

---

Wayne State University Dissertations

---


1-1-2018

## Identification Of Streptococcus Pyogenes Using Raman Spectroscopy

Ehsan Majidi

Wayne State University, ehsan.majidi\_ee@yahoo.com

Follow this and additional works at: [https://digitalcommons.wayne.edu/oa\\_dissertations](https://digitalcommons.wayne.edu/oa_dissertations)

 Part of the [Artificial Intelligence and Robotics Commons](#), [Biomedical Engineering and Bioengineering Commons](#), and the [Optics Commons](#)

---

### Recommended Citation

Majidi, Ehsan, "Identification Of Streptococcus Pyogenes Using Raman Spectroscopy" (2018). *Wayne State University Dissertations*. 2048.

[https://digitalcommons.wayne.edu/oa\\_dissertations/2048](https://digitalcommons.wayne.edu/oa_dissertations/2048)

This Open Access Embargo is brought to you for free and open access by DigitalCommons@WayneState. It has been accepted for inclusion in Wayne State University Dissertations by an authorized administrator of DigitalCommons@WayneState.



©COPYRIGHT BY

EHSAN MAJIDI

2018

All Rights Reserved

## DEDICATION

*To my father, mother, and sister*

## ACKNOWLEDGEMENTS

I would like to thank my advisor Prof. Auner for his great support throughout my Ph.D. work. I have always been appreciative of his continuous and precious support, mentorship, and encouragement.

I would like to thank my family who has unconditionally helped me all the way, my father for his endless guidance and counsel, and my mother and sister for their kindness and encouragement.

I would like to thank the Smart Sensors and Integrated Microsystems (SSIM) lab and my colleagues there who provided an excellent environment for research. Also, I am grateful for the support and empathy of all my friends.

I would like to thank Wayne State University and in particular the Electrical and Computer Engineering Department for providing me the opportunity to pursue my education and broaden my knowledge.

## TABLE OF CONTENTS

<b>Dedication</b> . . . . .	<b>ii</b>
<b>Acknowledgements</b> . . . . .	<b>iii</b>
<b>List of tables</b> . . . . .	<b>vii</b>
<b>List of figures</b> . . . . .	<b>viii</b>
<b>Chapter 1 Introduction</b> . . . . .	<b>1</b>
1.1 Background . . . . .	2
1.2 Identification of Bacteria Using Raman Spectroscopy . . . . .	4
1.3 Deep Learning . . . . .	7
1.3.1 Introduction to Deep Learning . . . . .	7
1.3.2 Convolutional Neural Networks . . . . .	12
1.3.3 Parallel Computing with Neural Networks . . . . .	15
1.4 Dissertation Scope . . . . .	16
<b>Chapter 2 Identification of <i>S. pyogenes</i> using Raman Spectroscopy: A comparative study on multivariate analyses</b> . . . . .	<b>17</b>
2.1 Introduction . . . . .	17
2.2 Material and Method . . . . .	20
2.2.1 Instrumentation and Sample Preparation . . . . .	20
2.2.2 Dataset . . . . .	21
2.2.3 Statistical Analysis . . . . .	21
2.3 Result and Discussion . . . . .	24
2.3.1 Data visualization . . . . .	24
2.3.2 HCA . . . . .	24

2.3.3	PCA . . . . .	27
2.3.4	Discriminant Function Analysis . . . . .	28
2.3.5	SVM . . . . .	30
2.3.6	The Effect of the Number of PCs on the Classification . . . . .	33
2.3.7	Random Forest . . . . .	38
2.3.8	Comparative Result . . . . .	42
<b>Chapter 3</b>	<b>Real-Time Deep Learning Approach for pathogen identification using Raman Spectroscopy: Identification of S. pyogenes . . . . .</b>	<b>46</b>
3.1	Introduction . . . . .	46
3.2	Material and Method . . . . .	48
3.2.1	Dataset . . . . .	48
3.2.2	Input . . . . .	49
3.2.3	Overview of the Model . . . . .	49
3.3	Result and Discussion . . . . .	54
3.3.1	Pre-processing Unit . . . . .	55
3.3.2	Classification Result . . . . .	55
3.3.3	Comparative Result . . . . .	59
<b>Chapter 4</b>	<b>Identification of Streptococcus pyogenes in confounding background using Raman Spectroscopy . . . . .</b>	<b>62</b>
4.1	Introduction . . . . .	62
4.2	Material and Method . . . . .	63
4.2.1	Instrumentation and Sample Preparation . . . . .	63
4.2.2	Dataset . . . . .	64
4.2.3	Input . . . . .	64

4.2.4	Model . . . . .	65
4.3	Result and Discussion . . . . .	65
4.3.1	Data visualization . . . . .	65
4.3.2	Training Result . . . . .	66
4.3.3	Realization of the Network: Macromolecules . . . . .	70
<b>Chapter 5</b>	<b>Conclusion and Future Works . . . . .</b>	<b>77</b>
5.1	Conclusion . . . . .	77
5.2	Future Works . . . . .	78
<b>References</b>	. . . . .	<b>81</b>
<b>Abstract</b>	. . . . .	<b>96</b>
<b>Autobiographical Statement</b>	. . . . .	<b>98</b>



## LIST OF TABLES

Table 1.	Major Raman bands used as a reference library in microbiological analysis [1, 2, 3]. . . . .	6
Table 2.	Dataset Summary. . . . .	22
Table 4.	Classification Accuracy on Testing Dataset. . . . .	43
Table 5.	Classification Result on Training, Validation, and Testing Dataset. . . . .	59
Table 6.	Result on the Benchmark Dataset with Water Background. The proposed approach (DPINN) achieves the lowest error and the highest sensitivity, specificity, and F1 score among others. . . . .	61
Table 7.	Dataset Summary with Confounding Background . . . . .	64
Table 8.	Classification Results on Training, Validation, and Testing Dataset. . . . .	66
Table 10.	Mean Probability of Macromolecules from Which an accurate identification with a probability above 0.65 can be yielded. . . . .	75

## LIST OF FIGURES

Figure 1.	Experimental Setup of Raman Spectroscopy [4]. . . . .	21
Figure 2.	Mean and Standard Deviation of Raw Data Normalized to Maximum Intensity. . . . .	25
Figure 3.	Mean and Standard Deviation of Background Removed Spectra Normalized to Maximum Intensity. . . . .	26
Figure 4.	Hierarchical Cluster Analysis Performed on Averaged Raman Spectra of Seven Species used in This Study. The red dot represents the distance between clusters calculated by ward method. The dendrogram shows a clear separation of the tap water from pathogens. E.coli is the least similar to S. pyogenes, and the spectrum of Legionella pneumophila and Pseudomonas aeruginosa is very close to the S. pyogenes. The MRSA and MSSA can be categorized in one cluster as they are strains of Staphylococcus. . . . .	27
Figure 5.	Cumulative Explained Variance Ratio for First 10 Principal Components. .	28
Figure 6.	Linear Function Analysis to Discriminate S. pyogenes from Not-S. pyogenes. The solid blue corresponds to the S. pyogenes pathogen and the red circles correspond to the Not- S. pyogenes species. The circles display two times the standard deviation for each class. The black dot is the mean value of each class. The misclassified samples are represented by dark blue and dark red corresponding to S. pyogenes and Not-S. pyogenes, respectively. . . . .	30
Figure 7.	Quadratic Function Analysis to Discriminate S. pyogenes from Not-S. pyogenes. The solid blue corresponds to the S. pyogenes pathogen and the red circles correspond to the Not- S. pyogenes species. The ellipsoids display two times the standard deviation for each class. The black dot is the mean value of each class. The misclassified samples are represented by dark blue and dark red corresponding to S. pyogenes and Not-S. pyogenes, respectively. . . . .	31
Figure 8.	Grid Search for Validation Accuracy. a) An initial logarithmic search on $C$ and $\gamma$ values where the kernel was 'rbf' and the 4-fold cross-validation is used. b) The fine-tuning for $10^{-4} < \gamma < 10^{-2}$ and $10^5 < C < 10^8$ . . . . .	32
Figure 9.	SVM Classification Accuracy for Different Kernels. . . . .	33
Figure 10.	Bias and Variance of PCA-LDA in Terms of Number of PCs Where Gaussian Noises Are Added to All Spectra . . . . .	35

Figure 11. Bias and Variance of PCA-QDA in Terms of Number of PCs Where Gaussian Noises Are Added to All Spectra . . . . .	36
Figure 12. Bias and Variance of PCA-SVM in Terms of Number of PCs Where Gaussian Noises Are Added to All Spectra . . . . .	37
Figure 13. Bias and Variance of PCA-LDA in Terms of Number of PCs Where Gaussian Noises Are Added to the Validation Set . . . . .	39
Figure 14. Bias and Variance of PCA-QDA in Terms of Number of PCs Where Gaussian Noises Are Added to the Validation Set . . . . .	40
Figure 15. Bias and Variance of PCA-SVM in Terms of Number of PCs Where Gaussian Noises Are Added to the Validation Set . . . . .	41
Figure 16. Grid Search on Average 4-Fold Cross-Validation Accuracy. 25 trees with 5 descriptors are the best parameter of Random Forest result in 94.46% accuracy. . . . .	43
Figure 17. ROC of Random Forest, Gaussian SVM, linear SVM, LDA, and QDA on the Common Test Set. Random Forest with an area under the curve of 0.997 results in the best method to identify <i>S. pyogenes</i> in terms of specificity and sensitivity. . . . .	44
Figure 18. Overview of the Model. . . . .	50
Figure 19. Architecture of the Pre-processing Network. . . . .	50
Figure 20. Architecture of the Identification Network. . . . .	54
Figure 21. Output of Pre-processing Unit Applied to Raw Spectrum of pathogen. It can be seen that pre-processing unit can estimate the ground-truth spectrum very firmly in almost all bands. . . . .	56
Figure 22. Output of Pre-processing Unit Applied to Raw Spectrum of Water. . . . .	57
Figure 23. Output of Preprocessed Unit for <i>S. pyogenes</i> Data. The mean and standard deviation of predicted and ground-truth pre-processed spectra are plotted in blue and green, respectively. . . . .	58
Figure 24. ROC of Training, Validation, and Testing Dataset. . . . .	60
Figure 25. Mean and Standard Deviation of Raw Data Acquired Using Throat Swab Normalized to Maximum Intensity. . . . .	67

Figure 26. Mean and Standard Deviation of Background Removed Spectra Acquired Using Throat Swab Normalized to Maximum Intensity. . . . .	68
Figure 27. Misclassified Sample. <i>S. pyogenes</i> Spectra Classified as Not- <i>S. pyogenes</i> . . .	69
Figure 28. ROC of three datasets. The AUC of each ROC is illustrated in the plot. . .	71
Figure 29. Result of True Negative and True Positive on All Macromolecules Participating in the Network. b-carotene, d-arabinose, d-fucose, and d-mannose have probability above 75% for accurate detection of <i>S. pyogenes</i> spectra and the l-histidine and amylopectin with a likelihood above 75% to be responsible for correct identification of Not- <i>S. pyogenes</i> spectra. . . . .	72
Figure 30. Result of True Positive and False Negative on All Macromolecules Participating in the Network. Adenine and d-xylose have the highest probability of false negative rates, above 0.65. . . . .	73
Figure 31. Result of True Negative and False Positive on All Macromolecules participating in the Network. b-carotene with a probability of 0.68 is the strongest macromolecule contributed to false positive rate. . . . .	74

## CHAPTER 1 INTRODUCTION

Raman Spectroscopy (RS) is a non-invasive technique that can provide a fingerprint of a molecule. The influence of water is minimal in Raman spectrum which makes it a suitable technique for biological application since biological samples have an abundance of water.

Currently, microbiological techniques can identify microorganisms on species and strain. However, these methods are labor-intensive as analysis concentrates on certain types of biomolecule analysis. Raman spectra of bacteria provide a molecular fingerprint pattern of bacteria which can be used for identification. Nevertheless, they contain the information of the cell's composition. Bacteria have a conventional structure, and as a result, their spectra have a typical pattern. However, it is complicated to assign each band to the cell's composition as Raman spectra are a superposition of contributions from all Raman active molecules in the cell.

Although data analysis techniques have developed rapidly in recent years, the evaluation of techniques for Raman spectra analyses has not been satisfactory. As a result, they have relied on expert effort and assignment of each band to specific vibrations or biomolecules that cause the identification of bacteria to be labor-intensive and less accurate, as they ignore possible vibrations of other molecules. On the other hand, contributions of background molecules in Raman spectra of bacteria are not well-understood, as they are challenging and complex which restricts the usage of the current approaches for clinical applications.

This study aims to develop data analysis techniques based on deep learning methods to identify bacteria, in particular, *S. pyogenes*, using their Raman spectra. It includes the study of current approaches for bacteria identification and whether these approaches provide

useful information to design a deep neural network. Although the proposed technique in this dissertation will be tested on *S. pyogenes* dataset, it can provide a framework to be extended to other bacteria as well.

## 1.1 Background

Bacteria are single-cell organisms that are associated with infectious disease. Although some bacteria are important for human health, pathogenic bacteria can be a threat to human life. For example, some staphylococci pathogens cause food poisoning [5] and some streptococci cause throat and ear infections [6].

Over decades these pathogens have been investigated, and methods have been developed to detect and characterize the pathogens. In vitro identification of the pathogens has had an enormous impact on patients with infections, and it has been shown that the mortality rates and health care costs can be reduced when bacteria are identified quickly [7].

Bacteria can be classified into three groups based on their morphological forms: rods (bacilli), spherical (cocci), and spirals (spirilla). Although their morphological form can be different, they have a typical structure. The cell envelope, cytoplasm, and nucleoid (DNA) are the primary structure of bacteria. The cell envelope is the most critical part of the bacteria that keeps it alive and is composed of the capsule, cell wall, and cell membrane. Gram-positive bacteria consist of a two-layer wall, a thick peptidoglycan sheet, and an internal membrane. Gram-negative bacteria have a cell wall of multilayer structure: a thin peptidoglycan sheet, an internal membrane, periplasm, and outer membrane. The outer membrane is a lipid bilayer composed of phospholipids and lipopolysaccharide (LPS). LPS is composed of two proteins and a lipid A tail. Gram-positive bacteria are classified as aerobic

cocci and bacilli based on their shape. Gram-negative bacteria are divided into four groups: cocci, enteric, nonfermenters, and pleomorphic bacteria. Gram-positive aerobic cocci has a thick cell wall and spherical shape and show aerobic action on glucose. There are many known Gram-positive aerobic cocci, such as Micrococcus, Staphylococcus, and Streptococcus.

Streptococcus bacteria are divided into four groups. *S. pyogenes* cause throat infections and can be treated with penicillin. *S. agalactiae* bacteria are responsible for urogenital infections. Type D Streptococcus includes two subgroups of enterococci and non-enterococci. The last group is the viridans group that includes *S. mitis* and *S. mutans*.

*Streptococcus pyogenes*, so-called strep A, can cause throat and skin infection and may vary from mild condition to life-threatening disease. The non-invasive GAS infections are more common and less severe, and bacteria usually colonize the throat area. Strep throat, so-called pharyngitis, causes 15-30% of childhood cases and 10% of adult cases. These infections can be treated with antibiotics [8]. The invasive infections of GAS are more severe and less frequent. The bacteria colonize in areas such as blood and organs [9]. These infections can cause diseases, such as streptococcal toxic shock syndrome (STSS), necrotizing fasciitis (NF), pneumonia, and bacteremia [10].

Microbiological methods for bacteria identification are based on the cultivation of bacteria from pure culture and determining the response of bacteria to environmental conditions. In these methods, an expert is needed to compare the test case with known microorganism by viable counting of the visible colonies [11].

Many tests are usually performed to identify a bacteria, such as the morphology test and Gram-strain reaction [12]. These tests include blood counts, urinalysis, and cultures of blood or fluid from a wound site. Gram-strain test is usually performed to identify Gram-positive

cocci in chains, and then it is cultured on blood agar. Moreover, a bacitracin antibiotic disk is added to show sensitivity for an antibiotic.

These methods are time-consuming and can take up to a few days, and accuracy is limited [13]. Early recognition and treatment of severe GAS infections are critical [9] as they may lead to shock, multisystem organ failure, or death.

Vibrational spectroscopy approaches have been developed because of the demands for rapid and accurate identification of pathogens. Infrared (IR) and Raman spectroscopy (RS) are based on the vibration modes of the molecules and provide a unique molecular fingerprint of bacteria. In IR, an infrared light is absorbed by the sample when the light's frequency is matched with the natural vibration frequency of the sample molecules and the absorbed radiation can be detected [14]. In RS, coherent light is focused on the sample and the scattered beam detected to identify the vibration modes of molecules. These methods are noninvasive and nondestructive and can detect bacteria rapidly and more accurately [15].

## **1.2 Identification of Bacteria Using Raman Spectroscopy**

RS is a molecular fingerprinting method that has been used in various applications, such as study of minerals, [16] characterization of polymers [17] and medicine [18]. Applications of RS in chemical characterization started a long time ago. However, it has been used lately in the study of biological samples intended to identify and characterize pathogenic organisms. RS instrumentation has recently become more powerful, fast, and portable. Moreover, data analysis techniques have developed rapidly. Due to its all optical and noninvasive nature, RS can be a robust detection method for bacteria.

A shift in the vibration of the nucleus occurs when the photon interacts with the nucleus.



Complementary to IR, which measures the dipole moment variation of a molecule, RS measures the polarizability variation of the molecule caused by the interaction of a photon and nucleus. Vibration scattering can be categorized into Stokes and anti-Stokes scattering. Stokes Raman scattering occurs when the emitted light has less energy than the incident light. Hence, Stokes lines have a more extended wavelength, and a photon is released during scattering. The loss in the energy of the scattered photon from the incident photon is converted to energy for a change of the shift in dipole moment. On the other hand, anti-Stokes scattering occurs when the scattered light has more energy than the incident light, and as a result, a photon is destructed leading to a change in the vibration state of a molecule [19]. In this case, the molecule is already in the excited state, and an incident photon absorbs its energy. Hence, the scattered photon emits more energy than the incident one. Stokes scattering is more intense than anti-Stokes scattering at a standard temperature as the probability of the lower states is more than the higher states.

Infrared absorption relies on the variation of intrinsic dipole moment as molecules vibrate. Raman scattering requires a change in the polarizability of functional groups occurring with atoms vibrations. As a result, polar groups such as C-O, N-H, and O-H have intense IR stretching vibration while non-polar groups such as C C, and S S have strong Raman bands. Table 1 shows the major Raman bands that are used as a reference library in microbiological analysis [19, 1, 2].

Vibrational spectroscopy classifies microorganisms based on the biochemical compositions of the biochemical cell membrane. Supervised and unsupervised chronometric models have been developed to study various bacterial cells using RS.

RS has been used for discriminating many bacteria including *Listeria monocytogenes*

Assignment of Raman spectra bands		
Raman frequency (cm <sup>-1</sup> )	fre-	Assignment
407		Skeletal modes of carbohydrates (glucose)
481		Skeletal modes of carbohydrates (starch)
520–540		S S str
540		COC glycosidic ring ref
620		Phenylalanine (skeletal)
640		Tyrosine (skeletal)
665		Guanine
720		Adenine
785		Cytosine, uracil
810–820		Nucleic acids (C–O–P–O–C in RNA backbone)
830		“Exposed” tyrosine
838		DNA
852		“Buried” tyrosine
858		CC str, COC 1,4 glycosidic link
897		COC str
1004		Phenylalanine
1061		C N and C C str
1085		C O str
1098		CC skeletal and COC str from glycosidic link
1102		>PO <sub>2</sub> <sup>-</sup> str (sym)
1129		C N and C C str
1230–1295		Amide III
1295		CH <sub>2</sub> def
1440–1460		C H <sub>2</sub> def
1575		Guanine, adenine
1573		C=C, N–H def, and C–N str (amide II)
1606		Phenylalanine
1614		Tyrosine
1650–1680		Amide I
1658		Unsaturated lipids
1735		>C O ester str
2870–2890		CH <sub>2</sub> str
2935		CH <sub>3</sub> and CH <sub>2</sub> str
2975		CH <sub>3</sub> str
3059		(C C H) <sub>(aromatic)</sub> str

Table 1: Major Raman bands used as a reference library in microbiological analysis [1, 2, 3].

[20], Salmonella enterica [21], Escherichia coli O157:H7 [22], Pseudomonas aeruginosa [23], Staphylococcus sp. [24].

As Raman signals are weak, some techniques have been developed to increase the signal intensity. Surface-enhanced Raman scattering (SERS) is based on the fact that if an analyte is close to a roughened surface (i.e. substrate) vibration mode can be enhanced [25, 26, 27]. SERS substrates that are suitable for biological samples consist of gold or silver nanoparticles. It has been shown that Raman intensity is increased by as much as  $10^{15}$  in-fold [28]. However, the data from SERS bands cannot be compared with the data of RS as there is a shift in bands.

A new method is presented to discriminate gram-positive (Enterococcus faecalis and Streptococcus pyogenes) and gram-negative (Acinetobacter baumannii and Klebsiella pneumonia) bacteria using SERS [29]. For this purpose, silver nanoparticles in solutions with highly concentrated chloride ions are used as the SERS substrate.

## 1.3 Deep Learning

### 1.3.1 Introduction to Deep Learning

The success of machine learning is based on the successful choice of features. The best result in machine learning cannot be achieved without an expert to determine which aspect of the problem should be considered more in the input. Natural learning systems such as human or animal brain can determine which aspects of high-dimension input are more worth focusing on with little guidance. This difference in feature representing and also algorithm learning between natural and artificial learning leads to a difficulty in creating learning systems which can respond to high-dimensional input flexibility and do "hard AI" tasks like

human level image understanding and communicating fluently in natural language.

Mapping one data representation to another as input is valuable to a machine learning algorithm. Input data typically are raw, low-level, unabstracted data like pixel intensity. Representing learning algorithm aims at determining the high-level properties of raw input. For example, this representation might be edging, shape, or color. For audio signal, it can be frequencies or compound sound from the dictionary. This representation algorithm may function on data independent of each other if the data are discrete, such as in a set of independent images. Alternatively, it might depend on the history of the signal up to that time such as in continuous audio.

There are techniques which can be considered as an automated representation, learning which aims at reducing the dimension of input such as Principal Components Analysis (PCA) [30], k-means clustering [31], and canonical correlation analysis (CCA) [32, 33].

These techniques are useful for pre-processing of learning algorithms. There are many other methods which are useful for feature extraction, such as manifold learning [34, 35], sparse coding [36, 37, 38], spectral clustering [39, 40], (single-layer) autoencoder networks and variations [41], and probabilistic latent factor models such as latent Dirichlet allocation [42], sigmoid belief networks [43], and restricted Boltzmann machines (RBMs) [44, 45].

These techniques produce a single new representation all at once. In other words, they do not use intermediate layers of representation. However, deep representation learning extract features on multiple intermediate layers where high-level features are functions of low-level features.

It was confirmed that useful representation for the hard problem might need multiple layers of representation [46, 47].

The primary visual cortex in mammals has been shown to be hierarchically organized. The earlier stages of processing, area V1, determine points, edges, and lines, where the later layer, area V2, uses this feature to identify more complex shapes [48]. Also, the visual cortex has shown that there is a feedback connection from a high level to low levels as well. Although some researchers were inspired by this to develop a Deep Attention Selective Network (dasNet) [49], many models nowadays use the feedforward-based network.

To get the best result on "hard AI" tasks, some properties of learning algorithms should be considered such as expressivity, disentangling factors of variation, and the most critical one abstraction [50]. Expressivity suggests that useless information should be omitted. Disentangling factors of diversification indicated that induced features should change independently from each other. Abstraction suggests identifying inputs share information with each other and find meaningful, predictive features. In other words, more abstract features recognize properties shared between instances which might look dissimilar. Hence, it enables us to use low-level features that lead to increasing the efficiency of representation and finally creating hierarchical representation.

These representation algorithms that create hierarchy features are referred to "deep learning". Deep learning is based on the principle that more abstract elements are a form of the more primitive ones. It has been shown that deep neural network can be scaled up to high-dimension data which are useful for many tasks compared to non-hierarchy techniques. Probabilistic models have been used in machine learning for a long time. Although Hierarchical representation based on a probabilistic approach might be appealing, the specific interface between multiple layers of unobservable variables is intractable. As a result, fitting those variables to data is more difficult.

Deep Neural Networks (DNN) are based on artificial neural networks and are descendant from feedforward neural network, so-called multilayer perceptron. The multilayer perceptron is a basic feedforward neural network architecture consisting of multiple layers and a single dimension output. The architectures of deep neural networks have been developed in recent years to improve the results where input or output of the network shows a particular structure. A deep learning algorithm automatically creates a hierarchical set of features. Recent development in the field has introduced novel architectures and new techniques of training data.

Feedforward neural networks with multiple layers were introduced long ago [51]. However, these networks were not suitable in practice. Finding the optimal parameters for the models has been a challenging problem. These networks consist of one or at most two hidden layers initializing the network with random parameters and then training them based on stochastic gradient descent technique [52]. However, this approach has failed in deep neural network as its cost function is highly non-convex and may have a pathological curvature that causes the failure of the gradient descent with random initialization [50, 53, 54].

Some methods were proposed to overcome these difficulties. The first method suggests initializing the parameters through optimization of an unsupervised, generative, layer-by-layer training criteria for multilayer networks [55, 56]. This technique, initializing before training, has been known as "pre-training". One of the successful methods of pre-training was called the deep belief network that learns parameters through a generative probabilistic model [57, 58].

Pre-training deep neural networks helps the performance of deep networks drastically as it might regularize them toward more general solutions. In other words, it abates the

influence of early examples [53]. Also, it may reduce pathological curvature around random initialization [54]. The error of gradient descent in random initialization remains unchanged for first layers, and it varies for final layers. Pre-training causes the deep neural network to have a reasonable initial point to start and also allows the deep neural network to propagate all the way down to first layers.

In addition to pre-training methods, applying new activation functions has shown an improvement in the result. The most common activation function is a sigmoid function, such as logistic sigmoid and hyperbolic tangent. Recently, It has been shown that good results can be obtained by using new functions, such as rectified liner units ReLU, which use  $\max(0, x)$  or maxout which uses the maximum over inputs [59].

These functions have an extensive range as they are linear while the sigmoid functions are saturated to a maximum or a minimum which causes the derivative to vanish. As a result, using piecewise linear activation functions helps the error of gradient descent to be propagated to low-level layers.

The other technique to improve the result is to use a more complicated optimization method instead of using first-order stochastic optimization. The motivation for this method is that training with a random initialization leads to a pathological curvature which makes first-order optimization slow. In [54], the Hessian-free algorithm approximates Newton's method which is applied and used by [60] to train deep networks.

Recently, deep learning networks could solve several difficult problems, such as speech recognition [61], image and character recognition [62, 63, 64], natural language parsing [65, 66], language modeling [67], machine translation [65], pixels-to-controls video-game playing [68], and playing the challenging game of go [69] among others. This reveals the fact that

learning multilayer representations is a successful approach for complex tasks.

