A machine learning approach to analysis and classification of measurements in cultural heritage

MSc Thesis

Student

Vasileios Sevetlidis

This work is devoted to all those of you who are the closest to me.
Your support inspires me.

PLAGIARISM STATEMENT I certify that this assignment is my own work, based on my personal study and/or research and that I have acknowledged all material and sources used in its preparation, whether they be books, articles, reports, lecture notes, and any other kind of document, electronic or personal communication. I also certify that this assignment has not previously been submitted for assessment in any other unit, except where specific permission has been granted from all unit coordinators involved, or at any other time in this unit, and that I have not copied in part or whole or otherwise plagiarised the work of other students and/or persons.

this is a LATEX document created with the Overleaf online editor, using the AAU report template.

Vasileios Sevetlidis
Omirou 36, GR-54638
Thessaloniki,
Greece

Title:
A machine learning approach to analysis and classification of measurements in cultural heritage

Project Period:
Spring Semester 2018

Student:
Vasileios Sevetlidis

Supervisor:
Dr. George Pavlides

Co-Supervisor:
Prof. Ioannis Liritzis

Committee:
Dr. Anestis Koutsoudis
Prof. Spiros Vosinakis

Copies:
5

Page Numbers: 59

Date of Completion:
JUNE 11, 2018

Abstract:

Treatment of spectral information is an essential tool for the examination of various cultural heritage materials. Raman Spectroscopy has become an everyday practice for compound identification due to its non-intrusive nature, but often it can be a complex operation. Spectral identification and analysis on artists' materials is being done with the aid of already existing spectral databases and spectrum matching algorithms. We demonstrate that with a machine learning method called Extremely Randomised Trees, we can learn a model in a supervised learning fashion, able to accurately match an entire-spectrum range into its respective mineral. Our approach was tested and was found to outperform the state-of-the-art methods on the corrected RRUFF dataset, while maintaining low computational complexity and inherently supporting parallelisation.

Contents

Chapter 1

Introduction

Treatment of spectral information is an active field of ongoing research. Mineral identification is a fundamental step in a wide range of analyses and applications, such as planetary exploration, geological field expeditions, materials research, medical diagnosis, etc. Especially in the field of cultural heritage, the identification of artists' materials of mineral origin, such as traditional paint pigments, is a valuable tool. One successful spectroscopy method in cultural heritage, Raman spectroscopy, has become a daily practice, due to its non-destructive and non-invasive nature [1].

Raman spectroscopy is a method that exploits the monochromatic light interaction with the vibrations in molecules that results in energy shifts of the source light, which is being detected to provide information about vibrational, rotational or other low-frequency modes [2, 3]. Conservation of energy during the interaction of the source light with the material in study describes the shift in energy of the source light in respect to the change of state of the material. Still, for the material to exhibit the desired Raman effect (or Raman shift), a change in its polarisability connected with the vibrational coordinate should occur, and the detected effect is proportional to this change. Typically, Raman shifts are described in wavenumbers, and conversion from spectral wavelength is based on

$$\Delta w = \left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1} \right) \tag{1.1}$$

Δw being the the Raman shift typically expressed in inverse centimetres (cm^{-1}), λ_0 the excitation wavelength, and λ_1 the spectrum wavelength. Raman spectra can be recorded over a range of $4000 - 10, \mathrm{cm}^{-1}$ [4].

Experts in heritage science use the technique of Raman spectroscopy under controlled experimental conditions, such that the materials under examination are not damaged in any way. A Raman spectrum can be considered as a fingerprint that could be used for compound identification, and is effective and useful when a database of standard spectra is available for comparison. However, the matching of spectra with respective minerals can be complex. Limitations can be introduced as a result of the equipment involved, leading in loss of focus during analysis in relatively light objects or in size restrictions due to the use of traditional microscopes that prohibit the analysis of large objects. In addition, there are limitations inherent within the Raman technique like the effect of fluorescence. Often, the laser beam can excite electronic transitions that may mask the Raman signal. Thus, analysis of spectra requires detailed knowledge of group theory and involves lengthy calculation. There is a growing volume of scientific publications regarding the application of Raman spectroscopy in heritage science that shows the great interest of the community on the method and the benefits it can bring to the domain [5, 6, 7, 8].

In practice, in order to analyse a spectrum's identity scientists rely either on matching software e.g., CrystalSleuth or on direct manual comparisons with a database of reference spectra. These types of identification have critical limitations, as there is not a unique measured spectrum that can be considered as a mineral's true identity, since the impurity of chemical mixtures can result in differences in the measured spectra. Thus, there is no database that can be entirely comprehensive, or no software that can unequivocally claim that any spectrum is a *perfect match*.

1.1 Motivation

Raman spectroscopy has advanced in recent years where its use in both industry and academia has increased significantly, and is drawing an increasing attention of experts in biomedical and pharmaceutical research. Only in the past decade there was an explosion of published research regarding this technique due to technological advancements in instrumentation has decreased the cost and enhanced the user-friendliness, such that there was not required a user to be a laser specialist any more. Alongside the instrumentation improvements, there was also the development of Raman techniques such as SERS and TERS (surface-enhanced Raman scattering and tip-enhanced Raman scattering, respectively) and many other exciting variants of the 'normal' Raman technique that assisted on the researchers' interest growth. Novel applications of Raman spectroscopy variants, illustrate the diverse nature of this technique and its extraordinary ability to be involved in the solutions of many biological problems that concern the pharmaceutical industry, forensics and medicine. Raman spectroscopy is now well established as a complementary technique to much of the analytical instrumentation currently available.

In cultural heritage the Raman spectroscopy presents several desirable attributes, which make this spectroscopic method as a routinely used tool by the experts. These attributes are summarized bellow [6]:

- Non-destructive analysis of materials

- No chemical, mechanical preparation or desiccation necessary for specimens prior to analysis
- Assessment of degradation for buried artefacts and human tissues which informs the previous depositional history
- Specific biomolecular spectral marker recognition for the identification of genuine and fake objects which may have been made for fraudulent purposes
- Geographical sourcing of ethnobotanical and biological materials, such as resins

Given this widespread use of Raman spectroscopy, we found our motivation in providing a method that accurately identifies a mineral from its Raman spectrum. Every compound has its own unique Raman spectrum which can be used for sample detection and quantification. The differences in energy between the incident photons (usually provided by a laser) and the scattered photons correspond to vibrations in the molecule or crystal, and provide a "fingerprint" of the sample's composition and molecular structure [7]. Ideally, all samples coming from different specimens but of the same mineral species should have similar fingerprints. However, in practice irregular compound mixture impurities found in samples' composition make the identification of minerals a difficult task.

1.2 Objectives

The scope of this project is to investigate the effectiveness of ensemble learning techniques within the domain of mineral spectral identification. We will examine the existing use of tree based algorithms and then use this established information to experiment with current methods and a novel approach. We wish to evaluate how different ensemble learning techniques are suited to the problem, how the various parameters involved affect the outcome and whether this leads to a framework that could be used as a daily practice tool by experts.

Overall, the very nature of Raman Spectra leads this research to face many interesting challenges, mentioning a few, the noise within the measured spectra with low intensities may mistakenly matched with other spectra in the database due to the similarities in the vibrations produced by that noise, the existing methods for mineral matching are either computationally demanding or trivial and assumptive. Thus, we aim at providing a method that is able to accurately discriminate the intensity irregularities due to induced noise and correctly identify the species or other label groupings using the mineral's spectrum as a fingerprint.

1.3 Contributions

In our work we propose the use of two off-the-shelf machine learning methods, namely the *Random Forest* [9] and the Extremely Randomized Trees [10] for the task of mineral recognition on already baseline corrected Raman spectra. The Random Forest (RF) and the Extremely Randomized Trees (XT) are *ensemble learning methods* based on multiple decision trees that they are particularly successful in tackling with the problem of overfitting. The ensemble learning methods gained a significant attention in the machine learning community over the last decade, prior to the widespread popularity of the CNN. Many works suggested that when compared with PCA-LDA and RBF SVM on Raman microspectroscopy data [11, 12, 13] it performed poorly. However, in this paper we demonstrate that with the proper treatment of the data (i.e., preprocessing and augmentation) XT becomes particularly efficient in correctly classifying the spectra into their respective minerals. In addition, we utilize the RRUFF database for the evaluation of our experiment. We tested our algorithm against a large number of spectra following the experimental setup of [12]. We compared Random Forest and Extremely Randomized Trees with the following methods: (i) 1-nearest neighbor classifier, which it is believed CrystalSleuth's matching software is based upon [12, 13]; (ii) weighted neighbor (WN) classifier; (iii) vector metric, the cosine similarity after normalization and squashing [12] and with the work of [13], which is the current state-of-the-art. We found that our method is not only competitive in terms of accuracy, but it also does not sacrifice computational speed during training or testing nor it is a computational demanding procedure.

1.4 Publications and Other Submitted Manuscripts

Parts of this thesis have been submitted for publication in a conference and a journal. The paper titled "**Hierarchical classification for improved compound identification in Raman spectroscopy**" has been **accepted** at the *CAA-Gr 2018-Spreading Excellence in Computer Applications for Archaeology and Cultural Heritage*. Moreover, a paper submitted to the *Journal of Cultural Heritage* is pending review.

1.5 Outline of Contents

We conclude this introduction with a brief outline of the contents of our paper.

- Chapter 2 will aim to introduce the relevant background material covering the Raman Effect, the Raman Spectroscopy, the relevant work in mineral identification via Raman Spectra, and the necessary machine learning concepts that we will make use of.

- In chapter 3, we will propose our novel approach to extending the ensemble learning framework into the domain of mineral identification.
- Chapter 4 provides a documentation of our design choices for the proposed systems, as well as we are justifying our design choices for the system.
- Chapter 5 concludes the project with discussions regarding the success of our objectives and a summary of our achievements.

Chapter 2

Background Material

2.1 Raman Spectroscopy

2.1.1 Basic Principles

Raman spectroscopy was discovered by C. V. Raman (Figure 2.1) and K.S. Krishnan in 1928, which gained the former the Nobel Prize for Physics in 1930. Naturally, the method, which was named after him in his honor, was praised by the scientific community as one of the most important discoveries in physics made up to that time.

Raman spectroscopy (RS) is a versatile method for analysis of a wide range of forensic samples. It resolves most of limitations of other spectroscopic techniques, for instance Water can be used as a solvent, the sample preparation is not very elaborate since it can be in any state, it gives an indication of covalent character in the molecule. It can be used for both qualitative as well as quantitative purpose. Qualitative analysis can be performed by measuring the frequency of scattered radiations while quantitative analysis can be performed by measuring the intensity of scattered radiations, and Raman Spectroscopy needs relative short time to perform detections [14, 15].

Figure 2.1: C. V. Raman

Primarily, Raman spectroscopy is a

scattering technique. It is based on Raman Effect, i.e., frequency of a small fraction of scattered radiation is different from frequency of monochromatic incident radiation. It is based on the inelastic scattering of incident radiation through its interaction with vibrating molecules [15, 16, 17]. A Raman spectrum is presented as an **intensity vs. wavelength shift** [16]. Raman spectra can be recorded over a range of 4000–10 cm^{-1} [18]. However, Raman active normal modes of vibration of organic molecules occur in the range of 4000– 400Δcm^{-1}. Depending on spectrophotometer's design and optical components, typical Raman spectra cover the wavenumber region between 400–5Δcm^{-1} and 4000– 3800Δcm^{-1} [17]. A Raman spectrum is significantly simpler than their Infrared (IR) counterparts because in normal Raman overtones, combination and difference bands are rare [15, 16]

When a sample's molecules are being illuminated by the beam of a monochromatic laser (radiation) with a wavenumber $\tilde{\nu_0}$ is incident on systems the phenomenon of scattering light (at all directions) is being observed. If the frequency content of the scattered radiation is analyzed, there will be observed to be present not only the wavenumber $\tilde{\nu_0}$ associated with the incident radiation but also, in general, pairs of new wavenumbers of the type $\nu^L = \tilde{\nu_0} + \tilde{\nu_M}$. Principally, the wavenumbers $\tilde{\nu_M}$ belong in the ranges associated with the transitions between rotational, vibrational, and electronic levels. Since the scattered light has different frequency from that of incident light (inelastic scattering), it is used to construct a Raman spectrum. Therefore, Raman spectra arise due to inelastic collision between incident monochromatic radiation and the sample's molecules.

The origin of the modified frequencies found in Raman scattering is explained in terms of energy transfer between the scattering system and the incident radiation. When a system interacts with radiation of wavenumber $\tilde{\nu_0}$, it makes an upward transition from a lower energy level E_1 to an upper energy level E_2. It must then acquire the necessary energy, $\Delta E = E_2 E_1$, from the incident radiation. The energy ΔE is expressed in terms of a wavenumber $\tilde{\nu_M}$ associated with the two levels involved, where:

$$\Delta E = hc\tilde{\nu_M} \tag{2.1}$$

This energy requirement is regarded as being provided by the absorption of one photon of the incident radiation of energy $hc\tilde{\nu_0}$ and the simultaneous emission of a photon of smaller energy $hc(\tilde{\nu_0}) - \tilde{\nu_M})$, so that scattering of radiation of lower wavenumber, $\tilde{\nu_0}) - \tilde{\nu_M}$, occurs. Alternatively, the interaction of the radiation with the system may cause a downward transition from a higher energy level E_2 to a lower energy level E_1, in which case it makes available energy

$$E_2 - E_1 = hc\tilde{\nu_M} \tag{2.2}$$

Again a photon of the incident radiation of energy $hc\tilde{\nu_0}$ and the simultaneous emission of a photon of higher energy $hc(\tilde{\nu_0}) + \tilde{\nu_M})$, so that scattering of radiation of higher wavenumber, $\tilde{\nu_0} + \tilde{\nu_M}$, occurs.

Much of this scattered radiation has a frequency which is equal to the frequency of the incident radiation and constitutes the **Rayleigh scattering**. Even though there is no effective difference in the energy state of the system, it still contributes directly in the scattering act, resulting to the simultaneous absorption and emission of the incident radiation's photon $hc\tilde{\nu_0}$ so that the scattering radiation of unchanged wavenumber, $hc\tilde{\nu_0}$, occurs.

