Characterization of cervid skin tissues with chronic wasting disease by Raman spectroscopy and machine learning

Binbin Zhu
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Recommended Citation

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Characterization of cervid skin tissues with chronic wasting disease by Raman spectroscopy and machine learning

by

Binbin Zhu

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Agricultural and Biosystems Engineering

Program of Study Committee:
Chenxu Yu, Major Professor
M. Heather Greenlee
Kurt Rosentrater
Yumou Qiu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this thesis. The Graduate College will ensure this thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University
Ames, Iowa
2020

Copyright © Binbin Zhu, 2020. All rights reserved.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>vii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>viii</td>
</tr>
<tr>
<td><strong>CHAPTER 1. GENERAL INTRODUCTION AND BACKGROUND</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Chronic Wasting Disease</td>
<td>1</td>
</tr>
<tr>
<td>1.1.2 The Prions</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Diagnosis of CWD</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1 Clinical signs</td>
<td>2</td>
</tr>
<tr>
<td>1.2.2 Laboratory diagnosis via biopsy of tissue samples</td>
<td>3</td>
</tr>
<tr>
<td>1.2.3 CWD diagnosis in living animals</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Raman spectroscopy</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Raman Spectral Data Processing</td>
<td>7</td>
</tr>
<tr>
<td>1.4.1 Preprocessing</td>
<td>7</td>
</tr>
<tr>
<td>1.4.2 Principal Component Analysis</td>
<td>10</td>
</tr>
<tr>
<td>1.4.3 Classification</td>
<td>11</td>
</tr>
<tr>
<td>1.5 Research Objectives</td>
<td>14</td>
</tr>
<tr>
<td>1.6 Thesis Overview</td>
<td>14</td>
</tr>
<tr>
<td>References</td>
<td>14</td>
</tr>
<tr>
<td><strong>CHAPTER 2. DETECTION AND CHARACTERIZATION OF DEER SKIN TISSUE FOR CHRONIC WASTING DISEASE WITH RAMAN MICROSCOPY AND MACHINE LEARNING</strong></td>
<td>17</td>
</tr>
<tr>
<td>2.0 Abstract</td>
<td>17</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>18</td>
</tr>
<tr>
<td>2.2 Materials and Methods</td>
<td>20</td>
</tr>
<tr>
<td>2.3 Results and Discussion</td>
<td>23</td>
</tr>
<tr>
<td>2.3.1 Analysis of Raman spectra collected from the skin samples</td>
<td>23</td>
</tr>
<tr>
<td>2.3.2 Analysis of Raman spectra from de-paraffined skin samples</td>
<td>27</td>
</tr>
<tr>
<td>2.4 Conclusion</td>
<td>32</td>
</tr>
<tr>
<td>References</td>
<td>32</td>
</tr>
</tbody>
</table>
CHAPTER 3. DETECTION AND CHARACTERIZATION OF DEER SKIN TISSUE FOR CHRONIC WASTING DISEASE WITH PORTABLE RAMAN SPECTROMETER AND MACHINE LEARNING ................................................................. 35

3.0 Abstract .................................................................................................................. 35
3.1 Introduction ............................................................................................................ 36
3.2 Materials and Methods ........................................................................................ 38
3.3 Results and Discussion ........................................................................................ 41
3.4 Conclusion ............................................................................................................. 45
References .................................................................................................................... 45

CHAPTER 4. GENERAL CONCLUSION AND FUTURE PERSPECTIVE ............... 47
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1.</td>
<td>Characteristics of transmissible spongiform encephalopathies (TSEs) or prion disease. Image credit to Kerry L. Helms, Scientific Illustrator (Public domain).</td>
<td>2</td>
</tr>
<tr>
<td>Figure 1.2.</td>
<td>Immunohistochemical detection of PrPres in retropharyngeal node lymphoid follicles (red, arrows) of a fawn exposed orally to CWD-positive brain inoculum (a, b). No PrPres staining was detected in the retropharyngeal node follicles (arrows) of fawns exposed to CWD-negative brain inocula (c, d). Bar, 100 µm (a, c) or 10 µm (b, d). Reproduced with permission.</td>
<td>4</td>
</tr>
<tr>
<td>Figure 1.3.</td>
<td>Three types of scattering processes</td>
<td>6</td>
</tr>
<tr>
<td>Figure 1.4.</td>
<td>Jablonski Diagram of energy states for Rayleigh, Stokes and Anti-Stokes Raman Scattering</td>
<td>6</td>
</tr>
<tr>
<td>Figure 1.5.</td>
<td>Simple Moving Average Example</td>
<td>7</td>
</tr>
<tr>
<td>Figure 1.6.</td>
<td>(a) Baseline value; (b) Spectrum after baseline correction</td>
<td>9</td>
</tr>
<tr>
<td>Figure 1.7.</td>
<td>Demonstration of a feedforward neural network in this research</td>
<td>13</td>
</tr>
<tr>
<td>Figure 2.1.</td>
<td>Demonstration of a feedforward neural network in this research</td>
<td>23</td>
</tr>
<tr>
<td>Figure 2.2.</td>
<td>Spectra of negative PSS and positive PSS with p-value of t-test</td>
<td>24</td>
</tr>
<tr>
<td>Figure 2.3.</td>
<td>The cumulative variance of 10 PCs, 50 PCs, and 200 PCs for PSS</td>
<td>25</td>
</tr>
<tr>
<td>Figure 2.4.</td>
<td>Clustering of PC0 and PC1 for negative and positive PSS</td>
<td>26</td>
</tr>
<tr>
<td>Figure 2.5.</td>
<td>Decision boundaries of PC0 and PC1 for the 50 PCs model</td>
<td>26</td>
</tr>
<tr>
<td>Figure 2.6.</td>
<td>Spectra of negative DSS and positive DSS with p-value of t-test</td>
<td>28</td>
</tr>
<tr>
<td>Figure 2.7.</td>
<td>The cumulative variance of 10 PCs, 50 PCs, and 200 PCs for DSS</td>
<td>30</td>
</tr>
<tr>
<td>Figure 2.8.</td>
<td>Clustering of PC0 and PC1 for negative and positive DSS</td>
<td>31</td>
</tr>
<tr>
<td>Figure 3.1.</td>
<td>Demonstration of a feedforward neural network in this research. N is the input dimensional number, which equals 10, 50 or 998 for different types training data.</td>
<td>41</td>
</tr>
<tr>
<td>Figure 3.2.</td>
<td>(a) Spectra of paraffin, negative and positive tissue block samples with p-value of t-test; (b) The first 5 PC loadings showing peaks related to the spectra in positive tissue block.</td>
<td>42</td>
</tr>
</tbody>
</table>
Figure 3.3. The cumulative variance of 10 PCs, 50 PCs for tissue block samples

Figure 3.4. Clustering of PC0 and PC1 for negative and positive tissue block samples
LIST OF TABLES

Table 2.1. Explained variance and cumulative variance of five PC scores ........................................ 25
Table 2.2. Support vector machine results for PSS ........................................................................... 27
Table 2.4. Values of raman shift, intensity, p-value, and assignment of peaks shown in Figure 2.6 ...................................................................................................................... 28
Table 2.5. Raman shifts of peaks and the characteristic assignments ............................................. 29
Table 2.6. Support vector machine results for DSS ........................................................................ 31
Table 2.7. Artificial Neural Network results for DSS ..................................................................... 31
Table 3.1. Support vector machine results for tissue block samples ......................................... 44
Table 3.2. Artificial Neural Network results for tissue block samples ......................................... 44
ACKNOWLEDGMENTS

I would like to thank my committee chair, Dr. Chenxu Yu and my committee members, Dr. M Heather Greenlee, Dr. Kurt Rosentrater, and Dr. Yumou Qiu, for their guidance and support throughout the course of this research.

In addition, I would also like to thank my friends, colleagues, the department faculty and staff for making my time at Iowa State University a wonderful experience. I want to also offer my appreciation to my family who offered me patience, respect, and love, without whom, this thesis would not have been possible.
ABSTRACT

Chronic wasting disease (CWD) is a contagious neurological disease in cervids that belongs to transmissible spongiform encephalopathies (TSEs). Its spread has threatened the healthy growth of wild and farm-raised deer and resulted in adverse population-level impacts. It also raised concerns over the possibility of infecting human beings like bovine spongiform encephalopathy (BSE). CWD is a prion disease that may take as long as two years for visible signs of the disease to appear. Currently, diagnostic tests approved for official CWD are postmortem tests (immunohistochemistry (IHC) and ELISA) which are not suitable for in vivo diagnosis.