The methods for training data and choosing architectures vary significantly from one task to another task suggesting that different problems should benefit from different algorithms. For example, conventional neural networks are used in image processing as using shared feature maps at different locations of an image proved to have the best outcome [62, 64]. Also, convolutional architectures have been designed in such a way to yield the best performance, for example, using max pooling to reduce possible overlapping regions and the use of fully connected layers before output.

In order to acquire the best result, different architectures and different training methods need to be used. For solving a given problem, a certain amount of experience is necessary to identify the best architecture and optimal methods for training data. As new techniques are introduced at a rapid pace, it will still be empirical to determine an optimal solution for a task.

The field of deep learning is a broad area. Some architectures like convolutional neural networks have been developed for a particular form of input; on the other hand, some techniques such as dropout or novel transfer functions, like rectified linear, are generally used. The empirical nature of the field prohibits proving a method dominates another in all situations. In the present work, we aim at providing evidence that our model has desirable properties to some extent for biological data of RS.

### **1.3.2 Convolutional Neural Networks**

In image processing, the input is visual images that can be represented as an unstructured vector where the intensity of pixels describes the vector. Such representation ignores the spatial relationship between pixels. Earlier progress in the image processing area attempted



to extract some features of the images such as edge [70], corner textures [71] and other features. These features all are dependent on the spatial arrangement of pixels.

Convolutional neural network (CNN) is a method based on the assumptions that each feature depends on pixels in a small window, and the same feature map should be applied to all patches of the image that help to extract a localized characteristic of the image. CNN is used in the Deep network by creating a deep hierarchy of multilayer CNN. This network attempts to determine localized feature values in each layer.

In many neural networks, all neurons in hidden layers are connected to all other neurons in the previous layer including inputs. It is possible to connect each neuron to a small number of neurons in the previous layer. The connection patterns can be specified for the structure in the input. For example, if an image is input, each neuron of the hidden layer can only consider adjacent pixels of input and extend it to a network locally deep connected. Due to implementation of this idea, in the forward pass, the weight of an unconnected neuron can be considered zero, and thus for backpropagation, the gradient descent does not need to be computed for empty connections.

Using local connection reduces the number of links. Weight sharing is another method that can even lessen connection further. In weight sharing, some of the weights can be considered the same for that layer. Convolution inspires the idea of weight sharing. In a convolutional neural network, a filter or feature map is applied to many locations of input. The convolution layer is usually connected to another layer named the pooling layer. For instance, max-pooling computes the max value of the outputs of a convolution layer. Then, the output of the pooling layer is used as input for the next layer. One advantage of applying max-pooling to the convolution layer might be that the output of max-pooling is invariant

to shift in inputs. In other words, the output is transnational invariance. Transnational invariance can be useful for many data as a measure of typical data translation. The max-pooling layer is known as a sub-sampling layer and reduces the input size drastically.

One variant of CNN is local contrast normalization (LCN). In this method, another layer is applied on the max-pooling layer, normalizing some set of the output of max-pooling. In other words, it subtracts the mean and divides the standard deviation of the input. This method helps the network to have brightness invariance. It is possible to stack another CNN network on top of the CNN network; the output of max-pooling becomes the input of the second CNN.

Backpropagation in CNN should be modified in a way that the gradient descent propagates from a higher level to lower level. These modifications can be in the forward pass, considering all weights, such as fully connected network while supposing some of them are the same. In backward pass, the gradient is computed for all the weights and during the update, using the average of the gradients of the shared weights. For max-pooling in backward pass, the gradient descent of the branch that gives the max value should be computed.

The other variant of CNN is for multiple channel input. In this method, the convolution is applied to various channels. However, the weights across channels are not shared. The other technique is to extend convolution architecture to have many filters or feature maps for one position. In this technique, the outputs of the feature maps can be considered as multiple channel input for the next layer.

In order to use CNN for more than one dimension, the filters can have the proportion of the input. For example, for an image, the filter can be two dimensions. In image processing application images usually do not have the same size. One method is to crop the image at

the center and resize the images to the original size. Usually, the outputs of the latest CNN in the network are connected to one or two fully connected neural networks which helps to see the input data as a whole. Also, dropout is usually applied to fully connected layers.

Depending on the size of images or data, many implementations of CNN consider using FFT techniques or step convolution. Sometimes using FFT methods for convolution makes the forward and backward pass faster. The breakthrough study is introduced in [72], and many new techniques have been presented after that which have resulted in rapid improvement of CNN [73, 74].

### 1.3.3 Parallel Computing with Neural Networks

Computation cost is one of the major drawbacks of deep neural network, especially when the training set is large. Parallelism is a method addressing this issue that can be in model level or data level. DistBelief [75] is a popular framework which has two levels of parallelism.

The model parallelism suggests partitioning a model to multiple machines while communication is established between them. Although there are communication costs, usually machines are locally connected; this cost is smaller than computation cost.

Data parallelism attempts to partition data into different sets and train the model on each set of data. Synchronizing between machines can be done by using another device, a so-called parameter server which stores the parameters. At each step, every copy of the model computes its parameters, such as gradient on its data, and then sends it to the parameter server. The parameter server updates the parameter as it is received from the different machine and broadcasts new parameters to the network after some iteration. This approach of parallelism is called data parallelism via asynchronous communication as this communication is asynchronous [76].

There is another level of parallelism that is based on multithreading within a machine. In this level, some cores are assigned to read data, and other cores are allocated to perform matrix calculation. In this way, as soon as computation is done, data is ready for another processing.

## 1.4 Dissertation Scope

The scope of this dissertation is to explore conventional machine learning techniques and deep learning methods for the identification of *S. pyogenes* from other selected pathogens and also control, not-pathogen using RS. The potential of RS as a novel diagnostic tool and the recent development in machine learning arena have motivated us to investigate a method which provides a real-time and end-to-end diagnostic without any expert intervention for analyzing the spectra.

In Chapter 2 the traditional machine learning techniques are explored for pathogen identification, and their performances are compared with each other due to provide a benchmark for a specific dataset. A unique deep neural network with an ability to provide end-to-end detection with zero expert intervention will be introduced in Chapter 3 to identify *S. pyogenes*. The performance of the proposed method will be discussed and compared with benchmark methods. In Chapter 4, the challenge of using such approach in the clinical application will be discussed to provide more insights into how this network performs.

## CHAPTER 2 IDENTIFICATION OF *S. PYOGENES* USING RAMAN SPECTROSCOPY: A COMPARATIVE STUDY ON MULTIVARIATE ANALYSES

### 2.1 Introduction

The rapid identification of pathogens and bacteria is a challenge for effective therapy of bacterial disease, and there is an urgent need for improving identification approaches. Traditional techniques require more effort and are time-consuming, and even some of these methods are destructive [77]. A real-time method that is not destructive to the matrix is desirable to detect pathogens. Raman Spectroscopy (RS) has been utilized to generate a spectrum of microorganisms and is considered a cellular-based phenotype system. The patterns generated by RS are based on molecular vibrations and can reveal the structure and composition of the samples. As a result, RS has been used to identify and characterize biological systems [78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91].

However, biological samples are complicated by the chemical composition that makes it challenging to interpret the patterns generated, and usually, the patterns have been compared with known spectra of the microorganisms. The bacteria are composed of biomolecules, such as proteins, lipids, carbohydrates and nucleic acids, and different bacteria often share similar biomolecules in their structure. Also, each biomolecule is composed of various molecules and macromolecules, and some of these molecules are common among other biomolecules.

Nonetheless, a molecule has a unique Raman spectrum, and the spectra of a bacterium are distinctive. Some studies have attempted to study the molecular fingerprints of some bacteria and distinguish critical bands or spectral markers to identify them [92, 83]. Also, a study has been attempted to distinguish different strains of the same species [93].

[29] attempted to characterize RS of *Streptococcus pyogenes* using SERS. The spectrum bands are identified for *Streptococcus pyogenes* and assigned to Ring breathing (adenine),  $\nu(\text{CNC})$  alanine, symmetric stretch CON, CC ring breathing (polysaccharide),  $\nu(\text{COC})$ , ring breathing, Amide III,  $\nu(\text{CO}_2)$  ( $\alpha$ -amino acids, COH (oligosaccharides), N-Acetyl related bands, and  $\delta(\text{NH})$  and  $\nu(\text{CN})$  amide II. The spectrum of other pathogens contains similar bands such as nucleic acids, protein, and carbohydrates. However, their relative intensities are different among various species. In Table 1, the assignment of Raman bands is shown, and can be used as a reference to analyze the *S. pyogenes* spectra. Such characterization methods are aimed at distinguishing spectral markers and then using them to detect or identify the pathogens.

Sometimes a small variation in the spectra might contain critical information of a species. The statistical analyses have been employed to interpret the Raman spectra. The multivariate statistical analysis techniques are grouped into two main categories of supervised and unsupervised methods. These techniques can be used individually or in combination. For instance, an unsupervised method such as principal component analysis (PCA) is used to reduce the high dimension raw spectra to a few variables such that the relevant information is preserved. Then, supervised methods are used to discriminate the principal components of different bacteria [94]. Unsupervised techniques rely on dividing species into distinct groups in the form of a cluster such as PCA [95] or a dendrogram such as hierarchical cluster analysis (HCA). These methods are based on the significant latent variables preserving the information of biochemical components within the bacterial cell which contribute to the identification of the bacteria [96]. Supervised methods provide qualitative or quantitative analysis and require reference values. These methods are usually built upon a prior

unsupervised model.

Discriminant function analysis (DFA) is a robust technique that is applied to spectra analysis and has been used in pathogen identification studies [97]. LDA and QDA use a set of independent variables to predict group membership of each spectrum given to maximize the separation between groups. In other words, it attempts to minimize the variance within groups and maximize the variance between groups [98]. Support vector machine (SVM) [99, 100, 101] has been widely applied to RS data to classify various pathogens and bacteria [102, 103]. SVM algorithms are based on finding the optimal boundary among classes by solving a constrained optimization problem. Using kernel trick enables the SVM models to perform non-linear classification. Decision trees can handle high-dimensional data better. However, it is reported that it can overfit to training. There have been many efforts to improve its primary drawback which has led to various tree-based algorithms [104, 105]. The ensembles of trees were one of the best ways proposed to improve prediction accuracy and overfitting habits of decision trees. Among these algorithms, Random Forest is one of the most attractive methods [104, 106]. It contracts each tree using a bootstrap sample of the training data set (bagging), and also each node is split using a subset of predictors (features) randomly selected at that node. It has been shown that this algorithm performs very well and is very robust against overfitting [107].

The goal of this chapter is to provide a comparative evaluation of traditional machine learning and multivariate algorithms including LDA, QDA, SVM, and Random Forest. Recent studies of RS have used various algorithms for pathogen identification, and the algorithms used in this study provide a representative set of discrimination and classification methods commonly used in RS analysis. The purpose is to obtain a concise understanding

of pros and cons of these algorithms for pathogen identification. Furthermore, an additional objective is to provide an understanding of the efficiency of PCA algorithms through the bias-variance analysis of these techniques.

## 2.2 Material and Method

### 2.2.1 Instrumentation and Sample Preparation

The bacterial specimen, *Streptococcus pyogenes* ATCC 49117 was prepared from bacteria plated on tryptic soy agar plates. A single isolated colony from the stock bacteria streaked nutrient agar plate was picked and added to 5 ml of tryptic soy broth in a 10ml culture tube. This culture tube was placed on a shaker in a 37 ° C incubator and incubated overnight (18 hours). The overnight culture was centrifuged at room temperature for 5 min @ 3500 rpms. The supernatant was removed, and the bacteria pellet was re-suspended in 5ml of filtered (sterilized) tap water. The bacteria were centrifuged, and the washing process was repeated once. After the final wash, filtered tap water was added to the bacteria pellet until the optical density (measured at a wavelength of 600 nm) of the solution was  $1.00 \pm 0.05$ . A 0.15 ml volume of the bacteria suspension was then placed on a mirror polished stainless steel substrate (alloy 304, Stainless Supply) for Raman analysis. Also, a 0.15 ml volume of the filtered tap water was prepared and then placed on a mirror polished stainless steel substrate (alloy 304, Stainless Supply). In addition to this data, other pathogens including *E. coli* (K99), MRSA, MSSA, *Legionella pneumophila*, and *Pseudomonas aeruginosa* were acquired in similar fashion.

RS data were acquired using an inVia Raman microscope (Renishaw) equipped with a 50 mW 514.5nm laser as the excitation source, an 1800 l/mm grating, a 576 x 400-pixel



thermoelectric cooled charge-coupled device, and WiRE 3.3 software. The laser light was focused onto the substrate through a 63X dipping objective (Leica HCX PL APO 1.2NA Corr/0.17 CS. Spectra were acquired with 37mW of laser power at the sample over a spectral range of 400-3200  $\text{cm}^{-1}$  with 40 accumulations at an integration time of 10s. Figure 1 shows the basic setup of RS.

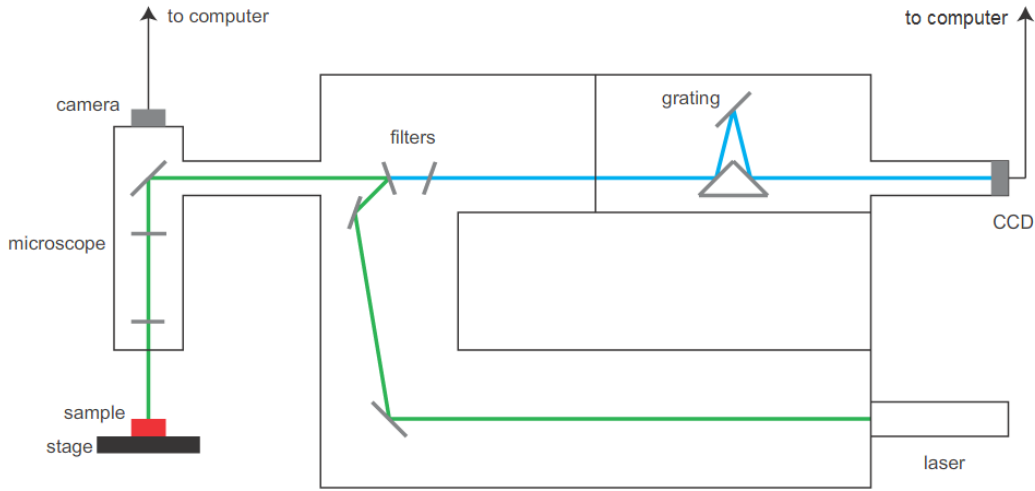


Figure 1: Experimental Setup of Raman Spectroscopy [4].

### 2.2.2 Dataset

The Raman spectra of the *S. pyogenes* with water background and other species including filtered water, MRSA, MSSA strain, *Legionella pneumophila*, *Pseudomonas aeruginosa* and *E.coli* (K99) were acquired with the method explained in section 2.2.1. The dataset summary is represented in Table 2.

### 2.2.3 Statistical Analysis

**Pre-processing** Raman spectrum is needed for pre-processing before further analysis. Pre-processing consists of removing cosmic rays, noise, and tissue fluorescence, and normalization of data. In order to remove fluorescence, different approaches have been proposed. For

Dataset	Biological Species	count #
S. pyogenes	S. pyogenes	101
	MRSA	78
Not-S. pyogenes	MSSA	53
	filtered water	29
	E. coli (K99)	50
	Legionella pneumophila	20
	Pseudomonas aeruginosa	8
Total:		339

Table 2: Dataset Summary.

example, a polynomial plot can be approximated to local minima of the spectra [108]. The background subtraction is based on the florescence-to-signal ratios to minimize the residual mean square (RMS) error. Such methods attempt to fit the best polynomial fit for each spectrum. In this section, morphological weighted penalized least squares (MPLS) was used to remove the baseline [109, 110]. Also, it is possible to use wavelet to remove noise and smooth the data.

Each spectrum has 1368 samples in the range of 400-2472  $\text{cm}^{-1}$ . Next, the data was min/max normalized to get the values between zero and one. The data was split into training data (80%) and test data (20%). For cross-validation (CV), the  $k$ -fold method is used where  $k = 4$  [111].

**Multivariate Analysis** In order to organize relative to intra-spectral similarities, the hierarchical cluster analysis (HCA) was performed to the data. HCA is an unsupervised method, and all mean spectra of the species were fed into the HCA.

Due to high-dimensionality of Raman spectra, PCA was applied to the data, so that the components with the most variability within dataset would be preserved [112]. It helps

to condense many samples in each spectrum to only a few variables where these variables contain at least 95% of the variance. It can reduce the variance by decreasing the complexity of the models.

Discriminant function analysis (DFA) is a multivariate data analysis technique used for classification of data. Although Raman Spectra cannot meet the requirements of DFA statistically [113], this method can be applied to principal components of the spectra.

DFA uses a set of independent variables to predict group membership of each spectrum given to maximize the separation between groups. The linear discriminant function attempts to minimize variance within groups and maximize variance between groups [97].

Following PCA, support vector machine (SVM) was used to classify the samples. SVM is a supervised classifier that provides a maximum margin of separation between classes [99, 100, 101]. Several kernel functions, including poly, linear, rbf, and sigmoid, were explored to map the data in a different space. A grid search was conducted on the training dataset to optimize the values of  $\gamma$  and  $C$ .  $C$  value is a parameter to control correctly classifying the samples and the smoothness of the decision boundary where higher  $C$  values result in reducing misclassification error and lower  $C$  values result in a smoother decision boundary. The  $\gamma$  value indicates how far the influence of a single training point reaches such that the classifier attempts to reach far samples when the  $\gamma$  value is higher.

A decision tree based algorithm, Random Forest, was explored since it has been presented as a powerful tool to ensemble decision trees [114]. Random Forest bootstraps samples of the training data and randomly chooses features to build trees. A grid search has been conducted to select the best model according to the number of trees, cost function and a minimum number of samples to split.

Finally, all models were evaluated on a test database where none of the spectra from this dataset were used in training the corresponding models.

## 2.3 Result and Discussion

### 2.3.1 Data visualization

Figure 2 illustrates the mean and standard deviation of raw data which each spectrum normalized to its corresponding maximum intensity. The background removal of the spectra is an essential step to distinguish *S. pyogenes* from other species. Figure 3 shows the background removed spectra of each species using MPLS. Although there is a significant difference between water and other bacteria, the bacteria have several bands in common. Nonetheless, the intensity of spectra is various among them.

### 2.3.2 HCA

The average spectra of each species were fed to the HCA in order to explore the intra-spectral similarities. Figure 4 illustrates the dendrogram and the corresponding distance pairs computed by ward method [115]. The plot shows that the spectra can be divided into two main groups of bacteria and non-Bacteria. Also, it shows that *E. coli* have a different spectrum from the rest of the species. The MRSA and MSSA, the two strains of staphylococcus pathogen, are in one subgroup revealing their high similarity. *P. aeruginosa* spectra show more similarity to *S. pyogenes* than the staphylococcus although the *P. aeruginosa* is a gram-negative pathogen. Also, the *Legionella pneumophila* illustrates the highest similarity to *S. pyogenes*. It can be concluded that the problem of discriminating the *S. pyogenes* from other species in this dataset is not only about the identification of a pathogenic species from a non-pathogenic species but also about the identification of a specific pathogen from other

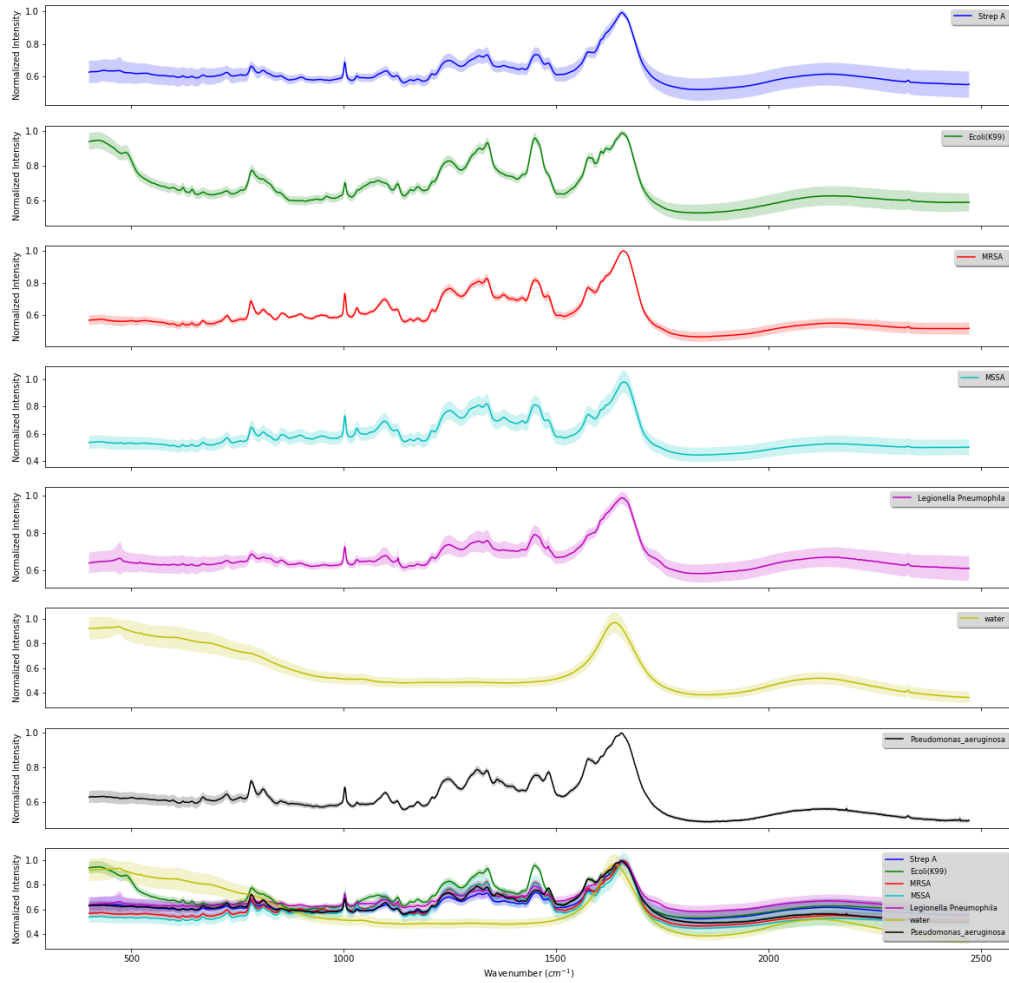


Figure 2: Mean and Standard Deviation of Raw Data Normalized to Maximum Intensity.

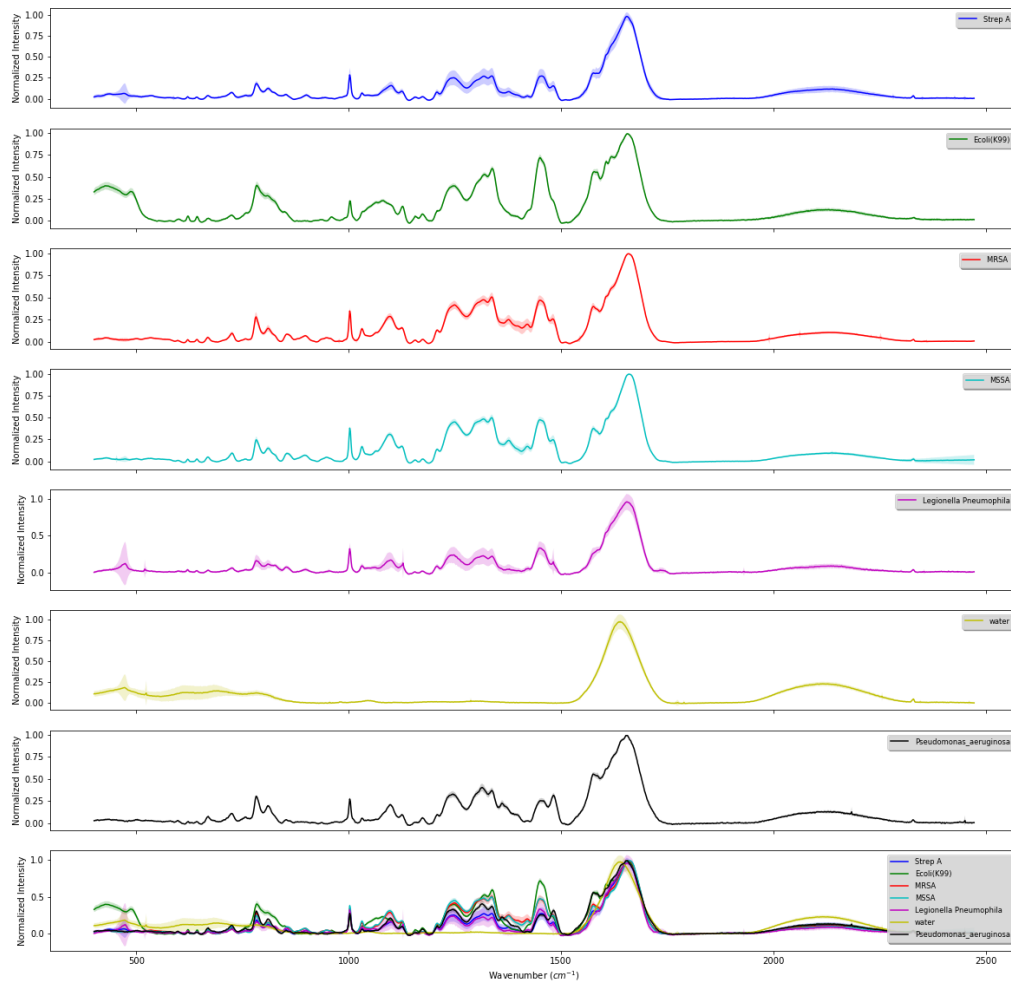


Figure 3: Mean and Standard Deviation of Background Removed Spectra Normalized to Maximum Intensity.

pathogenic species with very similar spectra.