It is clear that, as far as wavenumber is concerned, a Raman band is to be characterized not by its absolute wavenumber, $\nu^L = \tilde{\nu_0} + \tilde{\nu_M}$, but by the magnitude of its wavenumber shift $\tilde{\nu_M}$ from the incident wavenumber. The small fraction of scattered radiation which has different frequency than the incident's it constitutes the **Raman scattering** as such wavenumber shifts are often referred to as Raman wavenumbers. When the frequency of incident radiation is higher than the frequency of scattered radiation **Stokes lines** appear in the Raman spectrum. Otherwise, when the frequency of incident radiation is lower than the frequency of scattered radiation, **anti-Stokes lines** appear in the Raman spectrum. Scattered radiation is usually measured at right angle to incident radiation [14, 15]. Where it is necessary to distinguish Stokes and anti-Stokes Raman scattering we shall define $\Delta\tilde{\nu}$ to be positive for Stokes scattering and negative for anti-Stokes scattering, that is $\Delta\tilde{\nu} = \tilde{\nu_0} + \nu^L$, as shown in Figure 2.2[1].

In summary, the magnitude of Raman shifts does not depend on wavelength of incident radiation [14]. Raman scattering depends on wavelength of incident radiation [16]. A change in polarizability during molecular vibration is an essential requirement to obtain Raman spectrum of sample. The amplitude of the vibration is called the nuclear displacement. The monochromatic laser beam induces excitation of molecules which transforms them into oscillating dipoles emitting light of three different frequencies, as described bellow:

1. A molecule with no Raman-active modes absorbs a photon with the frequency $\tilde{\nu_0}$. The excited molecule returns back to the same basic vibrational state and emits light with the same frequency $\tilde{\nu_0}$ as an excitation source. This type if interaction is called an elastic *Rayleigh scattering*. About 99.999% of all incident photons in spontaneous Raman undergo elastic *Rayleigh* scattering. This type of signal is useless for practical purposes of molecular characterization. Only about 0.001% of the incident light produces inelastic Raman signal. Spontaneous Raman scattering is very weak and special measures should be taken to distinguish it from the predominant *Rayleigh* scattering. Instruments such as notch filters, tunable filters, laser stop apertures, double

[1] based on work of Moxfyre and User:Pavlina2.0 vectorization of File:Raman energy levels.jpg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=7845122

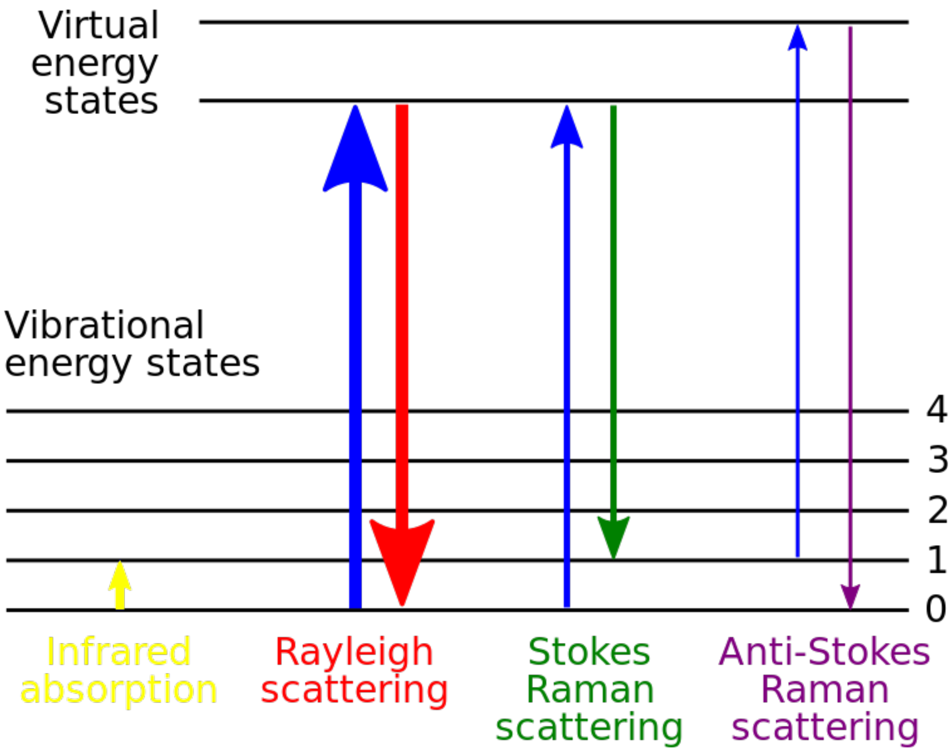

Figure 2.2: Energy transfer model of Rayleigh scattering, Stokes Raman and anti-Stokes Raman scattering. The blue arrows represent $\tilde{\nu_0}$, the red arrow $\tilde{\nu_0}$, the green arrow $\tilde{\nu_0} - \tilde{\nu_M}$ and the purple $\tilde{\nu_0} + \tilde{\nu_M}$ energy transfers, while the yellow arrow represents the infrared absorption $\tilde{\nu_M}$.

and triple spectrometric systems are used to reduce *Rayleigh* scattering and obtain high-quality Raman spectra.

2. A photon with frequency $\tilde{\nu_0}$ is absorbed by Raman-active molecule which at the time of interaction is in the basic vibrational state. Part of the photon's energy is transferred to the Raman-active mode with frequency $\tilde{\nu_M}$ and the resulting frequency of scattered light is reduced to $\tilde{\nu_0} - \tilde{\nu_M}$. This Raman frequency is called Stokes frequency, or just "Stokes". Stokes bands are more intense than *anti-Stokes* bands and hence are measured in conventional Raman spectroscopy [14, 19]
3. A photon with frequency $\tilde{\nu_0}$ is absorbed by a Raman-active molecule, which, at the time of interaction, is already in the excited vibrational state. Excessive energy of excited Raman active mode is released, molecule returns to the basic vibrational state and the resulting frequency of scattered light goes up to $\tilde{\nu_0} + \tilde{\nu_M}$. This Raman frequency is called AntiStokes frequency, or just "Anti-Stokes". *Anti-Stokes* bands are measured with fluorescing samples because

fluorescence causes interference with Stokes bands.

2.1.2 Instrumentation

A sample is normally illuminated with a laser beam in the ultraviolet (UV), visible (Vis) or near infrared (NIR) range. Scattered light is collected with a lens and is sent through interference filter or spectrophotometer to obtain Raman spectrum of a sample. A Raman system typically consists of four major components:

1. Excitation source (Laser)
2. Sample illumination system and light collection optics.
3. Wavelength selector (Filter or Spectrophotometer).
4. Detector (Photodiode array, CCD or PMT).

Excitation source

Raman's fascination about the phenomenon of light scattering lead him to an extensive series of measurements of scattered light. Much of the early work on the analysis of light scattered by a liquid was done by the visual observation of color rather than precise measurements of the light's wavelength, as shown in Figure 2.3

In 1927 Raman obtained a seven-inch (18 cm) refracting telescope, which enabled him to condense the sunlight and create a more powerful light source for his studies. Also, during the same time period, mercury arc lamps were commercially available, and he switched to this even more intense light source. This type of light source was used until the 60's. Nevertheless, it was evident that the more intense the light sources the better the observations they could make. A new invention of light sources at the end of the 60's, namely the laser, provided an even more intense source of light that not only could serve as a probe exploring the properties of the molecule but could also induce dramatically new effects. Laser sources completely replaced the mercury arc lamps, since they could provide a constant, stable and more intense beam of radiation. Table 2.1 summarizes the wavelengths provided by different kind of light sources.

Note that, short wavelength sources such as argon ion and krypton ion lasers can produce significant fluorescence and cause photodecomposition of the sample. However, long wavelength sources such as diode or Nd:YAG lasers can be operated at much higher power without causing photodecomposition of sample and eliminates or reduces fluorescence in most cases [20, 21].

As visual and qualitative observations alone would not be sufficient information there was a need to measure the exact wavelengths of the incident and Raman scattering. Thus, they replaced the observer with a pocket spectroscope, which was later replaced by a quartz spectrograph. The quartz spectrograph could photograph the spectrum of the scattered light and enable the measurement of its

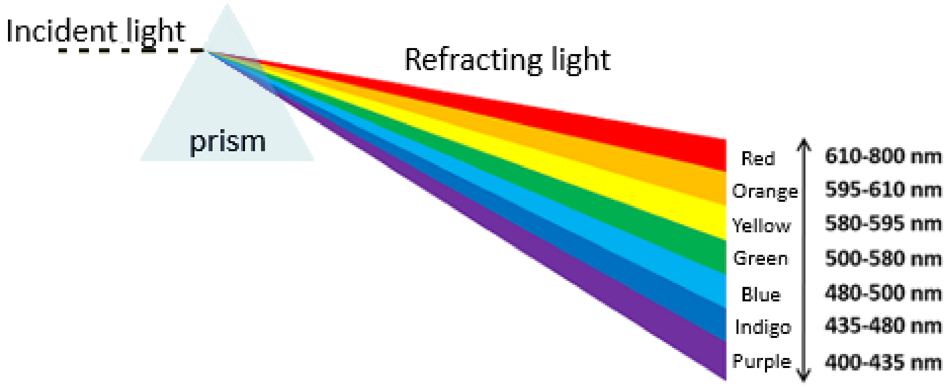

Figure 2.3: A triangular prism, dispersing light; waves shown to illustrate the differing wavelengths of light.

Table 2.1: Wavelengths for different light sources

Light Source	Wavelength (nm)
Mercury arc lamps	435.8
Argon ion laser	488 and 514.5
Krypton ion laser	530.9 and 647.1
Near Infrared (IR) diode lasers	785 and 830
Neodymium-Ytrium Aluminum Garnet (Nd:YAG)	1064
Neodymium-Ytrium Ortho-Vanadate (Nd:YVO4)	1064
Frequency doubled Nd:YAG diode lasers	522
Frequency doubled Nd:YVO4 diode lasers	532

wavelength. A quartz spectrograph is shown in Figure 2.4[2].

Light collection

It is usually necessary to "clean up" the laser output, which may consist of discrete plasma lines or broadband fluorescence background in addition to the principal laser line. If the laser is delivered to the sample with an optical fiber, it is especially critical to remove the Raman modes and fluorescence that are excited in the silica fiber by the laser.

[2] Courtesy the Indian Association for the Cultivation of Science

Figure 2.4: A quartz spectrograph

Background removal can be accomplished in several ways. If there is sufficient space between the laser output and sample, spatial filtering may be used. In this method, an optical element is used to disperse the laser output, which then travels some distance and passes through an aperture. Ideally, this physically blocks all but the laser line of interest. Another option for cleaning up the laser beam is to send it through a monochromator set to pass only the line of interest. This has the disadvantage of low throughput of the laser line. Perhaps the simplest and most cost-effective method is the use of an interference filter which passes only the laser line with transmission usually > 80%. Interference filters are also known as laser line filters or bandpass filters.

There are four basic types of filters to choose from: a long wave pass (LWP) edge filter, a short wave pass (SWP) edge filter, a notch filter, and a laser line filter (each shown below). Laser line filters transmit only the laser and block all other light, while notch filter block only the laser line while passing both long and shorter wavelengths. By using these two filters together, both Stokes and Anti-Stokes Raman scattering can be measured simultaneously.

People use commercially available interference (notch) filters which cut-off spectral range of $\pm 80 - 120cm^{-1}$ from the laser line. This method is efficient in stray light elimination but it does not allow detection of low-frequency Raman modes.

Wavelength selector

The wavelength selector is the most critical component in a Raman spectrometer, through which the intensity information of individual frequencies is extracted.

There are basically two types of wavelength selection mechanisms, **dispersive** and **non-dispersive**. A dispersive spectrometer uses prism or grating since it relies on e dispersive components to separate light spatially according to the wavelength. Stray light is generated in the spectrometer mainly upon light dispersion on gratings and strongly depends on grating quality. Raman spectrometers typically use holographic gratings which normally have much less manufacturing defects in their structure then the ruled once. Stray light produced by holographic gratings is about an order of magnitude less intense then from ruled gratings of the same groove density. Using multiple dispersion stages is another way of stray light reduction. Double and triple spectrometers allow taking Raman spectra without use of notch filters. In such systems Raman-active modes with frequencies as low as 3-5 cm-1 can be efficiently detected. The non-dispersive spectrometer selects light either with an interferometer such as Michelson interferometer in Fourier Transform Raman spectrophotometer or by an optical filter. The measured signal is the interferogram in time domain, and the Raman spectrum can be obtained by the Fourier transformation of the interferogram. Compared with the dispersive Raman spectrometer, the FT-Raman spectrometer has a higher throughput, excellent frequency accuracy and precision, and higher resolution. FT-Raman spectrometers are mainly used when samples fluoresce, such as in forensic analysis [22] and pharmaceutical applications [23] because Raman scattering efficiency in the longer wavelength (NIR) region is lower than that in the short wavelength (visible) region. The lower Raman scattering efficiency limits the sensitivity of the FT-Raman spectrometer, which is important in applications e. g., to detect water contaminants [24].

Commercial Fourier Transform-Raman spectrophotometers (FT-Raman) were introduced in late 1980's to improve the detection system capable of overcoming the limitations of CCD and other detectors for operating in the near-IR region when using 1064nm laser excitation [25]. FT-Raman spectrophotometer uses a Michelson interferometer and continuous wave laser such as Nd–YAG which emits the radiation at 1064 nm. In GaAs and germanium (Ge) detectors are operated at cryogenic temperatures in order to reduce noise and thus raise the signal-to-noise ratio [14]. Cryogenic temperature is a temperature at which molecular motion comes as close as theoretically possible to ceasing completely. At cryogenic temperature, materials are as close to a static and highly ordered state as is possible [15]. Since water absorbs in the 1000nm region, aqueous samples cannot be analyzed by FT-Raman spectrophotometer [16, 15].

Depending on the area of use, Raman spectrophotometers can be categorized into two broad classes: lab based spectrophotometers and in-field, in-situ or down-field use Raman spectrophotometers which include portable and hand-held devices or remote or stand-off systems [17]. The basic principle is same in each case and these systems are differentiated by versatility of an instrument and size and relative cost of its components. More compact components are used in on-site Raman spectrophotometers. Benchtop, handheld, portable, remote or stand-off Raman spectrophotometers are available for on-site analysis and research purpose [17].

Detectors

World War II was the point of turn for using the Raman spectroscopy as later it became quickly the basic analytical tool.Before that, infrared spectroscopy was improved with the use of sensitive detectors. Raman spectroscopy was not a match to the infrared technique until another technological advancement, namely the laser. The development of laser had a major impact to Raman spectroscopy as it had formed a new beginning for the method in the 1960s. Infrared measurements became daily practice even for non-experts. On the other hand the application of the Raman technique still required trained operators and darkroom conditions. Data handling with computers and the the Fourier transform (FT) technique, led to commercial FT-Raman spectrometers in the late 1980s, resulting in the rebirth of the original Raman Effect.