Raman spectroscopy offers a potential approach to detect and diagnose CWD rapidly in real time as a first screen onsite. With the Raman spectral data, machine learning algorithms could be utilized to extract meaningful information to differentiate the spectroscopic features that underline the signatures associated with the diseases effectively, even with a low signal-to-noise ratio (SNR) Raman spectral data acquired with a portable Raman spectrometer.

In this study, in order to evaluate the effectiveness of Raman spectroscopy on CWD diagnosis, Raman spectra were collected by a Raman microscope as well as a portable Raman spectrometer from cervid skin tissue samples collected from both healthy (i.e., control, CWD-negative) and diseased (i.e., CWD-positive) cervids. The spectral data were classified by two machine learning algorithms, support vector machine and artificial neural network. The results suggested that Raman spectroscopy in conjunction with Machine learning can indeed offer a rapid first screening for CWD, with the highest accuracy of 94.4%. It has the potential to become a useful tool for in-field diagnosis and detection of CWD.
CHAPTER 1. GENERAL INTRODUCTION AND BACKGROUND

1.1 Introduction

1.1.1 Chronic Wasting Disease

In 1967, the first case of Chronic Wasting Disease (CWD) was identified in Colorado. After that, CWD has emerged in wild or captive cervids in 25 states of the United States, 3 Canadian Provinces, Norway, and South Korea. It is a contagious neurological disease in cervid that belongs to transmissible spongiform encephalopathies (TSEs). The most widely accepted theory is that the disease is caused by an abnormal form of a cellular protein called a prion, most commonly found in the central nervous system and lymphoid tissue. The prion “infects” the host animal by promoting regular cellular protein convert to the abnormal form. Once the CWD infectious agent inserts into an animal, it results in emaciation, abnormal behavior, loss of bodily functions, and finally death. This process may take as long as two years before the animal begins to show visible signs of the disease. As of today, CWD is incurable and poses a fatal threat to animal health.

1.1.2 The Prions

The prion protein, PrP$^C$, as shown in Fig 1.1, is a small, cell-surface glycoprotein notable primarily for its critical role in the pathogenesis of the neurodegenerative disorders known as prion diseases, such as Creutzfeldt-Jakob disease in humans, bovine spongiform encephalopathy (BSE), scrapie in sheep and goats, and chronic wasting disease in cervids. (Castel and Gill, 2017) The accumulation of abnormal PrP$^C$, which is misfolded and rich in $\beta$- sheets, is supposed to be the cause of TSE. The pathologic PrP$^C$ was denoted as PrP scrapie (PrP$^{SC}$). (Gavier-Widen et al, 2005) The prion protein gene (PRNP) encoded PrP$^C$, which is present in mammalian species.
Studies of the PRNP gene in mule deer and white-tailed deer have identified many amino acid and nucleotide sequence mutations. Previous studies identified PRNP polymorphisms in elk and deer at codon positions 96, 132, 138, and 226. (O’Rourke et al., 1999; Raymond et al., 2000) Elk differs from white-tailed and mule deer at codon 226 and exhibits variability in codon 132. Both white-tailed and mule deer are polymorphic at codon 138, while heterogeneity at codon 96 is unique to white-tailed deer. Polymorphisms in PRNP could affect CWD susceptibility in cervids, impacting the degree to which cervid populations are vulnerable to disease transmission and progression.

![Characteristics of transmissible spongiform encephalopathies (TSEs) or prion disease.](image)

**Figure 1.1** Characteristics of transmissible spongiform encephalopathies (TSEs) or prion disease. Image credit to Kerry L. Helms, Scientific Illustrator (Public domain).

### 1.2 Diagnosis of CWD

#### 1.2.1 Clinical signs

Chronic Wasting Disease has a long incubation period, which usually lasts 18 months up to 3 years. After animals infected with CWD, there are no obvious symptoms until the late stage
of the disease. The first visible sign in this progress is emaciation. Then, animals lose bodily functions and show abnormal behavior, such as lack of control in bowel movements, listlessness, depression, aggression, lack of interaction with other animals, lowering of the head, and more. However, the only way to diagnose CWD definitively is through laboratory biopsy tests.

### 1.2.2 Laboratory diagnosis via biopsy of tissue samples

On microscopic examination, lesions of CWD in the central nervous system are similar to those of other spongiform encephalopathies. PrP<sup>Sc</sup> can be detected by immunohistochemistry (IHC), Western blotting, enzyme-linked immunosorbent assay (ELISA), prion misfolding cyclic amplification (PMCA), and real-time quaking induced conversion (RT-QuIC). (Gavier-Widen et al, 2005) According to USDA description, immunohistochemistry (IHC) testing of the obex area of the brain stem or the medial retropharyngeal lymph nodes is applied for definitive diagnosis. As shown in Fig 1.2, PrP<sup>Sc</sup> in retropharyngeal node lymphoid follicles was stained by IHC, and positive vs. negative samples with respect to CWD transfection could be clearly identified. (Sigurdson et al, 1999).

Emaciation and aspiration pneumonia are the most possible reasons to cause death from gross lesions in necropsy, which also correspond to the symptoms of CWD. In mule deer, a former study proved that PrP<sup>Sc</sup> could be detected by IHC as early as 6 months. (Sigurdson et al, 1999)
Figure 1.2. Immunohistochemical detection of PrPres in retropharyngeal node lymphoid follicles (red, arrows) of a fawn exposed orally to CWD-positive brain inoculum (a, b). No PrPres staining was detected in the retropharyngeal node follicles (arrows) of fawns exposed to CWD-negative brain inocula (c, d). Bar, 100 μm (a, c) or 10 μm (b, d). Reproduced with permission
1.2.3 CWD diagnosis in living animals

However, diagnostic assays for living animals are limited and inaccurate. The rectal mucosa biopsy sampling test was used to detect the CWD in white-tailed deer. The diagnostic sensitivity ranged from 63% to 100%, which was dependent on genotype at prion protein gene (Prnp) codon 96 and the stage of disease as assessed by obex grade. The diagnostic sensitivity reached 100% in the last 2 stages of the preclinical disease but was only 36% for the earliest stage. (Thomsen et al, 2012) Thus, USDA did not approve the rectal biopsy test for routine regulatory testing. More studies are greatly needed for the improvement of live-animal diagnostic tests for CWD.

1.3 Raman spectroscopy

Raman Spectroscopy is a non-destructive chemical analysis technique based on the scattering of light by molecules, which provides detailed information about chemical bond structures. There are three types of scattering processes that can occur when light interacts with an electron, shown in Figures 1.3 and Figure 1.4. The electrons have different vibrational levels. When an incident light interacts with an electron in the sample, the electron absorbs energy from the incident photon and rises to a virtual state of energy. Then, the electron falls back to an energy level by losing energy. If the energy lost equals the energy of the incident photon, the electron falls back to its initial level and therefore, the wavelength of the scattered photon is equivalent to that of the incident photon. The Rayleigh scattering or elastic scattering occurs. However, in a much rarer event (approximately 1 in 10 million photons), an inelastic scattering occurs that electrons will fall back to a different level. (Smith & G. Dent, 2005) If the electron gains energy from the incident photon during the scattering. The scattered photon loses energy, and its wavelength increases, which is called Stokes Raman scattering. Inversely, if the electron loses energy to a
lower vibrational level, the scattered photon gains the corresponding energy, and its wavelength decreases, which is called Anti-Stokes Raman scattering. Raman spectroscopy has been used for determining the biochemical changes in the bio-macromolecules of cells and tissues, including lipids, proteins and DNA, simultaneously. Moreover, the technique has been employed to analyze the body fluids, like blood plasma/serum and urine, for the diagnosis of the diseases. The potential of Raman spectroscopy for the diagnosis of diseases that are caused by viral infections, such as dengue and HPV, has also been demonstrated. (Ditta et al, 2018)

Figure 1.3. Three types of scattering processes

Figure 1.4. Jablonski Diagram of energy states for Rayleigh, Stokes and Anti-Stokes Raman Scattering
1.4 Raman Spectral Data Processing

In this study, data processing includes three parts, preprocessing, Principal Component Analysis (PCA), and Machine Learning for Discriminant Analysis.

1.4.1 Preprocessing

Data preprocessing is applied prior to the multivariate analysis to eliminate the noise caused by fluorescence, background, etc.

(a) Smoothing

The first step of preprocessing is smoothing, which removes the effect of random variation. Moving average, also called running or rolling average, is the most commonly used with time series data or spectral data. There are various types of moving average, simple moving average, cumulative moving average, weighted moving average, and exponential moving average. Simple Moving Average (SMA) uses a sliding window to take the average over a set number of time periods, as Figure 1.3 was shown. Suppose there is a spectrum with intensity Y and wavelength X. Two black lines set as a moving window, an interval \([X-H, X+H]\), where \(H\) is a small positive integer. Then, a new \(Y\) at \(X\) is obtained from an equally weighted mean of all \(Ys\) in the interval \([X-H, X+H]\). In this study, 10 points average was used for smoothing spectral data.

