Hokkaido University

Doctoral Dissertation

Raman Microscopic Histology Using Machine Learning Techniques for Non-Alcoholic Fatty Liver Disease

（非アルコール性脂肪肝疾患における機械学習技術を用いたRaman組織学）

---

by

Khalifa Mohammad Helal

Graduate School of Life Science
Hokkaido University
Japan

December 2019

This thesis is dedicated to
my beloved mother and grandmother,
my little princess Junaina and my lovely wife, Tasnuva Tarannum,
and the loving memory of my father.

Abstract

Histopathology is a standard means to diagnose the disease states of cells or tissues from their morphological features, but it requires the expertise of histopathologists and is, therefore, susceptible to human bias. Raman micro-spectroscopy can provide additional biochemical information that is not available by morphological examination, and has a large potential to assist histological inspection and to benefit diagnosis of disease as objectively as possible in fluorescent label-free manner. Detailed analysis of Raman microscopic data is essential to detect the spectral changes originating from the underlying biochemical changes in cells or tissues due to the progression of disease.

This thesis is concerned with the development of diagnostic tools by integrating Raman microscopic imaging with methods of machine learning and information theory, and the analysis of Raman hyper-spectral images of rat liver tissues comprising a dietary model of non-alcoholic fatty liver disease (NAFLD), with each liver tissue having been histopathologically diagnosed as normal, non-alcoholic fatty liver (NAFL), or non-alcoholic steatohepatitis (NASH).

In the first study, dimension reduction (manifold learning) and ensemble-learning-based random forest classification are performed on the Raman spectra obtained from the regular spatial grid averaging of Raman images for predicting different states (dietary and histological). I identify a set of important Raman bands in differentiating the Raman spectra arising from different states of tissues. Furthermore, I find that NAFL state is distinguished into two pahses, namely, 'slowly progressive NAFL' (NAFL-α) and 'rapidly progressive NAFL' (NAFL-β) in terms of Raman imaging, and main Raman shifts to separate these two NAFL models are identified. This enhances the diagnostic capabilities to distinguish the states of tissues at early stages of the disease.

In the second study, using the dietary model of NAFLD in rats, I apply machine learning and information theory to evaluate cellular-level information in liver tissue samples.

The method first increases the signal-to-noise ratio while maintaining spatial and spectral structures of Raman images as much as possible through extension of the simple linear iterative clustering superpixel algorithm developed in the area of image analysis. Second, using the unsupervised machine learning with rate distortion theory and the Poisson error arising from photon counting, it identifies a set of characteristic spectra having distinct Raman information across the tissues. I discover diverse chemical environments in the liver tissues, allowing for the quantification of representative biochemical components such as vitamin A, lipids, and cholesterols which can be very important insights into the disease states of cells or tissues. Third, armed with microscopic information about the biochemical composition of the liver tissues, I group tissues having similar chemical composition using agglomerative hierarchical clustering, providing a novel "descriptor" enabling us to infer tissue states, contributing valuable information to histological inspection. Excessive lipid deposition with the appearance of cholesterol signatures indicates the severity of the disease state of the NAFLD tissues.

Raman microscopy coupled with the proposed techniques will offer new clinical tools that will aid pathologists in more precise NAFLD diagnosis with molecular information about the liver tissue, and enable us to predict the progression of disease at some early stages where any morphological feature of diseases does not appear yet.

Key words: Raman Hyper-Spectral Imaging, Non-Alcoholic Fatty Liver Disease, Manifold Learning, Machine Learning, Superpixel Segmentation, Rate-Distortion Theory.

Acknowledgements

This thesis would not have been accomplished without the encouragement, assistance, and continuous support of many wonderful people. My heartfelt gratitude to all of them.

First and foremost, I would like to gratefully thank my advisor Professor Tamiki Komatsuzaki, for his excellent guidance, encouragement, patience, motivations, and cooperation during my whole period of study in Hokkaido University. His wonderful mentorship provided me with great opportunity to be introduced into the field of multi-disciplinary research. I am very much thankful for his contributions of incredible ideas and funding to make my PhD research productive and interesting. Thank you, Tamiki sensei, for everything you have done for me.

I belonged to a cheerful group of professor Komatsuzaki, having helpful and motivational attitude which is important for effective research. I would like to thank all the fellow lab members of this group who were very friendly and helpful. This group has contributed immensely to my daily life and my professional time at Hokkaido University.

Specially, I would like to express my sincere gratitude to assistant professor Dr. J. Nicholas Taylor (Nick), for supporting me during past four years with valuable scientific advice, many insightful discussions and suggestions. He introduced me the concepts of Raman microscopic images and analysis of Raman data on which this thesis is based. We worked together on the analysis of Raman data to accomplish the thesis. He was my primary resource to get the answer of my science questions regrading my PhD research, and he has done a major role to develop methodologies to obtain the results related to this thesis. I will forever be thankful to you Nick, for your ideas, time and patient that have been great contributors to accomplish the thesis.

I would like to thank Associate Professor Hiroshi Teramoto for his valuable scientific advice, fruitful discussions and useful guidance. In the first year of my PhD, I worked with him on collective cell motility. He helped me a lot with interesting conversations

and mentoring with great patient.

I sincerely thank my co-advisors Professor Hisashi Haga, Professor Makoto Demura, and Professor Yoshinori Harada for their valuable times for me with fruitful suggestions to improve the thesis.

I thank Koji Tabata, a very nice and generous person who helped me to learn ensemble learning algorithms and their implementation in Python. I also thank Udoy Sankar Basak, a genuinely nice guy who was always very supportive, friendly as a younger brother in my daily life in Sapporo. I thank Jean-Emmanuel Clement, a helpful, friendly, nice, and smart French guy, who was always encouraging and cheerful about everything. I am very much happy to have worked with this interesting group in wonderful environment. I would like to thank Dr. Sulimon Sattari, Dr. Yuta Mizuno, and other members to make this an interesting research group. I thank the past members of our group. I would like to thank former Associate Professor in our lab Chun Biu Li for his academic advice regarding the research. My special thanks to Dr. Yuji Tamiya, who helped me a lot, when I first arrived in Sapporo.

The Raman analysis of non-alcoholic fatty liver disease discussed in this dissertation would not have been possible without Raman image data obtained by the group of professor Yoshinori Harada at Kyoto Prefectural University of Medicine (KPUM), Kyoto, Japan. I would like to acknowledge their impressive collaboration. My special thanks to Harada-sensei for his continuous inspirational discussion with us regarding the analysis and histopathology.

This work is a collaborative study between our group and other three wonderful groups from KPUM, Osaka University, and information science and technology group, Hokkaido University. I would like to thank all of them, specially, Professors Yoshinori Harada, Katsumasa Fujita, and Atsuyoshi Nakamura for their valuable comments and discussions in the JST/CREST team meetings.

It would be inappropriate if I omit to acknowledge Maiko Ishida, an extremely helpful secretary in our lab. I thank her for being so helpful and friendly both in my daily family life and academic time. I will not forget her help regarding the admission of my daughter in the Kindergarten school in Sapporo. I also thank all other staffs for their helping attitude.

It is my pleasure to acknowledge my friends and colleagues who have generously offered

their help and friendship and provided some much needed humor and entertainment during the years of my PhD period.

I thankfully acknowledge all the funding sources that helped to accomplish my PhD research. In the framework of Japanese Government-financed scholarship under International Graduate Program (IGP), Hokkaido University, The Ministry of Education, Culture, Sports, Science and Technology, Japan (MEXT) provided funding for this study. Japan Science and Technology Agency (JST)/ Core Research for Evolutional Science and Technology (CREST) was also partially supported. I wish to acknowledge MEXT, and JST/CREST for the financial support during my stay in Japan.

I would also like to express deep gratitude to Comilla University, Bangladesh, for allowing me a study leave for the whole period of my study in Hokkaido University, Japan.

Most importantly, I would like to express my deep gratitude towards my family, my biggest source of strength, for their love and support. The blessings of my late father Abdus Shahid, who spent whole of his life to support me with great inspiration in all my pursuits, have made a great contribution to reach in this stage of my life. I badly miss him in my everyday life. I am also very much thankful to my mother for her unconditional love and pray. I would not have been where I am today without her encouragement, caring, and pray. I thank my grandparents, my parents-in-law, my maternal uncle, and my sisters for their constant support, pray and love. I woe a great deal to them.

Finally, there are no words to express my gratitude to my loving, caring, supportive, and patient soul-mate, my wife Tasnuva Tarannum, whose true love during my PhD journey made this work possible. Her inspiration and motivating speech during my hard times was always cheering me up. This work would not accomplish without her great sacrifice. I would like to express my thanks to our wonderful and precious gift, my beautiful daughter Junaina, who is always the reason of my happiness in any situation of my life.

Lastly and above all, I thank the Almighty Lord for providing me the strength to reach this stage of my life.

Khalifa Mohammad Helal

Hokkaido University, December 2019

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis aims to develop analytical methods using the combination of techniques from machine learning (ML) and information theory for the quantitative analysis, and interpretation of Raman micro-spectroscopic image data of biological samples to enhance diagnostic reliability of histopathology as objectively as possible. As an illustrative vehichle, we apply our methods to the Raman hypersepctral images of non-alcoholic fatty liver disease (NAFLD) [1,2] in rat livers.

NAFLD is a common liver condition that is associated with large lipid accumulation inside hepatic tissue [3,4] with negligible evidences of alcohol consumption, and affects $20 - 30\%$ of the total population worldwide [3]. NAFLD often co-occurs with various metabolic abnormalities, such as diabetes, obesity, and dyslipidemia [2,5], and is divided into two subcategories: non-alcoholic fatty liver (NAFL) and non-alcoholic steatohepatitis (NASH) [6]. NAFL is histopathologically indicated by the presence of lipid droplets in hepatocytes without ballooning or fibrosis. In NASH, the presence of abundant hepatocellular lipid is accompanied by the appearance of inflammatory cells and hepatocyte ballooning as the manifestation of liver injury; NASH often co-occurs with fibrosis [7]. While NAFL typically has stable and benign prognosis, NASH tends to have a poor prognosis with increased risk of progression to liver cirrhosis or even hepatocellular carcinoma [8]. Recent evidence has shown that the natural course of NAFLD is more variable than previously expected. NAFL is a heterogeneous condition, partly due to the accumulation of different types of lipids with different levels of cytotoxicity, and has the potential to rapidly progress to NASH with advanced fibrosis and liver cirrhosis [9–12].

Diagnosis of various disease including NAFLD/NASH is primarily dependent on mor-

phological analysis of tissues by pathologists. Histopathologic diagnosis requires the expertise of pathologists to inspect morphological features of a biopsy or surgical specimen, and can be affected by inter- and intra-observer variability resulting diagnostic variability [13–16]. Recently, several criteria for the histologic grading/staging of NAFLD have been developed, and are clinically useful [1,17–19]. However, uncertainty remains concerning histopathology to fully predict future fibrosis (which is regarded to be a determining factor in the prognosis of NAFLD [20]) in NAFL, and to find rapid progressors in NAFL patients [9,11,12]. Thus there is a strong need to develop diagnostic tools that assist the pathologists to diagnose the disease (e.g., NAFLD) more reliably reducing the human bias, and provide objectivity and precision through recognition of the underlying biochemical compositions of the samples.

Raman spectroscopy, widely used for label-free, direct identification of molecules in biological specimens [21–27], is an emerging tool having the potential to provide biomolecular fingerprint information for clinical diagnosis of various human diseases [28–31], and to quantify the underlying molecular contents in the liver tissues. As Raman spectroscopic measurement is a non-destructive method and does not necessitate any pretreatment, fresh pathological specimens can be reused after the measurement for the subsequent histological assessments [32]. Raman hyper-spectral imaging techniques (combining spectroscopy and imaging [33]) take into account structural and chemical information about not only specific molecules but also many molecules including small molecules to which fluorescent dye labelling is difficult. Differences among the states of cells or tissues, e.g., diseased or non-diseased, are expected to lead to significant Raman spectral differences.

Recent research indicates that Raman scattering light measurement can be applied for diagnosis of various diseases [28,29]. For instance, Raman spectroscopy techniques have been used to diagnose cervical cancer [34], skin cancer [35], and rapid fiber-optic Raman spectroscopy have been used for detection of gastric intestinal disorder [36]. Furthermore, these techniques have also been used to identify cell states, and to distinguish various cell lines such as cancer or non-cancer [28,37–39], based on Raman spectral changes originated from underlying biochemical changes in cells due to disease. Raman microspectroscopy produces spatial distributions of biomolecules, e.g., cytochrome, lipids and proteins, which can be very important insights into the disease states of cells or tissues.

Raman spectroscopy may provide a means to extract biochemical information to aid

in predicting whether different states of NAFL will advance to NASH with fibrosis. Several recent studies have also used Raman-spectroscopy-based techniques to investigate NAFL and NASH. Pathological changes in the biochemical profiles of vitamin A and lipids were investigated, and found to be potential biomarkers in liver tissue that provide information concerning states of NAFLD [3,40,41]. More recently, chemometric analysis of Raman spectra that were averaged across liver tissues was used to assess the diagnostic applicability of Raman spectroscopy in the context of NAFLD [4]. Although these studies are successful in their application of Raman spectroscopy to liver tissues, information is only available on an averaged spatial scale. Considering that the histopathologist examines the tissue on a microscopic level, it is advantageous if the Raman spectra provide information on smaller spatial scales. This is problematic, however, in that Raman scattering is very weak, in that only about 1 in every $10^6 - 10^8$ incident photons are scattered inelastically [42], and are contaminated with detection noise, autofluorescence, etc. This causes spectral differences from tissues in different states of NAFLD to be too subtle for conventional analytical methods. It is therefore important to continue to develop tools to efficiently analyze the Raman spectra at smaller spatial scales that will aid in histopathological diagnosis.

Recently, various supervised learning techniques from the field of machine learning have been applied for the quantitative analysis of Raman spectroscopic data to diagnose and classify normal and malignant cells/tissues. For example, artificial neural networks (ANN) have been used for the detection of brain cancer [43], support vector machines (SVM) have been used for the screening of prostate cancer [44], for lymph node diagnostics [45], to analyze dengue infection [46], and to estimate the macrophage activation state at single cell level [47]. More recently, Edgar et al. [48] used ANN and SVM with Raman spectroscopy to classify diabetic and healthy patients, and Taylor, J. N. et al. [49] employed unsupervised machine learning to investigate follicular thyroid carcinoma (FTC-133) cells, a well-know thyroid cancer. Most of these machine learning methods are used after projecting Raman spectra into low-dimensional space using principal component analysis or any other dimensional reduction techniques. Thus, the major aim of the present study is to develop theoretical tools to differentiate and predict the stage of NAFLD in rats by referencing molecular fingerprints buried in the Raman signals. Here, we demonstrate two strategies to analyze Raman hyperspectral images of dietary

models of NAFLD as a means to investigate the different states of NAFLD that may aid pathologists in more precise diagnosis.

In the first strategy, we employ dimensional reduction/manifold learning (ISOMAP) [50] and ensemble-learning-based classification (random forest) [51,52] on the Raman spectra obtained from the regular spatial grid averaging of Raman images for predicting different states (dietary/ histological) of NAFLD. Dimensional reduction enables us to explore the underlying structures of high-dimensional data in far fewer dimensions, and to visualize the data in terms of different groups and subgroups that correspond to different tissue states. Random forest classification predicts the disease state of a tissue sample based on training from histopathological assignments, and provides the most relevant set of spectral features corresponding to NAFLD state differentiation. We show that the Raman hyper-spectral data can be used to aid in classification of the histopathological states (judged by tests of fat deposition, ballooning degeneration, inflammation, and fibrosis in tissues) of liver tissue as normal, NAFL, or NASH, and also to uncover larger variability of spectral features in NAFL, reflecting larger variability in the corresponding chemical environment as compared to normal tissue and NASH. Under the assumption that diet is a determining factor for prognosis in the models used, and using spectral distance relationships and dimensional reduction, the rats having NAFL histology can be further classified into two subgroups. One indicates cases of NAFL that are not expected to rapidly progress to NASH, while the other indicates those cases in which NASH may be expected to rapidly develop. Random forest classification also extracts the relatively important Raman shifts associated with this distinction, many of which are associated with increased deposition of important compounds associated with NASH, such as fatty acids and cholesterol.

In the second approach, we apply machine learning and information theory for Raman microscopic histology of NAFLD with increased signal-to-noise ratio of Raman images via superpixel [53] transformation. We use Raman micro-spectroscopic images of dietary models of NAFLD as a means to reveal the variation in chemical components at cellular spatial resolution. Here, we develop a diagnostic tool using Raman micro-spectroscopy coupled with unsupervised machine learning techniques to produce distributions of varying biochemical content across the tissue samples with at spatial resolution ($\sim 200\ \mu m^2$) that is on the order of single cells. First, as a major preprocessing scheme, we extend the

idea of superpixel segmentation [53,54], namely, simple linear iterative clustering (SLIC), a perceptual grouping of pixels having similar characteristics [55], to accommodate the multi-channel measurements of Raman hyperspectral imaging, generating superpixels that increase signal-to-noise ratio in comparison to the single-pixel spectra, and maintain the spatial structure of the Raman images as much as possible. Next, we employ rate-distortion theory, an unsupervised clustering method based in information theory [56,57], which reveals the groups (clusters) of spectra having similar Raman information across all the tissues. Using the Poisson error arising from photon counting as a guide, we identify a minimal number of clusters that represent regions of the tissues having similar molecular environment. We examine the distributions of those clusters across the tissues, and assess how they capture the underlying chemical environment corresponding to different dietary models or NAFLD states, as an aid in understanding the development of NAFLD. Last, we classify the different groups of the tissues using the information obtained from the cluster populations with agglomerative hierarchical clustering (AHC), allowing us to compare groupings of tissues to histopathological NAFLD states and dietary models. This provides a novel "descriptor" enabling us to infer tissue states, contributing valuable information to histological inspection. Through this analysis of cellular level molecular variation, we obtain a detailed representation of the diverse chemical environments across the liver tissues, and discover that the groupings are most representative of dietary models. This suggests our approach may serve as an aid to pathologists in more precise NAFLD diagnoses by providing detailed molecular information about the liver tissue at a cellular spatial level.

An outline of the thesis is as follows: Chapter 2 briefly explains the overview of Raman micro-spectrscopic imaging and non-alcoholic fatty liver disease. In chapter 3, after short description on the experimental system and Raman microscopic data along with histological result, we focus on the preprocessing of Raman image data. Then we explore manifold learning, and ensemble-learning-based classification and feature selection techniques. Finally, we discuss our results based on these techniques. Chapter 4 is concerned with the Raman microscopic histology of NAFLD. First, the extended simple linear iterative clustering (SLIC) is explained to segment Raman hyperspectral images. Then, we describe the clustering algorithm based on the rate determining theory (RDT), and determine the most appropriate number of clusters of spectra by considering the Pois-

son error (shot noise) originating from the Raman scattering and background photons. The agglomerative hierarchical clustering (AHC) is introduced for the classification of the state of tissues using the information obtained from the cluster populations. Based on these techniques, we discuss our obtained results and compare with the histological results. Chapter 5 sketches the summary and outlook of this work.

Chapter 2

Raman Microspectroscopy and Non-Alcoholic Fatty Liver Disease: Overview

Introduction

This chapter provides the background of Raman microspectropcopy and the overview of non-alcoholic fatty liver disease.

2.1 Raman Spectroscopy

Raman spectroscopy is an optical technique based on the phenomenon of inelastic scattering of laser light during its interaction with molecules, where emitted photon leaves molecules with an altered energy to the incident photon.

2.1.1 Raman Scattering of Light (Raman Effect)

When monochromatic (laser) light is incident upon a substance, the incident photons on the molecules of the substance can undergo two types of scattering: elastic (Rayleigh) scattering and inelastic (Raman) scattering. The molecule gains energy during the interaction with an incident photon, and thus it is excited to a vibrational energy state before a photon is emitted from this molecule. From here, there are three different possibilities of this excited molecule in terms of energy levels which can be explained using the

Jablonski energy diagram (Fig. 2.1) [58,59].

- In the first and most common case, the excited molecule elastically returns back to the ground state emitting a photon in any direction (scattering) with the same energy as the incident photon, and this process is called Rayleigh scattering. Here, the emitted photon has the same energy as the light source (incident photon) which can be expressed as $E = h\nu$, where ν is the frequency of the incident photon and h is the Planck's constant. In this case, the emitted photon has the same wavelength as incident photon.

- Second, the excited molecule drops down to the vibrational state (real photon state) from the virtual state releasing a photon with less energy (therefore longer wavelength) than the incident photon; this process is known as Stokes Raman scattering. Here, the energy of the scattered photon is equal to $h\nu - \Delta E$, where ΔE is the energy difference (loss of energy) between incident and emitted photon.

- In the third case, the incident photon interacts with a molecule which is already in a vibrational energy state instead of ground state and raised the molecule to the virtual state. The molecule returns back to the ground state emitting a scattered photon having more energy (therefore shorter wavelength) than the incident photon; this process is known as anti-Stokes Raman scattering [60]. Here the increased energy of the scattered photon is equal to $h\nu + \Delta E$.

Only a small part of incident photons, about 1 in every $10^6 - 10^8$ is scattered inelastically. Since at room temperature, most molecules stay at the ground energy state, more photons undergo Stokes Raman scattering than anti-Stokes Raman scattering [60].

The inelastic scattering of light was first predicted theoretically by Adolf G. Smekel in 1923 [61]. Raman effect was first observed in liquid experimentally by professor C. V. Raman and K. S. Krishnan in 1928, and independently by G. Landsberg and Mandelstaron [62,63], and C. V. Raman received the Noble prize in Physics for the discovery of this Raman effect in 1930 [62,64].

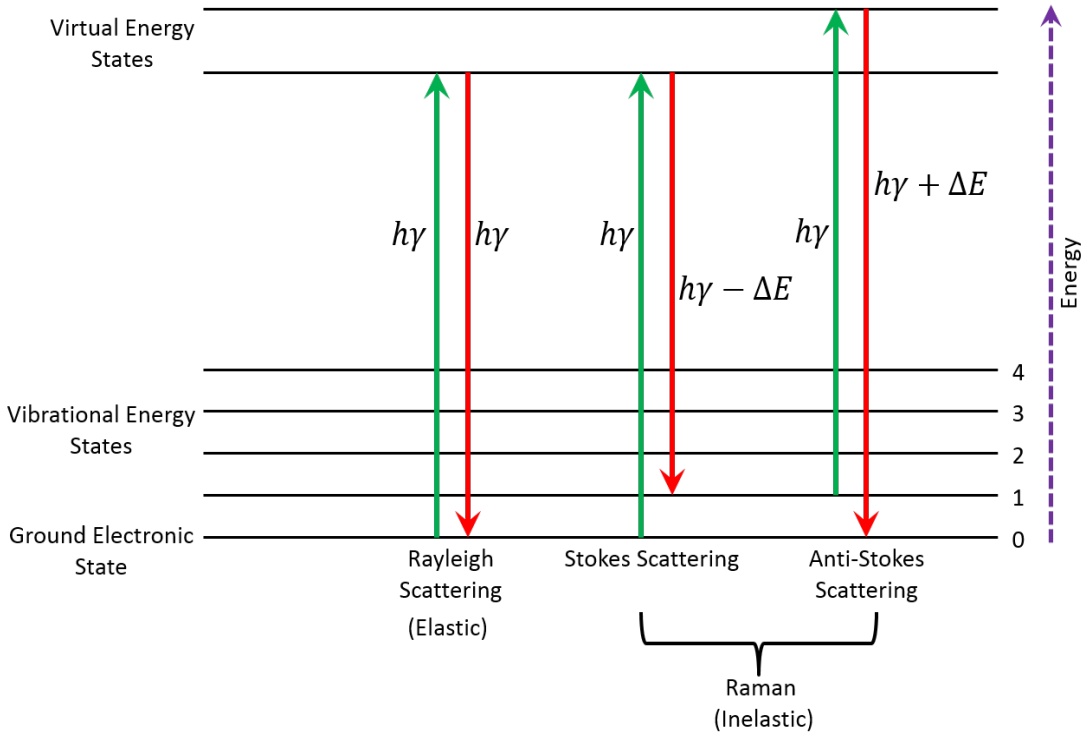


Figure 2.1: Jablonski energy diagram explaining three different types of photon scattering.

The frequency difference between incident photon and Raman scattered photon is known as Raman shift, and is defined by

$$\Delta(\mathrm{cm}^{-1}) = 10^{-7}\left(\frac{1}{\lambda_i} - \frac{1}{\lambda_s}\right), \tag{2.1}$$

where Δ is the Raman shift or wavenumber (cm^{-1}), (change in energy), λ_i is the wavenumber (nm) of the incident photon and λ_s is the scattered photon. The Raman shift is independent on the wavelength of the incident radiation, but Raman scattering depends on incident wavelength (excitation source) [65]. A plot of the intensity of scattered light versus Raman shift is known as Raman spectrum.

Raman spectroscopy uses Raman spectrometer which can detect three forms of scattering and filter out the Rayleigh, Stokes or anti-Stokes scattering. The sample is irradiated by a laser and the Raman scattering light is detected by the Raman spectrometer, and a Raman spectrum is produced for the molecule in the samples using the energy differences from initial to final state [66,67].

Raman microspectroscopy is the combination of Raman spectrometer and a standard optical microscope [60,68]. Raman spectroscopy combined with high resolution confocal microscopy can be used to image a biological samples, and to microscopic analysis. It uses

only light to probe a sample without the need of any foreign agent, and allows imaging of the samples which contain spatial and spectral information of molecular vibrations that are unique to specific cellular components [25].

Due to high resolution ($< 1\mu m$), Raman microspectroscopy provides the information at cellular and sub-cellular levels producing the images of cellular organelles such as nucleus, mitochondria, cytoplasm, lipid droplets. Raman microscopy also provides hyperspectral images of samples where every pixel across the sample provides a Raman spectrum, and therefore biological constituents and their spatial distributions along with spectral information can be identified which may be useful to detect the disease state of biological sample as diagnostic tools.

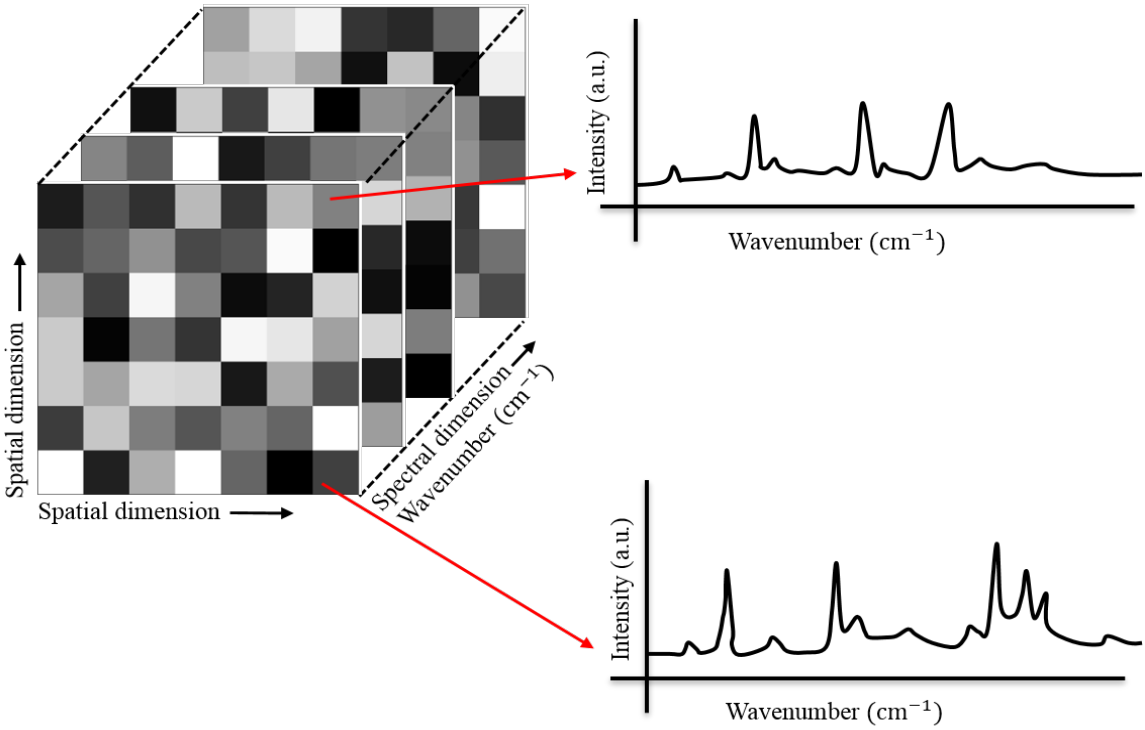


Figure 2.2: Raman hyperspectral imaging.

2.2 Non-Alcoholic Fatty Liver Disease

Non-alcoholic fatty liver disease (NAFLD) is a type of fatty liver disease having abnormal lipid accumulation in the liver of more than 5% that is not associated to significant alcohol consumption. NAFLD, a major public health problem, is one of the most common liver disease that affects men, women and children (around $20 - 30\%$ of the total population

worldwide [3]), and is identified by large lipid accumulation inside hepatic tissue [3], [4]. There are two main types of NAFLD, 'non-alcoholic fatty liver (NAFL)' in which there is lipid accumulation (simple steatosis) in the liver without liver cell damage (hepatocellular ballooning) nor fibrosis [69], and 'non-alcoholic steatohepatitis (NASH)' where there is lipid accumulation in the liver as well as inflammation and hepatocyte damage with or without fibrosis [6,69]. Steatosis is regarded as the benign condition, while NASH may lead to fibrosis, cirrhosis and/or hepatocellular carcinoma (HCC) [69,70].

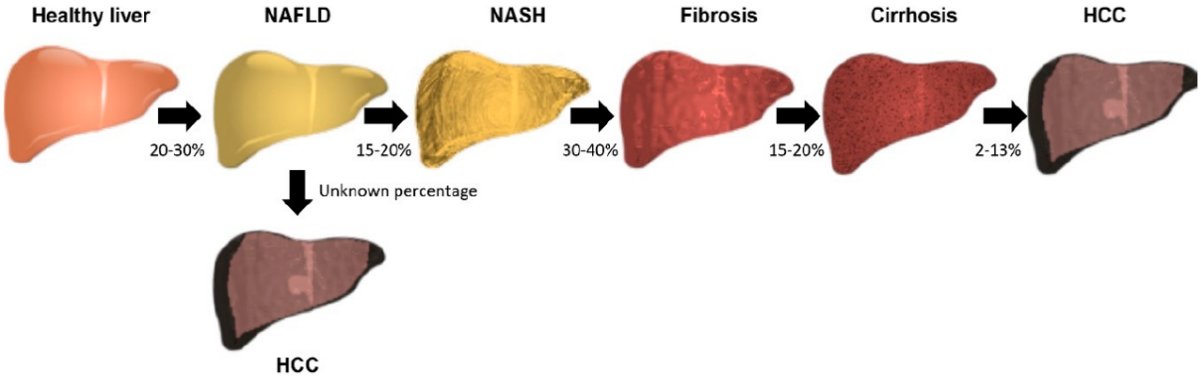


Figure 2.3: Schematic representation of the various stages of nonalcoholic fatty liver disease (NAFLD). The percentages below the arrows indicate the portions of patients progressing from one stage to the next stage. NAFLD: Non-alcoholic fatty liver disease, NASH: nonalcoholic steatohepatitis, HCC: hepatocellular carcinoma. This figure is adopted from [71].