Because of the low Raman scattering efficiency, detection of the Raman signal is very challenging, and the detector should be sensitive. The detectors exploit the photoelectric effect which uses the incoming light energy to generate charge carriers that are separated and can subsequently be measured as a current at the terminals. Two key parameters associated with a detector are the quantum efficiency (QE) and the noise. QE defines the efficiency of a detector to convert optical photons to free charges and noise refers to the dark current caused by the thermal generated charge carriers. Accordingly, to observe the weak Raman signal, the detector should have high QE in the related wavelength band, low noise level and high dynamic range. To date, several types of detectors have been successfully used in Raman spectrometers, and most of them are discussed in the following subsections [24].

Single-point detectors such as Thermoelectrically cooled photomultiplier tubes (PMT) was the mainstream component which early experts used to utilize. However, it was time consuming the procedure of collecting a single Raman spectrum with a PMT detector in wavenumber scanning mode. Nowadays, advances in instrumentation and technology replace these detectors with more sensitive charge transfer devices (CTDs) such as charge-coupled devices (CCDs) and charge-injection devices (CIDs) to detect the Raman scattered light. These devices act as a detector and used in the form of arrays. In CTD's arrays, photosite converts the incoming optical signal into charge which is integrated and transferred to readout devices.1 Multichannel CCD detectors are used with laser wavelengths of less than 1 lm while single element low band-gap semiconductor such as Germanium (Ge) or Indium–Gallium–Arsenic (InGaAs) detectors are used with laser wavelengths of greater than 1 lm [25, 19]. Nowadays, the factors of sensitivity and performance of CCD detectors are advancing technologically thus the CCD detectors are the component of choice for Raman spectroscopy.

2.1.3 Raman Variants

Raman spectroscopy has seen major recognition by the scientific and industrial communities resulting a wide range of diversity on its applications. Subsequently, different needs had to be met leading to the development of several variations of Raman spectroscopy. The usual purpose is to enhance the sensitivity (e.g., surface-enhanced Raman) or to improve the spatial resolution.

Surface Enhanced Raman Spectroscopy (SERS)

Surface enhanced Raman spectroscopy (SERS) is to date the most efficient Raman technique for very low concentration detection. Since its first observation in 1974, SERS has been broadly researched in academia. SERS is a modified technique in which a sample is adsorbed on a colloidal metallic surface (silver, gold or copper) such that it improves the intensity of Raman signals. Additionally, fluorescence caused by diluents and matrices is being eliminated [16, 26]. A detailed review of SERS including the fundamentals, active substrates and its application can be found in the literature broadly [27, 28, 29, 24].

Time-Gated Raman Spectroscopy

Noise reduction can enhance the Raman scattering by improving the signal-to-noise ratio. Sources of noise in Raman measurement can be either from the detector or from the incoming optical signal. Raman detectors, are responsible for most of the noise being introduced due to the CCDs and APDs. The most efficient way to reduce that effect is to cool the detector during measurement. On the other hand, Rayleigh scattering and fluorescence emission can produce additional noise alone or in combination. The difference between Raman scattering and Rayleigh scattering in terms of wavelength allows to eliminate the influence of Rayleigh scattering by using an optical filter. In contrast, the fluorescence emission band overlaps with the Raman peak for certain excitation wavelengths, which blurs or masks the Raman peaks. Overall, reducing background fluorescence and detector noise are of great importance for a high signal-to-noise ratio [24].

Kerr Gated Raman System

The Kerr gate is the most well-known optical shutter in time-gated Raman spectrometers for its fast response. A Kerr gate with picoseconds (25 ps) response and high repetition rate was proposed as early as the 1970s [30, 24]. As technology advanced the Kerr gate it achieved a response time of 3 ps [31, 24]. Currently, the Kerr gate has become a very popular optical shutter in time-gated Raman detection.

This Kerr gate system consisted of two crossed polarizers, and a Kerr medium was placed between the polarizers. A gating pulse was used to gate on and off the Kerr gate, by varying the polarization orientation of light passing through the Kerr medium. Otherwise, no light could pass the Kerr gate due to the crossed polarizers. In the Kerr gate system, if the short gating pulse temporal overlaps with the Raman signal, then the Raman signal would be able to pass the gate with the fluorescence signal being suppressed.

However, its benefit as a fast response system is also its disadvantage due to its complex setup. The fast response provided by the Kerr gate is sufficient to perform the function of fluorescence rejection limiting it to laboratory use. Nevertheless, this setup has also been used in a diverse spectrum of applications ranging from plant auto-fluorescence [32] to depth profiling of calcifications in breast tissue [33, 24].

Fast Time-Gated Raman Systems

Fluorescence rejection can also be met in fast gated detectors. The most commonly used detector is the intensified CCD (ICCD). Different from the normal CCD, the ICCD can be operated in the time-gated mode performing ultra-sensitive detections. A gain voltage controlled image intensifier tube is positioned in front of the CCD, such that incident photons are multiplied inside the intensifier before being focused onto the detector. The ICCD has been used as an alternative technique to the Kerr gate system. Although not as fast as the Kerr gate system, most of the modern ICCDs can achieve hundreds picoseconds gating width, which is adequate for normal Raman spectroscopy [34, 35, 36, 24].

Portable Raman Spectrometers for Field Applications

For purpose of field applications, a variety of portable Raman spectrometers have been developed in the industry. Comparing with benchtop Raman spectrometers, the portable Raman spectrometers are low cost, light weight, and more compact. These spectrometers can be battery powered with several hours operational time and fast acquisition can be achieved. The 785 nm laser is widely used in these instruments for general purposes of applications. These instruments provide wide spectrum range with˜$10cm^{-1}$ spectral resolution. They can be used for raw material identification or manufacturing process material validation.

2.1.4 Applications

As mentioned before, application of the Raman technique is extremely diverse. Detection and identification of single molecules represents the final goal of trace analysis and is of great scientific and practical interest in many fields, such as physics, chemistry, biology, medicine, pharmacology, materials, and environmental science. Bellow it is mentioned a few of its major applications and contributions.

Pharmaceutical analysis

Raman spectroscopy has evolved to include several variants of the normal dispersive technique. From the perspective of pharmaceutical analysis, Raman scattering has enabled the rapid non-invasive volumetric analysis of pharmaceutical formulations which could lead to many important applications in pharmaceutical settings, including imaging and the quantitative analysis of pharmaceutical tablets and capsules in process and quality control [37, 38]. They have been used in protein pharmaceutical characterization, raw material verification, manufacturing process monitoring and product quality control [39, 40, 41].

Biology

Recently there has been a particular interest in variants of Raman spectroscopy for the investigation of viruses and microorganisms24, in particular bacteria and yeasts for medical and pharmaceutical applications. In an interesting study by Harz et al. [42] Raman techniques such as Raman microscopy, Raman optical activity (ROA), UV-resonance Raman (UVRR)-spectroscopy, SERS and TERS were employed.

The application of these Raman techniques allowed for the analysis of chemical components of cells and sub-cellular regions along with monitoring chemical differences (characterisation) which arose as a result of the growth of microorganisms. The interaction of microorganisms with active pharmaceutical agents was demonstrated which, in combination with chemometric methods, showed that these techniques could be applied to identify microorganisms in microcolonies and on single cells.

Forensic science

Raman spectroscopic techniques have been introduced in forensic science. The portability of Raman spectrometers, enabled both the detection and identification of chemical and biological hazards. Within the contexts of forensic and homeland security, conducting analysis in the field while adapting a non-contact approach to the hazard is the preferred method. In many cases there is a need for non-contact/non-invasive chemical analysis of hazards concealed within non-transparent containers and packaging [43].

Disease diagnosis

A particularly novel variant of the SERS technique named SESORS (surface-enhanced spatially-offset Raman spectroscopy) has also been recently reported in the literature [44]. It enables the detection and identification of vibrational fingerprints within tissue and has obvious applications within medicine. In particular this provides an opportunity to adapt these particles and technique for potential clinical applications for disease diagnosis where a tumour may not be readily accessible, or surgery is too invasive. Furthermore, it has been demonstrated whole-body Raman imaging, nanoparticle pharmacokinetics, multiplexing, and in vivo tumour targeting, using an imaging system adapted for small-animal Raman imaging.

Manufacturing process monitoring

Raman spectroscopy has been used to monitor manufacturing processes also in the petrochemical Chemists can watch paint dry and understand what reactions are occurring as the paint hardens. Using a fiber-optic probe, they can analyze nuclear waste material from a safe distance. Surface-enhanced Raman spectroscopy is used for studying surfaces and reactions on surfaces. Additionally, the incorporation of short wavelength lasers in Raman spectrophotometers opens the doors for use of telecommunications-type optical fibers such as remote-fiber-optic probes which can be operated over long distances (>10 m in some instances) and are well suited for in-situ or on-site analysis of samples. These fiber optic probes can also be used to record the Raman spectra in locations remote from the sample site and thereby prevent the exposure of investigator to hazardous environment [15].

2.1.5 Raman Spectroscopy in Cultural Heritage

Until 1975 infrared spectroscopy was the preferred method for the analysis of art objects. The advent of of the MOLE (Molecular Optical Laser Examiner) [45] changed the scenery dramatically and the employment of the Raman technique was soon adopted for the determination of pigment composition on manuscripts. Not long the technique was employed for the molecular characterization on genuine archaeological materials.

The first reports came in the 90's with samples taken from biodeterioration exposed Renaissance frescoes [46, 47], biodeteriorated cave art [48], archaeologically excavated biomaterials (notably, the mummified skin of Otzi the Alpine Iceman) [49]. Technical advancements such as a wider range of options in the selection of excitation wavelengths, the portability and transportability led to a rapid growth in the field [50]. Hand-held instrumentation efficiently brought the laboratory to the specimen, art work or artefact taken from archaeological excavations and depositional environments [51, 52]. The non-destructive or minimally destructive nature of Raman spectroscopy established the technique as the practice of choice for researchers at the arts/science community. This has given rise to the field of 'forensic art' investigations, which are now seen as an essential prerequisite for the establishment of a holistic analytical portfolio of an art work [17]. In this context, Raman spectroscopic data have been involved in several high-profile case studies [53]. An important role played the capability for mobile Raman spectroscopy. Quite often there are large sample collections that an expert needs to examine at the collection's environment. Each artifact's uniqueness and frangibility might prohibit their movement (i.e., parts of architectural monuments [54]) or risk exposure to different environments. On-site measurements are requested for many objects which cannot be sampled and moved outside museums [55]. Certainly, it is easier to reach each object than to gather all in one lab. Nevertheless, Raman spectroscopy is used extensively for archaeometric purposes, such as color and pigment characterization, degradation of materials studies, forensic art investigations.

Colors, Pigments and Technologies

One of the most regular applications of Raman spectroscopy is the characterization of pigments and binding media. The synthesis and properties of pigments have been studied extensively using this technique, answering questions such as the structure and stability of pigments [56]. A great variety of studies both in terms of time span (e. g., rock art [57]) and cultural diversity have been carried out (e. g., paintings [58]). Often, Raman analysis is part of the general analytical strategy complementing other measurement techniques such as SEM-EDS, XRF, LIBS or XRD.

The study of color manufacturing technologies concerns also ceramics. Raman spectroscopy assists the documentation process of technical and production aspects in Mediterranean Antiquity [59], North America prehistory [60], renaissance [61] and contemporary art [7, 62]

Studies Dealing with Degradation Processes

Conservation of cultural heritage is of great importance. The field which studies the degradation processes is under development as it constitutes a technical and economical challenge for present and future societies. Additionally, pigment identification on partly or highly damaged materials reinforces the attempt to get a better idea of the original object [63]. Degradation of numerous materials is explored in both cases of organic to mineral substances. Waterlogged wood degradation and their conservation treatments optimisation was studied in [64]. Paper [65] and textile [66] degradation states can be dealt with the use of Raman spectroscopy as a diagnostic function. A study of complementary spectroscopy techniques (e. g., IR and Raman) has been presented for the comprehension of alteration mechanism in fossil samples like amber [67]. Raman spectroscopy has been used to revise iron atmospheric corrosion mechanisms [68]. Sharp description of material micro-heterogeneity and organization can be accessed by Raman structural imaging. These features are closely related to degradation mechanisms and structural imaging for low crystallized phases can suggest methods to improve mechanisms descriptions [7].

An other example of the use of Raman spectroscopy is on hair samples. Analysis of hair samples contibute on the evaluation of historical and archaeological biodeterioration in depositional environments [69]; hair consists of keratin based proteins, which can survive for an extensive time in adverse burial conditions. Other features present in the Raman spectra of the historical hair sample can be attributed to the identification of additives and cosmetics in use in the mid-nineteenth century.

Disambiguation has been gratified on the opinions regarding a unique specimen of an eye-bead from an Egyptian eighteenth dynasty cat-mummy (1350 BC), which it was believed to be be either amber or brown glass by the archaeologists. The Raman spectrum indicated to be neither [70], as it presented the characteristics of a keratin based material, which closely matched that of an animal claw or horn. Thus, the assumptions ware altered leading to the belief that the cat's claws had been heat-processed into a bead [6].

Forensic art investigations

Forensic archeology involves the use of archaeological principles and techniques for the location, recovery, and interpretation of evidence for past events within the constraints of the criminal justice system. The need for a relatively rapid, non-destructive analytical method is paramount for the enforcement and prevention of illegal trafficking. As an example, ivory has been highly valued for many centuries as a practical and decorative art medium. Raman spectroscopy can provide a discrimination analysis on mammalian species from which information on trade routes of ancient cultured can be acquired as well as it can hint geographical sources of seized contraband materials in modern forensic science scenarios. In the case of ivory, Raman spectra can be acquired from suspicious specimens leading to the discrimination between modern ivories and specimens of ivory which have been stored in museum collections from their Raman band signatures [71, 6].

The use of Raman spectroscopy as a botanical discriminator aids the geographical sourcing of ancient resins on museum artefacts, which can provide the archaeologist and historian with information about ancient trade routes and cultures [50, 6]. Dragon's blood resins have been known in decorative art and medicine in ancient cultures for several thousand years. initially coming from an East African source on the island of Socotra, the Dracaena cinnabari resin gives a significantly different spectrum to that of Daemonorops draco, which was used in medieval times. A familiar forensic application is generated here by the non-destructive Raman identification of a genuine resin and the fake or counterfeit alternatives being marketed [6].