![Figure 1.5. Simple Moving Average Example](image)

Moving Window \([X-H, X+H]\)
(b) Baseline Correction

Baseline correction is an imperative step in data preprocessing because it extracts the actual intensity of Raman peaks. Various methods are applied to the baseline correction, such as linear fitting, polynomial fitting, wavelet transform, asymmetric least squares. The polynomial fitting algorithm was utilized in this study, which iteratively performs a polynomial fitting in the data to detect its baseline and finally give the best estimation of the real baseline. At every iteration, the fitting weights on the regions with peaks are reduced to identify the baseline only.

In the process of baseline processing, the intensities of the Raman spectra of N points are expressed as \(y = (y_1, \ldots, y_N)\) and \(y[x] = s[x] + b[x]\), where \(N\) is the number of measured spectral data, \(y[x]\) is the measured spectrum, \(s[x]\) is the analytical spectrum, and \(b[x]\) is the background spectrum. The background is modeled as a piecewise polynomial function, which can be written as

\[
b = \begin{cases} 
    b_1 + \alpha_1 x_1 + \alpha_2 x_2^2 + \cdots + \alpha_{p_1} x_1^{p_1}, & x_1 \leq x \leq \text{seg}_1 \\
    b_2 + \beta_1 x_1 + \beta_2 x_1^2 + \cdots + \beta_{p_2} x_1^{p_2}, & \text{seg}_1 \leq x \leq \text{seg}_2 \\
    \vdots \\
    b_n + \gamma_1 x_n + \gamma_2 x_n^2 + \cdots + \gamma_{p_n} x_n^{p_n}, & \text{seg}_n \leq x \leq x_n
\end{cases} \tag{Equation 1.1}
\]

where \(x_1\) and \(x_n\) are the Raman shifts corresponding to the first point and the last point \(\text{seg}_1\) to \(\text{seg}_n\) are the Raman shifts of segmentation points, and \(p_1\) to \(p_n\) represent the highest order of polynomial in each fitting, respectively. (Hu et al, 2018) Fig 1.4 shows the changes of spectrum after applying baseline correction processing.
Figure 1.6. (a) Baseline value; (b) Spectrum after baseline correction

(c) Normalization

Normalization is the process of scaling real valued numeric attributes into the range 0 and 1. In basic terms, normalize data is essential when the algorithm predicts based on the weighted relationships formed between data points. Although there are many other ways to normalize data, the “scikit-learn” machine learning library in Python provides three norms: “max”, “l1”, and “l2”. When creating a new instance of the Normalizer class you can specify the desired norm under the norm parameter. The “max” norm uses the absolute maximum.

\[ x_{\text{normalized}} = \frac{x}{\max(x)} \]  \hspace{1cm} \text{(Equation 1.2)}

The l1 norm uses the sum of all the values as and thus gives equal penalty to all parameters, enforcing sparsity.

\[ x_{\text{normalized}} = \frac{x}{\sum(x)} \]  \hspace{1cm} \text{(Equation 1.3)}

The l2 norm uses the square root of the sum of all the squared values. This creates smoothness and rotational invariance. Some models, like PCA, assume rotational invariance, and so l2 will perform better.

\[ x_{\text{normalized}} = \frac{x}{\sqrt{\sum_{i}a_{i}x_{i}^{2}}} \]  \hspace{1cm} \text{(Equation 1.4)}
1.4.2 Principal Component Analysis

Principal component analysis (PCA) is a multivariate statistical technique to reduce the dimensionality of data, increase interpretation, and minimize information loss. PCA was first formulated in statistics by Pearson, who formulated the analysis as finding “lines and planes of closest fit to systems of points in space”. (Pearson, 1901) Hotelling further developed PCA to its modern instantiation. (Hotelling, 1933) In the recent decades, PCA has been used for statistical analysis and classification tool in a variety of human diseases, such as near-infrared Raman spectroscopy of cervical precancer (Mahadevan-Janssen et al., 1998), near-infrared fluorescence spectroscopy of Alzheimer disease (Hanlon et al., 1999), FT-Raman spectroscopy of atherosclerosis in human carotid artery (Nogueira et al., 2005), and Raman spectroscopy for human breast cancer diagnosis (Manoharam et al., 1998), parathyroid tissue pathology (Dans et al., 2006), and discrimination of normal and malignant mucosal tissues of the colon (Chowdary et al., 2007).

Algebraically, principal components are particular linear combinations of the p random variables $X_1, X_2, \ldots, X_p$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $X_1, X_2, \ldots, X_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure.

Let the random vector $X' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix $\Sigma$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$.

Consider the linear combinations

\[
\begin{align*}
Y_1 &= a_1'X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= a_2'X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \\
Y_p &= a_p'X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{align*}
\]

(Equation 1.5)

Then, variance and covariance are
\begin{align*}
    \text{Var}(Y_i) &= a'_i \Sigma a_i \quad i = 1, 2, \ldots, p \quad \text{(Equation 1.6)} \\
    \text{Cov}(Y_i, Y_k) &= a'_i \Sigma a_k \quad i, k = 1, 2, \ldots, p \quad \text{(Equation 1.7)}
\end{align*}

The principal components (PCs) are those uncorrelated linear combinations \( Y_1, Y_2, \ldots, Y_p \) whose variances in 1.5 are as large as possible.

The first PC is the linear combination with maximum variance, which maximizes \( \text{Var}(Y_1) = a'_1 \Sigma a_1 \). It is clear that \( \text{Var}(Y_1) = a'_1 \Sigma a_1 \) can be increased by multiplying and \( a_1 \) by some constant. To eliminate this indeterminacy, it is convenient to restrict attention to coefficient vector of unit length. Therefore, principal components define as below

\begin{align*}
    \text{First PC} &= \text{linear combination } a'_1 X \text{ that maximizes } \\
    &\quad \text{Var}(a'_1 X) \text{ subject to } a'_1 a_1 = 1 \quad \text{(Equation 1.8)} \\
    \text{Second PC} &= \text{linear combination } a'_2 X \text{ that maximizes } \\
    &\quad \text{Var}(a'_2 X) \text{ subject to } a'_2 a_2 = 1 \text{ and } \\
    &\quad \text{Cov}(a'_1 X, a'_2 X) = 0 \quad \text{(Equation 1.9)} \\
    \text{ith PC} &= \text{linear combination } a'_i X \text{ that maximizes } \\
    &\quad \text{Var}(a'_i X) \text{ subject to } a'_i a_i = 1 \text{ and } \\
    &\quad \text{Cov}(a'_i X, a'_k X) = 0 \text{ for } k < i \quad \text{(Equation 1.10)}
\end{align*}

In general, PCA differentiates the spectral data into two groups, positive and negative CWD skin samples. The correlated features are transformed to the uncorrelated variables, the principal components, to reduce the dimensionality. The first PC explains maximum variability in the data, while each successive PC explains the remaining variability.

1.4.3 Classification

Classification is a supervised learning process of predicting the class of given data points. Classification predictive modeling through learning the mapping function \( f \) from the input variables \( x \) to the output variables \( Y = f(x) \). There is a lot of classification algorithms available now, including linear and logistic regression, multi-class classification, decision trees, and support
vector machines, but it is not possible to conclude which one is superior to others. In this research, the main classification algorithms are Support Vector Machine and Neural Network.