As shown in Fig. 2.3, about 15–20% of NAFLD patients progress to NASH [71,72] and 30–40% of NASH patients progress to fibrosis [71]. About 15–20% of the patients having fibrosis may develop to advanced fibrosis or cirrhosis [71], and 2–13% of NASH patients having cirrhosis may lead to HCC [71]. Furthermore, NAFLD is remarkably associated with the people having type 2 diabetes, high blood pressure (cardiovascular disease), high levels of cholesterol and triglycerides, and specific metabolic abnormality including metabolic syndrome [73,74]. Thus, it is crucial important to diagnose the early stage of NAFLD, specifically to predict the early stage of NASH to improve the prevention of NAFLD from its severe stage.

Chapter 3

Machine Learning based Analysis of Raman Images to Evaluate Non-Alcoholic Fatty Liver Disease

In this chapter, after short description on the experimental system, Raman measurement[1], and Raman microscopic data along with histological results, we focus on the preprocessing of Raman image data. Then, we explore manifold learning, and ensemble-learning-based classification and feature selection techniques. Finally, we discuss our results based on these techniques.

3.1 Materials and Methods

Histological and Raman microscopic analysis were performed for three groups of rats fed with either a standard diet (SD), a high-fat diet (HFD), or a high-fat high-cholesterol diet (HFHC). A total of 48 liver tissues were collected after 2, 4, 8, and 16 weeks from $16(= 4 \times 4)$ rats fed with each diet for histological and Raman spectroscopic analyses.

[1]The experiments were performed by professors Hideo Tanaka and Yoshinori Harada's group at the Department of Pathology and Cell Regulation, Kyoto Prefectural University of Medicine, Kyoto, Japan. Histopathological diagnosis was performed by professors Yoshinori Harada, and Akira Okajima, Department of Gastroenterology and Hepatology, Kyoto Prefectural University of Medicine, Kyoto, Japan.

3.1.1 Animal Model and Sample Preparation

Adult Slc: Sprague-Dawley male rats (Shimizu Laboratory Supplies, Japan) at 8 weeks old were maintained under 12-hour light/dark cycles with *ad libitum* access to food and water. Rats were treated with standard diet (SD), high fat diet (HFD, 60% lard), or high fat, high cholesterol (HFHC) diet (60% lard, 1.25% cholesterol, 0.5% cholic acid). All diets were provided by Oriental Yeast Co., Tokyo. 4 animals from each group were euthanized after 2-, 4-, 8-, and 16-week feeding periods. All procedures for animal experiments were conducted under protocols approved by and in accordance with the guidelines from the Committee for Animal Research of Kyoto Prefectural University of Medicine. The liver organ of each rat was excised and dissected. A portion of the organ was sliced at $\sim$ 1 mm thickness using a handmade slicer, immediately transferred into 4°C Krebs-Henseleit buffer (KHB), and used within 2 hours. Slices were placed in glass-bottomed dishes filled with KHB during measurement. Other liver portions were fixed with 10% formalin (Wako Pure Chemical Industries, Ltd. Japan), paraffinized, sliced at 4 μm thickness, then deparaffinized for hematoxylin and eosin staining. Stained samples were evaluated by a pathologist and hepatologist conversant with liver histopathology, and classified as normal tissue, NAFL, or NASH [17,18].

3.1.2 Raman Microscopic Measurement

Raman images were acquired using a confocal Raman microscope, Raman-11 (Nanophoton, Japan). 532 nm excitation light was delivered through a 20x/0.75 dry-type objective lens (Olympus, Japan). The mode of Raman data acquisition was point excitation, and the illuminated light intensity was 68 mW/μm^2 at the sample plane. The epi-detected Raman signal was directed through the 50 μm entrance slit of a spectrometer and read by a thermoelectrically cooled charge-coupled device (CCD) camera, Pixis 400BR, at −70°C (Princeton Instruments, USA). The sample image was projected onto the slit of the spectrometer with the magnification of 18.4x when the objective lens of 20x is utilized. Raman images having area 95 μm × 345 μm (20 × 70 pixels) were obtained via point-by-point scanning with 5 μm step and 1 second exposure time. Samples were not used more than once for the Raman measurement.

3.1.3 Histological Staining of the Liver Tissues

Histological staining of the liver tissues from SD, HFD, and HFHC models after 2-, 4-, 8-, and 16-week feeding using HE and Sirius Red are depicted in Fig.3.1.

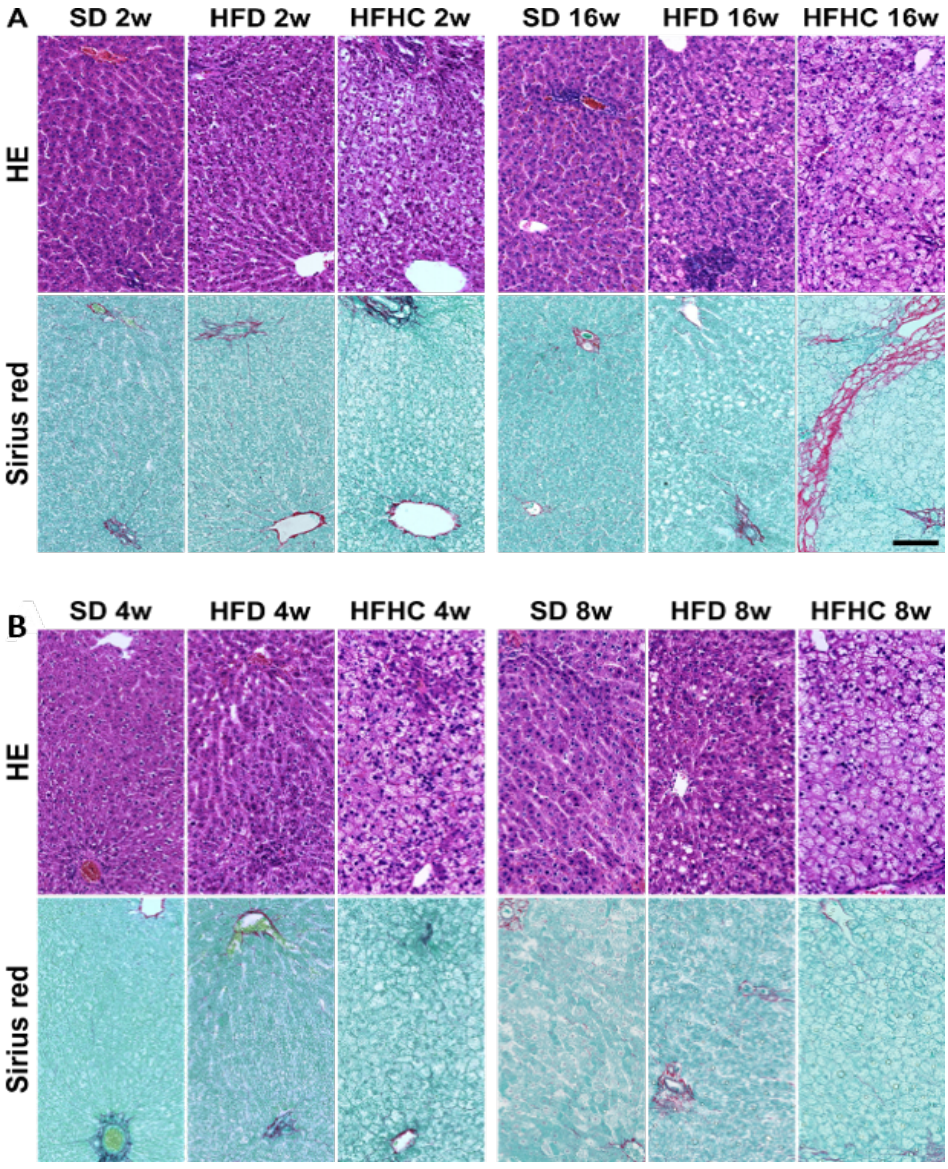


Figure 3.1: Histological staining of the liver tissues using HE and Sirius Red (A) after 2- and 16-week feeding, (B) after 4- and 8-week feeding.

Histological staining of liver tissues from SD, HFD, and HFHC models (after 4- and 8-week feeding) using HE and Sirius Red; SD: standard diet, HFD: high-fat diet, HFHC: high-fat high-cholesterol diet, HE: hematoxylin and eosin; scale bar = 100 μm.

3.1.4 Histological Evaluations of the Liver Tissues

Rats fed with HFD are reported to be a NAFLD model with slow fibrosis progression. In the HFD group, only minimal fibrosis is observed after long-term feeding (36-50 weeks) in some cases (30, 31). In contrast, rats fed with HFHC are known to be a NAFLD model with rapid fibrosis progression (32-34).

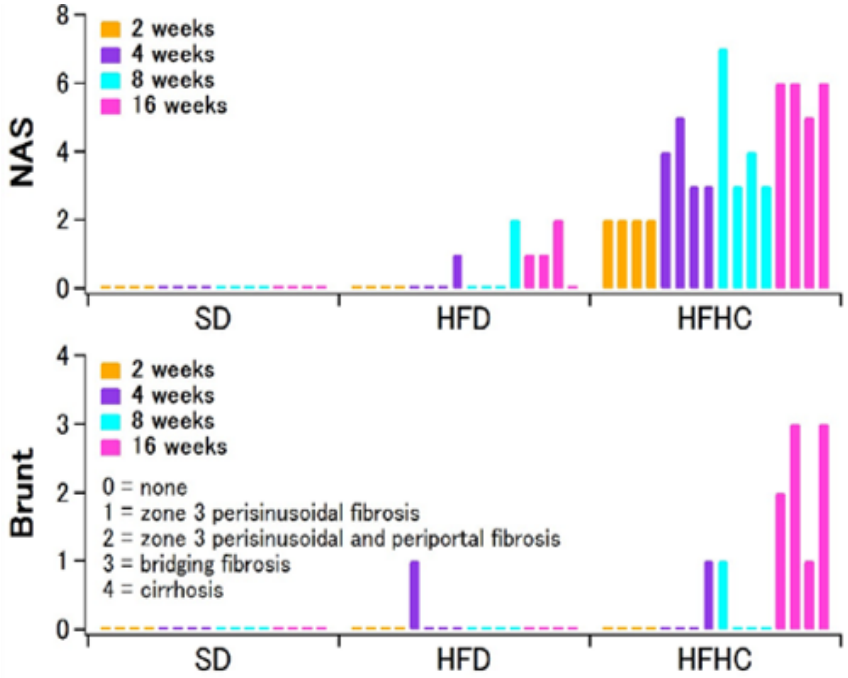


Figure 3.2: Results of histological assessments of the stained liver sections based on the NAFLD activity score (NAS) and Brunt stage criteria. Note the occurrence of NAFL in the HFD model and NASH with fibrosis in the HFHC model after 16-week feeding. (scale bar = 100 μm).

The hematoxylin and eosin (HE) and Sirius Red-stained sections from paraffin-embedded liver samples are depicted in Fig. 3.1. No accumulation of lipid droplets was observed in any liver tissue from the SD group model. Distinct lipid accumulation was observed in 16-week HFD as well as 2-, 4-, 8- and 16-weeks HFHC groups. Furthermore, the HFHC group showed strong indication of fibrosis after 16-week feeding. The results of NAFLD activity score (NAS) (25), which reflects the extent of steatosis, inflammation, and ballooning degeneration of tissues, and histological fibrosis stage classified by Brunt's criteria (24, 35) are shown in Fig. 3.2 and Table 3.1.

Table 3.1: Histological evaluation of the liver tissues. NAFLD activity score (NAS) and Brunt staging of hematoxylin-eosin-stained liver tissue sections from SD, HFD, and HFHC models (after 2-, 4-, 8- and 16-week feeding). Although NAS is reported as an indicator of disease activity, not for diagnostic purposes, the relationship between NAS and NAFLD diagnosis is reported as follows: biopsies with NAS of less than 3 were mostly considered to be NAFL; by contrast, NAS of ≥ 5 is related to a diagnosis of NASH. Biopsies with NAS of 3 and 4 correlated with both NAFL and NASH (1 of supple). The 15 NAFL rats are further divided into two groups, NAFL-α, and NAFL-β, by Raman data analysis (See the results and discussion in section 3.2).

Sample		Brunt staging	NAS			Diagnosis
			Steatosis	Inflammation	Ballooning	
	1	0	0	0	0	Normal
SD	2	0	0	0	0	Normal
2W	3	0	0	0	0	Normal
	4	0	0	0	0	Normal
	1	0	0	0	0	Normal
HFD	2	0	0	0	0	Normal
2W	3	0	0	0	0	Normal
	4	0	0	0	0	Normal
	1	0	1	1	0	NAFL-β
HFHC	2	0	2	0	0	NAFL-β
2W	3	0	2	0	0	NAFL-β
	4	0	2	0	0	NAFL-β
	1	0	0	0	0	Normal
SD	2	0	0	0	0	Normal
4W	3	0	0	0	0	Normal
	4	0	0	0	0	Normal
	1	1	0	0	0	Fibrosis, mild
HFD	2	0	0	0	0	Normal
4W	3	0	0	0	0	Normal
	4	0	1	0	0	NAFL-α
	1	0	3	1	0	NAFL-β
HFHC	2	0	3	2	0	NAFL-β
4W	3	0	2	1	0	NAFL-β
	4	0	3	0	0	NASH
	1	0	0	0	0	Normal
SD	2	0	0	0	0	Normal
8W	3	0	0	0	0	Normal
	4	0	0	0	0	Normal
	1	0	0	0	0	Normal
HFD	2	0	0	0	0	Normal
8W	3	0	0	0	0	Normal
	4	0	2	0	0	NAFL-α
	1	1	3	3	1	NASH
HFHC	2	0	2	1	0	NAFL-β
8W	3	0	3	1	0	NAFL-β
	4	0	2	1	0	NAFL-β
	1	0	0	0	0	Normal
SD	2	0	0	0	0	Normal
16W	3	0	0	0	0	Normal
	4	0	0	0	0	Normal
	1	0	1	0	0	NAFL-α
HFD	2	0	1	0	0	NAFL-α
16W	3	0	2	0	0	NAFL-α
	4	0	0	0	0	Normal
	1	2	3	3	0	NASH
HFHC	2	3	3	2	1	NASH
16W	3	1	3	2	0	NASH
	4	3	3	3	0	NASH

In HFD model three rats were developed to NAFL after 16-weeks while only one rat showed NAFL at 4- and 8-weeks feeding (i.e., HFD4w-4 and HFD8w-4). For rats on the HFHC diet, NAFL occurred even after a 2-week feeding period, and NASH with fibrosis was well-developed after 16 weeks of the HFHC diet.

As seen in Table 3.1, twenty-six rats were histologically diagnosed as normal: all SD rats, all HFD 2-week rats, and some rats on HFD for longer than two weeks. Fifteen and six rats were histologically diagnosed as having NAFL and NASH, respectively. Note that all histological indices are scored in terms of morphological features of tissues. One rat of the HFD 4 weeks model (HFD4w-1) showed a peculiar histology of the liver: while it had mild fibrosis, no hepatic steatosis was observed.

3.1.5 Preprocessing of the Raman Image Data

The raw Raman data must be pre-processed before analysis to eliminate effects of unwanted signals from auto-fluorescence, detection noise, cosmic rays, signal from cell media or glass substrate, etc., and to enhance subtle difference between different samples [75–77]. Since pure spectral features are hidden underneath noise and contamination, we need to extract the information relevant to the NAFLD states by reducing the effects of these contamination and noise in the Raman image data. We extracted the spectral features by performing some preprocessing schemes, e.g., by removing baseline drifts, reducing noise and other contaminations before quantitative analysis of the Raman image data. This section will give a detailed explanation of preprocessing of Raman microscopic data.

3.1.5.1 Bias and Baseline removal

Bias is estimated as the minimum intensity of a particular Raman tissue image, and is subtracted from each spectrum of that particular image. The baseline (arising from substrate and autofluorescence of the tissue) removal was carried out using recursive polynomial fitting [78], [79]. Each single pixel spectrum s_w is first fit to a polynomial function p_w ($= \sum_{i=0}^{m} c_i w^i$ (c_i: coefficient)) of order m, and then a modified spectrum $s'_w = \min(s_w, p_w)$ is constructed as the minimum intensities of original spectrum s_w and the fitted polynomial p_w. We counted the number of spectral points n_b of s_w below the fitted polynomial p_w, and continued the procedure until a terminative criterion holds. To make the terminative criterion consistent for all the spectra in our application we iterated

polynomial fitting until $n_b \leq 0.01 n_w$ holds, where n_w is the total number of wavenumbers. Polynomial of order $m = 8$ was chosen for fitting the spectra with the above mentioned stopping criterion. Finally, the baseline corrected spectrum was obtained by subtracting the baseline from the original spectrum.

3.1.5.2 Noise Reduction

Instrumental and photon shot noise can degrade Raman images, making it difficult to diagnose NAFLD tissues appropriately. Noise contribution in the Raman spectra has been reduced using singular value decomposition (SVD)(7, 8), which truncates smaller singular values that mostly represent noise in the data. Unfolding (unifying) the spatial dimensions of the hyperspectral image data, we obtain a 2D data matrix, where each row represents one spectrum. Applying SVD to this 2D data matrix, and reconstructing the matrix retaining the top 10 singular values (so that the sum of squares of the retained singular values captures at least 90% of the squares of all the singular values), and then reforming the 3D matrix from this reconstructed 2D matrix, the influence of the noise in the Raman spectra is reduced.

3.1.5.3 Segmentation Averaging (Regular Grid)

There are various kinds of cells, such as hepatocytes, hepatic stellate cells, and white blood cells, composing the liver tissue. Since about 6 pixels $\times$ 6 pixels (30 μm $\times$ 30 μm) is the typical size of a rat hepatocyte, we segmented the spatial dimension of Raman images into 10 pixels $\times$ 10 pixels image patches and computed the averaged spectrum over each patch. This minimizes the possibility to detect not hepatocytes but mainly the other cells, and reduces possible contributions of cosmic rays and other contamination covering 1 or 2 pixels.

3.1.5.4 Normalization of the Total Intensity (Area Normalization)

Due to the different concentration (variability of the amount of molecules) among the tissue samples, the Raman signal intensity fluctuates from sample to sample due to variable focus position producing intesity variation among the spectra. Therefore, each Raman spetrum is normalized by the area under the curve. For i-th spectrum $s_w(i)$, the

normalized spectrum $s'_w(i)$ is obtained by

$$s'_w(i) = \frac{s_w(i)}{\sum_{w=1}^{W} I_w(i)} \ . \tag{3.1}$$

3.1.5.5 Cropping Silent Region

We collected 14 averaged spectra from each image and brought them (672 spectra) together in the form of matrix for the analysis. All the spectra were cropped by removing the silent region ($1801 - 2800$ cm^{-1}), where there are no Raman signals from most of the biomolecules in the tissue [25].

3.1.5.6 Distance Matrix Calculation

To quantify differences among the spectra, we computed Euclidean (L_2) distance matrix (EDM) $D = (d_{i,j})$ [80] by measuring pairwise distance $d_{i,j}$ among the Raman spectra. The L_2 distance $d_{i,j}$ between the spectra i and j is defined as

$$d_{i,j} = \left(\sum_{w} \left[s_w(i) - s_w(j) \right]^2 \right)^{1/2} . \tag{3.2}$$

Here, w denotes the wavenumber and the summation is taken from 679 cm^{-1} to 3057 cm^{-1}, excluding the silent region. The distance matrix provides information about which (subsets of) spectra are similar to or different from other (subsets of) spectra.

3.1.6 Quantitative Analysis of the Raman Image Data

For the quantitative analysis of NAFLD data, we perform dimensional reduction and ensemble-learning-based Random forest classification. This section will give a detailed explanation of quantitative analysis of Raman microscopic data.

3.1.6.1 Dimensionality Reduction (DR)

The spectra of Raman images can be represented as vectors in a high-dimensional space, for instance, in 738-dimensional space (from 679 cm^{-1} to 3057 cm^{-1} excluding the cropping silent region 1801-2800 cm^{-1}, with increments 1 or 2 cm^{-1}) for our NAFLD data. Dimensionality reduction (DR) techniques are used as tools to construct a low-dimensional

representation, capturing the spectral variability and similarity, and to visualize the underlying hidden structures of the dataset in the low-dimensional embedded space. DR methods are classes of algorithms which can provide a two- or three-dimensional map representing the prominent structures of the data and give an embedded image of low-dimensional structures hidden in the high-dimensional information. DR methods map the high-dimensional data into a low-dimensional space that captures the geometric structure of the data as much as possible. DR methods can be of two types: linear (under the assumption that the data points lie on a linear subspace) and nonlinear (taking into account the nonlinear relationship among the variables and preserving the curved manifold). From the numerous linear DR methods, we apply classical multidimensional scaling (MDS) algorithm [81–83] to project the original high-dimensional data to a low-dimensional (embedded) space. But this linear method may not be appropriate for the analysis of the data that lies in a nonlinear manifold (curvilinear surface). For validation of the robustness of MDS, and to capture the high-dimensional nonlinear manifold appropriately, we applied a nonlinear DR method called Isometric Feature Mapping (ISOMAP)[50].

3.1.6.1.1 Multidimensional Scaling (MDS) Classical multidimensional scaling (MDS) maps the original high-dimensional data points $X = [x_1, x_2, ..., x_n]_{d\times n},\ x_i \in \mathcal{R}^d$ to data points $Y = [y_1, y_2, ..., y_n]_{p\times n},\ y_i \in \mathcal{R}^p,\ y << d$ in a low-dimensional space but preserves the relationship of pairwise distances as much as possible. It gives low-dimensional Euclidean coordinates of points by minimizing a cost function $E(X, Y)$ as the error between pairwise distances of the data points in original high-dimensional, and the low-dimensional space. That is, MDS finds configuration points in Y , by minimizing

$$E(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij}^{(X)} - d_{ij}^{(Y)})^2, \tag{3.3}$$

(S3) where $d_{ij}^{(X)} = \|x_i - x_j\|_{L_2}$ and $d_{ij}^{(Y)} = \|y_i - y_j\|_{L_2}$ are, respectively, the pairwise distances between points i and j in high- and low-dimensional spaces. Given a distance matrix D_X, MDS tries to find n data points $y_1, y_2, ..., y_n$ in a low-dimensional space with dimension p , such that D_Y is as close as possible to D_X.

3.1.6.1.2 Isometric Feature Mapping (ISOMAP) MDS can reveal the low-dimensional structure of the data as long as the data lie on a linear subspace. However, if the high-

dimensional data in question lie on a smooth submanifold (e.g., curved surface), MDS cannot always retain the structure of the underlying manifold properly, because it preserves the Euclidean distance which cannot appropriately measure the interpoint distances between the two points lying on such manifold. ISOMAP solves this problem by preserving the pairwise interpoint geodesic distance between two points constructing a neighborhood graph [50]. The steps for ISOMAP can be described briefly as:

- Using the input Euclidean distances, count nearest neighbors of every data point based either on fixed radius ϵ (ϵ-isomap) or number of nearest neighbors k, (k-isomap), where ϵ-isomap picks up all data points whose distance from the chosen point are within ϵ, and k-isomap picks up data points from the closest, to the k-th closest data points from the chosen point;
- Construct a distance-weighted neighborhood graph G (i.e., edge includes the distance information) by connecting every point to its nearest neighbors;
- Compute the shortest path on the graph between all pairs of points using Floyd's shortest path algorithm [84] which is the approximation of true geodesic distance;
- Compute geodesic distance matrix $D_G = d_G(i, j)$, where $d_G(i, j)$ = shortest path distance (approximation of geodesic distance);
- Construct the low-dimensional space Y by applying classical MDS to D_G.

ISOMAP has a single parameter ϵ or k, which needs to be tuned to find the optimal solution. In this paper, we employed ϵ-isomap algorithm that minimizes the cost function defined by

$$E(\epsilon) = \|D_G(\epsilon) - D_Y\|_{L^2}\,, \tag{3.4}$$

where $\|A\|_{L^2}$ is the matrix norm $\sqrt{\sum_{ij} A_{ij}^2}$. Choosing ϵ too large causes shortcut edges while too small ϵ makes a disconnected (sparse) graph. To select the suitable neighborhood size ϵ , we used the algorithm described by Samko et al. [85]. First, we determined the range of possible values of ϵ, $\epsilon_{\text{opt}} \in [\epsilon_{\min}, \epsilon_{\max}]$ where $\epsilon_{\min}$ is the minimum value, with which all the neighborhood graphs are fully connected in order to prevent any isolated graph, and $\epsilon_{\max}$ is $\max(d_X(i, j))$, where $d_X(i, j)$ is the input matrix in the original high-dimensional space. Second, we chose the embedding dimension of the low-dimensional

space Y as two for visual clarity. Third, we calculated all the minima of the cost function $E(\epsilon)$ with respect to ϵ , among which we obtained the optimal ϵ using the formula

$$\epsilon_{\text{opt}} = \arg\min_{\epsilon}(1 - \rho^2_{D_G\ D_y}), \tag{3.5}$$

where $\rho^2_{D_G\ D_y}$ is the standard linear correlation coefficient over entries of D_G and D_Y, where the quantity $1-\rho^2_{D_G\ D_y}$ is called residual variance; the smaller the residual variance, the better the low-dimensional distance matrix D_Y captures the structure of geodesic distance matrix D_G over the smooth submanifold lying in the original high-dimensional space.

3.1.6.2 Ensemble Learning Based Analysis

Raman spectral differences among different tissues are often very small, making it difficult to diagnose the disease by analyzing these differences visually. Random forest (RF) classifier, one of an ensemble-learning-based classifier, was used to classify the Raman spectra from different groups of tissues, and we also explored the relevant set of features (Raman shift) that are most important for that classification. Based on the difference in intensities at different Raman shifts in the spectra, the RF classifies the data sets, and also revealed the important Raman shift for discriminating the data set.

3.1.6.2.1 Classification and Feature Importance Extraction using Random Forest Random forest (RF) [51,52] is a classification method in which many decision trees are used as weak learners to classify the spectra based on their features (wavenumbers) by aggregating the class prediction of each tree using simple majority voting. In addition to classification, RF provides the information about the importance of individual variable for that classification. To estimate the relative importance of each wavenumber in the classification of the Raman spectra, we use the Random Forest (RF) classifier [51,52,86,87]. For classification purposes, RF uses information provided by a training data set, whose members have been labeled *a priori*, to assign class labels to the members of a test data set. The method utilizes decision trees as shown in Fig. 3.3 , i.e., hierarchical collections of decision nodes, each of which uses the values at a particular feature, e.g., intensities at a particular wavenumber, as a means to produce a binary partitioning of the input data set, e.g., a set of Raman spectra. The aim of RF is to produce

binary partitions that are of the highest possible purity; that is, RF aims to produce two subsets from the input data, each of which contains a large population of one class label and a small population of the other class label. We note that while the primary aim of RF is to produce a binary partitioning of a test data set, a secondary output is quantification of feature importance through estimation of the mean decrease in impurity arising from each feature in the training data set. This mean decrease impurity (MDI) is also known as the Gini importance, whose computation is described as follows.

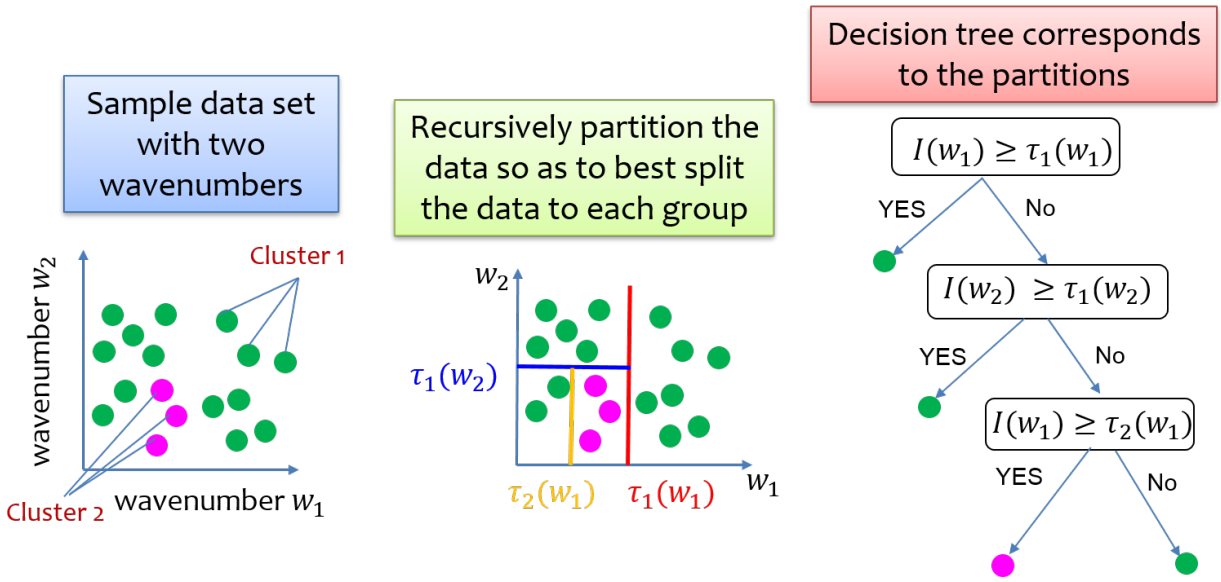


Figure 3.3: *The schematic of a simple decision tree for two wavenumbers (right) and the details of partitioning (left and middle). Decision nodes are rectangular boxes in the decision tree.*

During training, each decision tree is initialized through random selection of a single feature from the feature vector $\mathbf{w} = \{w_1, w_2, ..., w_W\}$, which contains the set of W wavenumbers at which measurements were taken. After a wavenumber w_k has been randomly selected, the input data set is divided into two groups through application of an intensity threshold τ at the selected wavenumber. This procedure represents the root node of the decision tree, producing two new nodes. Further subsets are then obtained through repetition of feature selection and binary partitioning on the two resulting subsets. Such repetition and production of new, branching nodes continues until all subsets produced by the decision nodes contain spectra originating from only a single class label, producing a single decision tree. Subsequent trees are initialized and propagated in the same manner, producing a collection (or forest) of decision trees. The importance of each wavenumber to a particular classification is assessed through a Gini importance that is

aggregated across many decision trees.

