FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Self-Learning Artificial Intelligence for Laser Induced Breakdown Spectroscopy: Data Analysis and System Control

Alberto Mesquita Santos

MASTER DISSERTATION

INTEGRATED MASTER IN ELECTRICAL AND COMPUTERS ENGINEERING

Supervisor: Doutor Rui Martins

Co-supervisor: Prof. Vladimiro Miranda

July 23, 2021

Resumo

Espectroscopia de Plasma Induzido por Laser (LIBS) é uma tecnologia que permite de um forma rápida e com pouca preparação, a leitura de amostras para determinação da sua composição. O LIBS extrai informação sobre a amostra, recolhendo as linhas espectrais formadas após uma rotura de ligações moleculares, permitindo identificar e quantificar cada constituinte presente. A análise destas diferentes emissões no plasma contém informação sobre os elementos químicos, isótopos ou moléculas e é utilizado para trabalhar amostras complexas.

O LIBS é uma ferramenta emergente, sendo atualmente utilizado em diversas áreas, desde as mais industriais às maioriamente tecnológicas assim como no setor da saúde e farmacêutica. O LIBS permite aceder com precisão elevada à essência das amostras, interpretando a sua composição química de uma forma pouco evasiva.

Apesar da praticabilidade do LIBS, analisar o seu sinal não é trivial devido à sua complexidade e dos fenómenos físicos presentes. A resolução finita dos aparelhos não permite assumir a exclusividade das linhas espectrais de cada elemento e a incerteza quântica não permite que o processo de identificação dos elementos seja direto às bases de dados existentes (p.e. NIST).

Esta dissertação enquadra-se no plano de investigação do INESC TEC – Centro de Fotónica Aplicada (CAP) e visa contribuir para a quantificação de elementos em amostras geológicas ou outras por técnicas, no CAP desenvolvidas, de processamento de sinal e inteligência artificial com auto-aprendizagem.

O principais objetivos desta dissertação contemplam a contribuição no pré-processamento do sinal LIBS, desenvolvimento de inteligência artificial e a realização de análises comparativas dos diversos métodos do estado da arte de inteligência artificial para os atualmente utilizados em laboratórios avançados. Primeiramente será abordada uma resumida revisão bibliográfica do LIBS e das diferentes técnicas de quantificação para processar o seu sinal. Numa segunda fase, uma exploração do equipamento laboratorial disponível, métodos a serem utilizados e dados considerados relevantes.

Com esta dissertação tem por objetivo demonstrar e testar melhorias no processamento de informação LIBS face ao estado da arte. Parte do material desenvolvido será utilizado na escrita de um artigo científico, a ser submetido para uma revista de especialidade indexada (Q1 ou Q2).

Palavras chave: amostras geológicas; auto-aprendizagem; inteligência artificial; inteligência artificial explicável; *laser-induced breakdown spectroscopy*.

Abstract

Laser-Induced Breakdown Spectroscopy (LIBS) is a technology that offers the possibility to determine fastly and without much preparation, the composition of a sample. LIBS extracts information from the sample by acquiring spectral lines emitted during the molecular breakdown process, allowing the detection and quantification of each present element. The analysis of the different emission lines contains information about chemical elements, isotopes and molecules, interpreting complex samples.

LIBS is an emerging tool that is used in many different areas, from most industrial to the most technical ones, being also used in the healthcare sector. LIBS allows access with high precision to the samples' essence interpreting their composition in a minimally invasive way.

Besides LIBS' practicability, the signal's analysis is not trivial due to its complexity and the physical phenomena present. The finite resolution of the devices does not allow the assumption of the lines' exclusivity and the quantum uncertainty exists, so the elements matching to the existing databases (e.g. NIST) is not direct.

This dissertation is inserted in the research plan of INESC TEC - Center for Applied Photonics (CAP) and aims to contribute to the quantification of the elements on geological samples and CAP's developed techniques for signal processing and application of self-learning artificial intelligence.

This dissertation's prime objectives comprise the contribution to the pre-processing of LIBS signal, development of artificial intelligence, and benchmarking the state-of-art techniques against methods used in cutting-edge laboratories. In a first stage, it will be addressed a resumed bibliography review on LIBS and its different quantification techniques and in a second stage, the exploration of the available laboratory devices, methods to be used and the relevant data.

This dissertation has as objectives the demonstration and improvement testing on LIBS data processing against state-of-art methods. Part of the materials developed will be used on a scientific article to be submitted to an indexed specialized journal.

Keywords: artificial intelligence; explainable AI; geological samples; laser-induced breakdown spectroscopy; self-learning; self-teaching.

Agradecimentos

Cinco anos passaram desde que pisei pela primeira vez o ensino superior e termina agora uma das minhas grandes etapas de vida, o Mestrado Integrado em Engenharia Electrotécnica e de Computadores. Esta foi uma jornada complexa e particular, a qual não seria possível sem a ajuda e motivação de algumas pessoas, às quais presto os meus agradecimentos:

Primeiro quero agradecer ao meu orientador, Dr. Rui Martins, que desde o meu primeiro dia no INESC TEC até à escrita desta dissertação me tem acompanhado sem falhar, passando todo o seu conhecimento e apoio, e pela disponibilidade em me ensinar tópicos valiosos, representando um complemento importantíssimo ao meu percurso académico.

Uma palavra de agradecimento também para o meu co-orientador, Prof. Vladimiro, para o investigador Dr. Sumaili e para o Tiago Rodrigues pela ajuda e pelo despertar para tópicos que representaram uma boa adição para esta dissertaçao.

De seguida um agradecimento para a minha família em especial para o meu pai, a minha mãe e a minha avó pelo apoio e conforto que me proporcionaram desde sempre.

Um grande obrigado aos meus colegas e amigos de curso, em especial ao Fábio, Diogo, Gonçalo, Ricardo, Samuel, Rolando, Bernardo e Bruno pelo espírito de entreajuda, camaradagem e partilha de bons momentos.

Um agradecimento enorme para o BEST Porto, associação de estudantes que abraçei no meu 2º ano de faculdade e mantenho até hoje, pelas experiências que me proporcionou tanto no pequeno escritório da FEUP como em viagens pela Europa.

Um obrigado muito especial para a minha namorada Ana Rita, companheira de todas as horas, rapariga que me motiva a ser uma pessoa melhor todos os dias, amiga que me conforta quando preciso, que celebra os bons momentos da vida e por todas as nossas brincadeiras.

Um muito obrigado também para a Ana Miguel, amiga desde os 6 anos que apesar de não manter um contacto regular, está sempre lá quando preciso e pelo apoio que sempre me deu.

Deixo para o fim, o agradecimento mais importante, para a minha bisavó Tété que tristemente faleceu durante a escrita desta dissertação, perdendo por poucos meses o sonho de um dia me ver com o curso acabado. Mulher de garra que me criou desde bebé, com quem dei os primeiros passos e que desde sempre foi uma inspiração para mim. Um grande beijinho e uma lágrima de saudade. Esta tese é para ti!

Alberto

Contents

List of Figures

List of Tables

Abbreviations

ANN Artificial Neural Network

ASDF Atomic state distribution function

BPNN Back-Propagation Neural Networks

CAP Center for Applied Photonics

CF-LIBS Calibration-Free Laser Induced Breakdown Spectroscopy

CNN Convolution Neural Networks

DBN Deep Belief Network

FWHM Full width at half-maximum

MAPE Mean Absolute Percentage Error

IUPAC International Union of Pure and Applied Chemistry

LDA Linear Discriminant Analysis

LIBS Laser-Induced Breakdown Spectroscopy

LocPLS Local Partial Least Squares

LOD Limit of Detection

LSC Laser-support combustion

LTE Local thermodynamic equilibrium

LV Latent variables

NIST National Institute of Standards and Technology

OSC Orthogonal Signal Correction

PCA Principal Component Analysis

PLSR Partial Least Squares Regression

QDA Quadratic Discriminant Analysis

R Person Correlation

ReLu Rectifier Linear Units

REP Relative Error of Prediction

RI Reference Interval

RMSE Root-Mean-Square Error

RSD Relative Standard Deviation

SIMCA Soft Independent Modelling by Class Analogy

SL-AI Self-Learning Artificial Intelligence

SVM Support Vector Machine

Chapter 1

Introduction

1.1 Context

This dissertation aims to develop an automated smart Laser-induced Breakdown Spectroscopy (LIBS) system. Some of the challenges of applying LIBS techniques include analysing complex samples by extracting the emission lines after a molecular breakdown and interpreting the time-course plasma emission dynamics as a molecular fingerprint for identification and quantification. This emitted plasma is usually not composed of pure elements but a mixture of them, making the identification process complicated. What is more, physical aspects must also be considered, such as the Stark effect, scattering, Doppler effect, straylight effect, baseline removal, the temperature of the plasma, etc. The purpose of applying self-learning AI is to analyse high-resolution spectroscopic data to understand the mechanics and quantum models behind the sample to assure the identification, almost in the near-deterministic category, to be as straightforward as possible.

1.2 Motivation

While other plasmas are induced by electric arc discharges or microwaves, in LIBS systems, as there is not enough energy to ionize the sample directly, it is done by the laser's thermal effect, which can heat it to several thousands of Kelvin. The plasma is produced by the laser pulse interaction in a given sample resulting in an ablation vapour that can be analyzed through its dispersion in the spectrometer and whose atomic emission lines are recognized. Besides the plasma initiation to its expansion and the associated shock waves, are present other phenomena. Many theoretical models have been developed in the last years but are only applicable under idealized conditions. Given the complexity, this kept investigators exploring new approaches, and it is, on that point of view, the starting-point of this dissertation. The challenge includes correction methods and Self-Learning AI to deal with the analysis's hurdle on Laser-Induced Plasmas. Nevertheless, in a very embryonal stage, this emerging tool has gained much popularity against other techniques and, given the promising results, it is being used by NASA on their current missions on Mars.

1.3 Objectives

This thesis is within the framework of the LIBS photonic Research and Development of the Centre for Applied Photonics (CAP) of INESC TEC, giving me, as a student of this Master's, the possibility to work with and to further acquire expertise in cutting-edge technology. The objetives for this dissertation include an extensive analysis of LIBS quantitative methods and data processing to implement SL-AI quantification of elements and molecules in geological samples.

1.4 Structure

The final document will be structured the following way:

- Brief bibliography review on LIBS;
- Possible quantification methods;
- Signal pre-processing;
- Benchmarking Deep Learning algorithms with and without the pre-processing stage;
- Benchmarking Self-Learning AI with state-of-art models.

Chapter 2

LIBS Literature Review

Laser-Induced Breakdown Spectroscopy is a method used to determine a sample's composition. Such is possible by using a high power laser pulse focused upon a certain surface that breaks the molecular structure and creates free electrons, absorbing the rest of the laser energy due to physical phenomena, and creating a crater on the surface of the sample (ablation).

The plasma, containing the removed matter, can be analysed through its emission lines to identify and determine the sample's composition. The apparatus of a typical LIBS application is straightforward, as it is possible to check on figure 1, turning this method more advantageous compared to other analytical methods. However, some drawbacks can also be addressed, such as matrix effects on choosing wrong emission lines for quantification and self-absorption due to in-plasma processes.

The LIBS implementation sample's preparation is not needed, which is significantly beneficial to using it on-site with portable devices.

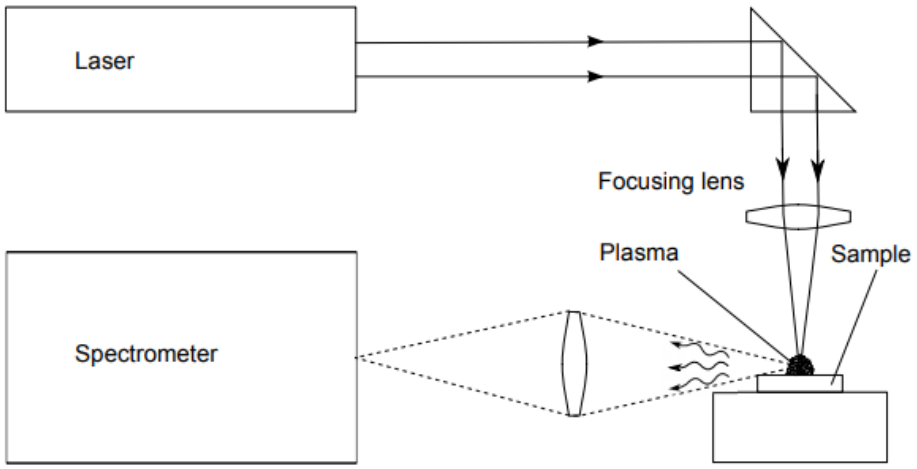

Figure 2.1: LIBS typical apparatus

2.1 Laser Induced Plasma

The breakdown, firstly characterized in the early sixties, results in a spark defined by the generation of ionized gas, as seen in equation 2.1

$$N_D = \frac{4\pi}{3} h^3 n_0 \geq 1 \tag{2.1}$$

where $n_0(m^{-3})$ represents the concentrations of ions and free electrons and h the Debye length, given by expression 2.2

$$h = \sqrt{\frac{\varepsilon_0 k T_e}{e^2 n_0}} \tag{2.2}$$

where $T_e(K)$ is the temperature of the plasma, ε_0 the vacuum's permittivity, k the Boltzmann's constant and $e(C)$ represents the fundamental charge. These expressions (2.1 and 2.2) assure that the plasma is ideal and Debye's ions shielding theory is applicable. On figure 2.2 it is possible to see the line separating the ideal plasma from the non ideal and, marked by a circle, the Laser Induced Plasma.

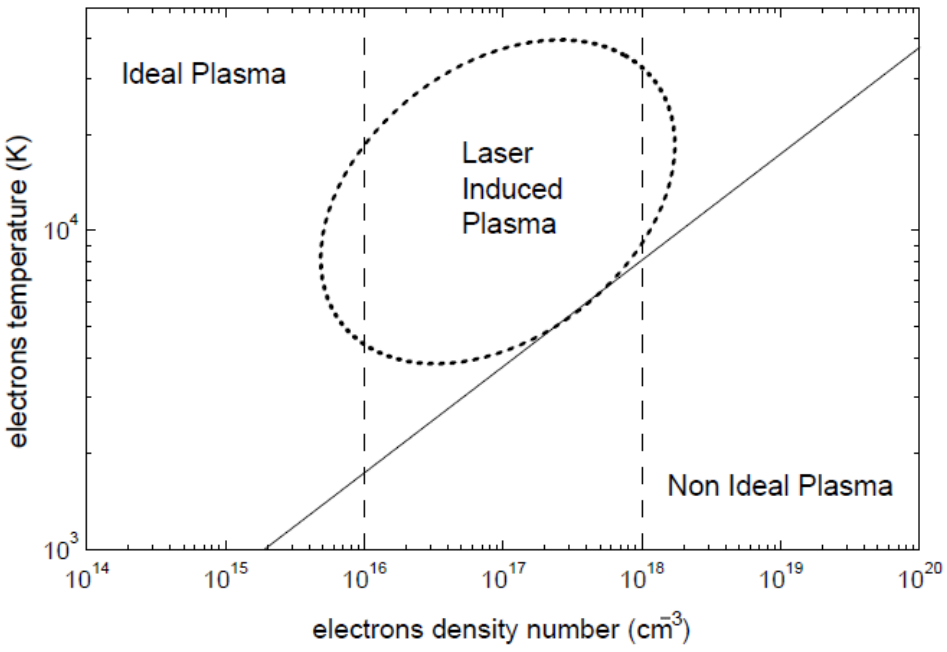

Figure 2.2: The ideal Laser Induced plasma. Extracted from Rakosky [1].

2.2 Laser Induced Breakdown

The energy distributed to the Laser Induced Plasma, similarly to other discharges, is due to kinetic energy generated by the electromagnetic field creating free electrons either by multi-photon

ionization effect or by collisions between electrons and neutrons. Numerically, it is possible to represent the photon's energy by the equation 2.3

$$E = h\frac{c}{\lambda} \tag{2.3}$$

where $h(m^2kgs^{-1})$ is the constant of Plank, $c(ms^{-1})$ the speed of light and $\lambda(m)$ the wavelength.

2.3 Post-breakdown

During the post-breakdown, free electrons are present within the plasma, absorbing the rest of the laser energy that was not consumed in the breakdown, via inverse Bremsstrahlung effect. This process leads to plasma heating and additional ionization [1]. The electron concentration also increases being described by the equation 2.4

$$\omega_p = \sqrt{\frac{e^2 n_0}{\varepsilon_0 m_e}} \tag{2.4}$$

where $\omega_p(Hz)$ is the frequency of the plasma, $e(C)$ the fundamental charge, $\varepsilon_0(Fm^{-1})$ the vacuum's permittivity, m_e the mass of the electron and n_0 the concentration [6].

Above a specific value, plasma loses its transparency. This is due to plasma shielding that impedes the laser pulse to reach the plasma's lower parts. When separated from the plasma, the formed shock wave can supply further heating for the laser to reach the inner parts. This phenomenon, referenced as laser-support combustion (LSC), happens when the irradiance is below a certain threshold, and the resulting vapour starts at the sample's surface [7].

2.4 Local Thermodynamic Equilibrium

To determine the atomic state distribution function (ASDF) collisional processes have to be taken into account, considering both excitation and deexcitation, ionization and recombination phases; and radiative processes comprising photo-excitation, radiative decay and Bremsstrahlung effects.

When the electrons' density is considerably high, the radiative processes are negligible compared to collisional processes. Thus, collisions between electrons are likely to be higher than by photon's absorption or emission. This state is described in the literature as a local thermodynamic equilibrium.

As collisional deexcitation is faster than the radiative one, it applies to the ASDF the Boltzmann distribution 2.5. The temperature will be represented by the Maxwell distribution function 2.6 for the velocity of the electron.

$$n_j = n\frac{g_j}{Z(T)}e^{\frac{-E_j}{kT}} \tag{2.5}$$

$$C_{ik} = n_e \langle v_e \sigma_{ik} \rangle \tag{2.6}$$

where v_e is the velocity of the electrons and σ Born's approximation for collisional cross-section.

Also, to access plasma's LTE, two criteria should be considered. For a certain T Temperature and n_e electron density the expression 2.7:

$$\frac{T(t) - T(t+\tau_{rel})}{T(t)} \ll 1; \frac{n_e(t) - n_e(t+\tau_{rel})}{n_e(t)} \ll 1 \tag{2.7}$$

where τ_{rel} is the relaxation time to set ionic equilibrium. The time can be given by different phases of the Laser-Induced Plasma evolution: directly on the breakdown, during plasma expansion or in the cooldown phase.

Additionally, other criteria 2.8 should also be addressed to deal with plasma inhomogeneity generated by strong gradients. This disturbance on the plasma's edge is a direct consequence of the particles diffusion.

$$\frac{T(x) - T(x+\lambda_D)}{T(x)} \ll 1; \frac{n_e(x) - n_e(x+\lambda_D)}{n_e(x)} \ll 1 \tag{2.8}$$

where

$$\lambda = \sqrt{D\tau_{rel}} \tag{2.9}$$

is the length of the diffusion during relaxation time [8]. The particles in which diffusion time is significantly high, are considered in these criteria, being given by the expression 2.10:

$$D \approx 3 * 10^{19} \frac{kT}{N_{II} M_A} \tag{2.10}$$

where kT is represented in eV, $N_{II}(cm^{-}3)$ is the number of ions and M_A its relative mass.

These criteria are critical to decide where LTE and Saha's equations apply to determine the emission lines, as seen on figure 2.3, showing a typical LIBS spectral emission since the ablation process until the ion emission phase. The method used in this dissertation, better described in section 3, uses several spectral sub-optical lines and projects it into a one-dimensional vector stored in a data matrix for further processing.