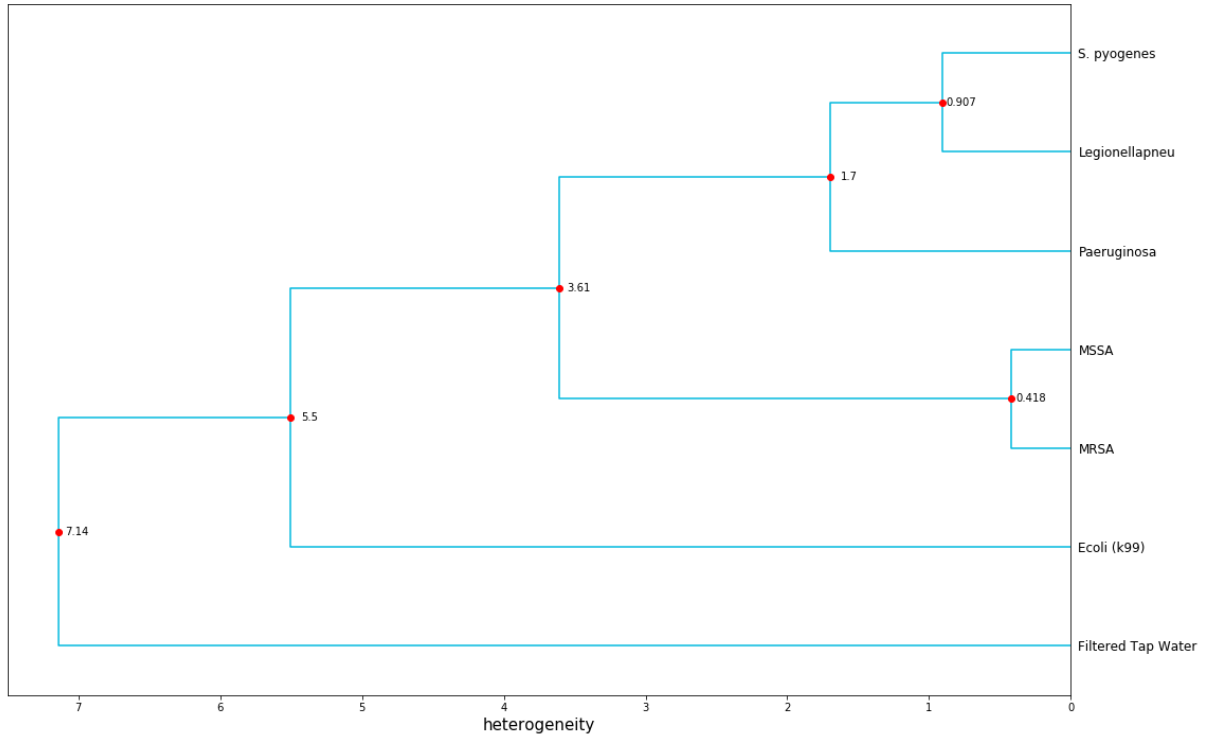


Figure 4: Hierarchical Cluster Analysis Performed on Averaged Raman Spectra of Seven Species used in This Study. The red dot represents the distance between clusters calculated by ward method. The dendrogram shows a clear separation of the tap water from pathogens. E.coli is the least similar to *S. pyogenes*, and the spectrum of *Legionella pneumophila* and *Pseudomonas aeruginosa* is very close to the *S. pyogenes*. The MRSA and MSSA can be categorized in one cluster as they are strains of *Staphylococcus*.

### 2.3.3 PCA

The principal component analysis (PCA) was applied to the training dataset. PCA reduces the dimensionality of the spectra from 1368 down to a few linearly uncorrelated PC scores. Figure 5 shows the cumulative explained variance ratio for the first 10 principal components. The first two PCs represent 80% of the explained variance, and it can be seen that after five components the variance changes insignificantly as the first five significant components preserve 95.89% of the variance. Thus, five first principal components were

selected to be fed to the classification methods of LDA, QDA, and SVM. Although selecting a higher number of PCs can decrease the bias, it also results in a high variant model. As the number of training data is not very large, it is critical to simplifying the model as a high-variance method is at risk of overfitting to the noise or unrepresentative training data (e.g., the fluorescent background).

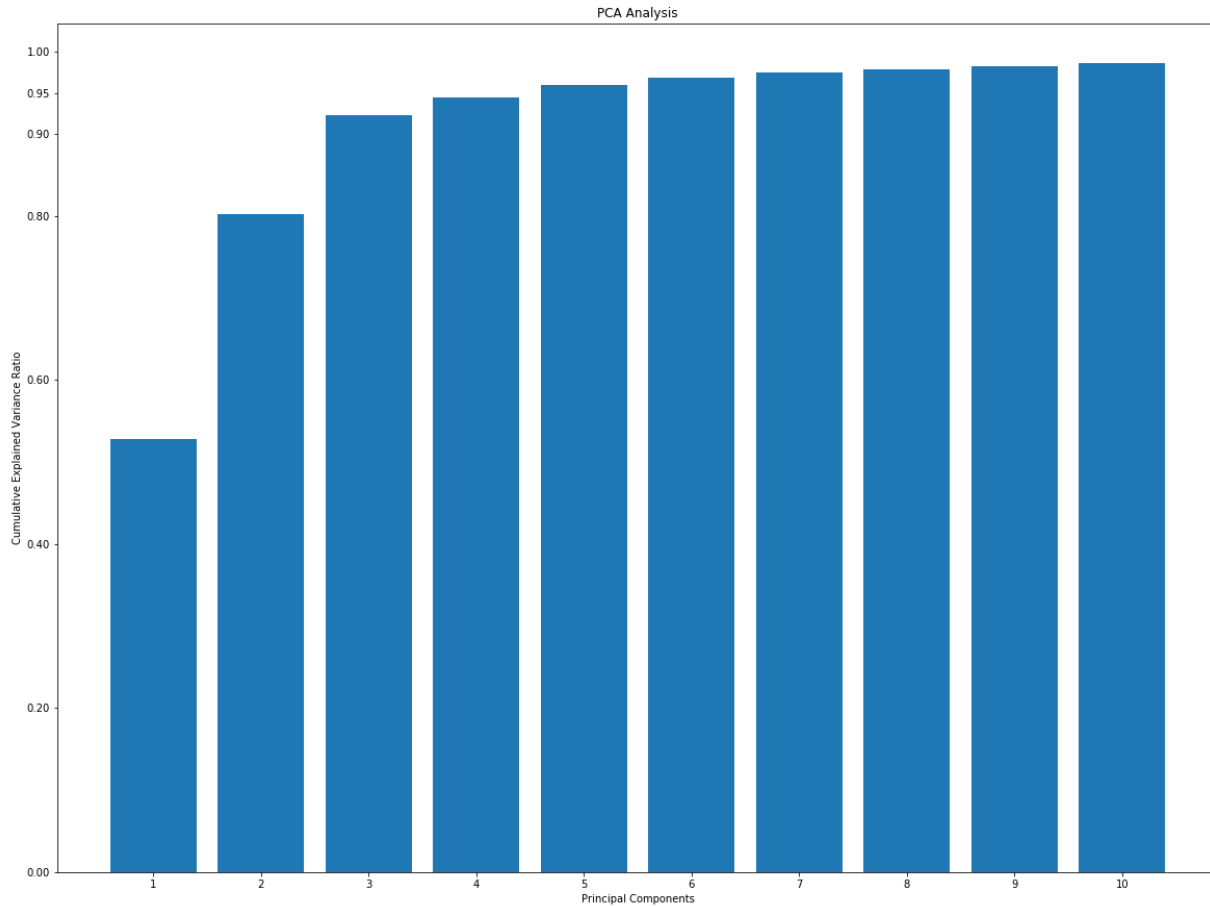


Figure 5: Cumulative Explained Variance Ratio for First 10 Principal Components.

### 2.3.4 Discriminant Function Analysis

A discriminant function analysis was applied to the PC scores of the dataset to discriminate *S. pyogenes*. The linear and quadric discriminant analyses were applied to the training dataset. The former one assumed the covariance of both groups to be the same while the



later one considered a different covariance for each group.

**LDA** The trained LDA classifier correctly classified 70.84% of the training spectra. Cross-validation was done based on  $k$ -fold, where  $k$  was 4, and resulted in 69.09% accuracy. The result reveals that the model is underfitting the data and better approaches need to be explored. The model was examined on the test dataset, from randomly sampled data and not used for obtaining model parameters, to test the robustness of the model, It classified 66.17% of the test spectra correctly. The reduction in the classification accuracy of the test dataset illustrates that the generalization error of the model is not low and the difference between cross-validation error and test misclassification error might originate from the sampling error.

In Figure 6, the mapped data on the first two PCs are shown. This plot illustrates the discrimination function results. It can be seen that some of the spectra were misclassified in the training dataset. The solid blue corresponds to the *S. pyogenes* pathogen, and the red circles correspond to the Not-*S. pyogenes* species. The circles display two times the standard deviation for each class. The black dot is the mean value of each class. The misclassified samples are represented by dark blue and dark red corresponding to *S. pyogenes* and Not-*S. pyogenes*, respectively.

**QDA** Using QDA, the classification accuracy on training, validation, and test dataset was 78.97%, 77.73%, and 76.47%, respectively. The results suggest that the QDA is better than LDA as its misclassification error is lower. Moreover, there is less over-fitting in QDA compared to LDA. Nevertheless, QDA is under-fitting the data, and more robust approach needs to be explored. Figure 7 illustrates the result of applying QDA on two PCs. It can be

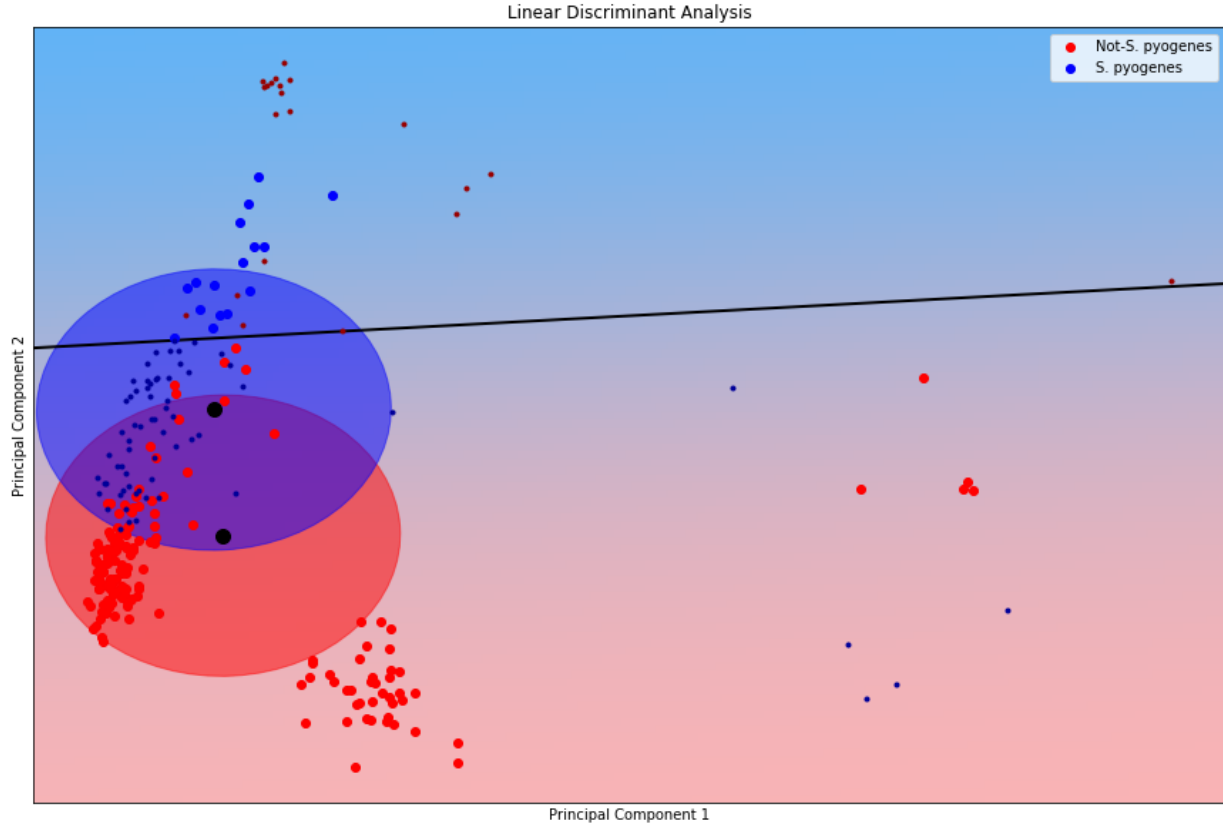


Figure 6: Linear Function Analysis to Discriminate *S. pyogenes* from Not-*S. pyogenes*. The solid blue corresponds to the *S. pyogenes* pathogen and the red circles correspond to the Not-*S. pyogenes* species. The circles display two times the standard deviation for each class. The black dot is the mean value of each class. The misclassified samples are represented by dark blue and dark red corresponding to *S. pyogenes* and Not-*S. pyogenes*, respectively.

seen that two times the standard deviation is in the shape of ellipsoids as it assumes that covariance of the classes is different.

### 2.3.5 SVM

In this subsection, the result of the SVM method on the dataset is presented. As mentioned a grid search has been conducted on  $C$ , and  $\gamma$  values and usage of the different kernels are explored. Figure 8 illustrates how the grid search was conducted for each kernel. An initial logarithmic search on  $C$  and  $\gamma$  values were applied, and an area with a better validation accuracy was selected, and then a fine-tuning was conducted to find the best parameters.

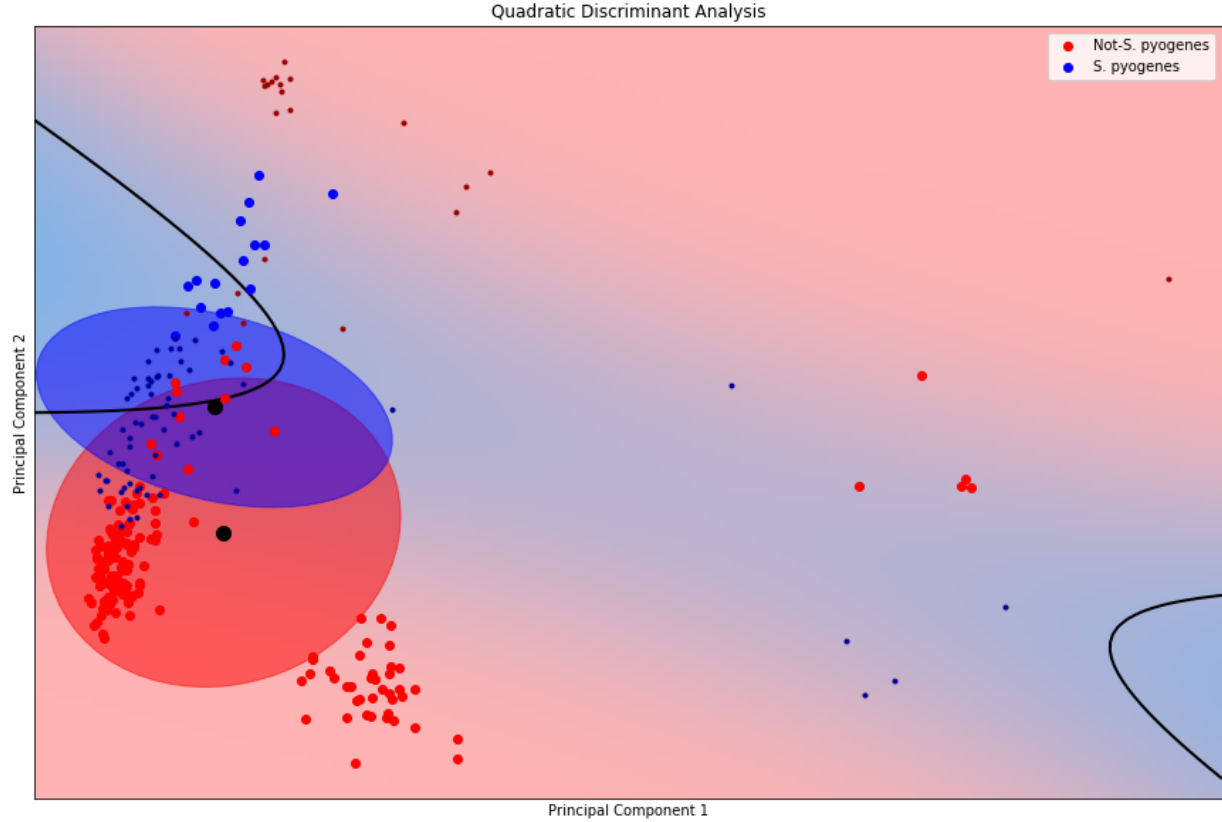


Figure 7: Quadratic Function Analysis to Discriminate *S. pyogenes* from Not-*S. pyogenes*. The solid blue corresponds to the *S. pyogenes* pathogen and the red circles correspond to the Not- *S. pyogenes* species. The ellipsoids display two times the standard deviation for each class. The black dot is the mean value of each class. The misclassified samples are represented by dark blue and dark red corresponding to *S. pyogenes* and Not-*S. pyogenes*, respectively.

Similarly, the best parameters of  $C$  and  $\gamma$  for poly and sigmoid kernels was found. The Figure 9 summarizes the classification accuracy for different kernels. The radial kernel results in the best classification accuracy with 93.4% and 91.17% accuracy on the validation and test dataset, respectively. Also, training accuracy is 95.94% revealing the fact that the under-fitting of the model to the data improved significantly compared to the QDA. The poly kernel shows the over-fitting to the data as the accuracy is decreased dramatically on the validation dataset although the training accuracy is the highest one among the other kernels, 97.04%. The linear and sigmoid kernels are under-fitting to the dataset with a training accuracy of

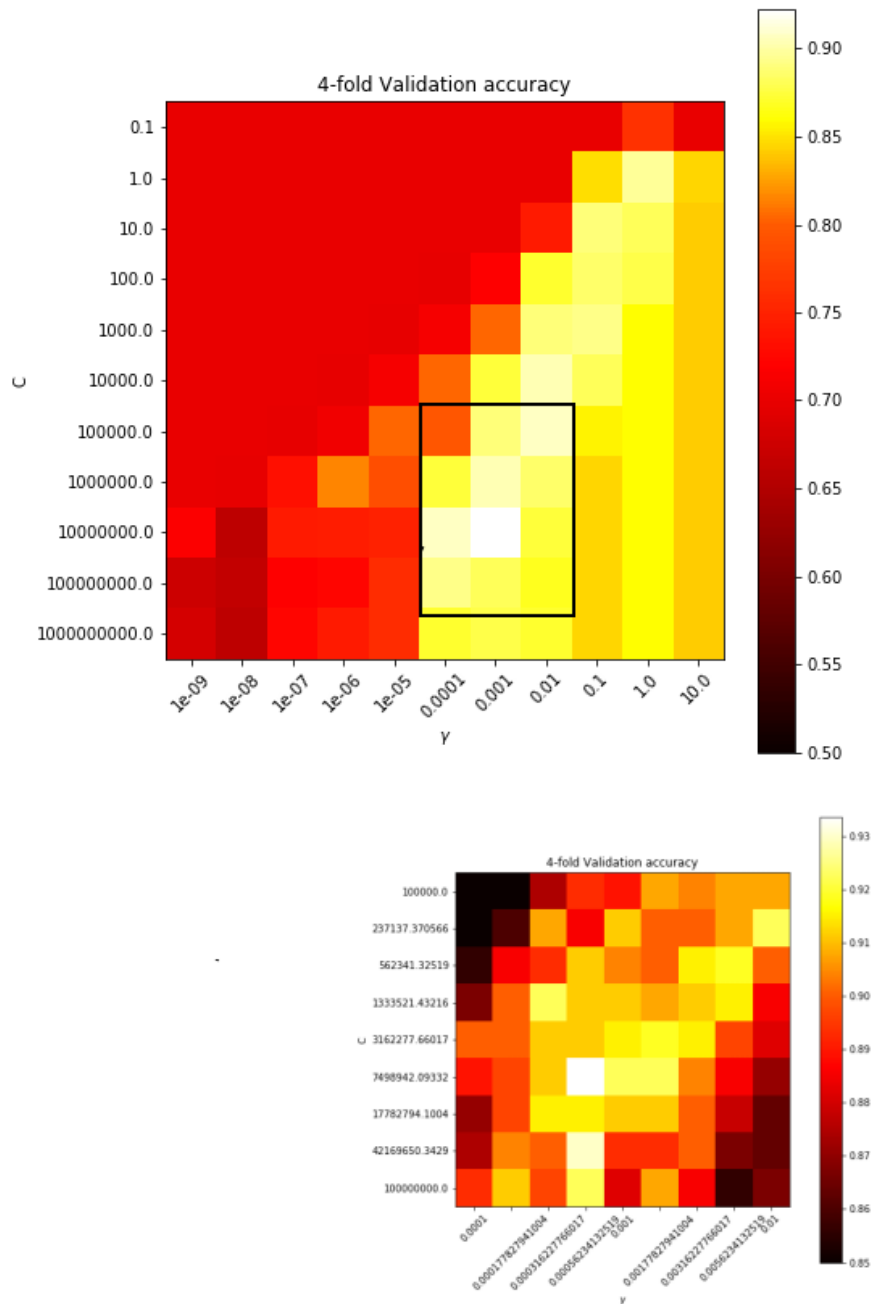


Figure 8: Grid Search for Validation Accuracy. a) An initial logarithmic search on  $C$  and  $\gamma$  values where the kernel was 'rbf' and the 4-fold cross-validation is used. b) The fine-tuning for  $10^{-4} < \gamma < 10^{-2}$  and  $10^5 < C < 10^8$ .

80.44% and 75.27%, respectively.

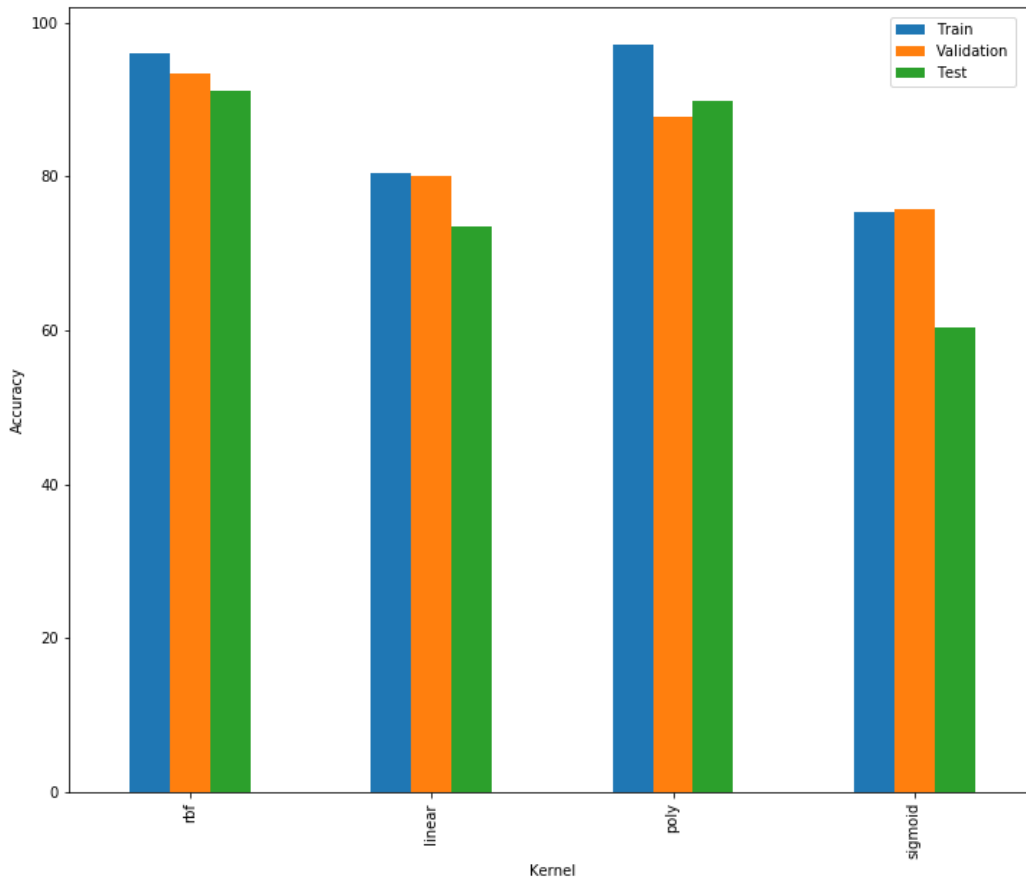


Figure 9: SVM Classification Accuracy for Different Kernels.

### 2.3.6 The Effect of the Number of PCs on the Classification

The number of PCs is usually determined by the variance within the original data each PC describes. So far, the number of PCs was considered as a fixed parameter in our analyses. In this section, the effect of selecting a different number of PCs on the classification results is explored in detail. The number of PCs is one of the leading parameters to control bias and variance of PCA-LDA, PCA-QDA, and PCA-SVM methods. Also, in SVM algorithms, the  $C$  and  $\gamma$  parameters can control the bias and variance of the model. The dimensionality

of the training data is an inevitable factor which can affect the model significantly. As it is desired to improve classification accuracy, it is beneficial to generalize the model as well such that the proposed model will be more robust to the variations of the spectra due to noises or other artifacts. The approach mentioned in [116], is used to compute the bias and variance of the models.

In order to analyze this effect, Gaussian noises with a different signal-to-noise ratio (SNR), from less noisy (high SNR) to high noisy (low SNR) are aggregated to the dataset. First, noises are added to all dataset, training and validation set, and bias and variance are computed for each model. Figure 10, 11, and 12 illustrate the bias and variance of PCA-LDA, PCA-QDA, and PCA-SVM in terms of the number of PCs respectively where noises with various SNR are added to all spectra. It can be seen that the fluctuation of bias and variance is reduced above 14 PCs in PCA-LDA, PCA-QDA, and PCA-SVM. Furthermore, bias and variance are reduced as the number of PCs is increased. The reduction in bias and variance simultaneously might indicate that the PCA can model the noise in data such that the general pattern of the bias and variance of the model is not changed although they are higher for lower SNR.