Gini importance is computed through the decrease in purity produced at the decision nodes. The purity of an arbitrary node t is assessed through the Gini impurity, which is expressed in terms of the population of each class at the node. Given the class labels of each of the N_t spectra in the (sub)set at node t, $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, ..., \gamma_{N_t}\}$, we obtain the population $p_t(c_i)$ of each class c_i,

$$p_t(c_i) = \frac{1}{N_t} \sum_{n=1}^{N} \delta(\gamma_n, c_i). \tag{3.6}$$

Here, $\delta(\gamma_n, c_i)$ is a Kronecker delta function. Considering only binary classification, we then obtain the Gini impurity at node t,

$$g(t) = \sum_{i=1}^{2} p(c_i)(1 - p(c_i)). \tag{3.7}$$

After randomly choosing a wavenumber $w_k^{(t)}$ for node t, the next task is to determine an intensity threshold such that the two subsets resulting from the division of the (sub)set at node t into child nodes t_l and t_r are as pure as possible. This is accomplished through maximization of the decrease in Gini impurity resulting from the binary partitioning of the (sub)set at node t. Given the (sub)set of spectral intensities at a randomly chosen wavenumber $w_k^{(t)}$, $\mathbf{I}_k = \{I_k^{(1)}, I_k^{(2)}, ..., I_k^{(N_t)}\}$ and arbitrary intensity threshold τ, we compute the decrease in Gini impurity at node t, $\Delta g_t(\tau)$, to be the difference of the Gini impurity at node t and the weighted sum of that produced at nodes t_l and t_r,

$$\Delta g(t) = \max_{\min \mathbf{I}_k < \tau < \max \mathbf{I}_k} \left[g(t) - \frac{N_l}{N_t} g(t_l, \tau) - \frac{N_r}{N_t} g(t_r, \tau) \right]. \tag{3.8}$$

Here, N_l and N_r are the numbers of spectra assigned to nodes t_l and t_r, respectively. The decrease in Gini impurity for a particular wavenumber w_j with a particular decision tree T is then the sum of the Gini impurity produced by that wavenumber over all nodes t in the tree T,

$$\Delta g(w_j, T) = \sum_{t \in T} \delta(w_j, w_k^{(t)}) \Delta g(t). \tag{3.9}$$

Note that $\delta(w_j, w_k^{(t)})$ is a Kronecker delta function. Finally, we compute the mean Gini

impurity over all trees to obtain the Gini importance for wavenumber w_j,

$$\Delta g(w_j) = \frac{1}{N_{trees}} \sum_{T}^{N_{trees}} \Delta g(w_j, T). \tag{3.10}$$

The second measure for feature importance we computed, permutation importance, elucidates how each feature influences the accuracy of random forest using test data. For each tree, the unused spectra (approximately 36%) , known as the out-of-bag (OOB) data, were used to compute the permutation importance, or the OOB error [51]. To measure the importance of each feature variable, new test data are made from the OOB data by randomly shuffling the values of the particular feature variable in the OOB data, while keeping the values of all other variables unchanged, and the prediction error of this new test data is again calculated by the same tree. The importance of the feature variable is computed from the difference between the two prediction errors. The score is calculated for each tree of random forest, and averaged over all the trees in the random forest [88].

3.2 Results and Discussions

3.2.1 Observations from the Distance Matrix.

How can Raman imaging identify the underlying state of NAFLD? We analyzed the Raman data based on the histological assignments and scores. After applying data pre-processing schemes (section 3.1.5), the Euclidean distances matrix $D = (d_{ij})$ is obtained by the pairwise distances d_{ij} among the Raman spectra $S(w)$ (w: wavenumber) averaged over the typical size of hepatocytes in the liver tissue Raman images. Here the distance d_{ij} is evaluated by Euclidean (L_2) distance $\sqrt{\sum_w |S_i'(w) - S_j'(w)|^2}$, where $S_i(w)$ is normalized to $S_i'(w)\big(= \frac{S_i(w)}{\sum_w S_i(w)}\big)$ (see section 3.1.5 for details). The distance d_{ij} quantifies dissimilarity between the normalized spectra $S_i^{'}$ and $S_j^{'}$ (see section 3.1.5 for details) ; smaller distances indicate similarity in the (averaged) chemical environment at the size of hepatocytes, while larger distances indicate different chemical environments.

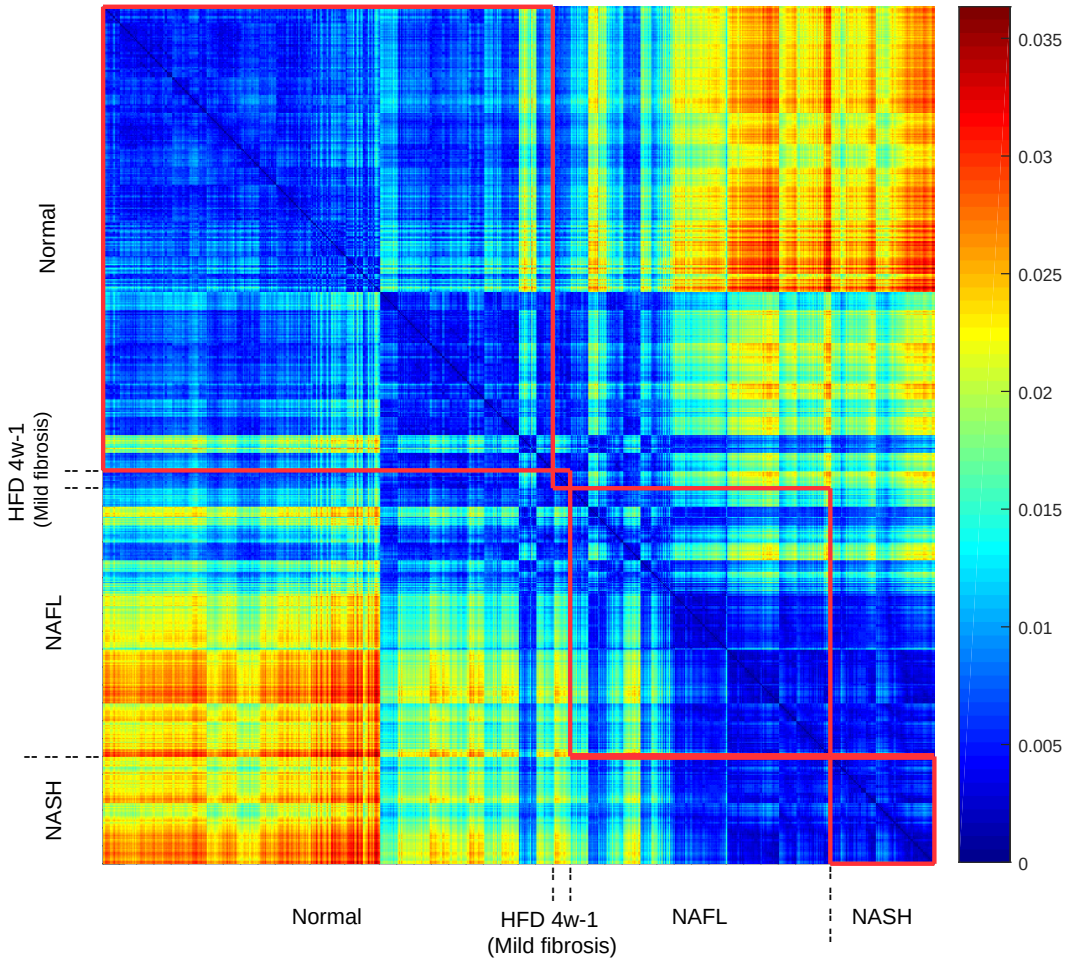


Figure 3.4: Distance matrix representing the pairwise distances among the spectra for all histological states. The distance matrix representing pairwise Euclidean (L_2) distance between the area-normalized spectra and histological diagnosis. The bluer (redder) the color in the matrix element, the more similar (different) the underlying chemical environment at the cell level. The column and row are ordered by referencing the histological states in Table 3.1 from Normal to "Fibrosis, mild (HFD 4w-1)", NAFL-α, NAFL-β, and NASH. The detailed order is as follows: Normal: SD2w-1, SD2w-2, SD2w-3, SD2w-4, HFD2w-1, HFD2w-2, HFD2w-3, HFD2w-4, SD4w-1, SD4w-2, SD4w-3, SD4w-4, HFD4w-2, HFD2w-3, SD8w-1, SD8w-2, SD8w-3, SD8w-4, HFD8w-1, HFD8w-2, HFD8w-3, SD16w-1, SD16w-2, SD16w-3, SD16w-4, HFD16w-4 (26 samples) Fibrosis, mild: HFD4w-1 (1 sample)
NAFL-α: HFD4w-4, HFD8w-4, HFD16w-1, HFD16w-2, HFD16w-3 (5 samples)
NAFL-β: HFHC2w-1, HFHC2w-2, HFHC2w-3, HFHC2w-4, HFHC4w-1, HFHC4w-2, HFHC4w-3, HFHC8w-2, HFHC8w-3, HFHC8w-4 (10 samples)
NASH: HFHC4w-4, HFHC8w-1, HFHC16w-1, HFHC16w-2, HFHC16w-3, HFHC16w-4 (6 samples)
For example, HFD4w-1 denotes a set of Raman spectra averaged over the size of hepatocytes (10 pixels $\times$ 10 pixels with 1 pixel being 5 μm) of the 1st rat among 4 individuals that were fed a highfat diet for 4 weeks.

Fig. 3.4 indicates large distances between rats diagnosed as normal and those diagnosed as having NASH, while distances between those and the rats from the NAFL group

show diverse behavior. The row and column of the distance matrix are ordered based on histological results, i.e., Normal followed by "Fibrosis, mild (HFD4w-1)", NAFL, and NASH. The block-diagonal matrices indicated by red lines represent the groups of Normal, "Fibrosis, mild (HFD 4w-1)", NAFL, and NASH, respectively.

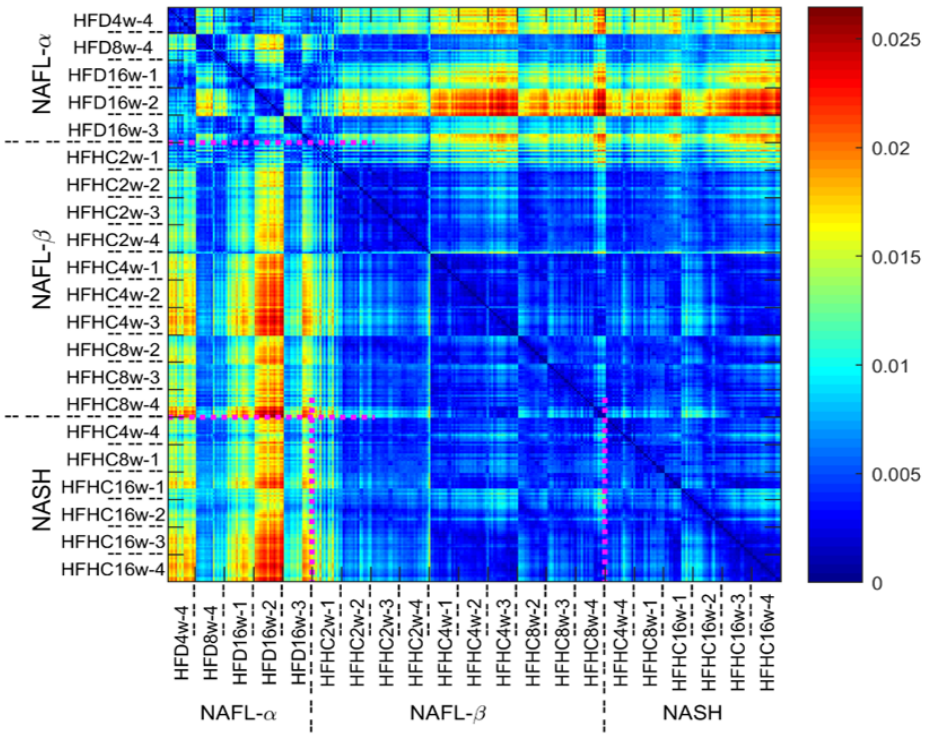


Figure 3.5: The distance matrix representing pairwise Euclidean (L_2) distance between the area-normalized spectra and histological diagnosis NAFL and NASH. For example, HFD8w-4 denotes a set of Raman spectra averaged over the size of hepatocytes (10 pixels $\times$ 10 pixels with 1 pixel being 5 μm) of the 4-th rat among 4 individuals that were fed high-fat diet for 8 weeks. The bluer (redder) the color in the matrix element, the more similar (different) the underlying chemical environment at the cell level.

Although histological analysis shows no clear distinction, Raman analysis indicates that the rats diagnosed as having NAFL from the HFD group mostly have large L_2 distances to those rats diagnosed as having NAFL from the HFHC diet group. Considering that HFD shows slow and/or rare fibrosis progression in comparison to rapid fibrosis progression for HFHC diet [89–92], we term the HFD group of NAFL-diagnosed rats to be NAFL-α, while NAFL-diagnosed HFHC rats are termed NAFL-β. Because the NAFL-β group has, on average, smaller L_2 distances to the histologically assigned NASH group than does the NAFL-α group, the degree of similarity between NASH and NAFL-β is larger than those between NASH and NAFL-α. It should be noted however, according to the spectral similarity as measured by L2 distance, HFD 8w-4 (NAFL-α) and HFHC2w-1

(NAFL-β) are close to each other, as compared to some of NAFL-α group and some of NAFL-β group, respectively. Fig. 3.5 for the distance matrix among NAFL-α, NAFL-β, and NASH confirms these results.

3.2.2 Dimensional Reduction.

To capture the underlying relationship among chemical environments of tissues (Raman spectra) and the disease state, we map the spectra to a two-dimensional space by performing a nonlinear isometric feature mapping (ISOMAP) [50,85] in Fig. 3.6 (details in section 3.1.6.1). We have used optimal parameter for ISOMAP. IWe have also compared the ISOMAP mapping with classical multidimensional scaling (MDS). The details of the parameter selection for ISOMAP, and comparison between ISOMAP and MDS is given in section 3.2.2.1. Each point in Fig. 3.6 corresponds to a single spectrum, and all spectra are colored according to their pathological states. Distances between points in the ISOMAP space are interpreted similarly to the L_2-distance above, i.e., nearby points are similar in chemical environment while distant points are less similar. One can see that normal (blue) and NASH (red) spectra are clearly separated. NAFL spectra (green and purple) are largely variable and overlapped with both the normal and NASH spectra, which indicates heterogeneity in the characteristics of NAFL [8].

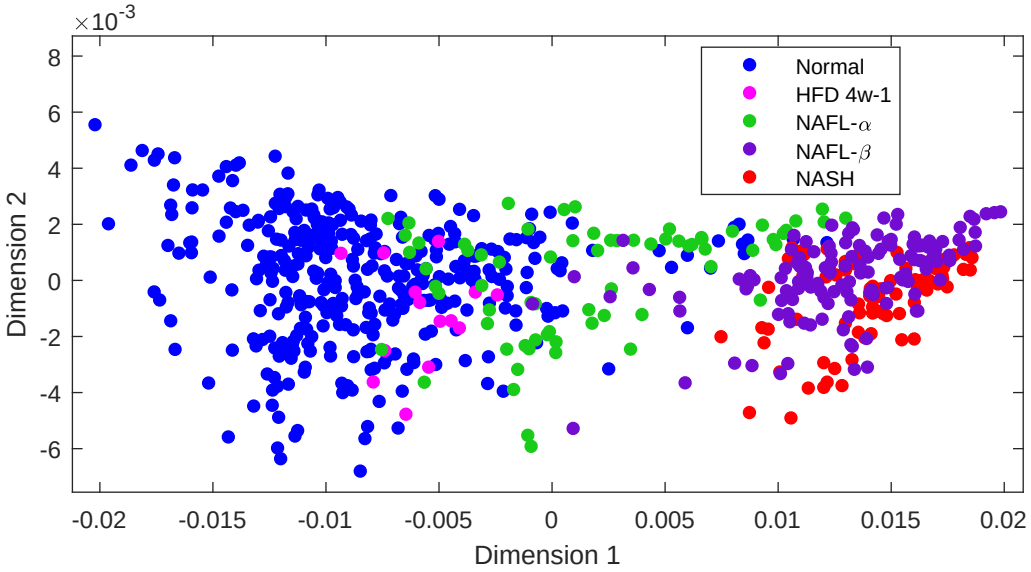


Figure 3.6: Scatterplot of the spectra in two-dimensional ISOMAP space.

In Table 3.1, out of 15 NAFL rats, 5 rats are labeled as 'NAFL-α' group, whose

total NAS score is 1 or 2, and 10 rats are labeled as 'NAFL-β', whose total NAS score is 2-5. This indicates a certain consistency between the Raman data and histological assignment, even though the former and the latter reflect different information concerning the (microscopic cell level's) chemical environments and morphological characteristics of tissues. Although the HFD 4w-1 rat showed a unique histological score (zero NAS score with mild fibrosis) of the liver, in Raman analysis its spectra overlap with the normal histology group. It should be noted again that our assignment of two subgroups of NAFL is not solely dependent on microscopic chemical environments of tissues (see, e.g., some NAFL-α is closer to NAFL-β and NASH in Raman spectral distance than the other NAFL-α in Fig. 3.6) but also on the fact that HFD shows slower and/or more rare fibrosis progression than that for HFHC diet [89–92]. In Fig. 3.7, we compared the standard linear mapping, classical multidimensional scaling (MDS), and the nonlinear ISOMAP for our data, showing that the two-dimensional classical MDS only qualitatively capture Raman spectral feature relationship among different histological groups.

3.2.2.1 Comparison between ISOMAP and Conventional Multidimensional Scaling (MDS) for NAFLD Data

In this Section, we compare our Isometric Feature Mapping (ISOMAP) to the conventional Multidimensional Scaling (MDS) that projects high-dimensional spectral data into a linear subspace of two dimensions.

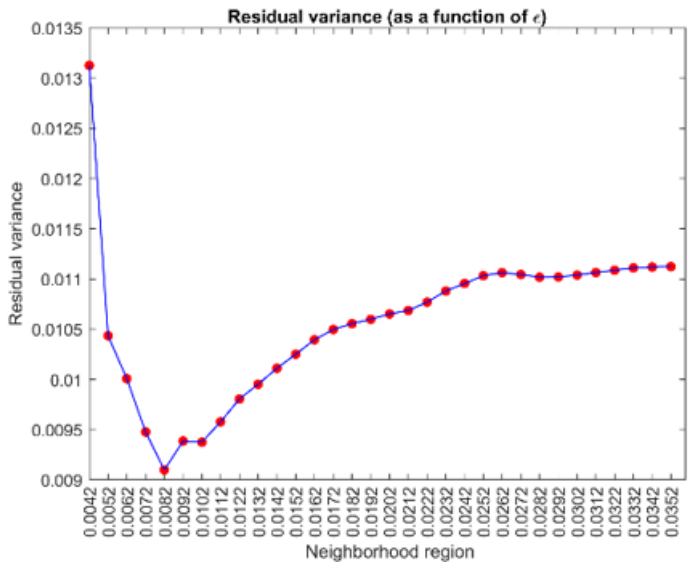


Figure 3.7: Residual variance as function of neighborhood size. Here $\epsilon_{\text{opt}} = 0.0082$.

The linear mapping MDS corresponds to a two-dimensional projection by using two principal components with the largest and the second largest variances if the distance relationship is evaluated in terms of L_2 distance. The optimal ISOMAP parameter selection from our NAFLD data is illustrated in Fig. 3.7. From the possible interval $\epsilon \in [0.0042, 0.036]$, $\epsilon_{\text{opt}} = 0.0082$ is chosen as optimal to find the smallest residual variance projection of datasets most accurately.

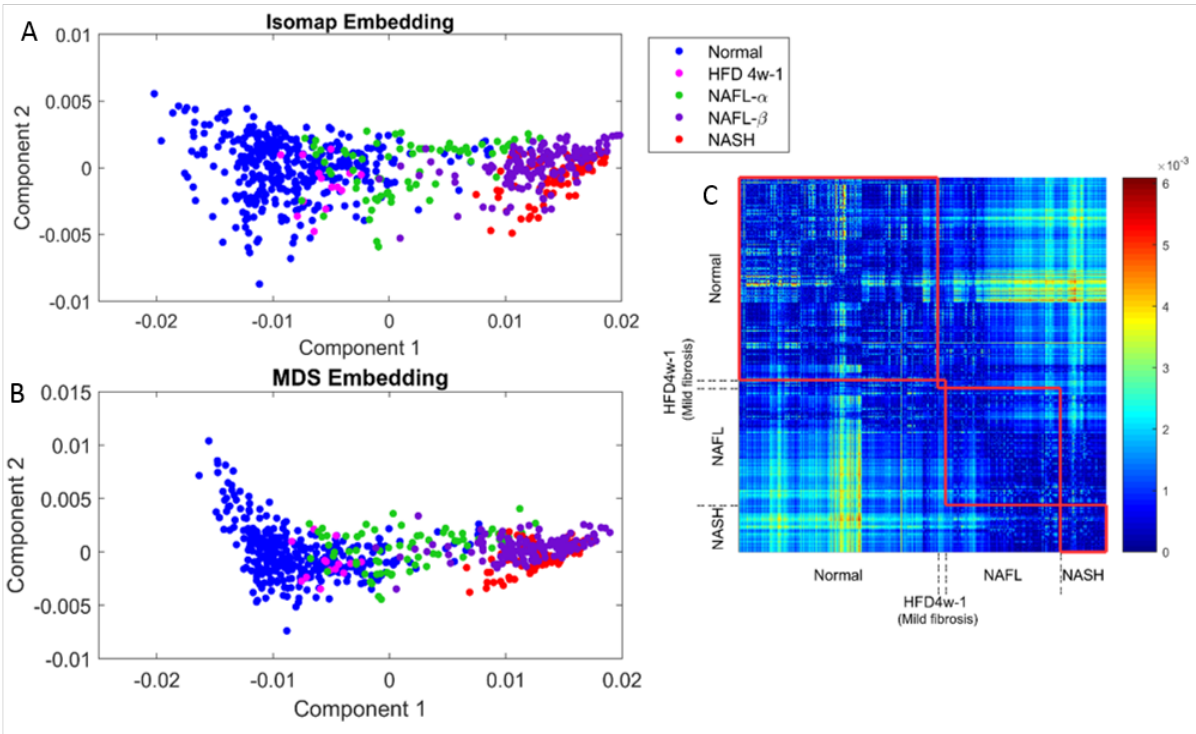


Figure 3.8: Scatterplots using ISOMAP and MDS, and difference in distance matrix in the projected space.

The two-dimensional scatterplots of the spectra using MDS and ISOMAP are given in Figs. 3.8A and 3.8B, which provides a visual comparison. Although visually there is no significant difference between the two projections in the perspective of spatial distributions of the spectra, the subtle difference is seen. For instance, the distances among the spectra are smaller in magnitude and more compact along the first component (horizontal axis) of MDS compared to those of ISOMAP. The absolute difference between the elements of distance matrix from two different projections is illustrated in Fig. 3.8C. The differences within 'normal', 'NAFL' and 'NASH' groups shown in the diagonal blocks of the distance matrix are very low indicating that MDS quantitatively captures the distance relationship by low-dimensional MDS space. In turn, most part of the off-diagonal blocks shows that the difference in distances are relatively high indicating that distance

relationship between the groups in ISOMAP is captured by the MDS only qualitatively.

As a consequence, two-dimensional classical MDS qualitatively capture Raman spectral features of NAFLD of different histological groups of rats, even though one should keep in mind that accuracy is dependent on which pair of Raman spectral features are compared. These two projection methods can be versatile tools to analyze and explore the high-dimensional data in order to prevent us from making a misleading interpretation by a linear projection because there exists no a priori reason why the linear subspace employed by MDS can capture the underlying relationship among the spectra of NAFLD and there may exist some possibility that our interpretation is due to the apparent projection of high-dimensional spectral data into the linear subspace of two dimensions.

3.2.3 Time Development of Pathological State of NAFLD Dependent on Diets.

How does the pathological state of rat liver progress in the chemical space represented by Raman data depending on diets? Note that the time from 2 to 16 weeks is not physical time because four rats are sacrificed every two weeks. Here we assume that it indicates the time progression of pathological states as an ensemble. In Fig. 3.9, a two-dimensional spectral feature space represented by the 2-D ISOMAP where x and y best describe the mutual relationship in a sense of L_2 distance among Raman spectra, we show the joint probability distributions for each diet across weeks to show the progression of the rat liver in this chemical space (here, the least/most populous regions are blue/red). To explore whether the distributions approach to some steady distributions, the progression of the joint probability distributions are constructed as follows: from Raman data of 2 weeks (at 2 weeks), 2 and 4 weeks (at 4 weeks), 2, 4, and 8 weeks (at 8 weeks), and all weeks (at 16 weeks). To trace the time development of the pathological state of the rat livers, the ISOMAP locations of the spectra, as they appear in Fig. 3.6, overlay the probability distributions and indicates pathological indices (i.e., normal, NAFL-α, NAFL-β, NASH).

The spectra of the SD and HFHC diet models are less dispersed in this feature space than are those of the HFD model, with SD and HFHC spectra being relatively localized in largely distinctive regions, while those of the HFD model are scattered across a broader region of the feature space. Looking into histological characteristics, one can see that the livers histologically assigned as normal from the SD model and those assigned as normal

from HFD model are located in different regions (e.g., see SD 2w, SD 4w, HFD 2w). This suggests that, even though no morphological differences are present in the normal liver tissues of the HFD model, the microscopic chemical environment is different for the two different diets. That is, some chemical changes occurred due to high fat diet consumption.

The HFHC diet rats are clearly separated from the SD rats in this feature space and are relatively localized after a small shift along the horizontal direction from two- to four-weeks. Note here that the tissues assigned as NAFL-β that have been on the HFHC diet for 4 weeks or longer and NASH tissues are located in almost identical regions of this Raman feature space, consistent with the observation from the distance matrix of Fig. 3.4). This indicates that the histopathological state of the liver cannot perfectly be differentiated by Raman spectra. However, it is suggested that, once the chemical environment of a liver tissue quantified by Raman image is transitioned to that region in the space, the liver tissue keeps its chemical circumstance for some time duration, and transitions to NASH. That is, Raman diagnosis is expected to predict the future development of NAFL-β to NASH before morphological features of NASH emerge.

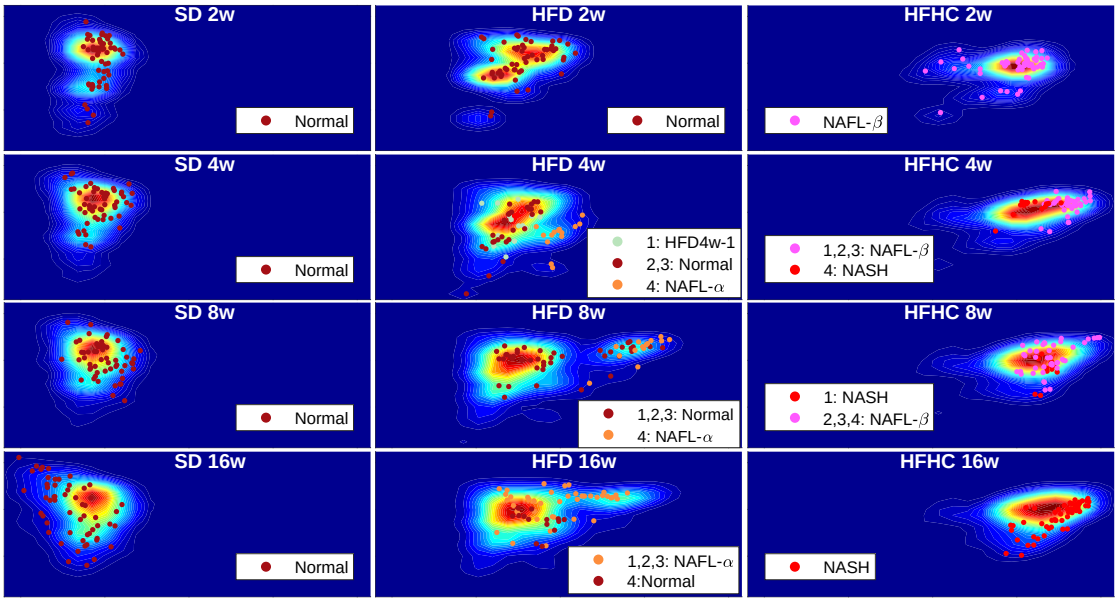


Figure 3.9: The "time" propagation of pathological states in the two-diemensional Raman spectral feature space dependent on diets superimposed with the contour plots of those points. The joint probabilities were filtered with a 2D Gaussian smoothing kernel [93].

For the HFD diet model, the dispersion of the Raman spectra of the liver tissue is larger as the "time" elapses, in comparison to the SD and HFHC diet models. This

may suggest that the liver tissue observed in the first two weeks of HFD feeding is more variable or unstable than those in the first two weeks of SD and HFHC feeding. Note that NAFL-α and normal tissue cannot be well differentiated in the Raman feature space but remain well separated from the NASH spectra (c.f., the distance matrix in Fig. 3.4).

Finally, let us comment on the progression of the joint probability distributions of the rat liver in the chemical space from Raman image data. Disease progression is out of equilibrium with input of different nutrition. At nonequilibrium steady state, the movement on energy landscape (population landscape) are triggered by both the potential of mean force and flux on that space at nonequilibrium steady state [94,95]. For the SD and HFHC models, the distributions might be regarded as "being steady" compared to that for HFD model at the timescale of sixteen weeks because the joint probability distributions of the SD and HFHC models became quickly localized after exploration for four weeks, while that of the HFD is still developing to bifurcate into two regimes. One might infer population landscape relevant to energy landscapes at equilibrium [94–97] on the fatty liver tissue states in terms of population taken over two to sixteen weeks at least for SD and HFHC diet models.

3.2.4 Random Forest (RF) based Classification and Feature Importance Extraction.

The success of the ISOMAP mapping suggests that only a few dimensions are needed to differentiate and predict the histological states from the Raman images. Here, we use the ensemble learning-based random forest (RF) classifier [51,52] to systematically identify a set of important wavenumbers in the discrimination of histological stages such as 'NAFL-α' versus 'NAFL-β'. The details of random forest (RF) classifier is given in section 3.1.6.2.

Briefly, RF is typically composed of a few hundred decision trees, each of which contains nodes that associate a subset of the training data with a histological identification, or label, such as normal, NAFL, or NASH. A node in the decision tree operates like an "IF" function, e.g., "if the intensity of a specific Raman shift is greater than a certain threshold." Each node is accompanied by two edges, corresponding to "true" or "false" outcomes in the application of the IF condition, along which the parent data set is split into two daughter subsets. The wavenumbers and the thresholds used at the nodes are

determined so as to split the parent data such that the daughter subsets contain either of the labels as uniquely as possible. The importance of each wavenumber is evaluated with reference to how often it was used over the few hundred decision trees, as well as by how successful it was in providing the correct label for the training data set.