2.5 Quantitative analysis methods

LIBS offers great potential for *in situ* and real-time analysis without sampling preparation step. This lack of preparation is translated in quickness and versatility on a qualitative analysis standpoint. For quantitative analysis, we have to consider other aspects as LIBS requires some standards

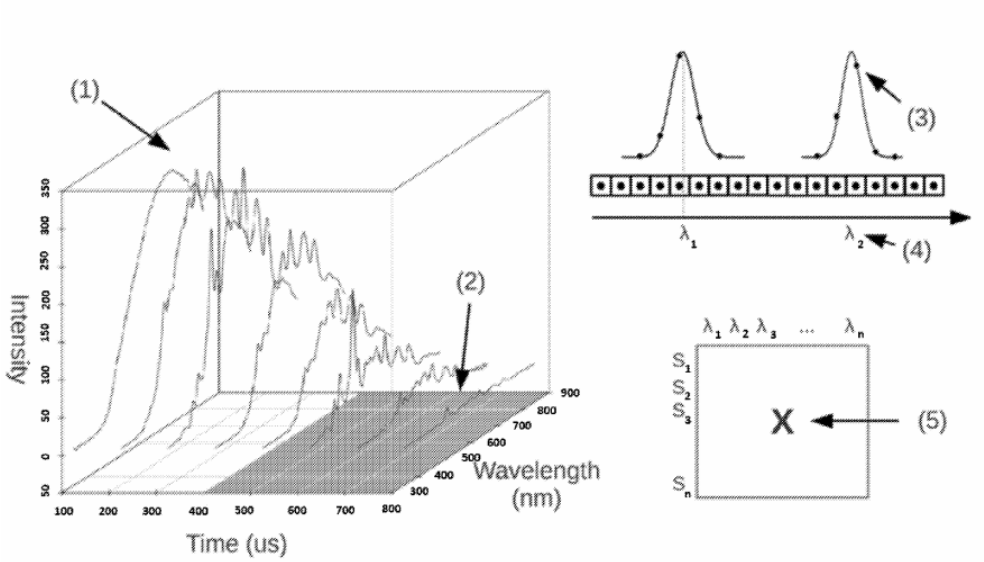

Figure 2.3: LIBS spectral emission in LTE: 1) Laser's ablation; 2) Ion emission spectra; 3) Sub-optical spetral lines; 4) Projection of multiple lines in a one dimensional vector; 5) Data storage of the multiple sub-optical spectral lines in a multiple dimensional vector; (Extracted from Martins, Rui et al. [2])

and calibration. Therefore matrix-matched are not available in some situations. As mentioned before, the matrix effect involves dealing with the ablation process, emission of free electrons, laser's plasma absorption and the resulted temperature.

2.5.1 Calibration Free - LIBS

2.5.1.1 Applications

Calibration Free methods are based on two different equations: Boltzmann's distribution, associating each intensity line to the corresponding temperature on a given ion's excitation state and Saha's distribution, grouping the ions population onto different clusters (the higher the temperature, the distribution changes to upper states). Saha-Boltzmann combines the previous distributions being used to predict plasma's temperature, electronic density and ion distribution for each element. Given the distribution, the concentrations can be estimated in percentage terms. CF-LIBS is a helpful vehicle for problem analysis, particularly the selection of unknown samples. More often, CF-LIBS techniques have been used in industry, namely aluminium [5, 9–12], steel and iron [11–13] and copper [14–19] alloys but also in glasses [20], soils [21–23], and meteorites [24].

The majority of the CF-LIBS mentioned rely on atmospheric air acquired spectra and nanosecond pulse lasers for excitation source. De Giacomo et al. [18] weighed the spectra analysis results of copper allows with nanosecond and femtosecond pulses. In general, and as it is possible to verify on table 2.1, the Nd:YAG laser with 1064 nm of wavelength was the preferred one, while other groups also have used ultraviolet or visible radiation. The variation of the laser's pulse energy varies from about 1 mJ to 100 mJ. Mostly, CF-LIBS has been used to analyse the spectra within

the UV-visible-near IR range. Still, vacuum UV range can also be measured. CF-LIBS analysis on acquired successive spectral windows with ample range spectra has also been proven with pretty acceptable results [12]. The critical aspects are the instrument's accuracy while characterising an entire wavelength range and the plasma's balance during all the measurement procedure.

It is important to mark that the effort to apply CF-LIBS methods was motivated by measuring on-site. Though, it is hard to determine the results' accuracy regarding unknown samples. For that, several statistical analysis to evaluate the performance can be applied [25,26]: multi-element quantitative analysis methods are widely used in spectroscopic activities.

2.5.1.2 CF-LIBS Approach

Ciucci et al. [27] and Tognoni et al. [28] describe the algorithm used to apply CF-LIBS and the underlying assumptions. The idea behind the approach resulted from the conclusion that the emitted plasma is highly sensitive to matrix' changes. The other way around, it is possible to measure matrix' changes through LIBS spectrum analysis. Strictly speaking, Ciucci et al. [27] hypothesized that the information contained in the spectrum is all that is necessary to derive the matrix' composition. Despite the high complexity of the laser-induced plasma which description is not achievable by simply utilising mathematical approaches, several assumptions are frequently adopted:

- The composition of the plasma is representative of the undisturbed target composition;
- The plasma is both temporally and spatially in local thermodynamic equilibrium (LTE);
- The plasma can be represented as a homogeneous source in a spatial measurement window;
- The plasma's spectral lines are individually optical thin.

These techniques have qualitative rightness degrees as Saha-Boltzmann approximations do not comprise all complex phenomena of the plasma.

2.5.1.3 Elements identification

When analysing unknown samples, spectral lines have to be confirmed and matched to a probable element, like so, a correct recognition of the elements is necessary. Also, it is important to choose the proper lines for quantitative analysis. De Giacomo et al. [24] gives four criteria to select the appropriate lines well:

- For a given element, no resonance lines if the concentration is higher than 6000 cm^{-1} to reduce the self-absorption effects;
- Not consider lines with spontaneous emission rate lower than $2x10^{6}s^{-1}$ as this is equal to the plasma's time scale variations;
- Not choose ionic lines according to Saha's equilibrium;
- Avoid high-intensity spectral lines to avert overestimation.

Table 2.1: CF-LIBS methods: quantitative analysis on different samples (Extracted from Tognoni et al. [5])

Target	**Buffer gas**	**Excitation source**	**Hardware**	**Reference**
Aluminium alloy	Air	Nd:YAG 1064nm	Echelle spectrograph and ICCD, spectral range 230–900 nm	[5]
Aluminium alloy	Air	Nd:YAG 1064nm, 1.5 mJ	Grating spectrometer and ICCD, spectral range 200–800 nm	[9]
Aluminium alloy	0.1 mbar air	Nd:YAG 1064 nm, 90 mJ	Grating spectrometer and ICCD, spectral range 220–700 nm	[10]
Aluminium alloy	Air	Nd:YAG 1064 nm, 200 mJ	Ocean Optics LIBS2500-7, spectral range 200–980 nm	[11]
Aluminium alloy	Air	Nd:YAG 1064 nm, 90 mJ	Grating spectrometer and ICCD, spectral range 220–700 nm	[12]
Steel, iron-based alloys	Air	Nd:YAG 1064 nm, 130 mJ	Ocean Optics LIBS2500-7, spectral range 200–980 nm	[11]
Steel, iron-based alloys	Air	Nd:YAG 1064 nm	Echelle spectrograph and ICCD, spectral range 200–1100 nm	[13]
Cooper alloys	Air	Ruby laser 694 nm, pulse width 25 ns	Spectrometer and CCD, spectral range 280–770 nm, spectral range 200–1100 nm	[14]
Cooper alloys	Air	Nd:YAG 355 nm, 20 mJ	Grating spectrometer and ICCD, spectral range 180–850 nm	[15]
Cooper alloys	Air	Nd:YAG 1064 nm	Echelle spectrograph and ICCD, spectral range 200–900 nm	[16]
Cooper alloys	Air	Nd:YAG 1064 nm, 200 mJ	Imaging spectrometer and ICCD	[17]
Cooper alloys	Air	Nd:YAG 532 nm, 280 mJ	Grating spectrometer and ICCD	[18]
Cooper alloys	Air	Nd:glass 527 nm, 2.5 mJ, 280 mJ	Grating spectrometer and ICCD	[19]
Soils and rocks	Air	Nd:YAG 1064 nm, 25 mJ	Echelle spectrograph and ICCD, spectral range 280–1000 nm	[22]
Soils and rocks	Air	Nd:YAG 532 nm, 22 mJ	Grating spectrometer and PMT	[23]
Meteorites	Air	Nd:YAG 532 nm	Grating spectrometer and ICCD	[24]

2.5.1.4 Discussion

According to the earlier literature, CF-LIBS methods approximate the plasma's state-of-art, with successful results on simple samples but fails on the most complex ones as it does not contemplate the other phenomena present in LIBS signal. Nonetheless, CF-LIBS is an acceptable analytical method for unknown samples characterization, despite only providing a qualitative estimation of the trace elements.

The investigated literature is generous on the quantity and quality of information. However, it still lacks statistical approaches for systematic analysis with large data sets and experimental variable settings which would be beneficial for confirmation and completion on the investigation led by Herrera et al. [12] and De Giacomo et al. [24].

2.5.2 Chemometrics-based methods

Unlike the calibration-free techniques mentioned and its local thermodynamic equilibrium assumption, classic chemometrics methods show up as an interdisciplinary procedure to extract valuable information from LIBS spectra in a very accurate, qualitative and quantitative way. These supervised methods [29] are based in a pre-calibration phase with known concentration samples. Some of the most advantageous features gather signal analysis, data processing and pattern recognition.

One of the chemometrics approach techniques is the standard calibration curve [30] that plots the selected line's intensity against known concentrations of a determined calibration set. It is also possible to apply more sophisticated methods like partial least-squares regression (PLSR) or neural networks. These procedures with proven reliability and accuracy on qualitative and quantitative analysis [31–33] have been widely used in LIBS, focusing on improving its performance.

Hernández et al. prove PLSR's efficiency and robustness [34] on matrix effects multivariate treatments when analysing complex spectra, providing good selectivity on LIBS spectra lines to predict elements' concentration [35]. Though it was still pointed out, due to PLSR nature and strong lines' self absorption, it was impossible to completely model the plasma's saturation effect totally. That being said, it is mandatory to use an additional method, as LIBS is characterized by its signal nonlinearity [36]: artificial neural networks (ANN) grant an adequate answer in this matter.

Sirven et al. [37] conducted a study by directly comparing, on the same spectra, the accuracy, precision and limit of detection between the PLSR and ANN techniques. It was used a Nd:YAG laser producing 7 ns pulses with a 3 Hz repetition of a 3 mJ shot. Among other experimental setup hardware, it was used the Oriel MS260i spectrometer and an Andor iStar CCD detector.

Principal component analysis (PCA) is a very common methodology to classify spectral information in Chemometrics [29]. Several authors [38–41] have used this method, giving the possibility to perform a dimension reduction by representing the data set in fewer variables that contain the majority of the relevant information.

Usually, orthogonal signal correction (OSC) is used before applying PLSR to remove uncorrelated concentration's information to improve its overall performance and to reduce the number of principal components to the optimal level [42].

Artificial neural networks (ANN) can be utilised to retrieve spectra's concentration. It is a valorous nonlinear optimisation system with the following mechanism: each neuron is preceded by one or more inputs dynamically weighted. The neuron relates the sum of the inputs to a certain threshold. The computed output is applied via a nonlinear function (sigmoid, tangent). The term "network" comes when the neurons are organised in layers to achieve the nonlinear classification. Sirven et al. [37] used a three-layer network whose inputs receive the wavelengths' intensity followed by a hidden layer and finishing with one neuron (one output) as it is only desired the analysis of one element. The calibration/optimisation phase consists of changing the weights between neurons, so it is possible to relate the predicted and the true concentration values. The validation phase comes right next to where a validation set is used to check the network's performance.

The models' comparison was performed by using an average relative prediction error (REP) to evaluate the accuracy:

$$REP(\%) = \frac{100}{N_v} \sum_{i=1}^{N_v} \left| \frac{\hat{c}_i - c_i}{c_i} \right| \tag{2.11}$$

where N_v is the spectra's validation number, c_i the real concentration and $\hat{c}_i$ its estimation. The precision of the prediction is returned by the relative standard deviation (RSD):

$$RSD(\%) = \frac{100}{N_{conc}} \sum_{i=1}^{N_{conc}} \frac{\sigma_{c_k}}{c_k} \tag{2.12}$$

with

$$\sigma_{ck}^2 = \sum_{i=1}^{p} \frac{(\hat{c}_{ik} - c_k)^2}{p-1} \tag{2.13}$$

where N_{conc} is the validation set's number of unique concentrations and p the sprectra for each concentration. σ_{ck} represents the standard deviation of the concentration σ_k. The limit of detection (LOD), according to the International Union of Pure and Applied Chemistry (IUPAC) definition, is determined by:

$$LOD(ppm) = \frac{3\sigma_a}{b} \tag{2.14}$$

where the signal $s = a + bc$ is obtained by the regression.

2.5.2.1 Calibration Curve

Firstly, to apply the standard calibration curve, several normalizations like spectra smoothing and residual baseline removal were performed.

- Normalization concerning the matrix by aligning the matrix species [43, 44];

- Normalization by background, offsetting plasma's fluctuations due to laser's emitted energy, spatial variations or irregular surfaces [37];
- Normalization by the area supporting the spectrum [45].

As Sirven et al. [37] demonstrates, normalize by the iron line shows to have acceptable results in relative error of prediction (REP) terms. On the other way around, normalizing by the background gives a better relative standard deviation (RSD). The best result for this experiment was obtained by the spectrum area. As the author explains, smoothing does not have a significant improvement in prediction accuracy.

2.5.2.2 Partial Least Squares Regression (PLSR)

Despite the good results obtained via calibration curve method, using a multivariate full-spectrum method like PLSR is foreseen to have better efficiency. To have a valid comparison, PLSR was applied to the same dataset as the one used for the calibration curve. The author selected the five calibration spectra that best represented it, leaving five others to validate the system. PLSR model was after calculated and further characterized. Via optimization concluded that three components were the ideal number to represent the spectra (neither missing valuable information nor overfitting the model). As reported, just a PLSR as a spectra bandwidth's function prove that validation's RSD was improved by more than 40%.

OSC was subsequently applied to the various datasets and comparing the changes again on accuracy and prediction's precision, backed improvements in the RSD, despite enhancing overfitting risk. This time, smoothing improved PLSR predictions as it, by definition, heightens spectral information and suppresses the residual baseline [46].

With this analysis, PLSR was proven not to be suited for a wide dynamic LIBS measurement since only intense lines let a considerable signal trace spectra's elements. That being said, neural networks might have the answer for predicting concentrations over a broader range.

2.5.2.3 Artificial Neural Networks (ANN)

This study used the same dataset, this time with a three-layer neural network with five calibration spectra and five others to validate. The hidden layer was optimized by testing between 1 to 10 neurons: four hidden nodes were proven to be the best performing. For the network calibration, the fittest learning speed was obtained for 50000 iterations. By deriving the spectra, the effect of smoothing was also tested, not showing to improve more than what was achieved on PLSR technique. It was confirmed that the derivative induced a better performance regarding accuracy and precision, but only showing a slightly higher limit of detection (LOD).

When compared to the calibration curve method, all three comparison parameters: REP, RSD and LOD are improved by 2. For PLSR method, they are divided by a factor of 3.

These results illustrate the power of ANN concerning nonlinear variables' relationships, so it is possible to extract valid spectra quantitative information. Even when the concentration range

reduced, the overall performance was hardly impacted, concluding Sirven et al. that it was the most powerful tool tested.

2.5.2.4 Discussion

As for the conducted research among chemometrics-based methods (calibration curve, partial least squares regression and neural networks methods), it was used a spectra dataset obtained in a controlled LIBS environment giving the possibility to investigate how the spectra normalization, smoothing or derivation can impact accuracy and precision of the results. Regarding the mentioned methods, neural networks were established as the best performing method, almost twice as good as the standard calibration technique. PLSR had the worst performance principally due to the nonlinearity within spectral intensities and plasmas' self-absorption.

2.5.3 Deep Learning methods

Notwithstanding the previous techniques, other types of networks can be addressed, i. e. convolutional neural network (CNN). CNN is a deep learning algorithm inspired by animals' visual perception and shows reliable results on image recognition and classification. Due to its success, CNN application has been extended to hyperspectral image classification in LIBS [47].

L.-N Li, et al. [48] study reports a deep CNN construction composed by diversified layers and using one-dimensional data for the quantitative analysis. CNN performance was compared with two alternative schemes: back-propagation neural networks (BPNN) and partial-least squares (PLS).

For that, a preprocessing stage was first applied: subtraction of the dark spectrum and the spectra's calibration. When applying BPNN scheme, generally to perform dimension reduction, it is fundamental to add feature selection, while for a mere CNN algorithm, it can recognise these features.

2.5.3.1 Build the CNN

Several built-in libraries (Matlab/Python or R packages) can be used to develop deep CNN. Specifically, L. -N Li, et al. used the sequential model with the following characteristic architecture and which graphical diagram can be seen in figure 2.4.

1. Normalization Layer;
2. Convolutional Layer;
3. Pooling Layer;
4. Convolutional Layer;
5. Pooling Layer;
6. Flatten Layer;

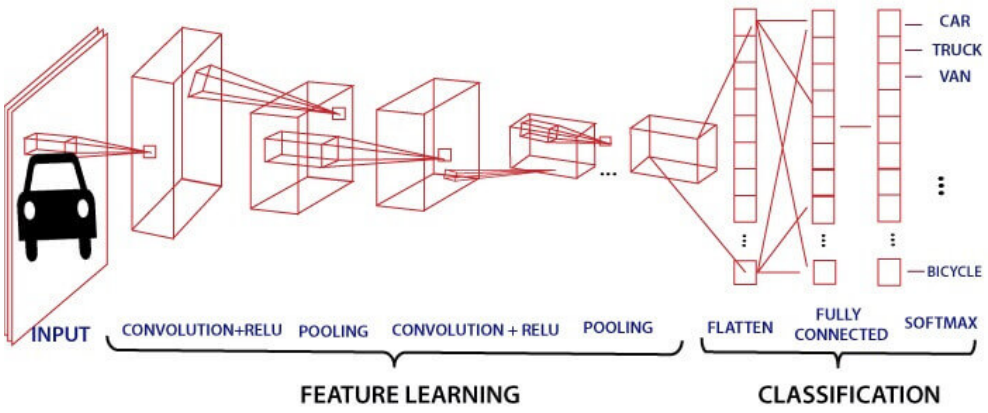

Figure 2.4: Schematic of a traditional CNN sequential model architecture [3]

7. Dense Layer;

From the figure 2.4 schematic it is possible to detect its core components: convolutional layers extract and recognize features onto a features map; pooling layers merge into one value the precedent convolution layers; the flatten layer converts the multi-dimensional matrices into one vector; the flatten layer, or as shown on figure 2.4, "fully connected layer" correlates the vector's features and the classification result. The output is computed by an activation function such as sigmoid or softmax that classifies the data.

The *modus operandi* for a CNN application usually consists in a division into a training and a validation set. The training data is introduced in the network, updating the network weights in every iteration. This is done by outputting the predictions on the already entered data and comparing to the real values to detect if it correctly fits or if it is needed a weight rebalance. Once the training process is completed, the validation data is applied to the network to monitor its predictive aptitude and evaluate in a pre-defined metric; for instance, the root mean square error (RMSE):

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(\hat{y}_i - y_i)^2}{N}} \tag{2.15}$$

where N represents the number of observations, $\hat{y}_i$ the predicted values and y_i the observed values. This performance test can be perfomed by applying a cross-validation strategy that leaves one out to examine the network's accuracy, as explained in [49].

2.5.3.2 Performance and discussion

CNN was compared with BPNN sequential model and the both PLS models: PLS-1 that regresses a single variable and PLS-2 that regresses two or more response variables [50, 51]. It was applied the same training and validation data and the same cross-validation performance evaluator for every method.

By this research, L. -N Li, et al. [48] closed that CNN was the best performing method for almost all the validation dataset and showing the best accuracy in RMSE terms. As for the BPNN and PLS-2 the results were quite similar although for PLS-1 came shorted than expected. The probable explanation given by Dyar [52] for the weak overall results of PLS-1 due to ignoring elements' correlations, which naturally pushes up the PLS-2 accuracy as it takes it into account.

2.5.4 Self-Learning Artificial Intelligence (SL-AI)

As presented before, the existing approaches aim to predict a sample's composition by applying a training dataset, having a high dependency on previous knowledge. Chemometrics or Deep Learning both fail on the full dynamic characterization when applied on plasma emission spectroscopy, the LIBS case. They rely on black-box approaches, giving a determined mathematical result when applying certain elements' input, based on algorithms without any reasoning or interpretability. These neural network-based methods that focus on creating complex feature maps are extremely inefficient in identifying and quantifying elements with very close wavelengths or coincidental band regions. Furthermore, when new data is compiled, the whole features map has to be rebalanced, making it computationally-intensive.

