval. An example of this PCA graph can be seen in Figure 25, where the red data points are from lymphocyte cells and the blue data points are from monocyte cells. The colored ovals represent 80% of the dataset for each species and thus the core significance of each dataset. It can be seen that some lymphocyte data points got misclassified as monocyte and thus are listed as FN. These values can be seen in Table 4. One best usage of PCA was a case where PCA identified some outlier datasets Figure 26, and upon closer inspection found that these cells were misclassified. Once the correct labels were set, the PCA improved in sensitivity.

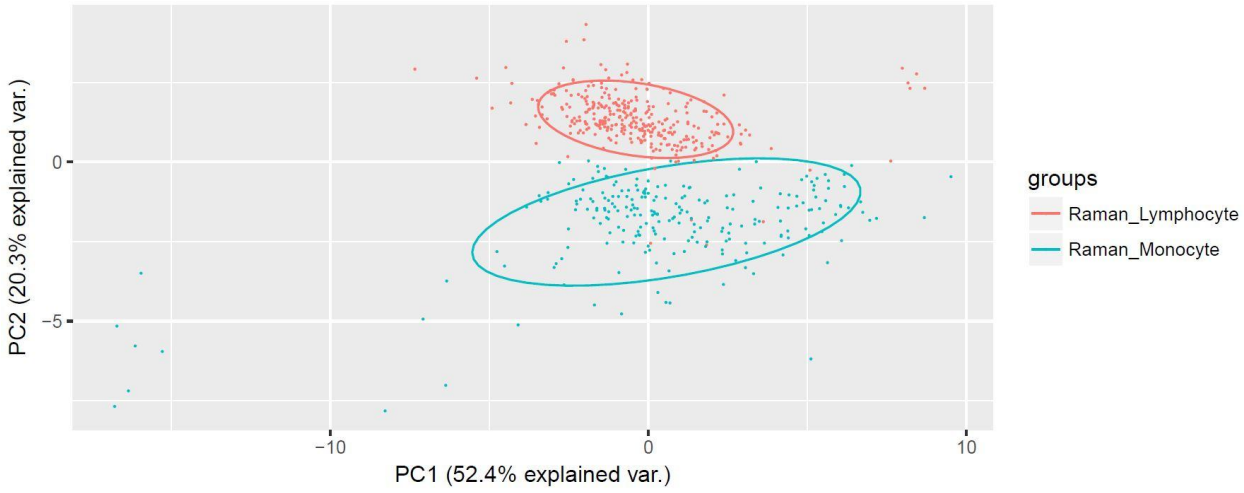



**Figure 25 – PCA graph of the two types of WBC datasets with the red representing lymphocyte and the blue representing monocyte. Clear separation can be seen for these two types with additional statistics of this PCA show in Table 4.**

**Table 4 – Data Analysis of the PCA Graph shown in Figure 25**

| Groups | TP/TN | FP/FN | Dots in Oval | Total in Each Group | Sensitivity % |
|---|---|---|---|---|---|
| Lymphocyte | 263 | 0 | 263 | 321 | 98 |
| Monocyte | 178 | 6 | 184 | 229 | 100 |

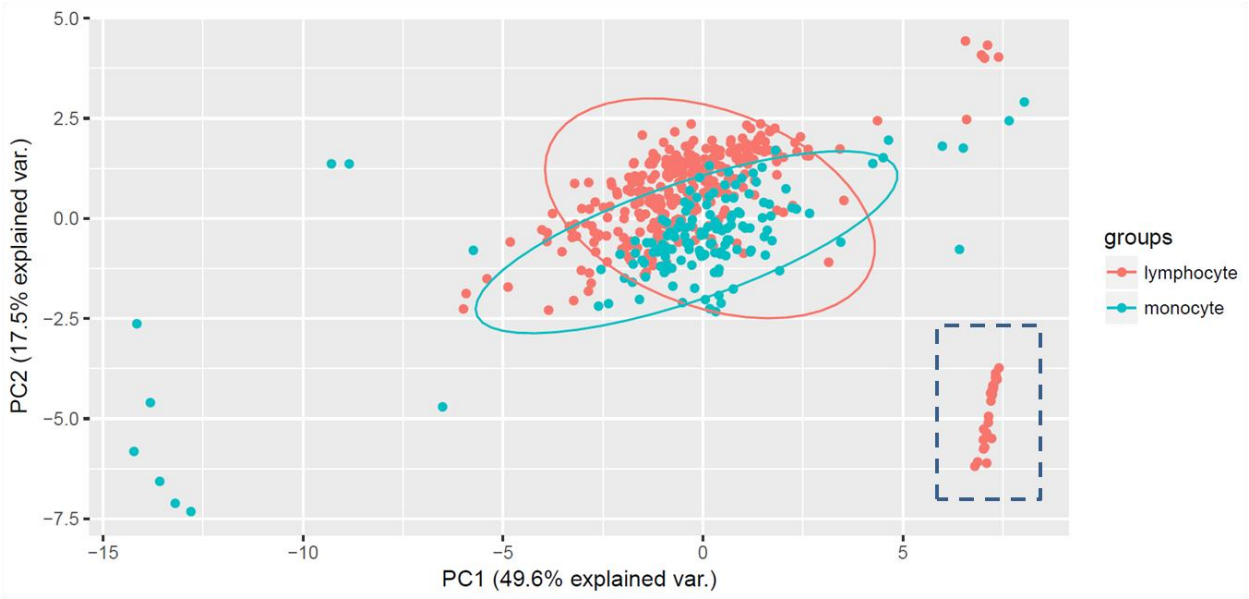


**Figure 26 – An example of a PCA when poorly selected biomarker Raman peaks are used for the PCA. It can be seen here that the two types of WBCs have more overlap. In addition, the dashed box shows an outlier dataset of lymphocyte, but upon further investigate, it was found to be mislabeled by human error.**

Based on the clinical scenario of large patient influx in Xijing Hospital in northwestern China, this study proposes a combination of Raman test, supervised learning and unsupervised learning for data processing, to identify monocyte and lymphocyte in CSF. The pros/cons of the protocols from Xijing Hospital, WHO and Raman method are analyzed. Raman protocol is quick enough and can reduce the burden of labor and avoid artificial errors, in which cells are identified based on lipid peaks at around 2900 $cm^{-1}$ and the types were verified by doctors in this research. Raman measures bonds of molecules vibrational. However, spectra for lymphocyte and monocyte look quite similar after normalized, so special data processing methods are needed. DNN was chosen as supervised learning method for successfully avoiding the inherent bias from the users. For unsupervised learning, PCA was chosen since it's a blind signal separation category with no over-modeling. The separation sensitivities for lymphocyte and monocyte by PCA are 98 % and 100 %, respectively. This Raman protocol is designed to avoid manual errors, with its potential to automate, doctors could liberate from the repeated works such as cell identification and counting.

While these results show off the ability of this Raman protocol to differentiate white cells in CSF, there is still future work to do in applying this Raman protocol to the real-world setting. To enhance the Raman signal and reduce measurement time, a SERS hybrid platform for Raman test needs to be built, which can also increase the sensitivity/specificity. It will also be necessary to investigate the antibiotic resistance proteome via Raman and its application in the clinical setting. This will reduce the need for susceptibility testing which can take days before a result; our technique can do this at the same time as species identification.

## 4.2. Usage of Convolutional Neural Network in WBC Type Differentiation

While the initial goal was to use the SERS platform for meningitis differentiation, it was very quickly determined that SERS differentiation is not practical in the clinical setting. The doctors, upon seeing the power of machine learning, asked if it was possible to leverage our machine learning algorithm for image-based differentiation based purely on their MGG stained slides. This was because the hospital has access to a biobank of a decade worth of MGG stained samples and this large cohort of patient data could be useful in seeing if an image-based machine learning algorithm has the ability to differentiate WBC types for clinical diagnosis. If this is true, then this will have a direct application to the hospitals current diagnosis pipeline.

### 4.2.1. Motivation and Background

The central nervous system (CNS) is one of the most important systems in the human body. One important aspect of the CNS is the cerebral spinal fluid (CSF), which is typically sterile and only contains around 1-5 white blood cells (WBCs) per microliter (μL). This value, however, is only for a healthy individual. When perturbed by an infectious disease, the human body responds by increasing WBC population leading to an inflammation of the CNS. The danger of these CNS infectious diseases is that if not treated within one or two days, the patient will most likely die so immediate medical response is necessary. The global burden of CNS infections in 2016 was tabulated in a recent study [35] and estimated to be 9.4 million incidences with a mortality rate of 5%, or 458,000 deaths annually. With such a high clinical priority and impact, there is always a need for improvement in this field.

The current diagnostic method for CNS infections consists a series of tests, with the gold standard being cell count differentials, culturing, and gram staining. In developing countries, the sensitivity of culturing and gram staining is low [36]. Treatment usually begins at the onset of signs of CNS inflammation, immediately after the cell count and differential cell count become abnormal. This

WBC identification is typically achieved with May-Grüwald Giemsa (MGG) staining of the CSF, which stains the nucleus and granules of the WBCs. In the case of one of the biggest hospital of the northwestern region in China where this study is conducted, the hospital annually treats 120,000 outpatients with neurological diseases and among this, 4000 patients are suspected with CNS infections [37]. Because of this number, the hospital employs a large amount of resources with an estimated 10 working hours per day dedicated just for CSF cell counting, cell staining, and cell identification alone.

Recent years have seen the boon of machine learning for analyzing large datasets and in particular, DNN has been used to help analyze and differentiate red blood cells (RBCs) and WBCs in whole blood [38]–[44]. These studies imply different tactics such as image segmentation, clustering, thresholding, local binary pattern, edge detection, etc. [40]. However, initial implementation of these strategies for this application resulted in low clinical accuracies, thus to accommodate a more generalized model, a generic object detection neural network like region-based convolutional neural network (R-CNN) was explored and found to be more successful [23]. To date, there have not been any studies for WBC differentiation in CSF using any machine learning algorithms to the best of our knowledge.

In this study, the objective is to explore the feasibility of letting DNN to completely replace the currently employed manual labor leading to significant improvement in cell counting accuracy and cost savings. DNN is used in the differentiation of lymphocyte, monocyte, neutrophil, and erythrocytes for CNS inflammation diagnosis. To highlight how DNN accomplishes this, there are three main pillars presented in this study: 1) systematic validation of the DNN to confirm similar quality of care to current standards, 2) analysis of accuracy and precision in automation, and 3) analysis of time savings if applied to the real case. The first section describes the laboratory process and training regimen of the DNN. This includes examples of the optical microscope (OM) pictures of MGG stained cells and description of

the DNN architecture and its implementation for cell analysis. Validation of the DNN software is done by ensuring the required minimum number of cell images for each patient case was inputted and the best complete training of the DNN was objectively analyzed with the lowest saturated value of the global loss function. In the second section, the DNN accuracy is explored by comparing it to the current hospital's gold standard procedure and with blind testing. In addition, the precision of computerized image output is examined to highlight the benefits of automation. The final section of time saving is scrutinized to determine the practicality of the DNN application in clinical settings. The achievements reported here are expected to greatly improve patient care when it comes to diagnosis of infectious CNS diseases.

### 4.2.2. DNN training method for identifying RBCs and WBCs

Images were organized into training and testing folders with a split in the database of 9:1 ratio. Each image was individually labeled with an open-source software called LabelImg [45] by trained technicians and labeled with each cell's classification. The LabelImg also helps to establish spatial locations of each cell by having the user draw boxes in each image. The DNN software is of a region-based convolution neural network so it has great edge detection and it uses the label mapping to separate labeled areas from the non-labeled background areas. Preprocessing scripts were written in Python to organize the data for use in Tensorflow and training was done until the loss function saturated and observed via Tensorboard. Once the newly trained model is frozen, validations were done on the test image folder and compared with the ground truths of the trained technicians. After a reasonable accuracy is achieved, additional unlabeled images were evaluated with the frozen DNN model.

### 4.2.3. Validating the Deep Neural Network

The application of the DNN in this study is in the identification of the 4 main types of cells found in infectious CNS disease patients' CSF. The four main types of cells typically found are lymphocyte,

monocyte, neutrophil, and erythrocytes. The routine procedure in the hospital when a doctor suspects a CNS infection is lumbar puncture and CSF withdrawal from the patient, which will be stained for clear cell identification by the hospital technicians. The MGG staining provides a red acidic stain, a blue basic stain, and a purple color for cellular components. This effectively gives the RBCs a dark grey or red-pink color, the WBCs a blue color with the lymphocyte a distinctive singular round purple nucleus, the monocyte with a large and bean-shaped purple nucleus, and finally the neurotphil with multi-lobed purple-colored nucleus[46]. An example of MGG staining is shown in Figure 27**a**, where all four types of the cells can be seen from one patient.