(a) Support Vector Machine

A support vector machine (SVM) is a supervised machine learning algorithm used for classification, regression, and outlier detection. SVM was developed by Vapnik and his group at AT&T Bell Laboratories. (Boser et al, 1992) The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. Hyperplanes are decision boundaries that help classify the data points constructed in a high or infinite dimensional space. Intuitively, a good classifier is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin), since, in general, the larger the margin, the lower the generalization error of the classifier. The SVM can be used to learn linear, polynomial, radial basis function (RBF), and multi-layer perceptron (MLP) classifiers. Two key elements in the implementation of SVM are the techniques of mathematical formulation and kernel functions. In this study, classifiers are constructed by sklearn.svm.SVC() in Python. Giving training vectors \( x_i \in \mathbb{R}^p, i = 1, \cdots, n \), in two classes, and a vector \( y \in \{1, -1\}^n \), the goal is to find \( \omega \in \mathbb{R}^p \) and \( b \in \mathbb{R}^p \) such that the prediction given by \( \text{sign}(\omega^T \phi(x) + b) \) is correct for most samples. SVC solves the following primal problem:

\[
\min_{\omega,b,\xi} \frac{1}{2} \omega^T \omega + C \sum_{i=1}^{n} \xi_i \quad \text{(Equation 1.11)}
\]

subject to \( y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \cdots, n \)

(Pedregosa et al, 2011)

(b) Neural Network

A neural network (NN) is a series of algorithms appropriate for complicated classification and pattern recognition problems in a set of data through a process that mimics the operation of
the human brain. A computational model for neural networks was first created based on mathematics and algorithm in 1943 by Warren McCullough and Walter Pitts. (McCullough and Pitts, 1943) NN models are implemented by “neurons,” which may interconnect with others forming complex processing networks. A neuron contains a mathematical function, termed as activation function, to provide output by applying the function on the inputs provided. There have been 5 primary activation functions, step, sigmoid, tanh, ReLU, and leaky ReLU. These neurons may be trained, adjusting the interconnection branch loads (synapse weights, $W_{ij}$) through training algorithms. The feedforward NN was utilized for the data classification in this research, shown in Fig 1.5. This class of network is composed of one input layer that receives the vector to be classified, one output layer that presents the classification results and a set of intermediate layers named hidden layers. In this model, the number of hidden layers is 8, and the number of neurons in each intermediate layer decreases to half of the previous layer.

![Figure 1.7. Demonstration of a feedforward neural network in this research](image)
1.5 Research Objectives

The effectiveness of detecting Chronic Wasting Disease by Raman spectroscopy is evaluated and discussed in this research, and there are two objectives:

1) To establish classifiers on the denoised Raman spectral data and certain principal components of the denoised Raman spectral data, and to compare the accuracy of two classification algorithms, SVM and NN.

2) To investigate the effectiveness of detection on the portable Raman spectrometer.

1.6 Thesis Overview

This dissertation comprises two major parts based on two objectives: In part one (Chapter 2), the major peaks in spectra were assigned, and the feasibility of differentiating tissues from healthy deer and ailing deer was studied. In part two (Chapter 3), it was investigated if spectra collecting by portable Raman spectrometer would generate a better classification accuracy by minimizing intra-tissue variations. In Chapter 4, summary and conclusion were presented for the entire research work, and future perspective was offered.

References


CHAPTER 2. DETECTION AND CHARACTERIZATION OF DEER SKIN TISSUE FOR CHRONIC WASTING DISEASE WITH RAMAN MICROSPECTROSCOPY AND MACHINE LEARNING

Binbin Zhu¹, Mary H. W. Greenlee², Chenxu Yu¹,*
¹ Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA 50011
² Department of Biomedical Sciences, College of Veterinary Medicine, Iowa State University, Ames, IA 50011
* Corresponding author

Modified from a manuscript to be submitted in Journal of Biomedical Optics

2.0 Abstract

Chronic wasting disease (CWD) is a contagious neurological disease in cervids that belongs to transmissible spongiform encephalopathies (TSEs). Since CWD was founded, it heavily threatened the healthy growth of wild deer and resulted in adverse population-level impacts. Meanwhile, the public and agency are concerned about if CWD is able to infect human beings like bovine spongiform encephalopathy (BSE). The pathogenic mechanism of CWD is the prion protein (PrP⁰⁰) infects the host animal by promoting regular cellular protein convert to the abnormal form. Once the CWD infectious agent inserts into an animal, symptoms include emaciation, abnormal behavior, loss of bodily functions, and finally death. This process may take as long as two years before the animal begins to show visible signs of the disease. Currently, two diagnostic tests approved for official CWD postmortem testing are immunohistochemistry (IHC) and ELISA. Therefore, an effective diagnostic method for living animals is crucial to supervise the physical condition of deer. In this study, Raman spectroscopy was utilized to detect the skin samples from healthy and sick white-tail deer. These spectra were analyzed by the machine learning algorithm to differentiate.
2.1 Introduction

Since the first diagnosis in 1967, Chronic Wasting Disease (CWD) cases have emerged in wild or captive cervids in 25 states of the United States, 3 Canadian Provinces, Norway, and South Korea. (USGS, 2020, Norwegian Veterinary Institute, 2016, Kim et al, 2005) It is a contagious neurological disease in cervids that belongs to transmissible spongiform encephalopathies (TSEs), which is believed to be caused by an abnormal form of the prion protein. The prion “infects” the host animal by promoting regular protein to fold into an irregular form. These CWD agents could cause emaciation, abnormal behavior, loss of bodily functions, and finally causing death in the infected animals. As of today, CWD is incurable and poses a fatal threat to cervid health. (Bian et al, 2019)

CWD has a long incubation period, which usually lasts 18 months up to 3 years. After animals are infected with CWD, there are usually no obvious symptoms until the late stage of the disease. The first visible sign in this progress is emaciation. Then, animals lose bodily functions and show abnormal behavior, such as lack of control in bowel movements, listlessness, depression, aggression, lack of interaction with other animals, lowering of the head, and more. However, the only way to diagnose CWD definitively is through laboratory biopsy tests. On microscopic examination, lesions of CWD in the central nervous system are similar to those of other spongiform encephalopathies. PrPSc can be detected by immunohistochemistry (IHC), Western blotting, enzyme-linked immunosorbent assay (ELISA), prion misfolding cyclic amplification (PMCA), and real-time quaking induced conversion (RT-QuIC). (Gavier-Widen et al, 2005) According to USDA description, immunohistochemistry (IHC) testing of the obex area of the brain stem or the medial retropharyngeal lymph nodes is applied for definitive diagnosis. Apparently, such diagnosis could only be carried out postmortem, which is not suitable to meet needs for preventative
measures being implemented to curtail the spreading of the infection, for which diagnosis in live animals is needed.

However, CWD diagnostic assays for live animals are limited and inaccurate. The rectal mucosa biopsy sampling test was used to detect the CWD in white-tailed deer. The diagnostic sensitivity ranged from 63% to 100%, dependent on genotype at prion protein gene (Prnp) codon 96 and the stage of disease as assessed by obex grade. The diagnostic sensitivity reached 100% in the last 2 stages of the preclinical disease but was only 36% for the earliest stage. (Thomsen et al, 2012) Thus, USDA did not approve the rectal biopsy test for routine regulatory testing. A method to quickly screen for possible CWD cases with reasonable accuracy hence is greatly needed.

Raman spectroscopy offers a potential approach to detect and diagnose CWD rapidly in real time as a first screen onsite. In Raman spectroscopy, chemical compositional makeup of biological samples can be revealed, which can lead to distinction between normal (healthy) tissues and diseased tissues (Yu et al., 2006, Wang et al., 2011, 2013). Diagnosis of viral infectious diseases, such as dengue and HPV, has been demonstrated with Raman spectroscopy (Ditta et al., 2018);

With the Raman spectral data, machine learning algorithms could be utilized to extract meaningful information to differentiate the spectroscopic features that underline the signatures associated with the diseases effectively, even with low signal-to-noise Raman spectral data. Raman spectroscopy armed with machine learning was used to obtain highly accurate differentiation (89%) between methicillin-resistant and -susceptible isolates of Staphylococcus aureus (MRSA and MSSA). (Ho et al, 2019). In this study, two machine learning algorithms, Support Vector Machine (SVM) and Artificial Neural Network (ANN), are applied to evaluate the effectiveness of Raman microspectroscopy as a means to differentiate skin biopsies from white-tailed deer infected with
CWD. CWD is a degenerative neural disease; traditional diagnostic methods were mostly focused on biopsy of internal organs/tissues, which rendered them not suitable for application in live animals. Prions can accumulate in fibroblasts in the skin tissue of infected animals. (Thomzig et al, 2007) Hence, a spectroscopic diagnostic method for CWD based on skin biopsy potentially could lead to a field-deployable fast screen to be developed, which would significantly improve the diagnosis and subsequent control of CWD in cervids.

2.2 Materials and Methods

Skin biopsy samples in this study were taken from white-tail deer during a herd depopulation. CWD disease was found in the herd, and so all of the deer were culled, and skin samples were collected. There were 11 positive and 13 negative skin samples collected from tail heads of deer. Each sample was put into 10% neutral buffered formalin and held in formalin for several months (about six months) until they were embedded in paraffin. After embedding, they were mounted on gold slides. These paraffined skin samples (PSS) were prepared in Dr. Greenlee’s lab at the college of veterinary medicine at Iowa State University. They were then transported in a sealed container to our lab for Raman microspectroscopic investigation.