Secondly, noises with various SNR are only added to the validation set in order to understand if these methods can predict spectra correctly where there is some aberration in the spectra. Thus, the bias and variance of the models are calculated for a different number of PCs. Figure 13, 14, and 15 reveals that there is a trade-off between bias and variance for SNR=1, 10dB (highly noised data) above 14 PCs for PCA-LDA and PCA-SVM and above 22 PCs for PCA-QDA, whereas for SNR above 20dB, the bias and variance remain constant. Thus, the PCA-based models have high ability to predict the spectra correctly where the

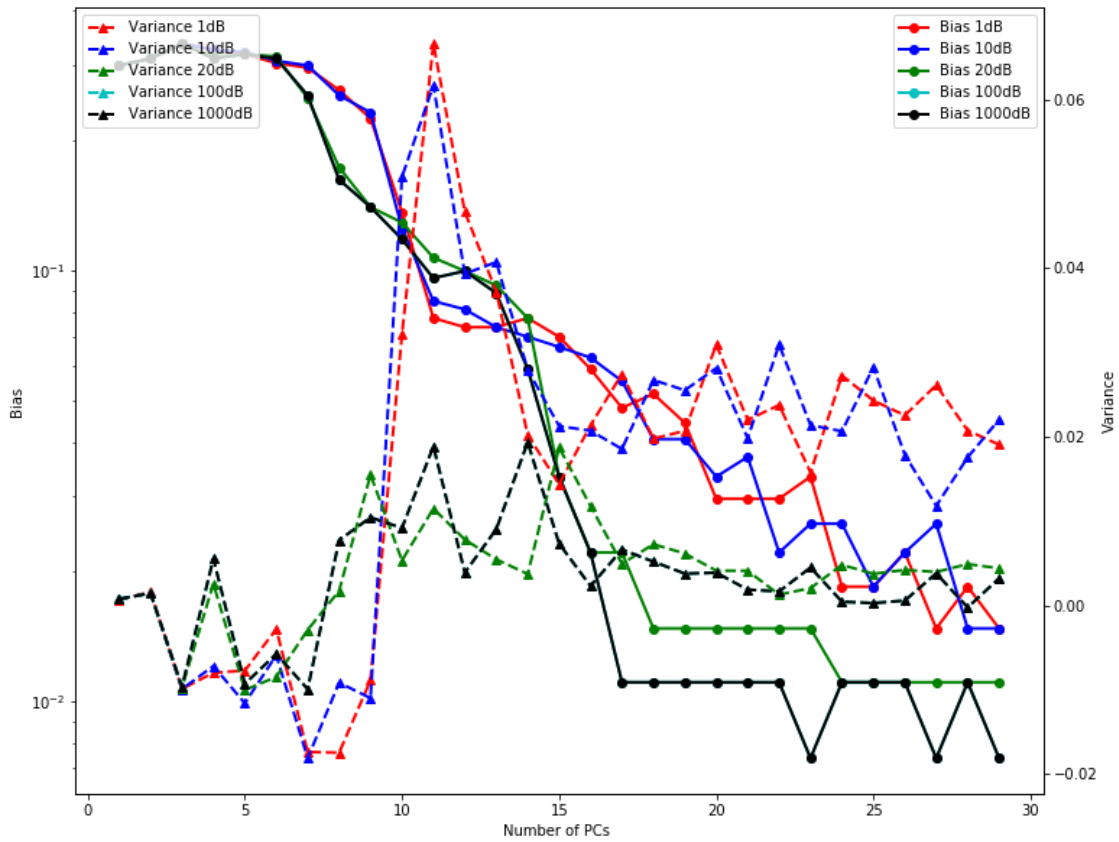


Figure 10: Bias and Variance of PCA-LDA in Terms of Number of PCs Where Gaussian Noises Are Added to All Spectra

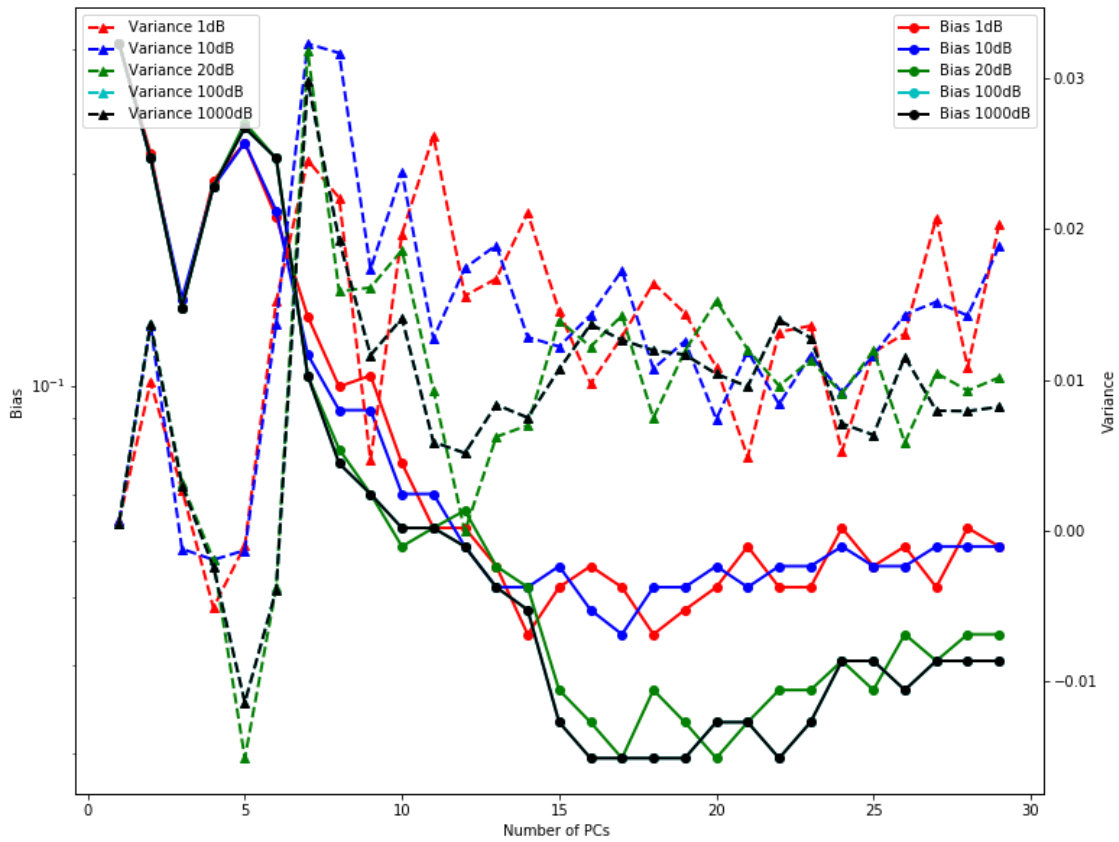


Figure 11: Bias and Variance of PCA-QDA in Terms of Number of PCs Where Gaussian Noises Are Added to All Spectra



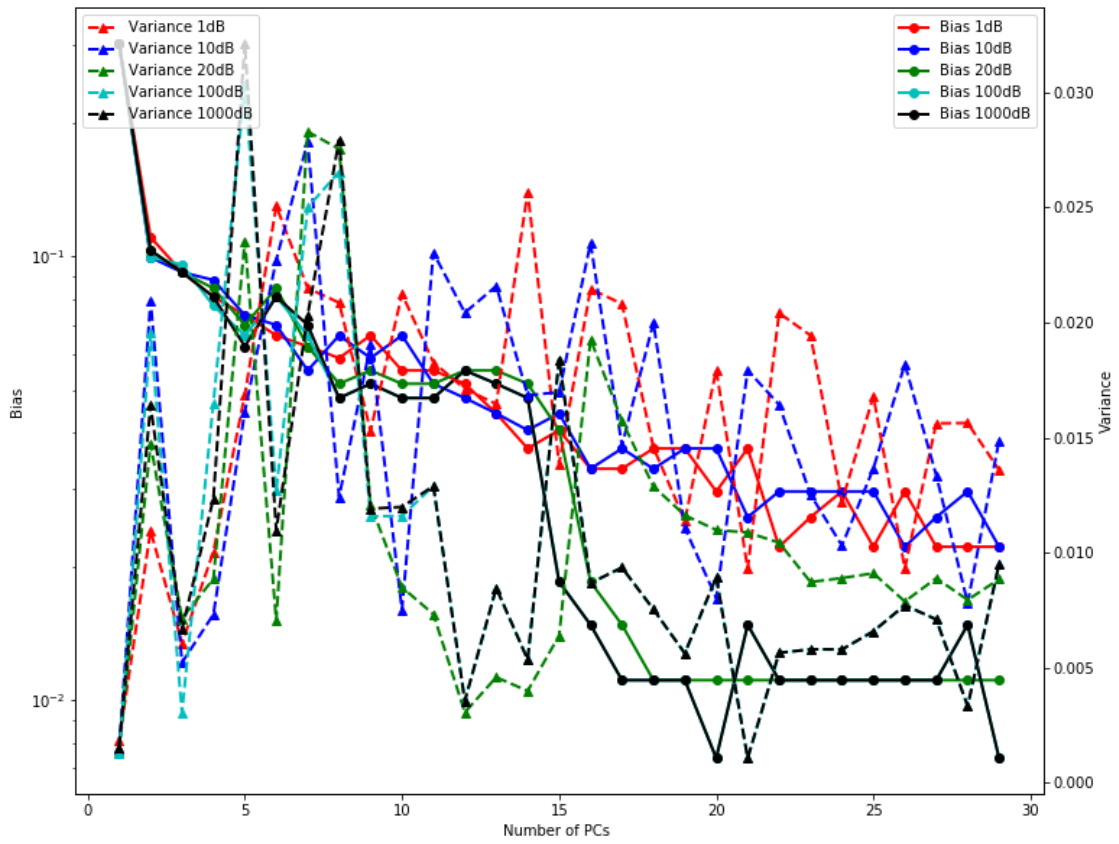


Figure 12: Bias and Variance of PCA-SVM in Terms of Number of PCs Where Gaussian Noises Are Added to All Spectra

testing or validation dataset has less variation from the original dataset. Also, it is notable that above 22 PCs a trade-off can be seen clearly in the PCA-QDA plot. However, the bias decreases for PCs above 6 until 14 and increases for PCs above 14. The variance increases drastically around 6 PCs and stays constant for PCs above that until the number of PCs reaches to around 22 PCs. It suggests that a selection of around 6 or 22 PCs would be good.

Furthermore, Figure 15 indicates that there is less variation of bias and variance for PCA-SVM above 5 principal components until 14 principal components. It suggests that any selection of PCs from 5 to 14 has a similar effect on variance and bias. Nevertheless, the model with a lower number of PCs has less complexity.

Consequently, the bias-variance analysis of the PCA-based model can reveal the significant amount of information where only relying on the contribution of each PCs within the variance of the data might be misleading as the overall performance of a well-trained model can drastically be reduced by increasing the complexity of the model.

In addition, the results suggest that the accuracy of these algorithms can be improved by selecting the optimal number of PCs such that there is less overfitting or underfitting to the data.

### **2.3.7 Random Forest**

The Random Forest method is applied on the wavenumber of the spectra where the training of trees used bootstrap aggregating or bagging. A grid search explores the number of trees or estimators in addition to a number of features or wavenumber used for training of each tree.

Figure the 16 shows the average 4-Fold cross-validation error for the various choices of the number of trees and descriptors. The ensemble of 25 trees with 5 descriptors resulted

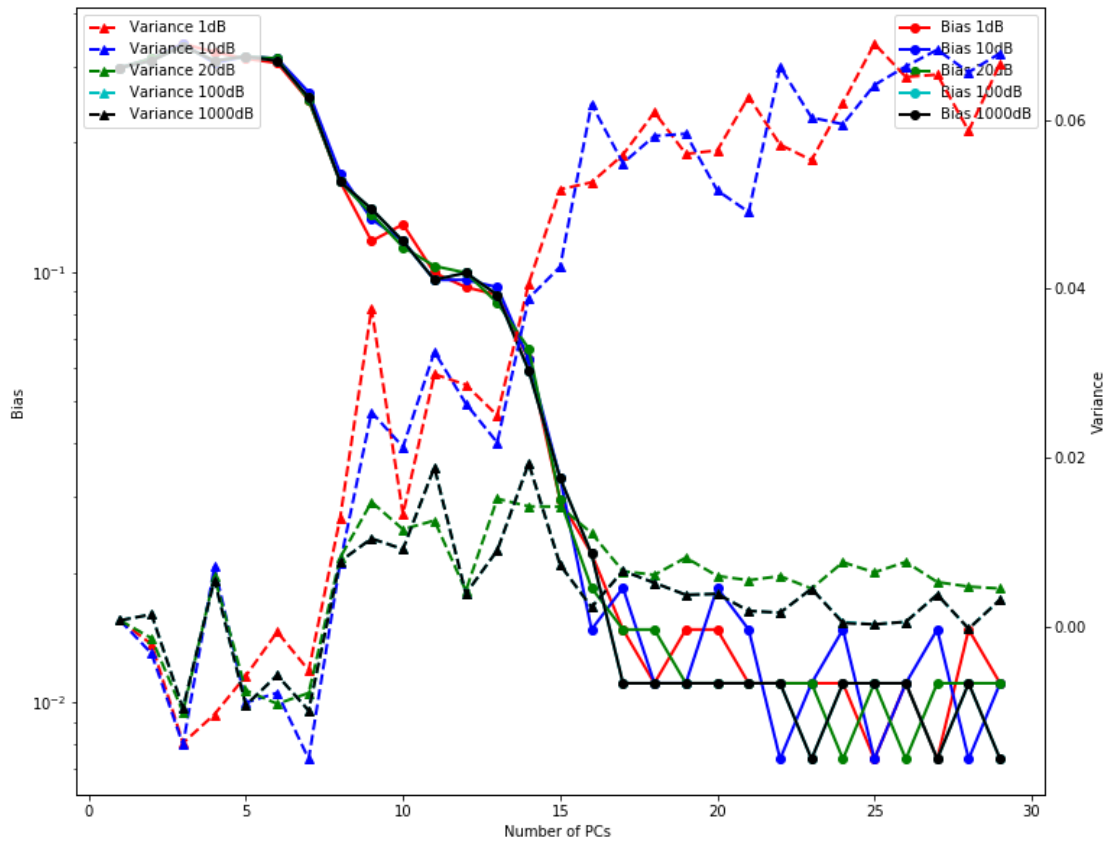


Figure 13: Bias and Variance of PCA-LDA in Terms of Number of PCs Where Gaussian Noises Are Added to the Validation Set

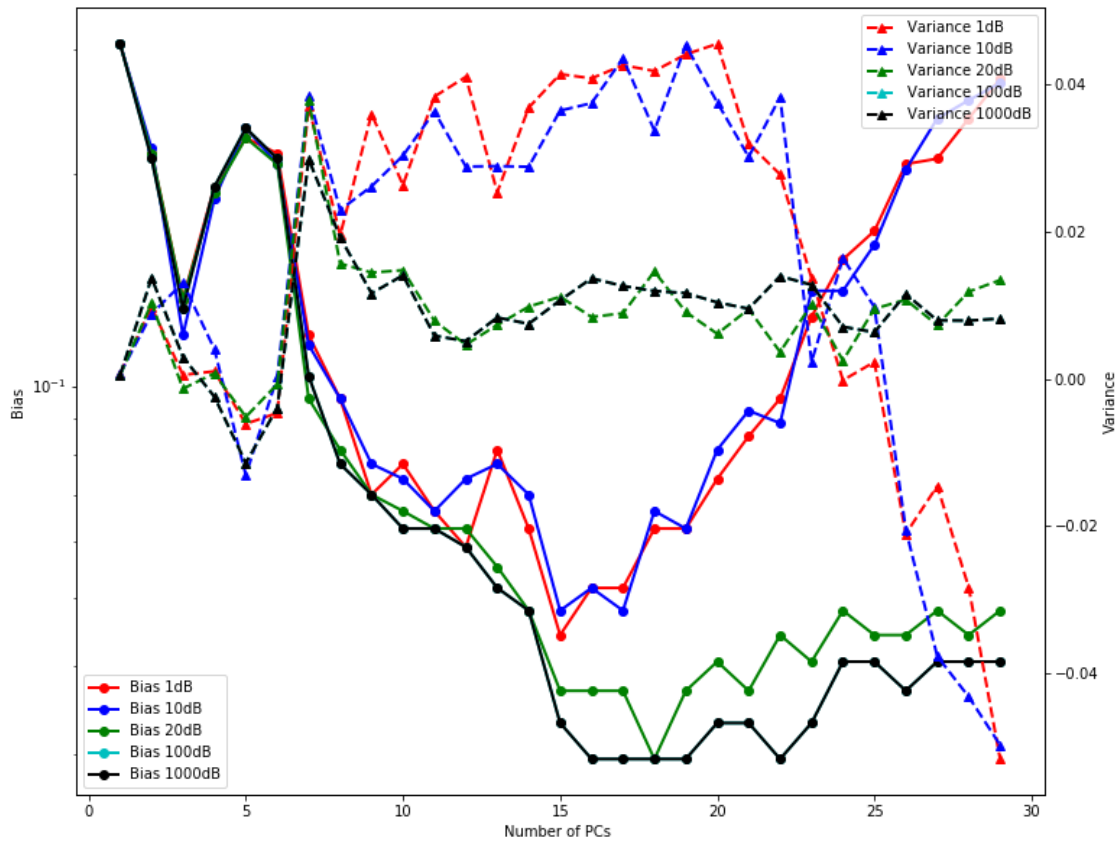


Figure 14: Bias and Variance of PCA-QDA in Terms of Number of PCs Where Gaussian Noises Are Added to the Validation Set

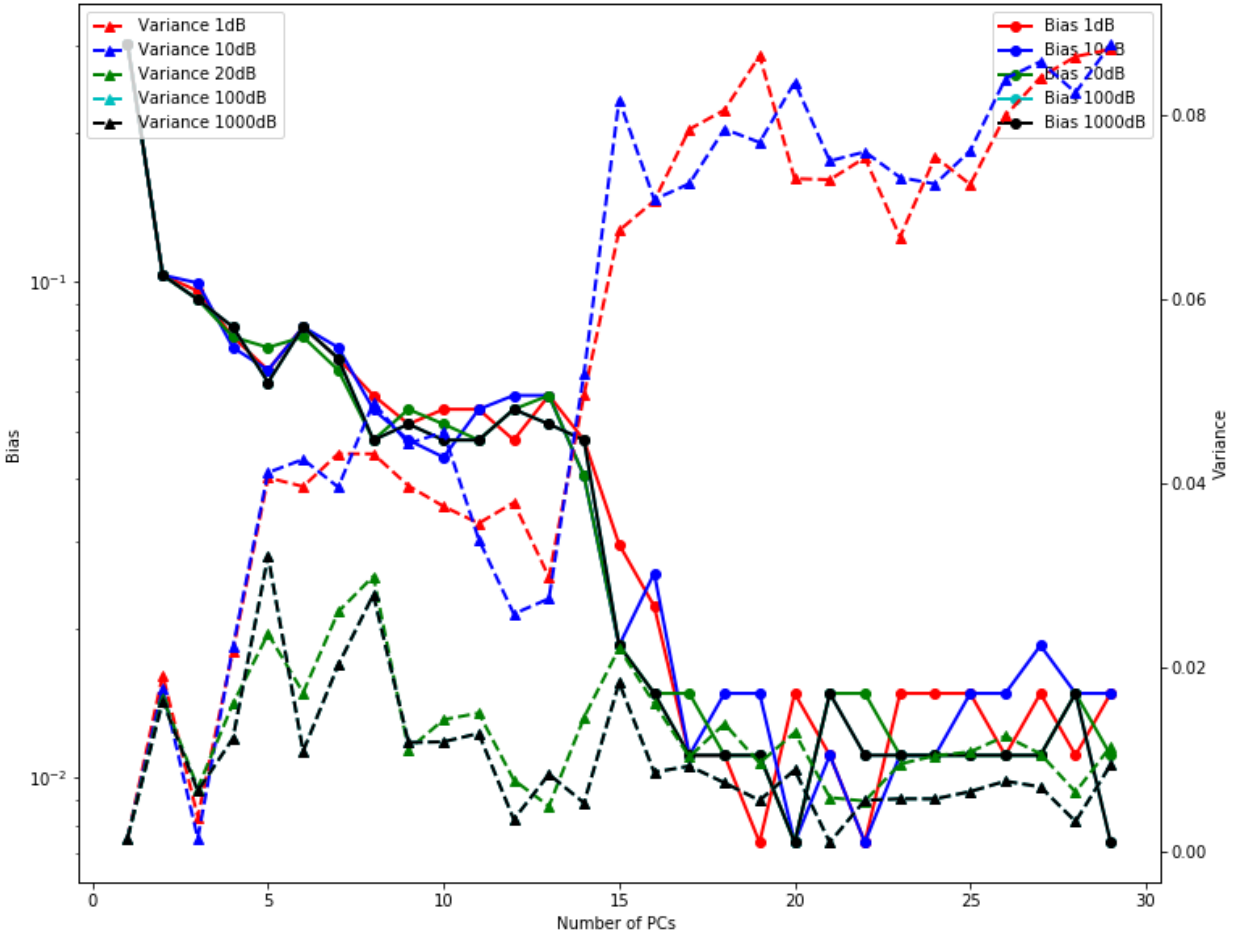


Figure 15: Bias and Variance of PCA-SVM in Terms of Number of PCs Where Gaussian Noises Are Added to the Validation Set

in the best cross-validation accuracy of 94.46%. Training accuracy was 100.00% suggesting that there is no under-fitting of the model to the dataset. It also illustrates that the number of descriptors is the primary parameter to tune the Random Forest performance once there is a sufficient number of trees (around 15).

The usage of different cost functions or changing minimum node size does not affect the result significantly where the best parameters for a number of estimators and descriptors are set. The default value of 1 for minimum node size is chosen in this study. Hence, the trees were grown to their maximum size.

### **2.3.8 Comparative Result**

In this section, the result of different methods on the testing dataset is studied. Table 4 shows the results of the testing dataset. The Random Forest algorithm showed the highest classification accuracy on the test dataset compared to the other method. SVM has higher accuracy compared to the QDA and LDA, and it can be concluded that SVM is the best candidate where the PCA is used to reduce the dimensionality of the spectra. However, the Random Forest showed that randomly choosing wavenumber can result in better accuracy on the test dataset, and it helps to generalize the algorithm for the more extensive or various data. QDA showed a better accuracy compared to LDA, revealing that the assumption of the same covariance for both classes is not correct and the nonlinear kernel can result in a better model when the complexity of the dataset is increased.

The ROC of the different methods is plotted in Figure 17 to explore the sensitivity and specificity of the models. The Random Forest shows the best area under the curve of 0.997 compared to the other methods, and its ROC is above the ROC of the others for all thresholds. It indicates the Random Forest is the best approach for pathogen identification

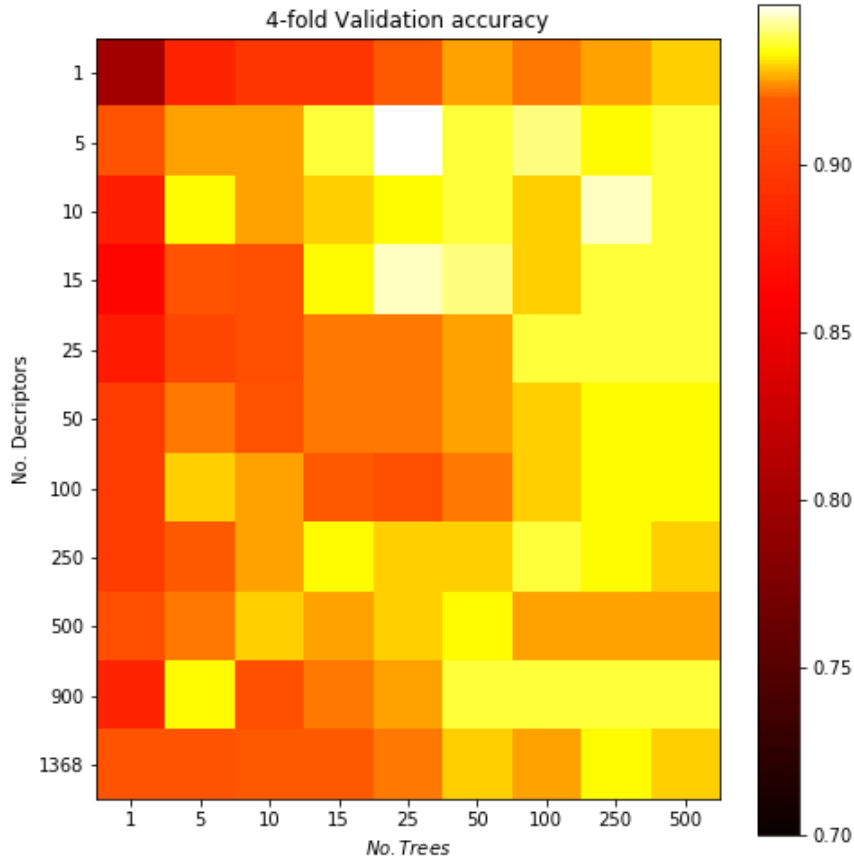


Figure 16: Grid Search on Average 4-Fold Cross-Validation Accuracy. 25 trees with 5 descriptors are the best parameter of Random Forest result in 94.46% accuracy.

Methods	Accuracy %
<b>Random Forest</b>	<b>94.11</b>
PCA-rbf-SVM	91.17
PCA-linear-SVM	73.52
PCA-poly-SVM	89.70
PCA-sigmoid-SVM	60.29
PCA-LDA	66.17
PCA-QDA	76.47

Table 4: Classification Accuracy on Testing Dataset.

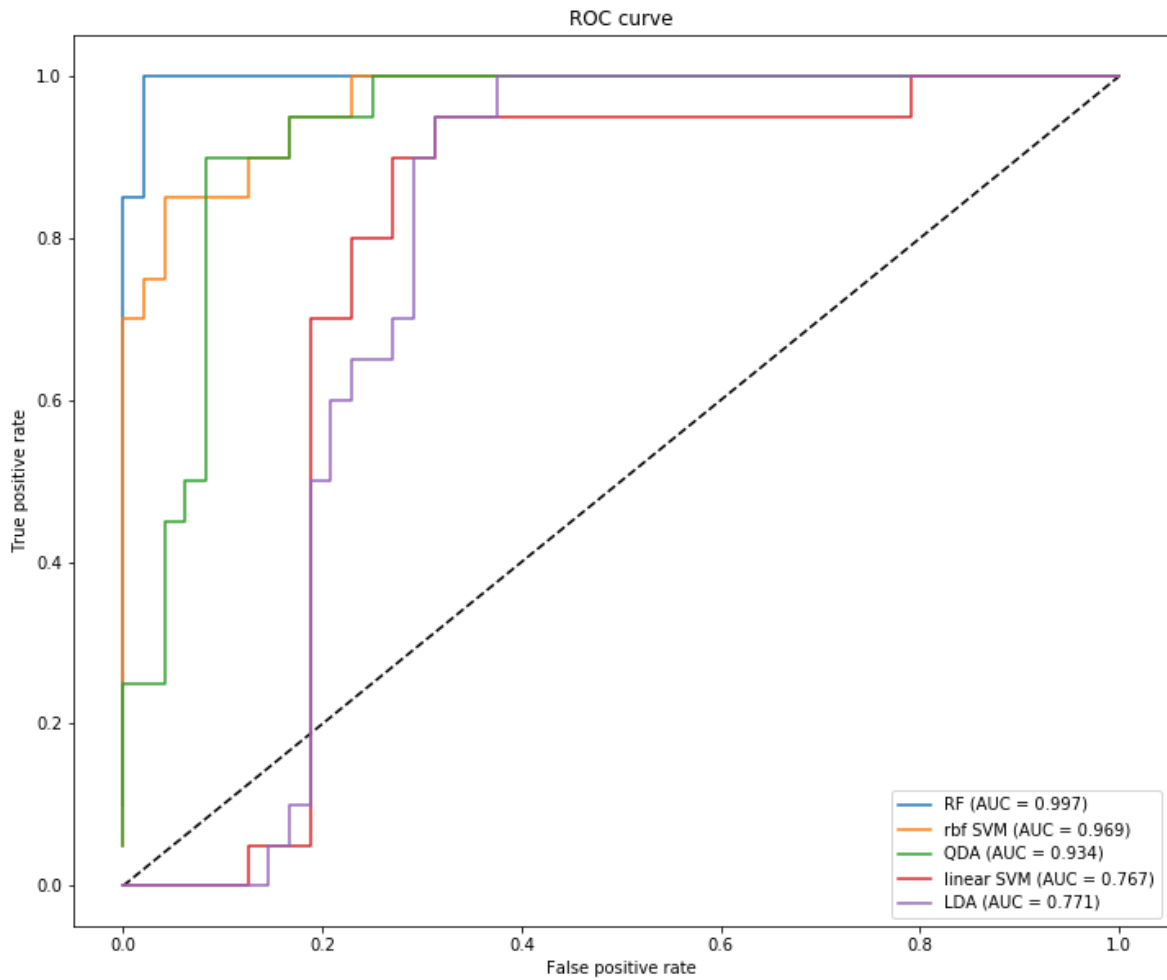


Figure 17: ROC of Random Forest, Gaussian SVM, linear SVM, LDA, and QDA on the Common Test Set. Random Forest with an area under the curve of 0.997 results in the best method to identify *S. pyogenes* in terms of specificity and sensitivity.



among other techniques. The ROC of the QDA and rbf SVM are similar although rbf kernel is slightly better than QDA. The ROC of the LDA and linear SVM are very close to each other. Nevertheless, they are below the ROC of QDA and rbf-SVM. It indicates that a nonlinear approach is better than a linear one. It can be concluded that Random Forest which uses bagging and random feature selection to grow trees is a robust approach regarding sensitivity and specificity.