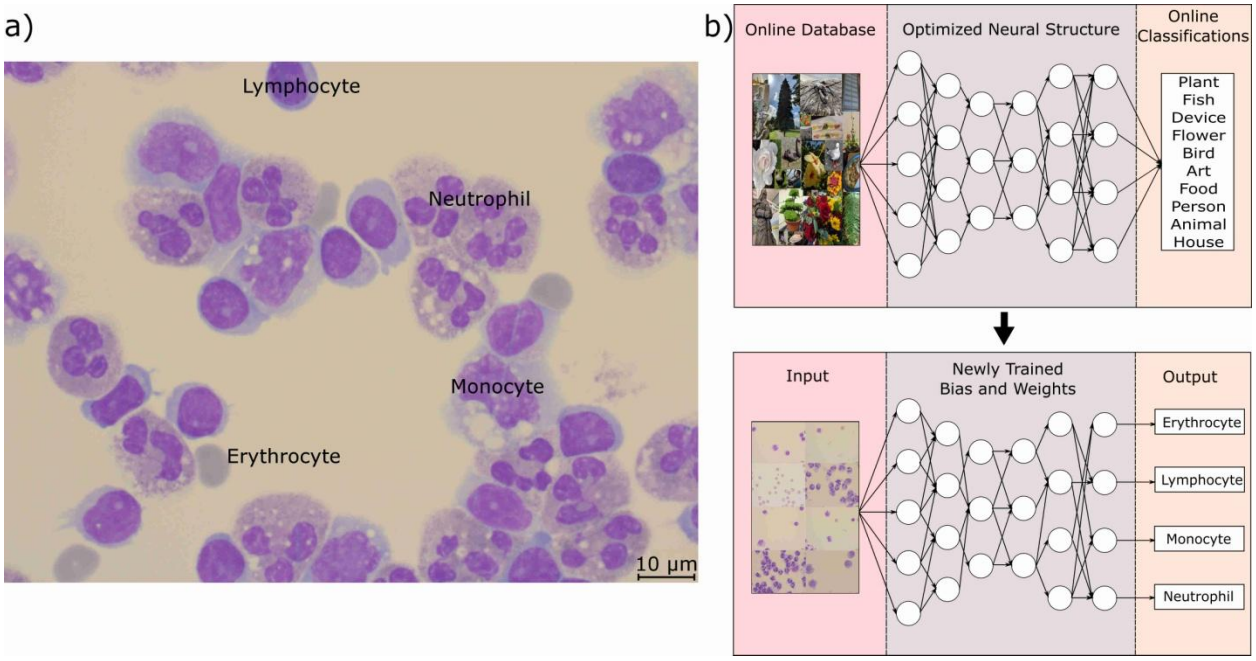



**Figure 27 – a) optical microscope image taken at 100x with scale bar of 10 µm. Cells were fixed and stained with MGG which provides a light color to the cell's cytoplasm and a purple color to the lobes of the nucleus. Labels for the three WBCs and the RBC can be seen in the picture. b) schematic of how the object-detection DNN model is trained to form its basic architecture. The structure along with an online database was used to train the DNN model and then with the basic architecture, the weights and biases are optimized for the MGG-stained cell images of each classification**

The DNN model employed for this study is based on an object detection image-based neural network built on Tensorflow and pre-trained on the COCO dataset [21]. The basics of a neural network can be considered as a repeating algorithm that classifies the importance of an input based on an

activation function. An activation function is similar to the action potential of a human neuron cell, where a necessary stimulus causes the firing of the neuron and this is an all or nothing process. This is analogous to artificial neural networks where the activation function is a mathematical threshold value and once that is met, the result is like the firing of a human neuron. There are additional nuances to this mathematical equation with coupling of weights and bias values, and the resultant firing is not a step function, but a specialized mathematical function containing in-between 0 and 1 activation values; an example is the sigmoid function. However, the main concepts translate to the idea that only the important characteristics of an image will be filtered through this activation function with each of these characteristics being represented as a neuron in one layer of the neural network. The addition of multiple layers gives rise to the non-linearly of a DNN and these features allow a DNN to recognize an image, similar to mimicking the image processing of a human brain. Coupled with the introduction of convolution neural networks (CNN), the processing requirement for image-based neural networks dropped significantly, paving the way for large advancements in the field [22]. However, the detailed description and workings of each of these improvements are beyond the scope of this study and a sample of this literature can be found in Ref. [23]–[25].

The application of the DNN to recognizing WBCs and RBCs was made possible by first applying the pre-trained DNN to a database of OM images labeled by doctors for each cell classification. The specifics of the Faster R-CNN model used can be found in this study [23], [25] and the training on the open-source image database, COCO by Microsoft [21], allowed for a DNN architecture to handle the complexities of the various cell types. As seen in Figure 27b, this pre-trained DNN model has already predetermined the number of layers and neurons are needed for an optimal score of the COCO database and by carrying out a process of transfer learning [47], this model has re-trained itself by adjusting its weights and biases for MGG-stained cell images.

The typical hospital protocol in WBC type classification involves checking around 200 cells per patient. This is known as the cell classification step and it is one of the most time-consuming processes for the hospital. As seen in Table 5, there is a significant amount of patients the hospital handles daily and as such, the hospital has the CSF Cytology Department to devote half-day daily to handle the suspected CSF samples. According to the hospital, the 200 cell minimum is an arbitrary standard set a while ago without much scientific basis, but has not led to failure. As such, an objective study was also done to determine the minimum number of cells needed per patient and also to determine the minimum number of images needed to be taken per patient. Figure 28 shows the result of this focused study where only the three main WBC types are compared with the total number of cells identified per patient. For a typical hospital CSF cytology report, the doctors base their diagnosis on the percentage of these WBCs. From the variety of possible CNS infectious diseases, they were categorized into 5 different cases: 1) low initial cell count (W= 0 - 4), 2) high neutrophil cell count, 3) high RBC count, 4) medium initial cell count (W = 5 – 50), and 5) high initial cell count (W ≥ 50). For Case 1, the low cell count typically means that the CSF of the patient is within the normal range and that the symptoms exhibited by the patient are from a different cause. However, Case 1 also has another difficulty where the entire cytospun sample contains typically less than 200 cells. As seen in Figure 28, the grey curves depict this and the saturation of the curves was not met. For cases 2-5, there are enough cells present and Figure 28 shows that saturation of the curves occurs after 315 cells are labeled. This number was calculated from an average of all the curves and from interpolations between data points after the minimum condition of saturation occurred. The onset of saturation can also be seen around 150 cells, but the error margin of 5% can be calculated.

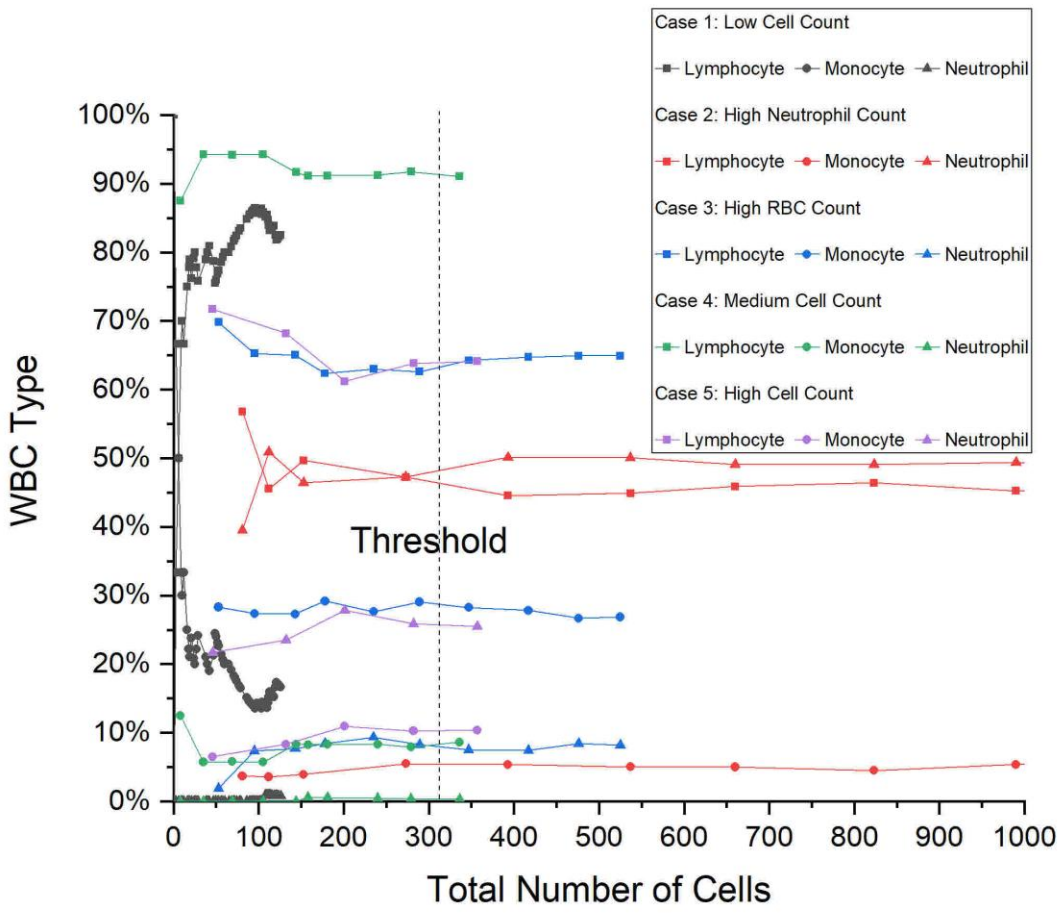
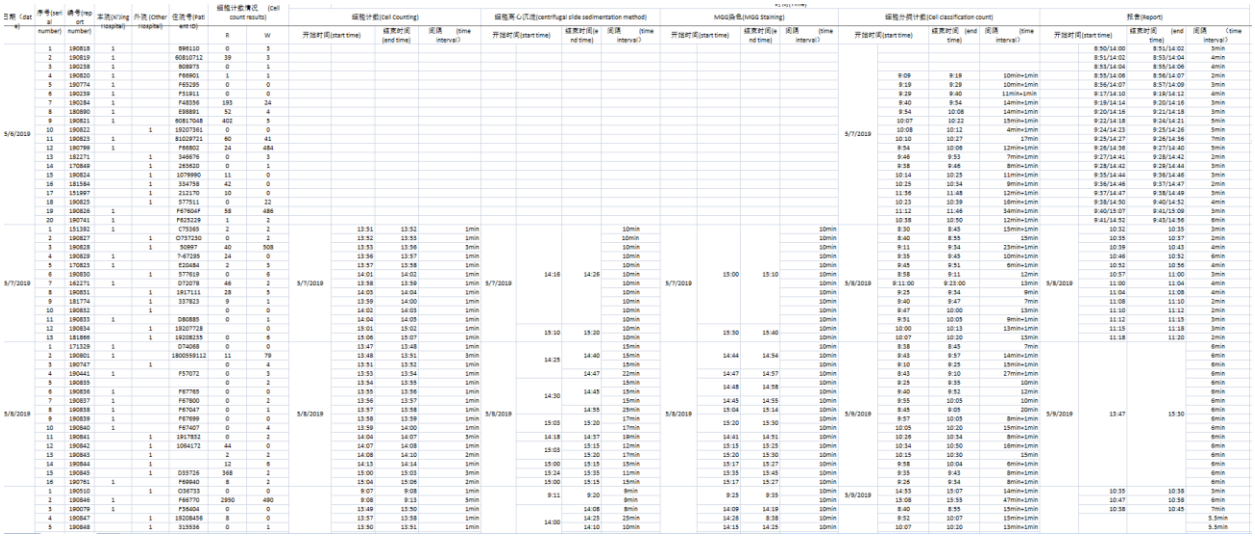



**Figure 28 – A trend graph of the number of cells per patient needed for certain patient condition examples. It can be seen that after a certain amount of cells the percentage of the WBC type saturates thus determining the amount of cells needed for a successful and accurate hospital report. The current hospital protocol requires at least 200 cells to be looked at and this trend graph shows that on average, 315 cells are needed with even the onset of saturation starting at around 150 cells.**

**Table 5 – A collection of the time needed per major step in the MGG staining process**



For the training of the DNN, 100x OM images were taken and every cell in each image was labeled by a trained technician and cross-checked with specialized doctors. For the training process of the DNN, 1300 images, or around 30,000 cells were individually labeled and fed into the program. To verify the effectiveness of the training process, Figure 29 shows the loss value plotted against the number of iterations. The lower the value of the loss function indicates the more fully trained the DNN model has become to an absolute limit of 0, which indicates that the model is perfectly trained. Generally, all DNN models are given trained values with a certain amount of noise, or in this case a variety of images of different situations, so that the DNN can have flexibility and not be over-fitted to a degree that it cannot identify images not perfectly matching its initial training dataset. Figure 29 shows the output loss values in grey along with a moving average for a better visual representation of the graph. An exponential decay function is also fitted to highlight the saturation of the loss function. The training of this DNN took around 20,000 iterations and around 2.5 days. However, once a DNN is trained, it only needs around 7 seconds for an output.

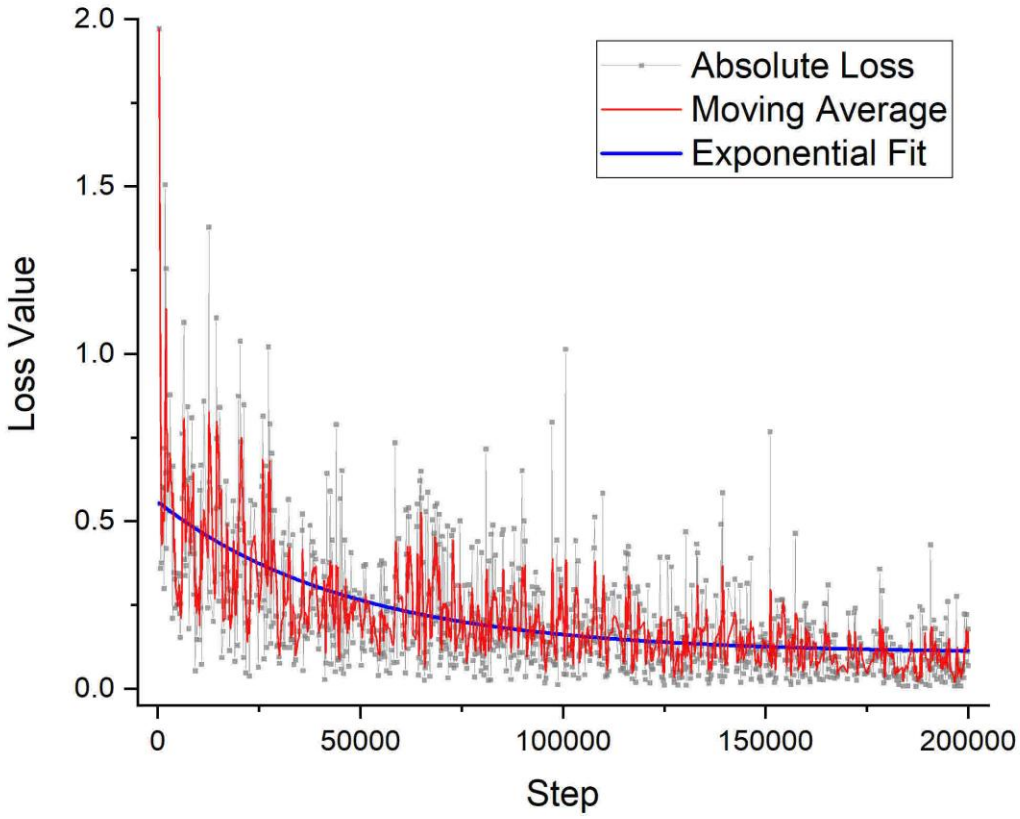



**Figure 29– The DNN model's training accuracy showing its precision vs. number of iteration steps. As can be seen that the graph takes on a 1/x, asymptotic relationship with saturation quickly established within the first few thousand steps. After 20,000 iterations, the precision % has not improved that much and the training of the model stopped, which took around 2 days of nonstop training.**

### 4.2.4. Precision of the Neural Network

Besides merely relying on the loss function plot, a cross-check of the validation was performed to verify the accuracy of the DNN model. A certain portion of the image dataset was kept from training as the testing validation set and the ratio amount chosen was 9:1. For a comparison, four trained technicians were also arranged to label the same validation dataset and then their results were compared with the DNN's prediction. Table 6 shows the labeling results of the validation dataset comparing the variations between the human labeling and the DNN's labeling. The immediate take-away is the confirmation that the multiple evaluation rounds of the DNN will produce the same result,

however, that is not always the case. As seen in Figure 30, there is a possibility for the DNN within the same validation round and with the same model version to produce two different image labeling outputs. In this case, the four validation rounds did not produce any variations. The other interesting factor comes from the human side with standard deviation amongst the technicians producing large variability. However such inaccuracy is suitable in the clinical setting where speed is more important and the WBC typing percentage can have a swing of +/- 10% as the MGG cell classification report is only one of the many diagnosis tests typically done in series on a patient's CSF. This further shows the importance of implementing the AI in cell classification to improve the accuracy of the clinical results to reduce the reliance of subsequent tests in aiding the doctor's diagnosis.