2.2 Mineral Identification

The first chemical analyses of artworks and archaeological artefacts were accomplished historically in the early 18^{th} century by Rene-Antoine Ferchault de Réaumur [?, 50] who focused on pottery, glass and porcelains. A century later Sir Humphry Davy [72] performed pigment identification on wall painting fragments. The identification of the pigments used by ancient artists, in both studies led to the complete destruction of the specimens was undertaken. [53] A highly regarded advantage of Raman spectroscopy is its non-destructive nature. The determination of the substances and their state can provide conservators with a warning notice for emergency restoration. Additionally, Raman spectra can provide a disambiguation mechanism for unique materials for which the experts' prior knowledge might be limited. Since the Raman spectra give information about the organic and inorganic components of a specimen, it has been possible to identify areas of wall-paintings, frescoes and rock art. However, in each case, characteristic Raman band signatures were provided for these minerals with established databases such as those in the literature [50, 73, 6].

2.2.1 Related Work

While early efforts relied on expert knowledge of spectral features, more recent approaches have made use of a wide range of statistical and machine learning tools. Specifically in the geosciences, pioneering work in automated mineral identification has been performed on ExoMars [74, 75, 76], building upon earlier insights [77]. Automated identification of minerals using univariate analysis has been studied [78, 79, 80], but it is not fully adaptable to mineral mixtures.

Some early approaches were limited to the task of identifying specific components [81, 82], while others attempted to cluster spectra into logical groups [83, 84]. The majority of methods perform a variety of spectrum preprocessing steps, in order to reduce the influence of noise and fluorescence [85, 86]. Additionally, many studies investigated spectra projections into lower-dimensional feature spaces, using Principal Components Analysis (PCA) [83, 87].

Recently, a wide use of Support Vector Machines (SVMs) [88] has been connected with small-scale medical related investigations. SVM, a powerful machine learning tool, outperformed linear discriminant analysis (LDA) and partial least squares discriminant analysis (PLS-LDA) [89, 90, 91]. Also, SVMs have been used for applications on mineral detection with near-infrared spectroscopy [82], as well as in composition prediction with Raman spectra [92].

Artificial neural networks (ANNs) have seen increasing use in applications in the domain of spectroscopy [84, 93]. Similarity-based methods for both peak-feature [74] and full-spectrum matching [94, 95] have also been explored.

Most of the aforementioned studies set the basis for mineral identification in pure phases, yet their applications were qualified to a specific domain using small-scale customized datasets for their respective tasks. In addition, they haven't utilized the full RRUFF Raman dataset [96] as a benchmark reference set, in order to test algorithms against its large number of spectra.

On the contrary, the most recent work [83, 12] has focused on identifying mineral species contained in RRUFF's larger-scale dataset[3] using nearest neighbor methods with different similarity metrics, such as cosine similarity and correlation. [12] achieved a species classification accuracy on a subset of the RRUFF database of 84.8% using a weighted neighbor (WN) classifier. Square root squashing, maximum intensity normalization, and sigmoid transformations were applied to the data prior to classification. Accuracy was determined using cross validation with semi-randomized splits over a number of trials. The WN classifier compared favorably with the $k = 1$ nearest neighbor (82.1% accuracy). Finally, [13] proposed a deep convolutional neural network (CNN) architecture for Raman spectra recognition and baseline correction. They performed one-leave-out cross-validation for their experiments, while they suggested an augmentation strategy for the training set, in order to meet the data volume requirements for the training of their architecture. The accuracy of spectra recognition on the already baseline corrected data reached 88.4%, while they achieved an accuracy of 93.3%, when they set up their architecture to perform baseline correction on raw data prior to the identification.

2.2.2 The Machine Learning Approach

The term Machine learning was coined by Arthur Samuel in 1959, and it refers to a field which has evolved from the study of pattern recognition and computational learning theory. It is a major branch of the artificial intelligence family [3] and it explores the study and development of algorithms that can learn from data and make predictions [4]. Machine learning is used in a vast range of tasks, where designing an otherwise explicit algorithm customized for solving a particular problem would wield poor performance. These tasks for example span the range of email filtering, to optical character recognition (OCR),[7] and computer vision.

A machine learning algorithm can be expressed as a function $y(x)$ which takes a new input vector x (mentioned also as feature vector) and generates an output vector y, encoded in the same way as the target vectors[REF BISHOP]. The behavior of the function of $y(x)$ is formed during the training phase on the basis of the

[3] it is believed that RRUFF has been also used in commercial software such as CrystalSleuth [12, 13]

training data. After the training phase the model is capable to meet its purpose by recognizing the identity of unseen feature vectors, which they belong in a test set. Usually, machine learning algorithms have the an advantageous characteristic, namely the ability to generalize, that is it has the capacity to perform well the task it was designed for at examples that differ from those that were used during the training phase.

Usually, the original input variables are preprocessed, in order to be transformed into a new space of variables. This serves the purpose of enabling the machine learning algorithm to better recognize the emergent patterns. For instance, a typical problem in computer vision is recognition of objects within images. As images come in many shapes and the objects within them are not always depicted in a similar fashion, the input images are pre-processed in terms of translation and scale so that meet a fixed size. This pre-processing stage is sometimes also called feature extraction. Pre-processing might also be performed in order to speed up computation. Often, there is a need for real-time predictions, thus the features being extracted should contain only the most relevant information in a meaningful way, so that they enable the model to learn how to discriminate them.

Supervised learning problem is known to be the setting where for each input vector its corresponding output vector is known during the training phase and is being used by the algorithm to model the relationship between them. Supervised learning is suitable for tasks such as classification or regression. When the aim is to assign each feature vector to one of a finite number of discrete categories, it called a classification problem. On the other hand, if the desired output consists of one or more continuous variables, then the task is called regression.

In contrast with the supervised learning problems, there is an other category in which the training data consist of a set of input vectors x without any corresponding target values. The goal in such **unsupervised learning problems** may be to discover groups of similar examples within the data, where it is called clustering, or to determine the distribution of data within the input space, known as density estimation, or to project the data from a high-dimensional space down to two or three dimensions for the purpose of visualization. [rEF BISHOP]

Finally, the technique of **reinforcement learning** (Sutton and Barto, 1998) is concerned with the problem of finding suitable actions to take in a given situation in order to maximize a reward. Here the learning algorithm is not given examples of optimal outputs, in contrast to supervised learning, but must instead discover them by a process of trial and error. Typically there is a sequence of states and actions in which the learning algorithm is interacting with its environment.

In this work we will deal with the problem of mineral identification as a supervised learning problem. We will regard as feature vector the measured spectrum itself, while for each spectrum there is a corresponding known a-priori label, namely its species name.

Mineral Identification as a Supervised Learning Problem

Supervised learning is simply a formalization of the idea of learning from examples. In supervised learning, the sample space is divided between two sets of data, a training set and a test set. The idea is for the algorithm to learn from a set of labeled examples in the training set, so that it can identify the label of unseen examples in the test set with the highest possible accuracy. The goal of the learning algorithm is to develop a rule (or a function) that classifies new examples (in the test set) by analyzing examples with class label it has been already given during the training phase.

For example, a training set might consist of images of different types of animals (say, dogs and cats), where the identity of the animal in each image is given to the learning algorithm. The test set would then consist of more unidentified images of animals, but from the same classes. The goal is for the learning algorithm to develop a function that can identify the elements in the test set. There are many different approaches that attempt to build the best possible method of classifying examples of the test set by using the data given in the training set.

Formally, in supervised learning, the training set consists of n ordered pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, where each x_i is some measurement or set of measurements of a single example data point, and y_i is the label for that data point. For example, an x_i might be a group of five measurements (a vector) for a swimmer in a team including height, weight, preferred swimming style, and lap records for different distances. The corresponding y_i might be a classification of the swimmer as "fast" or "not fast". The test data in supervised learning is another set of m measurements without labels: $(x_{n+1}, x_{n+2}, ..., x_{n+m})$. As described above, the goal is to make educated guesses about the labels for the test set (such as "fast" or "not fast") by drawing inferences from the training set.

Following the above example in this work we will consider that each x_i is a mineral's spectrum and the y_i gives the species name of the mineral. To train such a classifier, we would provide sample spectra (a training set) for each type of mineral. Then we would use the classifier by having it label new spectra of minerals (a test set).

Solving the Problem

In this work we propose a supervised learning approach for the problem of mineral identification. The proposed method belongs in the ensemble learning tree based classification algorithms family. As it is discussed in the next chapter a decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter/differentiator in input variables. All the splits are being made with the use of some criteria, such as the entropy or the information gain. At the end of the splitting process the tree can predict the possibility of a sample to belong in a given class. On the other hand, the combination of the opinions gathered from different trees, is called an ensemble. Where the tree models yield poor performance (weak learner), the ensemble is proved to perform better. Therefore, the class of ensemble learning methods is what we deem suitable for solving the aforementioned problem, since we can exploit its merits, while its disadvantages seem not to affect considerably our approach.

2.2.3 Organizational Systems for Minerals

Mineral classification can be an organizational nightmare. With over 3,000 different types of minerals a system is needed to make sense of them all. There are many ways which are in current use to help with the classification of minerals, such as: Dana's New Mineralogy [97], the Strunz classification [98], A Systematic Classification of Minerals [99] and the various volumes of Deer, Howie and Zussman (Rock-forming Minerals series), which use combinations of mineral structure and chemical composition to classify minerals. Mineral species can be grouped in a number of different ways, on the basis of chemistry, crystal structure, occurrence, association, genetic history, or resource, for example, depending on the purpose to be served by the classification. We have chosen to sort the minerals and synthetics using the Dana classification in Dana's New Mineralogy [97] devised by Professor James Dana of Yale University way back in 1848. The Dana system provides an hierarchical organization system as a four-part number that classifies minerals into classes, types, groups, and species accordingly [100].

The hierarchical division is nested, such that each class is described in types, the types into groups and the groups in species.

Chapter 3

Tree structured models

3.1 Decision Trees

A decision tree can be described in many ways depending from which perspective is examined. From graph theory perspective is directional graph, while from information theory is a constant binary division of the feature space. In general, a decision tree is a classifier expressed as a recursive partition of the sample space. Decision trees can be organized in two types. **Categorical Variable Decision Tree:** A Decision Tree which has categorical target variable then it called as categorical variable decision tree. Example:- In above scenario of student problem, where the target variable was "Student will play cricket or not" i.e., YES or NO. **Continuous Variable Decision Tree:** A Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree. The decision tree consists of nodes that form a rooted tree, meaning the root is the beginning of the tree structure. All other nodes have exactly one incoming edge and two outgoing edges leading to the children nodes. The nodes that have no children nodes connected to them are considered to be the terminal nodes (also called leaves). In a decision tree, each internal node splits the sample space into two (or more) sub-spaces according to a certain discrete function. In the simplest and most frequent case, each splitting test considers a single attribute, such that the instance space is partitioned according to the attribute's value.

Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. The samples are classified by traversing them through the tree from the root down to a leaf, according to the outcome of the tests along the path. FIGURE represents a tree, where the internal nodes are represented as black circles, whereas leaves are denoted as squares. Note that a decision tree can deal with both nominal and numeric attributes.

A geometrical interpretation of the numerical attributes in the internal nodes of a decision tree is that of a collection of hyperplanes, each orthogonal to one of the axes. The tree complexity affects its accuracy (breiman 87) and is explicitly influenced by the stopping criteria. The total number of nodes, total number of leaves, tree depth and number of attributes used are the usual measurements for the determination of the tree complexity.

The induction of a decision tree is closely related to rule induction. Each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to form the antecedent part, and taking the leaf's class prediction as the class value. (Quinlan, 1987).

Minimum samples for a node split Defines the minimum number of samples (or observations) which are required in a node to be considered for splitting. Used to control over-fitting. Higher values prevent a model from learning relations which might be highly specific to the particular sample selected for a tree. Too high values can lead to under-fitting hence, it should be tuned using CV. Minimum samples for a terminal node (leaf) Defines the minimum samples (or observations) required in a terminal node or leaf. Used to control over-fitting similar to min samples split. Generally lower values should be chosen for imbalanced class problems because the regions in which the minority class will be in majority will be very small. Maximum depth of tree (vertical depth) The maximum depth of a tree. Used to control over-fitting as higher depth will allow model to learn relations very specific to a particular sample. Should be tuned using CV. Maximum number of terminal nodes The maximum number of terminal nodes or leaves in a tree. Can be defined in place of max depth. Since binary trees are created, a depth of n would produce a maximum of 2^n leaves. Maximum features to consider for split The number of features to consider while searching for a best split. These will be randomly selected. As a thumb-rule, square root of the total number of features works great but we should check upto 30-40% of the total number of features. Higher values can lead to over-fitting but depends on case to case.

In summary, the basic terminology used with Decision trees:

- **Root Node:** It represents entire population or sample and this further gets divided into two or more homogeneous sets.
- **Splitting:** It is a process of dividing a node into two or more sub-nodes.
- **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
- **Leaf/ Terminal Node:** Nodes do not split is called Leaf or Terminal node.
- **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
- **Branch / Sub-Tree:** A sub section of entire tree is called branch or sub-tree.

- **Parent and Child Node:** A node, which is divided into sub-nodes is called parent node of sub-nodes where as sub-nodes are the child of parent node.

3.1.1 Splitting Criteria

The simplest case of discrete splitting functions is of univariate nature e. g., an internal node is split according to the value of a single attribute from the feature vector. Naturally, the inducer searches for the best attribute upon which to split. There are various univariate criteria. These criteria can be characterized in different ways, such as:

- According to the origin of the measure: information theory, dependence, and distance.
- According to the measure structure: impurity based criteria, normalized impurity based criteria and Binary criteria.

Bellow there are mentioned the criteria which are used most frequently.