3.2.5 Heterogeneity of NAFL: NAFL-α versus NAFL-β.

Since NAFL has the possibility to progress to NASH with fibrosis, early prediction is essential to prevent its progression to NASH, which may subsequently proceed to cirrhosis [9,11]. Since it is difficult to forecast progression to NASH by classical histology [8], we focus on the histologically-assigned NAFL state from which we extract the wavenumbers (molecular information) that lead to the diversity of the NAFL state. Fig. 3.10 shows the

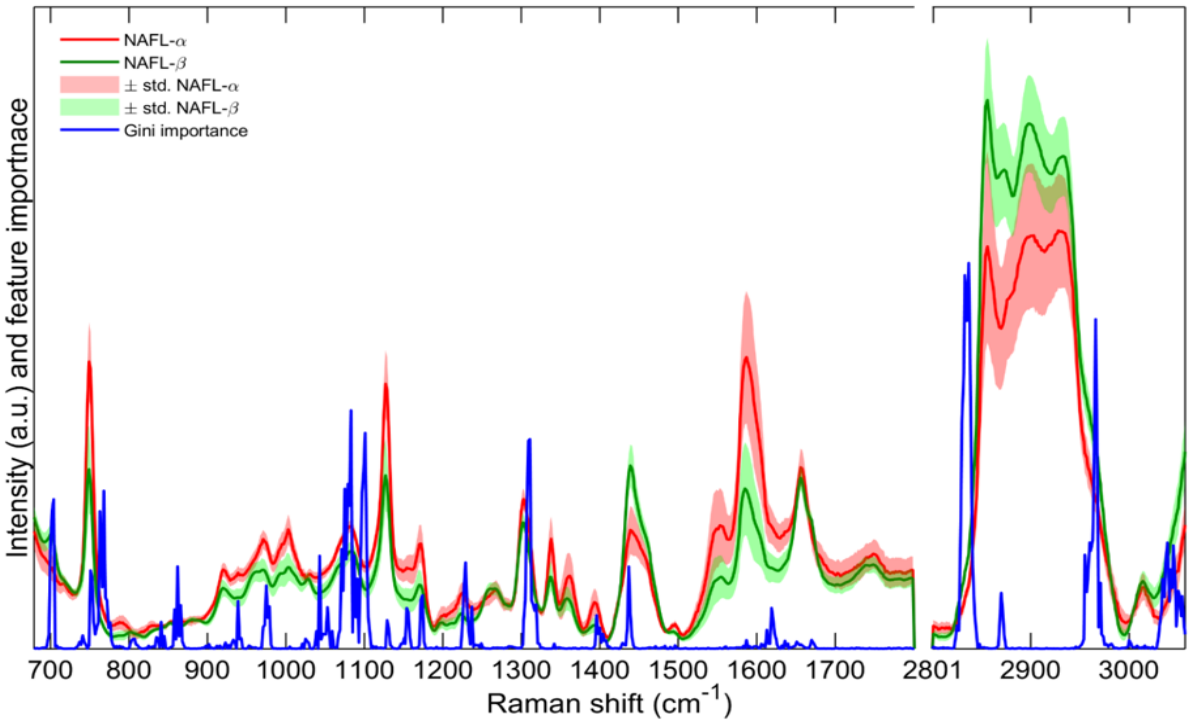


Figure 3.10: Averaged Raman spectra: NAFL-α (red-solid line); NAFL–β (green-solid line); standard deviation NAFL-α (red patch); standard deviation NAFL-β (green patch); Gini importance (blue-solid line). RF is comprised of 5,000 trees using full NAFL data as the training set. The silent region 1801-2800 cm^{-1} is excluded for visual purpose. .

averaged Raman spectra of NAFL-α (red) and NAFL-β (green) with standard deviations given by its shaded patches (enlarged figures are also given in Figs. 3.11A and 3.11B). One can see that the mean spectra of NAFL-α and NAFL-β exhibit significant Raman peaks at 750, 1127, 1441, 1591, 1655, and 2850-2950 cm^{-1}. The larger Raman peaks at 1441 and 2850-2950 cm^{-1} in NAFL-β than those in NAFL-α are considered to be

associated to accumulation of various types and quantities of lipids [98]. In addition, the Raman peak at 1591 cm^{-1}, due to vitamin A [41,99], and those at 750 and 1127 cm^{-1}, due to cytochrome c in mitochondria [24], show opposite tendencies from that of lipid, in that the relative intensities are suppressed in NAFL-β compared to NAFL-α.

The questions here are twofold. Which Raman shifts are essential in differentiation of these two different NAFL groups? How well can the RF machinery identify Raman images as the correct state of NAFL, i.e., either NAFL-α or NAFL-β?

RF analysis for the NAFL group was performed using 210 spectra from 15 rats. The data were split into 15 disjoint sets of equal size (based on the number of rats) where each set contains 14 spectra from each rat. One rat (14 spectra) was used for testing and the remaining 14 rats (196 spectra) were used for training the RF using several different numbers of trees (T=30, 50, 100, 200, 300). As shown in Table 3.2, the classification results of our NAFL data using RF classifier are not sensitive with respect to the total number of trees. It is observed that RF can classify NAFL-α and NAFL-β efficiently with accuracy of 91%. A total of 20 of 210 spectra are falsely classified: 12 spectra (out of 14) from the rat HFD 8w-4 and 6 spectra (out of 14) from the rat HFHC 2w-1. This result agrees with the distance matrix shown in Figs. 3.4 and 3.5, in that the degree of similarity of HFD 8w-4 rat to NAFL-β is higher (bluer) in comparison to NAFL-α. Conversely, some HFHC 2w-1 spectra show higher similarity to the NAFL-α group. That is, the classification of NAFL-α and NAFL-β by referring to diet model [89–92] shows overall consistency with the Raman analysis with high accuracy, but there exists some fuzziness in its classification of NAFL-α and NAFL-β.

Table 3.2: The performance of the RF classifier in predicting NAFL-α and NAFL-β.

Performance	Number of trees T				
	30	50	100	200	300
TN	60	60	59	59	59
FN	8	8	8	8	8
TP	132	132	132	132	132
FP	10	10	11	11	11
Accuracy	0.914	0.914	0.910	0.910	0.910

What chemical information can one learn from the RF classifier? The RF classifier is composed of 5,000 decision trees (for elucidation of importance measures) in which

each of the nodes has an IF function with respect to a specific Raman shift and some threshold. The importance of Raman shifts is naturally evaluated by referring to how often a Raman shift was employed in the decision trees, and by how uniquely it splits the data set into NAFL-α and NAFL-β.

In Fig. 3.10 the importance (Gini importance) [86,87] of Raman shifts in separating these two NAFL states is shown as a blue solid line (enlarged figures are also given in Figs. 3.11A and 3.11B). (We also confirmed that the importance analysis to our NAFLD data was almost independent of the choice of the measures of importance by recomputing the feature importance using permutation importance [52,86,87] (details in section 3.2.5.1)).

The important Raman shifts for the discrimination of NAFL-α and NAFL-β are mainly associated with the six most important spectral regions: 2829-2838 cm^{-1}, 2964-2968 cm^{-1}, 1079-1083 cm^{-1}, 1307-1311 cm^{-1}, 763-768 cm^{-1}, 700-704 cm^{-1}. The first two Raman regions (2829-2838 cm^{-1}, 2964-2968 cm^{-1}) are found to have large importance for the discrimination between NAFL-α and NAFL-β, appearing on the shoulders of the Raman bands 2850-2950 cm-1, whose relative intensity averaged over NAFL-β group is larger than that over NAFL-α group.

This is interpreted as follows: the standard deviations (shaded bands in Figs. 3.10 and 3.11A) are so large among each set of the spectra of NAFL-α and NAFL-β for this region that the intensities of these Raman shifts are overlapped between NAFL-α and NAFL-β. Although relative lipid accumulation is higher in NAFL-β than in NAFL-α on average, the diversity of chemical environments corresponding to these Raman shifts in NAFL-α and NAFL-β prevents us from differentiating NAFL-β and NAFL-α with high accuracy. Instead, the shoulders 2830-2840 cm^{-1} and 2960-2970 cm^{-1} have smaller standard deviations in each set of the spectra NAFL-α and NAFL-β that are found to be more versatile in differentiating NAFL-α and NAFL-β with higher accuracy (3.11A). Fig. 3.11 presents the Gini importance and permutation importance that quantify the importance of Raman shifts in differentiating NAFL-α and NAFL-β at (A) high, and (B) low wavenumber regions, with the averaged Raman shifts of NAFL-α and NAFL-β and the standard deviations.

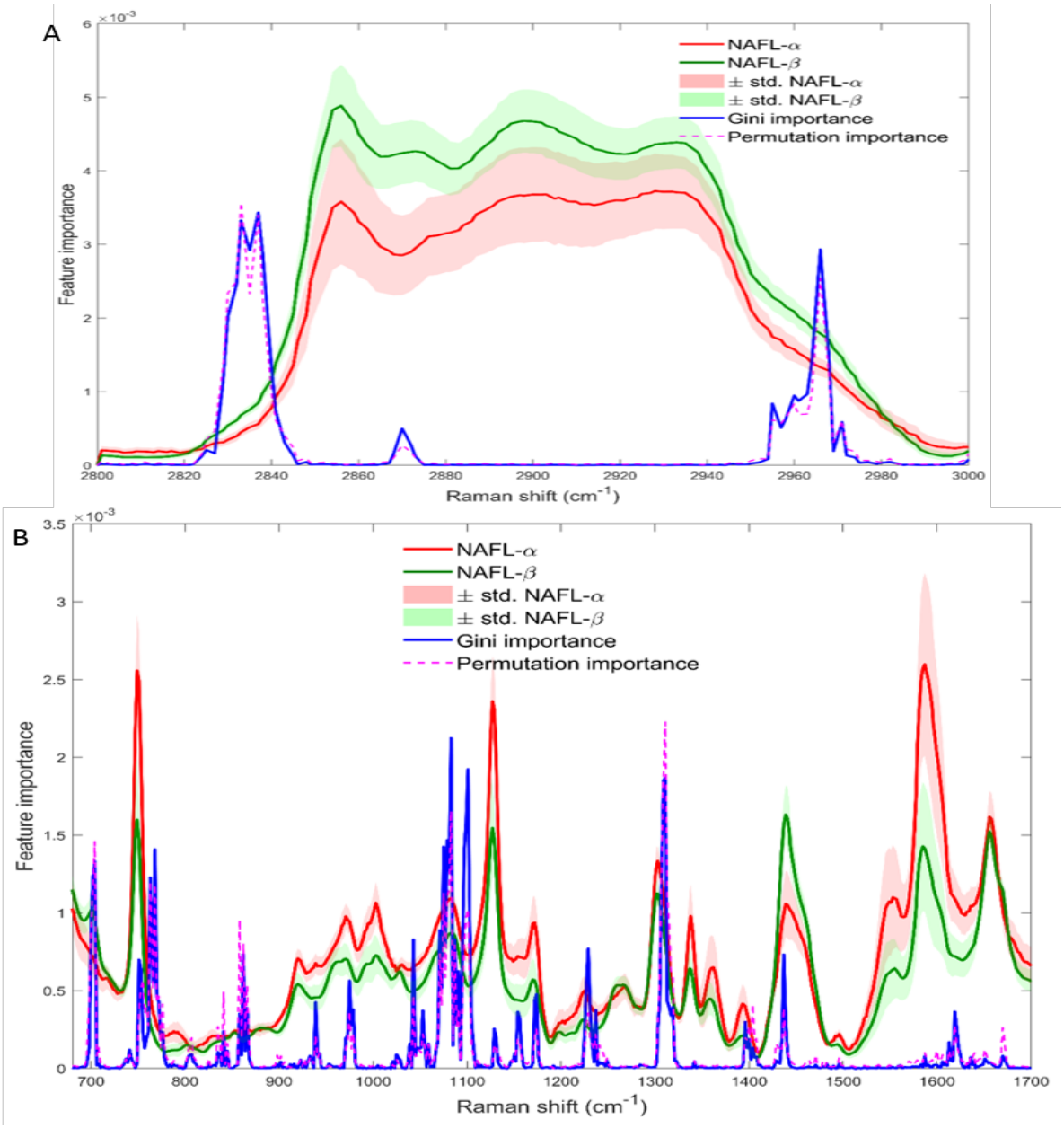


Figure 3.11: The Gini importance and permutation importance that quantify the importance of Raman shifts in differentiating NAFL-α and NAFL-β at (A) 2800-3000cm^{-1}, and (B) 700-1700 cm^{-1} regions, with the averaged Raman shifts of NAFL-α and NAFL-β and the standard deviations.

The spectral ranges of 2829-2838 cm^{-1} and 2964-2968 cm^{-1} can be attributed to Raman bands of the saturated fatty acids myristic acid and palmitic acid [98]. Myristic acid and palmitic acid, having Raman peaks at 2832 cm^{-1} and 2967 cm^{-1}, respectively, are reported to be increased in NASH [100]. Myristic acid also has a Raman peak at 2869 cm^{-1}, whose importance measures in Fig. 6 (see also Fig. S10) are relatively large among those in the region from 2850 cm^{-1} to 2950 cm^{-1} (the overlap caused by standard deviations ceases to exist at 2869 cm^{-1}). The accumulation of the saturated fatty acids

in the liver causes cell toxicity and plays important roles in NASH developing from NAFL [8,101].

In the two mean spectra of NAFL-α and NAFL-β (Figs. 3.10 and 3.11), significant Raman peaks exist at 750, 1127, 1441, 1591, 1655 cm^{-1}, with different behavior for NAFL-α and NAFL-β. For example, the Raman peaks at 1591 cm^{-1} (vitamin A), 750 cm^{-1}, and 1127 cm^{-1} (cytochrome c in mitochondria) are suppressed in NAFL-β compared to NAFL-α. However, these Raman shifts are not classified as being relatively important Raman shifts in differentiating NAFL-α and NAFL-β, because their intensity variation overlaps between NAFL-α and NAFL-β (see Fig. 3.11).

Raman peaks at 700-704 cm^{-1}, assigned to cholesterol [98], are also found to be relatively important for the differentiation between NAFL-α and NAFL-β. The mean intensity of the Raman peaks in NAFL-β is significantly larger than those in NAFL-α, and, as shown in Fig. 3.11, intensity variations at these wavenumbers do not overlap between NAFL-α and NAFL-β. This indicates that relative cholesterol presence is different in NAFL-β and NAFL-α, so that this Raman shift can be utilized for differentiating these NAFL groups. In NAFLD/ NASH patients, excess cholesterol intake influences the progression of the disease state [102,103]. Intrahepatic cholesterol is more abundant in NASH patients than healthy subjects [104]. Cholesterol is reported to be increased in hepatocytes, Kupffer cells and hepatic stellate cells [105]. When the cholesterol content of the mitochondrial membrane increases, production of cytotoxic reactive oxygen species is increased, leading to liver injury in NASH [106]. Thus, it is suggested that cholesterol in the livers affects NASH development. Note that NAFL-β comes from the HFHC diet group having higher lipid accumulation, and consequently the degree of similarity between NAFL-β and NASH is higher (in terms of lipid accumulation) compared to NAFL-α. Given the time course of the disease progression of the HFHC groups, the NAFL-β was regarded to be a nascent state of NASH, although morphological features have not yet emerged. In addition, note that the silent region 1801-2800 cm^{-1} was mostly found to be irrelevant for the differentiation (Fig. 3.14), which manifests the validity of this ensemble machine learning approach.

Fig. 3.12 visualized a plane made by the ISOMAP components from the five most important Raman shifts of the spectra from NAFL groups. The Raman banda are at 2837 cm^{-1}, 2833 cm^{-1}, 2966 cm^{-1}, 2835 cm^{-1} and 2838 cm^{-1}. From this projected scatter

plot of the spectra, we confirmed that NAFL-α and NAFL-β are well classified in terms of a small number of Raman shifts.

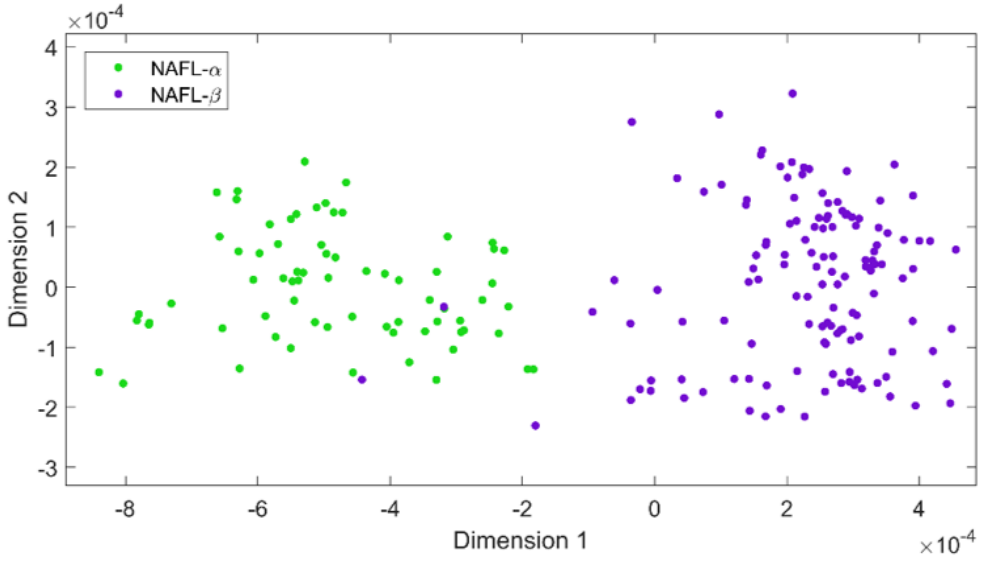


Figure 3.12: Scatter plot of the first two ISOMAP components built from the five most important Raman shifts of the spectra from NAFL groups.

The figures, from Figs. 3.13A to 3.13F, respectively, give the spatial intensity distributions of the of the 15 NAFL images at Raman shifts of 749 cm^{-1} , 1127 cm^{-1} (associated with cytochrome c), 1591 cm^{-1} (associated with vitamin A), 2837 cm^{-1} (relatively most important Raman shift), 2856 cm^{-1} (associated with lipid), and 2966 cm^{-1}. We see in Fig. 3.13D that the spatial intensity distribution of the NAFL-α images at 2837 cm^{-1} is relatively bluer than those of NAFL-β, with some spatial variation, which indicates that at these Raman shifts (associated with lipid), i.e., there is higher lipid accumulation in NAFL-β compared to NAFL-α. In Fig.3.13F, we observe that at the Raman shift 2966 cm^{-1}, the spatial intensity distribution of the NAFL-α images is also bluer than that of NAFL-β. The relative contrast in intensity between NAFL-α images to NAFL-β images are higher at the Raman shifts 2837 cm^{-1} and 2966 cm^{-1} compared to the relative contrast of these two groups of images at Raman shifts 749 cm^{-1}, 1127 cm^{-1}, 1591 cm^{-1}, and 2856 cm^{-1} (Figs. 3.13A, 3.13B, 3.13C, and 3.13E). This result validates the relative importance of the Raman shift 2837 cm^{-1} as the most important features to distinguish NAFL-α and NAFL-β. Furthermore, the Raman shifts 749 cm^{-1}, 1127 cm^{-1} (associated with cytochrome c), 1591 cm^{-1} (associated with vitamin A), and 2856 cm^{-1} are measured as less important by random forest, because, at those Raman shifts, there is no

clear contrast between the two groups of images, and consequently, standard deviations of all the averaged spectra from each 10 pixels × 10 pixels image patches overlaps to each other.

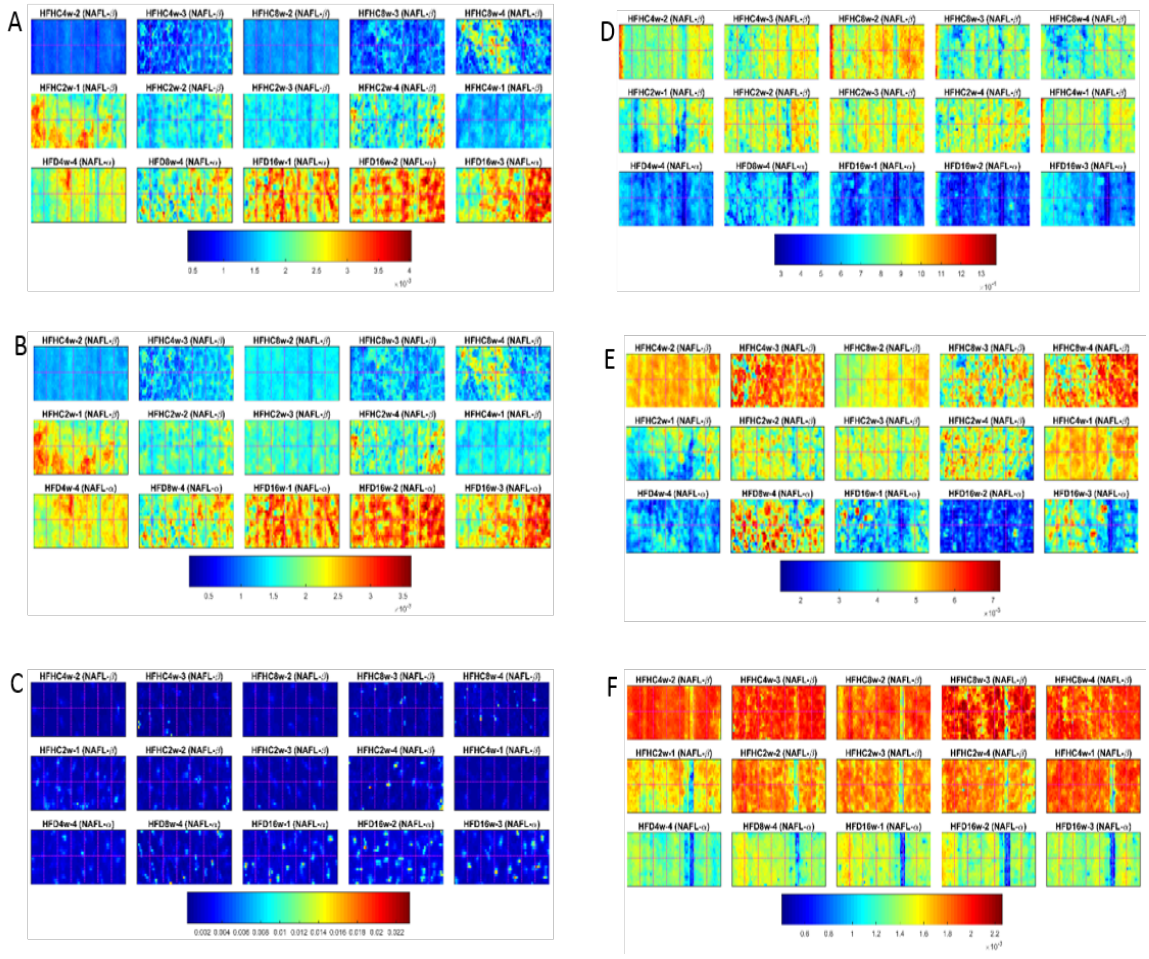


Figure 3.13: The spatial intensity distribution of the 15 NAFL images (NAFL-α and NAFL-β) at the Raman shifts, (A) 749 cm^{-1}, (B) 1127 cm^{-1}, (C) 1591 cm^{-1}, (D) 2837 cm^{-1}, (E) 2856 cm^{-1} and (F) 2966cm^{-1}. Colors denote the intensity of each pixel at these wavenumbers. Here, the first row from the left to the right: HFHC4w-2, HFHC4w-3, HFHC8w-2, HFHC8w-3, HFHC8w-4 and the second row from the left to the right: HFHC2w-1, HFHC2w-2, HFHC2w-3, HFHC2w-4, HFHC4w-1 (all assigned as NAFL-β), and the third row from the left to the right: HFD4w-4, HFD8w-4, HFD16w-1, HFD16w-2, HFD16w-3 (all assigned as NAFL-α).

3.2.5.1 Assessments of Insensitivity of the Methods of Feature Importance, Irrelevancy of Silent Region and Appropriateness of Relative Importances of Raman Shifts

In this paper, we have quantified the important wavenumbers in differentiating NAFL-α and NAFL-β from the Raman data by using two measures, i.e., Gini importance and permutation importance, both derived from ensemble machine learning random forest. In this section, we compare the results of these two measures and confirm that our important wavenumbers were insensitive irrespective of which measure we employed. Furthermore, we visualize the scatter plot in a space spanned by the top two important wavenumbers with the corresponding Raman images. Fig. 3.14 shows the Gini and permutation feature importances in differentiating NAFL-α and NAFL-β, where the blue solid line represents the Gini importance and the red dashed line represents the permutation importance. The sets of relatively important Raman shifts are mostly the same in the two measures. Note also that the silent region 1801-2800 cm^{-1} was almost irrelevant for the differentiation in both importance measures, indicating the validity of this ensemble machine learning approach for feature selection.

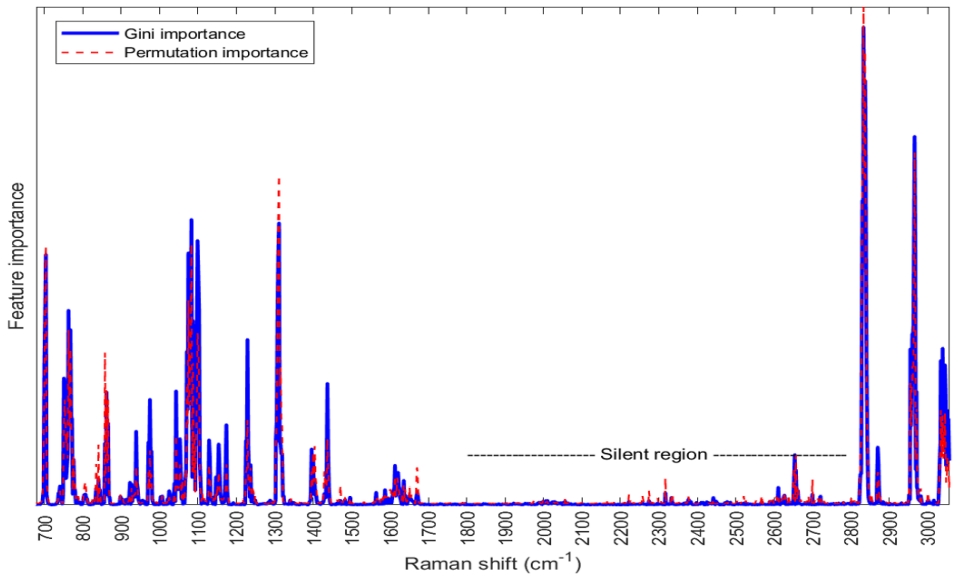


Figure 3.14: The feature importances for NAFL-α vs NAFL-β by two measures of importance: Gini importance (blue solid line); permutation importance (red dashed line): In order to compare the extents of the two different feature importances, the total extents of the feature importances over the wavenumbers are normalized.

The scatter plot of the Gini and permutation feature importances is shown in Fig 3.15, which shows a strong correlation (the Pearson correlation coefficient is 0.968). Especially

in both importance measures, the first and second important Raman shifts, 2837 cm^{-1} and 2833 cm^{-1} (arising from lipid), seem to be more important than the others in the degree of the importance.

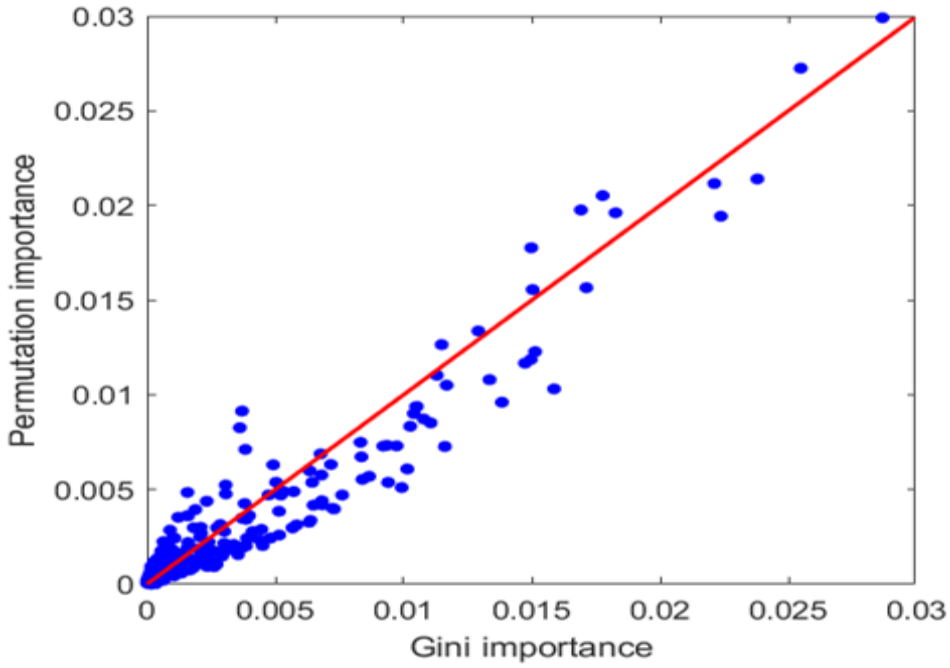


Figure 3.15: The scatter plot of the two feature importances: correlation coefficient = 0.9678.

3.2.5.2 Quantification of Spectral Variations in Raman Images.

There are various kinds of cells, such as hepatocytes, hepatic stellate cells, and white blood cells, composing the liver tissue. In our analysis, we segmented the Raman images into 10 pixels × 10 pixels patches (1 pixel is 5 μm), and computed the averaged spectrum over each patch. This is aimed at not only reflecting the difference of chemical environment at the typical size of hepatocytes and reducing the fluctuation in Raman spectra arising from other cells smaller than hepatocytes, but also reducing contributions of cosmic rays, shot noise, and other contaminations by the spatial averaging. To know how variable are the spectra acquired from each rat liver within each Raman image, and how does the internal spectral variation differ from that of other rats histologically assigned as normal or NAFL, and NASH, we computed the internal and external variation of NAFL images.

We calculated the internal spectral variation $\overline{d}_{int}(I)$ as the median of all pairwise Euclidean distances among the spectra belonging to a particular image I, and the external spectral variation $\overline{d}_{ext}(I)$ from the image I to the other images in its histologically assigned same group, i.e., either normal, NAFL, or NASH, as the median of all the pairwise

Euclidean distances among all the spectra in the image to all the spectra from the other images. We also calculate the 68% confidence interval of both $\overline{d}_{int}(I)$ and $\overline{d}_{ext}(I)$ (we simply omit the argument I hereinafter). The two-dimensional points $(\overline{d}_{int}(I), \overline{d}_{ext}(I))$ provide information about variable spectral behavior within a particular image of an individual versus variable spectral behavior in the same group.