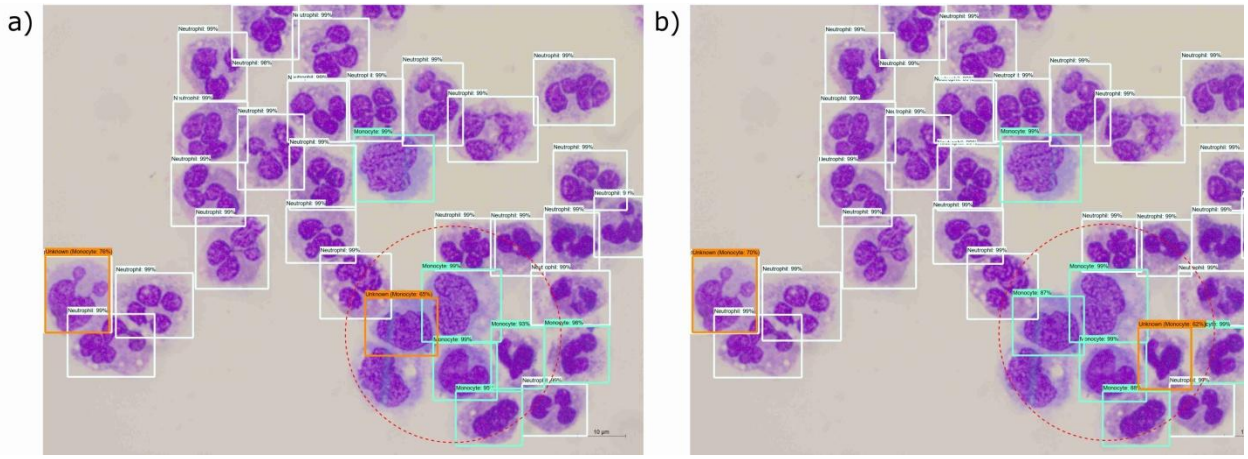



**Figure 30 – Two DNN output images showing the inconsistency of a model in determining cell classifications. These two images happened to be run during the same evaluation round and same model, yet the model still produced two different results. The red circle in a) and b) shows the difficulty in the model in identifying the two different monocyte cells as noted by the orange "unknown" label.**

**Table 6 – The number of cells labeled in the validation dataset between human and AI for each cell type**

| | Erythrocyte | Lymphocyte | Monocyte | Neutrophil |
|---|---|---|---|---|
| Human | | | | |
| Person 1 | 13 | 66 | 67 | 41 |
| Person 2 | 12 | 43 | 48 | 40 |
| Person 3 | 20 | 77 | 70 | 43 |
| Person 4 | 27 | 80 | 73 | 54 |

| | | | | |
|---|---|---|---|---|
| Human Std Dev | 7 | 17 | 11 | 6 |
| DNN | | | | |
| Round 1 | 28 | 77 | 66 | 50 |
| Round 2 | 28 | 77 | 66 | 50 |
| Round 3 | 28 | 77 | 66 | 50 |
| Round 4 | 28 | 77 | 66 | 50 |
| AI Std Dev | 0 | 0 | 0 | 0 |

There are two outputs of the DNN program: 1) a labeled image with each DNN-recognized cell boxed with its prediction percentage and 2) a report with the statistics of the recently run evaluation. An example of the output image can be seen in Figure 31 where the four major cell types are labeled by the DNN program. The program puts a predicted boxed area around the target cell and then gives each cell a classification prediction percentage. If that percentage falls under 80%, then the program will instead add another orange box over the original label and give it the label “unknown” so that a human technician can manually check the cell. Also coupled into the program is the label “unknown” with a light tan color to account for the more rare cell types (lymphoid, mitotic, basophil, etc.) and these will require the human technicians to check them as well. While the spatial location is information not currently being used for the hospital diagnosis report, the percentage of WBC types is important for diagnosis and the program calculates and outputs a statistical report of the three major WBC types.

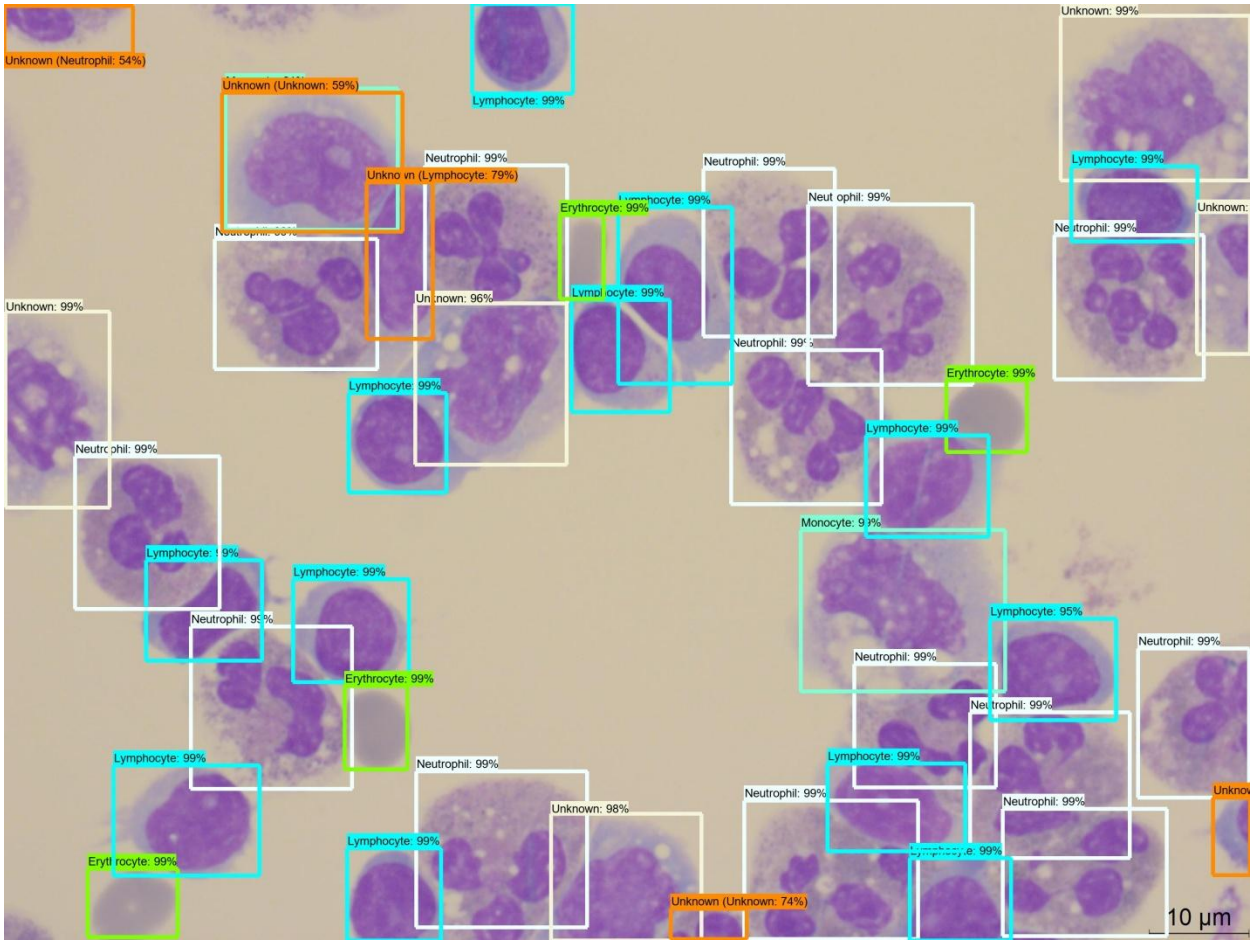



**Figure 31 – An example output of the DNN model with boxed labels along with the model's percentage prediction. One can see the predicted outputs of neutrophil, monocyte, lymphocyte, and erythrocyte with their respective colors along with the DNN model's percentage prediction. In addition, some cells are labeled with the "unknown" label tag (tan and orange boxes) when the prediction percentage is below 80% or when the shape of the cell indicates a possibility of a rare cell type (i.e. basophil, eosinophil, mitotic, etc).**

### 4.2.5. Accuracy of the Neural Network

To determine the effectiveness of the DNN in a real-world application setting, a blind test was performed and the comparison can be found in Table 7. During the blind test, the images were taken by operators without knowledge of the hospital report and given to a DNN operator, without any patient information except their ID number. The ID number is scrambled with the key being kept by a third party. From Table 7, the average differences show that the DNN model is fairly accurate when compared with the hospital report with the largest margin of error in cell classification with neutrophil

and the largest patient variability with Patient #1. Overall, the DNN was able to handle the various infectious disease cases presented to it: 1) high neutrophil count, 2) high RBC count, 3) even distribution of WBC types, and 4) high lymphocyte count. The average accuracy of this DNN for these three WBC types is 95%. Compared to similar studies done on whole blood, our result is on similar levels of accuracy [40]–[42], [44]. Upon closer inspection of the data discrepancy for Patient #1, it was found that the DNN had not previously encountered abnormal neutrophil images during its training phase. These abnormal neutrophil pictures had the individual nuclei lobes clustered together into a similar shape of the monocyte nuclei producing a false negative result; an example of this can be seen in Figure 32. These misclassifications led to the uneven monocyte/neutrophil percentage and thusly incorrect report. To better apply the DNN for future clinical situations, the training regime will have more of an emphasis on number of patients trained rather than the number of cells trained for each cell classification to account for the complex clinical patient situations.

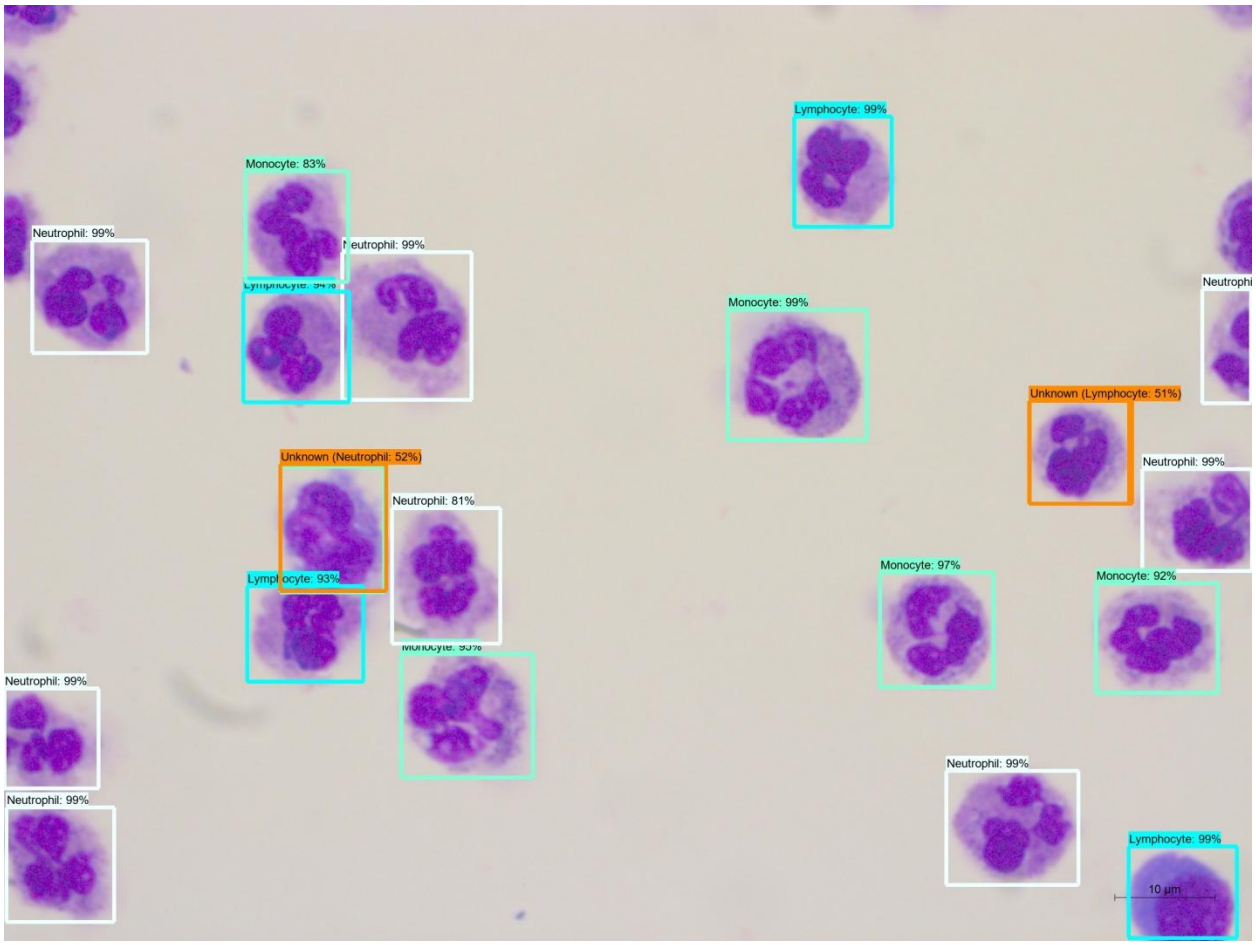



**Figure 32 – DNN output of the blind test with Patient 1 (ID# 190931). The mislabeling of the neutrophils can be seen in this image with the only true lymphocyte cell classification in the bottom right; the rest of the image should be all neutrophil classifications but the various granulate aggregation confuses the DNN to label these cells as lymphocyte or monocyte instead of neutrophil.**