Raman spectra were collected from each of the samples before/after de-paraffin. Samples were deparaffined to reduce the background interference from the paraffin. The procedure was as follows:

i. 10 min incubation in Xylene  
ii. 10 min incubation in 100% ethanol  
iii. 10 min incubation in 95% ethanol  
iv. 10 min incubation in 70% ethanol  
v. 10 min incubation in DI water
After deparaffinization, these de-paraffined skin samples (DSS) were analyzed by a dispersive Raman spectrometer (DXR Raman microscope) using 532 nm laser with 10 mW laser power. Aperture is 25 µm pinholes. 20 spectra were acquired from each sample. Each spectrum was obtained with 20 seconds of exposure time and 2 replicates. The data preprocessing was then conducted, including smoothing, polynomial baseline correction, and l2 normalization. Smoothing was performed with a 10-point moving average; baseline correction was implemented based on an iterative polynomial fitting (Hu et al, 2018); and l2 normalization was conducted based on the square root of the sum of all squared values using the function in the open source “scikit-learn” machine learning library in Python.

After preprocessing, principle component analysis (PCA) was conducted to reduce the dimensionality of data, increase interpretation, and minimize information loss. After the PCA transformation of the spectral data, classification modeling was constructed to differentiate disease positive vs. negative samples. The PCA transformed data were randomly assigned to training and testing groups to evaluate the classification accuracy. The variations between tissue block samples were neglected. The testing group contains 70 negative samples and 60 positive samples, which are randomly picked throughout all data, to evaluate the machine learning model fitted by training groups.

Two machine learning algorithms for classification modeling were evaluated for their effectiveness in this study, SVM and ANN. SVM is a supervised machine learning algorithm used for classification, regression, and outlier detection first developed by Vapnik and his group at AT&T Bell Laboratories (Boser et al, 1992). In SVM, hyperplanes are calculated in an N-dimensional space to serve as decision boundaries that distinctly classify data points in question. Intuitively, a good classifier will be achieved by the hyperplane that has the largest distance to the
nearest training data points of any class (so-called functional margin), since in general, the larger
the margin, the lower the generalization error of the classifier. SVM can be used to learn linear,
polynomial, radial basis function (RBF), and multi-layer perceptron (MLP) classifiers. Two
critical elements in the implementation of SVM are the techniques of mathematical formulation
and kernel functions. In this study, classifiers were constructed by sklearn.svm.SVC() in Python.
Giving training vectors $x_i \in \mathbb{R}^p, i = 1, \cdots, n$, in two classes, and a vector $y \in \{1, -1\}^n$, the goal
was to find $\omega \in \mathbb{R}^p$ and $b \in \mathbb{R}^p$ such that the prediction given by $\text{sign}(\omega^T \phi(x) + b)$ was correct
for most samples. SVC solves the following primal problem (Pedregosa et al, 2011):

$$\min_{\omega, b, \zeta} \frac{1}{2} \omega^T \omega + C \sum_{i=1}^n \zeta_i$$

subject to $y_i(\omega^T \phi(x_i) + b) \geq 1 - \zeta_i$, $\zeta_i \geq 0$, $i = 1, \cdots, n$

ANN is a series of algorithms appropriate for complex classification and pattern recognition
problems in a set of data through a process that mimics the operation of the human brain. The first
ANN model was created in 1943 by Warren McCullough and Walter Pitts (McCullough and Pitts,
1943). ANN models are implemented by “neurons”, which may interconnect with others forming
complex processing networks. A neuron contains a mathematical function, termed as activation
function, to provide output by applying the function on the inputs provided. There have been 5
major activation functions, step, sigmoid, tanh, ReLU, and leaky ReLU. These neurons may be
trained by adjusting the interconnection branch loads (synapse weights, $W_{ij}$) through training
algorithms. The feedforward ANN was utilized for the data classification in this research, shown
in Fig 2.1. This class of network is composed of one input layer that receives the vector to be
classified, one output layer that presents the classification results and a set of intermediate layers
named hidden layers. In this study, the number of hidden layers was 8, and the number of neurons
in each intermediate layer decreases to half of the previous layer.
2.3 Results and Discussion

2.3.1 Analysis of Raman spectra collected from the skin samples

Spectra of negative and positive PSS samples prior to de-paraffin exhibited five obvious peaks at 1067 cm\(^{-1}\), 1136 cm\(^{-1}\), 1300 cm\(^{-1}\), 1445 cm\(^{-1}\), and 1464 cm\(^{-1}\). However, spectrum of paraffin wax contains \(\nu\) (CC) band around 1107 cm\(^{-1}\) and \(\nu\) (CO) features around 1060 cm\(^{-1}\); CH\(_2\) deformation contributes peaks at 1445 cm\(^{-1}\) and 1464 cm\(^{-1}\) (Barry et al. 1992). Thus, spectra of negative and positive PSS showed a strong background signal of the paraffin wax. It was hard to define peaks contributed by cellular components, including PrP\(^{Sc}\) prions from these spectral data.

Figure 2.2 indicates that positive PSS cannot be differentiated from negative PSS by simple visual inspection of peaks in spectra. In order to highlight significant differences between the two sample groups, Welch’s t-test using a p-value less than 0.01 as a significance level was utilized to identify peaks that showed significant differences. Only the peak at \(~1464\) cm\(^{-1}\) had a significant
p-value of 0.00501, which, as mentioned, could be due to the paraffin wax and did not reflect any physiological differences between the two groups.

Figure 2.2. Spectra of negative PSS and positive PSS with p-value of t-test

Principal component analysis (PCA) was applied for dimensionality reduction and features extraction for these two sets of data (prior to de-paraffin). It is essential to estimate how many principal components (PCs) are needed to describe the spectral data. This can be determined by looking at the cumulative explained variance ratio as a function of the number of components. Figure 2.3 indicates that the cumulative explained variance of 10 PCs, 50 PCs, and 200 PCs is 67.00%, 83.45%, and 98.87% respectively. As shown in Table 2.1, within the first 5 PCs, PC0 and PC1 were responsible for 27.93% and 13.73% of the total variance, respectively. As shown in fig.2.4, a simple 2-D projection plot with PC0 and PC1 does not offer a clear separation between the P/N datasets.
The differentiation accuracies between P(+)/N(-) samples with SVM and ANN discriminant models fitted with the original spectral data were compared with that of models reconstructed from these PCA transformed datasets. The PC scores were defined as new intermediate quality variables and classified by Support Vector Machine (SVM) and Neural Network (NN). Classifiers were constructed by 10 PCs, 50 PCs, 200 PCs, and raw spectral data from the training sets, separately. After being fitted, models were used to predict new data entries from the test sets values in order to evaluate performances on prediction.

Table 2.1. Explained variance and cumulative variance of five PC scores

<table>
<thead>
<tr>
<th>% Explained variance</th>
<th>PC0</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Cumulative variance</td>
<td>27.93</td>
<td>13.73</td>
<td>8.92</td>
<td>4.51</td>
<td>3.30</td>
</tr>
<tr>
<td>% Explained variance</td>
<td>27.93</td>
<td>41.66</td>
<td>50.58</td>
<td>55.10</td>
<td>58.40</td>
</tr>
</tbody>
</table>

Figure 2.3. The cumulative variance of 10 PCs, 50 PCs, and 200 PCs for PSS
In SVM, a grid search cross-validation was used to determine the best training model by adjusting kernel functions and parameters in kernel functions. Table 2 listed the best training model corresponding to each data set. For visualization, two-dimensional decision boundaries of the best training model were plotted in Figure 2.5. The classifier built by PSS original spectral data can predict 0.585 of test data correctly, which is the lowest. The classifier with the best prediction accuracy was constructed with 50 principal components, at 0.685, as shown in Table 2.2.