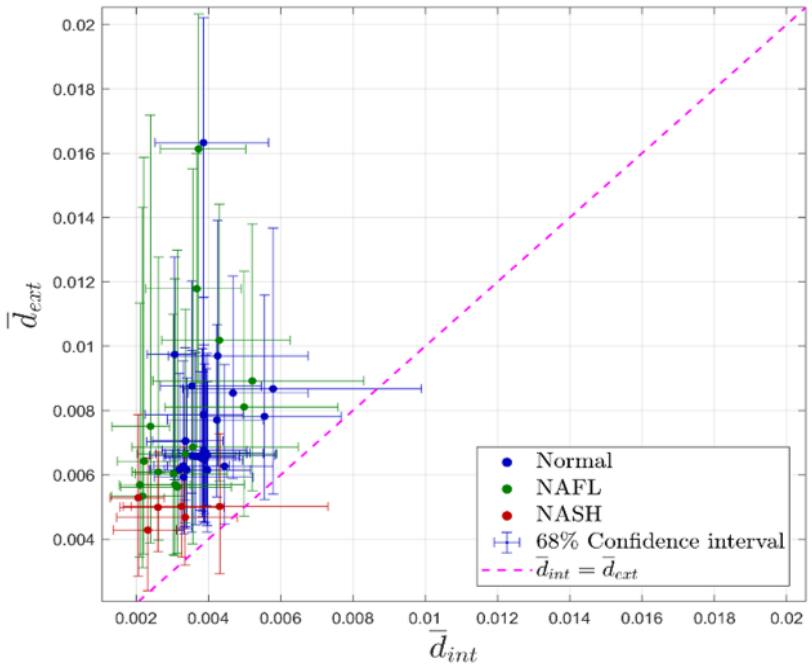


Figure 3.16: Scatter plot of the spectral variation of Raman images $(\overline{d}_{int}(I), \overline{d}_{ext}(I))$ from normal (blue dots); NAFL (green dots) and NASH (red dots), with the 68% confidence intervals of the corresponding median (bars with the same color). Magenta dashed line represents the locations where the internal, and external spectral variations are equal to each other, i.e.,$(\overline{d}_{int}(I) = \overline{d}_{ext}(I)$

Fig. 3.16 shows the 2D scatter plot of $(\overline{d}_{int}(I), \overline{d}_{ext}(I))$ for each image from the three histological groups: normal; NAFL and NASH, with the 68% confidence intervals. As the overall trend, the internal variation $\overline{d}_{ext}(I)$ of each image is smaller than its external variation $\overline{d}_{ext}(I)$ to all other images of the corresponding group. This indicates that for all histological states the internal spectral variations within each image of an individual rat (that may arise from fluctuation of microchemical environment including cells smaller than hepatocytes, the size of image patches, etc.) are smaller on average than the external spectral variation arising from different individuals in the same group. This implies that internal spectra variations are less significant compared to those from rat-to-rat variation over the individual rats.

Among these three histological states, NASH group has smaller external spectral

variations compared to normal and NAFL groups. The NASH group shows the smallest rat to rat variation, followed by the normal and NAFL groups. Likewise, the internal spectral variations of NASH are also relatively small compared to the other states. These are all consistent with Fig. 2 in which the distance among the group of NASH tends to be uniformly smaller than those among the other group (much bluer in the figure). On the contrary, of these quantities, normal tissue and NAFL have the large rat to rat variation, partially reflecting the observed heterogeneity of the NAFL condition [8].

3.3 Concluding Remarks

The diagnostic approach 'Raman imaging along with machine learning techniques', presented in this chapter, can differentiate the states of NAFLD in rats. Our approach split NAFL groups into two subgroups NAFL-α and NAFL-β, and it well distinguished NAFL-β from NAFL-α; it predicted the appearance of NASH after only two weeks feeding on a high-fat, high-cholesterol diet before histological signatures emerge, identifying a nascent state of NASH. Furthermore, the random forest classifier extracted the most relevant set of spectral features which led to the discrimination of NAFL-α and NAFL-β efficiently with high accuracy.

Chapter 4

Raman Microscopic Histology using Unsupervised Machine Learning

This chapter is concerned with the Raman microscopic histology of NAFLD using a combination of methods from machine learning and information theory. First, the extended simple linear iterative clustering (SLIC) is explained to segment Raman hyperspectral images. Then, we describe the clustering algorithm based on the rate determining theory (RDT), and determine the most appropriate number of the clusters of spectra by considering the Poisson error (shot noise) originating from the Raman scattering and background photons. The agglomerative hierarchical clustering (AHC) is introduced for the classification of the state of tissues using the information obtained from the cluster populations. Based on these techniques, we discuss our obtained results, and compare with the histological assessments.

4.1 Materials and Methods

As described in section 3.1 in chapter 3, a total of 48 liver tissues were collected after 2, 4, 8, and 16 weeks from $16(= 4 \times 4)$ rats fed with each diet (standard diet (SD), high-fat diet (HFD), and a high-fat high-cholesterol diet (HFHC)) for histological and Raman spectroscopic analyses. Animal model and sample preparation, Raman spectroscopic measurement, histological staining and evaluations of liver tissues including histological results are described in section 3.1 in chapter 3.

4.1.1 Preprocessing the Raman Image Data of NAFLD

In this study, after bias and baseline (background and autofluorescence contamination) correction using the methods described in section 3.1.5.1 in chapter 3, to further increase the signal to noise ratio in the spectra without degrading the spatial feature of the Raman images, we extended the ideas of simple linear iterative clustering (SLIC) [53,54], a superpixel segmentation algorithm designed for three channel images, to accomodate Raman hyper-spectral images having any number of channels (wavenumbers). Details of extended SLIC algorithm are explained in the next section.

Tissue Segmentation using Superpixel Approach

After bias and baseline correction, we segmented the Raman tissue images into superpixels (comprise of pixels having similar characteristics [55]) to further reduce possible remaining contributions of noise (superpixels reduce the susceptibility to noise and outliers, capturing image redundancy [112,113], to speed up the analysis.

A superpixel is a group of contiguous pixels having similar properties which forms irregularly shaped grids in 2D image. Achanta et. al. [53,54] presented Simple Linear Iterative Clustering (SLIC) algorithm to create superpixels in 2D images by iteratively performing local k-means clustering of pixels starting from regular grids based on gradient ascent method until some convergence criteria is satisfied. Instead of considering only spatial proximity as in regular griding, here the pixels are grouped in consideration of both the spatial proximity and color similarity. SLIC performs a local grouping of pixels in the 5-D space: three values L, a, b from CIELAB color space for color similarity and two values x, y from pixel coordinates of the image for spatial proximity [53].

Inspired by the approach of Achanta et al. [53], we have extended SLIC algorithm to generate superpixels in Raman hyper-spectral images. Instead of using only 5-D space as in SLIC, we use $(W+2)$-dimensional space defined by Raman intensities $\{I_w(i)\}$ in the spectral dimensions, and pixel positions x_i and y_i in the spatial dimensions, i.e., the i-th pixel in hyperspectral image γ_i is represented by $\gamma_i = [I_1(i), I_2(i), \ldots, I_w(i), \ldots, I_W(i), x_i, y_i]$ where $w(= 1, 2, \ldots, W)$ denotes the w-th wavenumber measured.

We calculate the set of superpixels $\mathbf{\Gamma} = \{\Gamma_k\}$ $(k = 1, 2, ..., N_{\text{sp}})$ by first fixing the number of superpixels per image N_{sp}, then grouping the N pixels in the image to form

approximately equal size superpixels through iteration of a local k-means clustering of pixels in $(W+2)$-D space, starting from initial guess of superpixels $\{\Gamma_k^{(0)}\}$ sampled on a regular spatial grid of size S pixels $\times$ S pixels. Here the grid step S is defined as $(N/N_{\mathrm{sp}})^{1/2}$ which corresponds to the division of a square $N^{1/2} \times N^{1/2}$ into a set of N_{sp} smaller squares $S \times S$.

The center of each superpixel $\Gamma_k^{(l)}$ at the l-th iteration, $\hat{\Gamma}_k^{(l)}$, is defined by

$$\hat{\Gamma}_k^{(l)} = \left[\hat{I}_1^{(l)}(k), \hat{I}_2^{(l)}(k), ..., \hat{I}_W^{(l)}(k), \hat{x}^{(l)}(k), \hat{y}^{(l)}(k)\right] = \frac{1}{N_k^{(l)}} \sum_{j=1}^{N_k^{(l)}} \left[I_1(j), I_2(j), ..., I_W(j), x(j), y(j)\right] \tag{4.1}$$

as the mean vector of all pixels belonging to corresponding superpixel $\Gamma_k^{(l)}$, where $N_k^{(l)}$ is the number of pixels belonging to the superpixel centered at $\hat{\Gamma}_k^{(l)}$, and is initialized as $S \times S$ for all k. We then estimate the closeness of each pixel γ_i to each superpixel $\Gamma_k^{(l)}$ by defining the following distance;

$$D(\gamma_i; \hat{\Gamma}_k^{(l)}) = \left[\left(\frac{d_{\mathrm{spectral}}(\gamma_i; \hat{\Gamma}_k^{(l)})}{\max\limits_{\Gamma_k^{(l)} \in \mathbf{\Gamma}^{(l)}} \left[\max\limits_{\gamma_i \in \Gamma_k^{(l)}} d_{\mathrm{spectral}}(\gamma_i; \hat{\Gamma}_k^{(l)}) \right]} \right)^2 + \left(\frac{d_{\mathrm{spatial}}(\gamma_i; \hat{\Gamma}_k^{(l)})}{\max\limits_{\Gamma_k^{(l)} \in \mathbf{\Gamma}^{(l)}} \left[\max\limits_{\gamma_i \in \Gamma_k^{(l)}} d_{\mathrm{spatial}}(\gamma_i; \hat{\Gamma}_k^{(l)}) \right]} \right)^2 \right]^{1/2}, \tag{4.2}$$

where the spectral distance $d_{\mathrm{spectral}}(\gamma_i; \hat{\Gamma}_k^{(l)})$ and the spatial distance $d_{\mathrm{spatial}}(\gamma_i; \hat{\Gamma}_k^{(l)})$ between the pixel γ_i and the superpixel center $\hat{\Gamma}_k^{(l)}$ are given, respectively, by

$$d_{\mathrm{spectral}}(\gamma_i; \hat{\Gamma}_k^{(l)}) = \left(\sum_{w=1}^{W} \left(I_w(i) - \hat{I}_w^{(l)}(k) \right)^2 \right)^{1/2}, \tag{4.3}$$

and

$$d_{\mathrm{spatial}}(\gamma_i; \hat{\Gamma}_k^{(l)}) = \left(\left(x_i - \hat{x}_k^{(l)} \right)^2 + \left(y_i - \hat{y}_k^{(l)} \right)^2 \right)^{1/2}. \tag{4.4}$$

As written above, we start by placing initial guess of superpixel as regular grids of size $S \times S$, and define a search region as $2S$ pixels $\times$ $2S$ pixels around the center of each superpixel. We measure the distance between superpixel centers to each of the pixel γ_i lying in the search region around that center by the distance metric defined in Eq. 4.2, and then each pixel γ_i is assigned to its nearest superpixel center. Note that the search region also includes all γ_i within $S \times S$ initial guess, and if the closest superpixel center from the γ_i in the sense of Eq. 4.2 is not that of the initial guess, those γ_i will belong to

the other superpixel. After all the isolated γ_i are assigned into the closest superpixels, i.e., one iteration is finished, we update the set of each superpixel, the number of pixels within each k-th superpixel $N_k^{(l)}$, and recalculate the superpixel centers by Eq. 4.1. We then compute the residual error E as the Euclidean distance between the locations of current and previous superpixel centers. We iterate the SLIC algorithm 20 times or until the locations of superpixels converge within a 1% tolerance, whichever occurs first. We reassign some disconnected pixels (which appeared mostly in cases in which iteration was terminated without convergence) to the largest neighboring superpixels. In this paper, number of superpixels N_{sp} was chosen so that $S = 3$, i.e., each superpixel contains approximately 9 individual pixels (one pixel size is 5 μm). We also note that this procedure may not be appropriate for very large W, as differences among the spectral distances obtained from Eq. 4.3 would converge in a very high dimension.

Using this extended SLIC algorithm, we generated superpixels containing an average of 9 idividual pixels. We chose SLIC instead of a rectangular gridding scheme to better preserve spatial features in the Raman images, as illustrated in Fig. 4.1A; 2-dimensional image frames representing the mean intensities at each pixel over all wavenumbers are shown for an original Raman image (left), averaged 3×3-pixel rectangular grid (middle), and extended SLIC superpixel segmentation (right). Visual inspection demonstrates that, extended SLIC superpixel segmentation preserves the structural features of the original image better than regular gridding. Fig. 4.1B represents the average Poisson error (from all the Raman images) vs. wavenumber; single-pixel case (red) and after superpixel segmentation (blue). We see that extended SLIC reduced the average error in the spectral dimension by a factor of approximately 3 when compared to the single-pixel case, as shown in Fig 4.1B.

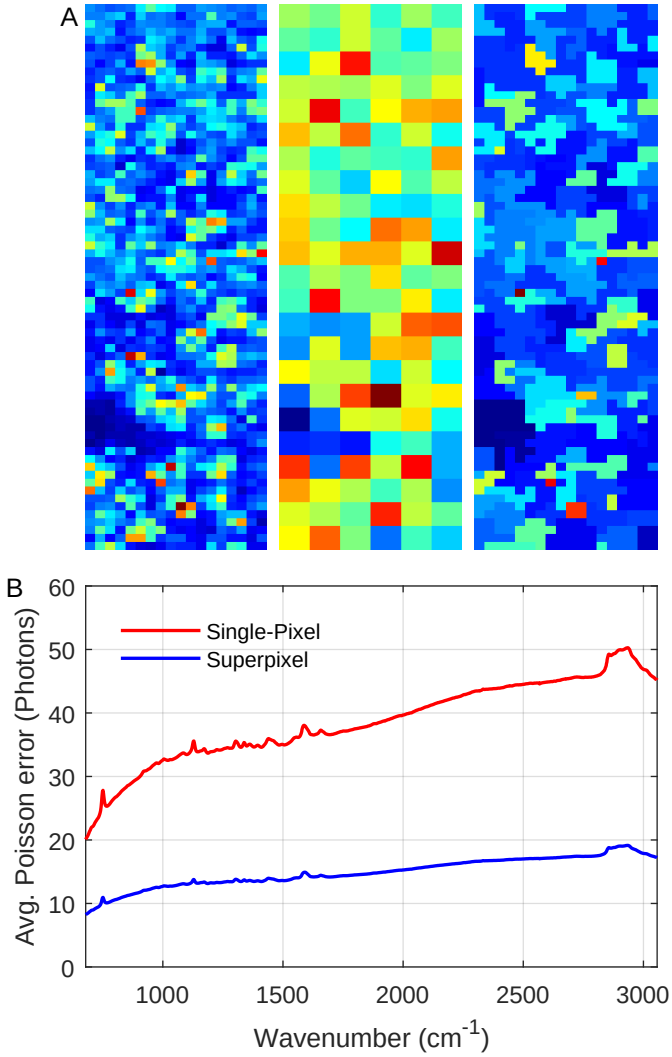


Figure 4.1: *(A) 2-dimensional images representing the mean intensities at each pixel over all wavenumbers for an original Raman image (left), averaged* 3×3*-pixel rectangular grid (middle), and extended SLIC superpixel segmentation (right). (B) Average Poisson error (from all the Raman images) vs. wavenumber; single-pixel case (red) and after superpixel segmentation (blue).*

4.2 Clustering the NAFLD data

After performing the preprocessing schemes described in section 4.1.1, we obtained a total of 7728 superpixel spectra in the 48 images, with each superpixel spectrum corresponding to the averaged spectrum over the individual pixels assigned to the superpixel. Before performing the clustering algorithm, each superpixel spectrum was area normalized (see section 3.1.5.4 in chapter 3); note that intensity in the silent region, $1801 - 2800$ cm^{-1}, in which there is no Raman signal from biomolecules, was excluded from all analyses. To

identify groups of spectra having similar Raman information, thus similar biochemical environment in the Raman images, we performed unsupervised clustering of the spectra with rate-distortion theory (RDT) [56,57], which has demonstrated good performance in situations involving detection noise and overlap among clusters [95,114]. The details of RDT clustering are explained in the following section

4.2.1 Rate-Distortion Theory (RDT) based Clustering

Clustering [115,116] is an unsupervised learning algorithm which divides the data (spectra) into groups (clusters) having the spectra with similar characteristics (similar Raman information) quantified by, e.g., L_2 distance between spectra. Since clustering method reduces the spectra into smaller number of representative sets (clusters), it compresses and distorts the information of the original data that results compression and distortion [114,117,118]. In this work, we employed an information theory ([56,57]) based clustering by using rate distortion theory (RDT) which classifies the spectra into different groups having similar Raman characteristics. RDT [56,57] mainly addresses this phenomenon "compressing the data keeping an acceptable level of distortion". Suppose that the set of N spectra is given by $\mathbf{S} = \{s_w(1), s_w(2), ..., s_w(N)\}$ and the set of N_c clusters is given by $\mathbf{C} = \{C_1, C_2, ..., C_{N_c}\}$. The RDT clustering algorithm clusters the sets of spectra $\mathbf{S}$ into set of clusters $\mathbf{C}$ through the minimization of the functional $\mathcal{F}[p(C_k|s_w(i))]$ with respect to the conditional probability $p(C_k|s_w(i))$ [95,114,117,119–121] defined by

$$\mathcal{F}[p(C_k|s_w(i))] = I(\mathbf{C};\mathbf{S}) + \beta \langle D(\mathbf{C},\mathbf{S})\rangle, \tag{4.5}$$

where the compression $I(\mathbf{C};\mathbf{S})$ known as rate is the average amount of information needed to specify the spectra $s_w(i)$ within the set of clusters $\mathbf{C}$ [114,117] which is defined as

$$I(\mathbf{C};\mathbf{S}) = \sum_{k=1}^{N_c}\sum_{i=1}^{N} p(C_k|s_w(i))p(s_w(i))\ \log\frac{p(C_k|s_w(i))}{p(C_k)}, \tag{4.6}$$

where $p(C_k|s_w(i))$ is the conditional probability of belonging the spectrum $s_w(i)$ to the cluster C_k, $p(s_w(i))$ is the probability of occurring the spectra $s_w(i)$, and $p(C_k)$ is the marginal probability of the cluster C_k; $\langle D(\mathbf{C},\mathbf{S})\rangle$ is the mean *distortion* among the spectra defined as the mean of the pairwise distance between all pairs of spectra within the set

of clusters $\mathbf{C}$, averaged over all clusters, which is expressed as [95,117,122]

$$\langle D(\mathbf{C},\mathbf{S})\rangle = \sum_{k=1}^{N_c} p(C_k)\left[\sum_{i,j=1}^{N} p(s_w(i)|C_k)p(s_w(j)|C_k)d_{i,j}\right] \tag{4.7}$$

and the parameter β controls the softness of the clustering as a trade-off between the rate and distortion.

The input parameters for RDT algorithm is both the number of class N_c and β. We choose β so that RDT provides a hard clsutering producing all the conditional probabilities $p(C_k|s_w(i))$ approaching 0 or 1. For a fixed β, minimizing the functional $\mathcal{F}$ in Eq. 4.5 at different values of N_c, we can obtain a set of possible models to describe the data, i.e., $\langle D(\mathbf{C},\mathbf{S})\rangle$ and $I(\mathbf{C},\mathbf{S})$ can be used as the model selection tools to choose the number of clusters N_c [117].

4.2.2 Error Propagation and Model Selection

Selection of an appropriate number of groups/clusters is a difficult part of any cluster analysis. Akaike information criterion [123], Bayesian information criterion [124], minimum description length principle [125], and other model selection approaches [95,114] are frequently used to find an appropriate model. Here we use a procedure analogous to the procedure described in [95,114], with which we find the smallest number of clusters describing the data with goodness-of-fit that is within a tolerance of the maximum allowed by measurement error.

4.2.2.1 Estimated Error of the Data

Taking the mean distortion, i.e., the average intracluster distance across all clusters in the model, returned by RDT (Eq. 4.7) to be a goodness-of-fit parameter, then the model returning the maximum achievable goodness-of-fit (minimum distortion) has each spectrum in its own cluster. For this model, if the data contains no error then the distortion is zero. On the other hand, if the data contain errors, nonzero distortion will arise from the errors, as there is uncertainty in the measured spectra. We estimate the Poisson errors in the single-pixel Raman measurements with a normal approximation, and use error propagation for sums of normally distributed random variables [126–128]

to obtain errors in the single-pixel spectra, and subsequently in the superpixel spectra.

After bias correction each pixel in the image frame at particular wavenumber contains a photon count I_0, which is the sum of two independent Poisson processes: Raman signal from the tissue I_w, and background contamination B_w, which is estimated from the raw Raman spectra by recursive polynomial fitting, as described in section 3.1.5.1 in chapter 3. Therefore, the observed intensity I_0 of each pixel at a particular wavenumber w can be expressed as the sum of these independent component processes, i.e., $I_0 = I_w + B_w$. After estimating B_w, we estimate I_w as $I_w = I_0 - B_w$. We are interested in the uncertainty of our estimation of I_w, ΔI_w. Using normal error propagation we obtain,

$$\begin{aligned}\Delta I_w^2 &= \left(\frac{\partial I_w}{\partial I_0}\right)^2 \Delta I_0^2 + \left(\frac{\partial I_w}{\partial B_w}\right)^2 \Delta B_w^2 \\ &= \Delta I_0^2 + \Delta B_w^2,\end{aligned}$$

where ΔI_w, ΔI_0, and ΔB_w are, respectively, the uncertainty of I_w, I_0, and B_w. Since I_0 and B_w are Poisson variables, $\Delta I_0^2 = I_0$ and $\Delta B_w^2 = B_w$. So the error (propagation uncertainty) of the Raman signal I_w is estimated as

$$\Delta I_w = (I_0 + B_w)^{1/2}. \tag{4.8}$$

The superpixel intensity at wavenumber w is the mean of the single-pixel intensities, $s_w = \frac{1}{m}\sum_{i=1}^{m} I_w(i)$, where m is the number of pixels inside a superpixel. Therefore, the error Δs_w in a superpixel (arising from Poisson noise) at the wavenumber w is calculated by

$$\Delta s_w = \frac{1}{m}\left(\sum_{i=1}^{m} \Delta I_w^2(i)\right)^{1/2}. \tag{4.9}$$

4.2.2.2 Exploring the Optimal Model using Distortion Cut-off

We select an appropriate number of clusters by defining 'distortion cutoff', which in our case is the amount of distortion arising from photon counting error. As discussed above, if the data is error-free, then the best fitting model has equivalent numbers of clusters and spectra, and the distortion vanishes; however, if there is uncertainty in the data then a nonzero distortion arises.

We estimate this distortion arising from error through incorporation of the uncertainties as calculated from Eqs. 4.8 and 4.9 as a normal approximation to a Poisson process, where for each Raman signal s_w (the mean intensity of a superpixel at a particular wavenumber w), a new signal is sampled from a normal distribution $\mathcal{N}(s_w, \Delta s_w)$. By randomly sampling at all wavenumbers, we generate a new spectral realization $s_w^1(i)$ ('1' denotes the first realization) for the original i-th superpixel spectrum $s_w(i)$. This new spectrum is a possible realization of the original spectrum considering uncertainty due to Poisson error.

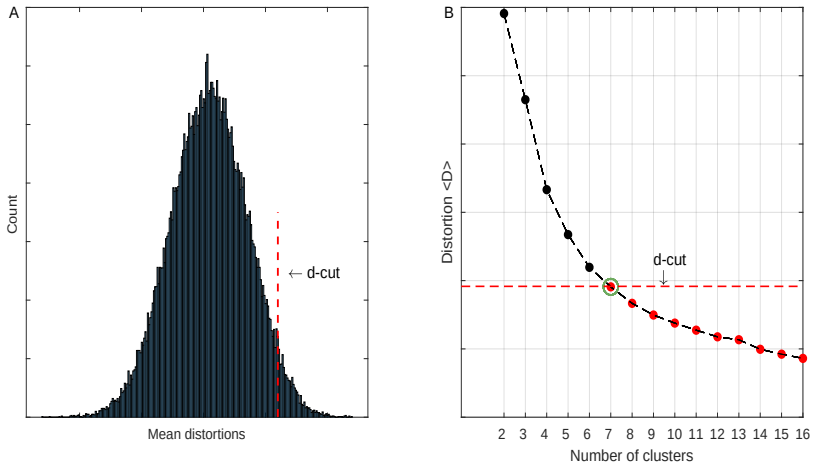


Figure 4.2: *A) Distribution of mean distortion due to error; B) Mean distortion $\langle D \rangle$ due to clustering versus the number of cluster. Dashed red line (red) represents distortion cut-off.*

Repeating this procedure generates another possible realization of the original spectrum $s_w^2(i)$, and we then calculate the pairwise Euclidean distance (Eq. 3.2 in chapter 3) denoted by d_i, thereby generating a distance for the i-th superpixel spectrum that may arise from the effects of Poisson error. Performing the same procedure for all N superpixel spectra, we generate the distances $d_i, i = 1, 2, ..., N$, and estimate the mean distortion arising from error over all superpixel spectra as $\langle d \rangle^1 = \sum_{i=1}^{N} p(i)\ d_i$, where $p(i)$ is weight of i-th superpixel spectrum defined by m_i/M where m_i, and M denote the number of pixels within i-th superpixel and the total pixels of the image, respectively.

The entire is then repeated a number of times (nb) to obtain $\langle d \rangle^1, \langle d \rangle^2, ..., \langle d \rangle^{nb}$, thus generating a distribution of mean distortions due to Poisson error, which is shown in Fig. 4.2A. We chose the 95% upper confidence bound on the distribution to be the 'distortion cutoff' at which the quality of the best fitting model cannot be separated from the distortion arising from the error. The distortion cutoff is displayed as a red dashed

line in Figs. 4.2A and 4.2B.

In the RDT framework, with a given parameter β, the number of clusters is determined within the allowed error in the data (see section 4.2.1 for details). Here β was chosen so that RDT produces a hard clustering; we performed RDT for different numbers of clusters, and computed the mean distortion $\langle D \rangle$ (defined in Eq. 4.7) due to clustering of the data. The plot of the mean distortion versus the number of clusters is shown in Fig. 4.2B. The model having the largest mean distortion $\langle D \rangle$ that is less than the distortion cutoff is chosen for further analysis, as it is the simplest model that achieves goodness-of-fit within the range of that allowed by Poisson error. This model has 7 clusters and is marked with a green circle in Fig. 4.2B, each of the clusters contains similar Raman characteristics with significant differences among clusters that are larger than those that may arise from Poisson error in a sense of L_2 distance.

4.3 Agglomerative Hierarchical Clustering (AHC)

AHC is a bottom-up clustering algorithm [129] which starts by assigning each data point (cluster population per tissue in our case) initially as individual singleton cluster, and then recursively combines (agglomerate) the closest pair of clusters by defining some proximity criteria until all data points belong to a single cluster. The results of AHC are visualized graphically by a multilevel hierarchy of clusters as a tree which is also called dendrogram. Starting from the bottom layer having each data point as individual cluster, the two clusters having the smallest pairwise distance are merged, and is represented by a horizontal line (binary tree) in the dendrogram. Likewise, treating the merged cluster as new singleton cluster, a new merge is obtained from these two clusters having smallest pairwise distances. This procedure repeats until all the data points lie into one cluster.

In the dendrogram, each of horizontal bar represents individual merging, and y-coordinates of those bars represent the distance between two clusters that were merged, and the data points are viewed as singleton clusters along x-axis (at the leaves of the dendrogram). By cutting the dendrogram at some threshold points, we can obtain different disjoint clusters. AHC is informative for displaying the data into different clusters and the dendrogram visualizes the hierarchical structure inside the cluster by reconstructing the history of the merges in the clustering.

In our data we choose average linkage technique to define the cluster proximity which measures the distance between two clusters as the average pairwise distance among all pairs of data points in two different clusters:

$$d_{12} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \text{dist}(x_i, y_j) \tag{4.10}$$

where $x_1, x_2, ..., x_{n_1}$ are the data points from cluster 1 and $y_1, y_2, ..., y_{n_2}$ are the data points from cluster 2 and $\text{dist}(x_i, y_j)$ is the L_1 distance between the points x_i and y_j.

Silhouette Coefficient

As stated in the main text, we measured the cluster average silhouette [130] to find the quality of grouping obtained by AHC on cluster population distribution. For a fixed number of groups of the data points (tissues), obtained by AHC, cluster average silhouette of all the points of that data set can evaluate the degree/strength of separation between groups which eventually give the suggestion about how appropriate the number of groups is. Comparing the cluster average silhouette obtained by different number of tissue groups, we can choose the model giving the highest average silhouette.

Silhouette coefficient for any point p in the data set is computed by the following equation:

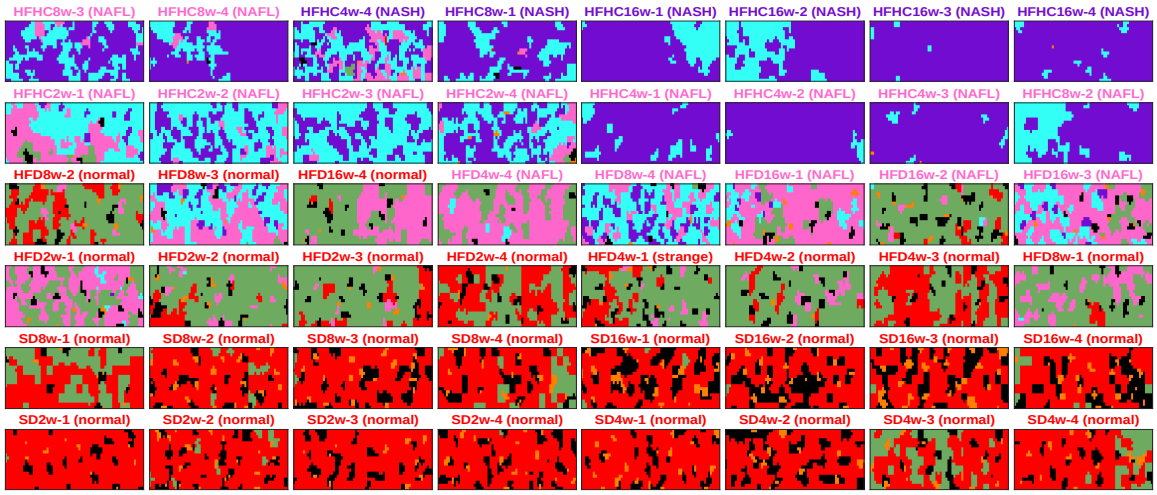
$$sil(p) = \frac{d_{ext}(p) - d_{int}(p)}{\max(d_{int}(p), d_{ext}(p))}, \tag{4.11}$$

where $d_{int}(p)$ is the average L_1 distance of p from all the data points in the same cluster (internal distance/cohesion); where $d_{ext}(p)$ is the average L_1 distance of p from all the data points in the nearest cluster (external distance/separation). The values of silhouettes can vary from -1 to 1, where the value 1 indicates that the point p is far from the nearest cluster, the value 0 indicates that the point p is very close to the nearest cluster. The average cluster silhouette is computed by taking the overall average of all silhouette coefficients obtained from all the data points.

4.4 Results and Discussions

4.4.1 Clusters are Distributed across the Tissues

We have extracted 7 clusters, each of which contains a set of similar superpixel spectra that corresponds to different parts/regions within the liver tissue images having similar biochemical environments.