**Table 7 – Blind testing results of the DNN vs the Hospital Diagnosis Report**

| | | % Lymphocyte | % Monocyte | % Neutrophil |
|---|---|---|---|---|
| Hospital Technician | | | | |
| | Patient 1 (ID# 190931) | 5 | 3 | 92 |
| | Patient 2 (ID# 191155) | 1 | 7 | 91 |
| | Patient 3 (ID# 191172) | 18 | 23 | 59 |
| | Patient 4 (ID# 191158) | 87 | 6 | 7 |
| DNN | | | | |
| | Patient 1 (ID# 190931) | 7 | 18 | 75 |

| | | | |
|---|---|---|---|
| Patient 2 (ID# 191155) | 4 | 7 | 90 |
| Patient 3 (ID# 191172) | 27 | 20 | 53 |
| Patient 4 (ID# 191158) | 89 | 8 | 3 |
| Comparison Between Human vs AI | | | |
| Patient 1 (ID# 190931) | 2 | 16 | 16 |
| Patient 2 (ID# 191155) | 3 | 0 | 1 |
| Patient 3 (ID# 191172) | 9 | 3 | 6 |
| Patient 4 (ID# 191158) | 2 | 2 | 3 |
| Average Difference | 4 | 5 | 7 |

### 4.2.6. Time Saving Potential

One of the main advantages of using the DNN program to replace the mundane task of cell type labeling is the time savings for the doctors so that their attention can be more focused on other tasks. To quantify this time savings, a short survey was conducted during a working week to estimate the time committed on per patient and daily basis. An example of the complete survey can be found in Table 5. Table 8 shows the time needed by the hospital personnel for the two time saving procedures that the DNN can contribute: 1) cell classification and 2) report writing. As seen in Table 8, the DNN can save around 16 minutes per patient and around 4 hours per day; this amounts to a doctor time reduction of 86%, daily. The DNN time was calculated from the validation dataset and extrapolated with average number of patients from the short survey. The minimum number of cells per patient, extrapolated from Figure 29, and the average number of cells per image were also factors used. In addition, the DNN processing time needed per image was also found to be independent to the number of cells present, with processing time slowing down as heat became more difficult to dissipate from the machine.

**Table 8 – The time saving potential when compared between the DNN and hospital technician**

| | Average Time Per Day (mins) | Average Time Per Patient (mins) | % |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Hospital Technician | | | |
| Cell Classification | 211 ± 25.3 | 13.4 ± 0.86 | N/A |
| Report Writing | 70 ± 13.4 | 4.4 ± 0.20 | N/A |
| Total Time | 281 ± 38.5 | 17.8 ± 0.92 | N/A |
| DNN | | | |
| Cell Classification | 34 ± 4.5 | 2.2 ± 0.04 | N/A |
| Report Writing | 3 ± 0.3 | 0.2 ± 0.00 | N/A |
| Total Time | 37 ± 4.8 | 2.4 ± 0.04 | N/A |
| Time Saved | 243 ± 38.8 | 15.5 ± 0.92 | 86 ± 4 |

### 4.2.7. Conclusion

This study presents a pioneering application of image-based DNNs to patient samples in clinical setting. Image analysis of MGG-stained patient samples is analyzed for CSF cytology. By applying the neural network technology to the clinical space of cell type classification, significant saving in time has been achieved. The daily saving in the time spent counting cells of hospital technician is estimated to be approximately 86% ±4%. DNN further rendered more consistent analyses capability against the large variability common to human classification analyses. Blind tests result in an average accuracy of 95% among the three WBC types, with the addendum being that the program's accuracy can always be improved further with additional training from a wider variety of patients. This report demonstrates clearly the promise of DNN in clinical practices pertaining to infectious diseases of central nervous systems.

## 4.3.MDR Tuberculosis Differentiation in Pakistan

Another important case study on the application of SERS is the rapid identification of Mycobacterium tuberculosis (MTB) in developing countries such as Pakistan. Tuberculosis (TB) is a disease that is one of the top 10 leading causes of death, worldwide, and also the leading cause of death

from a single infectious agent [48]. While TB affects all countries and age groups, it is predominately an issue in developing countries where the top five most burdened countries are India, China, Indonesia, Philippines, and Pakistan [48]. Below in Figure 33 is the worldwide incidence cases of TB in 2017 where the area of the circles represent countries that are most affected by TB.

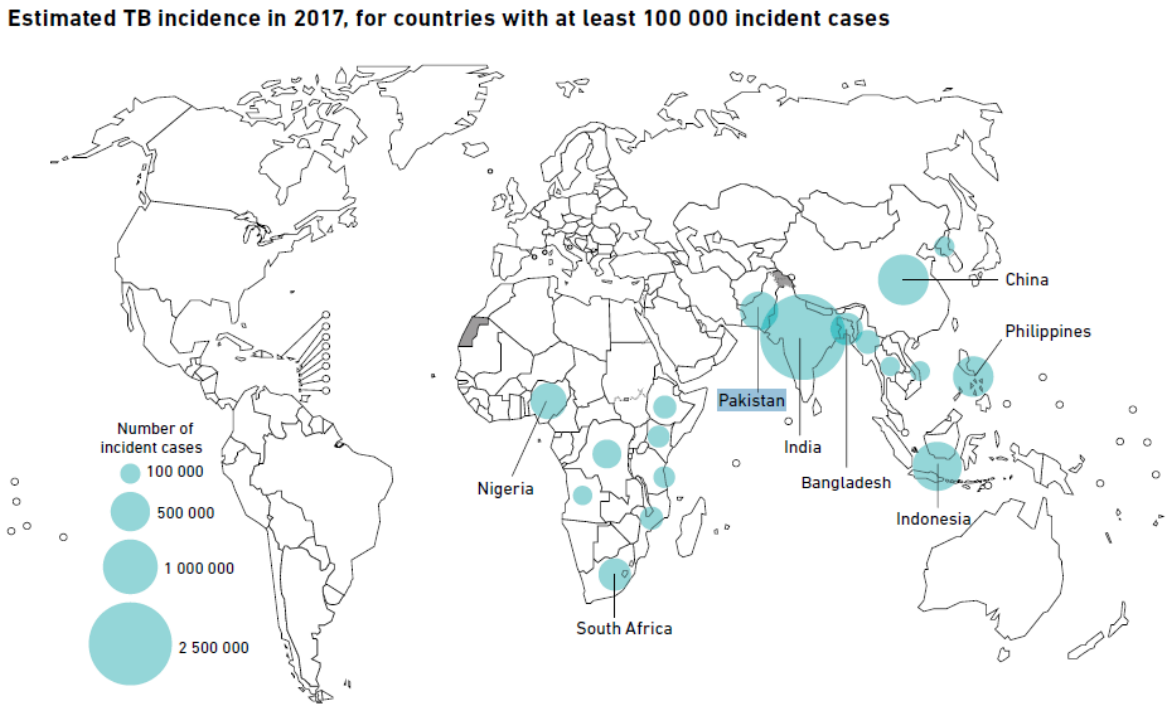



**Figure 33 – Estimated TB incidence worldwide in 2017** [48]. **Pakistan was rated one of the top ten countries greatly affected.**

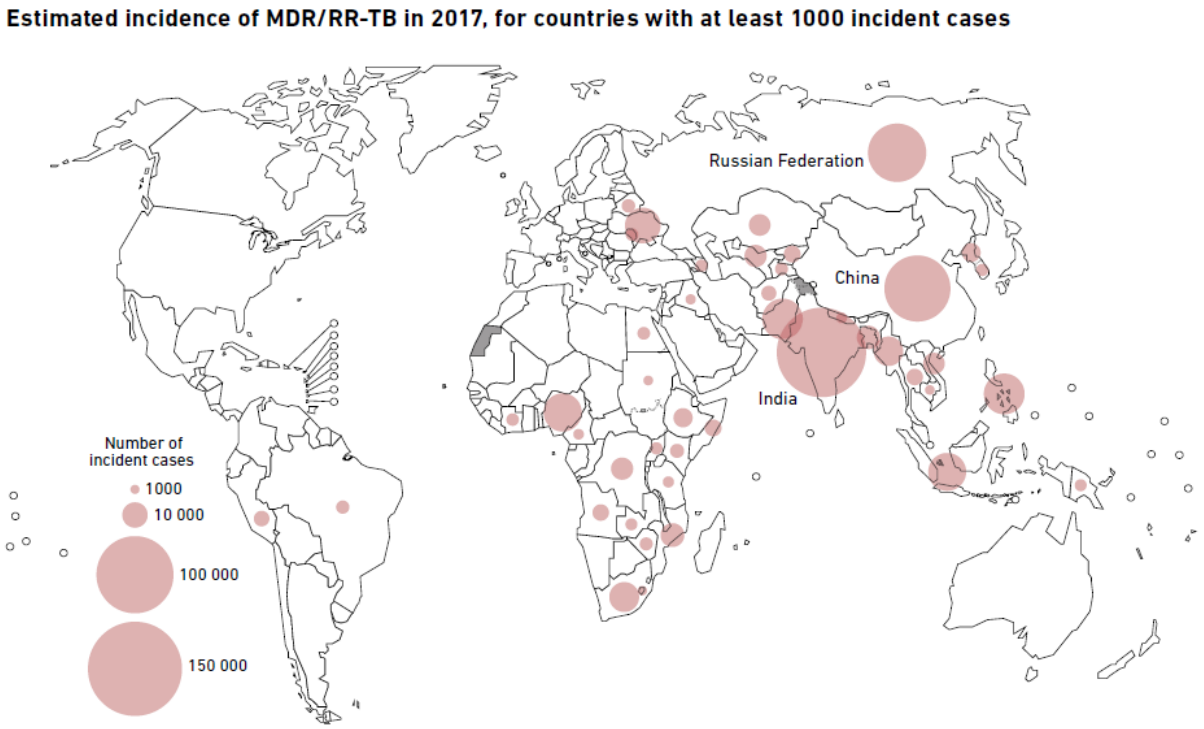



**Figure 34 – Estimated worldwide incidence of MDR MTB in 2017** [48]**. Once again, this shows the heavy burden of the disease in Pakistan as one of the top countries with a heavy burden.**

### 4.3.1. Pulmonary TB and Drug Resistance

Within the disease of TB, there can be complications such as HIV/AIDs and extrapulmonary TB, e.g. there are situations when a patient gets TB in their bones. However, the predominatelyz TB is in the lungs and is called pulmonary TB (henceforth will just be referred to as TB in this dissertation). As the name suggests, patients spread the disease through the air via coughing, sneezing and is transmitted when MTB is contained in the saliva bulbs. The MTB can survive for several days outside the human body and when it settles inside the host body, and also if successful in overwhelming the patient's immune system, the bacterium will grow itself new colonies.

### 4.3.2. Current Gold Standard – DST and GeneXpert

The WHO gold standard for diagnosing TB is through culturing of the MTB, however such a process is quite lengthy. This is because the MTB's doubling time is in the order of days thus it can take 4-6 weeks for a complete culture and strain identification. In recent years though, the WHO has made an effort to establish real-time polymerase chain reaction (rt-PCR) technology as a possible alternative. Specifically, there is a company that has a product called GeneXpert, and that is the machine the WHO champion and uses in the various Pakistani hospitals. However, one big weakness of the GeneXpert machine is that it can only detect antibiotic resistance for one drug, the main one rifampicin (RIF), even though there are a total of 4-5 main drugs as the 1$^{st}$ line of defense. Yet, conversations with the doctors in the Pakistani hospital say the GeneXpert is quite useful with same day turnaround and the underlying assumption is that if a patient has RIF-resistant MTB, then most likely the patient also has multi-drug resistance (MDR). Thus our group set out to meet this unmet need by providing the same rapid detection, when compared to the typical gold standard of culturing MTB strains, and the ability to measure all the antibiotic resistant strains, since we are not limited by the primers of rt-PCR.

### 4.3.3. SERS Feasibility of MDR Strain Separation

Building off the lessons of the Chinese collaboration for meningitis, the focus of this Pakistani collaboration was to build the project from the ground-up with the clinician's needs and usefulness, as well as balancing the successfulness of this project. From that, it was established that since this will be our platform's first handling of bacteria from patient samples, and the SERS measurements will be done by our Pakistani collaborator at Government College

University Faisalabad, GCUF, we did not shoot for the ultimate goal but an in-between to verify our success first before continuing. Thus, the scope of the project was reduced to a more manageable goal of using the SERS platform on cultured MTB isolated from MDR patients and simple TB patients, these are the control patients that have the regular TB. Culture was started early with a large focus on getting the MDR patient samples from the TB-specific hospitals in the rural regions. Meanwhile, I worked with my collaborators remotely through video conferencing and remote desktop control to calibrate and optimize their SERS measurement protocol, based on the lessons I learned from the previous studies.