INESC TEC's Center for Applied Photonics has the main research focus on cutting-edge laser's physics and machine learning approaches and, for the last years, advanced with a solution "Self-Learning Artificial Intelligence (SL-AI)", providing an accurate compositional prediction, LIBS-complex samples compatible. The success of these technological advancements led to the register of an international patent under the number WO/2020/026165 and which stated methods are the starting point for this master thesis. On section 3 this method will be further analysed.

2.5.5 Artificial Intelligence - Methods Benchmarking

Regarding several studies that have reported successful techniques to quantify and identify LIBS spectral elements correctly, a natural question arises: which reported method has the best-proven accuracy. This section will address a few studies that have been later reported on LIBS field of investigation for comparison and benchmarking purposes.

In 2019, in China, Zhao et al. [53] reported a study to predict lead-contaminated soil. It was purposely added distinct Pb levels to different soil samples in which was planted tobacco, for two-four weeks. To evaluate data's classification performance, the study was conducted comparing Deep Belief Network (DBN)'s accuracy with two alternative methods: Support Vector Machine (SVM) and Partial Least Squares Discriminant Analysis (PLS-DA). It was used a Chinese Q-Switch Vlite 200 laser with 8 ns of pulse duration, 300 mJ of maximum energy and wavelength of 1064 nm. The spectrometer was an English ME5000.

Concerning methods, SVM, a hyperplane separating algorithm, was chosen with a Gaussian radial basis function. The DBN consisted of two hidden layers with 100 and 30 neurons and applying a Rectifier Linear Units (ReLu) as the activation function.

The presented results for the accuracy of lead-contamination soils detection, as seen in Table , display great overall performance for all methods, putting DBN as the most accurate method but, attending the nature of the question, detection of contaminated soils, the results still are not acceptable.

In 2018, Zhang et al. [54] applied LIBS for geological purposes. It was adopted soft independent modelling by class analogy (SIMCA) and Principal Component Analysis (PCA) in order to classify several geological samples. The French Nd:Yag Laser, with its fundamental wavelength of 1064 nm and 10 ns of pulse width, was used. The spectrometer was a Mechelle 5000 with an ICCD camera. 25 types of geological samples were used in this experiment to identify differences within the laser spectra and trace the elements impurities.

Despite the high variability of certain rock elements, Zhang et al. [54] still got an overall classification accuracy of 94,1% which is pretty acceptable for geological exploration.

Metzinger et al. [55] conducted a study to measure coal's capability for quality control. Unlike the other authors, this investigation utilised a hand-held LIBS to study six different coal samples. The initial LIBS spectral lines were reduced to 18 variables of interest and classified with Linear Discriminant Analysis (LDA). All measurements were made with a portable LIBScan 25+ with a ND:Yag laser procucind 50 mJ of maximum energy, with 1064 nm of fundamental wavelength and 4 ns of pulse duration. The spectrometer was a fiber-optic AvaSpec-2048FT. As described, the cross-validation test of the produced model, reported a 95,33% of classification accuracy indicating the excellent results for LIBS applicability for coals' quality control.

LIBS was also implemented in the veterinary field to classify blood hemograms when analysing cytometry, namely, erythocytes, hemoglobin concentration, and hematocrit. Barroso TG et al. [56] investigation proves how efficient self-learning can be compared to statistical approaches based on correlations. The prototype, an Vis-NIR spectroscopy system, can measure the entire blood spectrum with a single blood drop.

Another example of Deep Learning algorithms on blood samples was presented in 2020 by Gaudioso et al. [57] to detect Alzheimer's diseases in patients above 65. Gaudioso et al. [57], in their research, used an apparatus composed of an Nd:YAG laser at 1064 nm of wavelength and 30 mm of focal length; and an Echelle spectrometer. As reported, the Quadratic Discriminant Analysis (QDA) method achieved an accuracy between 75% and 85%, obtained by a manual feature selection to differentiate the spectra.

Similarly, YanWu et al. [58] used LIBS to discriminate nasopharyngeal carcinoma serum in 2018 by Extreme Learning Machine - Random Forest, achieving an accuracy of 97,7%.

Among several methods used to benchmark the results, the most notable techniques, Local Partial Least Squares (LocPLS) and Self-Learning Artificial Intelligence (SL-AI) revealed excellent metrics: LocPLS with 90,4% within the reference interval (RI) for the count of red blood cells in canine species but SL-AI with killing results above 92% inside the reference interval for canine species and 100% of accuracy outside RI in feline species.

With the proviso of comparing different sample types, SL-AI on blood samples proves to be the most efficient one for identification and classification comparing to Deep Learning algorithms.

The excellent predictions even on small datasets and the reduced computing cost compared to other methods like chemometrics or ANN-based closes the bibliography review concluding that, for complex samples like on geology where a full dynamic fingerprint for the different elements is needed, Self-Learning AI is the most appropriate method to use and will be further discussed on the next section.

Table 2.2: SL-AI vs Deep Learning - Methods Benchmarking

Hardware	**Wavelength**	**Samples**	**Methods**	**Metrics/Results**	**Reference**
Q-Switch Vlite 200 & ME5000	1064 nm	soil samples	PLSDA	Acc: 78,63%	[53]
			SVM	Acc: 82,64%	
			DBN	Acc: 84,11%	
Nd:Yag Laser & Echelle 5000	1064 nm	geological samples	SIMCA	Acc: 94,1%	[54]
Nd:Yag Laser & AvaSpec-2048FT	1064 nm	coal samples	LDA	Acc: 95,33%	[55]
Vis-NIR System	200-1200 nm	blood samples	LocPLS	Acc: 90,4%	[56]
			SL-AI	Acc: 100%	
Nd:Yag Laser & Echelle 5000	1064 nm	blood samples	QDA	Acc: 75%	[57]
Nd:Yag Laser & Echelle 5000	532 nm	blood samples	ELM-RF	Acc: 97,7%	[58]

Chapter 3

Self-Learning Artificial Intelligence (SL-AI)

Self-Learning Artificial Intelligence (SL-AI), the technique used to quantify and identify geological elements in this dissertation, enters explainable AI space. Meanwhile, other approaches like ANNs and Deep Learning can have acceptable results for certain applications and even exceed humans performance they fail in depending on big data and computational power. Explainable AI allows the user to understand how the values are outputted and why the system's entered input gave a certain response back based on that given input's contribution. The explainable network grows in the form of an inverted tree shape grouping the results by explainable features. This is particularly interesting as the programmer can, in any time, understand what is happening, how it is happening, when it failed, when it succeeded, etc [59] and act in conformity with it.

3.1 SL-AI for LIBS

The process of applying SL-AI consists of a first characterization process based on the sample's composition. This characterization extracts the spectral information of the sample after the shot of an intense laser beam on one or more atoms/molecules/isotopes:

- Acquire electromagnetic spectral information's resolution by gathering one or more spectrums;
- Extract one or more spectral lines where its information is sub-optical;
- Project sub-optical lines into a deterministic feature space which is a multiple dimension vector that aggregates several spectral line.

The current CAP's state-of-art method extracts, very accurately, sub-optical lines that can be used as feature variables to further identify and quantify the sample's constituents. This method does not rely on big-data information. Instead, there is a search in the trained deterministic feature space for the matching spectral lines and an explainable model is given throughout each spectral

lines' contribution and are computed for the quantification and identification of the sample elements. These contributions can be used to understand the plasma emitted dynamics and quantum models to develop new instrumentation [56].

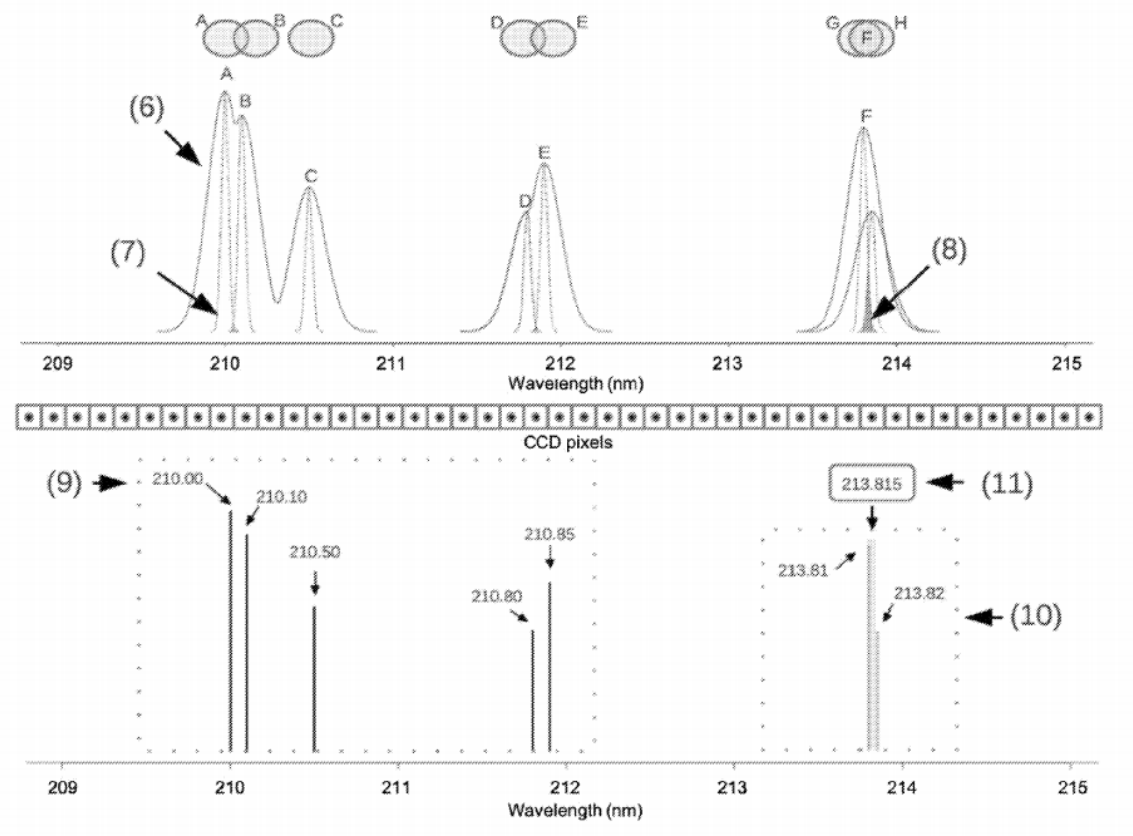

Figure 3.1: Extracion of LIBS spectral sub-optical lines (Extracted from Martins, Rui et al. [2])

Figure 3.1 presents how the LIBS spectral band is extracted and how the lines are deconvoluted. Represented by the number 9 are the lines recognised as exclusive and by number 11 the interferent lines.

On the contrary for Deep Learning solutions, SL-AI has the ability to add more data to the model without having to calculate the entire system. Moreover, as it is a self-supervising construction model, it can be continuously improved by further learning how the lines should be understood. This is a key characteristic for this thesis subject, as there is not much information on geology samples, i. e. big-data databases.

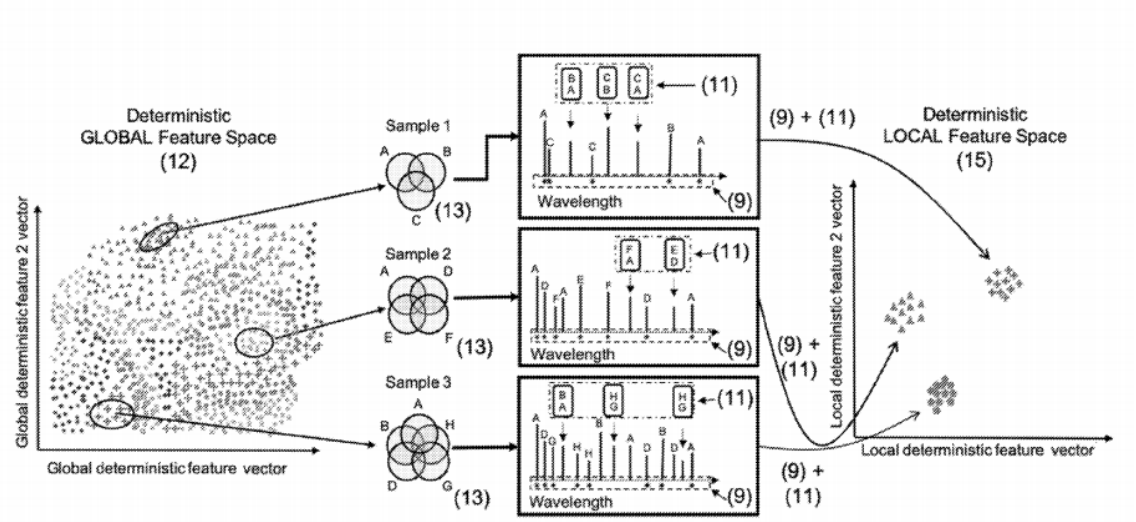

Figure 3.2: Construction of the constituents dynamical fingerprint (Extracted from Martins, Rui et al. [2])

Figure 3.2 shows the building process of a local feature space (15) from the analysis of various samples (13) from the global feature space (12). The process of self-learning consists of gathering different samples of both exclusive (9) and interferent (13) lines from different regions allowing to build a dynamic fingerprint of the constituent.

3.2 SL-AI algorithm

The process to achieve a correct identification and quantification of the sample elements pass through several steps. Firstly it goes through a pre-processing stage and afterwards the processing stage. As seen in figure 3.3, the emitted plasma spectra has its peaks extracted by a local maximum algorithm. For each maximum is performed a Gaussian fit. The independent peaks are used to calculate the median FWHM for determining the spectral resolution, defining interest lines. To build the global deterministic feature space, spectral resolution is used to search in the existing LIBS Databases (i.e. National Institute of Standards and Technology (NIST)) for the spectral lines that are specially tunned on that same spectral resolution, according to Richardson-Lucy deconvolution algorithm and extract both independent and interferent lines for the deterministic variables. Parallelly, samples' emission lines are used to deconvolute the plasma to minimize the LIBS's interferences, as explained in 2 section. The final lines, comprising both independent and interferent, generate the global deterministic feature space. Given the global space, the constituents of interest are selected to achieve the local deterministic feature space (constituent fingerprint) for identification and classification.

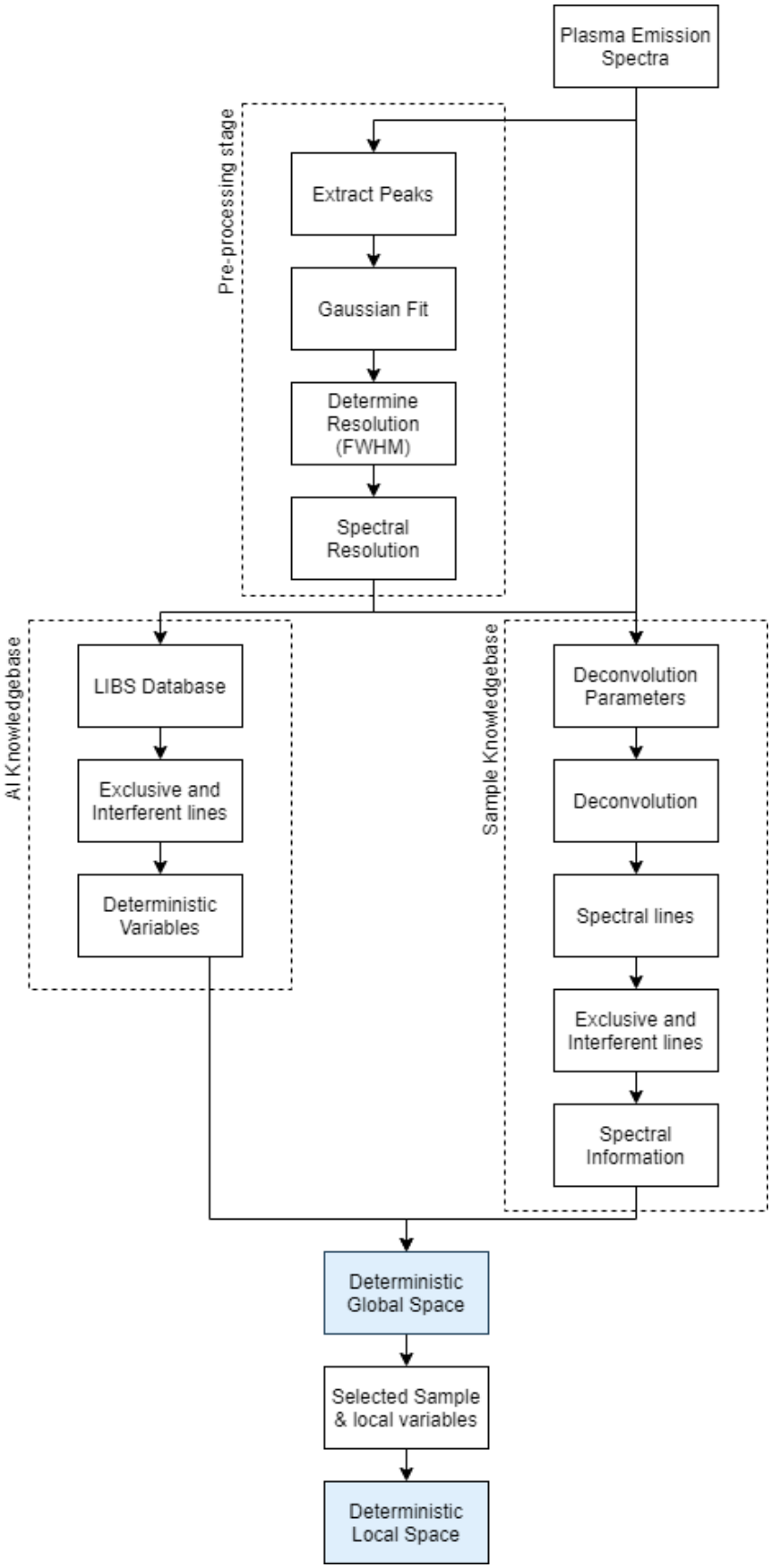

Figure 3.3: Workflow of the identification process (Adapted from Martins, Rui et al. [2])

Chapter 4

Materials and Methods

4.1 Data extraction and processing

From Barroso pegmatite mine, in the north-central side of Portugal, a drill hole 128 meters deep was studied to validate LIBS measurements for Lithium exploration. The powder from the hole was processed and kept meter by meter and analysed on INESC TEC - Center for Applied Photonics facilities.

It was used a 8 *ns* pulse duration and 200 *mJ* of maximum energy Q-switched Nd:Yag laser (Quantel CFR200) with a fundamental wavelength of 1064 *nm*. The beam (4,74 *mm* of diameter and a quality factor of 5.32) was focused on a plano-convex lens with 20 *cm* of focal distance followed by a sapphire window with 1.7545 refractive index, placed 10 *cm* away from the lens. The emitted light produced upon the fire on the sample was captured by an Avantes CCD Spectrometer covering a range of wavelengths comprehended between 180 and 920 *nm* spread over 8 channels, each one of them connected to its independent optical fiber for collection. Seven from these set of optical fibres are from the Sarspec Optical Fibers UV/VIS type with 200 *cm* of length and 200 μm of diameter and stainless steel shielded and the last 8th from Avantes FC-UV200-1.5-SR Optical Fiber type with 150 *cm* of length and 200 μm of diameter, solarization resistant. The data is collected through appropriate software, AvaSoft.

Two independent types of powder were analysed on LIBS, and the respective datasets were extracted. The standards for the experiment involve a Clean Shot with 410 μs delay and four shots with 340 μs Q-Switch delay. The data comprises 16361 wavelength observations (177,845 *nm* to 926,621 *nm*) for 124 meters on the first dataset and 76 meters on the second. The powder from the mine was also characterised by an Inductively Coupled Plasma Mass Spectrometer (ICP-MS), which is a crucial tool to evaluate quantifications accuracy, as in this methods, it is the equipment that gives the reference value. The Lithium lines considered for this study were the 610,5 *nm* and the 670,8 *nm* ones, based on the NIST database.