In ANN, the prediction accuracy improved in all four models. The classifier with the lowest accuracy was the 200 PCs’ model. The 50 PC classifier remained to be the best performer, with the prediction accuracy increased to 0.723 comparing to SVM, as shown in Table 2.3.
Table 2.2. Support vector machine results for PSS

<table>
<thead>
<tr>
<th>SVM Model kernel</th>
<th>10 PCs</th>
<th>50 PCs</th>
<th>200 PCs</th>
<th>Spectral Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Model parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly</td>
<td>rbf</td>
<td>rbf</td>
<td>rbf</td>
<td></td>
</tr>
<tr>
<td>degree=5, coef0=3</td>
<td>C=100, gamma=1</td>
<td>C=100, gamma=0.1</td>
<td>C=100, gamma=0.1</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.662</td>
<td>0.685</td>
<td>0.600</td>
<td>0.585</td>
</tr>
</tbody>
</table>

Table 2.3. Neural Network results for PSS

<table>
<thead>
<tr>
<th></th>
<th>10 PCs</th>
<th>50 PCs</th>
<th>200 PCs</th>
<th>Spectral Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.700</td>
<td>0.723</td>
<td>0.685</td>
<td>0.692</td>
</tr>
</tbody>
</table>

2.3.2 Analysis of Raman spectra from de-paraffined skin samples

Since paraffin wax seemed to dominate the Raman spectra collected from the tissue section samples, a de-paraffin procedure was used to remove the paraffin. Figure 2.6 plotted spectra of negative and positive deparaffined skin samples (DSS), which were significantly different from the spectra of paraffined skin samples (PSS). The welch’s t-test was again utilized to analyze the peaks showing significant differences between the P/N groups. Five peaks were identified (Figure 2.6), which were 863 cm$^{-1}$, 944 cm$^{-1}$, 1252 cm$^{-1}$, 1457 cm$^{-1}$, and 1673 cm$^{-1}$. None of these peaks belongs to paraffin. The amide I band with β-turn structure was identified by the characteristically around 1673 cm$^{-1}$, and a higher intense band at 1252 cm$^{-1}$ in the amide III region was observed in spectra which are related to its β-sheet secondary structure characteristics. (Malvern Panalytical, 2019) The band of CH$_2$ and CH$_3$ bending in proteins and lipids was observed around 1457 cm$^{-1}$. The peak at 944 cm$^{-1}$ assigned the C-C skeletal mode of α-helix structure. (Tan et al, 2017) The bands associated with tyrosine were observed around 863 cm$^{-1}$. Increases of these peaks on positive sample, as listed in table 2.4, appeared to be representing the structure of PrP$\text{Sc}$ prions in the skin samples. (Nobel et al, 2015, Lednev, I., Shashilov, V., & Xu, 2009)
Figure 2.6. Spectra of negative DSS and positive DSS with p-value of t-test

Table 2.4. Values of raman shift, intensity, p-value, and assignment of peaks shown in Figure 2.6

<table>
<thead>
<tr>
<th></th>
<th>Raman Shift (cm⁻¹)</th>
<th>Negative Intensity</th>
<th>Positive Intensity</th>
<th>P-value</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>862.929</td>
<td>0.000743</td>
<td>0.000809</td>
<td>0.000283</td>
<td>Tyrosine</td>
</tr>
<tr>
<td>2</td>
<td>943.924</td>
<td>0.000595</td>
<td>0.000672</td>
<td>0.000655</td>
<td>C-C α-helix in proteins</td>
</tr>
<tr>
<td>3</td>
<td>1251.515</td>
<td>0.00135</td>
<td>0.00141</td>
<td>0.000782</td>
<td>Amide III (b-sheet)</td>
</tr>
<tr>
<td>4</td>
<td>1456.896</td>
<td>0.00196</td>
<td>0.00199</td>
<td>0.000874</td>
<td>CH₂, CH₃ bending in proteins and lipids</td>
</tr>
<tr>
<td>5</td>
<td>1672.885</td>
<td>0.00125</td>
<td>0.00136</td>
<td>0.000581</td>
<td>Amide I (b-turn)</td>
</tr>
</tbody>
</table>
Table 2.5. Raman shifts of peaks and the characteristic assignments

<table>
<thead>
<tr>
<th>Raman Shift (cm$^{-1}$)</th>
<th>Assignment</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1670-1680</td>
<td>Admine I(β-Turn/random β-space)</td>
<td>Malvern Panalytical, 2019</td>
</tr>
<tr>
<td>1660-1670</td>
<td>Admine I (β-sheet)</td>
<td>Malvern Panalytical, 2019</td>
</tr>
<tr>
<td>1650-1655</td>
<td>Admine I(α-helix)</td>
<td>Malvern Panalytical, 2019</td>
</tr>
<tr>
<td>1602-1607</td>
<td>C = C band in Phenylalanine or Tyrosine</td>
<td>Tan et al, 2017</td>
</tr>
<tr>
<td>1541-1542</td>
<td>C-N stretching, Amide II</td>
<td>Tan et al, 2017</td>
</tr>
<tr>
<td>1445</td>
<td>CH2, CH3 bending in proteins and lipids</td>
<td>Tan et al, 2017</td>
</tr>
<tr>
<td>1371</td>
<td>Guanine in DNA, Tryptophan in proteins</td>
<td>Tan et al, 2017</td>
</tr>
<tr>
<td>1330-1340</td>
<td>Admine III (α-helix)</td>
<td>Malvern Panalytical, 2019</td>
</tr>
<tr>
<td>1328</td>
<td>CH vibration in DNA/RNA, CH2 twisting in lipids</td>
<td>Tan et al, 2017</td>
</tr>
<tr>
<td>1263</td>
<td>CH bending in lipids</td>
<td>Tan et al, 2017</td>
</tr>
<tr>
<td>1235-1250</td>
<td>Admine III(β-sheet)</td>
<td>Malvern Panalytical, 2019</td>
</tr>
<tr>
<td>933</td>
<td>C-C stretching mode, C-C αhelix in proteins</td>
<td>Tan et al, 2017</td>
</tr>
<tr>
<td>857</td>
<td>Tyrosine</td>
<td>Tan et al, 2017</td>
</tr>
<tr>
<td>745</td>
<td>Thymine in DNA</td>
<td>Tan et al, 2017</td>
</tr>
<tr>
<td>726</td>
<td>Hypoxanthine</td>
<td>Tan et al, 2017</td>
</tr>
</tbody>
</table>

Figure 2.7 shows the result of PCA on the DSS datasets. Ten principal components were responsible for less than 50% of the total variance, less than that of 10 PCs for PSS; and 50 components only retained about 70% of the variance. 200 components accounted for almost 100% of the variance. Discrimination between negative and positive still cannot be recognized easily with 2-D projection plot, as shown in Figure 2.8. Classifiers were constructed by 10 PCs, 50 PCs, 200 PCs, and original DSS spectral separately. The best classifier of SVM from grid search for each training data set is listed in Table 2.5. Comparing to the PSS classifiers, the prediction
accuracy increased by 5% to 10%. The classifier constructed by 50 principal components still yielded the best prediction accuracy, which reached 0.731.

In ANN, the prediction accuracy improved in all four models for the DSS data (Table 2.6). The classifier with the lowest accuracy was the 200 PCs’ model. 10 PC classifier and 50 PC classifier both yielded 0.762 accuracy, which was better than that of SVM models, respectively.

Figure 2.7. The cumulative variance of 10 PCs, 50 PCs, and 200 PCs for DSS
Figure 2.8. Clustering of PC0 and PC1 for negative and positive DSS

Table 2.6. Support vector machine results for DSS

<table>
<thead>
<tr>
<th>SVM Model kernel</th>
<th>10 PCs</th>
<th>50 PCs</th>
<th>200 PCs</th>
<th>Spectral Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Model parameters</td>
<td>degree=5, coef0=2</td>
<td>C=10, gamma=1</td>
<td>C=100, gamma=0.1</td>
<td>C=100, gamma=0.1</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.700</td>
<td>0.731</td>
<td>0.692</td>
<td>0.685</td>
</tr>
</tbody>
</table>

Table 2.7. Artificial Neural Network results for DSS

<table>
<thead>
<tr>
<th></th>
<th>10 PCs</th>
<th>50 PCs</th>
<th>200 PCs</th>
<th>Spectral Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.762</td>
<td>0.762</td>
<td>0.723</td>
<td>0.746</td>
</tr>
</tbody>
</table>

As the spectroscopic method was developed to eventually interrogate skins of live animals, the de-paraffined samples were better representation of what to be expected for live animal as no paraffin would be present on a live animal. The 76.2% classification accuracy hence was an indicator for potential diagnostic applications. It should be noted that microspectroscopic investigation was more susceptible to localized chemical variations within the skin biopsy samples, as the spectral acquisition was from a small area (a spot of 10 μm ×10 μm), which could further reduce the classification accuracy. It was expected that better classification accuracy may be
possible with larger spectral acquisition areas in fiber-optic based portable Raman systems, which would be more suitable for onsite deployment for in-field CWD screening.

2.4 Conclusion

Raman microspectroscopy in conjunction with PCA and machine learning data analysis was demonstrated to be a feasible way to differentiate CWD-positive deer skin tissues from CWD-negative ones. With de-paraffined skin tissues, which should offer a good mimic to in vivo situations, the ANN machine learning algorithm worked better than that of SVM algorithm, a 76.2% classification accuracy was obtained. The Raman microspectroscopic method used in this study offers microscopic spatial resolution for spectral acquisition, which renders it more susceptible to intra-tissue variations due to spot-to-spot differences. It is expected a larger scale spectral acquisition technique could generate better classification accuracy by minimizing intra-tissue variations. As a conclusion, the Raman spectroscopic technique was shown to have the potential of offering a useful tool for fast screening CWD via skin interrogation.