#### *4.3.3.1. Typical SERS Spectra*

Measuring the MDR MTB species is very similar to previous studies; at this point it is not about improving the SERS protocol but learning the ways to optimize the protocol for the current application. Figure 35 shows a typical good SNR SERS spectrum of a MDR MTB isolated from a patient. Figure 36 is a comparison of the two SERS spectra when comparing the MDR samples with the simple TB control samples. It can be seen that there are minute differences between the two spectra and thusly a ML algorithm is needed to separate the differences.

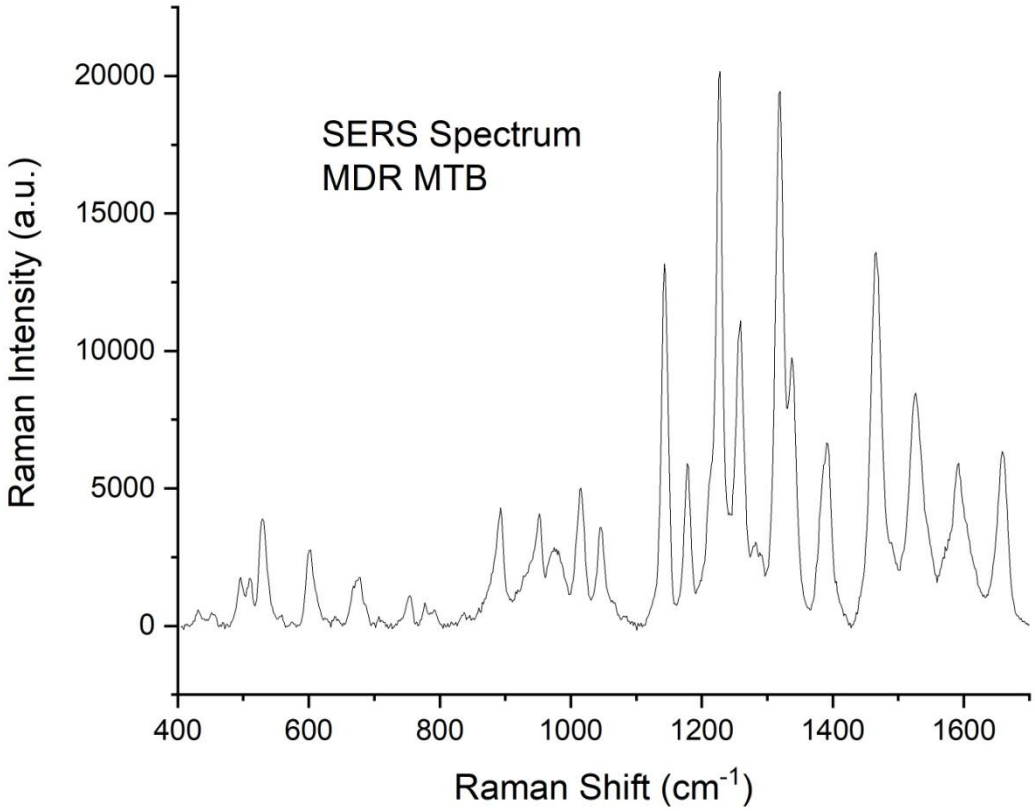



**Figure 35** – Example of the MDR MTB SERS spectra with high signal to noise ratio (SNR). We established a pre-processing filter to grab SERS spectra of this quality for the ML algorithm

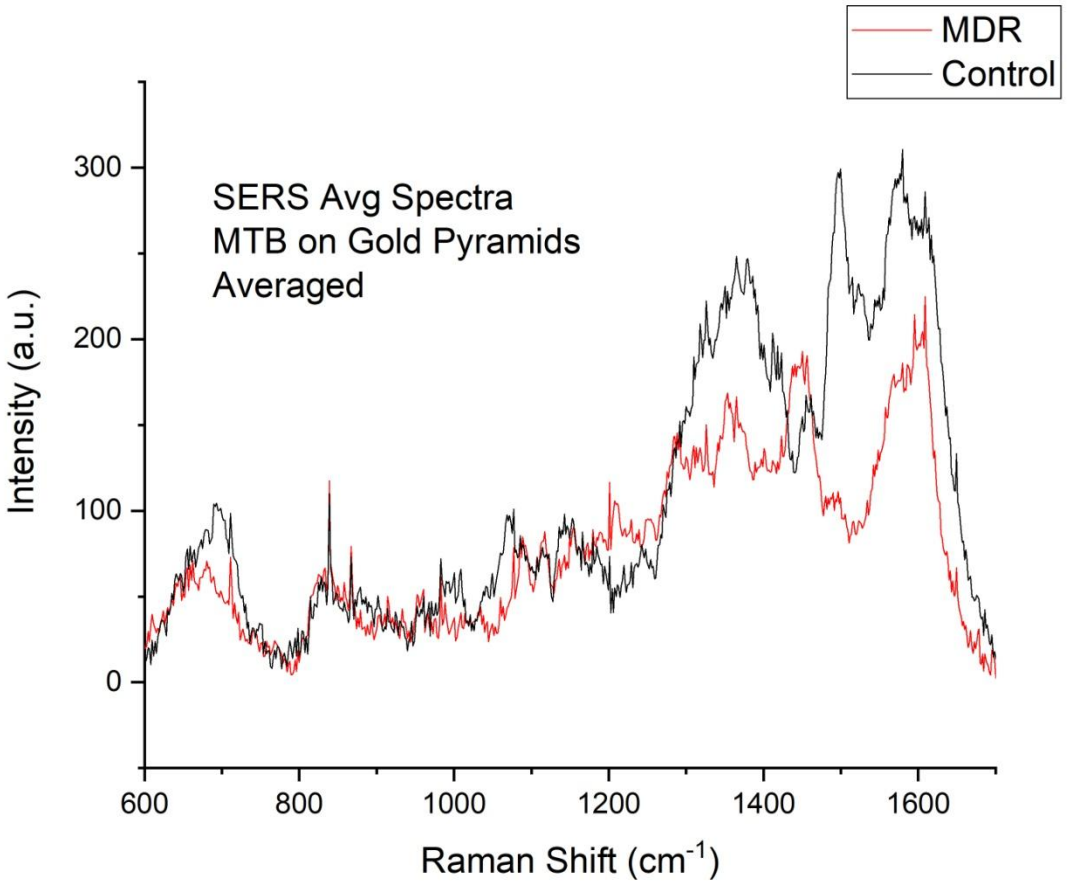



**Figure 36 – SERS spectra comparing the MDR and Simple TB. It can be seen that there are minute differences between each spectra and thusly needing ML to differentiate between the two.**

#### *4.3.3.2. DNN Results*

For the data analysis, we employed a deep neural network (DNN) to analyze the differences between the two types of samples. In total, 16 patients samples of MDR MTB were analyzed and 6 simple TB patients samples were also measured. We initially planned for more of the control samples to be measured to balance out the ratio between MDR and control but the COVID-19 pandemic hit around that time and all our experiments had to stop. Thankfully, we tackled the most difficult part of the study by measuring the MDR samples first so even if the balance is not equal, according to literature [49][50], a 3:1 ratio of patients to control is an acceptable number. Table 9 shows the result of the DNN among

the dataset with an average of 79% sensitivity and 52% specificity after 10 evaluation rounds. Adaboost was also done and showed similar results (not shown here). The methodologies of both ML algorithms differ although the result ended up the same; it was our attempt to try and improve the specificity by using a different analysis method. Upon closer review of the data, we concluded that the sample dataset for the simple TB was not sufficient enough to build a stable SERS fingerprint. It is unfortunate due to the pandemic that all hospitals in Pakistan are focused on COVID-19 and TB has been pushed to a lower priority so we are not able to get any more patient samples. Future iterations of this project will need more patient sample measurements to improve accuracy of this platform for TB.

**Table 9 – DNN results of MDR TB vs Simple TB. 10 cross-validation rounds were done and the average was taken.**

| | TP | FP | TN | FN | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Round1 | 236 | 28 | 27 | 42 | 85% | 49% |
| Round2 | 212 | 24 | 31 | 66 | 76% | 56% |
| Round3 | 209 | 20 | 35 | 69 | 75% | 64% |
| Round4 | 230 | 23 | 32 | 48 | 83% | 58% |
| Round5 | 238 | 30 | 25 | 40 | 86% | 45% |
| Round6 | 245 | 34 | 21 | 33 | 88% | 38% |
| Round7 | 197 | 32 | 23 | 81 | 71% | 42% |
| Round8 | 200 | 24 | 31 | 78 | 72% | 56% |
| Round9 | 230 | 27 | 28 | 48 | 83% | 51% |
| Round10 | 211 | 20 | 35 | 67 | 76% | 64% |
| **Average** | **221** | **26** | **29** | **57** | **79%** | **52%** |
| **Std Dev** | **17** | **5** | **5** | **17** | **6%** | **9%** |

# 5. Challenges Faced – Lessons to Bridge the Gap Between Laboratory and Clinic?

The last few sections showed great advancement of Surface Enhanced Raman Spectroscopy (SERS) in research laboratory and in clinical applications. However, such advancements were not without pain and were done through many trial and errors. The following subsections are key lessons we learned from this PhD journey and will hopefully enlighten the next researcher on potential pitfalls and make the translation to clinical applications much quicker.

## 5.1. SERS Substrate Fabrication

While the SERS fabrication technique is widely know with many in fabrication articles published in literature, the common theme is to use chemistry and fabricate metallic nanoparticles [51][52]. Although this is the most popular method in the SERS platform, it comes with reproducibility issues as this bottom-up approach rely on the consistency and regularity of the chemicals involved. Our group, coming from a semiconducting fabrication lab, looked at this problem and approached it from a top-down approach opting for a more controllable and repeatable process [19]. The process we finalized on utilizes top-down lithography so pattern creation is very controllable and repeatable. The one human element in the whole process is near the beginning where polystyrene spheres self-assemble to form a highly packed pattern, yet this whole procedure before my PhD was all done by hand. To remove this liability, we looked into the possibility of using a Langmuir Blodgett (LB) trough for a controllable polystyrene sphere packing. A LB trough is an equipment that specializes in packing hydrophobic or hydrophilic materials into a dense film on top of a thin pool of water. It accomplishes this via two moving horizontal bars that push the materials (polystyrene nanospheres in our case) towards each other and the target wafer substrate retracts vertically as the horizontal bars draw close to each other. The speed of the withdrawal of the target substrate and the speed that the horizontal bars draw close to

each other are determined by another submerged metallic plate, called a Wilhelmy plate, which measures the surface pressure of the water. This automated process, guarantees repeatable and easily reproducible polystyrene sphere packing for high density nanopyramid structures.

The reason why we went through such lengths in ensuring the reproducibility of our nanopyramids is because one of the biggest hurdles for a SERS platform's integration in real-life applications is the variability in SERS substrates. Most SERS papers deal with showing the high electromagnetic enhancement possible but these scenarios only occur when the gold nanoparticles are practically touching one another. As such, there is great variability in ensuring that nanoparticles pack as closely as possible and to ensure that the chemicals produce nanoparticles of equal sizes; otherwise different sized nanoparticles also produce different SERS hotspot volumes and that's another level of uncontrollability. Thusly, the switch to a top-down method, where our nanopyramids can be predictably located throughout the substrate, ensures equal areas of high-density nanofeatures for consistent SERS measurement. This is important as when we introduce our target analyte onto the substrate, we do not have a great deal of control in the final location of the cell on our substrate so consistent SERS hotspot density throughout the chip removes the chip as the failure point when troubleshooting. An image of the LB trough can be seen in Figure 37. When the densely packed polystyrene spheres are done by hand, a densely packed picture such as Figure 38 is shown, while the pattern when done by the LB machine is similar to Figure 39.