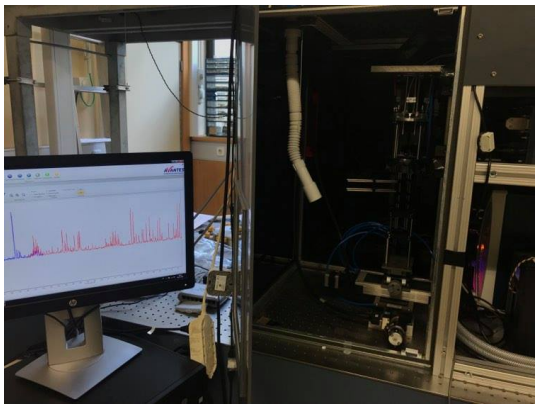

Figure 4.1: Setup LIBS e AvaSoft

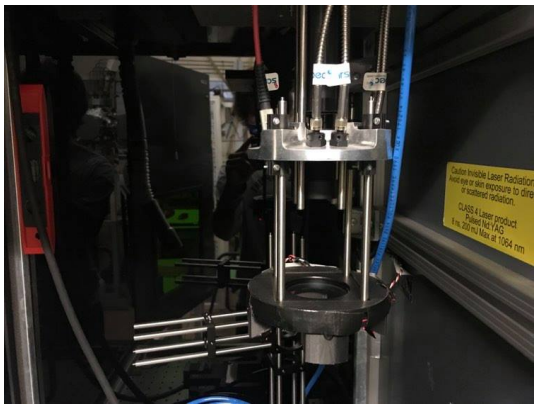

Figure 4.2: Optical Fiber apparatus

Given that it will be followed some data-mining on the data alongside with three different approaches to identify the different dataset clusters, firstly through a more empirical algorithm MeanShift and afterwards, the most implemented algorithms Support Vector Machine and Artificial Neural Networks. These will be further explained in the following sections.

4.2 Data-mining

On the data provided by the ICP-MS, considered the ground truth for LIBS methods, it was applied a Principal Component Analysis (PCA) as it is a vital tool to pre-analyse the data. PCA is a non-supervised data-mining technique that allows easy exploration of a batch of data, namely by the dimensional reduction without compromising the majority of the variance and identifying directions on the dataset (principal components). This is an essential tool that allows the distinction between required and discardable information.

The original dataset has a compositional classification in the many present elements from the periodic table (Li, Be, Ba, ...) along all the 124m depth of the ore, from which it was possible to measure 40 on the available equipment.

The original dataset has a compositional classification in the many present elements from the periodic table (Li, Be, Ba, ...) along all the 124m depth of the ore, from which it was possible to measure 40 on the available equipment. As 40 elements are many variables and arduous to analyse, the dataset was reduced to just a few by applying a PCA. PCA identifies the principal components as linear combinations of the original dataset. The first principal component aggregates most of the variance, followed by the second principal component, which is orthogonal to the previous, aggregating the most significant variation in that particular direction. The orthogonality of the principal components ensures the uncorrelation of the data in those dimensions. Once the samples are projected on the first components, it is possible to establish likelihoods and extract patterns.

4.3 Information Theory

Information Theory (Information-Theoretic Learning) uses concepts of entropy and Parzen windows to efficiently cluster data. One of the most common algorithms for this regard is Mean Shift, an algorithm with mode seeking properties for density probability functions that can be used to generate new sample points around the original data for better clustering. With sample densification, it is possible to generate a larger number of samples that belong to the original cluster mode. That gives the possibility to better train a neural network in a scenario that lacks data for technical difficulties, or cost-wise, are expensive.

4.3.1 Renyi's Entropy

The concept of entropy used in Mean Shift algorithm was brought by Rényi in 1950 as a generalization of other entropy previously presented. Rényi's non-parametric entropy estimator depends on a α parameter of a discrete aleatory variable P with the following probability density function p_k:

$$H_{R\alpha} = \frac{1}{1-\alpha} * \log[\int_P p_k dx], 0 < \alpha < \infty, \alpha \neq 1 \tag{4.1}$$

The authors of [60,61] associate Renyi's Entropy of Equation 4.1 with the Parzen windowing technique given by Equation 4.2, where G represents the Gaussian kernel and is a starting point for an easier non-parametric estimation:

$$G(x) = \frac{1}{(2\pi)^{k/2}\det(\Sigma_x^{1/2})} * \exp(-\frac{(x-x_i)\Sigma_x^{-1}(x-x_i)^T}{2}) \tag{4.2}$$

where Σ is the covariance matrix, that for simplicity can be assumed as $\Sigma = \sigma^2 I$. For a dataset $X(i) = \{x_1(1), x_2(2), ..., x_i(n)\} \,|\, i = 1, ...N$ the probability density function can be estimated as:

$$f_X(x) = \frac{1}{N}\sum_{i=1}^{N} G(x-x_i, \sigma^2) \tag{4.3}$$

4.3.1.1 In-cluster entropy

Substituting Equation 4.3 on Equation 4.5 it is possible to obtain what the authors of [60,61] define as Renyi's quadratic entropy:

$$H_R(x) = -\log \int f_X(x^2) dx, \alpha = 2 \tag{4.4}$$

The quantity, or Information Potential, is a decreasing function that takes into account the distance between the x_i and x_j samples in a similar way the potential energy physical particles do.

$$V(x) = \sum_{i=1}^{N}\sum_{i=1}^{N} G(x_i - xj, 2\sigma^2) \tag{4.5}$$

4.3.2 Cluster Evaluation Function (CEF)

4.3.2.1 Between-cluster entropy

Rather than using the Information Potential described on Equation 1, Jenssen et al. [62] compute a double sum on all dataset including the membership function $M(x_{ij})$ that takes the value 1 for a pair of x_i and x_j belonging to different clusters and the value 0 if within the same cluster. The entropy is given by:

$$H(C_1, C_2, ...C_N) = -\log V(C_1, C_2, ..., C_N) \tag{4.6}$$

$$V(C_1, C_2, ...C_N) = \frac{1}{2\prod_{k=1}^{N} K_N} \sum_{i=1}^{N} \sum_{i=1}^{N} M(x_{ij}) G(x_i - x_j, 2\sigma^2) \tag{4.7}$$

for each cluster C_n for n in a range $n = 1, ..., N$. When there is high separability, the values $V(C1, ..., C_N)$ are low and contrarily, $H(C1, ..., C_N)$, are large.

4.3.3 Information Theoretic Mean Shift - Algorithm

Information-Theoretic Mean Shift (IT-MS) is a non-parametric mode finding, and clustering algorithm that uses a kernel developed system for surface space searching. This algorithm has been employed in many ways, being the most common: the Gaussian Blurring Mean Shift (GBMS) and the Gaussian Mean Shift (GMS). On a specific data frame with different and independent samples $X = (x_i), i \in 1...N$, with the previously explained Parzen window technique, it is possible to express the probability density function for a positive, not null, bandwidth:

$$p(x) = \frac{1}{N} \sum_{i=1}^{N} G_\sigma(\|x - x_i\|^2) \tag{4.8}$$

where $G_\sigma(t) = \exp(-\frac{t}{2\sigma^2})$.

The mode of the probability density function can be adjusted from the stationary point theory $\bigtriangledown p(x) = 0$ to an iterative fixed-point like suggested by Sumaili et al. [63]:

$$x^{\tau+1} = m(x^{(\tau)}) = \frac{\sum_{i=1}^{N} G_\sigma(\|x - x_i\|^2) x_i}{\sum_{i=1}^{N} G_\sigma(\|x - x_i\|^2)} \tag{4.9}$$

The difference term $m(x) - x$ moves the points recursively to their corresponding cluster modes as described by Fukunaga et al. [64] and later named as Gaussian Blurring Mean Shift (GBMS). The variation to the Gaussian Mean Shift (GMS) slightly changes: the modified dataset on every iteration is no longer used, and it is only taken into account the original data frame X_0:

$$m(x) = \frac{\sum_{i=1}^{N} G(\|x - x_i\|^2) x_0}{\sum_{i=1}^{N} G(\|x - x_i\|^2)} \tag{4.10}$$

With the consideration of Equation 4.5 and the Gaussian result of the product between two Gaussians, with a variance corresponding the sum of variances, Renyi's quadratic entropy parametric estimator is:

$$H(X) = -\log(V(X)) \tag{4.11}$$

$$V(X) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i=1}^{N} G\left\|\frac{x_i - x_j}{\sigma'}\right\|^2 \tag{4.12}$$

where $\sigma' = \sqrt{2}\sigma$.

The Cauchy-Schwarz distance between the two probability density functions can be computed as suggested by Jenssen [62]:

$$D_{CS}(X;X_0) = \log\frac{\int p(x)q(x)dx}{\sqrt{\int p^2(x)dx \int q^2(x)dx}} \tag{4.13}$$

$$D_{CS}(X;X_0) = 2H(X;X_0) - H(X_0) - H(X) \tag{4.14}$$

being $p(x)$ the probability density function of X and $q(x)$ the probability density function of the original dataset X_0.

Also, it is possible to modulate the dataset $X = (x_i)_{i=1}^{N}$, minimizing the entropy while keeping a fixed Cauchy-Schwarz distance with a Lagrange multiplier:

$$J(X) = min_X H(X) + \lambda[D_{CS}(X;X0) - k] \tag{4.15}$$

The update rule with the differentiation of Equation 4.21 to each x_i:

$$x_i^{\tau+1} = \frac{c1\sum_{j=1}^{N} G(\left\|\frac{x_i^\tau - x_j^\tau}{\sigma'}\right\|^2)x_j^\tau + c_2\sum_{j=1}^{N} G(\left\|\frac{x_i^\tau - x_{oj}^\tau}{\sigma'}\right\|^2)x_{oj}^\tau}{c1\sum_{j=1}^{N} G(\left\|\frac{x_i^\tau - x_j^\tau}{\sigma'}\right\|^2) + c_2\sum_{j=1}^{N} G(\left\|\frac{x_i^\tau - x_{oj}^\tau}{\sigma'}\right\|^2)} \tag{4.16}$$

where $c1$ is defined as $c_1 = (1-\lambda)V^{-1}(X)$ and c_2 as $c_2 = 2\lambda V^{-1}(X;X_0)$. This λ Lagrange Multiplier can have different values accordingly to the case in interest.

4.3.3.1 Case: $\lambda = 0$

When selecting $\lambda = 0$ the cost function of Equation 4.21 becomes equal to:

$$J(X) = min_X H(X) \tag{4.17}$$

And the corresponding update function of Equation 4.16 for the fixed point:

$$x_i^{\tau+1} = \frac{\sum_{j=1}^{N} G(\left\| \frac{x_i^\tau - x_j^\tau}{\sigma'} \right\|^2) x_j^\tau}{\sum_{j=1}^{N} G(\left\| \frac{x_i^\tau - x_j^\tau}{\sigma'} \right\|^2)} \tag{4.18}$$

On this case, for $\lambda = 0$ the Mean Shift algorithm behaves as a Gaussian Blurring Mean Shift (GBMS), as it is possible to verify with Equation 4.9. This operation mode seeks to minimize the overall dataset entropy and converges to a single point which is the mode of the probability density function. One thing that has to be accounted is the stopping criteria in case no mode is found, to avoid a infinite and unstable loop.

4.3.3.2 Case: $\lambda = 1$

When selecting $\lambda = 1$ the cost function of Equation 4.21 becomes equal to:

$$J(X) = min_X H(X; X_0) \tag{4.19}$$

And the corresponding update function of Equation 4.16 for the fixed point:

$$x_i^{\tau+1} = \frac{\sum_{j=1}^{N} G(\left\| \frac{x_i^\tau - x_j^\tau}{\sigma'} \right\|^2) x_{oj}^\tau}{\sum_{j=1}^{N} G(\left\| \frac{x_i^\tau - x_j^\tau}{\sigma'} \right\|^2)} \tag{4.20}$$

On this case, for $\lambda = 1$ the Mean Shift algorithm behaves as a Gaussian Mean Shift(GMS). This operation mode seeks to minimize the cross entropy $H(X; X:0)$ and converges to the mode of the original probability density function without needing any stopping criteria.

4.3.3.3 Case: $\lambda > 1$

When selecting $\lambda > 1$ the cost function of Equation 4.21 becomes equal to:

$$J(X) = min_X H(X) + \lambda [D_{CS}(X; X0) - k] \tag{4.21}$$

When λ takes a higher value, the Cauchy-Schwarz distance $D_{CS}(X; X_0)$ gets more relevant, highlighting the most relevant features of the original dataset. For λ values comprehended between 1 and 2, the shape of the distribution converges to the primary data curve.

Rao et al. [61] suggest on their research stopping criteria for both GMS and GBMS algorithms. For GMS, a function that tracks the average distance of a particle being less than a predefined tolerance value can be implemented:

$$\frac{\sum_{i=1}^{N} \left\| x_i^\tau - x_i^{(\tau-1)} \right\|}{N} < tol \tag{4.22}$$

For GBMS algorithm, due to its nature instability it is suggested a stopping criteria based on Shannon's Entropy:

$$\left| H_S(d^{(\tau+1)}) - H_S(d^{\tau}) \right| < 10^{-8} \tag{4.23}$$

where $H_S(d) = -\sum_{i=1}^{B} f_i \log f_i$ is the Shannon Entropy.

4.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning technique firstly introduced by Vapnik [65] in 1998 and is mostly applied for pattern recognition or computer vision, which classifies data in two different regions. In a simple way, SVM inputs the data points and gives, as an output, a hyperplane that best suits the given data and acts as a decision barrier:

$$ax + b = 0 \tag{4.24}$$

The process of creating the hyperplane is achieved by selecting a group of support vectors, a subset of the vectors defined between the data points and the decision boundary that maximize the margin and whose training points $(x_i, y_i), i \in 1, ..., N$, satisfy the constraint:

$$y_i(a * x_i + b) \geq 1 \tag{4.25}$$

Although SVM is a technique that separates data linearly, it is possible to apply it to non-linear problems as it is possible to represent the data on spaces dimensionally larger and have a possible physically linear separation between groups.

The optimization of the algorithm is achieved by:

$$min\phi(a,b) = \frac{1}{2}\|a\|^2 + \lambda \sum_{i=1}^{N} \varepsilon_i \tag{4.26}$$

subject to restraint 4.25, where λ is the cost for the variable ε that represents the slack, allowing the algorithm to have some miss classifications.

For non-linear classifications it is possible to apply linear kernels but also polynomial functions or radial based:

- Linear: $x_i * y_i$
- Polynomial: $(x_i * y_i + z)^{\alpha}, z > 0, \alpha \geq 2$
- Radial Basis Function (RBF): $e^{-\theta \|x_i - x_j\|^2}$

4.5 Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a network sparked by the essence of biological neural networks that process, parallelly, information on its interconnections named neurons, shaped accordingly to perform certain mathematical operations.

Associated to every link is a specific weight attached, forwarding the information to the following nodes, similarly to synapses on biological brains, and heading to an activation function that activates the state of the neuron (sigmoid, ReLu). This "synapse" results in the sum of the inputs of the previous nodes with their corresponding weights.

This technique is highly versatile and can be employed for pattern recognition purposes, data classification, compliance detection, food and medical diagnosis,... The learning process comes from the result of adjusting the weights of the interconnections of the nodes.

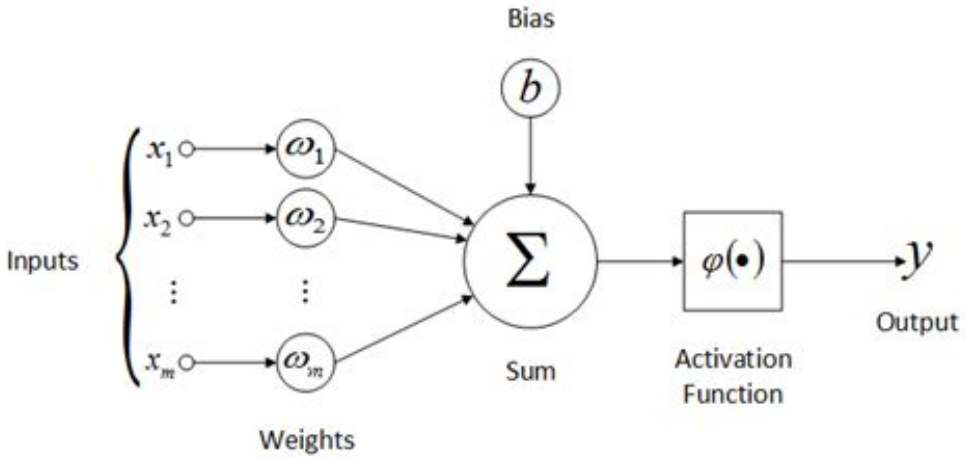

Figure 4.3: Artificial node/neuron

Each input $x_i, i \in 1,...,m$ is computed accordingly to their branch weight ω_i and summed to reach an activation function φ, as seen in Equation 4.27.

$$Sum = \sum_{i=1}^{m} (x_i * \omega_i) + b \tag{4.27}$$

Apart from the already mentioned applications of pattern extraction and trend detection, Artificial Neural Networks can also be employed to learn based on experience by providing the network a training dataset; Rearrange the network architecture based on new data during the learning step; And accept real-time problems by working on a stream line with more data for network training.

4.5.1 ANN structure

The arrangement of which the network nodes combine, the layered design, the number of neurons on each layer and its geometry, form the network structure or, as mentioned by some authors, network architecture.

4.5.2 Feed-forward network

The single-layer feed forward network (figure 4.4) is constitued by 2 layers: one for the input and another for the output. These layers are intra-connected to form a complete architecture, and, after the connections' weight assignment, it results in several results/outputs, one for each node.

A multi-layer feed-forward network (figure 4.5) is formed by several layers: one input layer that acts like a data buffer, one output layer that generates the output of the network and one or more hidden layers that have no association with the external atmosphere and are uniquely controlled by the connections weighting.

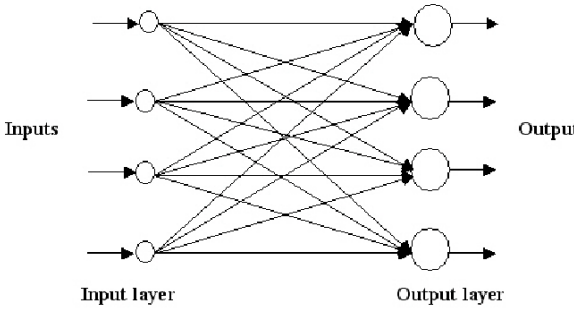

Figure 4.4: Single-layer network

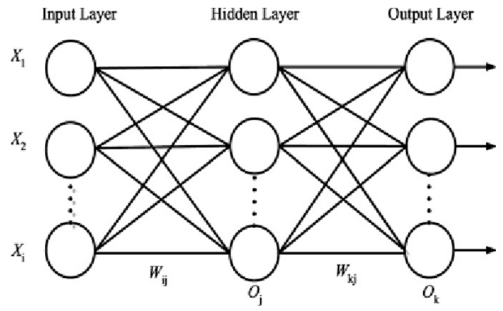

Figure 4.5: Multi-layer network

4.5.3 Activation Function

The activation function acts as an evaluation tool to provide the output of a specific neuron. It is an integration function that combines information either from network processing elements or by external responses, producing a controllable and delimited output. Some of the most popular activation functions comprise:

- Binary output:

$$\varphi(x) = \begin{cases} 1 & \text{if } x \geq \alpha \\ 0 & \text{if } x < \alpha \end{cases} \tag{4.28}$$

  where φ represents the threshold and the output is in binary value (0 or 1).

- Sigmoidal output:

$$sig(x) = \frac{1}{1+exp(-\beta x)} \tag{4.29}$$

  The activation function Sigmoid is highly implemented in artificial backpropagation networks as it relates, easily, the derivatives with the values of the nodes. It can be applied both as a binary function (range 0 to 1) or as a bipolar function (range from -1 to 1).

4.5.4 Training step

The training step for the artificial network can be achieved by updating the weights of the nodes' interconnections to accomplish a particular objective. This network adjustment can retain previous

training information and improve the network accuracy by changing the weights in an iterative way. The learning process can be categorised as supervised learning which training is targeted against the aspired output, the unsupervised learning which algorithm try to classify the data in clusters or group by identifying patterns and reinforcement learning which learning process is similar to the supervised learning with the proviso that the feedback is a reinforcement signal instead of the desired output. Only supervised learning will be covered as it was the method implemented in this dissertation.