## CHAPTER 3 REAL-TIME DEEP LEARNING APPROACH FOR PATHOGEN IDENTIFICATION USING RAMAN SPECTROSCOPY: IDENTIFICATION OF S. PYOGENES

### 3.1 Introduction

Raman spectroscopy (RS) has been widely used as a non-destructive tool to characterize chemical or biomedical samples. It can provide a structural fingerprint by identifying molecules using their specific vibration modes. RS has been applied in bacteria characterization as a rapid technique to discriminate pathogens by studying the modifications of biomolecules inside the cell. However, some issues remain in regards to utilizing RS to identify pathogens in real-time.

RS is a weak signal with a noise superimposed on a broad background. Removing this background noise is tedious work and usually needs direct supervision of an expert. Although during the last decades various studies have attempted to remove this background using multiple techniques including polynomial fit and morphology [117, 118, 119, 120, 121, 122, 108], these methods have not been efficient and usually need fine-tuning of some parameters. This issue prohibits them to be a candidate for a real-time detection or identification system.

Bacteria have the same structure and share similar macromolecules in their cell and membrane. Also, each bacteria is composed of various molecules and macromolecules which have significant bands in common. There are four central macromolecules in a bacteria cell.

Proteins are polymers of amino acids, and they are one of the primary macromolecule structures in the bacteria spectra. Amino acids differ by their side chain, and the Raman spectra bands of proteins are divided into three dominant groups. The amide I band includes 80 % C=O stretch, and its band is close to  $1650\text{ cm}^{-1}$ . The amide II band is nearly  $1550$

$\text{cm}^{-1}$  and consists of 60% N-H bend,  $\delta(\text{N-H})$ , and 40% C-N stretch,  $\nu(\text{N-C})$ . The amide III band is around  $1300 \text{ cm}^{-1}$  and includes 30% N-H bend,  $\delta(\text{N-H})$ , 40% C-N stretch,  $\nu(\text{N-C})$ , and skeleton stretches [1]. Other critical spectral features in protein spectra are Disulphide Bridges (S-S bonds) and Aromatic amino acids (Phenylalanine, tryptophan, tyrosine, histidine).

The primary structure of the cell membrane consists of lipids. Various lipids can be found in bacteria cell membrane and have been used for bacteria identification [123]. The leading bands of saturated fatty acids are at  $1295 \text{ cm}^{-1}$  and  $1440\text{-}1460 \text{ cm}^{-1}$  attributed to  $\text{CH}_2$  deformations, and  $1030\text{-}1130 \text{ cm}^{-1}$  assigned for C-C stretching vibration. The influential band of unsaturated fatty acids is lipids  $1658 \text{ cm}^{-1}$  attributed to C=C stretching [124].

Polysaccharides are the main component of the cell capsule, and lipopolysaccharide (LPS) is used to identify gram-negative bacteria [125]. Polysaccharides are made of different sugar monomers [126]. Also, the genetic information of the cell is represented by the sequence of the nucleotides in the nucleic acids. Nucleotides consist of a base linked to sugar by a glycosidic linkage and phosphate. Some of the bands associated with these macromolecules are summarized in Table 1.

Some studies have attempted to investigate the signature of some bacteria and identify some critical bands, and some other reviews have applied machine learning methods to discriminate different bacteria [127, 128, 129, 130]. Recently, methods based on deep learning have become very popular in supervised learning as it has been shown that deep learning is a robust and fast technique, and its accuracy has not reduced when the dataset is growing [131, 132, 133, 75].

Artificial Neural Networks (ANN) have been applied to bacteria identification studies

using vibrational spectroscopic methods. In [134], ANN is utilized to identify *Enterococcus faecium* with IR where neurons were the principal components of the spectra. Nevertheless, the application of neural network for spectra analysis remained dormant as the other methods overwhelmed the ANN. Also, as mentioned in Section 1.3 of Chapter 1, the addition of layers to the neural network can lead to exploding or diminishing gradient. In recent years, there has been a large number of various algorithms to improve the performance of the deep neural network and make it one of the most attractive methods for classification, regression, and machine learning.

In this paper, a new technique based on deep learning is presented to discriminate *Streptococcus pyogenes* using RS. This method is an end-to-end identification technique which does not require any pre-processing on the data, and as a result, identifies species rapidly. A deep neural network is trained such that it can estimate the background and subtract it from raw data, and another deep neural network is trained by considering the particular wavenumber of the spectra based on the biochemistry of macromolecules inside bacteria. The accuracy, sensitivity, and specificity of this method are 100%. We analyzed the performance of the aforementioned method on the dataset used in Chapter 2 to identify the *S. pyogenes*. It is worth noting that this dataset provides a representative, but not exhaustive, pathogen spectra. The objective here is to provide an end-to-end algorithm and unique deep neural architecture to discriminate pathogens, in particular *S. pyogenes*.

## 3.2 Material and Method

### 3.2.1 Dataset

We used the same dataset that was explained in section 2.2.2.

### 3.2.2 Input

After each spectrum was acquired, it was normalized to its maximum intensity. Following the normalization, 339 of total spectra were grouped into two separate datasets and divided randomly into three groups of training, validation, and testing dataset. Additionally, each spectrum was pre-processed and added to each group where it belonged. 60% of the data was used for training, 20% for validation, and 20% for testing.

We used  $x = (x_1, x_2, \dots, x_n)$  as an input where each band was considered as a neuron in the input. We limited each spectrum to a certain range from 400 to 2472  $\text{cm}^{-1}$  and a fixed length,  $|X| = 1368$ . As our RS system could acquire the spectrum from 400 to 3200  $\text{cm}^{-1}$ , each spectrum is truncated to the desired length. Given  $N_t$  as the sample size of the training set, the input is represented as a 2-D array with size of  $N_t \times |X|$ . The Theano toolbox [135] is used to implement the deep neural network.

### 3.2.3 Overview of the Model

The pipeline of simultaneous identification of *Streptococcus pyogenes* is comprised of three units as illustrated in Figure 18 : a real-time pre-processing unit based on convolution and deconvolution networks, a rearrangement unit based on prior knowledge of spectra biological macromolecules , and a real-time identification unit based on CNN [136, 64] and partially connected neural network (PCNN) [137, 138]. Firstly, the pre-processing unit is employed for background removal of the spectra using the raw spectra from RS. Then, the known spectral peaks of macromolecules inside the bacteria are used to generate a tensor from the pre-processed spectra. Finally, an identification unit is employed for feature generation and classification.

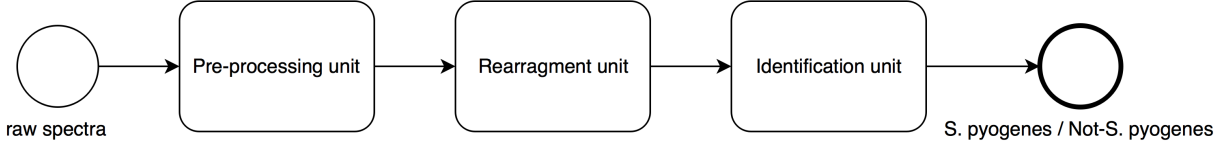


Figure 18: Overview of the Model.

**Pre-processing Unit** We have developed a deep neural network for real-time background removal of spectra. It is based on a Conv-DeConv network which is trained using raw spectral data as input, and the corresponding background removed spectra with the MPLS method as outputs [109, 110]. It is a real-time method which takes raw data as input,  $x$ , and generates the background removed spectra,  $z$ .

Figure 19 illustrates the architecture of the network. The whole network consists of 4 modules which are connected sequentially where raw spectra are input for the first module.

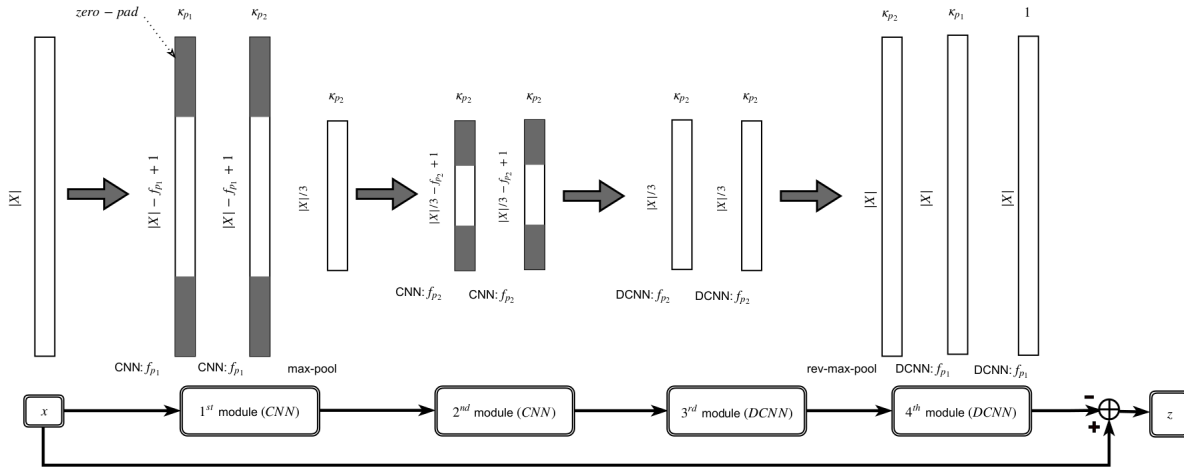


Figure 19: Architecture of the Pre-processing Network.

The first module contains convolution layers with a filter size of  $f_{p1} = 71$  and 16 feature maps are constructed for the first layer,  $\kappa_{p1} = 16$ , and the output is padded with zero to have

the same size of input. The second layer is a convolution layer with the same filter size and  $\kappa_{p_2} = 32$  feature maps, and the output is padded with zero again. The second layer is followed by a sub-sampling layer, max-pool, with a pool size of 3 and stride of 1.

The second module is similar to the first module with a difference in filter size that is  $f_{p_2} = 101$ , and both kernel sizes are equal to  $\kappa_{p_2} = 32$ . Also, there is no sub-sampling in this module.

The third module consists of deconvolution layers where the weights of the deconvolution layers are based on the weights of the convolution layers in the second module. The number of output kernels of the deconvolution layer is equal to the number of input kernels of the corresponding convolution layer. In same fashion, the input kernels of the deconvolution layer correspond to the output kernel of the convolution layer. Also, the weight of deconvolution filter is reversed as the operation order is reversed.

The fourth module is based on the weights of the first module. The first layer is an up-sampling layer based on the sub-sampling layer, which it assigns the same value for the indices on which the subsample filter is employed. Following the reverse of the sub-sampling layer, there is another deconvolution layer based on the weights of the second layer of the first module. The last layer is a deconvolution layer with the same filter size and number of the kernel of the first layer in the first module. The output kernel is 1 which yields the background of spectra. However, the parameters of this layer vary, and it is not constrained by the first layer in the first module. In other words, this layer can be learned independently. The sigmoid activation function is employed on the output of the last deconvolution layer. The goal is to minimize the mean of the least square root between the output of the network and background of spectra obtained using the MPLS method with direct supervision and

validation of an expert on the raw spectra.

This network is capable of generating features from raw data that can ultimately estimate the background of the Raman spectra. Finally, the output of the last layer is subtracted from the input signal  $x$  to construct the background removed spectra,  $z$ .

To train the pre-processing network, raw spectra are used as inputs, and the corresponding background removed spectra are used as outputs. The objective function was to minimize the mean square error (MMSE) on outputs and the stochastic gradient descent method was used to update parameters of the network. The learning rate was 0.001, and the network is trained for 10000 epoch where the earliest and best validation error is used to determine the optimal hyper-parameters of the net.

**Rearrangement Unit** We used a rearrangement scheme which maps the pre-processed spectra  $z$  to known peaks of biological macromolecules  $\Psi = (\Psi_1, \Psi_2, \dots, \Psi_\lambda)$  where  $|\Psi| = 60$  in our scheme. We used the reference database of biological molecules that is described in [89]. A macromolecule consists of known bands which are unique for that macromolecule,  $\Psi_j = (\xi_1, \xi_2, \dots, \xi_k)$  where the number of bands,  $|\Psi_j|$ , can be different for each macromolecule. The bands associated with the macromolecule vary in strength and might have a small shift depending on the spectroscopy system from which they are obtained. Nevertheless, all the bands involved in specifying the macromolecule are considered in our scheme. Additionally, a neighborhood window with width  $w$  of each band is chosen where the desired band is in the middle of the window. In other words,  $\xi_i = (\dots, \xi_{i,-2}, \xi_{i,-1}, \xi_{i,0}, \xi_{i,1}, \xi_{i,2}, \dots)$  where  $|\xi_i| = w$ . Finally, all of these bands are combined to generate a tensor with the shape of  $N_t \times N_\Psi \times |\xi|$ , where  $N_\Psi = \sum_{j=1}^{|\Psi|} |\Psi_j|$ . This tensor  $\gamma$  is a representation of the pre-processed spectra  $z$  on the



known bands of macromolecules.

**Identification Unit** Our neural network employed two layers of CNN followed by a PCNN.

Figure 20 shows the architecture of the network for the identification unit.

1. Convolutional Neural Networks: This network aims at generating feature from the local information of each band. It takes the tensor  $\gamma$ , rearranged tensor from pre-processed data  $z$ , as input and convolutes the  $|\xi|$ - axis of  $\gamma$  with a set of kernels. The number of filters for the first layer is  $\kappa_1 = 10$  and  $\kappa_2 = 5$  for the second layer. The kernel size for the first layer is 2 and  $w - 1$  for the second layer. As a result, the output shape of the CNN becomes  $N_t \times N_\Psi \times \kappa_2 \times 1$ .
2. Partially Connected Neural Network: PCNN, following CNNs, connected the generated features of bands associated with the corresponding macromolecule. A nonontogenic method was used for reduction of the connections as the input-output relationship is known for this case [139, 140, 141, 142]. The hyperbolic tangent is applied to the output as an activation function, and the shape of the output tensor is  $N_t \times N_\Psi$ . Finally, the output of this layer is connected to a logistic regression classifier with two nodes in the output.

The output of the pre-processing unit was fed to the rearrangement unit. After the spectra were rearranged, the identification network was trained by minimizing the negative log-likelihood (NLL) of the output of the logistic regression classifier on unseen samples as the zero-one loss is not differentiable. The stochastic gradient descent method was used to update the parameters of the identification network where the parameters of the pre-

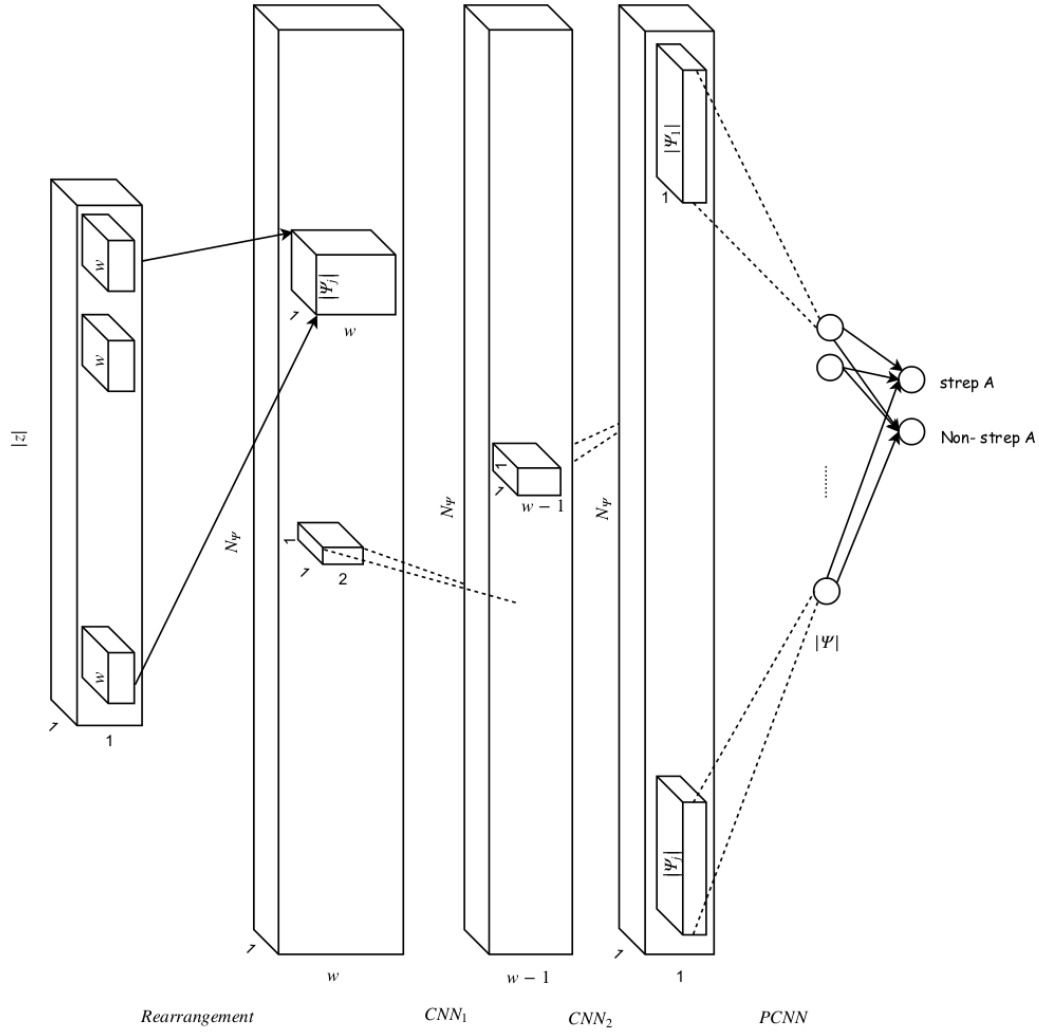


Figure 20: Architecture of the Identification Network.

processing unit were frozen during training identification unit. The epoch of training was set to 10000 such that the network explored the possible optimal points. Nevertheless, the validation score and cost were calculated at the end of each epoch, and the best and earliest one was considered as the optimal one.

### 3.3 Result and Discussion

In this chapter, a deep learning approach is used to generate the features automatically.

The result of this approach is presented when it is applied to the dataset explained in

section 2.2.2. This dataset includes different bacteria in the water background and contains 339 raw spectra and their corresponding pre-processed spectra which are divided into three sets of training (80%), validation (20%), and testing dataset (20%). The model was trained by using training dataset and validated during training by validation dataset. Eventually, the model was tested by feeding the testing database to the network, and further analyses such as specificity and sensitivity were performed to evaluate the model. The result of the pre-processing unit is reported and compared with the ground-truth data. Finally, the performance of the model is compared with the state-of-art machine learning approach presented in Chapter 2.

### 3.3.1 Pre-processing Unit

In this section, the output of the pre-processing unit is presented. Fig. 21 and 22 show the output of pre-processing unit which is applied on raw spectra. It can be seen that this unit is capable of predicting the ground-truth spectra very closely, and in each plot it can be seen that the predicted spectrum matches the ground-truth spectrum in almost all bands.

Figure 23 illustrates the mean and standard deviation of predicted and ground-truth spectra of *S. pyogenes* data. It can be seen that the standard deviation of the predicted spectra is similar to the standard deviation of the ground-truth spectra. The means of both plots are in better agreement with each other when their standard deviations are small. Nevertheless, the fluctuation pattern of the peaks in both plots, predicted and ground-truth, are similar.

### 3.3.2 Classification Result

The misclassification error, specificity, and sensitivity are evaluated on all three datasets in the end, and the result is illustrated in Table 5.

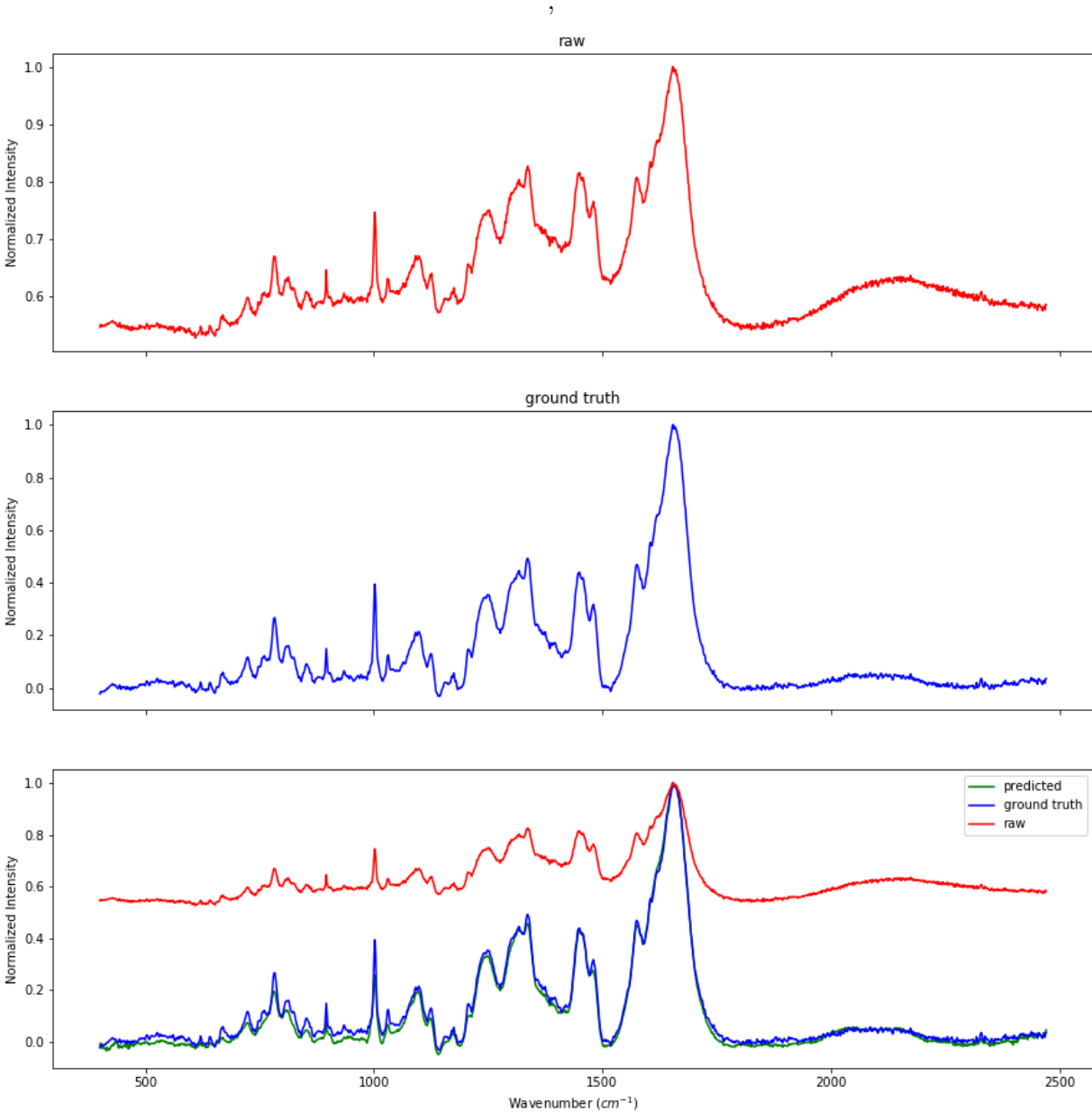


Figure 21: Output of Pre-processing Unit Applied to Raw Spectrum of pathogen. It can be seen that pre-processing unit can estimate the ground-truth spectrum very firmly in almost all bands.

After training the network on both *S. pyogenes* and other spectra, the network was tested on a test database, and accuracy of 100% for the test dataset was achieved. Sensitivity was 100%, and specificity was 100% as well.

The error for the training and validation sample were computed by feeding the input of

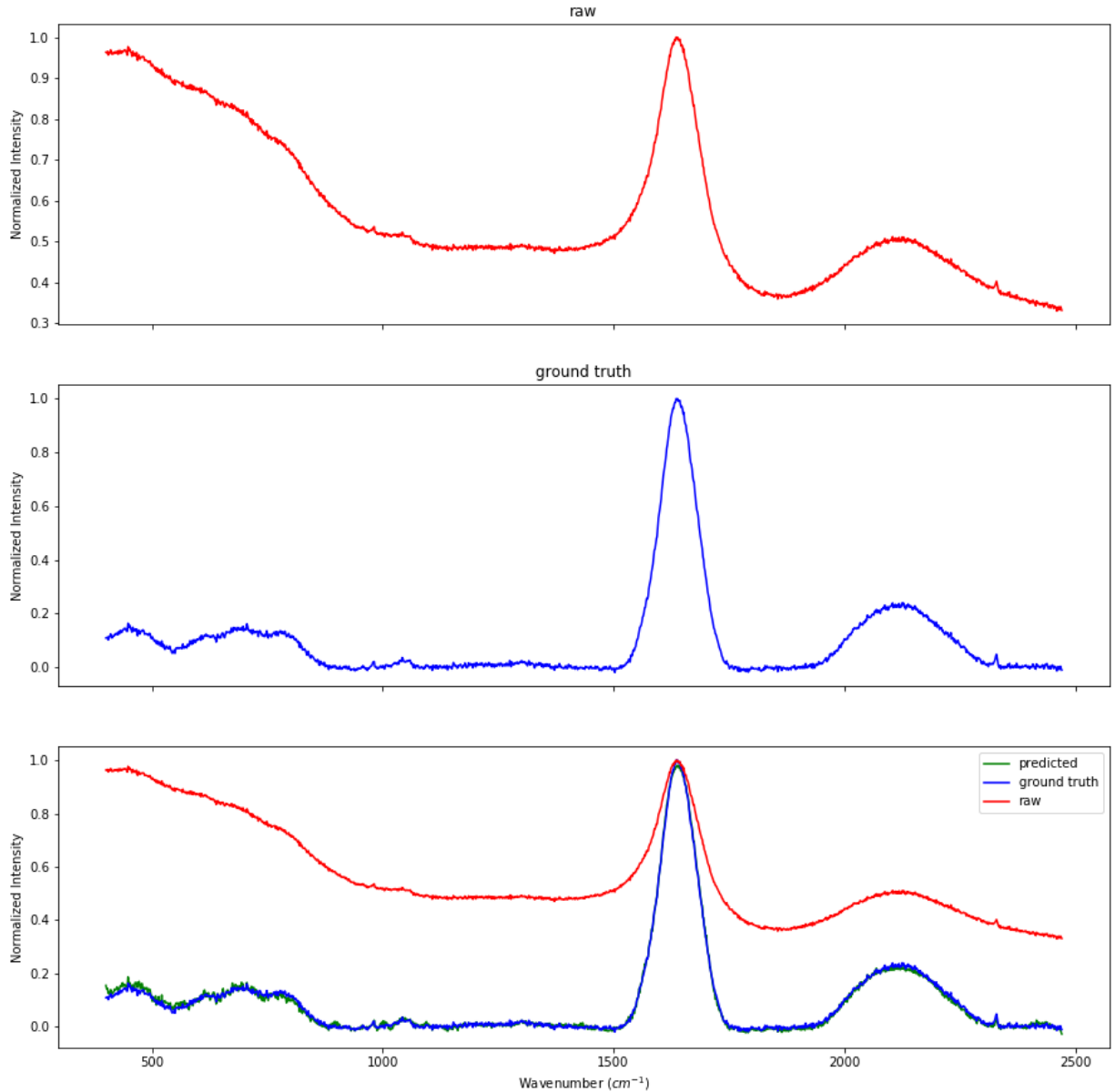


Figure 22: Output of Pre-processing Unit Applied to Raw Spectrum of Water.

the trained network with each spectrum in the database where it was 0%. In other words, there was no misclassification during training.