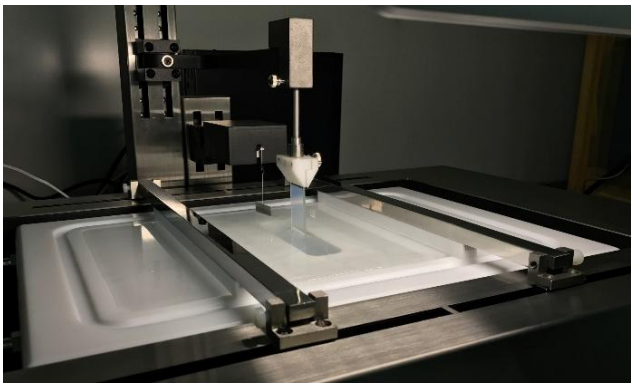


**Figure 37 – Image of the Langmuir Blodgett (LB) trough setup at Xi'an Jiaotong University. The horizontal bars help push the film together as the center rod raises to pull the polystyrene nanospheres into a closed packed density film on top of the target substrate. The Wilhelmy plate can be seen in the back behind the white holder attached to a force sensor. This whole process is automated by the computer with parameters set by the user.**

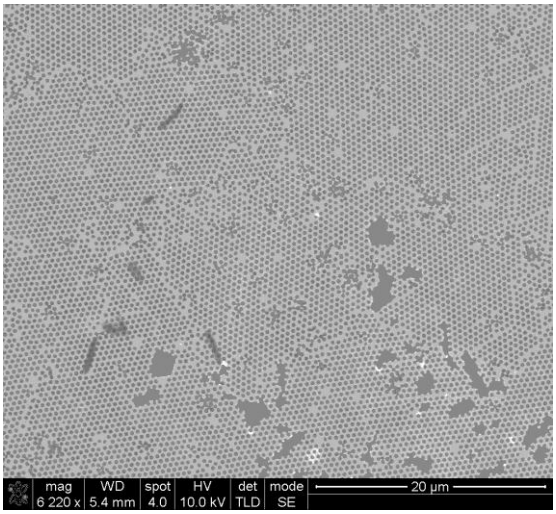



**Figure 38 – SEM of the close packing of the polystyrene sphere when done by hand. As can be seen, not every area is covered by the polystyrene sphere and hence became an art-form with variation of quality between students. It can also be seen that there are different regions of different grain orientation.**

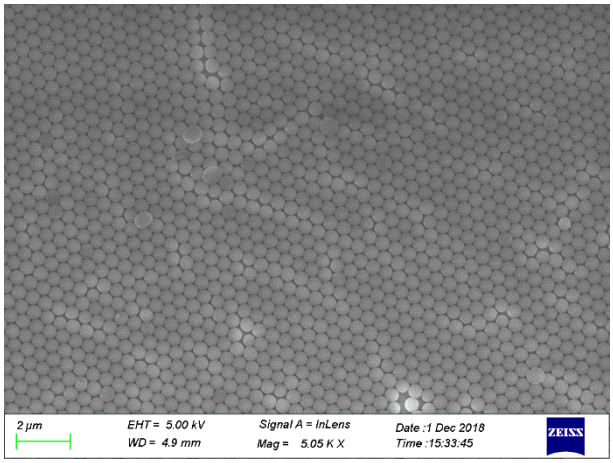



**Figure 39 – SEM image of closely packed polystyrene spheres done with the LB trough machine. As can be seen, the LB film has the same high density packing as the hand pulled substrate but the density areal coverage is far greater. Although orientational grains do exist (not depicted here), their areal size is far greater.**

## 5.2. Building a Repeatable SERS Protocol

### 5.2.1. Necessity of Graphene

As previously mentioned in Chapter 2, graphene serves multiple purposes and thusly is an essential part of the platform. While technically, SERS does not require this single atomic carbon layer for signal enhancement, graphene serves various other purposes to help eliminate the confounding factors during troubleshooting. Some of these include normalization of the datasets for direct comparison, additional chemical enhancement, and substrate calibration and quality check. An example of graphene's usage can be seen in the breast cancer cell line study in Chapter 3 as depicted in Figure 40. It can be seen in that figure that the bio-peak in that breast cancer cell happened to have a large SERS enhancement and only through the usage of graphene, can this spectrum be normalized to not render the rest of the mapping unusable.

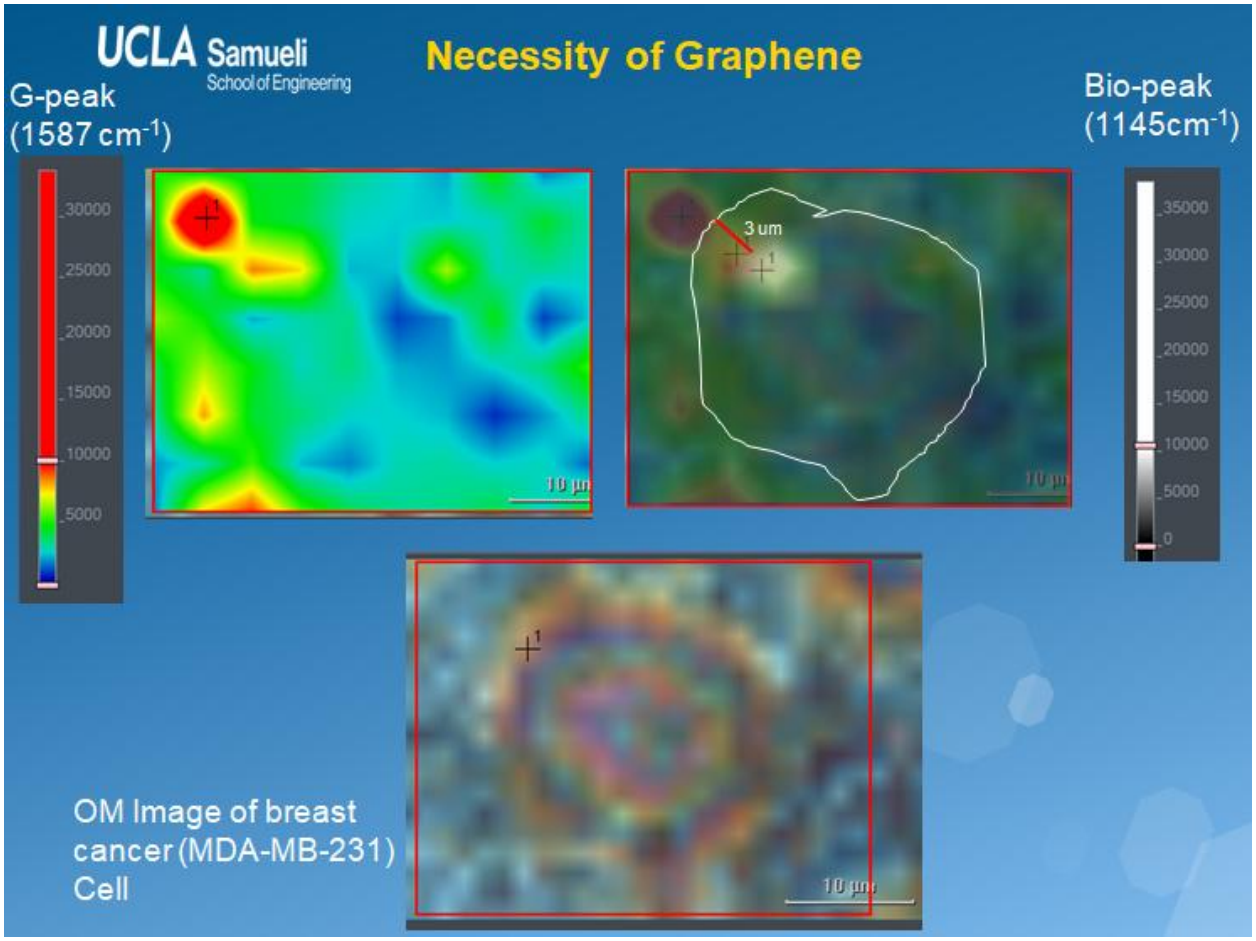



**Figure 40 – A composition of a Raman mapping of the breast cancer cell line MDA-MB-231. The bottom insert shows the OM image through the Raman microscope and the rainbow coloring that one observes when a cell is dried onto the SERS chip. Based on this, an outline of the cell can be seen as shown in white depicted in the top-right insert. The black crosshair labeled "1" shows a Raman point in this Raman mapping where the selected Raman peak at 1145 cm-1 is greatly enhanced if not normalized, will overpower the signal of the rest of the cell. The top-left insert shows the graphene enhancement and one can clearly see that this particular area happened to have a large hotspot enhancement. This shows that graphene is necessary for unbiased normalization unrelated to the cell proteome change.**

### 5.2.1. Signal to Noise Ratio

Another necessary step for a consistent SERS protocol is establishing a high quality SERS chip through the usage of graphene. The previous section explained graphene's usage in normalization of the dataset, but graphene is also needed to establish a quality control of the SERS substrate. In the beginning of the cell line studies, there was inconsistency in getting good signal to noise ratio (SNR) Raman spectra and it was unknown at the time if the cell was the issue, the pre-processing of the cell sample, the SERS substrate, or user error. To eliminate the SERS substrate as a variable, calibrating each

SERS substrate with large coarse SERS mapping throughout the chip and measuring the G and 2D peak for SERS enhancement calibration was done. It was found that for good differentiable SERS signal, the SNR of the G-peak needs to be at least 3.483 and the 2D-peak needs to be at least 6.030. Only with these SNR thresholds, will the SERS chip be used for cell proteome measurement.

## 5.3. PBS Salt and Protein Crystallization

One unexpected lesson learned is the presence of crystallization when the target analyte solution is allowed to dry on the SERS substrate. This is because the solution is also comprised of PBS salt and proteins from a patient's biofluid and upon evaporation of the solution's liquid, the salt ions crystallize with the proteins to form beautiful dendritic structures. However, this is not a new phenomena and is well documented in literature [53]. In fact, part of the literature search also confirmed an observation of ours that the liquid droplet gets pinned in its circular shape before completely drying, this is known in literature as the coffee ring stain [54][55]. What happens is that the particles in the solution, in this case salt ions and proteins, get pushed to the edge of the droplet and form a supersaturated concentration which pins the shape of the droplet or what the literature colloquially coins as a coffee ring stain. It is at this point, complex phenomena of evaporation, diffusion, and capillary force continuously push particles to the edge causing a buildup of material and making a salt/protein droplet wall. Continuous buildup of this material will cause another supersaturation and a dendritic growth forms driven by a temperature gradient and by the curvature of the dendrites.

While the previous observations have been reported before, what our group observed that is new is that cells migrate in the opposite direction of the salt/proteins. In effect, cells will migrate to the center of the droplet and then dendrites nucleate at the edges locating the cells in place. This probably won't be useful to other applications except our own because for SERS measurement, location of the cells is vital for SERS signal acquisition. An example of this experiment can be seen in Figure 41 where a

3 µL droplet of CSF is allowed to dry and the various panels are snapshots of a timelapse of this droplet drying on a flat gold substrate. The cells that migrate towards the center are white blood cells of a suspected inflammatory of the central nervous system patient, usually diagnosed for meningitis.

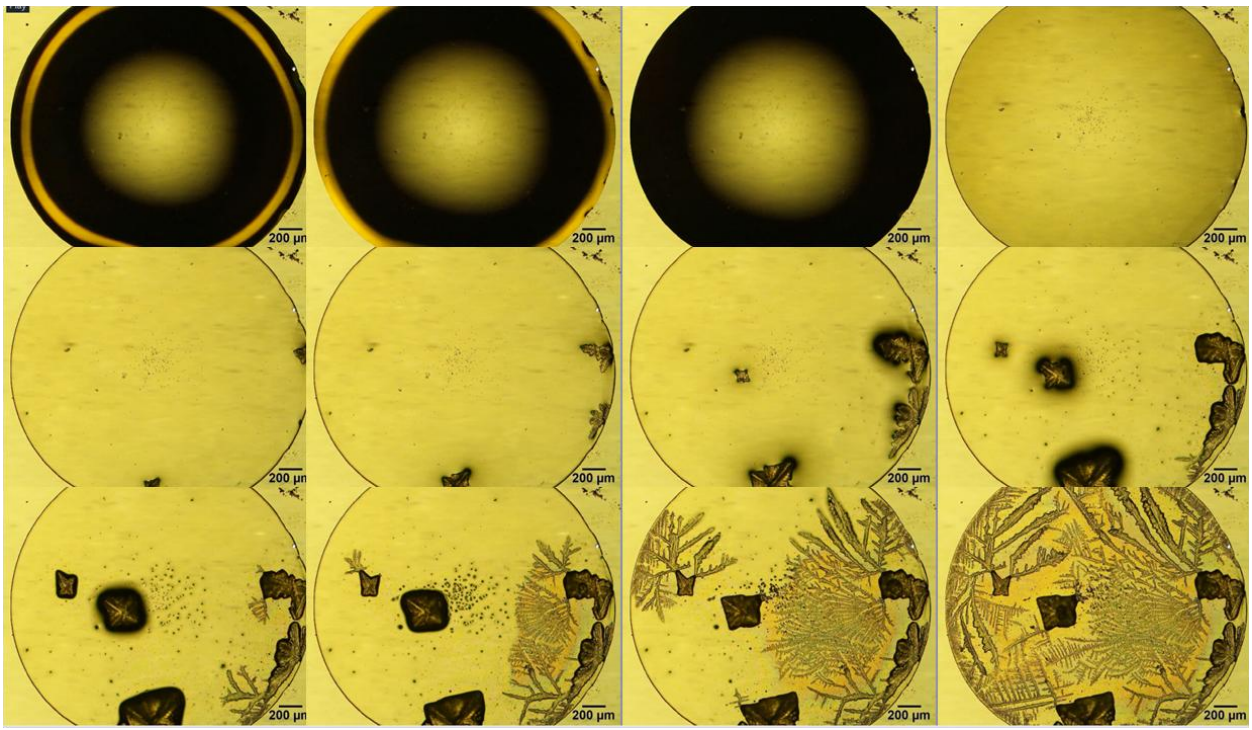



**Figure 41 – A video time-lapse of the cell droplet drying on a flat gold film substrate with the whole droplet drying within 30 mins. Each panel is taken for each significant event going from top left to right, middle left to right, and bottom left to right. The first row shows the hemispherical distortion of the droplet to the last panel with the outer droplet edge becoming defined (and already being pinned). The middle row shows the nucleation and growth process and all the migration of the cells towards the center. The bottom row shows the growth of large salt crystals and the eventual dendritic growth of the salt/protein mixture.**

### 5.3.1. Forced Directional Liquid Removal

While it is a nice discovery of the cellular migration and the effect of drying to form salt/protein dendritic growth, it is still unknown if the dendritic growth will crystallize underneath or over the cells. Knowledge of that is essential because if they crystallize underneath the cells, the SERS hotspot will become affected while if they crystallize above the cells, the change in refractive index will reduce the intensity of the laser and just reduces the signal. This is currently being done within our group, but to test workarounds, I tried to see if it was possible to just remove the salt directly. The thinking being that

after the cells have migrated towards the center, start sucking out the fluid and attempt to remove as much of the salt/protein as possible to for nucleation away from the center.

The following figure is the resultant liquid removal as seen in Figure 42. The experimental parameters are very similar to the previous one, with 3 µL of CSF being dried onto a flat gold film substrate but a filter paper cut into a triangle wedge is introduced to suck as much of the liquid away as possible. The idea is to remove as much of the salt/protein concentration to not introduce large crystal growth and widespread dendritic growth. Based on previous experiments, it was determined that there was a small window to introduce this filter paper after the 8 minute mark, when the cells have dramatically stopped moving, and before the 16 minute mark, when the salt/protein crystals start nucleating and growing. As such, the filter paper was introduced at the 11.5 minute mark as seen in the second row panel of Figure 42. One interesting observation to note is in the bottom panel where no longer present are the large salt crystallizations but very small faceted crystals nucleating and growing throughout the droplet. It can also be observed, that large dendrites no longer formed but an interweaving network of crystallization forming between these large crystals. This forced nucleation and growth is a typical controlled phenomena of material science and we observe that being done here for a biological application.