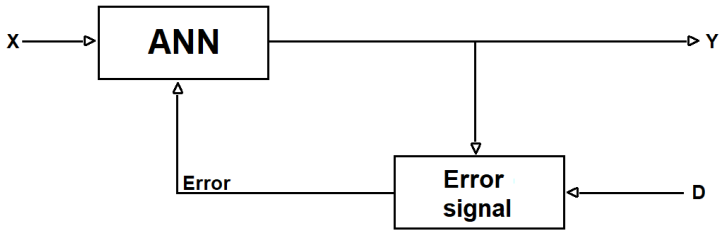

Figure 4.6: ANN-Supervised Learning

where X represents the input, Y the actual output, D the desired output and the Error is the result of the subtraction $D - Y$.

The learning process consists of comparing iteratively for the training points the actual output with the desired output. If the difference D-Y exists, it is produced an error signal that corrects the network weights for that point in matter. This kind of teaching seeks a classification error reduction where it is reckoned that the target values for the output are given to the input node.

4.5.5 Back propagation

The back-propagation process is correlated with multilayer artificial networks - back propagation network (BPN), where a training pair of input and output is used to update the connection weights. The associated error is backpropagated to the network's hidden layers during this process, and the weights are estimated throughout the learning process. This back-propagation process can be enumerated in the following steps:

1. Feed-forward the training sample input
2. Error Back-propagation
3. Weights adjustments

4.5.6 Code architecture

The neural network can be employed in many applications since classification to regression problems, following a similar code architecture shown in Figure 4.8. Initially, weights are set a random number in the range of -1 to 1 and the biases with small random numbers. After providing

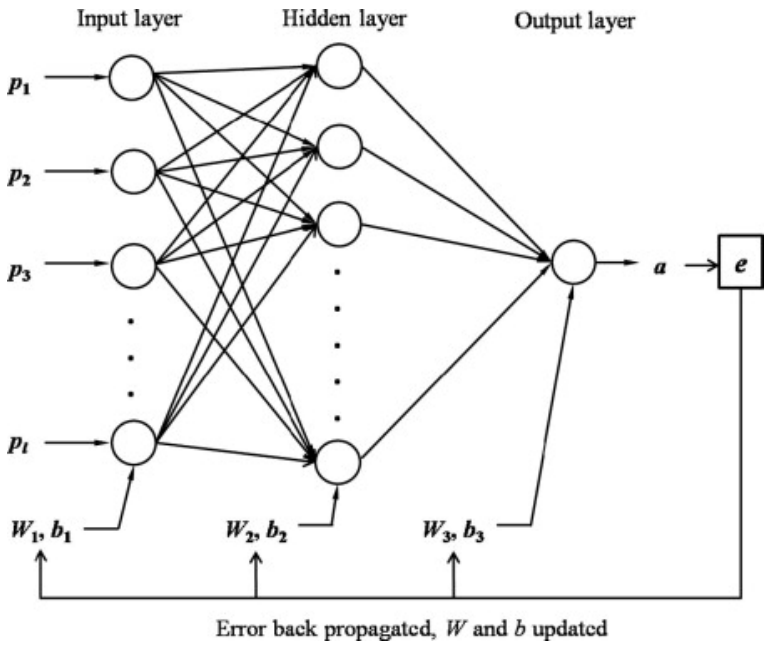

Figure 4.7: Back-propagation network (BPN). Extracted from Chen et. al. [4].

the training pair, the input passes along the network from the output of precedent layers to the input of succedent layers. The net input for one node results from the sum of the corresponding inputs multiplied by its associated weight. One of the activation functions mentioned previously is chosen to assign the output of the node. The error between the target value and the resulting one results in a back-propagation sequence that adjusts the weights and biases on precedent layers.

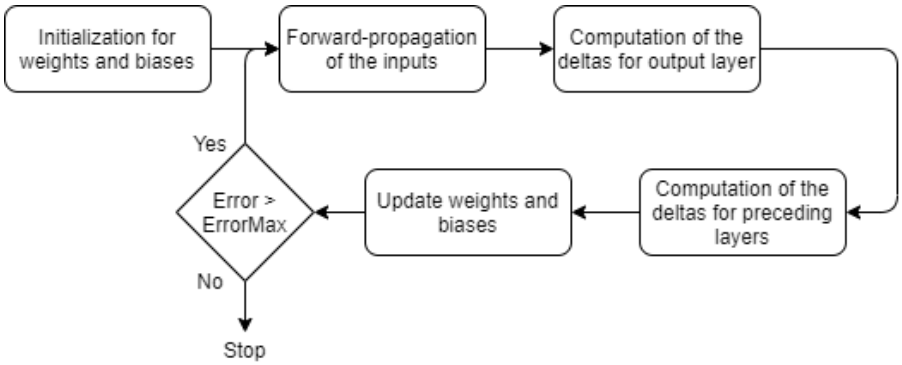

Figure 4.8: Back-propagation network algorithm

4.6 LIBS Spectral Corrections

Here, a new method is disclosed to deal with Laser-Induced Breakdown Spectroscopy (LIBS) uncertainties on geological samples. The sample matter is subject to interference and matrix effects during the signal acquisition, impeding a correct classification and quantification on state-of-the-art techniques like Chemometrics, Calibration-Free or Deep Learning methods. These pixel-based

approaches, along with black-box strategies, are proven that just a mere mathematical modelling is not enough to unscramble the plasma interactions neither mimic some of the well-known matrix effects like broadening, self-absorption/reversing, bremsstrahlung, black body or scattering. This present study seeks to introduce a paradigm shift by providing a method that considers the matrix effects and interferences to provide a probabilistic identification and quantification through a Covariance Mode procedure simultaneously identified with their unique eigenvectors. This method links the interferences between constituents to the eigenvectors, determining the relevant LIBS wavelengths with the attribution of a positive or negative weight to the interference in question, allowing a more straightforward LIBS spectral analysis. In this research, this method will be applied to a lithium ore for demonstration purposes, although this technique can be transposed to other spectra with its relevant adaptations. This procedure significantly improves LIBS systems as it contributes to the miniaturization of the equipment, allowing in situ, low-cost and low-resolution analysis.

4.6.1 Interference framework

Except for pure materials with simple emission spectral lines and backgrounds, LIBS readings will always be conditioned by interference and matrix effects, resulting in poorly optimized classification and quantification results when applying state-of-art methods. Real case scenarios confront some of the main characteristics that, up to date, have little to no scientific response: LIBS finite optical resolution and the complex multi-position and convoluted signals. These phenomena raise multi-scaled and multi-variable interference on the spectral information that enhances the hurdle on quantifications for any mainstream method.

Some of these state-of-art methods operate under very particular conditions and try to minimize the plasma effects, relying on pressure and laser parameters manipulation or atmospheric modification [66, 67].

Performing signal to noise ratio (SNR) corrections on Latent Thermodynamic Equilibrium (LTE) along with pixel-based approaches do not work on complex geological or biological LIBS samples, nor environments where having uncertainty is not an option (medical and pharmaceutical) and which quantification procedure is probabilistic based.

On Chemometrics methods, the latent variables data projection approach fails on non-ideal, non-pure, and complex samples (p.e. low compositional variability on pharmacology), on which the modelling is forced into a linear relationship like Partial Least Squares Regression (PLSR). These methods cannot shape the dynamical nature of LIBS like gas temperature, pressure, free-electron number density and signal convolutions.

4.6.2 Spectral complexity and inner effects

The analysed data is not different from the theoretic one, and it is characterised by its complexity and the progressive modification of the constituents inner-relations along the ore depth. This complexity is both shown by the laser spectra and the mass spectrometer (ICP-MS). In the mid

samples, around fifty meters deep, most components have several interfering lines, making the process highly complex for precise quantifications. As it is possible to see in images 4.9 to 4.12, these interferences are mandatory to be resolved as better as possible, so it is conceivable to have a clearer outlook for the elements that matter - in this case, the Lithium lines.

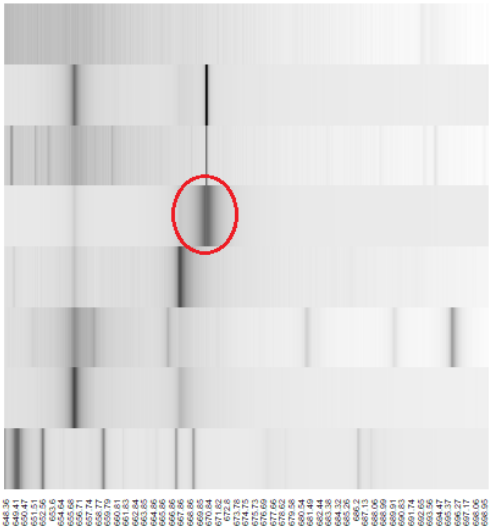

Figure 4.9: Interference band on the 670 *nm* Lithium line

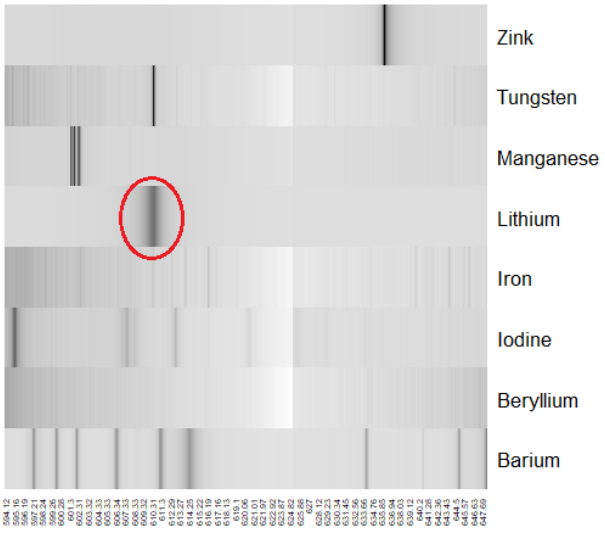

Figure 4.10: Interference band on the 610 *nm* Lithium line

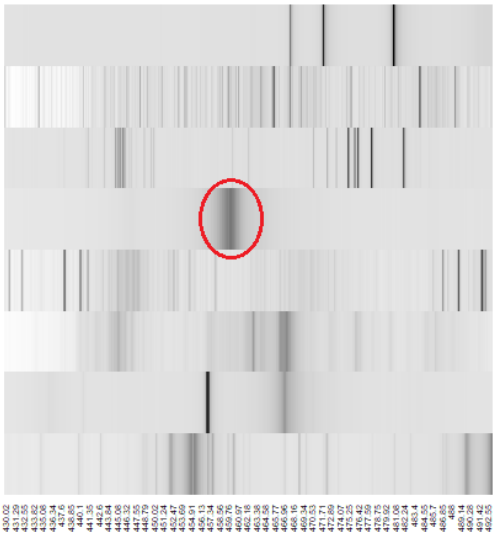

Figure 4.11: Interference band on the 460 *nm* Lithium line

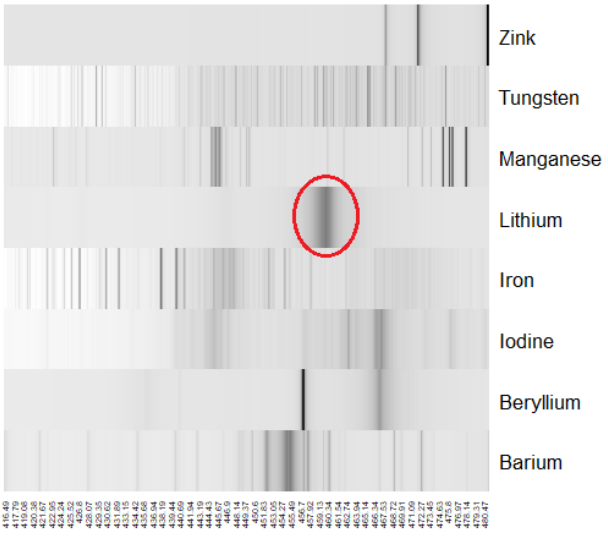

Figure 4.12: Interference band on the 413 *nm* Lithium line

For this purpose, on the four relevant Lithium lines - 670 *nm* (figure 4.9), 610 *nm* (figure 4.10), 460 *nm* (figure 4.11) and 413 *nm* (figure 4.12), it is clear that, in the same near spectral band, Lithium lines are contaminated with other elements lines - Beryllium, Iodine, Iron, Zink, Tungsten, Manganese and Barium (but applicable to the rest of the elements, even the ones that are not quantifiable by the equipment), which deteriorates any chance of quantifying Lithium just by using peak extraction techniques.

Direct peak extraction techniques also fail as the lines are not obtained on infinite resolution equipment, so Stark Broadening and Dopler effects are present and are signalled on a red circle, where it is possible to see the enlarging of the Lithium lines. Other quantification attempts without further processing also become highly ineffective as often spectra are acquired convoluted. One of the solutions that might be plausible to apply and will be further discussed are the application of a Fourier Transformation to re-construct the spectra in its covariance modes, where no longer it is necessary to directly looking for Lithium lines to quantify the underlying element, as it can do so by probabilistic relations with other elements.

4.6.3 Interference handling and Covariance Modes application

As for now, there are few explainable physics answers, namely via Saha-Boltzmann applications which do not take into account the matrix effects that occur on the plasma. An experimental approach is disclosed under the patent number WO2020/026165, where it is suggested the use of matrix effects (which are natural effects and hardware-based weaknesses that will always be associated with the equipment in use) to quantify one or more constituent.

This method extracts the spectral information of the sample by obtaining the spectral information resolution on one or more spectra firstly; extracts information where the spectrum is sub-optical; and schemes the spectral lines into a deterministic feature space with multiple dimension vector spaces.

This information can be saved on a constituents database, obtained by sub-optical extraction resolution, and be used to quantify the desired elements, even without accessing the spectral lines directly.

4.6.4 Notable metrics

The performance obtained on Lithium quantifications is determined and quantifiable by the techniques benchmarking and using the Mean Average Percentage Error (MAPE), similar to a Mean Distance from laboratory value in percentage terms and relative to each sample, and the Pearson Correlation (R) which gives the measure of linearity between laboratory and forecasted values.

$$MAPE(\%) = \frac{1}{N} * \sum \frac{|laboratory - prediction|}{|laboratory|} * 100 \tag{4.30}$$

where laboratory is assumed to be the reference values obtained via ICP-MS and the prediction the forecasted values for the different methods.

$$R = \frac{cov(x,y)}{S_x * S_y} \tag{4.31}$$

where S_x and S_y are the standard deviation for x and y, respectively.

Chapter 5

Results and discussion

5.1 Data-mining

5.1.1 Principal Components projections

As it is possible to see on the following figures 5.1 to 5.4, PCA can give some preliminary answers about the data and help the process of later applying machine learning techniques.

In figure 5.1, which represents the data projection on the first two principal components, it is possible to notice the most significant principal component (loading) of the dataset, Lithium (Li), as it is the direction where there is most variability of the data. Orthogonally, the second principal component, Manganese (Mn), explains the second largest aggregation of variance. From the direct analysis of the figure, it is immediately possible to tell that samples on the hundreds like 95, 104, 109, 110, 111 are the locations (meters depth) where it is more Lithium, oppositely to samples 15, 34, 35, 99 and so that are in the negative direction.

In figure 5.2, which represents the data projection on the second and third principal components, it is possible to notice, alongside Manganese (PC2) direction, samples 93, 94, 95 aggregate as well as, on the Barium direction (PC3), samples on the thirties like 29, 30, 31, 32, 33 also form a cluster.

In figure 5.3, that represents the component four and five, in other words, the element Barium (Ba) and Rubidium (Rb), there is a similarity between the first samples of the ore (the first meters) - samples 1 to 5.

As we go deeper, searching on smaller principal components, as seen in figure 5.4, it becomes harder to analyse the data as there is less explainability on those directions than for the first principal components.