Additionally, for the Receiver Operating Characteristic, ROC, the curve is plotted for three datasets in Figure 24. The ROC provides a comparison of two operating characteristics, true positive rate (TPR) against the false positive rate (FPR). The results suggest our

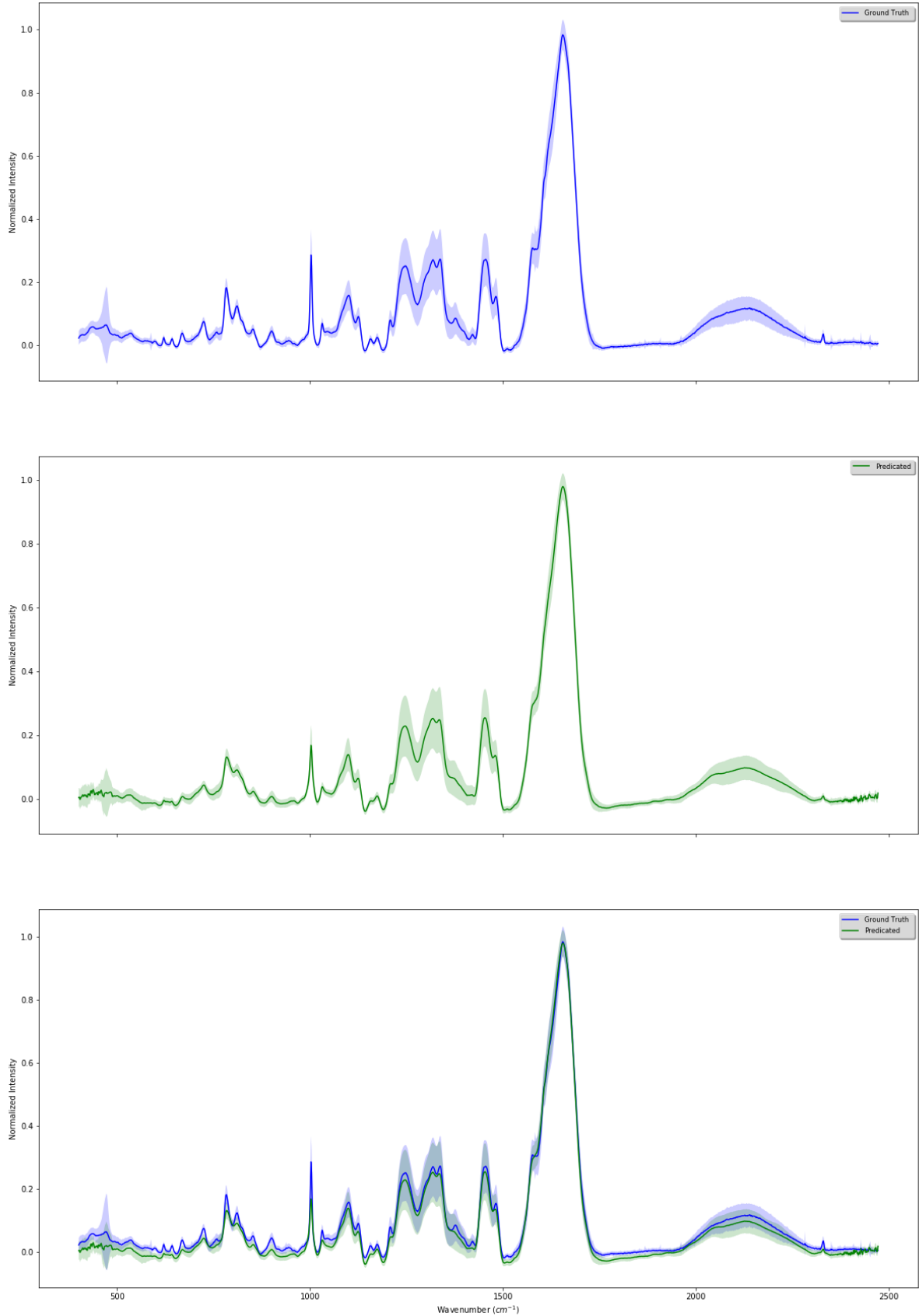


Figure 23: Output of Preprocessed Unit for *S. pyogenes* Data. The mean and standard deviation of predicted and ground-truth pre-processed spectra are plotted in blue and green, respectively.

Dataset	Error	Sensitivity or true positive rate	Specificity or true negative rate
Training	0%	100%	100%
Validation	0%	100%	100%
Testing	0%	100%	100%

Table 5: Classification Result on Training, Validation, and Testing Dataset.

model provides the optimal solution to distinguish *S. pyogenes* from other species in a water background.

As the error was 0% in each training, validation, and testing dataset, it suggested that the model is not over-fitting or under-fitting the result. Although our dataset was restricted to the limited number of pathogens, this model can be exploited to identify the pathogens using RS which can be achieved by fine-tuning the hyper-parameters or adding and training additional layers or kernels for the feature extraction unit. Moreover, the proposed model can be trained using a custom set of wavenumbers which suggests a great opportunity to embed the biological knowledge to such model which is data driven.

### 3.3.3 Comparative Result

In order to address the performance of the proposed model, the results produced by our approach are compared to other machine learning methods such as SVM and Random Forest. This study has been done on the same dataset, and it is discussed in detail in Chapter 2. Table 6 shows the summary of the results. It is noticeable that our approach, Deep Pathogen Identification Neural Network (DPINN) achieves the best result according to classification error, sensitivity, specificity, and F1 score. Furthermore, the PCA-LDA, PCA-QDA, and PCA-SVM (linear) have a very low sensitivity which makes them an unsuitable choice to be used for spectra analyses and classification. The specificity of the SVM with rbf kernel and

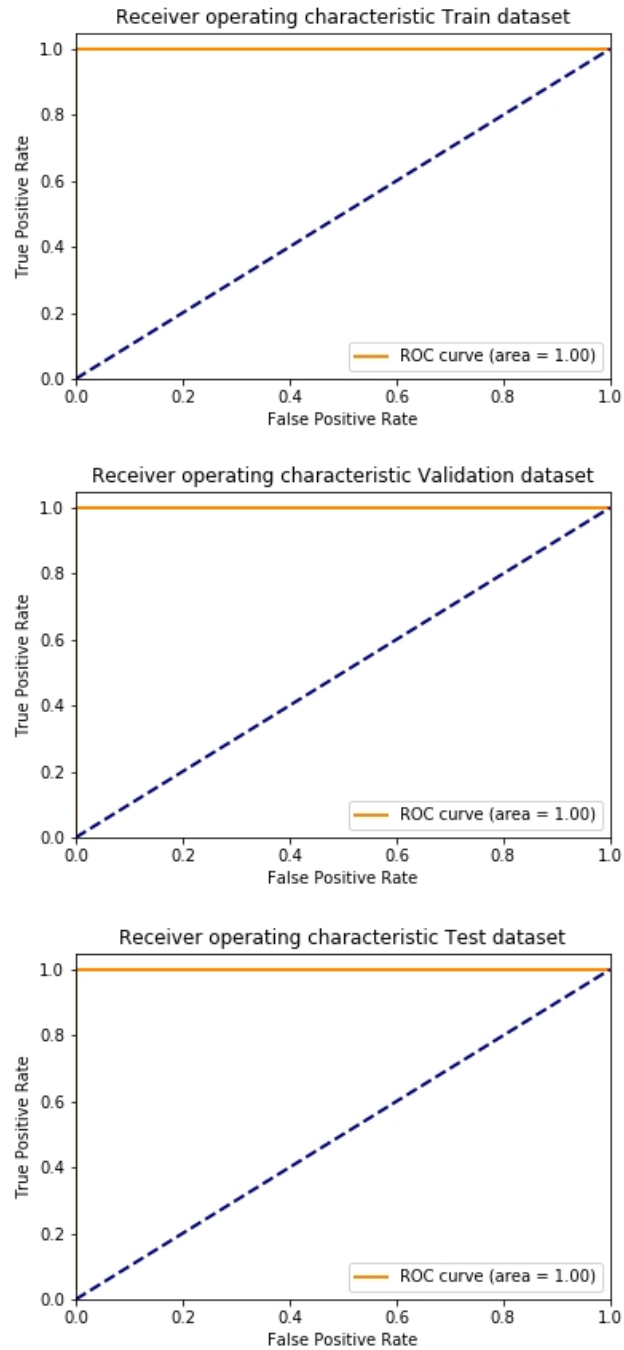


Figure 24: ROC of Training, Validation, and Testing Dataset.

the Random Forest is similar; however, Random Forest results in better sensitivity and F1 score.



Method	Error	Specificity	Sensitivity	F1
PCA-LDA	33.82%	81.25%	30.00%	0.343
PCA-QDA	23.53%	95.83%	30.00%	0.429
PCA-SVM(linear)	26.48%	85.42%	50.00%	0.710
PCA-SVM(rbf)	8.83%	97.92%	70.00%	0.800
Random Forest	5.89%	97.92%	85.00%	0.895
DPINN	0.00%	100.00%	100.00%	1.000

Table 6: Result on the Benchmark Dataset with Water Background. The proposed approach (DPINN) achieves the lowest error and the highest sensitivity, specificity, and F1 score among others.

## CHAPTER 4 IDENTIFICATION OF STREPTOCOCCUS PYOGENES IN CONFOUNDING BACKGROUND USING RAMAN SPECTROSCOPY

### 4.1 Introduction

In Chapter 3, it has been shown that a deep neural network is a robust method to identify *S. pyogenes* in water background. However, in real clinical use, a test on *S. pyogenes* involving a swab from the throat of a patient introduced a new challenge in identifying the bacteria in confounding background. The swab might comprise other chemical components depending on the patient or test conditions leading to the contribution of various backgrounds in the Raman spectra of the sample. As a result, the acquired spectra were composed of not only the molecular fingerprint of the bacteria but also the background bands and baseline of the sample.

Raman spectra is a representation of molecular vibrations of different molecules, and each band can be assigned to different molecular structures of cell or background macromolecules. The spectra of bacteria acquired in media such as water whose spectra is known are unique and represent molecular vibrations of different molecules inside the cell. When media is changed, or the spectra are acquired in confounding background, the molecular vibration of macromolecules inside the bacteria alters slightly. Moreover, the molecules in confounding background vibrate and thus contribute in the acquired spectra. As a confounding background might have similar fragments of bacteria, it can modify the obtained spectra significantly.

In Chapter 3, a deep learning method based on known spectra of some macromolecules is studied and has been shown that this method can be used to discriminate *S. pyogenes* from

other selected bacteria and also from water. In this paper, we aim at training a deep neural network to identify *S. pyogenes* in the confounding background using Raman Spectroscopy (RS). This method is shown to be robust to the modification of background and can be used in clinical application as a rapid identification technique to identify *S. pyogenes*.

## 4.2 Material and Method

### 4.2.1 Instrumentation and Sample Preparation

The instrumentation for data acquisition was the same Renishaw system described in section 2.2.1. Furthermore, the preparation of samples with water background and also pathogen culture were described in detail in section 2.2.1 .

The protocol for samples with throat swab background was as follows: The subjects were examined to make sure they had not used mouthwash or antibiotics and also that there was no sign of redness, swelling and especially white streaks or pus in their throat. Then with the use of a sterile cotton swab, the back of the throat (posterior pharynx) and both tonsils (tonsillar arches) were stroked several times. The swab was placed in a sterile 1.5 mL Eppendorf container. The swab was agitated in a tube with 0.4 mL of filtered sterilized tap water and then put in Nanofuge for 2 seconds. Finally, the swab was removed from the tube and prepared for testing and pathogen spike. To spike, the sample with a pathogen, the tube of the pathogen in filtered tap water, with final optical density (measured at a wavelength of 600 nm) of the  $1.00 \pm 0.05$ , was pipetted into the prepared throat swab sample and mixed appropriately.

### 4.2.2 Dataset

The spectra were acquired in two different backgrounds: water background and confound background (throat swab). The water background dataset was the same dataset used for Chapter 2 and explained in detail in Section 2.2.2. The confounding dataset consists of the throat swab background, *S. pyogenes* with throat swab background, and *P. aeruginosa* with throat swab background spectra. The total dataset is illustrated in Table 7.

Dataset	background	Biological Species	count #
S. pyogenes	water	<i>S. pyogenes</i>	101
	Throat swab	<i>S. pyogenes</i>	110
Not-S. pyogenes	water	MRSA	78
	water	MSSA	53
	water	<i>E. coli</i> (K99)	50
	water	<i>Legionella Pneumophila</i>	20
	water	<i>Pseudomonas aeruginosa</i>	8
	water	filtered water	29
	Throat swab	<i>Pseudomonas aeruginosa</i>	107
	Throat swab	Throat swab	117
Total:			673

Table 7: Dataset Summary with Confounding Background

### 4.2.3 Input

In similar fashion of Chapter 3, the spectra are normalized to their maximum intensity and cropped to a range of 400-2472  $\text{cm}^{-1}$  with a fixed length,  $|X| = 1368$ . The dataset is grouped into three datasets of training, validation, and testing dataset. Also, the pre-processed spectra are accumulated in each dataset.

We used  $x$  as an input where each band was considered as a neuron in the input. Thus, the input is a two-dimensional array with size of  $(N_t \times |X|)$  where  $N_t$  is the sample size. The implementation of the deep neural network was done using the Theano framework [135].

#### 4.2.4 Model

The network architecture is similar to the one explained in detail in Section 3.2.3. It consists of three units of pre-processing, rearrangement, and identification units.

The pre-processing unit takes raw data as input,  $x$ , and produces the background removed spectra,  $z$ . The whole network consists of 6 modules compared to 4 modules used in Chapter 3. The first and last two modules are similar, and another module based on the convolution network is added at the end of module 2, and subsequently, another module based on deconvolution network is inserted where its output is connected to the input of module 3.

The networks are trained similarly as explained in Chapter 3. The stochastic gradient descent is used to update the parameters and the earliest and best validation error is used to determine the optimal hyper-parameters of the network.

### 4.3 Result and Discussion

In this section, the classification result is presented on the confounding background dataset illustrated in section 4.2.2. This dataset consists of the raw spectra of the different pathogens with a different background. Also, their pre-processed spectra are augmented on the dataset. This dataset is randomly split into three sets: training (80%), validation (20%), and testing (20%).

#### 4.3.1 Data visualization

In Chapter 3, the raw and pre-processed spectra acquired in water background has been illustrated. The mean and standard deviation of raw and pre-processed spectra of *S. pyogenes* and *P. aeruginosa* acquired in throat swab background as well as only throat swab background

Dataset	Error	Sensitivity or true positive rate	Specificity or true negative rate
Train	0.37%	100.0%	99.46%
Validation	4.07%	98.84%	94.57%
Test	3.70%	94.44%	97.22%

Table 8: Classification Results on Training, Validation, and Testing Dataset.

are shown in Figures 25 and 26, respectively. The spectra are normalized to their maximum intensity. It can be noted that there is a strong baseline in this background compared to the water background and that it is essential to remove this baseline before data analysis. Also, the *S. pyogenes* and *P. aeruginosa* have similar peaks and pattern and also that the spectra of the throat swab have identical peaks in common with both pathogens.

A deep learning approach is presented in Chapter 3 as a robust method to distinguish the pathogens in water background. In this chapter we have applied this method on the confounding background dataset by extending some layers and tweaking some hyper-parameters of the networks as explained in section 4.2.4.

### 4.3.2 Training Result

Table 8 illustrates the results of the misclassification error of three datasets of training, validation, and testing.

The results show that the training error is around 0.37%. Thus, the model is not underfitting to the training dataset. The validation error is slightly above the training error. Nevertheless, this difference is insignificant such that there is no overfitting of the model to the training dataset. The testing error is 3.70%, and this error is similar to the validation set indicating the minimal difference between data distribution of the testing and validation dataset. Figure 27 illustrates one of the *S. pyogenes* spectra which is misclassified as Not-S.

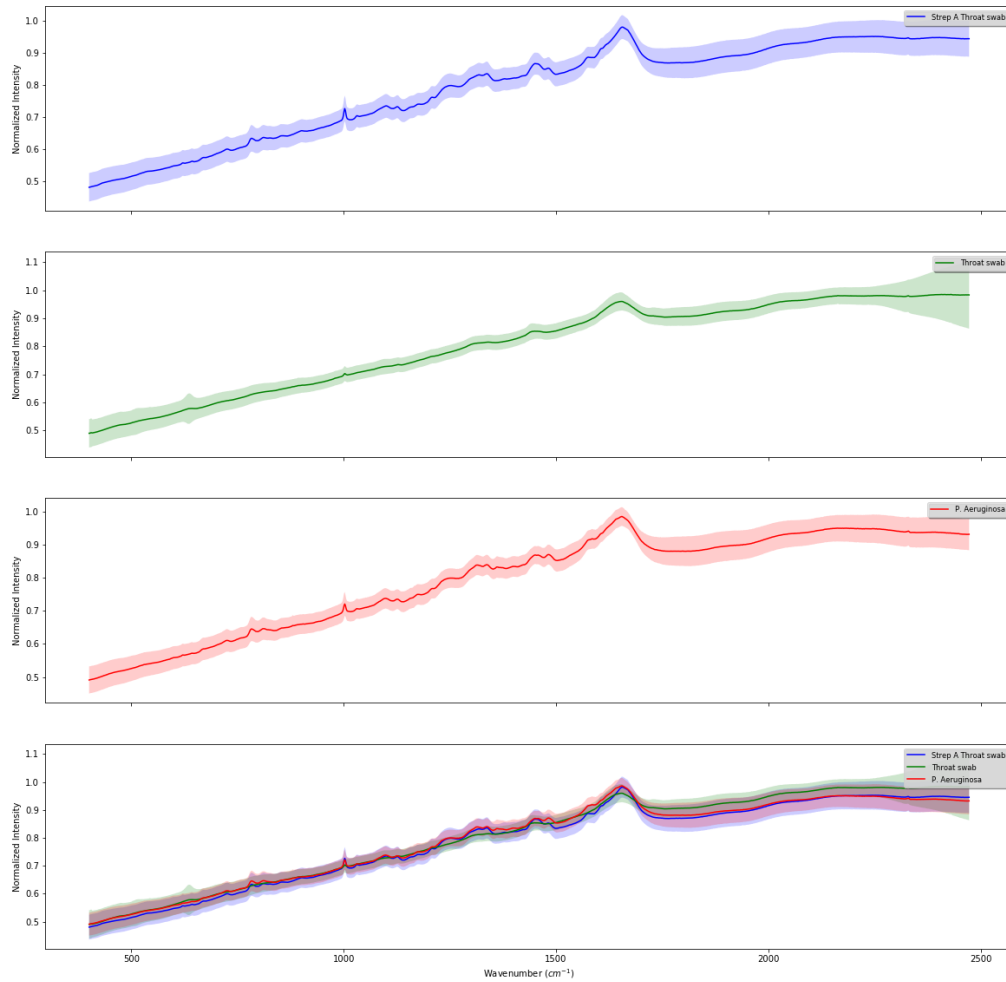


Figure 25: Mean and Standard Deviation of Raw Data Acquired Using Throat Swab Normalized to Maximum Intensity.

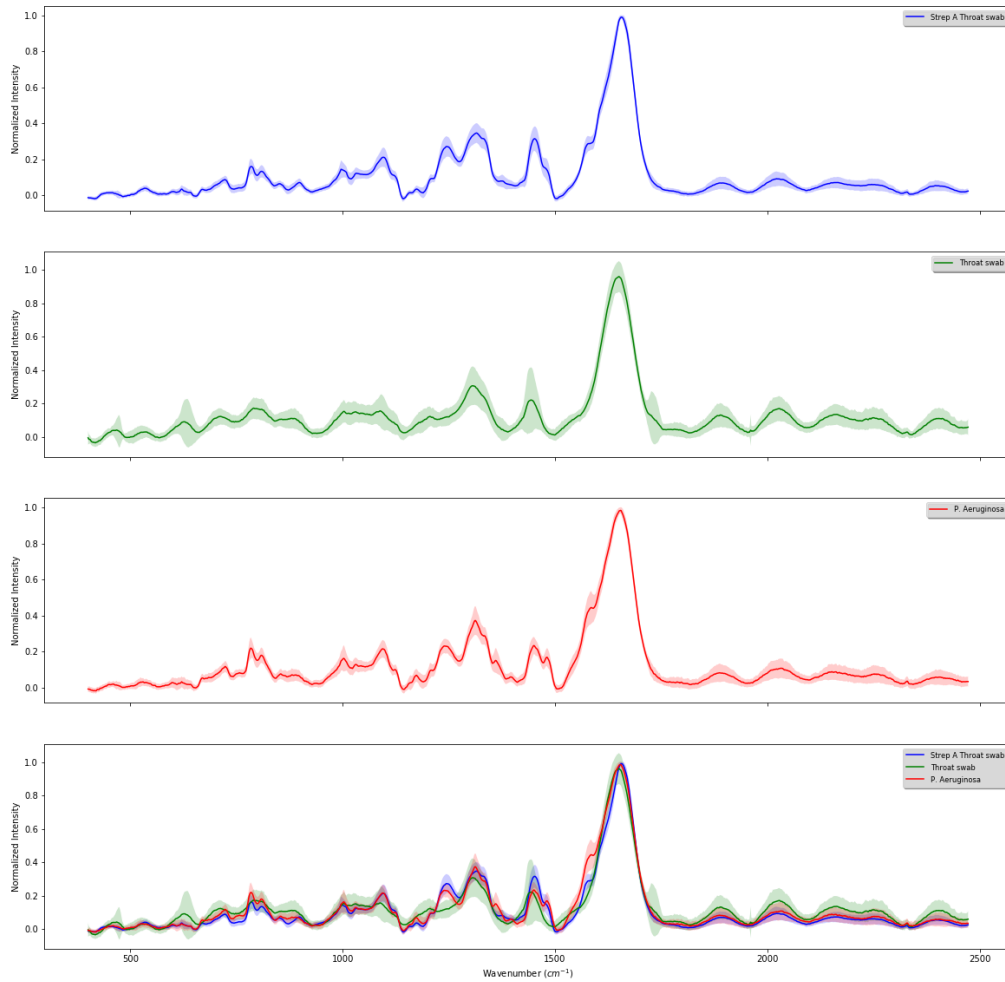


Figure 26: Mean and Standard Deviation of Background Removed Spectra Acquired Using Throat Swab Normalized to Maximum Intensity.



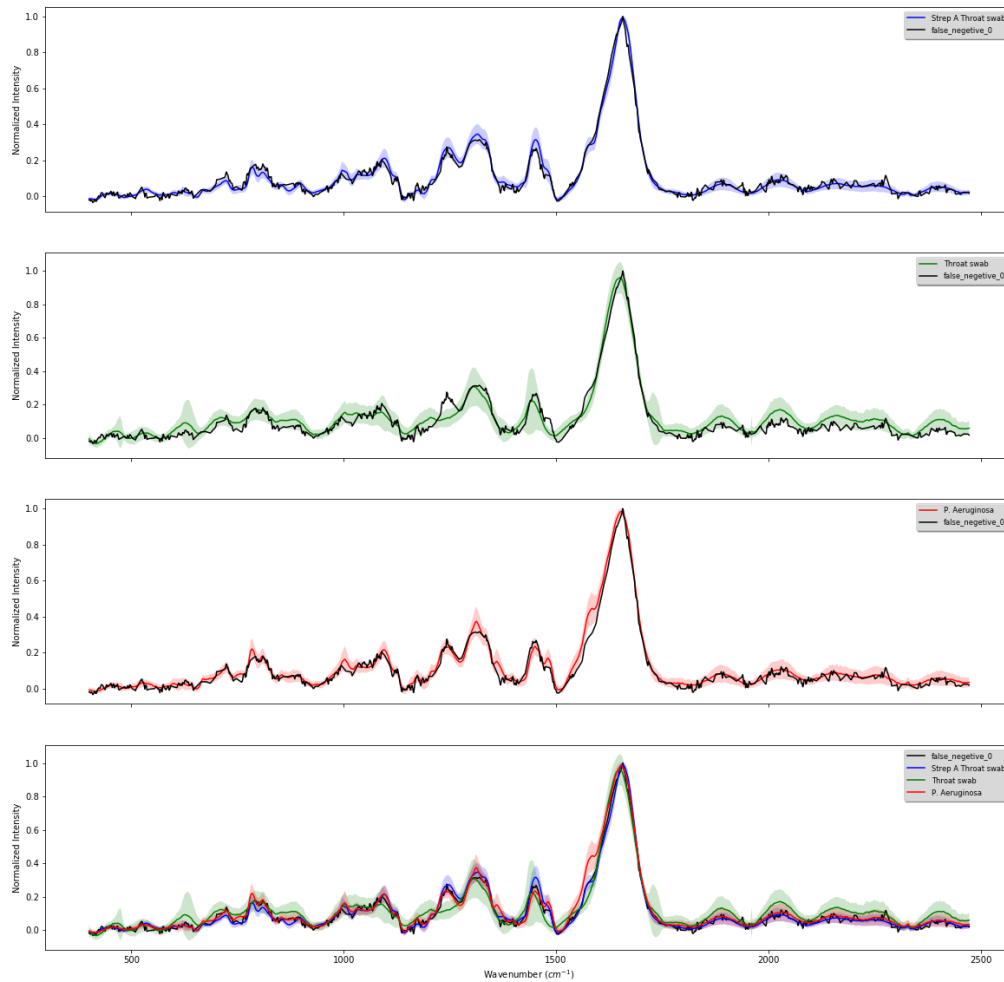


Figure 27: Misclassified Sample. *S. pyogenes* Spectra Classified as Not-*S. pyogenes*.