(A) Cluster maps

(B) Cluster population

Figure 4.3: (A) *Cluster maps of 7 clusters across all the images. Each of two consecutive rows at the bottom, middle and top are, respectively, from SD, HFD, and HFHC diet model. The diet (SD, HFD, HFC) and histological (normal, NAFL, NASH) states of each of the images are labeled at the top of each image; (B) The population distributions of 7 clusters in each of the images. The images are arranged in similar order as cluster maps and the clusters are ordered from left to right according to increasing order of lipid content as described in Fig. 4.4a below.*

The spatial distributions of these 7 clusters across the tissue images, termed as 'cluster

maps' hereinafter, are shown in Fig. 4.3A with 7 colors, each corresponding to a different cluster. The cluster maps visualize the distributions of the 7 clusters representing the underlying chemical environments across the images.

Fig. 4.3B shows the population distributions of each of the clusters within each of the Raman tissue images. The clusters are ordered from left to right according to increasing relative lipid intensities in the corresponding cluster mean spectra (shown in Fig. 4.4A below), and the colors correspond to those used in the cluster maps in Fig. 4.3A. Visual inspection of Fig. 4.3A shows that most clusters are not unique to a single diet or histologically assigned group, but are allocated across different diets and histological states. Additionally, Fig. 4.3B shows that, although the clusters are distributed across multiple diet/histological states, the population distributions vary from state to state. From Figs. 4.3A and 4.3B, we can observe that, on average, purple and cyan clusters ('cluster 7' and 'cluster 6') are highly populous in the HFHC diet model, while red, black and orange clusters ('cluster 3', 'cluster 2', and 'cluster 1') are more populous in the SD model. Though the green ('cluster 4') and pink ('cluster 5') clusters are the most populous groups, all clusters possess nonzero populations in the HFD diet model. This is in contrast to the SD and HFHC diet models, indicating a larger degree of microchemical diversity in the HFD dietary model of liver tissue.

The average spectra of the individual 7 clusters are shown in Fig. 4.4A, where the solid lines (with colors corresponding to those used in Fig. 4.3) represent the mean spectra of the clusters, and the shaded areas corresponding to $\pm$ one standard deviation. As described above, the clusters are labeled as 'cluster 1' to 'cluster 7' according to the increasing order of relative intensity in the corresponding mean spectra at the integrated lipid-rich wavenumnber bands $1425 - 1473$ cm^{-1} and $2801 - 2971$ cm^{-1}, corresponding to CH_2 moieties [98]. We note that an identical ordering is found if we arrange the clusters according to decreasing relative intensity of vitamin A regions ($1591 - 1600$ cm^{-1}, $1156 - 1160$ cm^{-1} and $1195 - 1200$ cm^{-1} [40]). That is, the cluster containing the highest intensity in the mean spectrum at lipid-rich regions is numbered as 'cluster 7' (purple color) while the cluster containing the lowest intensity in the mean spectrum at lipid regions is labeled 'cluster 1' (orange color). The other clusters, black, red, green, pink, cyan, are respectively labeled 'cluster 2' to 'cluster 6' according to ascending order of lipid intensity.

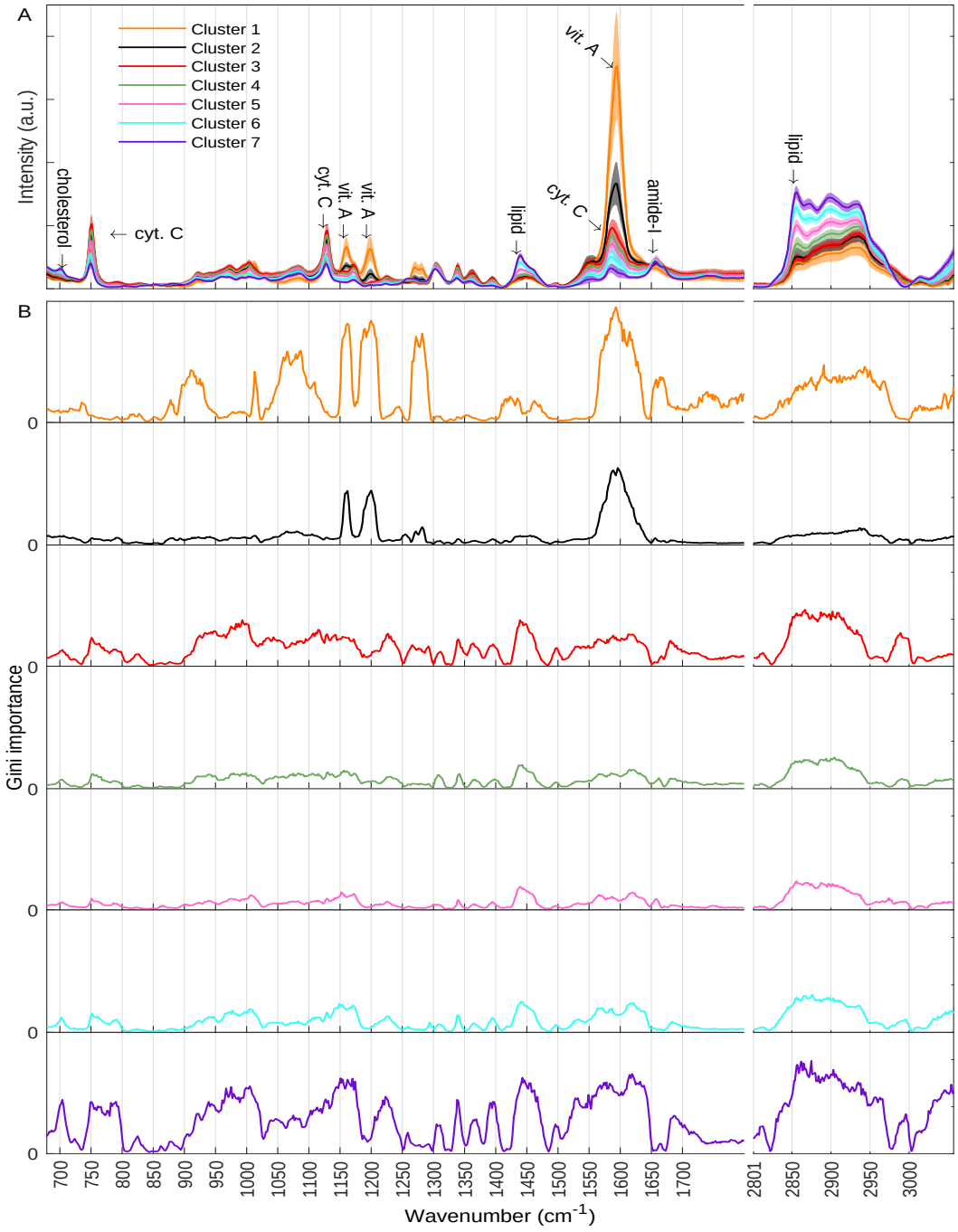


Figure 4.4: *(A) Mean spectra of* 7 *clusters with* ±*one standard deviations (shades patch). The spectra colored as orange, black, red, green, pink, cyan and purple are, respectively, numbered from* 1 *to* 7 *according to the increasing order of relative intensity at the integrated lipid-rich bands (integration of the wavenumbers* $1425 - 1473\ cm^{-1}$ *and* $2801 - 2971\ cm^{-1}$*); (B) Gini importance for each of the cluster relative to all other clusters (each of the* 7 *solid lines correspond to distinctive features of one cluster to all others). We used completely randomized features in decision trees of random forest to reduce bias in feature importances. The small gaps along wavenumber axes represent the silent regions* $(1801 - 2800\ cm^{-1})$ *which were cropped from the spectra.*

The detailed view of the mean cluster spectra (each in one subplot) is shown in Fig. 4.5.

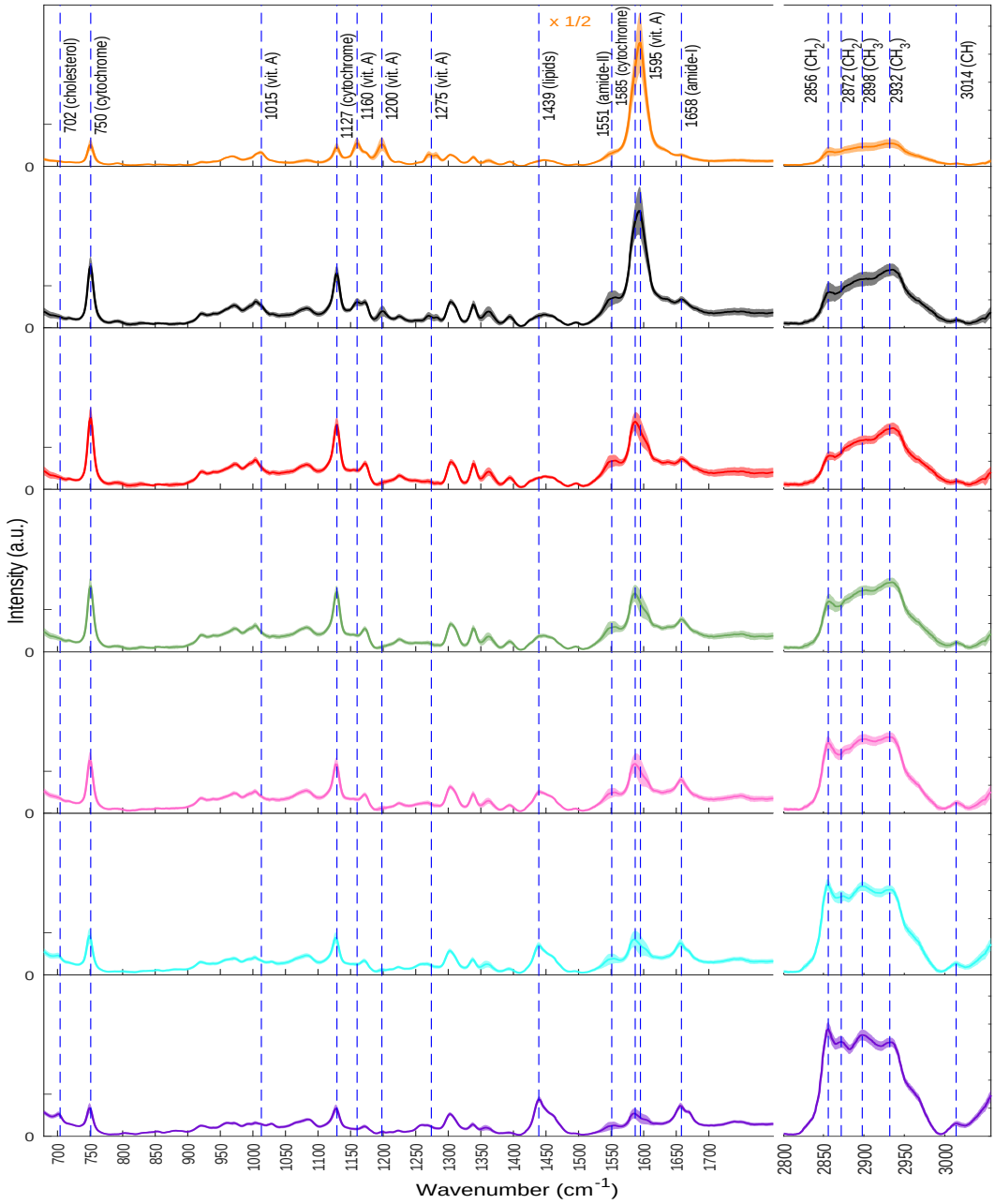


Figure 4.5: *Detailed of mean cluster spectra shown in Fig. 4.4 with ±one standard deviations (shades patch). The colors correspond to the 7 clusters. Due to very high intense peak at vitamin A (∼ 1591 cm^{-1}), the difference spectra for cluster 1 (orange) is scaled (multiplied) by 1/2 for visual clarity. The small gaps along wavenumber axes represent the silent regions. The significant Raman peaks are labeled with chemical assignments details of which are shown in Table 4.1.*

The mean spectra (Fig. 4.4A and 4.5) provide valuable information in understanding the chemical features of the clustering results. Significant Raman peaks are located at dis-

tinct wavenumber ranges associated to different chemical assignments such as cytochrome c, vitamin A, proteins, cholesterol, and different types of lipids. Representative peaks are located at the following wavenumber ranges: $700-705$ cm^{-1} (cholesterol), $740-760$ cm^{-1} (cytochrome c/heme), $997-1007$ cm^{-1} (phenylalanine; symmetrical ring breathing [4]), $1123-1133$ cm^{-1} (cytochrome c/heme), around 1156cm^{-1} and 1200 cm^{-1} (vitamin A), $1300-1311$ cm^{-1} (cytochrome c/heme), $1429-1460$ cm^{-1} (lipid), $1583-1590$ cm^{-1} (cytochrome c), $1591-1600$ cm^{-1} (vitamin A), $1653-1662$ cm^{-1} (C=C , amide-I), and $2833-2970$ cm^{-1} (lipids and proteins) [4,40,41,98,131]. See Table 4.1 for detailed assignment and tabulation of prominent peaks in the mean cluster spectra shown in Fig. 4.4A and 4.5 [4,40,41,98,131]:

Raman band locations (cm^{-1})	Chemical assignments
$700-705$	Cholesterol
$740-760$	Cytochrome/heme
~ 1015	Vitamin A
$1123-1133$	Cytochrome/heme
~ 1160	Vitamin A
~ 1200	Vitamin A
~ 1275	Vitamin A
$1429-1460$	Lipids
~ 1550	Amide-II
$1583-1590$	Cytochrome/heme
$1591-1600$	Vitamin A
$1653-1662$	amide-I
~ 2856	CH_2
~ 2872	CH_2
~ 2898	CH_3
~ 2932	CH_3
~ 3014	CH

Table 4.1: Wavenumber ranges for the prominent Raman peaks observed in the cluster mean spectra (Figs. 4.4A and 4.5).

To understand which wavenumber regions have significant contributions to the distinction of a particular cluster from other clusters, we estimated the importance of each wavenumber with the ensemble-learning-based random forest (RF) classifier [51,52] (see section 3.1.6.2 in chapter 3 for details). The estimation of feature importance was formulated as a *post hoc* classification in which the cluster labels for each spectrum output by RDT are input to a random forest classification as training labels. The random forest is then trained with the labeled spectra, and two different measures of feature importance,

the Gini and permutation importance measures [86,87] (see section 3.1.6.2 in chapter 3) are estimated in a *post hoc* manner.

Additionally, due to variation among the populations of the clusters, we note that the raw Gini importance measures have been rescaled such that the maximum importance, resulting from perfect classification at a single wavenumber, is unity. That is, to make the Gini importance measures as computed for the distinction of one particular cluster directly comparable to those generated for the distinction of other clusters, each measure is scaled according to the maximum possible value it can take for that particular cluster.

The Gini importance spectra, i.e., the Gini importance measures as a function of all wavenumbers, associated with all spectral clusters are shown in order of increasing lipid contribution from top to bottom in the subplots of Fig. 4.4B. To quantify differences in chemical contributions among the clusters, and to reinforce the results of feature importance estimation, we include the mean difference spectrum of each cluster as Fig. 4.6. Details of the calculation are provided in section 4.4.1.1. The shaded envelope in each of the subplots of Fig. 4.6 correspond to 68% confidence intervals around mean difference spectra. Positive and negative differences, respectively, represent increased and decreased Raman intensity of the chemical constituents associated to particular wavenumbers in the spectra assigned to the cluster.

The solid lines represent the cluster mean difference spectra while the shaded envelops correspond to the 68% upper and lower confidence bounds around mean difference spectra (colors of the lines correspond to the clusters used in Fig. 4.4A).

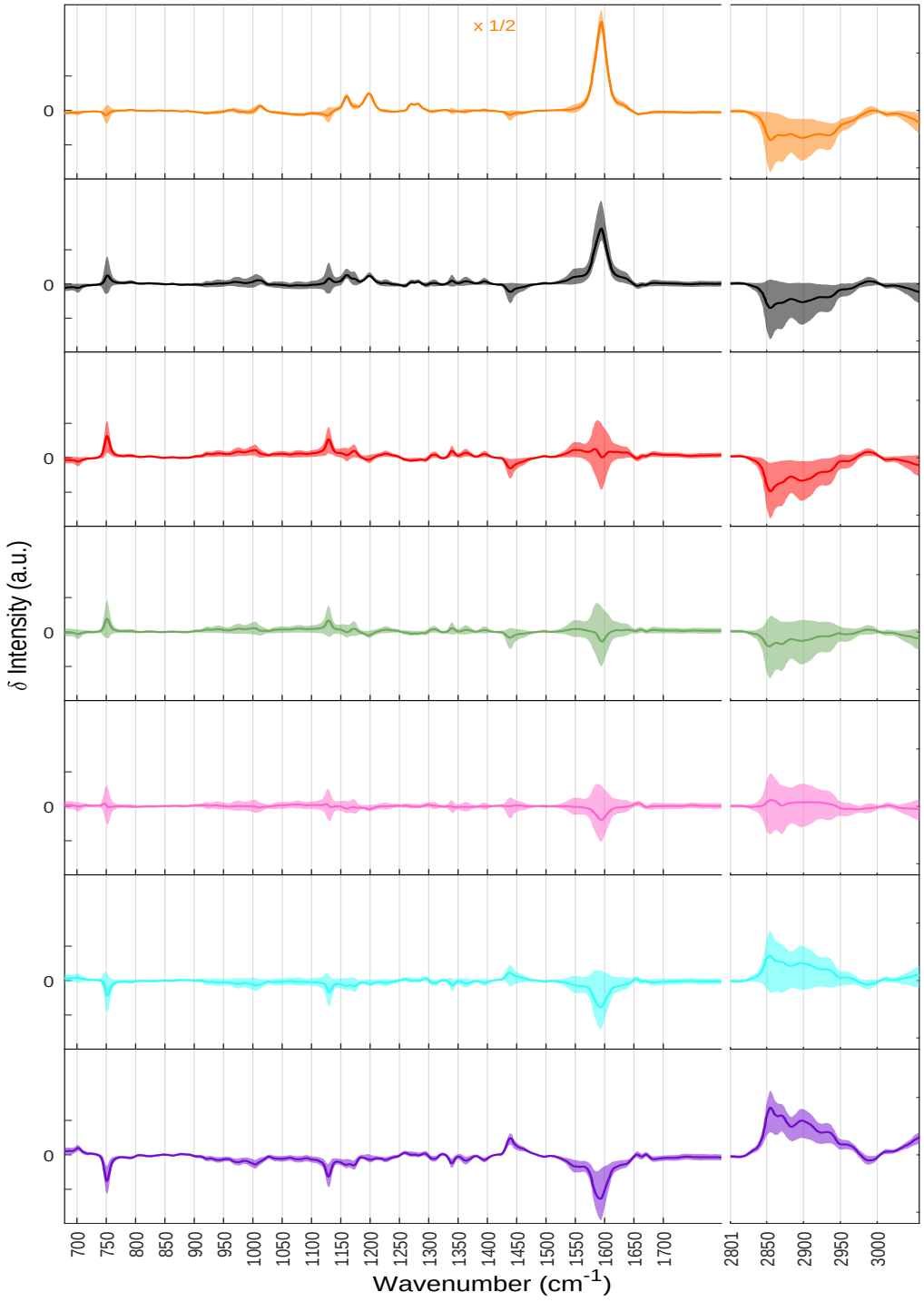


Figure 4.6: *Cluster mean difference spectra with* 68% *lower and upper confidence bound (shaded envelop). Due to very high intense peak at vitamin A* ($\sim$ 1591 cm^{-1})*, the difference spectra for cluster 1 (orange) is scaled (multiplied) by* 1/2 *for visual clarity. The small gaps along wavenumber axis represent the silent regions* ($1800 - 2800$ cm^{-1})*.*

Recent studies of liver tissue using Raman imaging have found that lipid profiles and vitamin A distributions in the liver tissue are two main features (biomarkers) in the

distinction of healthy and diseased tissue [40]. Accumulation of lipids due to high fat consumption is the first step in the pathogenesis of NAFLD [4]. The extent of the accumulation of triacylglycerols (TAGs) has been the basis for the quantification of severity of NAFLD [4,8]. Using information from these studies along with observations from Figs. 4.4, 4.3a and supporting Fig. S2, we may interpret the most likely associations between regions of the liver tissues and the spectral clusters.

Relative to those of other clusters, the mean spectra (Figs. 4.4A and 4.5) of cluster 1 (orange) and cluster 2 (black) possess remarkably high intensities in the 1590-1600 cm^{-1} wavenumber region associated with vitamin A [40]. Similarly high magnitude is observed in the same region of the Gini importance spectra (Fig 4.4B) corresponding to these two clusters, along with additional vitamin A contributions near $1015, 1160, 1200,$ and 1275 cm^{-1}, strongly suggesting that vitamin A is important to the distinction of these two clusters. Because Vitamin A is known to be mostly stored inside hepatic stellate cells (HSCs) in healthy tissue [40,41], we deduce that the regions of the tissue containing these two clusters are most likely associated with the quiescent HSCs. Considering the extremely high vitamin A intensities, regions of the tissues associated with cluster 1 (orange) most likely contain (regions of) HSCs with high concentrations of vitamin A while the regions associated with cluster 2 (black) may indicate (regions of) HSCs with less concentrated vitamin A.

We note that, while clusters 1 and 2 show small, disconnected regions of spatial contiguity in the cluster maps (Fig. 4.3A), clusters 3 through 7 show different behavior, having large regions of spatial contiguity that can span entire tissue samples. The Gini importance spectra of these clusters (3 to 7) are similar in shape and have large magnitudes in the wavenumber regions associated with lipids ($1425 - 1475,\ 2800 - 2900$ cm^{-1}), though significant contributions across the entire spectral range are also observed. As shown in Fig. 4.4A, the mean spectra of clusters 3 through 7 of similar shape but are marked by successive increases in intensity at wavenumbers associated with lipids accompanied by successive decrease in vitamin A and cytochrome ($745 - 755, 1125 - 1135, 1300 - 1315, 1580 - 1590$ cm^{-1}) regions. This trend is clearer in the difference spectra (Fig. 4.6), in which difference intensities are negative at wavenumbers associated with lipids for the clusters 3 (red) and 4 (green) and successively more positive for clusters 5 (pink), 6 (cyan), and 7 (purple). Oppositely, difference intensities at wavenumbers associated

with cytochrome are positive for clusters 3 and 4, near zero for cluster 5, and negative for clusters 6 and 7. These trends indicate increases in relative lipid contributions and decreases in relative cytochrome contributions, respectively, from cluster 3 to cluster 7. Also observed in cluster 7 is a slight positive contribution at ~ 702 cm^{-1}, suggesting an increase in cholesterol contribution in this cluster. Finally, each of clusters 3 through 7 exhibits zero or negative vitamin A intensity, which, when considered in concert with the large degree of spatial contiguity, suggests that these clusters represent regions of the tissues containing hepatocytes having various levels of lipid deposition.

Because cluster 3 (red) is most populous in SD and normal histological tissues, and is accompanied by numerous vitamin A rich (orange and black) regions, it most likely represents tissue regions containing healthy hepatocytes. Clusters 4 (green) and 5 (pink) are more populous in HFD and NAFL tissues, suggesting that they represent hepatocytes with increase lipid accumulation. Finally, clusters 6 (cyan) and 7 (purple) are most populous in tissues from the HFHC dietary group, containing both NAFL and NASH diagnoses, indicating that these clusters most likely represent hepatocytes having relatively high level of lipid accumulation. In the case of cluster 7, this accumulation is accompanied by an observable increase in cholesterol contribution and marked reduction of vitamin A, which has been attributed to significant growth in the number and size of lipid droplets in the progression of NAFLD [4].

4.4.1.1 Derivation of Cluster Mean Difference Spectra

The cluster mean difference spectra are obtained according to the following procedures: at a particular wavenumber w, the mean difference in intensities $\delta_k(w)$ between the spectra from cluster k to all the spectra from other clusters is obtained by the expression

$$\delta_k(w) = \frac{1}{N_k(N - N_k)} \sum_{i \in c_k} \sum_{j \in c_k^c} (s_w(i) - s_w(j))$$

where c_k is the set of all the superpixel spectra from cluster k, c_k^c is the set all spectra not belong cluster k; N_k is the total number of spectra in cluster k, and N is the total number of spectra in all the clusters. We have also computed 68% upper and lower confidence bounds around the mean difference spectrum obtained as piled from all the pairwise difference spectra.

4.4.2 Classifying the Liver Tissues based on Cluster Population

Because each of the 7 clusters represents a different underlying chemical environment, the cluster maps and populations shown in Fig. 4.3 indicate differences in biological and chemical compositions across the liver tissues and populations of clusters are not unique to a particular dietary model or histologically assigned group, suggesting that the cluster populations may be useful as a 'descriptor' of tissue state in distinguishing groups of tissues having different diets or different states of NAFLD. We use this information to investigate how the clustering results can be used to identify the groups of tissues in terms of statistical distances among population distributions of the 48 Raman images. Through the application of agglomerative hierarchical clustering (AHC) with average linkage algorithm [129] using the L_1 distance metric, we investigate the tissues and identify which of them are closest in terms of cluster population distributions. See section 4.3 for full details of AHC.

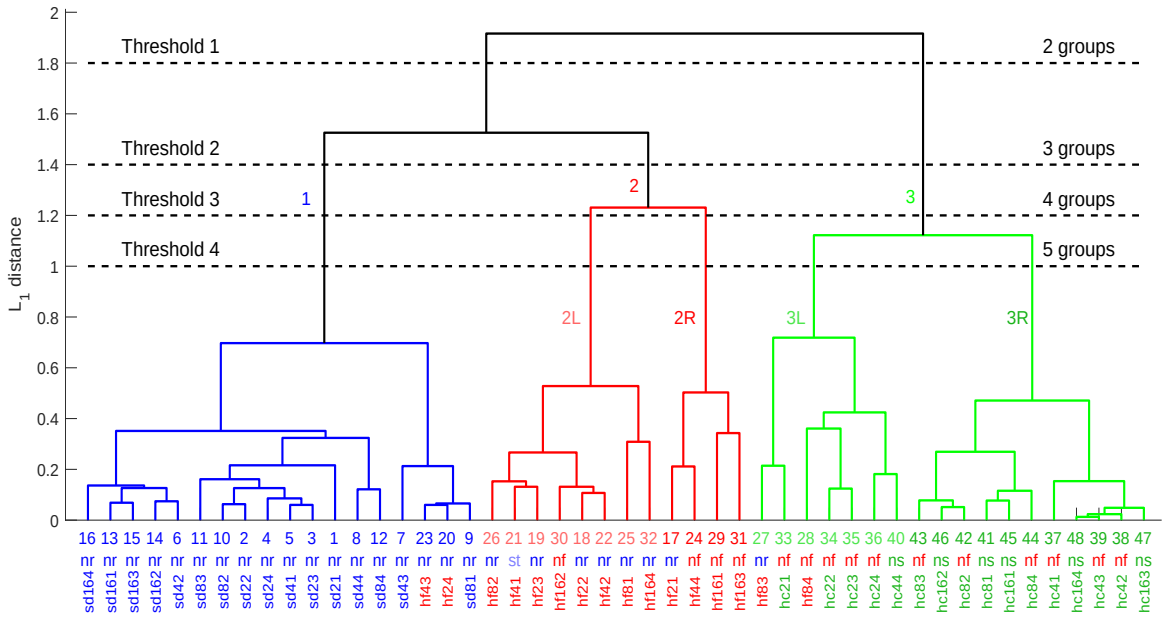


Figure 4.7: *Dendrogram resulting from AHC for tissue classification in terms cluster population. The images are numbered from 1 to 48 based on the diet order, i.e., SD images are labeled from 1 to 16, HFD images are labeled from 17 to 32, and HFHC images are labeled from 33 to 48. Diet and histological assignments are labeled at bottom of the leaf nodes. Here normal, NAFL and NASH histology are, respectively, represented by 'nr', 'nf' and 'ns' while standard diet, high fat diet and high fat high cholesterol diets are, respectively, represented by 'sd', 'hf' and 'hc', each of which followed by a numerical index representing the feeding weeks and rat's number (in last index), e.g., sd164 represents standard diet 16 week 4th rat etc. One tissue (HFD4w-1) with strange histology having mild fibrosis is represented by 'st'. We note that group 2R is nearly equidistant from groups 2L and 3L.*

The AHC results for the grouping of 48 tissues are presented in Fig. 4.7 as a dendrogram. An AHC dendrogram visualizes the results by showing how the individual tissues merge into groups. Single liver tissues (48 data points), viewed as initial singleton groups, are represented by nodes along the horizontal axis with corresponding indices, and diet and histological labels. Groupings among tissues are represented by horizontal connections on the vertical axis. The vertical positions of the horizontal connections represent the average L_1 distances (Eq. 4.10) at which two tissues or groups are merged to form a larger group.

Although AHC does not require a pre-specified number of groups, as the hierarchical behavior of AHC is adequately represented in the dendrogram, it is useful to analyze the properties of merged groups of tissues. Tissue groupings of various sizes and granularity can be obtained by merging the connected (groups of) nodes below a specified L_1 distance threshold, below which further branching is ignored. Fig. 4.7 indicates four such thresholds as dashed horizontal lines. To assess the quality of various numbers of groups we use the average cluster silhouette [130], a measure that compares the similarity within tissue groups to the similarity between adjacent tissue groups. Details of the cluster silhouette are discussed in section 4.3, and the values of the cluster silhouette as the number of tissue groups is increased is shown in Fig. 4.8; here the plot of the average cluster silhouette versus the number of tissue groups ranging from 2 to 8 obtained by AHC is shown.