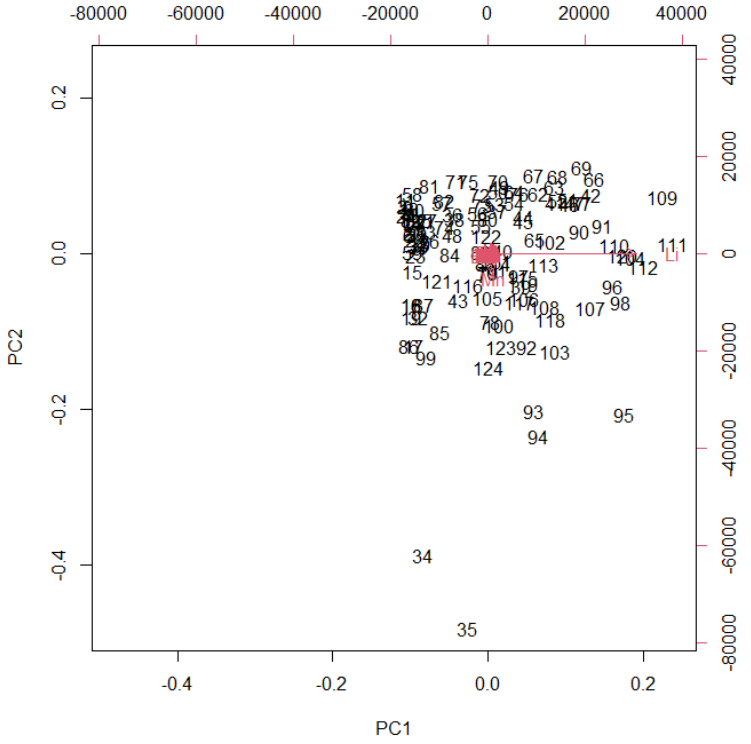

Figure 5.1: Plot of PC1 against PC2 and its loadings representation

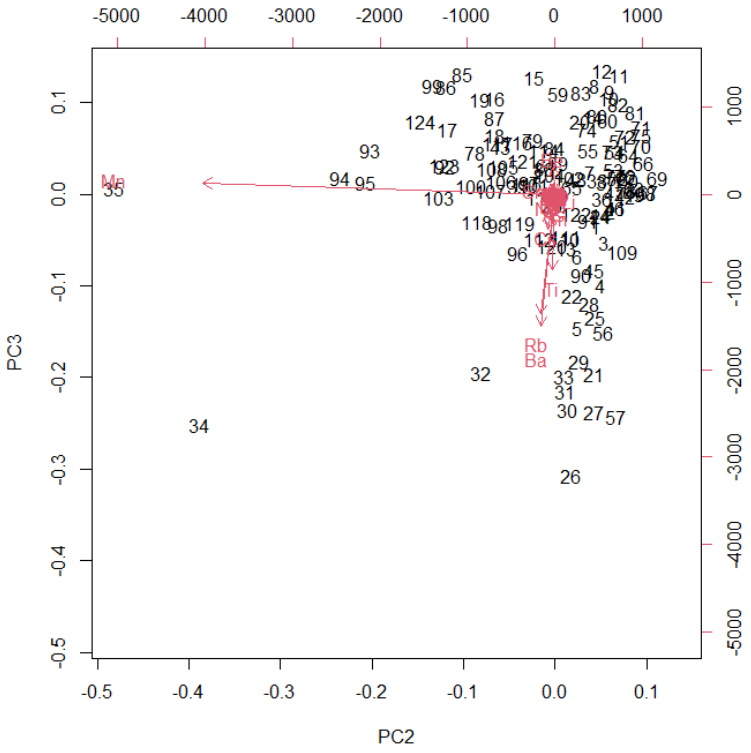

Figure 5.2: Plot of PC2 against PC3 and its loadings representation

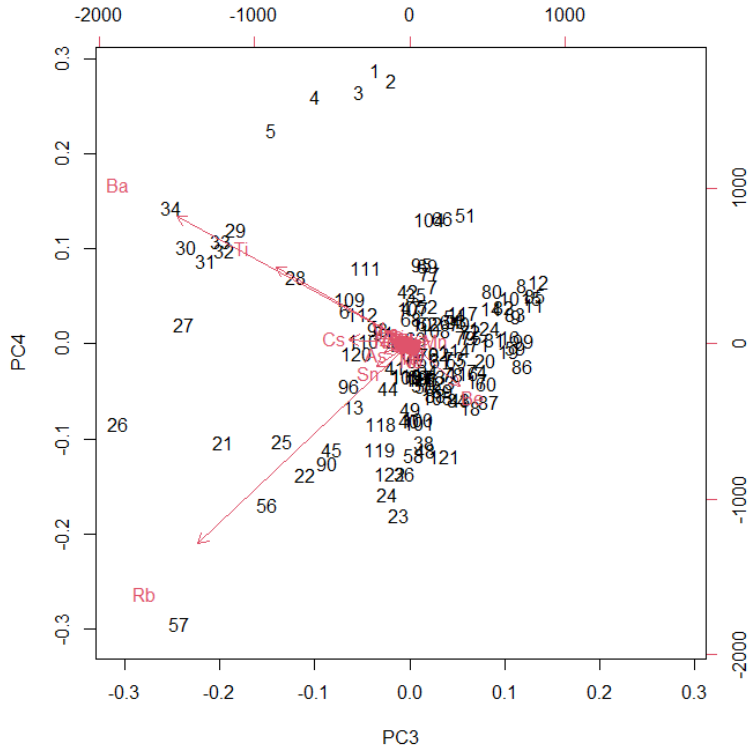

Figure 5.3: Plot of PC3 against PC4 and its loadings representation

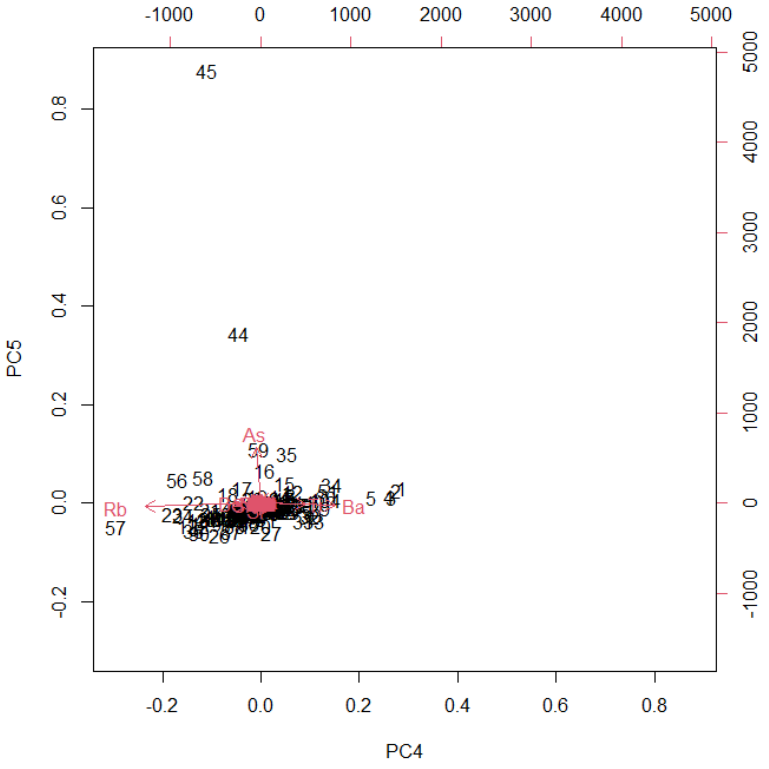

Figure 5.4: Plot of PC4 against PC5 and its loadings representation

5.1.2 Hierarchical clustering

More information can be extracted from the Principal Component Analysis, such as the cluster dendrogram, or the cluster tree (hierarchical clustering) on figure 5.5, that relates the variables one with each other, based on their levels of similarity or dissimilarity. The leafs of the tree that group are considered a cluster while the vertical lines represent higher separability.

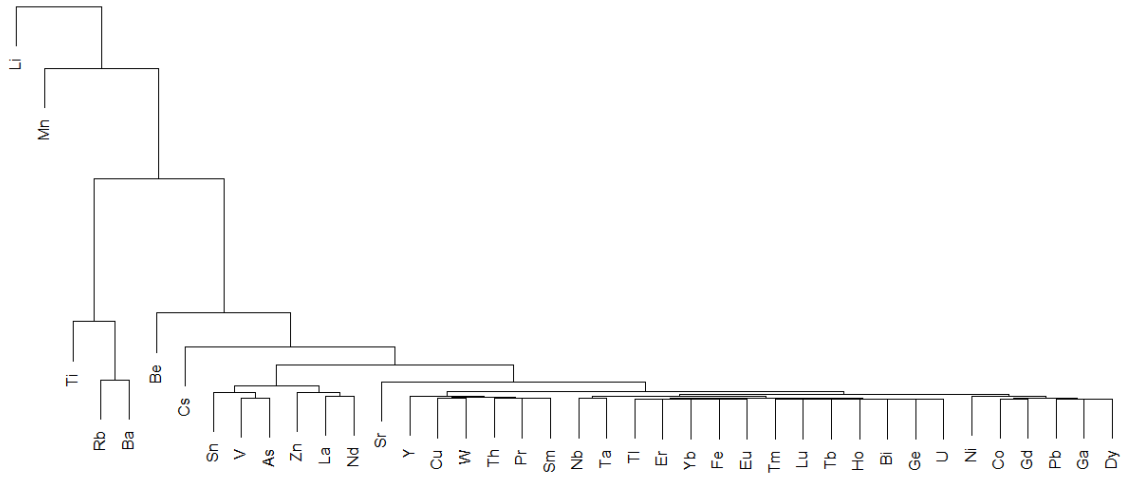

Figure 5.5: Cluster variables dendrogram

5.1.3 PCA plot

Principal Component Analysis can be plotted, on its directions, both with or without scalling (figures 5.7 and 5.6). When the PCA is not scaled, the projections are made accordingly to their variance aggregation, allowing the visualization of the top components of the data batch (components that aggregate the majority of the variance), oppositely to when it is scaled, the matrix is scaled to have a zero mean and unit variance and each component is divided by its variance and so, the plot shows the inner relations between the components.

On figure 5.7 it is possible to verify that the data is mostly spreaded under two components (Lithium and Manganese) that together aggregate 99,03% of the variance. On figure 5.6 it is possible to see a large data cluster that corresponds to the Pegmatite cluster, where Lithium is highly present, as it is spreaded on its direction.

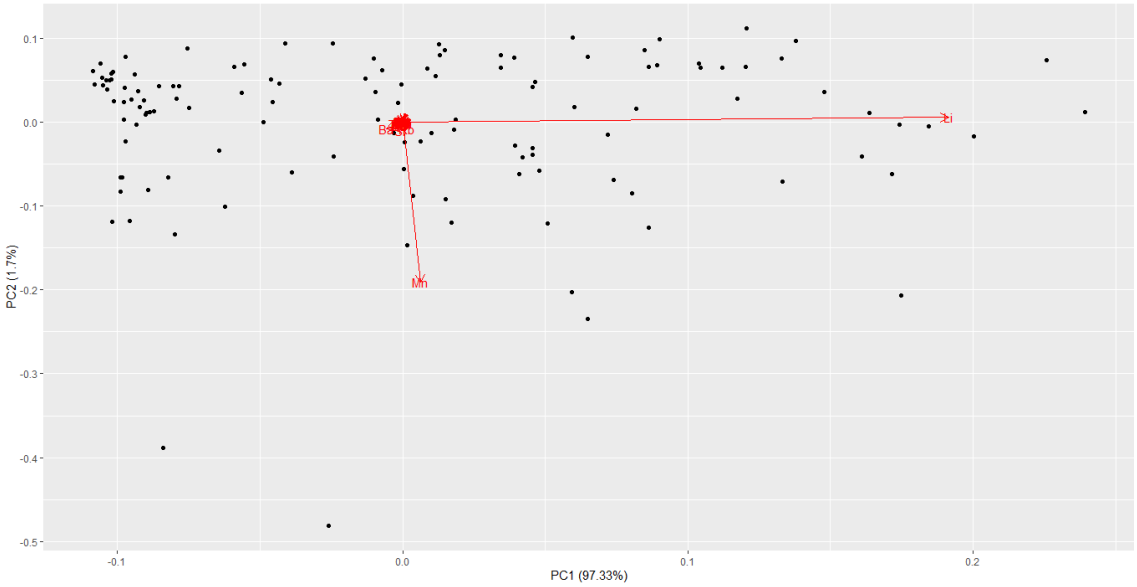

Figure 5.6: Principal Component Analysis not-scaled plot

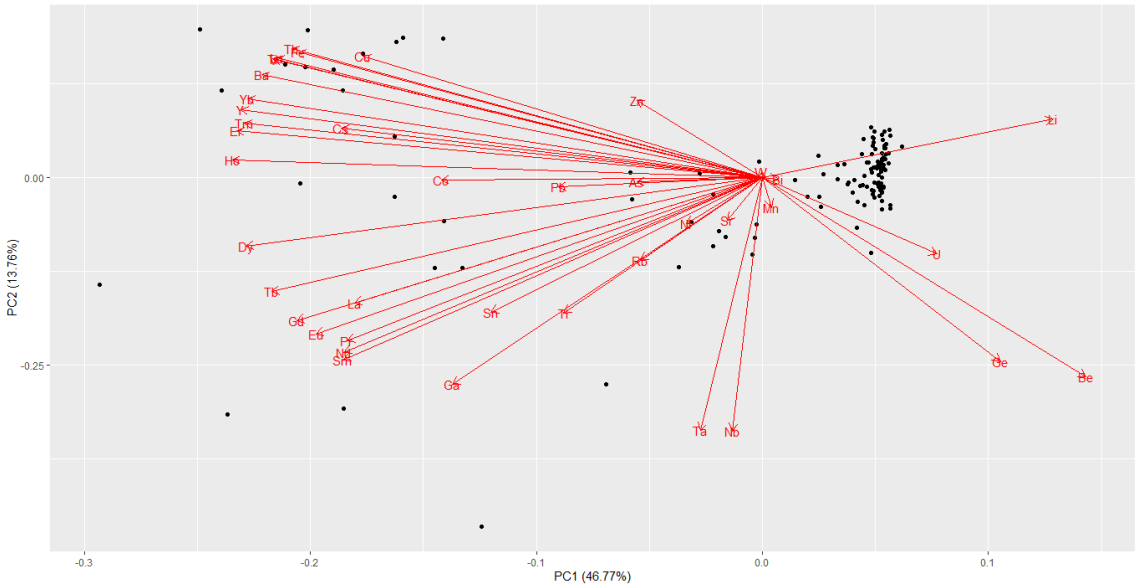

Figure 5.7: Principal Component Analysis scaled plot

5.1.4 3D PCA plot

The same study can be made in 3 dimensions where it is possible to span PCA over three components, as shown on figures 5.8 and 5.9. Here it is clear the separability between clusters, namely between the pegmatite cluster (FGP) and the others.

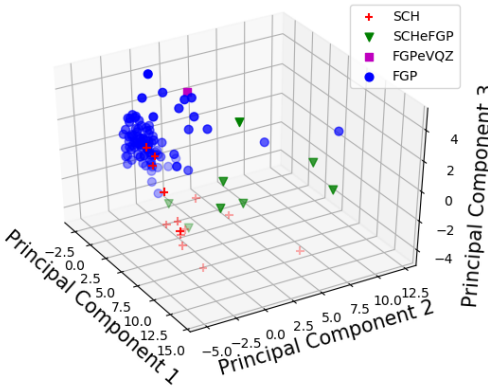

Figure 5.8: 3D Principal Component Analysis plot - perspective 1

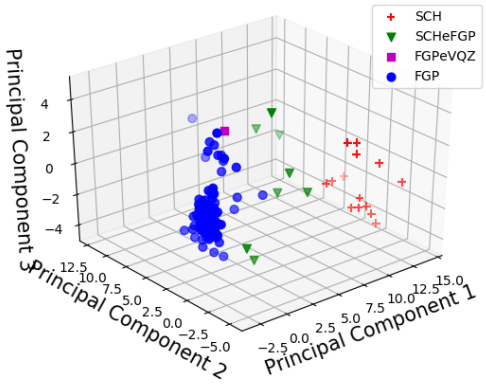

Figure 5.9: 3D Principal Component Analysis plot - perspective 2

It is possible to add that, by plotting the quantified Lithium against the top two principal components, as shown on figure 5.10, there are high concentrations of Lithium around PCA1 around -2 and PCA2 between -2 and 2, which verifies how accurate can be this preliminary method to identify the clusters that matter and advance to more precise quantification techniques to further understand the data in question.

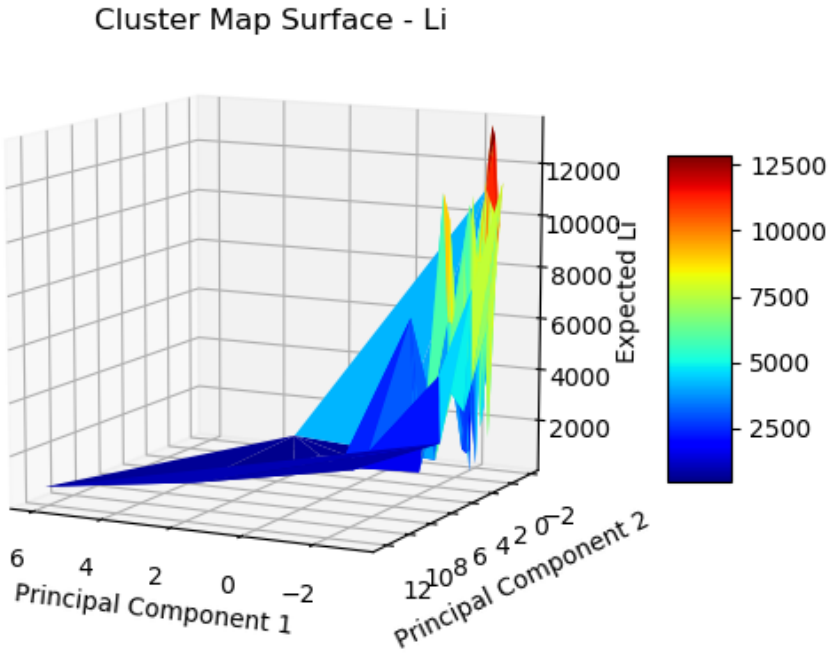

Figure 5.10: Quantified Lithium on a PCA representation

5.2 Classification - clustering

As it was mentioned, the dataset that was acquired from Barroso Lithium ore and afterwards analysed on LIBS and ICP-MS, has a pre-classification that was made based on geological support and will be considered as the reference value for LIBS, as if there was no prior knowledge there would not be the chance of testing the techniques for this data.

The explored clusterings were based on the 124m dataset, provided by Dr. Rui Martins and applied to the extracted PC1 and PC2 values of the Principal Component Analysis (figure 5.11).

The data is classified into four different clusters: Schist (SCH), Schist + Pegmatite (SCH+FGP), Pegmatite (FGP) and Pegmatite + Quartz (FGP+VQZ) and will be tested in several algorithms, since the most theoretical one (MeanShift) to the most nowadays applied (Support Vector Machine and ANN).

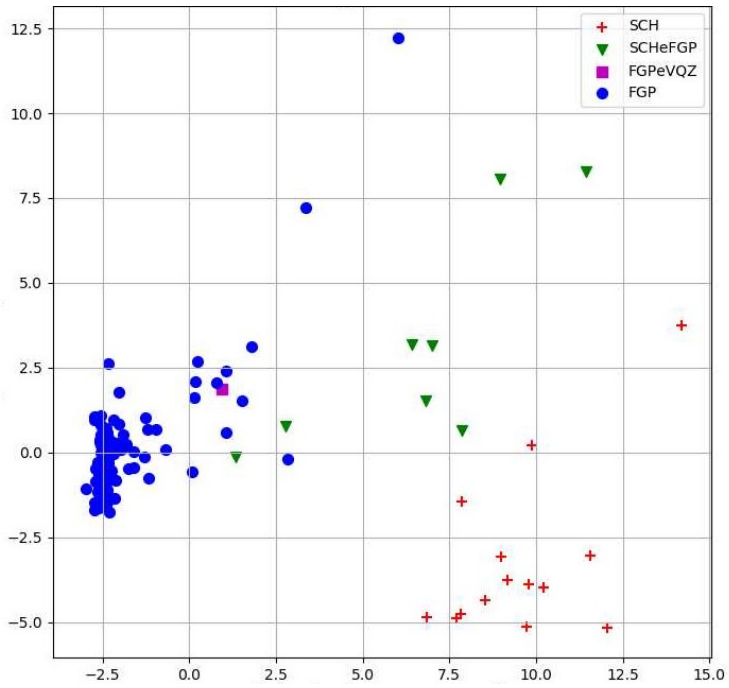

Figure 5.11: Geological classification initial scenario

5.2.1 IT-MeanShift

It was applied the probability density function (PDF) estimator using the Parzen window methodology. The PDFs are estimated correctly on the defined kernel bandwidth which is chosen between 0.4 and 1.6. Later the Information Theoretic MeanShift (IT-MS) is applied to find the modes of the distribution.

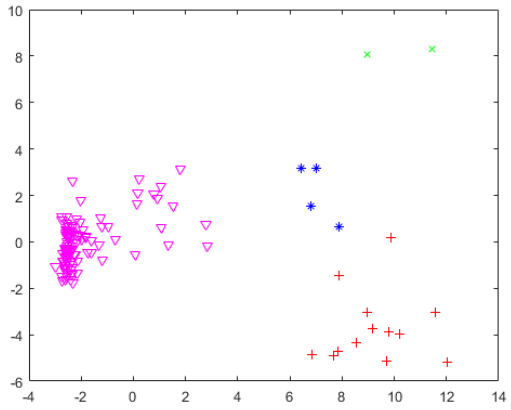

Figure 5.12: Clustering $\lambda = 0.4$, 4 clusters

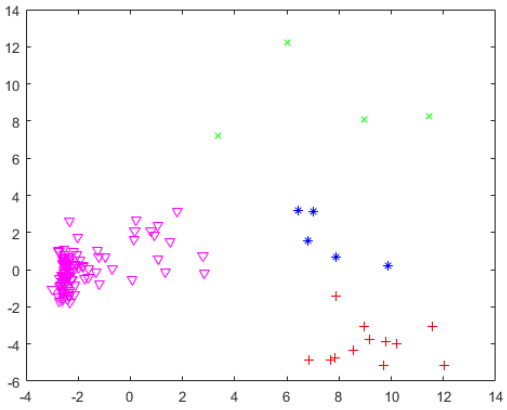

Figure 5.13: Clustering $\lambda = 0.8$, 4 clusters

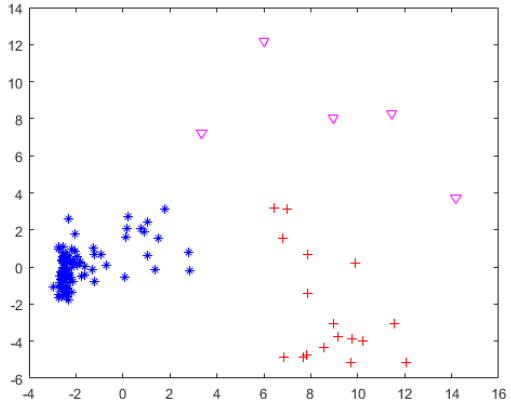

Figure 5.14: Clustering $\lambda = 1.2$, 3 clusters

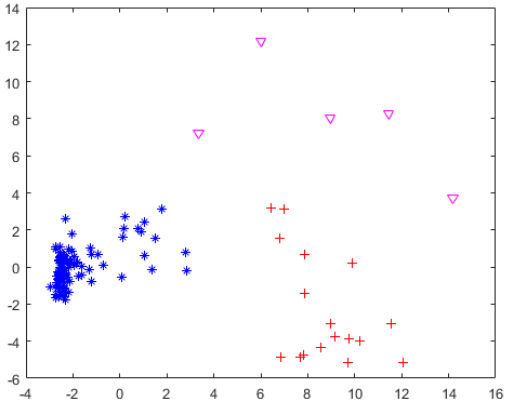

Figure 5.15: Clustering $\lambda = 1.6$, 3 clusters

For the clustering obtained, visible on figures 5.12 to 5.15, it is possible to see that the algorithm worked as intended achieving a very high accuracy on the given dataset, although the classification for some values is not valid. Commonly, for all values, it is not possible to distinguish the Schist + Pegmatite (SCH+FGP) from the Pegmatite (FGP) ones, as for a 2D-analysis it is not possible for the algorithm to spot the differences between the clusters in question.