*S. pyogenes* (False negative). It can be seen that there are differences between this spectrum and *S. pyogenes* (mean and standard deviation) where there is a similarity in pattern and intensity of the spectra with *P. aeruginosa* in some of the major bands, such as bands around 1400–1500. In other words, the intensity of such spectrum is not in the range of the standard deviation of *S. pyogenes* spectra but rather that of *P. aeruginosa*.

The ROC curve is plotted for three datasets in Figure 28. The ROC curve suggests that the best sensitivity and specificity score can be achieved using the threshold 0.35. The area under the curve (AUC) of training, validation, and testing dataset are 1, 1, and 0.99, respectively, revealing that there is a balance between the complexity and generalization of the model.

### 4.3.3 Realization of the Network: Macromolecules

The network is designed such that the features from each individual macromolecule can be extracted, and then the classifier distinguishes the sample based on the partially connected networks of different macromolecules. The study of this partially connected network for each macromolecule can yield an understanding of the trained network. For this purpose, the weights for a macromolecule network were set as trained weights, whereas for others, they were set to zero. Then, the dataset feed to this network and the mean probability for both groups were calculated. In this fashion, this process is repeated for every macromolecule.

In Figure 29, the probabilities of true positive and true negative for all macromolecules are illustrated. It can be seen that b-carotene, d-arabinose, d-fucose, and d-mannose have the highest probability for accurate detection of positive samples and the l-histidine and amylopectin yield the highest true negative rates.

It is possible that a macromolecule can yield a correct identification for a positive dataset along with a false detection for the other dataset. Figure 30, shows the mean probability of the network when the *S. pyogenes* dataset was fed to it. The blue bars indicate the likelihood of the true detections and green bars shows the likelihood of the false detections (false negative) ones. It can be seen that adenine and d-xylose have the highest probability of false negative rates, above 0.65. These macromolecules might be responsible for the reduction

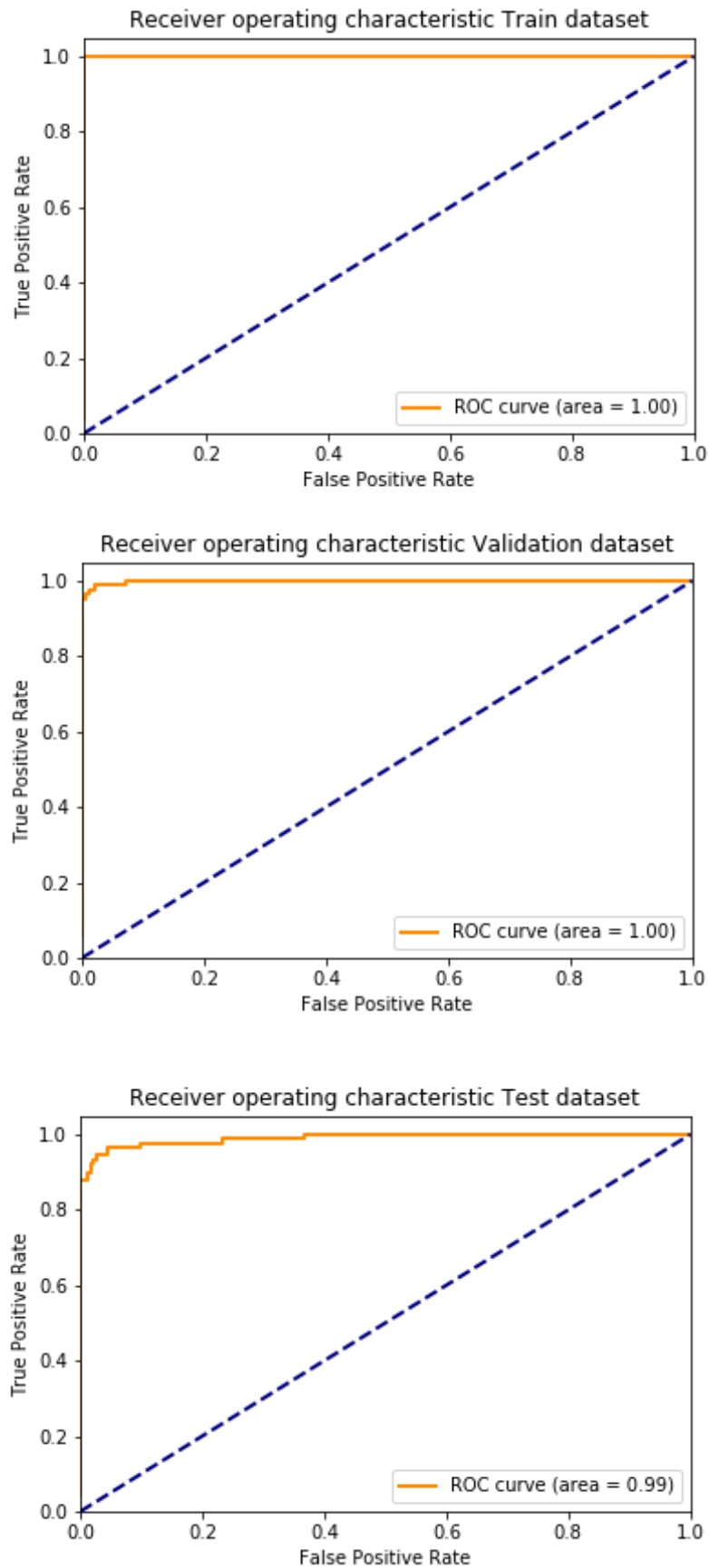


Figure 28: ROC of three datasets. The AUC of each ROC is illustrated in the plot.

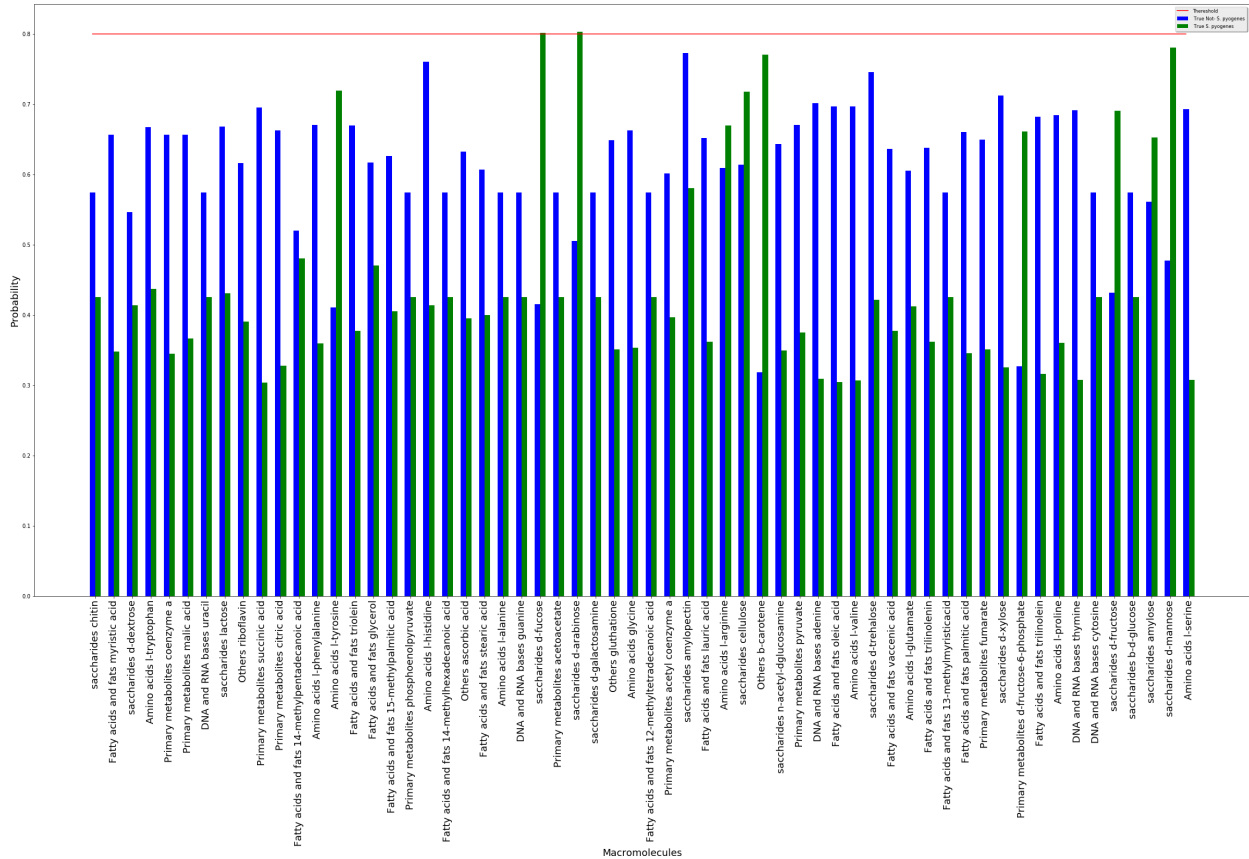


Figure 29: Result of True Negative and True Positive on All Macromolecules Participating in the Network. b-carotene, d-arabinose, d-fructose, and d-mannose have probability above 75% for accurate detection of *S. pyogenes* spectra and the l-histidine and amylopectin with a likelihood above 75% to be responsible for correct identification of Not-*S. pyogenes* spectra.

of the sensitivity of the model.

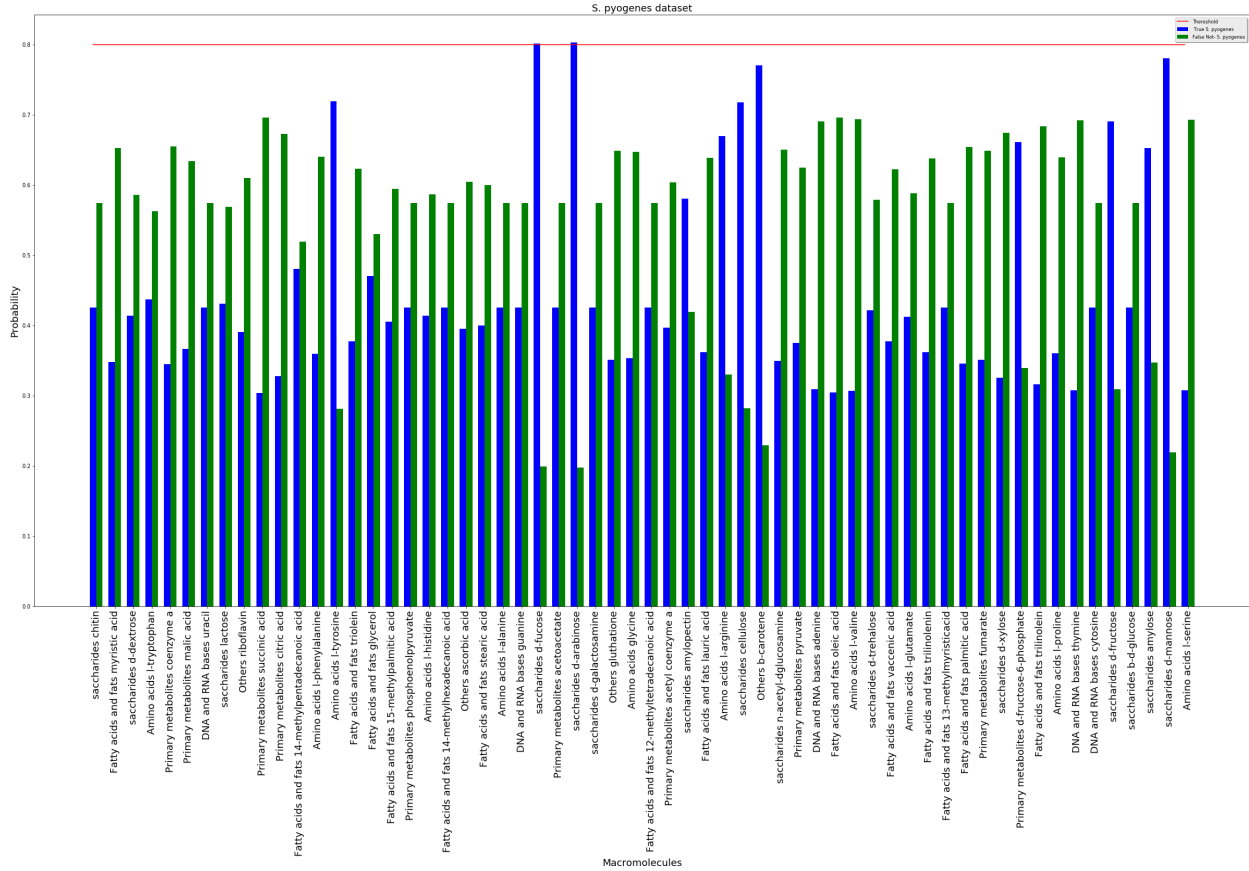


Figure 30: Result of True Positive and False Negative on All Macromolecules Participating in the Network. Adenine and d-xylose have the highest probability of false negative rates, above 0.65.

Similarly, Figure 31 shows the mean probability of the network when the Not-*S. pyogenes* dataset feeds to the network. It can be noted that b-carotene with a probability of 0.68 is the strongest macromolecule which contributed to the false positive rate. It can be concluded that this macromolecule reduces the specificity of the network.

Table 10 illustrates the mean probability of the macromolecules from which an accurate identification with a probability above 0.65 can be yielded. For each macromolecule, four probabilities are calculated where two different datasets of *S. pyogenes* and Not-*S. pyogenes* were the inputs of the network. As a result, the false positive rate and false negative rate in

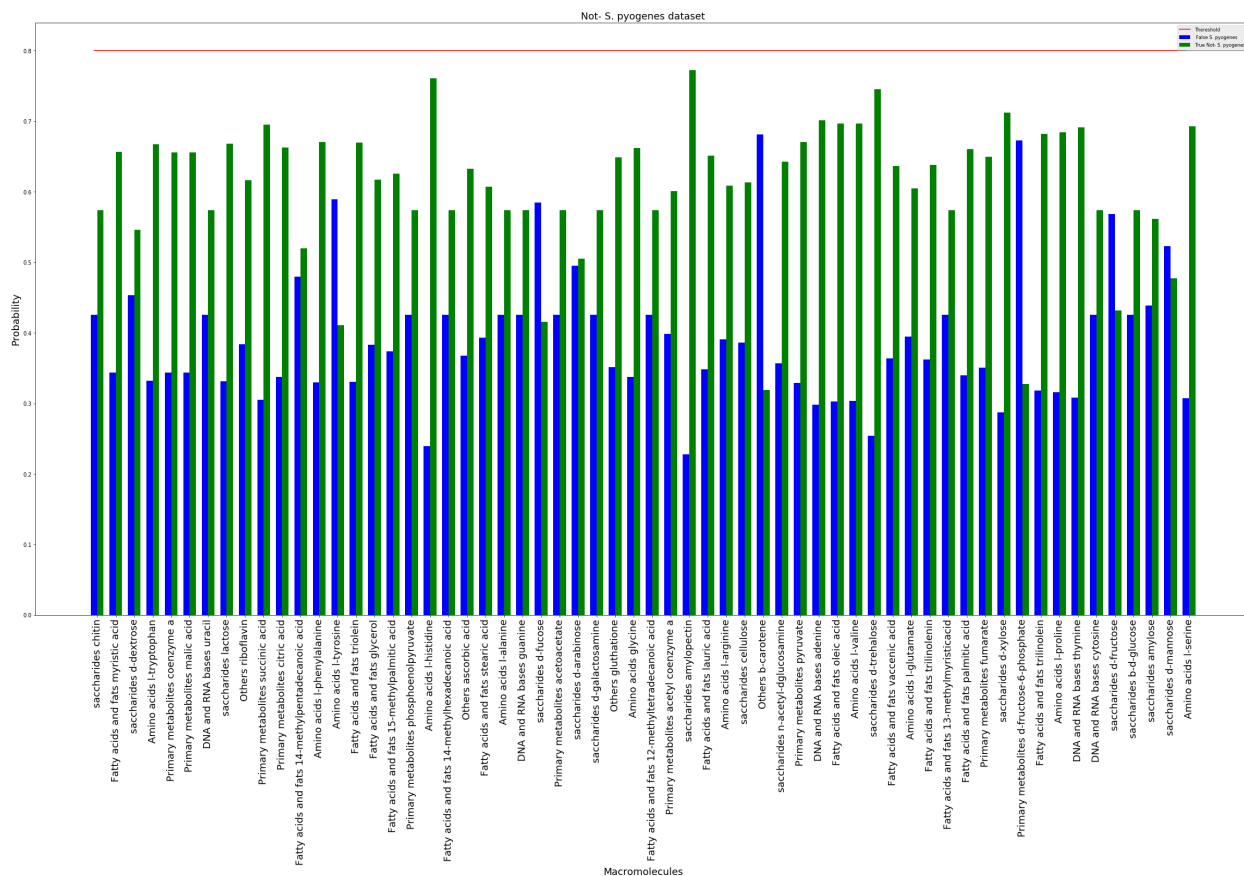


Figure 31: Result of True Negative and False Positive on All Macromolecules participating in the Network. b-carotene with a probability of 0.68 is the strongest macromolecule contributed to false positive rate.

In addition to true positive rate and true negative rate can be computed to understand which macromolecules can lead to the correct identification of negative or positive samples and which ones can be the reasons for the false negative or false positive detections.

It can be seen that b-carotene has a high probability of true positive and also false positive rates. Thus, a network of this macromolecule is not trained well to discriminate some of the samples which are not *S. pyogenes* from *S. pyogenes* samples. It can be concluded that the precision of the overall network can be affected by this macromolecule network. It also affects the specificity of the network as the b-carotene network tends to result positively, independent of the incoming spectra.

	Not-S. pyogenes- False pyogenes	S. True Not-S. pyogenes	S. pyogenes- True S. pyo- genes	S. pyogenes- False Not-S. pyogenes
Amino acids l-histidine	0.239312	0.760688	0.413524	0.586476
Amino acids l-tyrosine	0.589092	0.410908	0.719055	0.280945
DNA and RNA bases adenine	0.298337	0.701663	0.309394	0.690606
Others b-carotene	0.681207	0.318793	0.770675	0.229325
saccharides amylopectin	0.227511	0.772489	0.580468	0.419532
saccharides cellulose	0.386449	0.613551	0.717649	0.282351
saccharides d-arabinose	0.494961	0.505039	0.802654	0.197346
saccharides d-fucose	0.584357	0.415643	0.801040	0.198960
saccharides d-mannose	0.522542	0.477458	0.780666	0.219334
saccharides d-trehalose	0.254268	0.745732	0.421229	0.578771
saccharides d-xylose	0.287641	0.712359	0.325485	0.674515

Table 10: Mean Probability of Macromolecules from Which an accurate identification with a probability above 0.65 can be yielded.

Adenine and d-xylose are the macromolecules which yield not only the high false negative rates but also high true negative rates that affects the recall or sensitivity of the network as this network is less sensitive to the incoming spectra and tends to have negative results for both *S. pyogenes* and Not-*S. pyogenes*. So these macromolecules cannot identify *S. pyogenes* samples from Not-*S. pyogenes* samples accurately. In other words, they have very low sensitivity and can decrease the overall sensitivity of the network.

In summary, it can be concluded that the l-tyrosine, cellulose, d-arabinose, d-fucose, and d-mannose can yield the detection of the positive sample with the highest probability and the l-histidine, amylopectin, and d-trehalose are responsible for true negative rates. Although the realization of the network can help to understand the network and potentially help to

design a robust one, it cannot guarantee the result will wholeheartedly concur with the biological understanding of the pathogens. In other words, the network is trained based on the data it has received. As a result, a more extensive dataset that includes different species might yield a trained network close to the bacterial structure.



## CHAPTER 5 CONCLUSION AND FUTURE WORKS

### 5.1 Conclusion

In this study, we aimed at identifying *S. pyogenes* from other species using Raman Spectroscopy. The various multivariate algorithms including PCA-LDA, PCA-QDA, SVM, and Random Forest have been applied to the dataset where the background was filtered water. The different metrics including classification accuracy, specificity, sensitivity, and area under the curve (AUC) of the ROC were computed, and their performances were summarized in Chapter 2. It has been shown that Random Forest results in the best performance in terms of AUC of ROC and misclassification error. Also, the SVM with 'rbf' kernel and PCA-QDA using nonlinear kernel trick had a better performance than the linear methods such as PCA-LDA and linear PCA-SVM. This suggests that usage of linear methods is not the best approach when the dataset has high complexity.

In Chapter 3, a unique end-to-end deep neural network architecture was introduced to identify *S. pyogenes*. As current methods of analyzing Raman spectra are based on pre-processing methods composed of removing the fluorescent background, a multilayer neural network is introduced for training on raw and pre-processed spectra, acquired by an expert. Then, known bands of different biological macromolecules are embedded in the network to enable network learning from pre-known bands or wavenumbers. Hence, the identification network was presented and trained on the same dataset used in the previous chapter. This unique deep neural network resulted in the best classification error, 0%, and showed that using deep learning methods can achieve higher performance than using traditional classification methods used in Chapter 2.

In Chapter 4, in order to reach one step closer to generalize our approach for clinical applications, we determined to apply the proposed method presented in Chapter 3 to a dataset with confounding background acquired in two different media: the water and throat swab background. The confounding background has higher complexity than the water background in terms of fluorescent background and variation of the patterns of the species. The former affects the performance of the pre-processing units, and the latter one affects the performance of the identification network. Nevertheless, the proposed networks were trained using this dataset, and the results were illustrated. It was concluded that the misclassification error for test dataset was 3.7% . Furthermore, the realization of the identification network was presented to provide a better understating of the use of the bands of the known macromolecules in the network. For this purpose, each macromolecule was considered individually to determine its contribution to the true and false detections for positive and negative samples. It was shown that some macromolecules result in higher probability than others. These findings were summarized in Table 10.

In conclusion, this dissertation has attempted to contribute to identifying pathogens using Raman Spectroscopy by introducing and justifying new methods to improve state-of-the-art in the analysis of Raman spectra.

## 5.2 Future Works

In conclusion to Chapters 2 through 4, it would be valuable to discuss and provide possible extensions and future directions. As discussed, this dissertation was restricted to a dataset with a limited number of pathogens due to time and the expensive labor process of data acquisition. We have attempted to abate this issue by choosing various ranges of pathogens

which have a similar characteristic to *S. pyogenes*. Notwithstanding, it is my hope that the proposed technique will continue to be tested on new pathogens and also various strains of pathogens and will perhaps inspire new research toward more efficient and generalized deep neural networks to identify pathogens using Raman spectroscopy and background removal of spectra.

There is a potential for vast applications of real-time identification of pathogens in the clinical and biological arena. One such application could be identifying particular pathogens in food, water, blood, and mucus. Utilization of a real-time identification method can not only save time and costs but also save lives by identifying the pathogen quickly or prevent pathogens from spreading in case of an epidemic disease.

Furthermore, the method presented in this dissertation can be extended to address antibiotic usage for the elimination of pathogens. As some pathogens are resistant to antibiotics, monitoring the effect of antibiotic use in the real-time or short interval can prohibit the unnecessary usage of antibiotics in some cases.

This work can be extended to the multi-classification problem as well that I did not have a chance to investigate in this dissertation. However, I believe the most important improvement in the future of the pathogen identification will come from the area of deep reinforcement learning. Deep reinforcement learning with a capability to determine actions in sequence can be learned to predict the diagnostic decisions sequentially which enable us to use biological knowledge of pathogens effectively. Nevertheless, given the difficulty of reinforcement deep learning on complex tasks, it might take years for such methods to outperform deep learning models, and thus, in the meantime improving the deep learning models will continue to be beneficial. Also, deep learning methods can be used as a part of

such deep reinforcement learning methods.

## REFERENCES

- [1] W. E. Huang, M. Li, R. M. Jarvis, R. Goodacre, and S. A. Banwart, "Shining light on the microbial world: the application of raman microspectroscopy," *Advances in applied microbiology*, vol. 70, pp. 153–186, 2010.
- [2] X. Lu, H. M. Al-Qadiri, M. Lin, and B. A. Rasco, "Application of mid-infrared and raman spectroscopy to the study of bacteria," *Food and Bioprocess Technology*, vol. 4, no. 6, pp. 919–935, 2011.
- [3] A. C. S. Talari, Z. Movasaghi, S. Rehman, and I. U. Rehman, "Raman spectroscopy of biological tissues," *Applied Spectroscopy Reviews*, vol. 50, no. 1, pp. 46–111, 2015.
- [4] [http://www.doitpoms.ac.uk/tlplib/raman/raman\\_microspectroscopy.php](http://www.doitpoms.ac.uk/tlplib/raman/raman_microspectroscopy.php).
- [5] J. Todd, M. Fishaut, F. Kapral, and T. Welch, "Toxic-shock syndrome associated with phage-group-i staphylococci," *The Lancet*, vol. 312, no. 8100, pp. 1116–1118, 1978.
- [6] F. Dobbs, "A scoring system for predicting group a streptococcal throat infection.," *Br J Gen Pract*, vol. 46, no. 409, pp. 461–464, 1996.
- [7] G. V. Doern, R. Vautour, M. Gaudet, and B. Levy, "Clinical impact of rapid in vitro susceptibility testing and bacterial identification.," *Journal of clinical microbiology*, vol. 32, no. 7, pp. 1757–1762, 1994.
- [8] A. M. Fine, V. Nizet, and K. D. Mandl, "Large-scale validation of the centor and mcisaac scores to predict group a streptococcal pharyngitis," *Archives of internal medicine*, vol. 172, no. 11, pp. 847–852, 2012.
- [9] J. R. Carapetis, A. C. Steer, E. K. Mulholland, and M. Weber, "The global burden of group a streptococcal diseases," *The Lancet infectious diseases*, vol. 5, no. 11, pp. 685–694, 2005.
- [10] S. T. Shulman, A. L. Bisno, H. W. Clegg, M. A. Gerber, E. L. Kaplan, G. Lee, J. M. Martin, and C. Van Beneden, "Clinical practice guideline for the diagnosis and management of group a streptococcal pharyngitis: 2012 update by the infectious diseases society of america," *Clinical Infectious Diseases*, p. cis629, 2012.