Impurity-based Criteria

Given a random variable x with k discrete values, distributed according to $P = (p_1, p_2, ..., p_k)$, an impurity measure is a function $\phi : [0,1]^k \rightarrow R$ that satisfies the following conditions:

- $\phi(P) \geq 0$
- $\phi(P)$ is minimum if i such that component $p_i = 1$
- $\phi(P)$ is maximum if $\forall i, 1 \leq i \leq k, p_i = \frac{1}{k}$
- $\phi(P)$ is symmetric with respect to components of P
- $\phi(P)$ is smooth (differentiable everywhere) in its range

Note that if the probability vector has a component of 1 (the variable x gets only one value), then the variable is defined as pure. On the other hand, if all components are equal, the level of impurity reaches maximum. Given a training set S, the probability vector of the target attribute y is defined as:

$$P_y(S) = (\frac{|\sigma_{y=c_1} S|}{|S|}, ..., \frac{|\sigma_{y=c_{d(y)}} S|}{|S|}) \tag{3.1}$$

The goodness–of–split due to discrete attribute a_i is defined as reduction in impurity of the target attribute after partitioning S according to the values $v_{i,j}$ in $d(\alpha_i)$:

$$\Delta\phi(\alpha_i, S) = \phi(P_y(S)) - \sum_{j=1}^{d(\alpha_i)} \frac{|\sigma_{\alpha_i = v_{i,j}} S|}{|S|} \phi(P_y(\dot{\sigma}_{\alpha_i = v_{i,j}} S)) \tag{3.2}$$

Information Gain

Information gain is an impurity-based criterion that uses the entropy measure (origin from information theory) as the impurity measure (Quinlan, 1987).

$$InformationGain(\alpha_i, S) = Entropy(y, S) - \sum_{j=1}^{d(\alpha_i)} \frac{|\sigma_{\alpha_i=v_{i,j}} S|}{|S|} Entropy(\dot{y}, \sigma_{\alpha_i=v_{i,j}} S) \quad (3.3)$$

where:

$$Entropy(y, S) = \sum_{c_j \in d(y)} -\frac{|\sigma_{y=c_j} S|}{|S|} \cdot \log_2 \frac{|\sigma_{y=c_j} S|}{|S|} \quad (3.4)$$

Gini Index

Gini index is an impurity-based criterion that measures the divergences between the probability distributions of the target attribute's values. The Gini index has been used in various works such as (Breiman et al., 1984) and (Gelfand et al., 1991) and it is defined as:

$$Gini(y, S) = 1 - \sum_{c_{j \in d(y)}} (\frac{|\sigma_{y=c_j} S|}{|S|})^2 \quad (3.5)$$

Consequently the evaluation criterion for selecting the attribute a_i is defined as:

$$G(\alpha_i, S) = G(y, S) - \sum_{v_{i,j} \in d(y)} \frac{|\sigma_{\alpha_i=c_j} S|}{|S|} Gini(y, \sigma_{\alpha=v_{i,j}} S) \quad (3.6)$$

3.1.2 Stopping Criteria

The growing phase continues until a stopping criterion is triggered. The following conditions are common stopping rules:

1. All instances in the training set belong to a single value of y.
2. The maximum tree depth has been reached.
3. The number of cases in the terminal node is less than the minimum number of cases for parent nodes.
4. If the node were split, the number of cases in one or more child nodes would be less than the minimum number of cases for child nodes.
5. The best splitting criteria is not greater than a certain threshold.

3.1.3 Advantages vs. Disadvantages of Decision Trees

Advantages

Several advantages of the decision tree as a classification tool have been pointed out in the literature:

1. Decision trees are self–explanatory and when compacted they are also easy to follow. In other words if the decision tree has a reasonable number of leaves, it can be grasped by non professional users. Furthermore decision trees can be converted to a set of rules. Thus, this representation is considered as comprehensible.
2. Decision trees can handle both nominal and numeric input attributes.
3. Decision tree representation is rich enough to represent any discrete value classifier
4. Decision trees are capable of handling datasets that may have errors.
5. Decision trees are capable of handling datasets that may have missing values.

Disadvantages

Decision trees are considered to be a nonparametric method. This means that decision trees have no assumptions about the space distribution and the classifier structure.

On the other hand, decision trees have such disadvantages as (Quinlan, 1993):

1. Most of the algorithms (like ID3 and C4.5) require that the target attribute will have only discrete values.
2. As decision trees use the "divide and conquer" method, they tend to perform well if a few highly relevant attributes exist, but less so if many complex interactions are present. One of the reasons for this is that other classifiers can compactly describe a classifier that would be very challenging to represent using a decision tree. A simple illustration of this phenomenon is the replication problem of decision trees (Pagallo and Huassler, 1990). Since most decision trees divide the instance space into mutually exclusive regions to represent a concept, in some cases the tree should contain several duplications of the same sub-tree in order to represent the classifier. For instance if the concept follows the following binary function: then the minimal univariate decision tree that represents this function is illustrated in Figure 9.3. Note that the tree contains two copies of the same subtree.

3. The greedy characteristic of decision trees leads to another disadvantage that should be pointed out. This is its over–sensitivity to the training set, to irrelevant attributes and to noise.

3.2 Random Forest

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Some random forests reported in the literature have consistently lower generalization error than others. For instance, random split selection (Dieterrich [1998]) does better than bagging. Breiman's introduction of random noise into the outputs (Breiman [1998c]) also does better. But none of these three forests do as well as Adaboost (Freund and Schapire [1996]) or other algorithms that work by adaptive reweighting (arcing) of the training set (see Breiman [1998b], Dieterrich [1998], Bauer and Kohavi [1999]).

To improve accuracy, the randomness injected has to minimize the correlation ρ while maintaining strength. The forests studied here consist of using randomly selected inputs or combinations of inputs at each node to grow each tree. The resulting forests give accuracy that compare favorably with Adaboost. This class of procedures has desirable characteristics:

1. Its accuracy is as good as Adaboost and sometimes better.

2. It's relatively robust to outliers and noise.

3. It's faster than bagging or boosting.

4. It gives useful internal estimates of error, strength, correlation and variable importance.

5. It's simple and easily parallelized.

Following the pioneering work of Breiman [9], bellow it is described the definition of Random Forest in a formal manner as well as the theorem which gives the convergence characteristic of the algorithm .

The common element in all of these procedures is that for the k^{th} tree, a random vector Θ_k is generated, independent of the past random vectors $\Theta_1, ..., \Theta_{k-1}$ but with the same distribution; and a tree is grown using the training set and Θ_k, resulting in a classifier $h(x, k)$ where x is an input vector. For instance, in bagging the random vector Θ is generated as the counts in N boxes resulting from N darts thrown at random at the boxes, where N is number of examples in the training set. In random split selection Θ consists of a number of independent random integers between 1 and K. The nature and dimensionality of Θ depends on its use in tree construction.

After a large number of trees is generated, they vote for the most popular class. We call these procedures random forests.

Definition 3.1 *A random forest is a classifier consisting of a collection of tree-structured classifiers $h(x, \Theta_k), k = 1, ...$ where the Θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.*

Given an ensemble of classifiers $h_1(x), h_2(x), ..., h_K(x)$, and with the training set drawn at random from the distribution of the random vector Y, X, define the margin function as

$$mg(X, Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y}(h_k(X) = j) \tag{3.7}$$

where I is the indicator function. The margin measures the extent to which the average number of votes at X, Y for the right class exceeds the average vote for any other class. The larger the margin, the more confidence in the classification. The generalization error is given by

$$PE^* = P_{X,Y}(mg(X, Y) \leq 0) \tag{3.8}$$

where the subscripts X, Y indicate that the probability is over the X, Y space. In random forests, $h_k(X) = h(X, \Theta k)$. For a large number of trees, it follows from the Strong Law of Large Numbers and the tree structure that:

Theorem 3.2 *As the number of trees increases, for almost surely all sequences $\Theta_1, ..., \Theta_k$, PE^* converges to*

$$P_{X,Y}(P_\Theta(h(X, \Theta) = Y) - max_{j \neq q} P_\Theta(h(X, \Theta) = j) \leq 0) \tag{3.9}$$

This result explains why random forests **do not overfit** as more trees are added, but produce a limiting value of the generalization error.

3.3 Extremely Randomized Trees

The Extra-Trees algorithm builds an ensemble of unpruned decision trees according to the classical top-down procedure. Its two main differences with other tree-based ensemble methods are that it splits nodes by choosing cut-points fully at random and that it uses the whole learning sample (rather than a bootstrap replica) to grow the trees [10].

The Extra-Trees splitting procedure for numerical attributes has two parameters: K, the number of attributes randomly selected at each node and n_{min}, the minimum sample size for splitting a node. It is used several times with the (full) original learning sample to generate an ensemble model (we denote by M the number of trees of this ensemble). The predictions of the trees are aggregated to yield the final prediction, by majority vote in classification problems.

From the bias-variance point of view, the rationale behind the Extra-Trees method is that the explicit randomization of the cut-point and attribute combined with ensemble averaging should be able to reduce variance more strongly than the weaker randomization schemes used by other methods. The usage of the full original learning sample rather than bootstrap replicas is motivated in order to minimize bias. From the computational point of view, the complexity of the tree growing procedure is, assuming balanced trees, on the order of $N \log N$ with respect to learning sample size, like most other tree growing procedures. However, given the simplicity of the node splitting procedure we expect the constant factor to be much smaller than in other ensemble based methods which locally optimize cut-points.

The parameters K, n_{min} and M have different effects: K determines the strength of the attribute selection process, n_{min} the strength of averaging output noise, and M the strength of the variance reduction of the ensemble model aggregation. These parameters could be adapted to the problem specifics in a manual or an automatic way (e.g. by cross-validation). However, it is suggested the use of default settings for them in order to maximize the computational advantages and autonomy of the method [10].

3.4 Considerations

Despite the merits of the Random Forest technique, there are several aspects of the data which affect the performance of the proposed method and need to be taken into consideration. Our work is based on the following assumptions, which can also be found commonly in the literature. First, we consider that the spectra are not raw measurements but they are baseline corrected. Second, we exclude all the spectra samples that exhibit overwhelming fluorescence masking. Third, we observed that the spectra samples span different wavenumber ranges as well as different sampling ratios. Therefore, we apply a re-sampling technique. Finally, we reduct the wavenumber range between some limits, so that only the features drawn from them contain meaningful information. The aforementioned assumptions and the strategies we followed regarding them are described throughout the next chapter.

3.5 Proposed Approach

The systematic study of the spectra dataset reveals various characteristics, which make the classification task extremely challenging even for a human expert. Consequently, we devised a data engineering workflow, to carefully select and transform the data in order to be able to harness the (otherwise) expected classification power of the Random Forest method in such problems. The proposed method is referenced as *Random Forest P&A* (from *Preprocessing* and *Augmentation*), and is explained in the following paragraphs.

3.5.1 Preprocessing

In the field of machine learning, it is a common approach and usually desired to pre-process the data, in such a way that the learning algorithm is encouraged to learn better the discrimination between the samples. Pre-processing can be any transformation process that maps a feature space into an another.

The Raman spectra act as unique fingerprints themselves. Hence, the measurements of a spectrum, namely the intensity-wavenumber pair, can be used directly as a feature vector. In practice though, differences in equipment or the acquisition process can introduce a number of artifacts (sampling rates, noise, fluorescence effect or the Raman effect is weak) that prevent the direct comparison of such spectra. Therefore, a pre-processing step is required before shaping the feature vector. The pre-processing approach we found in the literature [12], which enhances the nature of the spectra, without projecting them in a substantially different space is the following:

- Re-sampling
- Square root transformation: $f(x) = \sqrt{x}$
- Sigmoid transformation: $f(x) = \frac{1-\cos\pi x}{2}$
- Intensity normalization with L_∞ norm: $\|x\|_\infty = max(|x_i|)$

Re-sampling

As a result of differences in the acquisition sources during the collection of the data, the spectra correspond to different sampling rates and measurement ranges. Therefore, it is crucial to resample the spectra into a common set of *wavenumbers*. For that reason, it is recommended the use of interpolation. Interpolation is a process of finding a formula (often a polynomial) whose graph will pass through a given set of points (x, y), in such a way that it replaces a set of data points (x_i, y_i) with a function given analytically. Since the measurement rates in the spectra are dense (despite the differences between them in separate spectra), linear interpolation is applied, in order to map the intensity, wavenumber pair into a common axis.

$$y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0} = \frac{y_0(x_1 - x) + y_1(x - x_0)}{x_1 - x_0} \tag{3.10}$$

3.5.2 Signal enhancing

Another difference between the spectra is the intensity range. The Raman intensity is a function of the polarizability and symmetry and therefore probes the bonding covalence and structure. Since the spectrum's intensity is concentration dependent, there might be subtle or severe differences in two different spectra acquired from separate samples of the same species. However, if one checks the ratio of two peaks just by height or area and compare the ratio of the same peaks and if both these ratios are roughly the same then one can infer that the difference in intensity is just an artifact in sample preparation and the concentration of the sample to be properly adjusted before measurement, as shown in Figure **??**. Therefore, the following signal enhancing strategy deals with the shape of the peaks, while it brings them into the same intensity range.

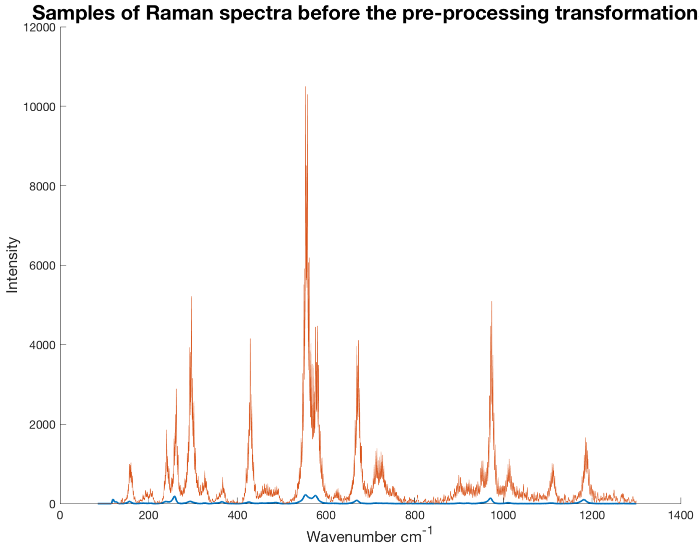

Figure 3.1: Example of samples before pre-processing. Note that both samples belong in the same family of species, despite the differences in the peak-intensity values.

First, intensity squashing is performed by transforming the signal as a function of its square root $f(x) = \sqrt{x}$, as shown in Figure 3.2

Subsequently, a sigmoidal transformation is performed $f(x) = \frac{1-\cos \pi x}{2}$, such that the signal's noise is moderated and smoothed. Finally, in Figure 3.3, it is shown the last step of the pre-processing (along with all the aforementioned transformations), which is intensity normalization. Intensity normalization, brings all spectra into the same intensity range, that is between 0 and 1, using the L_∞ norm.

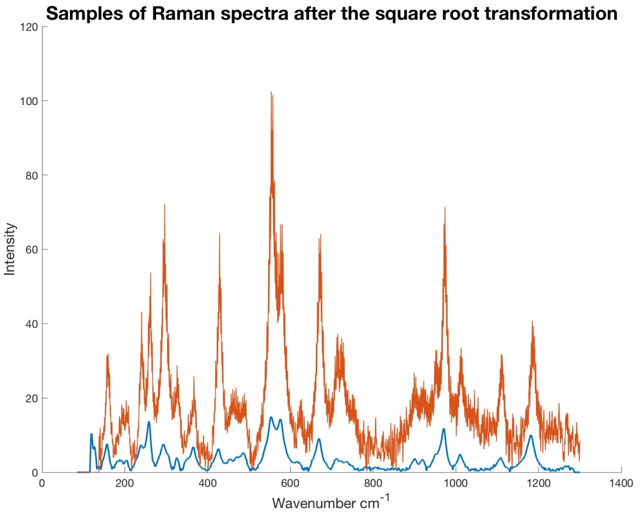

Figure 3.2: Example of samples after square root transformation.