Table 5.1: Accuracy for different kernel bandwidth λ values

	Valid	**Invalid**	**%**
$\lambda = 0.4$	118	6	95.2%
$\lambda = 0.8$	117	7	94.3%
$\lambda = 1.2$	115	9	92.7%
$\lambda = 1.6$	116	8	93.5%

5.2.2 Support Vector Machine (SVM)

As for the support vector machine, the same 124m dataset was considered on this method. It was explored both linear and quadratic support vectors in clustering identification. The data was splitted into 70/30 training and testing sets and afterwards applied in the SVM algorithm of python scikit library.

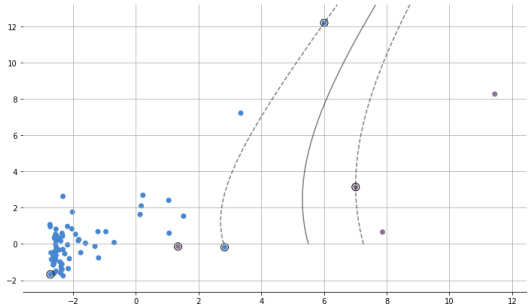

Figure 5.16: Clustering with quadratic SVM

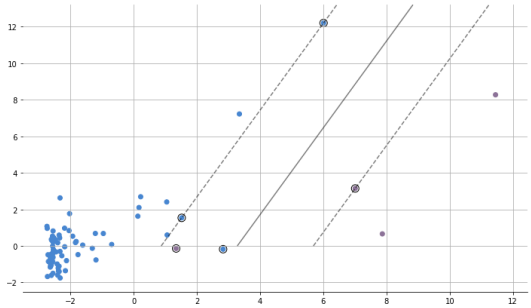

Figure 5.17: Clustering with linear SVM

As this algorithm achieves a direct classification (0 or 1/A or B), this classification was accomplished by classifying each cluster against the others, on a pair-wise manner. Given the present dataset and the existing clusters. The highest classification accuracy was achieved on the Pegmatite (FGP) cluster against Schist (SCH) with a result of 96,7% for the two linear and quadratic support vectors. That corresponds to 120 valid classifications and 4 miss-classifications.

Table 5.2: Accuracy on Support Vector Machine

	Valid	**Invalid**	**%**
Quadratic SVM	120	4	96,7%
Linear SVM	120	4	96,7%

5.2.3 Artificial Neural Network (ANN)

The same study on the same dataset was made with the application of a neural network. It was built a two hidden-layer neural network using the sklearn python library and a 80/20 data split. Again, the classification was achieved on a similar way as for the support vectors, classifying each cluster against the others. In order to avoid bad learning iterations due to the lack of samples of a certain type, ANN method was applied several times with a pre-shuffling stage and leave-one-out validation, resulting in an overall classification between 98 - 100% which corresponds to 1 or 2 miss-classifications. This last 100% result was achieved between the Schist (SCH) and Pegmatite (FGP) clusters. That is due to the low dispersity and high separability on those particular clusters that make the learning process highly efficient.

Table 5.3: Accuracy on Artificial Neural Network clustering

	Valid	**Invalid**	**%**
SCH+FGP / FGP	122	2	98,4%
SCH+FGP / SCH	123	1	99,2%
SCH / FGP	124	0	100%

5.3 Quantification

The quantification of Lithium was made on the same 124m dataset with all the 39 variables (40 variable minus the Lithium variable) and with the application of a Random Forest regression algorithm. The dataset was separated into the target (Lithium) and features (other variables) in a 75/25 training/testing sets using the Python Sklearn library. The algorithm converges for a random state of 42, and its variables importance is shown in Figure 5.18 and the respective regression tree visible in Figure 5.19.

As shown on figure 5.18, the algorithm attributes high importance to Lead (Pb), Yttrium (Y) and Beryllium (Be) to the quantification of Lithium with a overall 39,95% of MAPE error. Despite the high explainability of the Random Forest algorithm (figure 5.19 where it is possible to see all the algorithm decisions and results convergence), this algorithm lacks to understand how the different variables co-exist one with each other and the geological principals behind the support of Lithium existance.

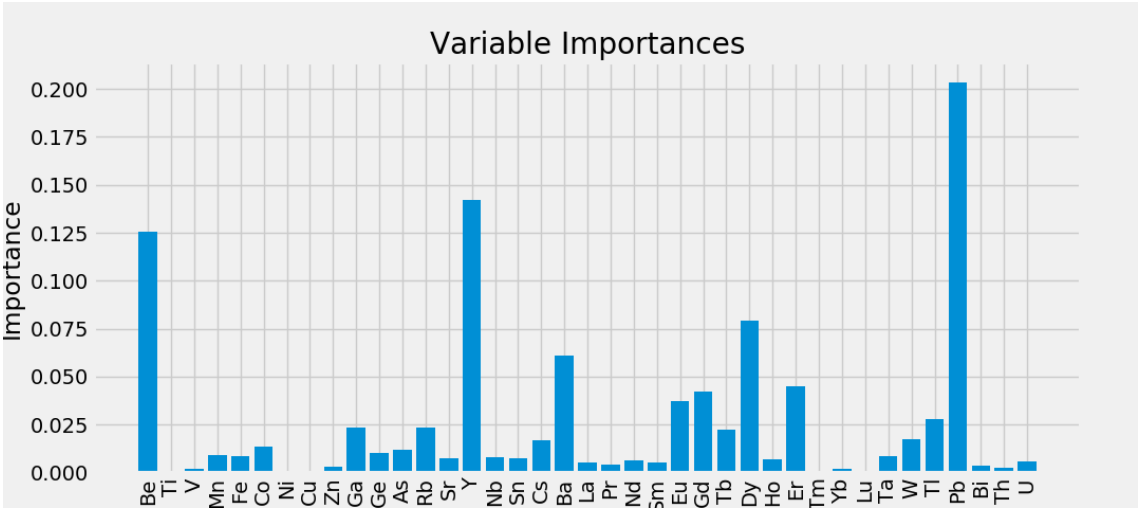

Figure 5.18: Random Forest variables importance

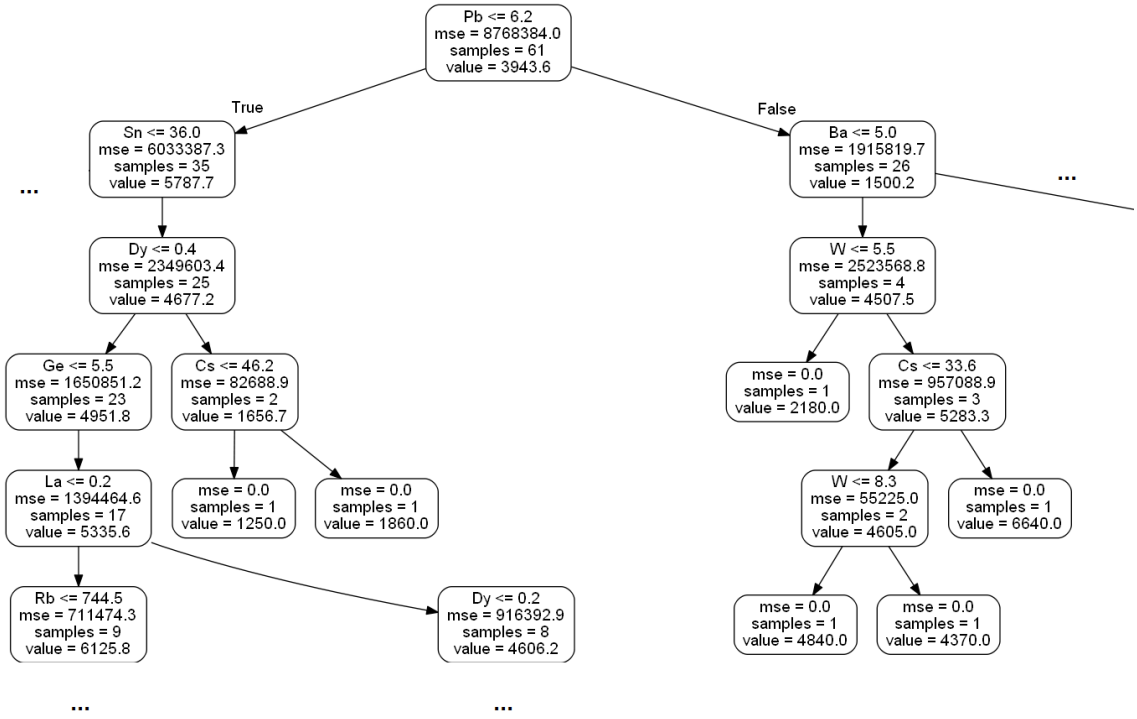

Figure 5.19: Short representation of the Random Forest tree - fragment (full version on appendix A)

5.4 Unscrambling interference

5.4.1 Scattering

In order to unscramble the LIBS interference, empirical scattering corrections were made on the two datasets provided for the two known Lithium lines from the NIST database. The process started by choosing several points (meters of depth) that describe well a wide variety of Lithium quantities, since the points that have little to no Lithium to the prominent pegmatite veins where it is abundant. Next, LIBS readings for those particular points were extracted and kept apart for comparison. Following, a CAP-based scattering correction technique was applied where the points are corrected according to the median reference value and shifted with a least-squares solver.

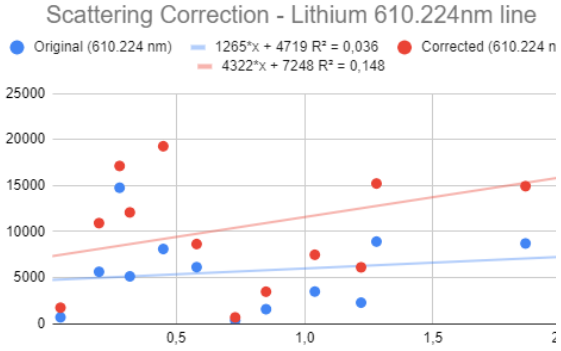

Figure 5.20: Scattering correction on the 76 *m* dataset - 610,22 *nm* Lithium line

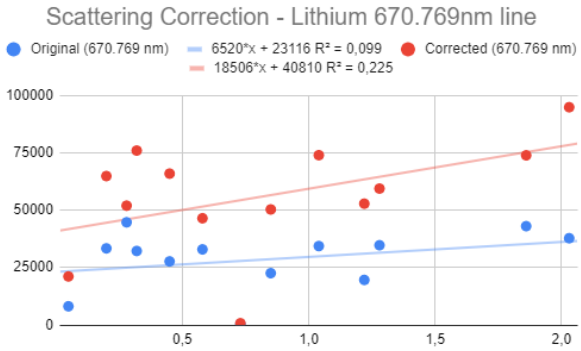

Figure 5.21: Scattering correction on the 76 *m* dataset - 670,77 *nm* Lithium line

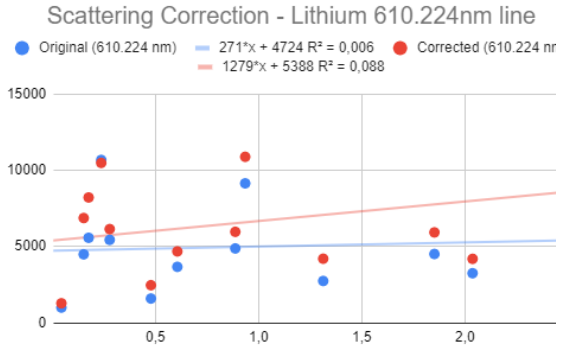

Figure 5.22: Scattering correction on the 124 *m* dataset - 610,22 *nm* Lithium line

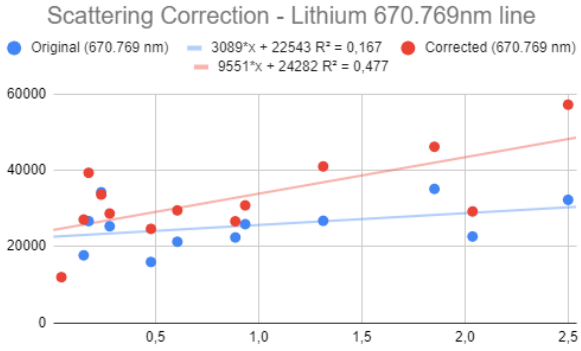

Figure 5.23: Scattering correction on the 124 *m* dataset - 670,77 *nm* Lithium line

It is possible to see that, although the empirical approach, the results on the Lithium regression are notable, passing on the 76m dataset from a 0,036 of R^2 to 0.148 on the first Lithium line and from 0,099 of R^2 to 0,225 on the second line. The results for the 124m are similar with approximate increases of R^2 value, which denotes the better predictability on Lithium after the scattering corrections all made.

The results, despite promising, only prove the effectiveness of the method. Far from over, as it is possible to see on tables 5.4 to 5.7, the Mean Absolute Percentage Error (MAPE) is very high, as

it needed to implement other LIBS signal corrections in order to output a valid and more accurate Lithium prediction.

Table 5.4: Scattering correction on the 76m dataset for the 610,22 *nm* Lithium line

i (*m*)	**Li**	**Original (610.224** *nm*)	**Corrected (610.224** *nm*)	**Li pred**	**MAPE (%)**
1	0,05	680	1725,389	7464,1	332,60
13	0,2	5625	10912,32	8112,4	25,66
21	0,28	14754	17121,43	8458,16	50,60
26	0,32	5129	12075,58	8631,04	28,52
38	0,58	6129	8639,316	9754,76	12,91
39	1,28	8905	15214,32	12780,16	16,00
42	2,03	9444	23693,51	16021,66	32,38
47	1,86	8703	14926,25	15286,92	2,42
50	1,04	3477	7473,323	11742,88	57,13
57	0,45	8103	19264,52	9192,9	52,28
64	1,22	2272	6110,529	12520,84	104,91
72	0,85	1560	3461,811	10921,7	215,49
75	0,73	338	670,0043	10403,06	1452,69

with an average MAPE of 183,35%, R of 0,384 and R^2 of 0,148.

Table 5.5: Scattering correction on the 76m dataset for the 670,77 *nm* Lithium line

i (*m*)	**Li**	**Original (670.769** *nm*)	**Corrected (670.769** *nm*)	**Li pred**	**MAPE (%)**
1	0,05	8145	21139,52	41735,3	97,43
13	0,2	33362	64822,57	44511,2	31,33
21	0,28	44706	51952,59	45991,68	11,47
26	0,32	32226	75927,28	46731,92	38,45
38	0,58	32922	46460,3	51543,48	10,94
39	1,28	34720	59382,01	64497,68	8,61
42	2,03	37775	94788,78	78377,18	17,31
47	1,86	43022	73859,9	75231,16	1,86
50	1,04	34363	73920,59	60056,24	18,76
57	0,45	27692	65900,26	49137,7	25,44
64	1,22	19617	52815,34	63387,32	20,02
72	0,85	22576	50299,65	56540,1	12,41
75	0,73	383	787,2637	54319,38	6799,77

with an average MAPE of 545,68%, R of 0,474 and R^2 of 0,225.

Table 5.6: Scattering correction on the 124m dataset for the 610,22 *nm* Lithium line

i (*m*)	**Li**	**Original (610.224** *nm*)	**Corrected (610.224** *nm*)	**Li pred**	**MAPE (%)**
1	0,044	996,25	1278,48	5444,276	325,84
13	0,176	5569	8211,1	5613,104	31,64
21	0,237	10669,5	10475,12	5691,123	45,67
22	0,152	4485,5	6864,7	5582,408	18,68
26	0,278	5433,25	6139,71	5743,562	6,45
35	0,606	3666,25	4677,03	6163,074	31,77
36	0,478	1589,5	2462,11	5999,362	143,67
37	0,887	4870,5	5960,33	6522,473	9,43
40	0,935	9137	10875,05	6583,865	39,46
91	1,852	4507,75	5922,89	7756,708	30,96
95	2,037	3247,5	4188,54	7993,323	90,84
111	2,5	7609,75	13501,14	8585,5	36,41
113	1,313	2739,25	4198,41	7067,327	68,33

with an average MAPE of 67,63%, R of 0,297 and R^2 of 0,088.

Table 5.7: Scattering correction on the 124m dataset for the 670,77 *nm* Lithium line

i (*m*)	**Li**	**Original (670.769** *nm*)	**Corrected (670.769** *nm*)	**Li pred**	**MAPE (%)**
1	0,044	9291.5	11958,81	24702,244	106,56
13	0,176	26644,25	39311,31	25962,976	33,96
21	0,237	34198,25	33600,52	26545,587	21,00
22	0,152	17681,25	27054,32	25733,752	4,88
26	0,278	25307,5	28635,39	26937,178	5,93
35	0,606	21245,5	29460,56	30069,906	2,07
36	0,478	15933,75	24619,39	28847,378	17,17
37	0,887	22369,25	26601,72	32753,737	23,13
40	0,935	25847	30768,88	33212,185	7,94
91	1,852	35094,5	46128,75	41970,452	9,01
95	2,037	22611,5	29174,68	43737,387	49,92
111	2,5	32225,75	57162,74	48159,5	15,75
113	1,313	26729,75	40973,65	36822,463	10,13

with an average MAPE of 23,65%, R of 0,690 and R^2 of 0,477.

5.5 Benchmarking Self-Learning Artificial Intelligence (SL-AI) against state-of-art methods

Self-Learning AI, the method to systematically extract co-variances between elements by optimizing the correct features and transforms them onto a feature space, was applied to two datasets: Avantes high-resolution signal and SciApps portable LIBS. On the configurable parameters: maximum components (directions) and number of clusters, the algorithm was set to a maximum value

of 2 components and 15 clusters. The vector space comprising the spectral lines will be, at the most, 2-dimensional.

To prove the algorithm's reliability, several case studies were made, namely for quantifications of other elements rather than Lithium and tested against the state-of-art methods like Similarity, Partial Least Squares and Artificial Neural Network. Some other elements have more or less interfering lines or their spectra distributed over a wide range of wavelengths so, a subset of random constituents will also be tested - Lithium (Li), Beryllium (Be), Barium (Ba), Manganese (Mn), Iron (Fe), Rubidium (Rb), Caesium (Cs) and Zinc (Zn).

5.5.1 SL-AI - Case study on multiple constituents

As it is possible to verify on figures 5.24 to 5.31, the co-variance modes for the several elements are detected and the predictions for those feature spaces successful. The best regression model is obtained for Rubidium as it is the element which can be better identified due to its dispersion on all the spectral band.