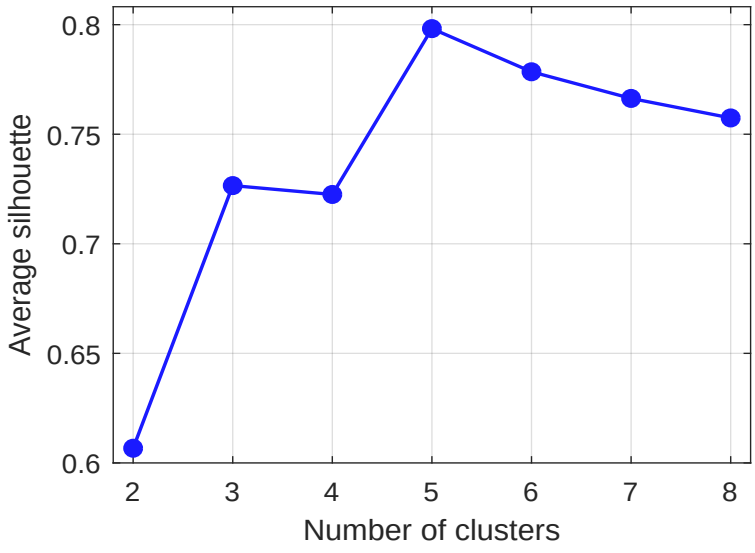


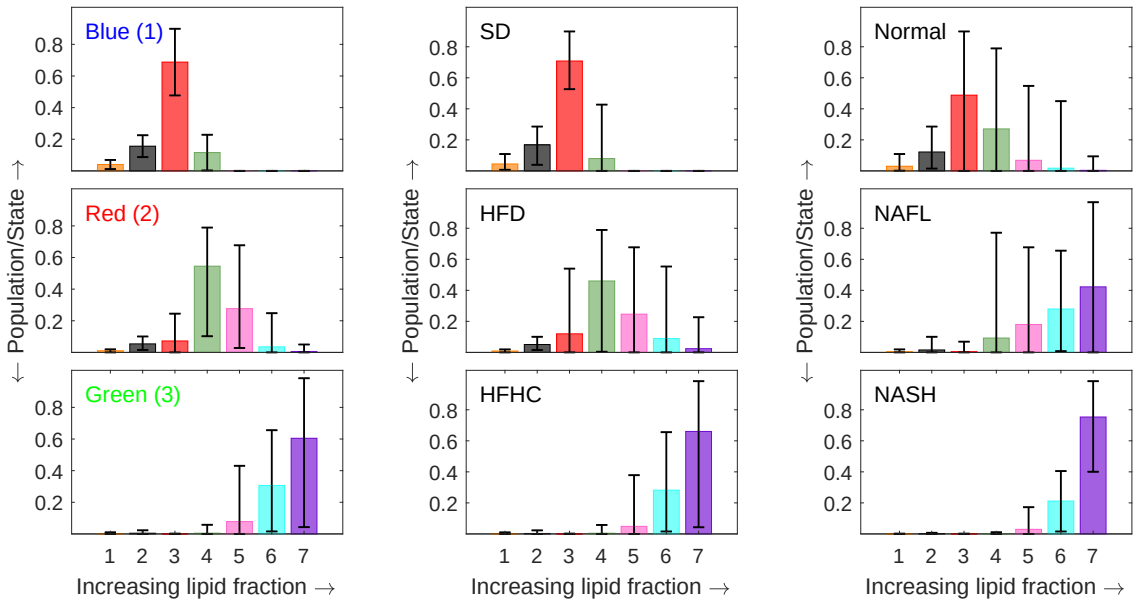
Figure 4.8: *The average cluster silhouette as a function of the number of tissue groups ranging from 2 to 8 by AHC.*

Though Fig. 4.8 indicates that five tissue groups produce the largest cluster silhouette, in our case it is advantageous to analyze multiple groupings as a means to better understand the hierarchical nature of the set of tissue samples. For the convenience, each of the 5 tissue groups is labeled as '1', '2L', '2R', '3L' and '3R' in Fig. 4.7.

Defining Threshold 1 to be $\sim 1.8\ L_1$ distance units, the dendrogram splits into two main branches, producing 2 groups of tissues. The group to the left contains all 16 SD tissues and 14/16 HFD tissues (red and blue labels), while the group to the right contains all 16 HFHC tissues and 2 HFD tissues. This result strongly suggests the presence of a distinct change in the Raman images of the liver tissues of the rats on the HFHC diet, relative to SD rats, and, to a lesser extent, HFD rats. It should be noted that, while the AHC results track nicely with diet, histological assignments are less precisely represented, with the group to the left containing tissues diagnosed as normal and NAFL, while the group to the right contains tissues diagnosed mostly as NAFL and NASH, but also one that was diagnosed as normal tissue.

Upon decreasing the L_1 distance threshold to $\sim 1.4\,L_1$ distance units (Threshold 2), the dendrogram splits into three main branches, yielding the blue, red and green groups, labeled as group 1, group 2, and group 3, respectively. As indicated in the labels corresponding to diet along the horizontal axis, this AHC results closely match the grouping by dietary models, with group 1 (blue), group 2 (red), and group 3 (green) each containing tissues belonging mostly to SD, HFD, and HFHC diet, respectively. Using the dietary models as a ground truth for AHC classification, we obtain 16/16 correct assignments for the SD model, 12/16 for the HFD model, and 16/16 for the HFHC diet model, yielding a classification accuracy by diet of 44/48 tissues, or 91.7%. In contrast, while all tissues in group 1 (blue) were histologically assigned to be normal liver tissue, groups 2 (red) and 3 (green) both contain relatively equal mixtures of histological assignments, with group 2 containing tissues assigned to be normal and NAFL, while group 3 contains mostly NAFL and NASH assignments.

The average population distributions of each of the 7 clusters of spectra (obtained by RDT clustering) with the 3-group AHC results, the dietary models, and histological states, are shown in Figs. 4.9A, 4.9B, and 4.9C, respectively. The clusters in Fig. 4.9 are arranged from left to right according to increasing lipid content. The error bars indicate the minimum and maximum population of a particular cluster within that par-



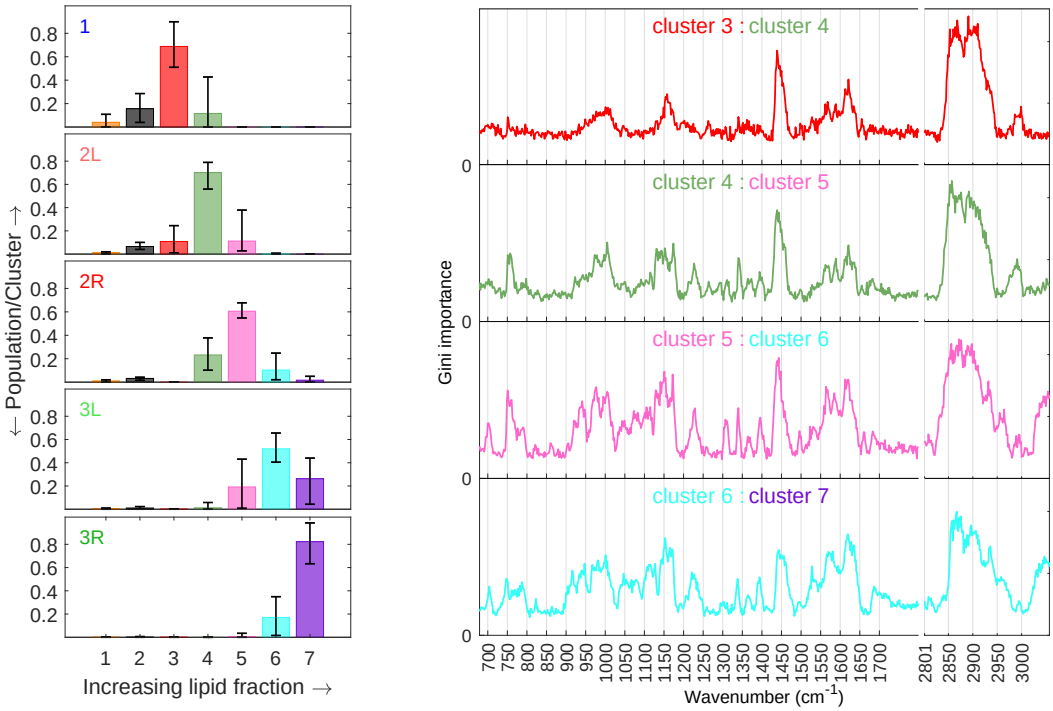
(A) 3-group AHC (B) Diet states (C) Histological states

Figure 4.9: *Average population distribution of each cluster within each: (A) 3-group AHC; (B) diet states; (C) histological sates. Error bar indicates the minimum and maximum population of a particular cluster within that state/AHC group.*

ticular state/AHC group, and colors of clusters correspond to those used in Fig. 4.3. Visual comparison of the average population distributions in Figs. 4.9A, 4.9B, and 4.9C confirms that 3-group AHC results show good alignment with dietary models, with the cluster populations of group 1 (blue) and SD, group (red) 2 and HFD, and group 3 (green) and HFHC, being comparable to one another. Again in contrast to the dietary model, the cluster populations of the histological model do not align with 3-group AHC results, suggesting that the AHC results are more indicative of dietary behavior than of histologically assigned NAFLD state. To further corroborate these results, we also compare cluster populations through a similarity score, which reflects the similarity of a particular tissue to a particular ground truth state, such as the SD dietary state. See section 4.4.3 for complete details. We find that the AHC results are mostly consistent with those obtained by the similarity score plot shown in supporting Fig. 4.13.

We note that in the 3-group AHC results, the group 2R is nearly equidistant from groups 2L and 3L. This is reflected in the next decrease of the AHC threshold to $\sim 1.2\, L_1$ distance units (Threshold 3), producing 4 groups in which group 2 (red) has further divided into two subgroups, group 2L and group 2R, while group 1 and 3 remain unchanged. As suggested above in the discussion of the cluster maps of Fig. 4.3A, this result is further indicative of a larger degree of diversity in group 2, and thus the HFD

dietary model, considering group 2 is fully composed of HFD tissues. Groups 2L and 2R contain 8 and 4 tissues, respectively, with 6/8 group 2L tissues having been assigned normal histology while 3/4 group 2R tissues were diagnosed as NAFL. This result suggests a loose association with histological assignments, however no firm conclusions can be drawn owing to the small sample size.



(A) 5-group AHC (B) Gini importance

Figure 4.10: *(A) Average population distribution of RDT clusters within each 5-group AHC, error bar indicates the minimum and maximum population of a particular cluster within a particular 5-group AHC; (B) Gini importance for each pair of adjacent clusters from cluster 3 (red) to cluster 7 (purple).*

For the final grouping having the maximum cluster silhouette, we produce five tissue groups according to Threshold 4, with group 3 (green) being further divided into two subgroups: groups 3L and 3R. The results of the 5-group AHC model are summarized in Fig. 4.10. As shown in Fig. 4.10A, the average population distributions for each of the 5 tissue groups display a notable progression in their shapes, with the population centers beginning to the left of the figure with group 1, and moving incrementally to the right for groups 2L, 2R, 3L, and 3R, toward clusters having higher lipid contributions. In tracking the progression of the most populous spectral cluster in each tissue group, we observe a sequential progression, as cluster 3 (red) is the most populous in group 1, cluster 4 (green) in group 2L, cluster 5 (pink) in group 2R, cluster 6 (cyan) in group 3L,

and cluster 7 (purple) in group 3R. As discussed in section 4.4.1, spectral clusters 3 to 7 are most likely hepatocytes of containing progressively increased lipid, and, in the case of clusters 6 and 7, cholesterol accumulation. This result suggests then, that each tissue group in the 5-group AHC is a coarse-grained region in the progression of healthy liver tissue to more diseased states like NAFL and eventually NASH.

To better understand the biochemical features underlying this sequential behavior, we computed the Gini importance spectra for each pair of adjacent spectral clusters from cluster 3 (red) to cluster 7 (purple), shown in Fig. 4.10B. Accompanying Fig. 4.10B is Fig. 4.11, which displays the mean difference spectra between the adjacent pairs of spectral clusters, and Fig. 4.12, in which adjacent pairwise difference spectra are superimposed to highlight changes in the progression from cluster 3 to cluster 7.

The top panel of Fig. 4.10B, showing the Gini importance spectrum for the cluster 3 : cluster 4 adjacent pair, indicates that the main contributions to the distinction of these two spectral clusters are in wavenumber regions corresponding to lipid contributions ($1425-1475\ \mathrm{cm}^{-1}$, $2800-2900\ \mathrm{cm}^{-1}$). The corresponding difference spectrum (Fig. 4.11) indicates an increase in intensity in these regions, suggesting an increase in relative lipid concentration in moving from cluster 3 to 4. Similar behavior is observed in moving from cluster 4 to cluster 5, with the additional observation of decreased cytochrome intensity. Although similar behavior is again observed in moving from cluster 5 to cluster 6, in that increased lipid and decreased cytochrome contributions are again observed, there are distinct changes in the shapes of the Gini importance and mean difference spectra. Most notably, peaks appearing in the Gini importance spectrum at $702, 1671$, and $2958\ \mathrm{cm}^{-1}$ and having positive intensity in the mean difference spectrum indicate an increase in contribution of cholesterol and/or cholesterol esters in moving from cluster 5 to cluster 6. Additional peaks appears near 2870 and $2932\ \mathrm{cm}^{-1}$, indicating increased presence of either fatty acids or triacylglycerols. This distinctive behavior is most obvious in supporting Fig. 4.12, in which the difference spectra in question are superimposed in the same axes. Lastly, moving from cluster 6 to cluster 7 displays similar behavior intensity, indicating similar chemical characteristics at further increased relative concentrations [98].

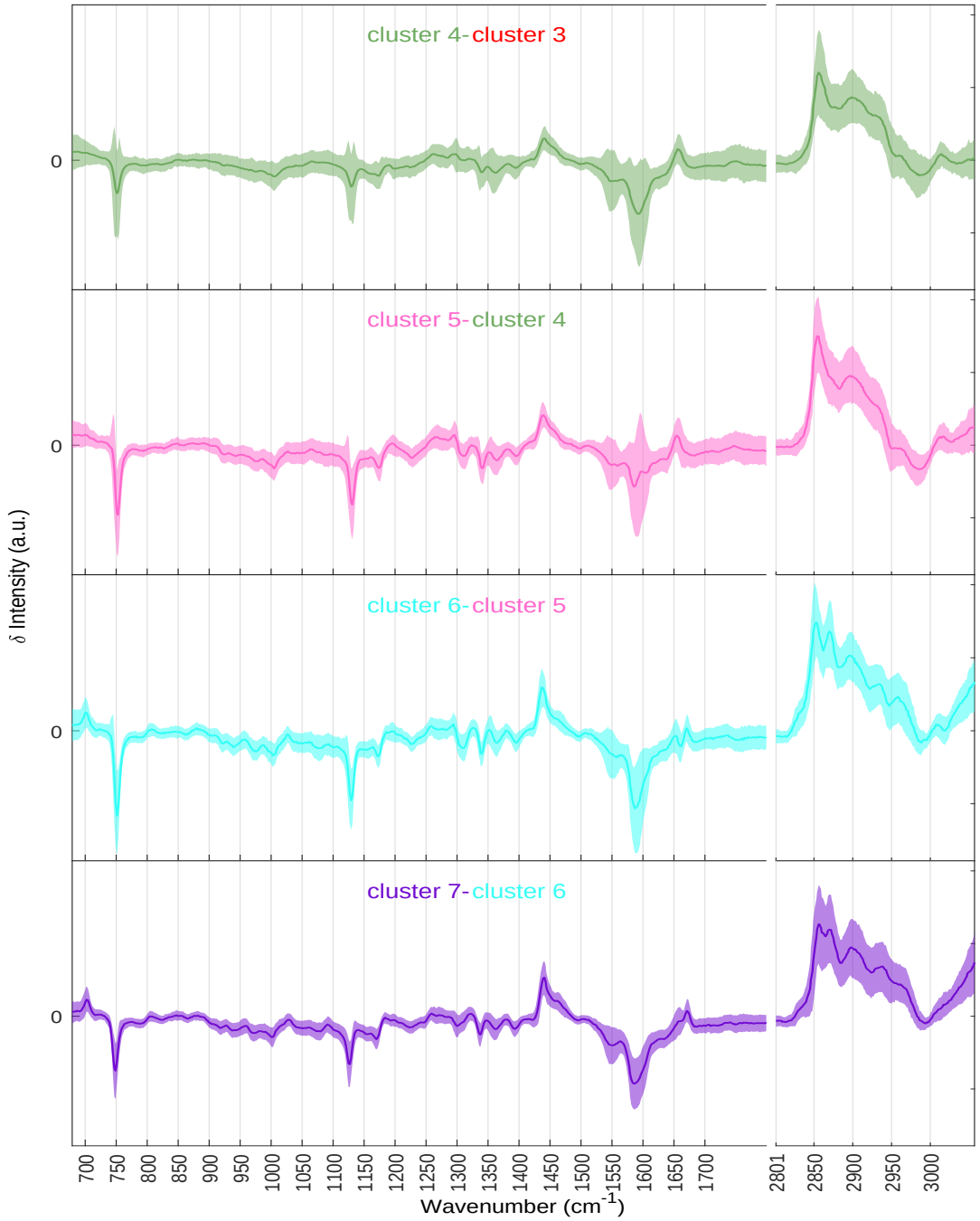


Figure 4.11: *The mean difference spectra of 5 adjacent clusters (cluster 3 (red) to cluster 5 (purple))*

The superimposed plot of the mean difference spectra of 5 adjacent spectral clusters (cluster 3 (red) to cluster 5 (purple)) are shown in following Fig. 4.12:

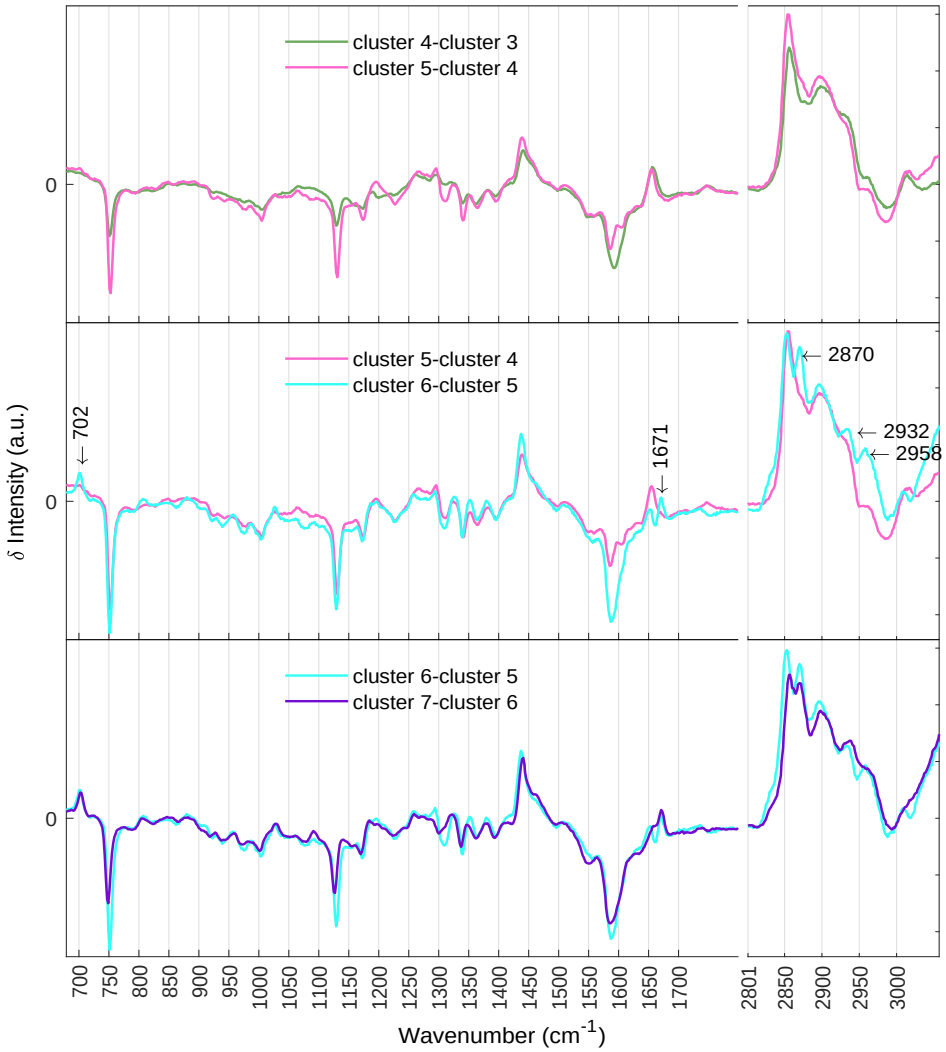


Figure 4.12: *The superimposed plot of mean difference spectra of 5 adjacent spectral clusters (cluster 3 (red) to cluster 5 (purple))*

Based on these observations, group 1, containing mostly SD and histologically normal tissues, and having population distributions indicating healthy hepatocytes and large presence of vitamin A in HSCs, most likely represent healthy liver tissue. Tissues in groups 2L and 2R, which are dominated by clusters 4 and 5, respectively, indicate increased lipid contribution and decreased vitamin A and cytochrome contributions without notable changes in lipid composition, which may suggest an early stage of NAFLD [4]. In contrast, groups 3L and 3R are dominated by clusters 6 and 7 in which notable changes in the spectral profile in regions associated with lipids are observed. This may indicate both

an increase in lipid accumulation and a reorganization of lipid composition, which, along with the appearance of cholesterol, cholesterol esters, and triacylglycerols and reduction in vitamin A contribution (vitamin A is absent in pathological states [40,41,132,133]), has been associated with more severe states of NAFLD [4,40]. To summarize, the 5 groups of tissues obtained by AHC may be coarse-grained reflections of different states of the liver tissues, with group 1 being a healthy state, and the progression through groups 2L, 2R, 3L, and 3R being different states of NAFLD ranging sequentially from mild to severe.

4.4.3 Analogy between Diet/Histological States and Raman Assignment

To know how the diet/histological assignment (disease states of NAFLD) for each of the tissue correlate with their Raman characteristics, we use the information of the cluster population distributions (shown in Fig. 4.3B) and mean population distributions in diet/histological state shown in Fig. 4.9B and 4.9C. The average population distribution is the mean characteristics of each histological/diet states, which provides the average behavior of the three histological/states expressed in terms of 7 clusters. These average population distributions are used as references to classify each tissue (rat) by measuring the degree of similarity of cluster population of individual rat (shown in Fig. 4.3B) to the mean population distribution. First, we calculate the L_1 distance of the population vectors of each rat from the average population distribution in a particular state g as

$$d(i,g) = \sum_{k=1}^{7} \left| P_k^i - \bar{P}_k^g \right|, \ i = 1, 2, .., 48 \text{ (no. of images)}; \ g = 1, ..., 3 \text{ (no. of states)},$$

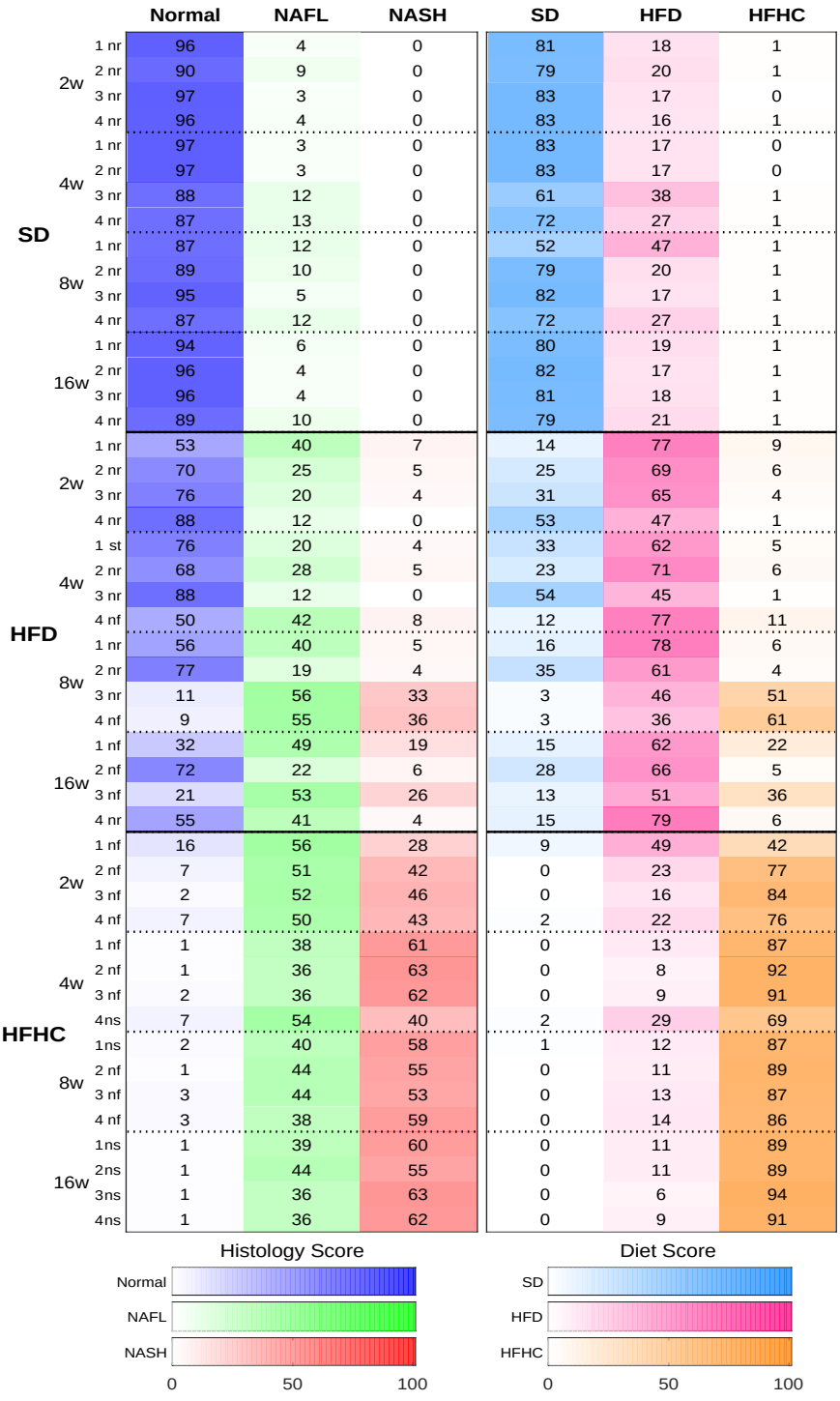
where P_k^i is the population of the spectral cluster k in the image i and $\bar{P}_k^g$ is the average population of cluster k in the state g. Then, the similarity of the cluster population of each rat to average population distribution is computed as

$$sm(i,g) = \max(d) - d(i,g), \text{ where } d = \{d(i,g)\}, \ i = 1, 2, ..., 48.$$

Finally, the similarity score of each of individual rat to the average population distribution is computed by

$$sc(i,g) = 100 \times \frac{sm(i,g)}{\sum_{g=1}^{3} sm(i,g)}.$$

The similarity score plot is shown in Fig. 4.13.



		Normal	NAFL	NASH	SD	HFD	HFHC
SD	2w	1 nr	96	4	0	81	18	1
		2 nr	90	9	0	79	20	1
		3 nr	97	3	0	83	17	0
		4 nr	96	4	0	83	16	1
	4w	1 nr	97	3	0	83	17	0
		2 nr	97	3	0	83	17	0
		3 nr	88	12	0	61	38	1
		4 nr	87	13	0	72	27	1
	8w	1 nr	87	12	0	52	47	1
		2 nr	89	10	0	79	20	1
		3 nr	95	5	0	82	17	1
		4 nr	87	12	0	72	27	1
	16w	1 nr	94	6	0	80	19	1
		2 nr	96	4	0	82	17	1
		3 nr	96	4	0	81	18	1
		4 nr	89	10	0	79	21	1
HFD	2w	1 nr	53	40	7	14	77	9
		2 nr	70	25	5	25	69	6
		3 nr	76	20	4	31	65	4
		4 nr	88	12	0	53	47	1
	4w	1 st	76	20	4	33	62	5
		2 nr	68	28	5	23	71	6
		3 nr	88	12	0	54	45	1
		4 nf	50	42	8	12	77	11
	8w	1 nr	56	40	5	16	78	6
		2 nr	77	19	4	35	61	4
		3 nr	11	56	33	3	46	51
		4 nf	9	55	36	3	36	61
	16w	1 nf	32	49	19	15	62	22
		2 nf	72	22	6	28	66	5
		3 nf	21	53	26	13	51	36
		4 nr	55	41	4	15	79	6
HFHC	2w	1 nf	16	56	28	9	49	42
		2 nf	7	51	42	0	23	77
		3 nf	2	52	46	0	16	84
		4 nf	7	50	43	2	22	76
	4w	1 nf	1	38	61	0	13	87
		2 nf	1	36	63	0	8	92
		3 nf	2	36	62	0	9	91
		4ns	7	54	40	2	29	69
	8w	1ns	2	40	58	1	12	87
		2 nf	1	44	55	0	11	89
		3 nf	3	44	53	0	13	87
		4 nf	3	38	59	0	14	86
	16w	1ns	1	39	60	0	11	89
		2ns	1	44	55	0	11	89
		3ns	1	36	63	0	6	94
		4ns	1	36	62	0	9	91

Figure 4.13: *Left column: The similarity score plot of individual image to each of the histological sate. Right column: The similarity score plot of individual image to each of the diet sate. Here the similarity scores ranges from 0 to 100 for normal, NAFL and NASH are, respectively, represented by the three different color maps: blue, green and red, while those for SD, HFD and HFHC are, respectively, represented by the color maps: cyan, pink and orange. The true tissue labels are presented vertically at the left of the figure; normal, NAFL, NASH and strange histology are, respectively, represented by 'nr', 'nf', 'ns' and 'st'.*

The similarity score provides the information about each of the images: how their Raman characteristics match with histological (/or diet based) assignment (i.e., disease state of tissue).

In case of dietary model, we see that SD and HFHC tissues are accurately classified while 4 HFD (HFD: 2w-4, 4w-3, 8w-3, 8w-4) tissues are misclassified which harmonizes with the results obtained by 3-group AHC results in Fig. 5; suggesting that Raman information can accurately classify the tissue states of dietary model (ground truth).
In case of histological assignments, we have found that, out of 26 normal rats, one (HFD 8w-3) is classified as NAFL. Rat with peculiar histology (mild fibrosis) (HFD 4w-1) is similar to normal. Out of 15 NAFL rats, two rats are classified as normal: HFD 4w-4 and HFD 16w-2, and six NAFL rats are classified as NASH: HFHC 4w-1, HFHC 4w-2, HFHC 4w-3, HFHC 8w-2, HFHC 8w-3, and HFHC 8w-4. Finally, out of six NASH, one is similar to NAFL: HFHC 4w-4.

Thus, we see that normal and NASH states are mostly in agreement with the Raman assignment. But in case of NAFL, the histological assignment and Raman characteristics are quite different, which may indicate that there is some uncertainty in the histological diagnosis for NAFL due to its heterogeneous nature, which can be extracted by Raman micro-spectroscopy.

4.5 Concluding Remarks

The approach presented in this chapter provided a scheme for Raman image-based histology in which low signal-to-noise ratio (SNR) inherent to Raman measurements are taken into account in terms of superpixel segmentation of the images based on spatial and spectral proximity. The superpixel segmentation algorithm, which can be used to segment any type of hyper-spectral image, produced a visually superior image (reducing noise while preserving spatial features as much as possible) in comparison to traditional rectangular griding. Cluster maps allowed for the visualization of the molecular distributions across the liver tissues, elucidating possible chemical components that have vital roles in the progression of the diseases. In agreement with previous studies [3], [4], [41], we observed that the progression of NAFLD is most likely associated to deposition of fats in the liver along with the reduction of vitamin A and cytochrome presence. We

also found that Raman information can extract the diversity in the different diet and histological states, and help to categorize each rat in terms of severity.