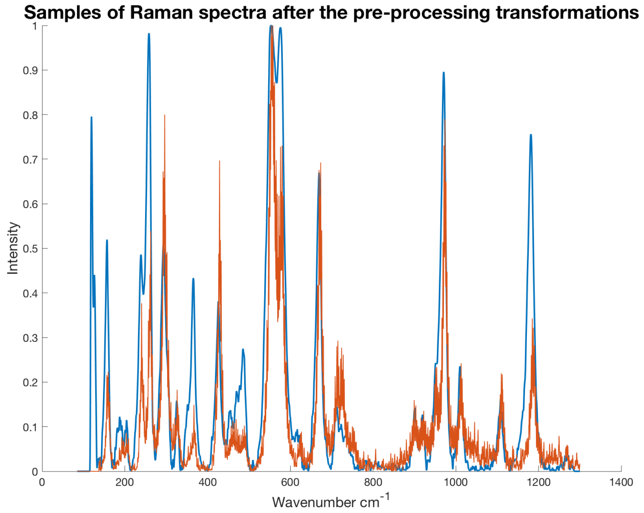

Figure 3.3: Example of samples after all pre-processing steps. The peak intensities are normalized and belong in the same range.

3.5.3 Data Augmentation

As described in [13] data augmentation can be applied in cases in which the training data volume is not enough for a successful training of a machine learning method. We performed the following augmentation procedure: (1) we shifted each spectrum left or right a few wavenumbers randomly; (2) we added random noise, proportional to the magnitude at each wavenumber. The result of the augmentation is illustrated by an example shown in Figure 3.4.

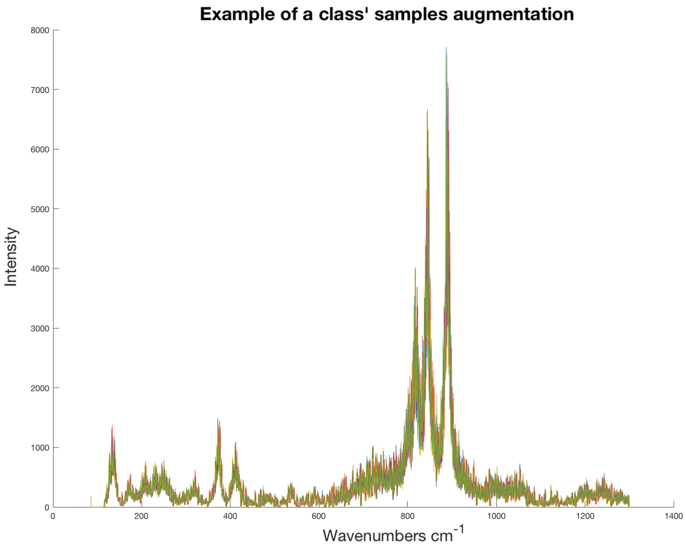

Figure 3.4: An example of sample data augmentation.

Chapter 4

Experimental Design and Implementation

4.1 Motivation

The principal aim of this research is to investigate the effectiveness of the ensemble learning methods on the task of mineral identification. To achieve that, we wish to discover how well the ensemble learning methods perform on an individual species basis, as well as in other particular mineral description groupings following the Dana Mineral Organizational Scheme. Therefore, two kind of experiments will be demonstrated.

In the first experiment, we show the effectiveness of various methods regarding the task of identifying the individual mineral species given their respective Raman spectra. Since the goal of this experiment is for a given method to learn how to distinguish and classify each spectrum to the correct mineral name, the samples set of the RRUFF database was split in two subsets, one for training the methods and the other for testing them respectively. Both the training and testing subsets contained the Raman spectra as features and their corresponding species names as labels. The splitting process is described in detail in this Chapter.

The second experiment, on the other hand, aims at measuring the effectiveness of the compared methods in identifying in which class, type and group a Raman spectrum belongs to alongside with its respective species label. We follow a similar experimental structure as the first experimental design, namely we use the Raman spectra as feature but the labels change from species to mineral class, type and group according to the Dana Mineral Organizational Scheme. Thus, in this experimental design we show the effectiveness of each method in classifying correctly a mineral's spectrum to more than one corresponding label.

4.2 Choice of Dataset

In today's world, many people rely on the availability of databases to perform daily activities such as checking meteorological forecast, finding a recipe, or verifying the spelling of a word. These are examples of actions that anyone can do quickly, routinely, and often at little cost, because of access to the Internet and the development of extensive databases.

The science community has always relied on databases, which in the earlier years were only available through collations of journals, books and data records. The analog nature of these data resources dramatically limits the process of searching records and establishing relations between datasets.

Mineral identification using Raman spectroscopy is normally performed by search/match routines that compare the acquired spectrum with reference spectra from a database. The purpose of the RRUFF project is to develop such a Raman database by measuring the chemistry, X-ray diffraction patterns, Raman, and infrared spectra of the known minerals and to make these data readily and freely available to the scientific community, industry, and the general public. RRUFF database currently contains about 7000 mineral samples representing 3500 mineral species.

As of this moment, 3527 of the 4985 known mineral species have been incorporated into the RRUFF sample collection. As data from a sample is collected, it is posted into the database with password restricted access, referred to as non-public access. After a review process, if the data appears to be representative of the sample, then the password restriction is removed and the data becomes publicly accessible. As a consequence, measurements from only 2128 mineral species are currently publicly accessible. When possible, data from at least two samples of each species, ideally from different localities, are included in the database. Having multiple samples provides a means to confirm data and capture the chemical variability frequently observed in minerals. For instance, the database currently contains 42 records on the important olivine forsterite-fayalite series, with associated Mg-Fe chemical variations. In total, data from 3791 samples are publicly accessible through the RRUFF database.

The RRUFF project includes an extensive reference library of publications directly linked to their associated minerals. For the most part, these articles are focused on spectroscopy, structure, and chemistry of minerals. The complete list of collaborators can be found at: `http://rruff.info/about/about_publishers.php`.

The RRUFF database contains both raw and already baseline corrected spectra. In our study, we used the set of spectra that were already baseline corrected with piecewise linear interpolation between smoothed off-peak segments, and further preprocessed for artifact removal.

4.2.1 Dataset Statistics

The RRUFF database contains both raw and already baseline corrected spectra. In our study, we used the set of spectra that were already baseline corrected with piecewise linear interpolation between smoothed off-peak segments, and further preprocessed for artifact removal. The acquisition of these spectra was performed at random orientations. Similar to [12], spectra with overwhelming specimen fluorescence were removed from our selection from the dataset. Even though performing piece-wise interpolation is a trivial procedure, it is our intention to leave the data untampered and use the provided preprocessed data, since implementing an in house solution for the raw spectra might introduce deviations from the initial dataset and the comparison with others' works would have been unfair. Hence, we adopted the provided baseline corrected data unaltered. Additionally, according to the aforementioned study, it is a common practice that experts of the Raman community, who use the data provided by the RRUFF database and the Crystal-Sleuth software, avoid further modification and work with the standard RRUFF processed data.

The subset of the RRUFF database we used it had more than 5000 spectra, which according to the DANA organization scheme except for the species label, also it could be described in groups, types and classes. Table 4.1 summarizes the dataset's statistical information from the DANA perspective.

Table 4.1: Dataset description according to DANA

Organization level	# labels
Class	79
Type	307
Group	977
Species	1683

4.2.2 Grooming the Dataset

As mentioned before, the RRUFF dataset has numerous samples for a wide range of species, but not all of them were suitable for the purpose of our experiments. A grooming procedure was employed, such that irrelevant samples to the purpose of his work were excluded. First, each mineral name was matched with its four-part Dana classification number, some minerals had no match, thus they were excluded.[33] Many minerals in the RRUFF database are represented by a small amount of spectra, as depicted in Figure 4.1, thus, performing random splits in our dataset would result to some subsets in the testing sets, in which the classifier would have no knowledge of their label. To avoid this problem we follow the

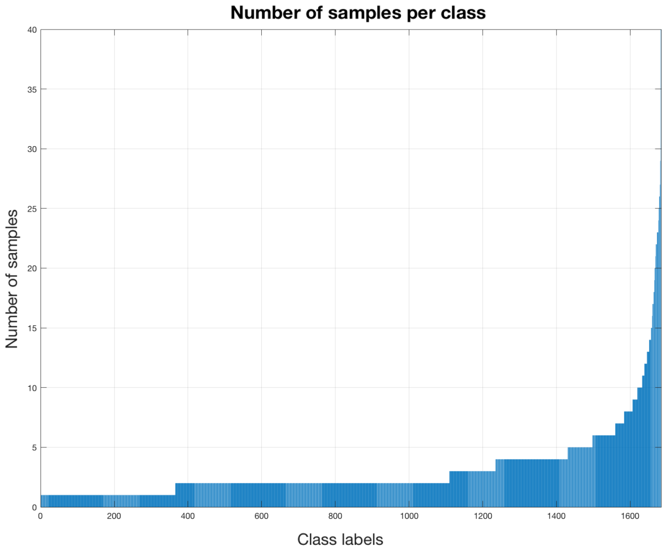

Figure 4.1: Statistical profile of the dataset.

same semi-randomized split strategy as in [12], where for each query spectrum to have at least one true match in the reference set, the reference set for each trial is constructed by selecting three spectra per mineral species at random. All remaining spectra are assigned to the trial's query set. As a result, species with three or fewer total spectra are not present in the query set for any trial and always appear in the reference set. According to this strategy, there are 2934 reference spectra and 1356 query spectra for each trial. Note, that our numbers differ from [12] because the RRUFF dataset has been updated since that study.

After the dataset's grooming procedure, its DANA description was as in Table 4.2

Table 4.2: Dataset description according to DANA (after grooming)

Organization level	# labels
Class	78
Type	306
Group	966
Species	1419

4.3 Parameter Selection

After extensive experimentation with the RRUFF dataset and our first results using multiple scenarios it has become clear that the application of Decision Trees or their extension, the Random Forest, although intuitively fitting the problem at hand, are expected to have a limited discriminative power and thus, classification efficiency. In addition, the systematic study of the dataset reveals various characteristics, which make the classification task extremely challenging even for a human expert. Consequently, we devised a data engineering workflow, to carefully select and transform the data in order to be able to harness the (otherwise) expected classification power of the Random Forest method in such problems. The proposed method is referenced as *Random Forest P&A* (from *Preprocessing* and *Augmentation*), and is explained in the following paragraphs.

4.3.1 Baseline correction

Variations in intensity can significantly affect the success of the search/ match routines. In certain cases, especially for rock-forming minerals, Raman spectra collected in additional different orientations with respect to the polarization direction of the incident laser are collected. CrystalSleuth is used to correct the spectral baseline, using a subroutine from the Razor library `http://www.spectrumsquare.com` by Spectrum Square Associates, to remove cosmic rays events from patterns [28], to trim edges, to reverse X-axis display, and to visualize and compare multiple spectra. Both raw and processed data are included on RRUFF as XY-ASCII files. In figure we provide a comparison between a raw measurement and a baseline corrected signal of the same species.

The acquisition of these spectra was performed at random orientations. Similar to [12], spectra with overwhelming specimen fluorescence were removed from our selection from the dataset. Even though performing piece-wise interpolation is a trivial procedure, it is our intention to leave the data untampered and use the provided preprocessed data, since implementing an in house solution for the raw spectra might introduce deviations from the initial dataset and the comparison with others' works would have been unfair. Hence, we adopted the provided baseline corrected data unaltered. Additionally, according to the aforementioned study, it is a common practice that experts of the Raman community, who use the data provided by the RRUFF database and the CrystalSleuth software, avoid further modification and work with the standard RRUFF processed data.

4.3.2 Pre-processing parameters

As a result of differences in the acquisition sources during the collection of the RRUFF data, the spectra correspond to different sampling rates and measure-

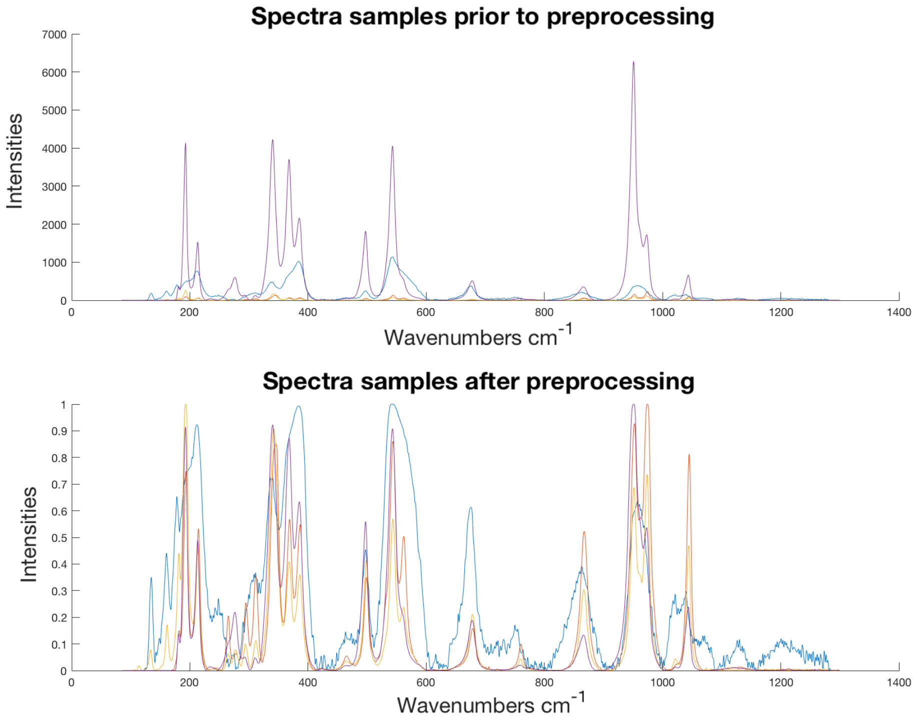

Figure 4.2: Several samples prior to preprocessing (top) and the same samples after preprocessing (bottom); the arrangement of peaks remain the same as the preprocessing aims at altering only the intensities.

ment ranges. Therefore, it is crucial to resample the spectra into a common set of *wavenumbers*. Using linear interpolation we opted for a measurement range between 85 and 1315 cm^{-1} with uniform sampling of 2048 intensity values.

4.3.3 Data Augmentation Parameters

The augmentation parameters for the spectrum shifting were $\alpha = 0.5$ for the random chance of the signal shifting left or right, $\delta = 5$ for the shifting range limit, and $\mu = 0.1$ as the proportion of noise imposed to the signal's magnitude.