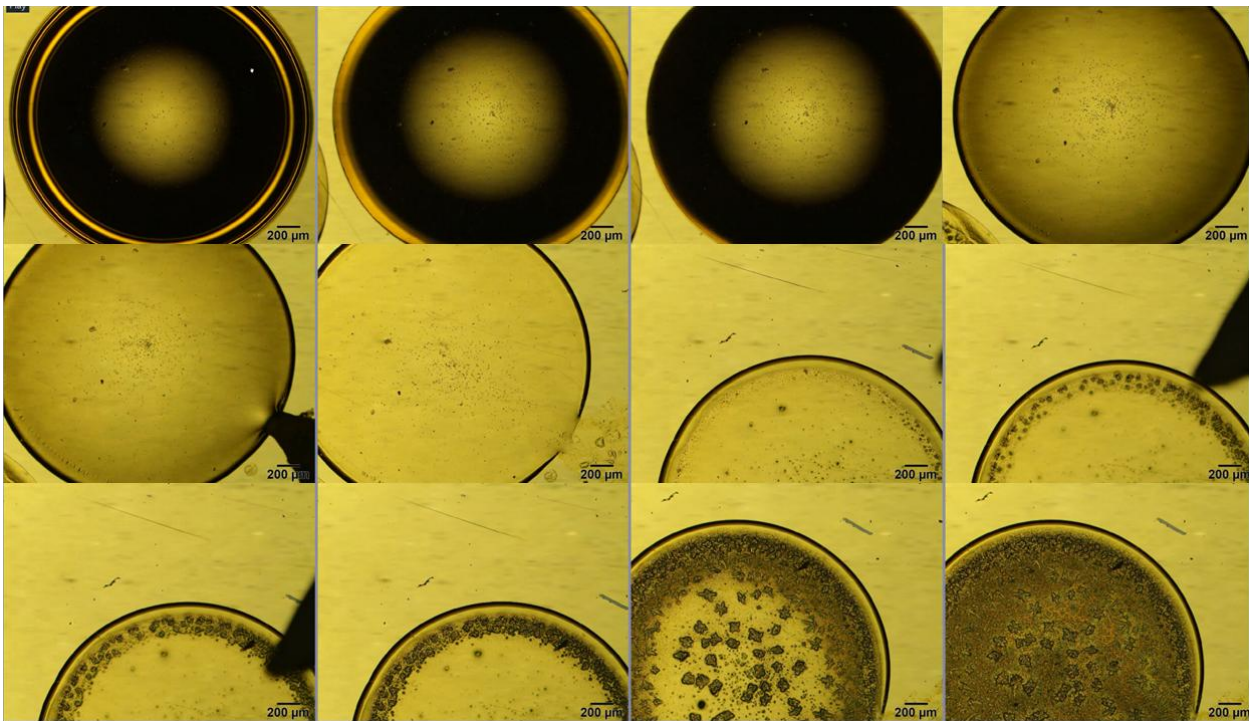



**Figure 42 – Video timelapse snapshots of a 3 µL CSF droplet drying on a flat gold film when a filter paper is introduced at the edge. The top row is very similar to the previous drying of the CSF droplet as seen in Figure 41, where the ring patterning are due to the droplet shape. The second row is when a filter paper is introduced and used to suck as much of the liquid away. The first snapshot happens at the 11.5 minute mark and immediately afterwards, salt/protein crystallization started to occur. The last panel of the second row is another attempt to suck the liquid but most of it has already been sucked or evaporated as the last row shows, no additional imprint is left on the droplet.**

# 6. Summary and Exciting New Frontiers of SERS Clinical Application

This dissertation covered various aspects of SERS and looks at its potential into clinical application. The guiding theme of my PhD is two-fold: 1) what are the SERS limits for cell-level bio-sensing application, and 2) what is necessary to translate the SERS platform for clinics. Chapter 1 provides the clinical landscape and the unmet medical needs of our time. It also highlights the potential of the SERS platform as a disease diagnostics. Chapter 2 provides the background knowledge into this interdisciplinary field: the biology, the physics, and the data science. Chapters 3, 4, 5 are the backbone of this dissertation and are the three main takeaways listed in the following section. Chapter 7 explores the current and future research of the group and the new possibilities that SERS can also provide for the medical community.

## 6.1. Three Main Takeaways

Below are the three main takeaways I have found to help answer the question of "what is needed to translate SERS from academic research to clinical application?" The details of each are covered in chapters 3, 4, 5, respectively.

### 6.1.1. Research Laboratory Environment

One of the first steps on this journey is testing the SERS platform on controlled research laboratory environment. The overarching theme for Chapter 3 is this notion of a "systematic approach on the SERS fingerprint limit". The first study involving lung cancer cell lines showed the SERS platform differentiation for different cell size and shape. The second study involving skin cancer cell lines showed a SERS differentiation when the cell size and shape are similar. The last cancer cell line study, with breast cancer, showed SERS differentiation taken from the same disease, cell morphology, cell type, but originating from different patient origin. Following the success of cancer differentiation, the cell structures of bacteria and fungus were also explored to see if the cell membrane proteome could still be

used for SERS differentiation with the added complication of a cell wall. In all these cases, SERS differentiation was possible with high sensitivity and specificity. Fi

### 6.1.2. Clinical Setting

Following the success of the controlled environment, the next step was to see how the SERS platform can handle small clinical studies. The main theme for Chapter 4 is this idea of "how theory actually differs in practice". Following the successful preliminary data of meningitis cultured cells, the first implementation of the SERS platform to clinics was done. Collaboration was done with Xijing hospital in China to get patient samples and check for meningitis using SERS. However, right away, it became apparently that actual diagnosis of meningitis is not a strong enough need as doctors tend to diagnose at the onset of inflammation. The chance of actually getting a patient with a bacterial infection was far too rare so the scope of the problem changed to see if the SERS platform can help with the diagnosis of central nervous system inflammation. As part of a side project to the SERS platform, the ML part of it was converted for image-recognition and was used for white blood cell identification using the hospital's biobank of stained cells. The culmination of all lessons learned was applied to the antibiotic resistance strain detection of tuberculosis patients in Pakistan. Since multi-drug resistance is rampant in rural Pakistan, an alternative diagnostic modality is desperately needed. The SERS platform was done on 16 patients compared with 6 healthy patients to show that the cell membrane proteome is enough for this antibiotic resistance strain differentiation.

### 6.1.3. Challenges Faced

Finally, with the success of both the research laboratory and clinics, Chapter 5 goes into challenges that I faced to obtain the success of those situations. The guiding principle for Chapter 5 is this philosophy of "what lessons were learned to bridge this gap from the research laboratory to the clinics". One of the first technical challenges to overcome was for the reproducible SERS substrate

manufacturing for consistent SERS hotspot density per chip. The second major lesson is the necessity of graphene in reducing the effect of inconsistent SERS hotspot as this single atomic layer provided a built-in intensity calibrator for fingerprint comparison. Next, the simple drying process of the analyte solution induced unique PBS/protein crystallization that formed dendritic patterns throughout the SERS chip. Finally, the evolution of the SERS data analysis suite from the initial simple dimension reduction to the more complex machine learning algorithms provided the much needed push in analyzing the vast amount of SERS data and its complexities.

# 7. Exciting Future of SERS in Healthcare

While my PhD journey comes to an end, there are still fascinating new topics that came to light from the conclusions of my research. Here are some of the examples presented below.

## 7.1.Investigating Exosomes and Early Cancer Detection

For cancer disease diagnostics, early stage diagnosis is the "holy grail" for cancer patient survival. There are many different approaches to this problem, e.g. circulating tumor cells (CTCs), circulating tumor free-DNA (cDNA), antibody or antigen biomarker, each with varying results. One new biomarker gaining research traction is through the usage of exosomes. Exosomes are extracellular vesicles of 30-140 nm that contain molecular cargo, i.e. nucleic acids and proteins, with very efficient delivery from the original cell to a target cell. In addition, studies have already shown that cancerous cells release more exosomes than healthy cells so exosomes make a good candidate for a potential new biomarker as well as the convient patient sample extraction, i.e. blood, saliva, cerebral spinal fluid [56]. In regards to colorectal cancer, several studies have already shown the presence of colorectcal cancer specific proteins through exosomal secretion, such as cadherin-17 and EpCAM [57][58]. While typically exosomal studies are done through proteomic techniques, SERS can also be used to analyze exosomes. One advantage of using SERS is the ability to analyze exosome contents individually as opposed to the typical mass requirement of tens of thousands of cells usually seen in mass spectrometry. Using PCA, exosomes from different parent sources were differentiated with clear sensitivity and specificity [59][60].

## 7.2. Establishing Proteomic Credibility Comparable to Mass Spectrometry

While the main portion of this dissertation highlights the amazing potential of SERS as an emerging technology, one main question that still remains is why is SERS not considered a "true" proteomic technique? The answer to that question can be split into two parts: 1) the (typically small) patient cohort size leading to insufficient (for clinical implementation) validation in reproducibility and

rigor and 2) lack of correlation between SERS spectra signature to specific protein expression pathways. The former is rooted in biostatistics and requires clinical trials with large scale funding, and the latter represents pending (or missing) research that could potentially be associated with fundamental biological challenges yet to be met.

For the first reason, reproducibility of the above-mentioned SERS studies for clinical applications is a big question mark in the field. Usually a statistically relevant clinical studies involves around 40 patients per measurable group depending on the disease type. To date, we have not come across a substantial amount of SERS clinical studies with this sort of rigor [61] & [62]. An example of this can be seen in the previously mentioned study where only 12 healthy patients were measured compared to 12 patients with colorectal cancer. Of those 12 cancer patients, 7 were male and 5 were female with 3, 7, and 2 patients split between cancer stage II, stage III, and stage IV, respectively [63]. Such patient cohort size is far too small to make a distinguishable clinical case, but it is enough to present the technology as a feasible possible future. As mentioned before, translation of a technology from laboratory to clinics does require consorted efforts often organized by federal agencies such as the US National Institutes of Health.

In the second explanation, all SERS studies we have come across tackle just diagnosis of the clinical disease without relating to the possible mechanism. Although SERS is touted as a proteomics technique, it technically is not since there is still difficulty correlating the observed SERS peaks with protein expression pathways. A good proteomic technique for comparison is with mass spectrometry where individual proteins are broken down to peptides and the subsequent amino acid sequence can be analyzed to form a conclusion with the corresponding gene and protein expression pathway. The working principle of SERS, on the other hand, dictates that the spectral peak heights are directly correlated to the abundance of one or more types of amino acids. With individual proteins being

composed of numerous amino acids, it remains an unanswered question as to whether the information about the abundance of amino acids could allow for unique connections to the abundance of proteins. This is in our opinion a fundamental challenge confronting SERS. It should be stressed here that even if such a direct correlation with proteome be impossible, it does not diminish the potential of SERS being used as a "molecular fingerprint", in other words clinically worthy biomarkers for disease diagnosis. However, the path to its acceptance into clinical practice will be expected to be much more treacherous because it's lacking of a clear proteomic or genomic foundation.

### 7.2.1. Controversy in -omics "fingerprint" diagnoses

Without a strong academic foundation for solidifying SERS as a proteomic technique, it will indeed face challenges integrating itself as a viable clinical technique. This is because, doctors would like have some sort of mechanistic pathway that is given from established proteomic techniques like mass spectroscopy. But even mass spectroscopy was laden in controversy as seen in this commentary article written by Dr. David Ransohoff, published in the Journal of the National Cancer Institute [64].

The argument is brought forward that –omics field techniques, e.g. proteomics, genomics, make claim to accurate diagnosis of cancer when in fact there is an issue in reproducibility and validation. Specifically, Dr. Ransohoff mentioned the root of the problem is in regards to poor experimental setup in the 2002 ovarian cancer study published in Lancet[65], where bias and chance were not fully removed from the study due to a lack of a double-blind. However, it is this author's opinion that while that part is indeed true, the heart of the issue is not the rigor of the published study, but Dr. Ransohoff and many similar scientists/clinicians in believing that a biomarker or "fingerprint" study does not constitute as enough evidence for diagnosis and prognosis. This line of thinking comes from the article, "when investigators trained in molecular biology or biochemistry start to conduct research in diagnosis and prognosis, they are undertaking, perhaps unwittingly, observational epidemiology research that involves

serious threats from chance and bias." [64] In particular, this statement is hinting to the field of proteomics and the fingerprinting nature of mass spectroscopy, where the claim of diagnosis is dependent on peak ratios without a clear biological mechanism.

While that approach is indeed valid, it is in this author's opinion that the situation can be viewed from another angle. An analogous example to this can be taken from forensic science where fingerprinting is routinely recognized by society as virtually full-proof. Yet, the science behind this is inherently empirical; there is not a governing theory that fundamentally claims a person's fingerprint is unique. However, even with this hand-wavy argument, criminal court cases use fingerprints as sufficient evidence. Another example of this similar logic can be applied to pathology in the field of WBC typing; here, WBC types are separated based purely on morphology instead of a scientifically rigor metric, such as genetically different. How can such instances work in the clinical field? This is because, empirically, diagnosis are successful and many of these diagnoses can be made without knowing the exact mechanistic origin. So bringing the whole picture back to Dr. Ransohoff's arguments: yes, it does seem that the initial 2002 study could have benefited from a double-blind to make it more rigorous and yes, the 2003 PNAS article [66] analyzing the 2002 study's data [65] with machine learning also could of used more rigor with separation of its test dataset. However, to claim that proteomics or genomics without a clear mechanistic protein expression pathway basis makes a discovery meaningless, is a far stretch.