- [11] A. S. McKee, A. S. McDermid, D. Ellwood, and P. Marsh, "The establishment of reproducible, complex communities of oral bacteria in the chemostat using defined inocula," *Journal of applied bacteriology*, vol. 59, no. 3, pp. 263–275, 1985.
- [12] P. Tille, *Bailey & Scott's Diagnostic Microbiology-E-Book*. Elsevier Health Sciences, 2015.
- [13] W. C. Evans, *Trease and Evans' Pharmacognosy E-Book*. Elsevier Health Sciences, 2009.
- [14] L. Mariey, J. Signolle, C. Amiel, and J. Travert, "Discrimination, classification, identification of microorganisms using ftir spectroscopy and chemometrics," *Vibrational spectroscopy*, vol. 26, no. 2, pp. 151–159, 2001.
- [15] H. Yang and J. Irudayaraj, "Rapid detection of foodborne microorganisms on food surface using fourier transform raman spectroscopy," *Journal of Molecular Structure*, vol. 646, no. 1, pp. 35–43, 2003.
- [16] P. D. Taylor, O. Vinn, A. Kudryavtsev, and J. W. Schopf, "Raman spectroscopic study of the mineral composition of cirratulid tubes (annelida, polychaeta)," *Journal of structural biology*, vol. 171, no. 3, pp. 402–405, 2010.
- [17] B. Chen, "Raman spectroscopy studies of carbon nanotube-polymer composites," *Dekker Encyclopedia of Nanoscience and Nanotechnology*, vol. 1580, p. 3267, 2004.
- [18] I. Nabiev, I. Chourpa, and M. Manfait, "Applications of raman and surface-enhanced raman scattering spectroscopy in medicine," *Journal of Raman Spectroscopy*, vol. 25, no. 1, pp. 13–23, 1994.
- [19] J. Ferraro, K. Nakamoto, and C. W. Brown, "Introductory raman spectroscopy. 2003."
- [20] G. C. Green, A. D. Chan, B. S. Luo, H. Dan, and M. Lin, "Identification of listeria species using a low-cost surface-enhanced raman scattering system with wavelet-based signal processing," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 10, pp. 3713–3722, 2009.
- [21] A. Männig, N. A. Baldauf, L. A. Rodriguez-Romo, A. E. Yousef, and L. E. Rodríguez-Saona, "Differentiation of salmonella enterica serovars and strains in cultures and food using infrared spectroscopic and microspectroscopic techniques combined with soft independent modeling of class analogy pattern recognition analysis," *Journal of Food Protection®*, vol. 71, no. 11, pp. 2249–2256, 2008.

- [22] S. Efrima and L. Zeiri, "Understanding sers of bacteria," *Journal of Raman Spectroscopy*, vol. 40, no. 3, pp. 277–288, 2009.
- [23] W. E. Huang, M. J. Bailey, I. P. Thompson, A. S. Whiteley, and A. J. Spiers, "Single-cell raman spectral profiles of pseudomonas fluorescens sbw25 reflects in vitro and in planta metabolic history," *Microbial ecology*, vol. 53, no. 3, pp. 414–425, 2007.
- [24] N. M. Amiali, M. R. Mulvey, J. Sedman, M. Louie, A. E. Simor, and A. A. Ismail, "Rapid identification of coagulase-negative staphylococci by fourier transform infrared spectroscopy," *Journal of microbiological methods*, vol. 68, no. 2, pp. 236–242, 2007.
- [25] A. Sengupta, M. Mujacic, and E. J. Davis, "Detection of bacteria by surface-enhanced raman spectroscopy," *Analytical and bioanalytical chemistry*, vol. 386, no. 5, pp. 1379–1386, 2006.
- [26] A. Sujith, T. Itoh, H. Abe, K.-i. Yoshida, M. S. Kiran, V. Biju, and M. Ishikawa, "Imaging the cell wall of living single yeast cells using surface-enhanced raman spectroscopy," *Analytical and bioanalytical chemistry*, vol. 394, no. 7, pp. 1803–1809, 2009.
- [27] W. Premasiri, D. Moir, M. Klempner, N. Krieger, G. Jones, and L. Ziegler, "Characterization of the surface enhanced raman scattering (sers) of bacteria," *The journal of physical chemistry B*, vol. 109, no. 1, pp. 312–320, 2005.
- [28] H. Chu, Y. Huang, and Y. Zhao, "Silver nanorod arrays as a surface-enhanced raman scattering substrate for foodborne pathogenic bacteria detection," *Applied spectroscopy*, vol. 62, no. 8, pp. 922–931, 2008.
- [29] R. Pucek, V. Ranc, L. Kvítek, A. Panáček, R. Zbořil, and M. Kolář, "Reproducible discrimination between gram-positive and gram-negative bacteria using surface enhanced raman spectroscopy with infrared excitation," *Analyst*, vol. 137, no. 12, pp. 2866–2870, 2012.
- [30] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

- [31] J. MacQueen *et al.*, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.
- [32] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.
- [33] T. W. Anderson, *An introduction to multivariate statistical analysis*, vol. 2.
- [34] J. B. Tenenbaum, V. De Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
- [35] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [36] D. L. Donoho, M. Elad, and V. N. Temlyakov, “Stable recovery of sparse overcomplete representations in the presence of noise,” *IEEE Transactions on information theory*, vol. 52, no. 1, pp. 6–18, 2006.
- [37] H. Lee, A. Battle, R. Raina, and A. Y. Ng, “Efficient sparse coding algorithms,” in *Advances in neural information processing systems*, pp. 801–808, 2007.
- [38] B. A. Olshausen and D. J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, no. 6583, p. 607, 1996.
- [39] M. Meila and J. Shi, “Learning segmentation by random walks,” in *Advances in neural information processing systems*, pp. 873–879, 2001.
- [40] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
- [41] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, ACM, 2008.
- [42] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.



- [43] R. M. Neal, "Connectionist learning of belief networks," *Artificial intelligence*, vol. 56, no. 1, pp. 71–113, 1992.
- [44] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," tech. rep., COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE, 1986.
- [45] N. Le Roux and Y. Bengio, "Representational power of restricted boltzmann machines and deep belief networks," *Neural computation*, vol. 20, no. 6, pp. 1631–1649, 2008.
- [46] G. E. Hinton, "Connectionist learning procedures," in *Machine Learning, Volume III*, pp. 555–610, Elsevier, 1990.
- [47] P. E. Utgoff and D. J. Straczuzi, "Many-layered learning," *Neural Computation*, vol. 14, no. 10, pp. 2497–2529, 2002.
- [48] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.
- [49] M. F. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber, "Deep networks with internal selective attention through feedback connections," in *Advances in neural information processing systems*, pp. 3545–3553, 2014.
- [50] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [51] R. D. Joseph, *Contributions to perceptron theory*. Cornell Univ., 1961.
- [52] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," tech. rep., DTIC Document, 1985.
- [53] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 625–660, 2010.
- [54] J. Martens, "Deep learning via hessian-free optimization," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 735–742, 2010.

- [55] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, *et al.*, “Greedy layer-wise training of deep networks,” *Advances in neural information processing systems*, vol. 19, p. 153, 2007.
- [56] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [57] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [58] G. E. Hinton, “Deep belief networks,” *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.
- [59] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio, “Maxout networks,” *ICML (3)*, vol. 28, pp. 1319–1327, 2013.
- [60] J. Martens and I. Sutskever, “Learning recurrent neural networks with hessian-free optimization,” in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1033–1040, 2011.
- [61] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.
- [62] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [63] D. C. Ciresan, U. Meier, J. Masci, L. Maria Gambardella, and J. Schmidhuber, “Flexible, high performance convolutional neural networks for image classification,” in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, p. 1237, Barcelona, Spain, 2011.
- [64] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- [65] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

- [66] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, “The stanford corenlp natural language processing toolkit,” in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60, 2014.
- [67] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, “Recurrent neural network based language model,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
- [68] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.
- [69] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, “Mastering the game of go with deep neural networks and tree search,” *nature*, vol. 529, no. 7587, pp. 484–489, 2016.
- [70] J. Canny, “A computational approach to edge detection,” *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.
- [71] R. M. Haralick, K. Shanmugam, *et al.*, “Textural features for image classification,” *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.
- [72] M. B. Blaschko and C. H. Lampert, “Correlational spectral clustering,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.
- [73] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, “cudnn: Efficient primitives for deep learning,” *arXiv preprint arXiv:1410.0759*, 2014.
- [74] D. Ciregan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on*, pp. 3642–3649, IEEE, 2012.
- [75] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, *et al.*, “Large scale distributed deep networks,” in *Advances in neural information processing systems*, pp. 1223–1231, 2012.

- [76] C.-C. Cheng, F. Sha, and L. K. Saul, "A fast online algorithm for large margin training of continuous density hidden markov models," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.
- [77] M. Zourob, S. Elwary, and A. P. Turner, *Principles of bacterial detection: biosensors, recognition receptors and microsystems*. Springer Science & Business Media, 2008.
- [78] K. Rebrošová, M. Šiler, O. Samek, F. Růžička, S. Bernatová, V. Holá, J. Ježek, P. Zemánek, J. Sokolová, and P. Petráš, "Rapid identification of staphylococci by raman spectroscopy," *Scientific reports*, vol. 7, no. 1, p. 14846, 2017.
- [79] N. Tien, H.-C. Chen, S.-L. Gau, T.-H. Lin, H.-S. Lin, B.-J. You, P.-C. Tsai, I.-R. Chen, M.-F. Tsai, I.-K. Wang, *et al.*, "Diagnosis of bacterial pathogens in the dialysate of peritoneal dialysis patients with peritonitis using surface-enhanced raman spectroscopy," *Clinica Chimica Acta*, vol. 461, pp. 69–75, 2016.
- [80] M. Wulf, D. Willemse-Erix, C. Verduin, G. Puppels, A. van Belkum, and K. Maquelin, "The use of raman spectroscopy in the epidemiology of methicillin-resistant staphylococcus aureus of human-and animal-related clonal lineages," *Clinical Microbiology and Infection*, vol. 18, no. 2, pp. 147–152, 2012.
- [81] A. Martinelli, "Effects of a protic ionic liquid on the reaction pathway during non-aqueous sol-gel synthesis of silica: A raman spectroscopic investigation," *International journal of molecular sciences*, vol. 15, no. 4, pp. 6488–6503, 2014.
- [82] E. Brauchle and K. Schenke-Layland, "Raman spectroscopy in biomedicine—non-invasive in vitro analysis of cells and extracellular matrix components in tissues," *Biotechnology journal*, vol. 8, no. 3, pp. 288–297, 2013.
- [83] U. Neugebauer, P. Rösch, and J. Popp, "Raman spectroscopy towards clinical application: drug monitoring and pathogen identification," *International journal of antimicrobial agents*, vol. 46, pp. S35–S39, 2015.
- [84] O. Samek, A. Jonáš, Z. Pilát, P. Zemánek, L. Nedbal, J. Tříška, P. Kotas, and M. Trtílek, "Raman

- microspectroscopy of individual algal cells: sensing unsaturation of storage lipids in vivo,” *Sensors*, vol. 10, no. 9, pp. 8635–8651, 2010.
- [85] K. C. Schuster, E. Urlaub, and J. Gapes, “Single-cell analysis of bacteria by raman microscopy: spectral information on the chemical composition of cells and on the heterogeneity in a culture,” *Journal of Microbiological Methods*, vol. 42, no. 1, pp. 29–38, 2000.
- [86] K. Rebrošová, M. Šiler, O. Samek, F. Růžička, S. Bernatová, J. Ježek, P. Zemánek, and V. Holá, “Differentiation between staphylococcus aureus and staphylococcus epidermidis strains using raman spectroscopy,” *Future microbiology*, vol. 12, no. 10, pp. 881–890, 2017.
- [87] N. K. Afseth, M. Bloomfield, J. P. Wold, and P. Matousek, “A novel approach for subsurface through-skin analysis of salmon using spatially offset raman spectroscopy (sors),” *Applied spectroscopy*, vol. 68, no. 2, pp. 255–262, 2014.
- [88] S. Pahlow, S. Meisel, D. Cialla-May, K. Weber, P. Rösch, and J. Popp, “Isolation and identification of bacteria by means of raman spectroscopy,” *Advanced drug delivery reviews*, vol. 89, pp. 105–120, 2015.
- [89] J. De Gelder, K. De Gussem, P. Vandenabeele, and L. Moens, “Reference database of raman spectra of biological molecules,” *Journal of Raman spectroscopy*, vol. 38, no. 9, pp. 1133–1147, 2007.
- [90] S. Bernatová, O. Samek, Z. Pilát, M. Šerý, J. Ježek, P. Ják, M. Šiler, V. Krzyžánek, P. Zemánek, V. Holá, *et al.*, “Following the mechanisms of bacteriostatic versus bactericidal action using raman spectroscopy,” *Molecules*, vol. 18, no. 11, pp. 13188–13199, 2013.
- [91] R. Mathey, M. Dupoy, I. Espagnon, D. Leroux, F. Mallard, and A. Novelli-Rousseau, “Viability of 3 h grown bacterial micro-colonies after direct raman identification,” *Journal of microbiological methods*, vol. 109, pp. 67–73, 2015.
- [92] R. Gautam, A. Samuel, S. Sil, D. Chaturvedi, A. Dutta, F. Ariese, S. Umaphathy, *et al.*, “Raman and mid-infrared spectroscopic imaging: Applications and advancements,” 2015.
- [93] S. Sil, R. Mukherjee, N. Kumar, S. Aravind, J. Kingston, and U. Singh, “Detection and classification of bacteria using raman spectroscopy combined with multivariate analysis,” *Defence Life Science Journal*, vol. 2, no. 4, pp. 435–441, 2017.

- [94] R. Goodacre, “Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules,” *Vibrational Spectroscopy*, vol. 32, no. 1, pp. 33–45, 2003.
- [95] X. Lu and B. Rasco, “Investigating food spoilage and pathogenic microorganisms by mid-infrared spectroscopy,” *Handbook of Vibrational Spectroscopy*, 2010.
- [96] R. M. Jarvis, A. Brooker, and R. Goodacre, “Surface-enhanced raman scattering for the rapid discrimination of bacteria,” *Faraday discussions*, vol. 132, pp. 281–292, 2006.
- [97] B. F. Manly, *Multivariate statistical methods: a primer*. CRC Press, 1994.
- [98] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, and R. Tatham, “Multivariate data analysis (7th eds.),” *NY: Pearson*, 2010.
- [99] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [100] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
- [101] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
- [102] Y. Seo, B. Park, A. Hinton, S.-C. Yoon, and K. C. Lawrence, “Identification of staphylococcus species with hyperspectral microscope imaging and classification algorithms,” *Journal of Food Measurement and Characterization*, vol. 10, no. 2, pp. 253–263, 2016.
- [103] S. Stöckel, S. Meisel, B. Lorenz, S. Kloß, S. Henk, S. Dees, E. Richter, S. Andres, M. Merker, I. Labugger, *et al.*, “Raman spectroscopic identification of mycobacterium tuberculosis,” *Journal of Biophotonics*, vol. 5, no. 10, pp. 727–734, 2017.
- [104] W. Tong, H. Hong, H. Fang, Q. Xie, and R. Perkins, “Decision forest: combining the predictions of multiple independent decision tree models,” *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 2, pp. 525–531, 2003.
- [105] T. Dietterich, “Ensemble learning, the handbook of brain theory and neural networks, ma arbib,” 2002.

- [106] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [107] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [108] A. Cao, A. K. Pandya, G. K. Serhatkulu, R. E. Weber, H. Dai, J. S. Thakur, V. M. Naik, R. Naik, G. W. Auner, R. Rabah, *et al.*, "A robust method for automated background subtraction of tissue fluorescence," *Journal of Raman Spectroscopy*, vol. 38, no. 9, pp. 1199–1205, 2007.
- [109] Z.-M. Zhang, S. Chen, and Y.-Z. Liang, "Baseline correction using adaptive iteratively reweighted penalized least squares," *Analyst*, vol. 135, no. 5, pp. 1138–1146, 2010.
- [110] Z. Li, D.-J. Zhan, J.-J. Wang, J. Huang, Q.-S. Xu, Z.-M. Zhang, Y.-B. Zheng, Y.-Z. Liang, and H. Wang, "Morphological weighted penalized least squares for background correction," *Analyst*, vol. 138, no. 16, pp. 4483–4492, 2013.
- [111] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, pp. 1137–1145, Montreal, Canada, 1995.
- [112] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical Methods*, vol. 6, no. 9, pp. 2812–2831, 2014.
- [113] W. R. Klecka, *Discriminant analysis*. No. 19, Sage, 1980.
- [114] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and qsar modeling," *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.
- [115] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011.
- [116] G. Valentini and T. G. Dietterich, "Bias-variance analysis of support vector machines for the development of svm-based ensemble methods," *Journal of Machine Learning Research*, vol. 5, no. Jul, pp. 725–775, 2004.

- [117] J. Zhao, H. Lui, D. I. McLean, and H. Zeng, "Automated autofluorescence background subtraction algorithm for biomedical raman spectroscopy," *Applied spectroscopy*, vol. 61, no. 11, pp. 1225–1232, 2007.
- [118] B. D. Beier and A. J. Berger, "Method for automated background subtraction from raman spectra containing known contaminants," *Analyst*, vol. 134, no. 6, pp. 1198–1202, 2009.
- [119] V. Mazet, C. Carteret, D. Brie, J. Idier, and B. Humbert, "Background removal from spectra by designing and minimising a non-quadratic cost function," *Chemometrics and intelligent laboratory systems*, vol. 76, no. 2, pp. 121–133, 2005.
- [120] C. A. Lieber and A. Mahadevan-Jansen, "Automated method for subtraction of fluorescence from biological raman spectra," *Applied spectroscopy*, vol. 57, no. 11, pp. 1363–1367, 2003.
- [121] Z.-M. Zhang, S. Chen, Y.-Z. Liang, Z.-X. Liu, Q.-M. Zhang, L.-X. Ding, F. Ye, and H. Zhou, "An intelligent background-correction algorithm for highly fluorescent samples in raman spectroscopy," *Journal of Raman Spectroscopy*, vol. 41, no. 6, pp. 659–669, 2010.
- [122] T. J. Vickers, R. E. Wambles, and C. K. Mann, "Curve fitting and linearity: data processing in raman spectroscopy," *Applied Spectroscopy*, vol. 55, no. 4, pp. 389–393, 2001.
- [123] B. Bendinger, R. M. Kroppenstedt, S. Klatte, and K. Altendorf, "Chemotaxonomic differentiation of coryneform bacteria isolated from biofilters," *International Journal of Systematic and Evolutionary Microbiology*, vol. 42, no. 3, pp. 474–486, 1992.
- [124] J. De Gelder, *Raman spectroscopy as a tool for studying bacterial cell compounds*. PhD thesis, Ghent University, 2008.
- [125] H. Brade, L. Brade, and E. T. Rietschel, "Structure-activity relationships of bacterial lipopolysaccharides (endotoxins): current and future aspects," *Zentralblatt für Bakteriologie, Mikrobiologie und Hygiene. Series A: Medical Microbiology, Infectious Diseases, Virology, Parasitology*, vol. 268, no. 2, pp. 151–179, 1988.



- [126] N. Bergström, P.-E. Jansson, M. Kilian, and U. B. Skov Sørensen, “Structures of two cell wall-associated polysaccharides of a streptococcus mitis biovar 1 strain,” *European Journal of Biochemistry*, vol. 267, no. 24, pp. 7147–7157, 2000.
- [127] X. Zhang, M. A. Young, O. Lyandres, and R. P. Van Duyne, “Rapid detection of an anthrax biomarker by surface-enhanced raman spectroscopy,” *Journal of the American Chemical Society*, vol. 127, no. 12, pp. 4484–4489, 2005.
- [128] R. M. Jarvis and R. Goodacre, “Discrimination of bacteria using surface-enhanced raman spectroscopy,” *Analytical Chemistry*, vol. 76, no. 1, pp. 40–47, 2004.
- [129] P. Rösch, M. Harz, M. Schmitt, K.-D. Peschke, O. Ronneberger, H. Burkhardt, H.-W. Motzkus, M. Lankers, S. Hofer, H. Thiele, *et al.*, “Chemotaxonomic identification of single bacteria by micro-raman spectroscopy: application to clean-room-relevant biological contaminations,” *Applied and environmental microbiology*, vol. 71, no. 3, pp. 1626–1637, 2005.
- [130] C. Xie, J. Mace, M. Dinno, Y. Li, W. Tang, R. Newton, and P. Gemperline, “Identification of single bacterial cells in aqueous solution using confocal laser tweezers raman spectroscopy,” *Analytical chemistry*, vol. 77, no. 14, pp. 4390–4397, 2005.
- [131] R. Collobert and J. Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, ACM, 2008.
- [132] X. Glorot, A. Bordes, and Y. Bengio, “Domain adaptation for large-scale sentiment classification: A deep learning approach,” in *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 513–520, 2011.
- [133] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.
- [134] R. Goodacre, E. M. Timmins, P. J. Rooney, J. J. Rowland, and D. B. Kell, “Rapid identification of streptococcus and enterococcus species using diffuse reflectance-absorbance fourier transform infrared

- spectroscopy and artificial neural networks,” *FEMS Microbiology Letters*, vol. 140, no. 2-3, pp. 233–239, 1996.
- [135] R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, A. Belopolsky, Y. Bengio, A. Bergeron, J. Bergstra, V. Bisson, J. Blecher Snyder, N. Bouchard, N. Boulanger-Lewandowski, X. Bouthillier, A. de Brébisson, O. Breuleux, P.-L. Carrier, K. Cho, J. Chorowski, P. Christiano, T. Cooijmans, M.-A. Côté, M. Côté, A. Courville, Y. N. Dauphin, O. Delalleau, J. Demouth, G. Desjardins, S. Dieleman, L. Dinh, M. Ducoffe, V. Dumoulin, S. Ebrahimi Kahou, D. Erhan, Z. Fan, O. Firat, M. Germain, X. Glorot, I. Goodfellow, M. Graham, C. Gulcehre, P. Hamel, I. Harlouchet, J.-P. Heng, B. Hidasi, S. Honari, A. Jain, S. Jean, K. Jia, M. Korobov, V. Kulkarni, A. Lamb, P. Lamblin, E. Larsen, C. Laurent, S. Lee, S. Lefrancois, S. Lemieux, N. Léonard, Z. Lin, J. A. Livezey, C. Lorenz, J. Lowin, Q. Ma, P.-A. Manzagol, O. Mastropietro, R. T. McGibbon, R. Memisevic, B. van Merriënboer, V. Michalski, M. Mirza, A. Orlandi, C. Pal, R. Pascanu, M. Pezeshki, C. Raffel, D. Renshaw, M. Rocklin, A. Romero, M. Roth, P. Sadowski, J. Salvatier, F. Savard, J. Schlüter, J. Schulman, G. Schwartz, I. V. Serban, D. Serdyuk, S. Shabanian, E. Simon, S. Spieckermann, S. R. Subramanyam, J. Sygnowski, J. Tanguay, G. van Tulder, J. Turian, S. Urban, P. Vincent, F. Visin, H. de Vries, D. Warde-Farley, D. J. Webb, M. Willson, K. Xu, L. Xue, L. Yao, S. Zhang, and Y. Zhang, “Theano: A Python framework for fast computation of mathematical expressions,” *arXiv e-prints*, vol. abs/1605.02688, May 2016.
- [136] P. Y. Simard, D. Steinkraus, J. C. Platt, *et al.*, “Best practices for convolutional neural networks applied to visual document analysis,” in *ICDAR*, vol. 3, pp. 958–962, 2003.
- [137] A. Canning and E. Gardner, “Partially connected models of neural networks,” *Journal of Physics A: Mathematical and General*, vol. 21, no. 15, p. 3275, 1988.
- [138] P.-C. Chang *et al.*, “A novel model by evolving partially connected neural network for stock price trend forecasting,” *Expert Systems with Applications*, vol. 39, no. 1, pp. 611–620, 2012.
- [139] D. Elizondao, E. Fiesler, and J. Korczak, “Non-ontogenic sparse neural networks,” in *Neural Networks, 1995. Proceedings., IEEE International Conference on*, vol. 1, pp. 290–295, IEEE, 1995.

- [140] S. Kang and C. Isik, "Partially connected feedforward neural networks structured by input types," *IEEE transactions on neural networks*, vol. 16, no. 1, pp. 175–184, 2005.
- [141] E. Fiesler, "Comparative bibliography of ontogenic neural networks," in *ICANN'94*, pp. 793–796, Springer, 1994.
- [142] D. Elizondo and E. Fiesler, "A survey of partially connected neural networks," *International journal of neural systems*, vol. 8, no. 05n06, pp. 535–558, 1997.

**ABSTRACT****IDENTIFICATION OF STREPTOCOCCUS PYOGENES  
USING RAMAN SPECTROSCOPY**

by

**EHSAN MAJIDI****JUNE 2018****Advisor:** Prof. Gregory W. Auner**Major:** Electrical Engineering**Degree:** Doctor of Philosophy

Despite the attention that Raman Spectroscopy has gained recently in the area of pathogen identification, the spectra analyses techniques are not well developed. In most scenarios they rely on expert intervention to detect and assign the peaks of the spectra to specific molecular vibration. Although some investigators have used machine-learning techniques to classify pathogens, these studies are usually limited to a specific application, and the generalization of these techniques is not clear. Also, a wide range of algorithms have been developed for classification problems, however, there is less insight to applying such methods on Raman spectra. Furthermore, analyzing the Raman spectra requires pre-processing of the raw spectra, in particular background removing. Various techniques are developed to remove the background of the raw spectra accurately and with or without less expert intervention. Nevertheless, as the background of the spectra varies in the different media, these methods still require expert effort adding complexity and inefficiency to the identification task. This dissertation describes the development of state-of-the-art classification techniques to identify *S. pyogenes* from other species, including water and other confounding background pathogens. We compared these techniques in terms of their classification accuracy, sensitivity, and specificity in addition to providing a bias-variance insight in selecting the number of principal components in a principal component analysis (PCA). It was observed that Random Forest provided a better result with an accuracy of 94.11%.

Next, a novel deep learning technique was developed to remove background of the Raman spectra and then identify the pathogen. The architecture of the network was discussed and it was found that this method yields an accuracy of 100% in our test samples. This outperforms other traditional machine learning techniques as discussed. In clinical applications of Raman Spectroscopy, the samples have confounding background creates a challenging task for the removal of the spectral background and subsequent identification of the pathogen in real-time. We tested our methodology on datasets composed of confounding background such as throat swabs from patients and discussed the robustness and generalization of the developed method. It was found that the misclassification error of the test dataset was around 3.7%. Also the realization of the trained model is discussed in detail to provide a better understating and insight into the efficacy of the deep learning architecture. This technique provides a platform for general analysis of other pathogens in confounding environments as well.

## AUTOBIOGRAPHICAL STATEMENT

**Ehsan Majidi** received his Bachelor of Science degree in Electrical and Computer Engineering from Shiraz University in 2008, and his Master of Science degree in Electrical and Biomedical Engineering from the University of Tehran in 2011. His master thesis was focused on pattern recognition and signal processing where he studied the connectivity and patterns of cortex sources from EEG signals. Overall, the medical imaging, statistical signal processing, control theory, brain-computer interface, and Machine Learning were the areas he was involved in and studied during his undergraduate and graduate programs.

Ehsan joined Wayne State University as a Ph.D. student in 2013. He found Machine Learning and Deep Neural Networks astonishing areas to work on and investigated the pathogen identification with Machine and Deep Learning using Raman Spectroscopy as part of his dissertation project. Also, he taught laboratory courses in the Computer Science and Electrical and Computer Engineering department.

Although Artificial Intelligence is his primary passion, he is interested in various fields of technology and science. One of his hobbies is to broaden his knowledge in other areas such as chemistry, biology, genetics, and finance.