Chapter 5

Conclusions and Outlook

The main goal of this work was the systematic analysis of Raman hyperspectral image data to increase diagnostic reliability of histopathology using the combination of analytical tools from machine learning and information theory. Using a dietary model of non-alcoholic fatty liver disease in rats, we have coupled Raman imaging with methods of machine learning and information theory to assist histological inspection that benefit disease diagnosis. This work was divided into two parts. In the first study, through dimensional reduction and ensemble-learning-based random forest classification of the spectra within the Raman images, we have found good agreement with the classification of the spectra and the histological assignments of livers judged by morphological changes. Furthermore, we have found enhancement of diagnostic capabilities in the distinction of the states of tissues in the early stages of disease in which histological characteristics had not yet been observed. Our approach well distinguished the two phases of non-alcoholic fatty liver (NAFL), 'slowly progressive NAFL' (NAFL-α) and 'rapidly progressive NAFL' (NAFL-β), whilst the histological test could not. It has predicted the appearance of NASH after only two weeks feeding on a high-fat, high-cholesterol diet before histological signatures emerge, identifying a nascent state of NASH. Furthermore, the random forest classifier explored the most relevant set of spectral features which led to the discrimination of NAFL-α and NAFL-β efficiently with high accuracy which may accelerate the Raman diagnosis.

In the second approach, we have developed a scheme for Raman image-based histology in which low signal-to-noise ratio (SNR) inherent to Raman measurements are taken into account in terms of superpixel segmentation of the images based on spatial and spectral

proximity. Low SNR was also considered in the extraction of chemically distinct micro-environments (clusters) distance among Raman spectra through the determination of the number of clusters with the quantification of Poisson noise in the Raman signals. This superpixel segmentation algorithm, which can be used to segment any type of hyper-spectral image, produced a visually superior image in comparison to traditional rectangular gridding. In this application, we chose 9 pixels per superpixel so the superpixels were roughly cellular in size, but one can also choose the size of superpixels with respect to a particular aim such as prediction accuracy. Rate-distortion theory, as an unsupervised clustering algorithm, identified a minimal number of spectral clusters as determined by the magnitude of the Poisson error arising from photon counting. Each cluster was a representation of regions of the tissue containing distinct biochemical composition, allowing for the exploration of the biochemical structures composing the liver tissues. Mapping spectral clusters to their locations in the tissues also allowed for the visualization of the molecular distributions across the liver tissues, elucidating possible chemical components that have vital roles in the progression of the diseases. In agreement with previous studies [3], [4], [41], we have observed that the progression of NAFLD is most likely associated to deposition of fats in the liver along with the reduction of vitamin A and cytochrome presence. We have also found that Raman information can extract the diversity in the different diet and histological states, and help to categorize each rat in terms of severity. We note that a point-detection method was employed with relatively large single pixels ($\sim 25\ \mu m^2$). Future considerations involve the use of nanometer-resolution Raman imaging of the liver tissues. In combination with the proposed machine learning framework, we expect even finer-scale insights into the biochemical distribution at cellular level inside the liver tissues.

Finally, we aimed that our analysis to be a diagnostic aid to histopathologists, assisting in NAFLD diagnosis, and even discerning molecular origins through the detection of subtle Raman spectral changes due to the disease-related changes in chemical constituents.

Bibliography

1. C. A. Matteoni, Z. M. Younossi, T. Gramlich, N. Boparai, Y. C. Liu, and A. J. McCullough, "Nonalcoholic fatty liver disease: a spectrum of clinical and pathological severity.," *Gastroenterology*, vol. 116, pp. 1413–9, Jun 1999.

2. P. Angulo, "Nonalcoholic fatty liver disease," *New England Journal of Medicine*, vol. 346, no. 16, pp. 1221–1231, 2002.

3. K. Kochan, E. Maslak, C. Krafft, R. Kostogrys, S. Chlopicki, and M. Baranska, "Raman spectroscopy analysis of lipid droplets content, distribution and saturation level in non-alcoholic fatty liver disease in mice," *Journal of biophotonics*, vol. 8, no. 7, pp. 597–609, 2015.

4. M. Z. Pacia, K. Czamara, M. Zebala, E. Kus, S. Chlopicki, and A. Kaczor, "Rapid diagnostics of liver steatosis by raman spectroscopy via fiber optic probe: a pilot study," *Analyst*, vol. 143, no. 19, pp. 4723–4731, 2018.

5. E. Buzzetti, M. Pinzani, and E. A. Tsochatzis, "The multiple-hit pathogenesis of non-alcoholic fatty liver disease (nafld)," *Metabolism*, vol. 65, no. 8, pp. 1038–1048, 2016.

6. E. Hashimoto, K. Tokushige, and J. Ludwig, "Diagnosis and classification of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis: Current concepts and remaining challenges," *Hepatology research*, vol. 45, no. 1, pp. 20–28, 2015.

7. E. M. Brunt, "Nonalcoholic steatohepatitis: definition and pathology.," *Seminars in liver disease*, vol. 21, no. 1, pp. 3–16, 2001.

8. M. V. Machado and A. M. Diehl, "Pathogenesis of nonalcoholic steatohepatitis," *Gastroenterology*, vol. 150, no. 8, pp. 1769–1777, 2016.

9. R. Pais, F. Charlotte, L. Fedchuk, P. Bedossa, P. Lebray, T. Poynard, and V. Ratziu, "A systematic review of follow-up biopsies reveals disease progression in patients with non-alcoholic fatty liver.," *Journal of hepatology*, vol. 59, pp. 550–6, Sep 2013.

10. S. Singh, A. M. Allen, Z. Wang, L. J. Prokop, M. H. Murad, and R. Loomba, "Fibrosis progression in nonalcoholic fatty liver vs nonalcoholic steatohepatitis: a systematic review and meta-analysis of paired-biopsy studies," *Clinical Gastroenterology and Hepatology*, vol. 13, no. 4, pp. 643–654, 2015.

11. S. McPherson, T. Hardy, E. Henderson, A. D. Burt, C. P. Day, and Q. M. Anstee, "Evidence of nafld progression from steatosis to fibrosing-steatohepatitis using paired biopsies: implications for prognosis and clinical management," *Journal of hepatology*, vol. 62, no. 5, pp. 1148–1155, 2015.

12. L. A. Adams and V. Ratziu, "Non-alcoholic fatty liver–perhaps not so benign," *Journal of hepatology*, vol. 62, no. 5, pp. 1002–1004, 2015.

13. S. Cramer, L. Roth, S. Mills, T. Ulbright, D. Gersell, C. Nunez, and F. Kraus, "Sources of variability in classifying common ovarian cancers using the world health organization classification. application of the pathtracking method.," *Pathology annual*, vol. 28, p. 243, 1993.

14. S. C. Hollensead, W. B. Lockwood, and R. J. Elin, "Errors in pathology and laboratory medicine: consequences and prevention," *Journal of surgical oncology*, vol. 88, no. 3, pp. 161–181, 2004.

15. S. S. Raab, "Variability of practice in anatomic pathology and its effect on patient outcomes," in *Seminars in diagnostic pathology*, vol. 22, pp. 177–185, Elsevier, 2005.

16. J. Yan, Y. Yu, J. W. Kang, Z. Y. Tam, S. Xu, E. L. S. Fong, S. P. Singh, Z. Song, L. Tucker-Kellogg, P. T. So, *et al.*, "Development of a classification model for non-alcoholic steatohepatitis (nash) using confocal raman micro-spectroscopy," *Journal of biophotonics*, vol. 10, no. 12, pp. 1703–1713, 2017.

17. E. M. Brunt, C. G. Janney, A. M. Di Bisceglie, B. A. Neuschwander-Tetri, and B. R. Bacon, "Nonalcoholic steatohepatitis: a proposal for grading and staging

the histological lesions," *The American journal of gastroenterology*, vol. 94, no. 9, p. 2467, 1999.

18. D. E. Kleiner, E. M. Brunt, M. Van Natta, C. Behling, M. J. Contos, O. W. Cummings, L. D. Ferrell, Y.-C. Liu, M. S. Torbenson, A. Unalp-Arida, *et al.*, "Design and validation of a histological scoring system for nonalcoholic fatty liver disease," *Hepatology*, vol. 41, no. 6, pp. 1313–1321, 2005.

19. E. M. Brunt, D. E. Kleiner, L. A. Wilson, P. Belt, B. A. Neuschwander-Tetri, and N. C. R. N. (CRN), "Nonalcoholic fatty liver disease (nafld) activity score and the histopathologic diagnosis in nafld: distinct clinicopathologic meanings," *Hepatology*, vol. 53, no. 3, pp. 810–820, 2011.

20. P. Stål, "Liver fibrosis in non-alcoholic fatty liver disease-diagnostic challenge with prognostic significance," *World Journal of Gastroenterology: WJG*, vol. 21, no. 39, p. 11077, 2015.

21. K. Hamada, K. Fujita, N. I. Smith, M. Kobayashi, Y. Inouye, and S. Kawata, "Raman microscopy for dynamic molecular imaging of living cells," *Journal of biomedical optics*, vol. 13, no. 4, p. 044027, 2008.

22. G. Puppels, F. de Mul, C. Otto, J. Greve, M. Robert-Nicoud, D. Arndt-Jovin, and T. Jovin, "Studying single living cells and chromosomes by confocal raman microspectroscopy," *Nature*, vol. 347, pp. 301–303, 1990.

23. Y. Harada, P. Dai, Y. Yamaoka, M. Ogawa, H. Tanaka, K. Nosaka, K. Akaji, and T. Takamatsu, "Intracellular dynamics of topoisomerase i inhibitor, cpt-11, by slit-scanning confocal raman microscopy.," *Histochemistry and cell biology*, vol. 132, pp. 39–46, Jul 2009.

24. M. Okada, N. I. Smith, A. F. Palonpon, H. Endo, S. Kawata, M. Sodeoka, and K. Fujita, "Label-free raman observation of cytochrome c dynamics during apoptosis," *Proceedings of the National Academy of Sciences*, vol. 109, no. 1, pp. 28–32, 2012.

25. A. F. Palonpon, M. Sodeoka, and K. Fujita, "Molecular imaging of live cells by

raman microscopy," *Current opinion in chemical biology*, vol. 17, pp. 708–15, Aug 2013.

26. K. Klein, A. M. Gigler, T. Aschenbrenner, R. Monetti, W. Bunk, F. Jamitzky, G. Morfill, R. W. Stark, and J. Schlegel, "Label-free live-cell imaging with confocal raman microscopy.," *Biophysical journal*, vol. 102, no. 2, pp. 360–8, 2012.

27. J. W. Chan, D. S. Taylor, T. Zwerdling, S. M. Lane, K. Ihara, and T. Huser, "Micro-raman spectroscopy detects individual neoplastic and normal hematopoietic cells.," *Biophysical journal*, vol. 90, no. 2, pp. 648–56, 2006.

28. A. S. Haka, Z. Volynskaya, J. A. Gardecki, J. Nazemi, J. Lyons, D. Hicks, M. Fitzmaurice, R. R. Dasari, J. P. Crowe, and M. S. Feld, "In vivo margin assessment during partial mastectomy breast surgery using raman spectroscopy.," *Cancer research*, vol. 66, pp. 3317–22, Mar 2006.

29. M. S. Bergholt, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. Yan So, A. Shabbir, and Z. Huang, "Fiberoptic confocal raman spectroscopy for real-time in vivo diagnosis of dysplasia in barrett's esophagus," *Gastroenterology*, vol. 146, no. 1, pp. 27–32, 2014.

30. L. M. Almond, J. Hutchings, G. Lloyd, H. Barr, N. Shepherd, J. Day, O. Stevens, S. Sanders, M. Wadley, N. Stone, and C. Kendall, "Endoscopic raman spectroscopy enables objective diagnosis of dysplasia in barrett's esophagus.," *Gastrointestinal endoscopy*, vol. 79, pp. 37–45, Jan 2014.

31. J. Wang, K. Lin, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, and Z. Huang, "Simultaneous fingerprint and high-wavenumber fiber-optic raman spectroscopy improves in vivo diagnosis of esophageal squamous cell carcinoma at endoscopy," *Scientific reports*, vol. 5, p. 12957, 2015.

32. Y. Harada and T. Takamatsu, "Raman molecular imaging of cells and tissues: towards functional diagnostic imaging without labeling," *Current pharmaceutical biotechnology*, vol. 14, no. 2, pp. 133–140, 2013.

33. Z.-Q. Wen, "Raman spectroscopy of protein pharmaceuticals.," *Journal of pharmaceutical sciences*, vol. 96, pp. 2861–78, Nov 2007.

34. F. Lyng, D. Traynor, I. Ramos, F. Bonnier, and H. Byrne, "Raman spectroscopy for screening and diagnosis of cervical cancer," *Anal. Bioanal. Chem.*, vol. 2407, p. 8279–8289, 2015.

35. H. Lui, J. Zhao, D. McLean, and H. Zeng, "Real-time raman spectroscopy for in vivo skin cancer diagnosis," *Cancer research*, vol. 72, no. 10, pp. 2491–500, 2012.

36. K. Lin, J. Wang, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, and Z. Huang, "Rapid fiber-optic raman spectroscopy for real-time in vivo detection of gastric intestinal metaplasia during clinical gastroscopy," *Cancer Prevention Research*, vol. 9, no. 6, pp. 476–483, 2016.

37. L. Austin, S. Osseiran, and C. Evans, "Raman technologies in cancer diagnostics," *Analyst*, vol. 141, pp. 476—-503, 2016.

38. E. Brauchle, S. Thude, S. Y. Brucker, and K. Schenke-Layland, "Cell death stages in single apoptotic and necrotic cells monitored by raman microspectroscopy," *Scientific reports*, vol. 4, p. 4698, 2014.

39. F. M. Lyng, D. Traynor, T. N. Q. Nguyen, A. D. Meade, F. Rakib, R. Al-Saady, E. Goormaghtigh, K. Al-Saad, and M. H. Ali, "Discrimination of breast cancer from benign tumours using raman spectroscopy," *PloS one*, vol. 14, no. 2, p. e0212376, 2019.

40. K. Kochan, K. M. Marzec, K. Chruszcz-Lipska, A. Jasztal, E. Maslak, H. Musiolik, S. Chłopicki, and M. Baranska, "Pathological changes in the biochemical profile of the liver in atherosclerosis and diabetes assessed by raman spectroscopy.," *Analyst*, vol. 138, no. 14, pp. 3885–3890, 2013.

41. K. Kochan, K. M. Marzec, E. Maslak, S. Chłopicki, and M. Baranska, "Raman spectroscopic studies of vitamin a content in the liver: a biomarker of healthy liver," *Analyst*, vol. 140.

42. M. B. Fenn and V. Pappu, "Data mining for cancer biomarkers with raman spectroscopy," in *Data Mining for Biomarker Discovery*, pp. 143–168, Springer, 2012.

43. M. Jermyn, J. Desroches, J. Mercier, M.-A. Tremblay, K. St-Arnaud, M.-C. Guiot, K. Petrecca, and F. Leblond, "Neural networks improve brain cancer detection with

raman spectroscopy in the presence of operating room light artifacts," *Journal of biomedical optics*, vol. 21, no. 9, p. 094002, 2016.

44. S. Li, Y. Zhang, J. Xu, L. Li, Q. Zeng, L. Lin, Z. Guo, Z. Liu, H. Xiong, and S. Liu, "Noninvasive prostate cancer screening based on serum surface-enhanced raman spectroscopy and support vector machine," *Applied Physics Letters*, vol. 105, no. 9, p. 091104, 2014.

45. M. Sattlecker, C. Bessant, J. Smith, and N. Stone, "Investigation of support vector machines and raman spectroscopy for lymph node diagnostics," *Analyst*, vol. 135, no. 5, pp. 895–901, 2010.

46. S. Khan, R. Ullah, A. Khan, N. Wahab, M. Bilal, and M. Ahmed, "Analysis of dengue infection based on raman spectroscopy and support vector machine (svm)," *Biomedical optics express*, vol. 7, no. 6, pp. 2249–2256, 2016.

47. N. Pavillon, A. J. Hobro, S. Akira, and N. I. Smith, "Noninvasive detection of macrophage activation with single-cell resolution through machine learning," *Proceedings of the National Academy of Sciences*, vol. 115, no. 12, pp. E2676–E2685, 2018.

48. E. Guevara, J. C. Torres-Galván, M. G. Ramírez-Elías, C. Luevano-Contreras, and F. J. González, "Use of raman spectroscopy to screen diabetes mellitus with machine learning tools," *Biomedical optics express*, vol. 9, no. 10, pp. 4998–5010, 2018.

49. J. N. Taylor, K. Mochizuki, K. Hashimoto, Y. Kumamoto, Y. Harada, K. Fujita, and T. Komatsuzaki, "High-resolution raman microscopic detection of follicular thyroid cancer cells with unsupervised machine learning," *The Journal of Physical Chemistry B*, 2019.

50. J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction.," *Science (New York, N.Y.)*, vol. 290, pp. 2319–23, Dec 2000.

51. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

52. A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

53. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

54. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," tech. rep., 2010.

55. P. Neubert and P. Protzel, "Superpixel benchmark and comparison," in *Proc. Forum Bildverarbeitung*, vol. 6, 2012.

56. C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

57. T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

58. `http://bwtek.com/raman-theory-of-raman-scattering/`.

59. `https://www.doitpoms.ac.uk/tlplib/raman/raman_scattering.php/`.

60. R. L. McCreery, *Raman spectroscopy for chemical analysis*, vol. 225. John Wiley & Sons, 2005.

61. T. Tolstik, *Development of new classification models based on Raman spectroscopy and MALDI spectrometry as novel tools for liver cancer diagnostic*. PhD thesis, Friedrich-Schiller-Universität Jena, 2016.

62. I. R. Lewis and H. Edwards, *Handbook of Raman spectroscopy: from the research laboratory to the process line*. CRC Press, 2001.

63. C. V. Raman and K. S. Krishnan, "A new type of secondary radiation," *Nature*, vol. 121, no. 3048, p. 501, 1928.

64. A. Zoladek, *Confocal Raman imaging of live cells*. PhD thesis, University of Nottingham, 2011.

65. G. S. Bumbrah and R. M. Sharma, "Raman spectroscopy–basic principle, instrumentation and selected applications for the characterization of drugs of abuse," *Egyptian Journal of Forensic Sciences*, vol. 6, no. 3, pp. 209–215, 2016.

66. R. Haydock, *Multivariate analysis of Raman spectroscopy data*. PhD thesis, University of Nottingham, 2015.

67. A. F. Palonpon, J. Ando, H. Yamakoshi, K. Dodo, M. Sodeoka, S. Kawata, and K. Fujita, "Raman and sers microscopy for molecular imaging of live cells.," *Nature protocols*, vol. 8, pp. 677–92, Apr 2013.

68. `http://www.horiba.com/us/en/scientific/products/raman-spectroscopy/raman-academy/raman-faqs/what-is-a-confocal-raman-microscope/`.

69. N. Malhotra and M. D. Beaton, "Management of non-alcoholic fatty liver disease in 2015," *World journal of hepatology*, vol. 7, no. 30, p. 2962, 2015.

70. N. Chalasani, Z. Younossi, J. E. Lavine, A. M. Diehl, E. M. Brunt, K. Cusi, M. Charlton, and A. J. Sanyal, "The diagnosis and management of non-alcoholic fatty liver disease: Practice guideline by the american association for the study of liver diseases, american college of gastroenterology, and the american gastroenterological association," *Hepatology*, vol. 55, no. 6, pp. 2005–2023, 2012.

71. A. Hanson, D. Wilhelmsen, and J. DiStefano, "The role of long non-coding rnas (lncrnas) in the development and progression of fibrosis associated with nonalcoholic fatty liver disease (nafld)," *Non-coding RNA*, vol. 4, no. 3, p. 18, 2018.

72. C. D. Williams, J. Stengel, M. I. Asike, D. M. Torres, J. Shaw, M. Contreras, C. L. Landt, and S. A. Harrison, "Prevalence of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis among a largely middle-aged population utilizing ultrasound and liver biopsy: a prospective study," *Gastroenterology*, vol. 140, no. 1, pp. 124–131, 2011.

73. K.-C. Sung and S. H. Kim, "Interrelationship between fatty liver and insulin resistance in the development of type 2 diabetes," *The Journal of Clinical Endocrinology & Metabolism*, vol. 96, no. 4, pp. 1093–1097, 2011.

74. H. Ma, C.-f. Xu, C.-h. Yu, and Y.-m. Li, "Application of machine learning techniques for clinical predictive modeling: A cross-sectional study on nonalcoholic fatty liver disease in china," *BioMed Research International*, vol. 2018, no. 4304376, 2018.

75. R. Gautam, S. Vanga, F. Ariese, and S. Umapathy, "Review of multidimensional data processing approaches for raman and infrared spectroscopy," *EPJ Techniques and Instrumentation*, vol. 2, no. 1, p. 8, 2015.

76. T. Bocklitz, A. Walter, K. Hartmann, P. Rösch, and J. Popp, "How to pre-process raman spectra for reliable and stable models?," *Analytica chimica acta*, vol. 704, no. 1-2, pp. 47–56, 2011.

77. G. Srinivasan, *Vibrational spectroscopic imaging for biomedical applications*. McGraw Hill Professional, 2010.

78. 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.

79. S. Lohumi, S. Lee, H. Lee, M. S. Kim, W.-H. Lee, and B.-K. Cho, "Application of hyperspectral imaging for characterization of intramuscular fat distribution in beef," *Infrared Physics & Technology*, vol. 74, pp. 1–10, 2016.

80. I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, "Euclidean distance matrices: essential theory, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 12–30, 2015.

81. J. B. Kruskal and M. Wish, "Multidimensional scaling. 1978," *Beverly Hills, CA*, 1978.

82. T. F. Cox and M. A. Cox, *Multidimensional scaling*. Chapman and hall/CRC, 2000.

83. I. Borg and P. Groenen, "Springer series in statistics," *Modern multidimensional scaling: Theory and applications (2nd ed.). New York, NY, US: Springer Science+ Business Media*, 2005.

84. R. W. Floyd, "Algorithm 97: shortest path," *Communications of the ACM*, vol. 5, no. 6, p. 345, 1962.

85. O. Samko, A. D. Marshall, and P. L. Rosin, "Selection of the optimal parameter value for the isomap algorithm," *Pattern Recognition Letters*, vol. 27, no. 9, pp. 968–979, 2006.

86. C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, p. 25, 2007.

87. G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in neural information processing systems*, pp. 431–439, 2013.

88. T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomedical Signal Processing and Control*, 2017.

89. M. Van Herck, L. Vonghia, and S. Francque, "Animal models of nonalcoholic fatty liver disease—a starter's guide," *Nutrients*, vol. 9, no. 10, p. 1072, 2017.

90. K. Omagari, S. Kato, K. Tsuneyama, C. Inohara, Y. Kuroda, H. Tsukuda, E. Fukazawa, K. Shiraishi, and M. Mune, "Effects of a long-term high-fat diet and switching from a high-fat to low-fat, standard diet on hepatic fat accumulation in sprague-dawley rats," *Digestive diseases and sciences*, vol. 53, no. 12, p. 3206, 2008.

91. Y. Okada, K. Yamaguchi, T. Nakajima, T. Nishikawa, M. Jo, Y. Mitsumoto, H. Kimura, T. Nishimura, N. Tochiki, K. Yasui, H. Mitsuyoshi, M. Minami, K. Kagawa, T. Okanoue, and Y. Itoh, "Rosuvastatin ameliorates high-fat and high-cholesterol diet-induced nonalcoholic steatohepatitis in rats.," *Liver international : official journal of the International Association for the Study of the Liver*, vol. 33, pp. 301–11, Feb 2013.

92. M. Ichimura, M. Kawase, M. Masuzumi, M. Sakaki, Y. Nagata, K. Tanaka, K. Suruga, S. Tamaru, S. Kato, K. Tsuneyama, *et al.*, "High-fat and high-cholesterol diet rapidly induces non-alcoholic steatohepatitis with advanced fibrosis in sprague-dawley rats," *Hepatology Research*, vol. 45, no. 4, pp. 458–469, 2015.

93. M. K. Chung, "Gaussian kernel smoothing," *Lecture notes*, pp. 1–10, 2012.

94. C.-B. Li, H. Yang, and T. Komatsuzaki, "Multiscale complex network of protein conformational fluctuations in single-molecule time series," *Proceedings of the National Academy of Sciences*, vol. 105, no. 2, pp. 536–541, 2008.

95. J. N. Taylor, C.-B. Li, D. R. Cooper, C. F. Landes, and T. Komatsuzaki, "Error-based extraction of states and energy landscapes from experimental single-molecule time-series," *Scientific reports*, vol. 5, p. 9174, 2015.

96. D. Wales, *Energy landscapes: Applications to clusters, biomolecules and glasses*. Cambridge University Press, 2003.

97. A. Baba and T. Komatsuzaki, "Construction of effective free energy landscape from single-molecule time series.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 19297–302, Dec 2007.

98. K. Czamara, K. Majzner, M. Pacia, K. Kochan, A. Kaczor, and M. Baranska, "Raman spectroscopy of lipids: a review," *Journal of Raman Spectroscopy*, vol. 46, no. 1, pp. 4–20, 2015.

99. N. Failloux, I. Bonnet, M.-H. Baron, and E. Perrier, "Quantitative analysis of vitamin a degradation by raman spectroscopy," *Applied spectroscopy*, vol. 57, no. 9, pp. 1117–1122, 2003.

100. F. Chiappini, A. Coilly, H. Kadar, P. Gual, A. Tran, C. Desterke, D. Samuel, J.-C. Duclos-Vallée, D. Touboul, J. Bertrand-Michel, *et al.*, "Metabolism dysregulation induces a specific lipid signature of nonalcoholic steatohepatitis in patients," *Scientific reports*, vol. 7, p. 46658, 2017.

101. F. Marra and G. Svegliati-Baroni, "Lipotoxicity and the gut-liver axis in nash pathogenesis," *Journal of hepatology*, vol. 68, no. 2, pp. 280–295, 2018.

102. K. Yasutake, M. Kohjima, K. Kotoh, M. Nakashima, M. Nakamuta, and M. Enjoji, "Dietary habits and behaviors associated with nonalcoholic fatty liver disease," *World Journal of Gastroenterology: WJG*, vol. 20, no. 7, p. 1756, 2014.

103. K. Yasutake, M. Nakamuta, Y. Shima, A. Ohyama, K. Masuda, N. Haruta, T. Fujino, Y. Aoyagi, K. Fukuizumi, T. Yoshimoto, *et al.*, "Nutritional investigation of non-obese patients with non-alcoholic fatty liver disease: the significance of dietary cholesterol," *Scandinavian journal of gastroenterology*, vol. 44, no. 4, pp. 471–477, 2009.

104. P. Puri, R. A. Baillie, M. M. Wiest, F. Mirshahi, J. Choudhury, O. Cheung, C. Sargeant, M. J. Contos, and A. J. Sanyal, "A lipidomic analysis of nonalcoholic fatty liver disease," *Hepatology*, vol. 46, no. 4, pp. 1081–1090, 2007.

105. G. Arguello, E. Balboa, M. Arrese, and S. Zanlungo, "Recent insights on the role of cholesterol in non-alcoholic fatty liver disease," *Biochimica Et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1852, no. 9, pp. 1765–1778, 2015.

106. M. Marí, F. Caballero, A. Colell, A. Morales, J. Caballeria, A. Fernandez, C. Enrich, J. C. Fernandez-Checa, and C. García-Ruiz, "Mitochondrial free cholesterol loading sensitizes to tnf-and fas-mediated steatohepatitis," *Cell metabolism*, vol. 4, no. 3, pp. 185–198, 2006.

107. M. Goetz, C. Weber, J. Bloecher, B. Stieltjes, H.-P. Meinzer, and K. Maier-Hein, "Extremely randomized trees based brain tumor segmentation," *Proceeding of BRATS challenge-MICCAI*, pp. 006–011, 2014.

108. P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

109. R. Marée, P. Geurts, and L. Wehenkel, "Random subwindows and extremely randomized trees for image classification in cell biology," *BMC Cell Biology*, vol. 8, no. 1, p. S2, 2007.

110. P. Geurts, M. Fillet, D. De Seny, M.-A. Meuwis, M. Malaise, M.-P. Merville, and L. Wehenkel, "Proteomic mass spectra classification using decision tree based ensemble methods," *Bioinformatics*, vol. 21, no. 14, pp. 3138–3145, 2005.

111. M. Soltaninejad, G. Yang, T. Lambrou, N. Allinson, T. L. Jones, T. R. Barrick, F. A. Howe, and X. Ye, "Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in flair mri," *International journal of computer assisted radiology and surgery*, vol. 12, no. 2, pp. 183–203, 2017.

112. O. Csillik, "Fast segmentation and classification of very high resolution remote sensing data using slic superpixels," *Remote Sensing*, vol. 9, no. 3, p. 243, 2017.

113. C. Shi and L. Wang, "Incorporating spatial information in spectral unmixing: A review," *Remote Sensing of Environment*, vol. 149, pp. 70–87, 2014.

114. J. N. Taylor, M. Pirchi, G. Haran, and T. Komatsuzaki, "Deciphering hierarchical features in the energy landscape of adenylate kinase folding/unfolding," *The Journal of chemical physics*, vol. 148, no. 12, p. 123325, 2018.

115. L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley and Sons, 2009.

116. A. Jain and R. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

117. M. Tavakoli, J. N. Taylor, C.-B. Li, T. Komatsuzaki, and S. Pressé, "Single molecule data analysis: an introduction," *arXiv preprint arXiv:1606.00403*, 2016.

118. F. E. Ruiz, P. S. Pérez, and B. I. Bonev, *Information theory in computer vision and pattern recognition*. Springer Science & Business Media, 2009.

119. R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *Information Theory, IEEE Transactions on*, vol. 18, no. 4, pp. 460–473, 1972.

120. S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

121. S. Still and W. Bialek, "How many clusters? an information-theoretic perspective.," *Neural computation*, vol. 16, no. 12, pp. 2483–506, 2004.

122. N. Slonim, G. S. Atwal, G. Tkačik, and W. Bialek, "Information-based clustering," *Proceedings of the National Academy of Sciences*, vol. 102, no. 51, pp. 18297–18302, 2005.

123. H. Akaike, "A new look at the statistical model identification," in *Selected Papers of Hirotugu Akaike*, pp. 215–222, Springer, 1974.

124. G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

125. J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

126. J. Taylor, "An introduction to error analysis: the study of uncertainties in physical measurements. univ," *Science, Sausalito, CA*, vol. 45, p. 92, 1997.

127. K. O. Arras, "An introduction to error propagation: derivation, meaning and examples of equation cy= fx cx fxt," tech. rep., ETH Zurich, 1998.

128. M. Rouaud, "Probability, statistics and estimation: Propagation of uncertainties in experimental measurement," 2017.

129. P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to data mining," 2006.

130. P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

131. A. Rygula, K. Majzner, K. M. Marzec, A. Kaczor, M. Pilarczyk, and M. Baranska, "Raman spectroscopy of proteins: a review," *Journal of Raman Spectroscopy*, vol. 44, no. 8, pp. 1061–1076, 2013.

132. S. L. Friedman, "Hepatic stellate cells: protean, multifunctional, and enigmatic cells of the liver," *Physiological reviews*, vol. 88, no. 1, pp. 125–172, 2008.

133. A. Geerts, "History, heterogeneity, developmental biology, and functions of quiescent hepatic stellate cells," in *Seminars in liver disease*, vol. 21, pp. 311–336, Copyright© 2001 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA, 2001.