References


CHAPTER 3. DETECTION AND CHARACTERIZATION OF DEER SKIN TISSUE FOR CHRONIC WASTING DISEASE WITH PORTABLE RAMAN SPECTROMETER AND MACHINE LEARNING

Binbin Zhu¹, Mary H. W. Greenlee², Chenxu Yu¹,*

¹ Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA 50011
² Department of Biomedical Sciences, College of Veterinary Medicine, Iowa State University, Ames, IA 50011
* Corresponding author

Modified from a manuscript to be submitted in Journal of Biomedical Optics

3.0 Abstract

Chronic wasting disease (CWD) is a contagious neurological disease in cervids that belongs to transmissible spongiform encephalopathies (TSEs). Its spread has threatened the healthy growth of wild and farm-raised deer and resulted in adverse population-level impacts. It also raised concerns over the possibility of infecting human beings like bovine spongiform encephalopathy (BSE). CWD is a prion disease that may take as long as two years for visible signs of the disease to appear. Currently, diagnostic tests approved for official CWD are postmortem tests (immunohistochemistry (IHC) and ELISA), which are not suitable for in vivo diagnosis. In this study, Raman spectroscopy in conjunction with machine learning was utilized to differentiate CWD-positive skin tissues of white-tail deer from that of CWD-negative ones using a portable Raman spectrometer. Two machine learning algorithms were investigated, with artificial neural network (ANN) generating the best classification accuracy at 94.4%. The results of this study suggested that Raman spectroscopic screening has the potential to offer a fast yet quite accurate screening of CWD in live animals for onsite diagnosis of the disease.
3.1 Introduction

Since the first diagnosis in 1967, Chronic Wasting Disease (CWD) cases have emerged in wild or captive cervids in 26 states of the United States, 3 Canadian Provinces, Norway, and South Korea. (USGS, 2020, Norwegian Veterinary Institute, 2016, Kim et al, 2005) It is a contagious neurological disease in cervids that belongs to transmissible spongiform encephalopathies (TSEs), which is believed to be caused by an abnormal form of the prion protein. The prion “infects” the host animal by promoting regular protein to fold into the abnormal form. These CWD agents could cause emaciation, abnormal behavior, loss of bodily functions, and finally death in the infected animals. As of today, CWD is incurable and poses a fatal threat to cervid health (Bian et al, 2019).

CWD has a long incubation period, which could last 18 months to 3 years. After animals are infected with CWD, there are usually no obvious symptoms until the late stage of the disease. The first visible sign in this progress is emaciation. Then, animals lose bodily functions and show abnormal behavior, such as lack of control in bowel movements, listlessness, depression, aggression, lack of interaction with other animals, lowering of the head, and more. However, the only way to diagnose CWD definitively is through laboratory biopsy tests. On microscopic examination, lesions of CWD in the central nervous system are similar to those of other spongiform encephalopathies. PrP^Sc can be detected by immunohistochemistry (IHC), Western blotting, enzyme-linked immunosorbent assay (ELISA), prion misfolding cyclic amplification (PMCA), and real-time quaking induced conversion (RT-QuIC). (Gavier-Widen et al, 2005) According to USDA description, immunohistochemistry (IHC) testing of the obex area of the brain stem or the medical retropharyngeal lymph nodes can be applied for definitive diagnosis. Apparently, such diagnosis could only be carried out postmortem, which is not suitable to meet needs for
preventative measures being implemented to curtail the spreading of the infection, for which diagnosis for live animals is needed.

However, CWD diagnostic assays for live animals are limited and inaccurate. The rectal mucosa biopsy sampling test was used to detect the CWD in white-tailed deer. The diagnostic sensitivity ranged from 63% to 100%, dependent on genotype at prion protein gene (Prnp) codon 96 and the stage of disease as assessed by obex grade. The diagnostic sensitivity reached 100% in the last 2 stages of the preclinical disease but was only 36% for the earliest stage. (Thomsen et al, 2012) Thus, USDA did not approve the rectal biopsy test for routine regulatory testing. A method to quickly screen for possible CWD cases with reasonable accuracy hence is greatly needed.

CWD is a neural degenerative disease, traditional diagnostic methods were mostly focused on biopsy of internal organs/tissues, which rendered them not suitable for application in live animals. Prions can accumulate in fibroblasts in skin tissue of infected animals (Thomzig et al, 2007). Hence, a spectroscopic diagnostic method for CWD based on skin biopsy potentially could lead to a field-deployable fast screen to be developed, which would greatly improve the diagnosis and subsequent control of CWD in cervids.

Raman spectroscopy offers a potential approach to detect and diagnose CWD rapidly in real time as a first screen onsite. In Raman spectroscopy, chemical compositional makeup of biological samples can be revealed, which can lead to distinction between normal (healthy) tissues and diseased tissues (Yu et al., 2006, Wang et al., 2011, 2013). Diagnosis of viral infectious diseases, such as dengue and HPV, has been demonstrated with Raman spectroscopy (Ditta et al., 2018). With the Raman spectral data, machine learning algorithms could be utilized to extract meaningful information to differentiate the spectroscopic features that underline the signatures associated with the diseases effectively, even with low signal-to-noise Raman spectral data. Raman
spectroscopy armed with machine learning was used to obtain highly accurate differentiation (89%) between methicillin-resistant and -susceptible isolates of Staphylococcus aureus (MRSA and MSSA) (Ho et al, 2019).

In this study, two machine learning algorithms, Support Vector Machine (SVM) and Artificial Neural Network (ANN), were applied to evaluate the effectiveness of Raman spectroscopy as a means to differentiate skin biopsies from white-tailed deer infected with CWD. A portable Raman spectrometer equipped with a fiber optic probe was utilized for spectral acquisition, as it could be readily applied for onsite screening. With earlier results of spectral data measured by Raman micro-spectrometer, skin sample infected CWD can be differentiated from the skin sample of healthy animals with ~76% accuracy. In this study, the portable Raman spectrometer was shown to yield even better classification accuracy, which renders it a good candidate for on-site detection in practice.

3.2 Materials and Methods

The same skin biopsy samples of white-tail deer in Chapter 2’s study were used in this study. Comparing to samples prepared for micro-spectrometer, biopsy specimens were maintained in the Paraffin-embedded tissue blocks, which usually were referred to as formalin fixed paraffin embedded (FFPE) tissue specimens. There were 6 tissue blocks, 3 healthy (CWD-negative) and 3 infectious tissue (CWD-positive), all were prepared in Dr. Greenlee’s lab at the college of veterinary medicine at Iowa State University. They were then transported in a sealed container to Dr. Yu’s lab for portable Raman spectroscopic investigation.

All tissue blocks were analyzed by a Portable Raman spectrometer (i-Raman Plus) using laser excitation of 785 nm with 340 mW nominal power at exiting probe and 445 mW nominal power at laser port. The range of Raman shift is between 65 and 2800 cm⁻¹. 15 spectra were
acquired from each tissue block. Each spectrum was obtained with 1000 milliseconds integration time and took an average of two measurements. The data preprocessing was then conducted with the same procedure in chapter 2, including smoothing, polynomial baseline correction, and l2 normalization. Smoothing was performed with a 5-point moving average to avoid reduction of small peaks; baseline correction was implemented based on an iterative polynomial fitting (Hu et al, 2018); and l2 normalization was conducted based on the square root of the sum of all squared values using the function in the open source “scikit-learn” machine learning library in Python.

After preprocessing, principle component analysis (PCA) was conducted to reduce the dimensionality of data, increase interpretation, and minimize information loss. After PCA transformation of the spectral data, classification modeling was constructed to differentiate disease positive vs. negative samples. The PCA transformed data were randomly assigned into training and testing groups to evaluate the classification accuracy. The variations between tissue block samples were neglected. The classification model was constructed by the train data set contained 36 randomly selected spectra from negative and positive, and left 9 spectra from each group were used for testing the classifiers.