4.3.4 Decision Tree Parameters

The parameters concerning the DT were the number of features randomly selected at each node $m = \sqrt{2048}$, the maximal depth was defined so as to let the nodes expand until all leaves are pure, or until all leaves contain less than the minimal number of samples, which was set to 3

4.3.5 Random Forest Parameters

The parameters concerning the RF were $K = 1000$ trees, the number of features randomly selected at each node $m = \sqrt{2048}$, the maximal depth was defined so as to let the nodes expand until all leaves are pure, or until all leaves contain less than the minimal number of samples, which was set to 3.

4.3.6 Extra Randomized Trees

The parameters concerning the RF were $K = 1000$ trees, the number of features randomly selected at each node $m = 30$, the maximal depth was defined so as to let the nodes expand until all leaves are pure, or until all leaves contain less than the minimal number of samples, which was set to 3.

4.4 Implementation Details

For traceability and repeatability we note that all the experiments run on the same Apple Macbook with a 2 GHz Intel Core i7 and 8 GB RAM, in which and where applicable (i.e., during the random forest training), we opted for GPU parallelization. We used Matlab for the analysis and visualization of the dataset, as well as for its grooming and the aforementioned transformation processes. The development of the machine learning framework situated in the environment of Python, as it offers many well established libraries for the training and testing of such algorithms.

Chapter 5

Results and Discussion

5.1 Performance

5.1.1 Species Identification

We evaluated our methods (Random Forest P&A and Extra Trees P&A) using the RRUFF database as in [12, 13]. Since our experimental setup is similar to the work in [12], we included their results for comparison in Table 5.1, where also the results by [13] (only for the case of the corrected RRUFF data) are included for a fair comparison. The table includes the results obtained by k-Nearest Neighbour classification after normalisation with $k = 1$ and $k = 10$ respectively; an accuracy of 82.1% has been reported for $k = 1$[1], whereas increasing the number k up to the 10 nearest neighbours didn't actually have a significant impact. The Weighted Neighbours method, after the square root, sigmoidal and maximum intensity normalisation transformations, reported an accuracy score of 84.8%, indicating the importance of the preprocessing step. A simple multilayer perceptron artificial neural network is reported to have achieved a low accuracy of 35.6%.

More similar to our approach are the Decision Tree and the Random Forest methods. As expected, the Decision Tree, being a weak learner in its nature, won't be able to learn well the classification. That has much to do with the complexities of the Raman spectra, where a weak model yields poor discriminating power when it comes to capturing the particular nuances of each mineral's spectrum, due to a sample's irregular chemical mixtures, which ultimately leads to misclassifications.

Prior to applying the preprocessing transformations and the aforementioned augmentation strategy, we tested the Random Forest with the unprocessed Raman spectra. We found that the method performs better than the Decision Tree, as it has achieved a higher classification performance, scoring 67.5%. However, after we applied the preprocessing transformations and augmentation of the training

[1] It is mentioned as CrystalSleuth's internal matching algorithm [12, 13]

Table 5.1: Summary of results of Mineral Species Identification

Method	Accuracy (%)
Decision Tree	31.6
Multilayer Perceptron	35.6
Random Forest	67.5
1-NN, norm	82.1
10-NN, norm	82.2
WN, sqrt+norm+sigmoid	84.8
Deep CNN (corrected data)	88.4
Random Forest P&A	81.3
Extra Trees P&A	**88.8**

dataset, the Random Forest has presented a significantly improved accuracy of 81.3%. Finally, the Extra Trees modification achieved an accuracy score of 88.8%, outperforming all of other methods, and also the CNN approach, which achieved an accuracy score of 88.4% on the already baseline corrected RRUFF subset.

5.1.2 Mineral Class, Type and Group Identification

In this section we present the results of the compared methods regarding their ability to classify the Raman spectra according to their Class, Type, Group, and Species classification label based on the Dana Organizational Scheme [97]. For each sample's spectrum instead of matching it with their respective species label, we provided the aforementioned additional labels. This experimental approach was presented previously in the literature [12]. The Decision Tree and the Multilayer Perceptron scored low at all cases. However, we observe an accuracy increase as the organizational hierarchy ascends. Note, that the Dana Scheme is a nested ontology of a mineral's identity description. Therefore, the top of the hierarchy being a Class may contain many Types, a Type may contain many Groups, and consequently a Group may contain many Species, which is the bottom tier of this hierarchy.

Similar improvement present all the methods, shown in Table 5.2. Interestingly, the results provided by the method of k-Nearest Neighbour classification after normalisation with $k = 1$ and $k = 10$ respectively; reached at an accuracy performance, abov 90%, at the Dana Class and Type tiers. Again, increasing the number k up to the 10 nearest neighbours didn't actually have a significant impact. The Weighted Neighbours method, after the square root, sigmoidal and maximum intensity normalisation transformations, presented an accuracy score of 94.8%, 93.3%, and 92.0% for the Dana tiers of Class, Type and Group respectively, placing this method at the top of the reported accuracy scores.

Our methods present comparable accuracy scores. From the two ensemble learning methods the weaker was the Random Forest. Even though there was an

Table 5.2: Summary of results of Mineral Identification in respect to Class, Type, Group, and Species Classification labels

	Accuracy (%)			
Method	Class	Type	Group	Species
Decision Tree	44.3	40.5	37.3	31.6
Multilayer Perceptron	59	51.9	46	35.6
1-NN, norm	94.9	92.9	90.7	82.1
10-NN, norm	93.8	91.7	89.9	82.2
WN, sqrt+norm+sigmoid	94.8	93.3	92.0	84.8
Random Forest P&A	86.7	85.3	83.5	81.3
Extra Trees P&A	**95.7**	**94.3**	**92.5**	**88.8**

improvement for each ascending Dana tier, it seems that this method reached a performance plateau. We speculate, that this was due to the method's splitting selection process, at its split criterion potentially satisfied the same requirements irrespective to the label's tier change. On the other hand, the Extra Randomized Trees, present an outstanding performance at any Dana hierarchical tier outperforming all other methods.

5.2 Achievements

This research met successfully the following goals:

- The proposed methods are off-the-shelf machine learning methods, intuitive and easy to understand by experts from fields other than Computer Science.
- Our proposed methods present accuracy scores that outperform the current state of the art.
- In terms of computational intensity, our proposed methods are less demanding than training a Deep Neural Network.
- The Extra Randomized Trees are able to discriminate between vibration noise and produce accurate classifications.
- Both presented ensemble learning methods are able to make computations in parallel, making them computationally faster.

5.3 Conclusion

The ensemble learning methods have seen an increased attention in the machine learning community, prior the wide spread use of CNN architectures. Naturally, it has already been compared against other powerful methods such as SVMs in many domains, including that of spectroscopy applications. In projects with binary classification tasks or small customized datasets, Random Forest showed decent, yet second-rate results. We believe that was due to the small number of training examples that didn't allow RF to reach its potential. However, in this thesis we use a large-scale dataset, which contained numerous Raman spectra, we found that RF scales well. Furthermore, an other ensemble of trees method, namely the Extremely Randomized Trees, showed a remarkable performance as it surpassed all previously reported methods and the state-of-the-art. Also, we demonstrated that the dataset engineering through data augmentation and preprocessing made it possible to boost the performance of both the Random Forest and the Extremely Randomized Trees, without sacrificing any computational speed or imposing additional computational demanding procedures. Overall, we found that our proposed method outperforms the best results previously reported by at least **4%** in classification accuracy at species identification, once again recognizing that the ensemble learning methods are able to learn and discriminate efficiently the subtle discrepancies of the chemical mixtures in spectral samples without computational burden.

Bibliography

[1] A. No and A. M. Committee, "Raman spectroscopy in cultural heritage: background paper," *Analytical Methods*, vol. 7, no. 12, pp. 4844–4847, 2015.

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

[3] D. Gardiner, P. Graves, and H. Bowley, "Practical raman spectroscopy. 1989."

[4] G. Gauglitz and T. Vo-Dinh, *Handbook of spectroscopy*. John Wiley & Sons, 2006.

[5] K. Janssens and R. Van Grieken, *Non-destructive micro analysis of cultural heritage materials*, vol. 42. Elsevier, 2004.

[6] H. G. Edwards and T. Munshi, "Diagnostic raman spectroscopy for the forensic detection of biomaterials and the preservation of cultural heritage," *Analytical and bioanalytical chemistry*, vol. 382, no. 6, pp. 1398–1406, 2005.

[7] L. Bellot-Gurlet, C. Coupry, *et al.*, "Raman spectroscopy in art and archaeology," *Journal of Raman Spectroscopy*, vol. 37, no. 10, pp. 962–965, 2006.

[8] F. Casadio, C. Daher, and L. Bellot-Gurlet, "Raman spectroscopy of cultural heritage materials: overview of applications and new frontiers in instrumentation, sampling modalities, and data processing," *Topics in Current Chemistry*, vol. 374, no. 5, p. 62, 2016.

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

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

[11] A. Maguire, I. Vega-Carrascal, J. Bryant, L. White, O. Howe, F. Lyng, and A. Meade, "Competitive evaluation of data mining algorithms for use in classification of leukocyte subtypes with raman microspectroscopy," *Analyst*, vol. 140, no. 7, pp. 2473–2481, 2015.

[12] C. Carey, T. Boucher, S. Mahadevan, P. Bartholomew, and M. Dyar, "Machine learning tools for mineral recognition and classification from raman spectroscopy," *Journal of Raman Spectroscopy*, vol. 46, no. 10, pp. 894–903, 2015.

[13] J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, and S. J. Gibson, "Deep convolutional neural networks for raman spectrum recognition: a unified solution," *Analyst*, vol. 142, no. 21, pp. 4067–4074, 2017.

[14] D. A. Skoog, F. J. Holler, and S. R. Crouch, *Principles of instrumental analysis*. Cengage learning, 2017.

[15] 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.

[16] F. A. Settle, *Handbook of instrumental techniques for analytical chemistry*. Prentice Hall PTR,, 1997.

[17] J. M. Chalmers, H. G. Edwards, and M. D. Hargreaves, *Infrared and Raman spectroscopy in forensic science*. John Wiley & Sons, 2012.

[18] B. M. Cullum and T. Vo-Dinh, "Sample collection and preparation of liquid and solids," *Handbook of Spectroscopy*, vol. 2, p. 1, 2003.

[19] E. Smith and G. Dent, *Modern Raman spectroscopy: a practical approach*. John Wiley & Sons, 2013.

[20] M. Havel, D. Baron, and P. Colomban, "'smart'raman/rayleigh imaging of nanosized sic materials using the spatial correlation model," *Journal of materials science*, vol. 39, no. 20, pp. 6183–6190, 2004.

[21] P. Colomban and A. Slodczyk, "Raman intensity: an important tool in the study of nanomaterials and nanostructures," *Acta Physica Polonica-Series A General Physics*, vol. 116, no. 1, p. 7, 2009.

[22] H. G. Edwards, S. E. J. Villar, J. Jehlicka, and T. Munshi, "Ft–raman spectroscopic study of calcium-rich and magnesium-rich carbonate minerals," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 61, no. 10, pp. 2273–2280, 2005.

[23] G. Vergote, T. De Beer, C. Vervaet, J. P. Remon, W. Baeyens, N. Diericx, and F. Verpoort, "In-line monitoring of a pharmaceutical blending process using ft-raman spectroscopy," *European Journal of Pharmaceutical Sciences*, vol. 21, no. 4, pp. 479–485, 2004.

[24] Z. Li, M. J. Deen, S. Kumar, and P. R. Selvaganapathy, "Raman spectroscopy for in-line water quality monitoring—instrumentation and potential," *Sensors*, vol. 14, no. 9, pp. 17275–17303, 2014.

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

[26] J. R. Lombardi and R. L. Birke, "A unified approach to surface-enhanced raman spectroscopy," *The Journal of Physical Chemistry C*, vol. 112, no. 14, pp. 5605–5617, 2008.

[27] P. L. Stiles, J. A. Dieringer, N. C. Shah, and R. P. Van Duyne, "Surface-enhanced raman spectroscopy," *Annu. Rev. Anal. Chem.*, vol. 1, pp. 601–626, 2008.

[28] M. Fan, G. F. Andrade, and A. G. Brolo, "A review on the fabrication of substrates for surface enhanced raman spectroscopy and their applications in analytical chemistry," *Analytica chimica acta*, vol. 693, no. 1-2, pp. 7–25, 2011.

[29] R. A. Halvorson and P. J. Vikesland, "Surface-enhanced raman spectroscopy (sers) for environmental analyses," 2010.

[30] E. Ippen and C. Shank, "Picosecond response of a high- repetition- rate cs2 optical kerr gate," *Applied Physics Letters*, vol. 26, no. 3, pp. 92–93, 1975.

[31] P. Matousek, M. Towrie, A. Stanley, and A. Parker, "Efficient rejection of fluorescence from raman spectra using picosecond kerr gating," *Applied Spectroscopy*, vol. 53, no. 12, pp. 1485–1489, 1999.

[32] F. Knorr, Z. J. Smith, and S. Wachsmann-Hogiu, "Development of a time-gated system for raman spectroscopy of biological samples," *Optics express*, vol. 18, no. 19, pp. 20049–20058, 2010.

[33] R. Baker, P. Matousek, K. L. Ronayne, A. W. Parker, K. Rogers, and N. Stone, "Depth profiling of calcifications in breast tissue using picosecond kerr-gated raman spectroscopy," *Analyst*, vol. 132, no. 1, pp. 48–53, 2007.

[34] E. V. Efremov, J. B. Buijs, C. Gooijer, and F. Ariese, "Fluorescence rejection in resonance raman spectroscopy using a picosecond-gated intensified charge-coupled device camera," *Applied spectroscopy*, vol. 61, no. 6, pp. 571–578, 2007.

[35] Y. Fleger, L. Nagli, M. Gaft, and M. Rosenbluh, "Narrow gated raman and luminescence of explosives," *Journal of Luminescence*, vol. 129, no. 9, pp. 979–983, 2009.

[36] F. Ariese, H. Meuzelaar, M. M. Kerssens, J. B. Buijs, and C. Gooijer, "Picosecond raman spectroscopy with a fast intensified ccd camera for depth analysis of diffusely scattering media," *Analyst*, vol. 134, no. 6, pp. 1192–1197, 2009.

[37] S. C. Pînzaru and I. E. Pavel, "Sers and pharmaceuticals," *Surface Enhanced Raman Spectroscopy: Analytical, Biophysical and Life Science Applications*, pp. 129–154, 2010.