As mentioned in the rebuttal by the 2002 authors published in 2005 by the Journal of National Cancer Institute [67], their study was to show feasibility of this new technique and further studies are needed to see the reproducibility of mass spectroscopy of serum proteins in other clinics. It could also be possible that the proteins they detected are far too small in concentration in the serum[64], however, such practicality issues are problems for startups and companies, not in novel academic research.

## 7.3. New Frontiers Summary

It can be seen that SERS still has a long path ahead before it can be fully integrated into clinics, however its potentials are quite exciting. While the focus of my dissertation largely focused on application of SERS to cells of any kind, this new chapter has shown that SERS is not limited to this. With the first section showing the applications of SERS to Exosomes shows great promise as Exosomes can be easily isolated directly from patient blood. This is particularly exciting for SERS application because this provides an inherent isolation of the biomarker from the background noise. The second section goes into the problems of SERS not being fully recognized as a proteomic tool. Generally, mass spectrometry is considered the gold standard tool and SERS can be a complimentary tool in proteomics with its own advantages and disadvantages. To solidify SERS as a proteomic tool, it is exciting to see SERS come up with its own online database of which select proteins can be quantified much similar in fashion to mass spectrometry. This will help eliminate the controversies and skepticism currently seen in SERS and its consideration as a viable tool for proteomics and in clinical applications. With these obstacles eliminated, it will only be short matter of time before SERS because a widely used tool in medicine and in healthcare.

# 8. References


[1] M. Heron, “Deaths: Leading Causes for 2017,” 2019.

[2] “US Inflation Calculator,” *U.S. Department of Labor Bureau of Labor Statistic*, 2019. .

[3] American Cancer Society, “Cancer Facts & Figures 2019,” Atlanta, 2019.

[4] B. L. Secretan *et al.*, “Breast-Cancer Screening — Viewpoint of the IARC Working Group,” *N. Engl. J. Med.*, vol. 372, no. 24, pp. 2353–2358, 2015.

[5] R. D. Rosenberg *et al.*, “Effects of age, breast density, ethnicity, and estrogen replacement therapy on screening mammographic sensitivity and cancer stage at diagnosis: Review of 183,134 screening mammograms in Albuquerque, New Mexico,” *Radiology*, vol. 209, no. 2, pp. 511–518, 1998.

[6] M. B. Coyle *et al.*, *Manual of Antimicrobial Susceptibility Testing*. American Society for Microbiology (ASM), 2005.

[7] M. Lobanovska and G. Pilla, “Penicillin’s discovery and antibiotic resistance: Lessons for the future?,” *Yale J. Biol. Med.*, vol. 90, no. 1, pp. 135–145, 2017.

[8] W. H. Organization, *Antimicrobial stewardship programmes in health-care facilities in low- and middle-income countries: a practical toolkit*, vol. 1, no. 3. Geneva, 2019.

[9] J. M. Janda and S. L. Abbott, “16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls,” *J. Clin. Microbiol.*, vol. 45, no. 9, pp. 2761–2764, 2007.

[10] I. Clarridge, Jill E., “Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases,” *Clin. Microbiol. Rev.*, vol. 17, no. 4, pp. 840–862, 2004.

[11] L. H. Sun, “Superbug known as ‘phantom menace’ on the rise in U.S.,” *The Washington Post*, Washington, pp. 1–3, 03-Dec-2015.

[12] Editorial-Opinion, “The superbug threat is morphing,” *The Washington Post*, Washington D.C., p. A.20, 28-Nov-2019.

[13] M. Ikeda and S. Takeno, “Genetics , Strain Improvement , a N D Proteins,” in *Manual of Industrial Microbiology and Biotechnology*, 3rd ed., A. L. Baltz, Richard H.; Davies, Julian E.; Demain, Ed. American Society for Microbiology (ASM), 2010, pp. 223–237.

[14] C. L. Glick, Bernard R.; Delovitch, Terry L.; Patten, “Proteomics,” in *Medicial Biotechnology*, American Society for Microbiology (ASM), 2014, pp. 2–70.

[15] P. G. Le Ru, E. C. & Etchegoin, *Recent developments*, First. Amsterdam, The Netherlands: Elsevier, 2009.

[16] P. Wang, O. Liang, W. Zhang, T. Schroeder, and Y. H. Xie, “Ultra-sensitive graphene-plasmonic hybrid platform for label-free detection,” *Adv. Mater.*, vol. 25, no. 35, pp. 4918–4924, 2013.

[17] N. Jung, A. C. Crowther, N. Kim, P. Kim, and L. Brus, “Raman enhancement on graphene: Adsorbed and intercalated molecular species,” *ACS Nano*, vol. 4, no. 11, pp. 7005–7013, 2010.

[18] P. Wang *et al.*, “Giant optical response from graphene-plasmonic system,” *ACS Nano*, vol. 6, no. 7, pp. 6244–6249, 2012.

[19] P. Wang *et al.*, “Label-Free SERS Selective Detection of Dopamine and Serotonin Using Graphene-Au Nanopyramid Heterostructure,” *Anal. Chem.*, vol. 87, no. 20, pp. 10255–10261, 2015.

[20] D. Mohammed, P. J. Matts, J. Hadgraft, and M. E. Lane, “In vitro-in vivo correlation in skin permeation,” *Pharm. Res.*, vol. 31, no. 2, pp. 394–400, 2014.

[21] T. Y. Lin *et al.*, “Microsoft COCO: Common objects in context,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014.

[22] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[23] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[24] W. Liu *et al.*, “SSD: Single shot multibox detector,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016.

[25] J. Huang *et al.*, “Speed/accuracy trade-offs for modern convolutional object detectors,” *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 3296–3305, 2017.

[26] Yangli-ao Geng, Ming Liu, Qingyong Li, and Ruisi He, “Introduction of machine learning,” in *Applications of Machine Learning in Wireless Communications*, 2019, pp. 1–65.

[27] B. Lantz, “Improving Model Performance,” in *Machine Learning with R*, 3rd ed., 2019, pp. 347–374.

[28] S. Rl, M. Kd, and A. Jemal, “Cancer statistics , 2015 .,” *CA Cancer J Clin*, vol. 65, no. 1, p. 21254, 2015.

[29] Invitrogen, “Cell Culture Basics Handbook,” *ThermoFisher Sci. Inc.*, pp. 1–61, 2010.

[30] O. Liang *et al.*, “Label-free distinction between p53+/+ and p53 -/- colon cancer cells using a graphene based SERS platform,” *Biosens. Bioelectron.*, vol. 118, no. July, pp. 108–114, 2018.

[31] CDC, “Antibiotic resistance threats in the United States, 2019, Atlanta, GA: U.S. Department of Health and Human Services,” 2019.

[32] L. D. Gray and D. P. Fedorko, “Laboratory diagnosis of bacterial meningitis.,” *Clin. Microbiol. Rev.*, vol. 5, no. 2, pp. 130–45, 1992.

[33] H. Balkhy, “Antimicrobial stewardship programmes in health-care facilities in low- and middle-income countries: a WHO practical toolkit,” 2019.

[34] G. Ajello *et al.*, *Manual for the Laboratory Identification and Antimicrobial Susceptibility Testing*

*of Bacterial Pathogens of Public Health Importance in the Developing World*. Geneva, Switzerland: World Health Organization, 2003.

[35] V. L. Feigin *et al.*, "Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016," *Lancet Neurol.*, vol. 18, no. 5, pp. 459–480, 2019.

[36] S. Nagarathna, H. B. Veenakumari, and A. Chandramuki, "Laboratory diagnosis of meningitis," in *Meningitis*, G. Wireko-Brobby, Ed. Croatia: InTech, 2012, pp. 185–208.

[37] L. Zhou *et al.*, "Simultaneous detection of five pathogens from cerebrospinal fluid specimens using Luminex technology," *Int. J. Environ. Res. Public Health*, vol. 13, no. 2, 2016.

[38] A. I. Shahin, Y. Guo, K. M. Amin, and A. A. Sharawi, "White blood cells identification system based on convolutional deep neural learning networks," *Comput. Methods Programs Biomed.*, vol. 168, pp. 69–80, 2019.

[39] A. Khashman, "Investigation of different neural models for blood cell type identification," *Neural Comput. Appl.*, vol. 21, no. 6, pp. 1177–1183, 2012.

[40] M. C. Su, C. Y. Cheng, and P. C. Wang, "A neural-network-based approach to white blood cell classification," *Sci. World J.*, vol. 2014, no. 1, 2014.

[41] M. Othman, T. Mohammed, and A. Ali, "Neural Network Classification of White Blood Cell using Microscopic Images," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 5, pp. 99–104, 2017.

[42] S. Çelebi and M. Burkay Çöteli, "Red and white blood cell classification using Artificial Neural Networks," *AIMS Bioeng.*, vol. 5, no. 3, pp. 179–191, 2018.

[43] M. Jiang, L. Cheng, F. Qin, L. Du, and M. Zhang, "White Blood Cells Classification with Deep Convolutional Neural Networks," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 09, p. 1857006, 2018.

[44] J. Zhao, M. Zhang, Z. Zhou, J. Chu, and F. Cao, "Automatic detection and classification of

leukocytes using convolutional neural networks," *Med. Biol. Eng. Comput.*, vol. 55, no. 8, pp. 1287–1301, 2017.

[45] Tzutalin, "LabelImg," 2015. [Online]. Available: https://github.com/tzutalin/labelImg. [Accessed: 10-Apr-2019].

[46] J. Rahimi and A. Woehrer, "Overview of cerebrospinal fluid cytology," *Handb. Clin. Neurol.*, vol. 145, pp. 563–571, 2017.

[47] D. S. Kermany *et al.*, "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, 2018.

[48] World Health Organisation, "Global Health TB Report," Geneva, 2018.

[49] P. J. Kelly, A. C. Webster, and J. C. Craig, "How many patients do we need for a clinical trial? Demystifying sample size calculations," *Nephrology*, vol. 15, no. 8, pp. 725–731, 2010.

[50] B. Röhrig, J. B. Du Prel, D. Wachtlin, R. Kwiecien, and M. Blettner, "Sample Size Calculation In Clinical Trials," *Dtsch. Arztebl.*, vol. 107, no. 31–32, pp. 552–556, 2010.

[51] M. Fan, G. F. S. Andrade, and A. G. Brolo, "A review on the fabrication of substrates for surface enhanced Raman spectroscopy and their applications in analytical chemistry," *Anal. Chim. Acta*, vol. 693, pp. 7–25, 2011.

[52] P. L. Stiles, J. A. Dieringer, N. C. Shah, and R. P. Van Duyne, "Surface-enhanced Raman spectroscopy," *Annu. Rev. Anal. Chem.*, vol. 1, no. 1, pp. 601–626, 2008.

[53] Y. J. P. Carreón, M. Ríos-Ramírez, R. E. Moctezuma, and J. González-Gutiérrez, "Texture analysis of protein deposits produced by droplet evaporation," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, 2018.

[54] R. D. Deegan, O. Bakajin, T. F. Dupont, G. Huber, S. R. Nagel, and R. A. Witten, "Capillary flow as the cause of ring stains fromdried liquid drops," *Nature*, vol. 389, pp. 827–829, 1997.

[55] R. E. Goldstein, "Coffee stains, cell receptors, and time crystals: Lessons from the old literature," *Phys. Today*, vol. 71, no. 9, pp. 32–38, 2018.

[56] O. A. Tovar-Camargo, S. Toden, and A. Goel, “Exosomal microRNA Biomarkers: Emerging Frontiers in Colorectal and Other Human Cancers,” *Expert Rev. Mol. Diagn.*, vol. 16, no. 5, pp. 553–567, 2016.

[57] O. K. Bernhard, D. W. Greening, T. W. Barnes, H. Ji, and R. J. Simpson, “Detection of cadherin-17 in human colon cancer LIM1215 cell secretome and tumour xenograft-derived interstitial fluid and plasma,” *Biochim. Biophys. Acta - Proteins Proteomics*, vol. 1834, no. 11, pp. 2372–2379, 2013.

[58] M. S. Ostenfeld *et al.*, “miRNA profiling of circulating EpCAM + extracellular vesicles: promising biomarkers of colorectal cancer,” *J. Extracell. Vesicles*, vol. 5, no. 1, 2016.

[59] Z. Yan *et al.*, “A Label-free platform for identification of exosomes from different sources,” *ACS Sensors*, vol. 4, no. 2, pp. 488–497, 2019.

[60] O. Galamb *et al.*, “Diagnostic and progostic potential of tissue and circulating long non-coding RNAs in colorectal tumors,” *World Jounral Gastroenterol.*, vol. 25, no. 34, pp. 5026–5048, 2019.

[61] B. Röhrig, J. B. Du Prel, D. Wachtlin, R. Kwiecien, and M. Blettner, “Sample Size Calculation in Clinical Trials,” *Dtsch. Arztebl.*, vol. 107, no. 31–32, pp. 552–556, 2010.

[62] P. J. Kelly, A. C. Webster, and J. C. Craig, “How many patients do we need for a clinical trial? Demystifying sample size calculations,” *Nephrology*, vol. 15, no. 8, pp. 725–731, 2010.

[63] H. Ito *et al.*, “Use of surface-enhanced Raman scattering for detection of cancer-related serum-constituents in gastrointestinal cancer patients,” *Nanomedicine Nanotechnology, Biol. Med.*, vol. 10, no. 3, pp. 599–608, 2014.

[64] D. F. Ransohoff, “Lessons from controversy: Ovarian cancer screening and serum proteomics,” *J. Natl. Cancer Inst.*, vol. 97, no. 4, pp. 315–319, 2005.

[65] I. Petricoin, Emanuel F *et al.*, “Use of proteomic patterns in serum to identify ovarian cancer,” *Lancet*, vol. 359, pp. 572–577, 2002.

[66] W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm, and J. S. Kovach, “Detection of cancer-specific markers amid massive mass spectral data,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 25, pp. 14666–14671, 2003.

[67] L. A. Liotta *et al.*, “Importance of Communication Between Producers and Consumers of Publicly Available Experimental Data,” *J. Natl. Cancer Inst.*, vol. 97, no. 4, pp. 310–314, 2005.