Two machine learning algorithms for classification modeling were evaluated for their effectiveness in this study, SVM and ANN. SVM is a supervised machine learning algorithm used for classification, regression, and outlier detection first developed by Vapnik and his group at AT&T Bell Laboratories (Boser et al, 1992). In SVM, hyperplanes are calculated in an N-dimensional space to serve as decision boundaries that distinctly classify data points in question. Intuitively, a good classifier will be achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general, the larger the margin, the lower the generalization error of the classifier. SVM can be used to learn linear,
polynomial, radial basis function (RBF), and multi-layer perceptron (MLP) classifiers. Two key elements in the implementation of SVM are the techniques of mathematical formulation and kernel functions. In this study, classifiers were constructed by sklearn.svm.SVC() in Python. Giving training vectors $x_i \in \mathbb{R}^p, i = 1, \cdots, n$, in two classes, and a vector $y \in \{1, -1\}^n$, the goal was to find $\omega \in \mathbb{R}^p$ and $b \in \mathbb{R}^p$ such that the prediction given by $\text{sign} (\omega^T \phi(x) + b)$ was correct for most samples. SVC solves the following primal problem (Pedregosa et al, 2011):

$$\min_{\omega, b, \xi} \frac{1}{2} \omega^T \omega + C \sum_{i=1}^n \xi_i$$

subject to $y_i (\omega^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \cdots, n$

ANN is a series of algorithms appropriate for complex classification and pattern recognition problems in a set of data through a process that mimics the operation of the human brain. The first ANN model was created in 1943 by Warren McCullough and Walter Pitts (McCullough and Pitts, 1943). ANN models are implemented by “neurons”, which may interconnect with others forming complex processing networks. A neuron contains a mathematical function, termed as activation function, to provide an output by applying the function on the inputs provided. There have been 5 major activation functions, step, sigmoid, tanh, ReLU, and leaky ReLU. These neurons may be trained adjusting the interconnection branch loads (synapse weights, $w_{ii}$) through training algorithms. The feedforward ANN was utilized for the data classification in this research, shown in Fig 3.1. This class of network is composed of one input layer that receives the vector to be classified, one output layer that presents the classification results and a set of intermediate layers named hidden layers. In this study, the number of hidden layers was 8, and the number of neurons in each intermediate layer decreases to half of the previous layer.
Figure 3.1. Demonstration of a feedforward neural network in this research. N is the input dimensional number, which equals 10, 50 or 998 for different types training data.

3.3 Results and Discussion

Since the tissue blocks were deeply embedded in paraffin, a de-paraffin process as described in chapter 2 could not be effectively used for them. However, by avoiding the paraffin layers, spectra could be obtained from tissue areas which appeared not to be affected by the paraffin wax greatly. Majority peaks of the spectrum of paraffin wax were different from the spectrum of the tissue, as shown in Figure 3.2 (a). Spectra of negative and positive samples exhibited five strong peaks at 1004 cm\(^{-1}\), 1066 cm\(^{-1}\), 1135 cm\(^{-1}\), 1303 cm\(^{-1}\), 1443 cm\(^{-1}\). Only the peak at 1303 cm\(^{-1}\) overlaps with the spectrum of paraffine wax, which is attributed to CH\(_2\) deformation. The band at 1004 cm\(^{-1}\) is attributed to the C–C aromatic ring stretching (Phenylalanine) (Kahraman et al, 2009). Apparently, the spectral features of CWD-positive samples still cannot be distinguished from that of CWD-negative samples by simple visual inspection of peaks. In order to highlight major differences between the two sample groups,
Welch’s t-test using p-value less than 0.01 as a significance level was utilized to identify peaks that showed significant differences. And the major differences between negative and positive is around 1235 cm\(^{-1}\) to 1270 cm\(^{-1}\), which is the region of amide III. Figure 3.2(b) shows the loadings of the first 4 PCs for the positive tissue blocks. 1135 cm\(^{-1}\), 1066 cm\(^{-1}\), and 1303 cm\(^{-1}\) peaks mainly contributed to PC0. PC1 explained by four peaks, 1004 cm\(^{-1}\), 1066 cm\(^{-1}\), 1135 cm\(^{-1}\), and 1303 cm\(^{-1}\). The region between 1235 cm\(^{-1}\) and 1270 cm\(^{-1}\) contributed to the PC2 mostly. 1004 cm\(^{-1}\) peak was the main part of PC3.

Figure 3.2 (a) Spectra of paraffin, negative and positive tissue block samples with p-value of t-test; (b) The first 5 PC loadings showing peaks related to the spectra in positive tissue block.
Figure 3.3 shows the result of PCA on the tissue block samples datasets. Ten principal components were responsible for 94.3% of the total variance, significantly larger than that of 10 PCs for PSS and DSS in Chapter 2; and 50 components accounted for almost 100% of the variance. Discrimination between negative and positive still can be partially recognized with 2-D projection plot, as shown in Figure 3.4. Classifiers were constructed by 10 PCs, 50 PCs, and original spectral data of tissue block samples separately.

Figure 3.4 Clustering of PC0 and PC1 for negative and positive tissue block samples
Table 3.1. Support vector machine results for tissue block samples

<table>
<thead>
<tr>
<th>SVM Model kernel</th>
<th>10 PCs</th>
<th>50 PCs</th>
<th>Spectral Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Model parameters</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>degree=6, coef0=1</td>
<td>0.833</td>
<td>0.833</td>
<td>0.722</td>
</tr>
</tbody>
</table>

Table 3.2. Artificial Neural Network results for tissue block samples

<table>
<thead>
<tr>
<th></th>
<th>10 PCs</th>
<th>50 PCs</th>
<th>Spectral Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.944</td>
<td>0.889</td>
<td>0.833</td>
</tr>
</tbody>
</table>

The best classifier of SVM from grid search for each training data set is listed in Table 3.1. Comparing to the DSS classifiers, the prediction accuracy increased by 10% to 15%. Both classifiers constructed by 10 and 50 principal components yielded the best prediction accuracy, which reached 0.833.

The prediction accuracy of ANN algorithm improved in all three models for the tissue block data (Table 3.2). The classifier with the highest accuracy was the 10 PCs’ model, which reached 0.944. 50 PCs’ classifier yielded 0.889 accuracy, and the lowest accuracy conducted by spectral data reached 0.833.

The portable Raman spectrometer represented a better discrimination than the microspectrometer. The laser spot diameter at focal plane of the portable device is 85 μm (comparing to 1-2 μm on the Raman microscope used in chapter 2), which can collect spectral signals from a larger area on the tissue sample and significantly reduce the variations due to local compositional differences. The increase of the classification accuracy was hence obtained. Thus,
the portable Raman spectrometer system would be more suitable for onsite deployment for in-field CWD screening.

3.4 Conclusion

Using Raman spectrometer to diagnose CWD on living animals’ skin is feasible according to the experiment results of this work. Nonetheless, good detection results could be obtained via statistical (i.e., SVM and ANN) discriminant analysis. Further improvements could be made on the current experiment design to increase the sensitivity and stability. Overall, Raman spectroscopy can deliver a rapid detection of CWD in field on living animals, which in the future may offer an effective supervision for cervids health.

References


CHAPTER 4. GENERAL CONCLUSION AND FUTURE PERSPECTIVE

In this research, two Raman spectrometer systems, a Thermo XTR Raman micro spectrometer and a Matrohm i-Raman plus portable Raman spectrometer, were utilized to differentiate CWD-positive skin tissues from CWD-negative samples of white-tail deer and elk. After spectral collection, support vector machine (SVM) and artificial neural network (ANN) discriminant models were generated based on PCs calculated from the model data sets of original spectra to classify the spectra kept in testing sets into positive/negative groups. The results in Chapter 2 has indicated that with the microscopic data from de-paraffined skin tissues, a 76.2% classification accuracy was achieved with the ANN model. In comparison, the discrimination of paraffined tissues using Raman microspectroscopic data was not as good, suggesting for microscopic spectral acquisition, the presence of paraffin would interfere with the signals from the actual biological samples to complicate the differentiation/classification. Nonetheless, the de-paraffined samples were a better representation of the real biological samples. In addition, the microscopic spectra data were more susceptible to intra-tissue variations due to spot-to-spot differences. A larger scale spectral acquisition technique was hence expected to generate better classification accuracy by minimizing intra-tissue variations, which was discussed in Chapter 3. The portable Raman spectrometer smooths some intra-tissue variants due to a bigger laser spot diameter at focal plane. Meanwhile, the effects of paraffin wax were eliminated. By using the portable Raman spectrometer, artificial neural network (ANN) generated the best classification accuracy at 94.4%. These results suggested that Raman spectroscopic screening potentially can offer a fast yet quite accurate screening of CWD in live animals for onsite diagnosis of the disease.

Nonetheless, the results reported in this thesis are not without weakness. The results reported in chapter 3 were obtained from 6 tissue blocks (3 for each group). It was a small sample
size, which limited the statistical robustness of the analysis. More samples need to be analyzed to overcome this weakness. Ideally, fresh deer skin samples should be investigated to be compared to the Paraffin-embedded tissue blocks, which would be a much more practical model to simulate the situation of onsite detection for living cervids. Meanwhile, surface-enhanced Raman spectroscopy (SERS) can be utilized on either animal bodies or tissue blocks to improve the accuracy of diagnosis results. With these further improvements, Raman spectroscopy, armed with machine learning, can deliver a rapid detection of CWD in field on living animals and offer an effective supervision for cervid health.