[38] K. Buckley and P. Matousek, "Recent advances in the application of transmission raman spectroscopy to pharmaceutical analysis," *Journal of pharmaceutical and biomedical analysis*, vol. 55, no. 4, pp. 645–652, 2011.

[39] Z.-Q. Wen, "Raman spectroscopy of protein pharmaceuticals," *Journal of pharmaceutical sciences*, vol. 96, no. 11, pp. 2861–2878, 2007.

[40] Z.-Q. Wen, X. Cao, and A. Vance, "Conformation and side chains environments of recombinant human interleukin-1 receptor antagonist (rh-il-1ra) probed by raman, raman optical activity, and uv-resonance raman spectroscopy," *Journal of pharmaceutical sciences*, vol. 97, no. 6, pp. 2228–2241, 2008.

[41] X. Cao, Z.-Q. Wen, A. Vance, and G. Torraca, "Raman microscopic applications in the biopharmaceutical industry: in situ identification of foreign particulates inside glass containers with aqueous formulated solutions," *Applied spectroscopy*, vol. 63, no. 7, pp. 830–834, 2009.

[42] M. Harz, S. Stöckel, V. Ciobotă, D. Cialla, P. Rösch, and J. Popp, "Applications of raman spectroscopy to virology and microbial analysis," in *Emerging Raman Applications and Techniques in Biomedical and Pharmaceutical Fields*, pp. 439–463, Springer, 2010.

[43] E. L. Izake, "Forensic and homeland security applications of modern portable raman spectroscopy," *Forensic science international*, vol. 202, no. 1-3, pp. 1–8, 2010.

[44] N. Stone, M. Kerssens, G. R. Lloyd, K. Faulds, D. Graham, and P. Matousek, "Surface enhanced spatially offset raman spectroscopic (sesors) imaging–the next dimension," *Chemical Science*, vol. 2, no. 4, pp. 776–780, 2011.

[45] M. Delhaye and P. Dhamelincourt, "Raman microprobe and microscope with laser excitation," *Journal of Raman Spectroscopy*, vol. 3, no. 1, pp. 33–43, 1975.

[46] H. Edwards, D. Farwell, M. Seaward, and C. Giacobini, "Preliminary raman microscopic analyses of a lichen encrustation involved in the biodeterioration of renaissance frescoes in central italy," *International Biodeterioration*, vol. 27, no. 1, pp. 1–9, 1991.

[47] H. Edwards, D. Farwell, and M. Seaward, “Raman spectra of oxalates in lichen encrustations on renaissance frescoes,” *Spectrochimica Acta Part A: Molecular Spectroscopy*, vol. 47, no. 11, pp. 1531–1539, 1991.

[48] J. Russ, R. L. Palma, D. H. Loyd, D. W. Farwell, and H. G. Edwards, “Analysis of the rock accretions in the lower pecos region of southwest texas,” *Geoarchaeology*, vol. 10, no. 1, pp. 43–63, 1995.

[49] A. C. Williams, H. G. Edwards, and B. W. Barry, “The ‘iceman’: molecular structure of 5200-year-old skin characterised by raman spectroscopy and electron microscopy,” *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, vol. 1246, no. 1, pp. 98–105, 1995.

[50] H. G. Edwards, “Probing history with raman spectroscopy,” *Analyst*, vol. 129, no. 10, pp. 870–879, 2004.

[51] P. Vandenabeele, H. G. Edwards, and L. Moens, “A decade of raman spectroscopy in art and archaeology,” *Chemical reviews*, vol. 107, no. 3, pp. 675–686, 2007.

[52] H. Edwards and P. Vandenabeele, *Analytical archaeometry: selected topics*. Royal Society of Chemistry, 2016.

[53] H. G. Edwards and P. Vandenabeele, “Raman spectroscopy in art and archaeology,” 2016.

[54] M. A. Ziemann, “In situ micro-raman spectroscopy on minerals on-site in the grotto hall of the new palace, park sanssouci, in potsdam,” *Journal of Raman Spectroscopy*, vol. 37, no. 10, pp. 1019–1025, 2006.

[55] S. Centeno, V. Buisan, and P. Ropret, “Raman spectrosc. 2006; 37: 1111.”

[56] M. Sánchez del Río, M. Picquart, E. Haro-Poniatowski, E. Van Elslande, and V. Hugo Uc, “On the raman spectrum of maya blue,” *Journal of Raman Spectroscopy*, vol. 37, no. 10, pp. 1046–1053, 2006.

[57] P. Buzzini and E. Suzuki, “Forensic applications of raman spectroscopy for the in situ analyses of pigments and dyes in ink and paint evidence,” *Journal of Raman Spectroscopy*, vol. 47, no. 1, pp. 16–27, 2016.

[58] M. Cañamares, D. Reagan, J. Lombardi, and M. Leona, “Tlc-sers of mauve, the first synthetic dye,” *Journal of Raman Spectroscopy*, vol. 45, no. 11-12, pp. 1147–1152, 2014.

[59] A. Zoppi, C. Lofrumento, E. Castellucci, C. Dejoie, and P. Sciau, “Micro-raman study of aluminium-bearing hematite from the slip of gaul sigillata wares,” *Journal of Raman Spectroscopy*, vol. 37, no. 10, pp. 1131–1138, 2006.

[60] J. Striova, C. Lofrumento, A. Zoppi, and E. M. Castellucci, “Prehistoric anasazi ceramics studied by micro-raman spectroscopy,” *Journal of Raman Spectroscopy*, vol. 37, no. 10, pp. 1139–1145, 2006.

[61] C. Sandalinas, S. Ruiz-Moreno, A. López-Gil, and J. Miralles, “Experimental confirmation by raman spectroscopy of a pb sn sb triple oxide yellow pigment in sixteenth-century italian pottery,” *Journal of Raman Spectroscopy*, vol. 37, no. 10, pp. 1146–1153, 2006.

[62] M. Marco de Lucas, F. Moncada, and J. Rosen, “Micro-raman study of red decorations in french faiences of the 18th and 19th centuries,” *Journal of Raman Spectroscopy*, vol. 37, no. 10, pp. 1154–1159, 2006.

[63] M. Aceto, A. Agostino, E. Boccaleri, F. Crivello, and A. C. Garlanda, “Evidence for the degradation of an alloy pigment on an ancient italian manuscript,” *Journal of Raman Spectroscopy*, vol. 37, no. 10, pp. 1160–1170, 2006.

[64] M. Christensen, M. Frosch, P. Jensen, U. Schnell, Y. Shashoua, and O. F. Nielsen, “Waterlogged archaeological wood—chemical changes by conservation and degradation,” *Journal of Raman Spectroscopy*, vol. 37, no. 10, pp. 1171–1178, 2006.

[65] M. Bicchieri, A. Sodo, G. Piantanida, and C. Coluzza, “Analysis of degraded papers by non-destructive spectroscopic techniques,” *Journal of Raman Spectroscopy*, vol. 37, no. 10, pp. 1186–1192, 2006.

[66] H. Edwards, N. Nik Hassan, D. Farwell, P. Garside, and P. Wyeth, “Raman spectrosc. 2006; 37: 1193.”

[67] Y. Shashoua, D. Berthelsen, M.-B. Lund, and O. F. Nielsen, “Raman and atr-ftir spectroscopies applied to the conservation of archaeological baltic amber,” *Journal of Raman Spectroscopy*, vol. 37, no. 10, pp. 1221–1227, 2006.

[68] J. Monnier, D. Neff, S. Reguer, P. Dillmann, L. Bellot-Gurlet, E. Leroy, E. Foy, L. Legrand, and I. Guillot, “A corrosion study of the ferrous medieval reinforcement of the amiens cathedral. phase characterisation and localisation by various microprobes techniques,” *Corrosion science*, vol. 52, no. 3, pp. 695–710, 2010.

[69] A. Hernanz, M. Mas, B. Gavilán, and B. Hernández, “Raman microscopy and ir spectroscopy of prehistoric paintings from los murciélagos cave (zuheros, córdoba, spain),” *Journal of Raman Spectroscopy*, vol. 37, no. 4, pp. 492–497, 2006.

[70] P. Colomban and V. Milande, “Raman spectrosc. 2006; 37: 606,” *Web of Science® Times Cited*, vol. 19.

[71] R. H. Brody, H. G. Edwards, and A. M. Pollard, “Chemometric methods applied to the differentiation of fourier-transform raman spectra of ivories,” *Analytica Chimica Acta*, vol. 427, no. 2, pp. 223–232, 2001.

[72] H. G. Edwards, “Forensic applications of raman spectroscopy to the non-destructive analysis of biomaterials and their degradation,” *Geological Society, London, Special Publications*, vol. 232, no. 1, pp. 159–170, 2004.

[73] H. G. Edwards, D. W. Farwell, E. M. Newton, and F. R. Perez, “Minium; ft-raman non-destructive analysis applied to an historical controversy,” *Analyst*, vol. 124, no. 9, pp. 1323–1326, 1999.

[74] P. Sobron, F. Sobron, A. Sanz, and F. Rull, “Raman signal processing software for automated identification of mineral phases and biosignatures on mars,” *Applied Spectroscopy*, vol. 62, no. 4, pp. 364–370, 2008.

[75] G. Lopez-Reyes, P. Sobron, C. Lefebvre, and F. Rull, “Multivariate analysis of raman spectra for the identification of sulfates: Implications for exomars,” *American Mineralogist*, vol. 99, no. 8-9, pp. 1570–1579, 2014.

[76] G. Lopez-Reyes, F. Rull, G. Venegas, F. Westall, F. Foucher, N. Bost, A. Sanz, A. Catalá-Espí, A. Vegas, I. Hermosilla, A. Sansano, and J. Medina, “Analysis of the scientific capabilities of the exomars raman laser spectrometer instrument,” *European Journal of Mineralogy*, vol. 25, no. 5, pp. 721–733, 2013.

[77] M. A. Mischna, M. I. Richardson, R. J. Wilson, and D. J. McCleese, “On the orbital forcing of martian water and co2 cycles: A general circulation model study with simplified volatile schemes,” *Journal of Geophysical Research: Planets*, vol. 108, no. E6, 2003.

[78] R. Perez-Pueyo, M. Soneira, and S. Ruiz-Moreno, “A fuzzy logic system for band detection in raman spectroscopy,” *Journal of Raman Spectroscopy*, vol. 35, no. 8-9, pp. 808–812, 2004.

[79] E. Kriesten, F. Alsmeyer, A. Bardow, and W. Marquardt, “Fully automated indirect hard modeling of mixture spectra,” *Chemometrics and Intelligent Laboratory Systems*, vol. 91, no. 2, pp. 181–193, 2008.

[80] I. H. Rodriguez, G. Lopez-Reyes, D. Llanos, and F. R. Perez, “Automatic raman spectra processing for exomars,” in *Mathematics of Planet Earth*, pp. 127–130, Springer, 2014.

[81] M. Paradkar and J. Irudayaraj, "Discrimination and classification of beet and cane inverts in honey by ft-raman spectroscopy," *Food Chemistry*, vol. 76, no. 2, pp. 231–239, 2002.

[82] M. S. Gilmore, B. Bornstein, M. D. Merrill, R. Castaño, and J. P. Greenwood, "Generation and performance of automated jarosite mineral detectors for visible/near-infrared spectrometers at mars," *Icarus*, vol. 195, no. 1, pp. 169–183, 2008.

[83] S. T. Ishikawa and V. C. Gulick, "An automated mineral classifier using raman spectra," *Computers & geosciences*, vol. 54, pp. 259–268, 2013.

[84] T. L. Roush and R. Hogan, "Automated classification of visible and near-infrared spectra using self-organizing maps," in *Aerospace Conference, 2007 IEEE*, pp. 1–10, IEEE, 2007.

[85] J. A. Jaszczak, "Word to the wise: Raman spectroscopy in the identification and study of minerals," *Rocks & Minerals*, vol. 88, no. 2, pp. 184–189, 2013.

[86] K. Carron and R. Cox, "Qualitative analysis and the answer box: a perspective on portable raman spectroscopy," 2010.

[87] V. Baeten, P. Hourant, M. T. Morales, and R. Aparicio, "Oil and fat classification by ft-raman spectroscopy," *Journal of Agricultural and Food Chemistry*, vol. 46, no. 7, pp. 2638–2646, 1998.

[88] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[89] 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.

[90] E. Widjaja, W. Zheng, and Z. Huang, "Classification of colonic tissues using near-infrared raman spectroscopy and support vector machines," *International journal of oncology*, vol. 32, no. 3, pp. 653–662, 2008.

[91] A. Kyriakides, E. Kastanos, and C. Pitris, "Classification of raman spectra using support vector machines," in *Information Technology and Applications in Biomedicine, 2009. ITAB 2009. 9th International Conference on*, pp. 1–4, IEEE, 2009.

[92] U. Thissen, M. Pepers, B. Üstün, W. Melssen, and L. Buydens, "Comparing support vector machines to pls for spectral regression applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 73, no. 2, pp. 169–179, 2004.

[93] M. Gallagher and P. Deacon, "Neural networks and the classification of mineralogical samples using x-ray spectra," in *Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on*, vol. 5, pp. 2683–2687, IEEE, 2002.

[94] S. Bayraktar, B. Labitzke, J. Bader, R. Bornemann, P. H. Bolivar, and A. Kolb, "Efficient, robust, and scale-invariant decomposition of raman spectra," in *Signal and Image Processing Applications (ICSIPA), 2013 IEEE International Conference on*, pp. 317–321, IEEE, 2013.

[95] S. Lowry, D. Wieboldt, D. Dalrymple, R. Jasinevicius, and R. T. Downs, "The use of a raman spectral database of minerals for the rapid verification of semiprecious gemstones," *Spectroscopy*, vol. 24, no. 5, pp. 1–7, 2009.

[96] B. Lafuente, R. Downs, H. Yang, and N. Stone, "The power of databases: the rruff project," *Highlights in mineralogical crystallography*, pp. 1–30, 2015.

[97] R. Gaines, H. Skinner, E. Foord, B. Mason, and A. Rosenzweig, "Dana's new mineralogy, john wiley & sons," *New York*, 1997.

[98] H. Strunz and E. H. Nickel, *Strunz mineralogical tables: Chemical-structural mineral classification system*. Schweizerbart, 2001.

[99] R. B. Cook, "A systematic classification of minerals," *Rocks and Minerals*, vol. 79, no. 6, p. 423, 2004.

[100] S. J. Mills, F. Hatert, E. H. Nickel, and G. Ferraris, "The standardisation of mineral group hierarchies: application to recent nomenclature proposals," *European Journal of Mineralogy*, vol. 21, no. 5, pp. 1073–1080, 2009.