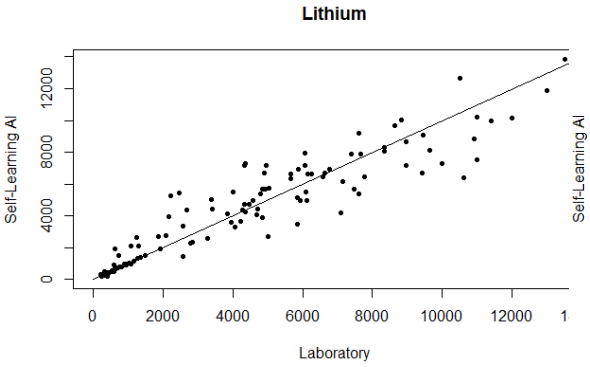

Figure 5.24: SL-AI on Lithium

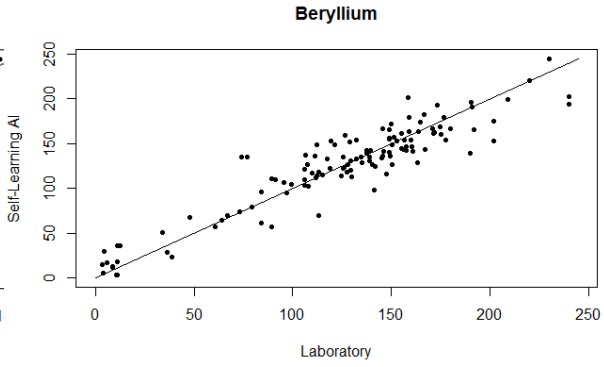

Figure 5.25: SL-AI on Beryllium

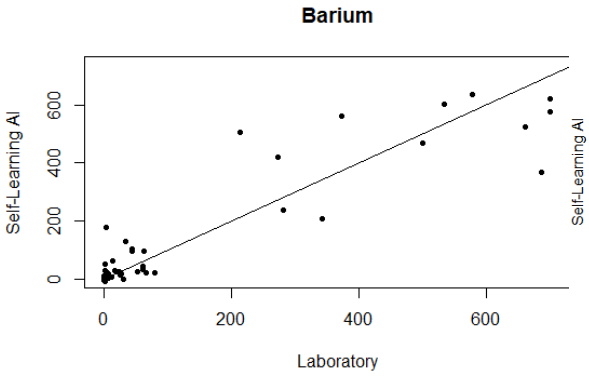

Figure 5.26: SL-AI on Barium

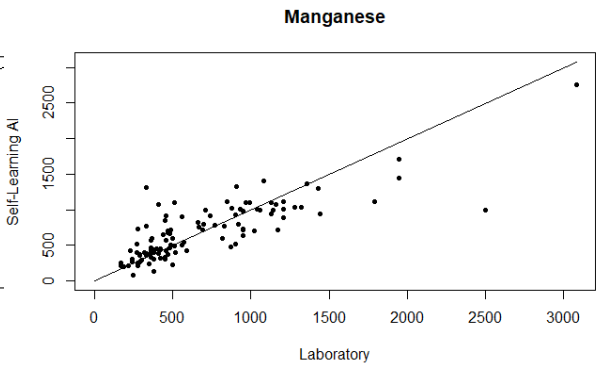

Figure 5.27: SL-AI on Manganese

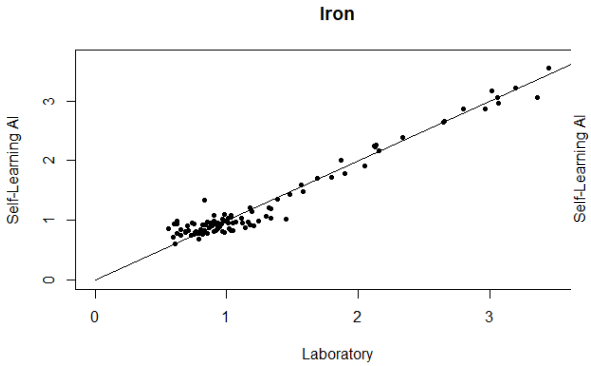

Figure 5.28: SL-AI on Iron

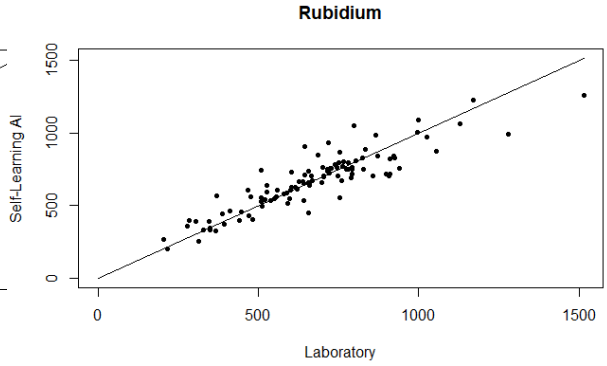

Figure 5.29: SL-AI on Rubidium

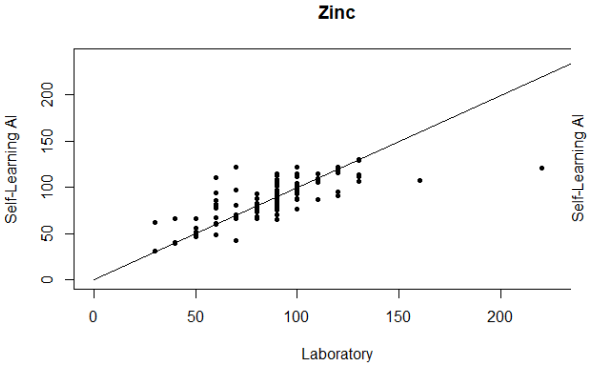

Figure 5.30: SL-AI on Zinc

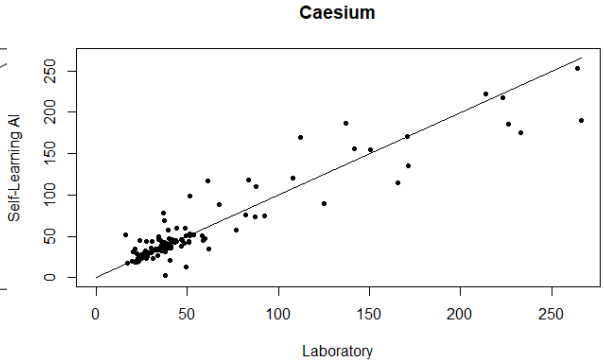

Figure 5.31: SL-AI on Caesium

The same study was applied to the SciApps portable LIBS system where it was only tested the Lithium element, showing a final $R = 0,865$ and $MAPE = 51,781\%$ for two latent variables.

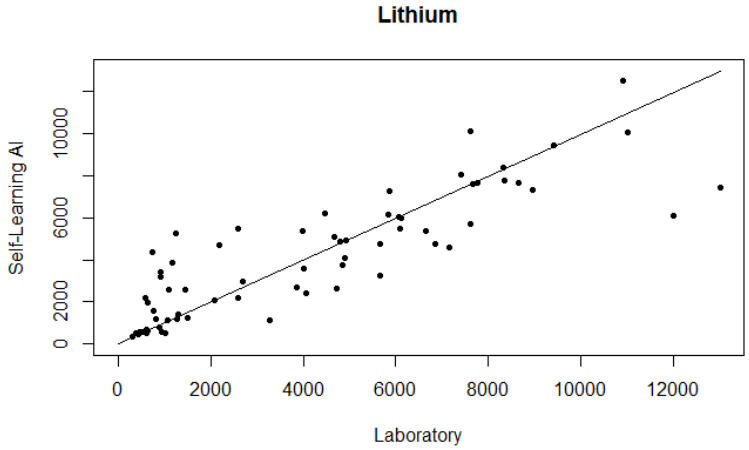

Figure 5.32: SL-AI on Lithium - Portable LIBS

5.5.2 SL-AI against state-of-art methods - multiple constituents

In table 5.8, it is exhibited a summary of all methods chosen to benchmark against the proposed Self-Learning technique in this dissertation: Similarity, Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Partial Least Squares (PLS). This study was applied to the Avantes high-resolution spectrometer dataset on the selected mentioned variables and using the metrics: Pearson Correlation (R), Number of latent variables (LV), and Mean Average Percentage Error (MAPE) for comparison.

Table 5.8: Benchmarking of SL-AI against state-of-art methods - multiple constituents

Method	Metric	Li	Be	Ba	Mn	Fe	Rb	Zn	Cs
Similarity k=3	**R**	0,622	NA	NA	0,552	NA	0,468	0,959	NA
	LV	-	-	-	-	-	-	-	-
	MAPE (%)	182,76	NA	NA	75,08	NA	70,13	95,91	NA
ANN	**R**	0,863	0,061	0,913	NA	0,875	0,896	0,122	0,899
	LV	-	-	-	-	-	-	-	-
	MAPE (%)	53,19	19,33	39,43	NA	42,64	12,64	30,47	60,54
CNN	**R**	0,663	0,761	0,628	0,129	0,457	0,431	0,706	0,654
	LV	-	-	-	-	-	-	-	-
	MAPE (%)	45,16	29,09	43,25	52,58	55,64	36,88	41,92	52,07
PLS	**R**	0,831	0,829	0,869	0,714	0,883	0,696	0,383	0,829
	LV	5	5	3	5	5	5	3	5
	MAPE (%)	81,29	52,2	55,3	38,3	23,6	20,8	27,18	42,49
SL-AI	**R**	0,919	0,953	0,932	0,813	0,979	0,898	0,809	0,973
	LV	3	2	1	2	2	2	2	2
	MAPE (%)	23,16	15,91	25,53	31,77	11,34	10,85	13,74	16,91

The first method is a similarity regression, tested with the application of the nearest neighbours technique, where the algorithm was tuned for three k-neighbours. There is no convergence of the algorithm for most constituents as it cannot find any similarities in terms of eigenstructure; those are represented with a Not Available (Na) sign. For the elements that the algorithm does find a proper similarity for clustering, which are the Lithium, Manganese, Rubidium and Zinc elements, the regression fit is shallow, and the MAPE considerable high. Despite Rubidium having a moderate correlation, it still is the element for which it is possible to have a quantification with 70,13% of MAPE.

The Artificial Neural Network was chosen as it is one of the widely adopted methods on state-of-art for spectral processing. It was optimized for two hidden layers using backpropagation as a regression algorithm with a tangent activation function. The results on the ANN are more satisfactory as it can map the spectral features on a piecewise sequence allowing it to interpret the nonlinearities of the dataset. The correlations vary between R$\sim$(0,86 to 0,89), yet some elements, such as Manganese, were not possible to model, as well as, Beryllium or Zinc, which despite the low MAPE$\sim$(19,33 and 12,64) present no correlation between the predicted values and the laboratory ones.

Convolutional Neural Network (CNN), contrarily to ANN, does not use neurons or weights; the process passes through several layers (math, rectified, and fully connected layer) that try to capture the patterns and provide a vector output. Like ANN, CNN relies on error measurements for learning improvement and epochs utilization for model effectiveness control. In MAPE terms, CNN manages to provide better value results than ANN with MAPE∼(29,09 to 52,07) but fails to deliver a fitting regression better than ANN or PLS with R∼(0,13 to 0,76).

Partial Least Squares, in most cases, outperforms the ANN results presenting a slight improvement in correlation terms for most of the elements by using between 3 and 5 latent variables. The MAPE values are moderate, making this algorithm well qualified for being one of the most used Chemometrics methods for spectra calibration.

The top results are comprised among all elements when using the Self-Learning Artificial Intelligence (SL-AI) method, which is the algorithm that better understand the LIBS matrix effects and spectral interference to deal with the geological variability. SL-AI presents better MAPE values when compared to any other method on every element tested with just 1 to 3 latent variables. The correlations are high, varying between R∼(0,81 to 0,98), which are outstanding values for the research field - geological exploration.

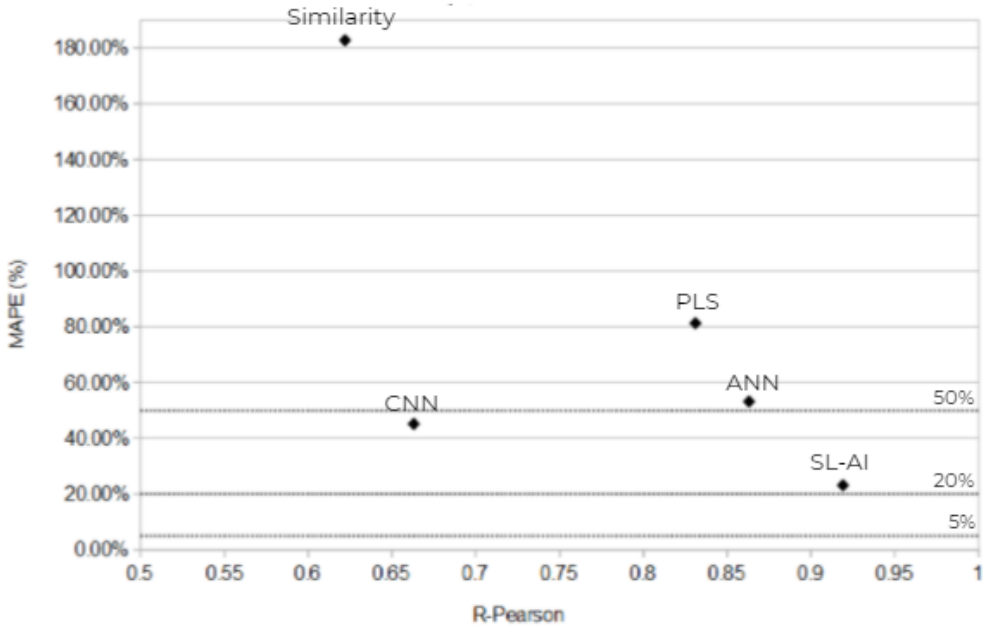

Figure 5.33: Benchmarking of SL-AI against state-of-art methods - Lithium

Chapter 6

Conclusions and future work

In Conclusion, LIBS prove to be an excellent tool for geological exploration due to its versatility on several elements and practicability on low-cost and low-preparation analysis. This technique still demonstrated the possibility of creating geological maps to study and measure samples directly from the powder extracted.

The first part of this dissertation approached several state-of-art methods and discriminated against their characteristics, applications, and limitations. Chemometrics, namely via Principal Component Analysis (PCA) and Partial Least Squares (PLS), are the mainstream methods employed on spectral calibrations to maximize the covariance between the extracted spectra and the sample composition. Artificial Neural Networks are widely used in machine learning problems and have also gained traction on spectral processing for years now. Although these methods can, in certain circumstances, output reasonable predictions with several optimizations, they still lack the most critical part: the explainability or interpretability of the results. The high complexity of LIBS spectra and the subject for interference and matrix effects creates a problem among the unscrambling of the numerous plasma interactions that pixel-based methods cannot mimic. These so-called black box approaches rely on big-data databases and mathematical modelling that can only output a probabilistic result that is doubtful and unacceptable in critical areas like pharmaceutical or medical research.

Assuming this problem on spectra interpretation, the second part of this dissertation rises as a novel solution, Self-Learning Artificial Intelligence (SL-AI), recently patented by INESC TEC - Center for Applied Photonics that shifts the paradigm on LIBS quantifications by providing a method that considers the matrix effects and interferences to provide an almost deterministic identification and quantification through an iterative Covariance Mode seeking that identifies the constituents according to their unique eigenvectors. The algorithm extracts information where the spectrum is sub-optical and schemes the spectral lines into a multiple dimension deterministic feature space. These feature spaces can be stored on a database and be used to quantify elements even without accessing the spectral lines directly, as long as the patterns for those elements are present. This is a significant advancement for in-site analysis as it can provide an accurate quantification even with low-resolution LIBS field equipment and output the results within seconds with low

computation effort.

The results obtained by benchmarking SL-AI against state-of-art methods are flawless, showing very high Pearson Correlation values R∼(0,81 to 0,98) and low Mean Average Percentage Error (MAPE) at MAPE∼(10,85 to 31,77%) for most of the studied constituents, which proves the high sensibility of this algorithm for geological variability that still did not have an answer with monolithic approaches like Partial Least Squares and Artificial Neural Networks.

There are still some aspects that can be improved to increase the performance of the LIBS systems and some other analysis that can be made and are comprised in a list for further research:

- Acquisition of more sample data: The algorithm works accurately on the used datasets, yet, with more data, it would be capable of constructing better the covariance modes for each element, as it would have more information to create each feature space by relating the lines in study with the other elements' ones.

- Code improvement: Some aspects in the SL-AI algorithm can be improved, such as eliminating the manual selection of the number of clusters as forcing this parameter with no further interference corrections can lead the algorithm to learn and map the feature space with noisy data. On the same note, an optimal number of latent variables can also be smartly chosen without assigning that parameter.

- Spectral Corrections: More spectral corrections can be employed in order to ease the data for the SL-AI application. Besides the implemented scattering baseline corrections, other corrections for plasma temperature and electron density with inverse Saha-Boltzmann equations and corrections for the Doppler effect (lines enlargement) can also be employed.

- Spectral Corrections Bypass: Once the corrections are applied, and the feature maps are created, a database can be designed to store the corrections necessary for each element, so via pattern recognition, it is possible to speed run the quantifications on field equipment by providing a direct match between the extracted spectra and the most probable necessary spectral correction. This step further fully decentralizes LIBS from the cutting-edge high-resolution equipment to produce a low-cost, low-resolution, on-field, immediate quantification analysis.

References

[1] Jozef Rakosky. Qualitative-portable to quantitative-laboratory LIBS. Sep 2012.

[2] Rui Martins, Pedro Jorge, Eduardo Silva, Almeida, and Alfredo Martins. WO2020/026165 A1. Feb 2020.

[3] Convolutional Neural Network - javatpoint, Feb 2021. [Online; accessed 2. Feb. 2021]. URL: `https://www.javatpoint.com/pytorch-convolutional-neural-network`.

[4] Guangying Chen, Kaiyun Fu, Zhiwu Liang, Teerawat Sema, Chen Li, Paitoon Tontiwachwuthikul, and Raphael Idem. The genetic algorithm based back propagation neural network for MMP prediction in CO2-EOR process. *Fuel*, 126:202–212, Jun 2014.

[5] Elisabetta Tognoni, Gabriele Cristoforetti, S Legnaioli, V Palleschi, Antonio Salvetti, Maike Müller, Ulrich Panne, and Igor Gornushkin. A numerical study of expected accuracy and precision in calibration-free laser-induced breakdown spectroscopy in the assumption of ideal analytical plasma. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 62(12):1287–1302, 2007.

[6] Division of Environmental Physics - Viktor Martisovits, Feb 2021.

[7] N. Arnold, J. Gruber, and J. Heitz. Spherical expansion of the vapor plume into ambient gas: An analytical model. *Appl. Phys. A*, 69(7):S87–S93, Dec 1999.

[8] G. Cristoforetti, A. De Giacomo, M. Dell'Aglio, S. Legnaioli, E. Tognoni, V. Palleschi, and N. Omenetto. Local Thermodynamic Equilibrium in Laser-Induced Breakdown Spectroscopy: Beyond the McWhirter criterion. *Spectrochim. Acta, Part B*, 65(1):86–95, Jan 2010.

[9] VN Lednev and SM Pershin. Plasma stoichiometry correction method in laser-induced breakdown spectroscopy. *Laser Physics*, 18(7):850–854, 2008.

[10] Kathleen K Herrera, Elisabetta Tognoni, Igor B Gornushkin, Nicoló Omenetto, Benjamin W Smith, and JD Winefordner. Comparative study of two standard-free approaches in laser-induced breakdown spectroscopy as applied to the quantitative analysis of aluminum alloy standards under vacuum conditions. *Journal of Analytical Atomic Spectrometry*, 24(4):426–438, 2009.

[11] Lanxiang Sun and Haibin Yu. Correction of self-absorption effect in calibration-free laser-induced breakdown spectroscopy by an internal reference method. *Talanta*, 79(2):388–395, 2009.

[12] Kathleen K Herrera, Elisabetta Tognoni, Nicoló Omenetto, Benjamin W Smith, and James D Winefordner. Semi-quantitative analysis of metal alloys, brass and soil samples by calibration-free laser-induced breakdown spectroscopy: recent results and considerations. *Journal of Analytical Atomic Spectrometry*, 24(4):413–425, 2009.

[13] D Bulajic, M Corsi, G Cristoforetti, S Legnaioli, V Palleschi, A Salvetti, and E Tognoni. A procedure for correcting self-absorption in calibration free-laser induced breakdown spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 57(2):339–353, 2002.

[14] VS Burakov, VV Kiris, PA Naumenkov, and SN Raikov. Calibration-free laser spectral analysis of glasses and copper alloys. *Journal of Applied Spectroscopy*, 71(5):740–746, 2004.

[15] Francesco Colao, Roberta Fantoni, Violeta Lazic, Luisa Caneve, and V. E. Fermi. LIBS as a diagnostic tool during the laser cleaning of copper based alloys: Experimental results. *J. Anal. At. Spectrom*, 19(4), Apr 2004.

[16] L. Fornarini, F. Colao, R. Fantoni, V. Lazic, and V. Spizzicchino. Calibration analysis of bronze samples by nanosecond laser induced breakdown spectroscopy: A theoretical and experimental approach. *Spectrochim. Acta, Part B*, 60(7):1186–1201, Aug 2005.

[17] S. M. Pershin, F. Colao, and V. Spizzichino. Quantitative analysis of bronze samples by laser-induced breakdown spectroscopy (LIBS): A new approach, model, and experiment. *Laser Phys.*, 16(3):455–467, Mar 2006.

[18] A. De Giacomo, M. Dell'Aglio, O. De Pascale, R. Gaudiuso, and G. P. Parisi. ns- and fs-LIBS of copper-based-alloys: A different approach. *Applied Surface Science - APPL SURF SCI*, 253(19):7677–7681, Jul 2007.

[19] J. A. Aguilera, C. Aragón, G. Cristoforetti, and E. Tognoni. Application of calibration-free laser-induced breakdown spectroscopy to radially resolved spectra from a copper-based alloy laser-induced plasma. *Spectrochim. Acta, Part B*, 64(7):685–689, Jul 2009.

[20] V. S. Burakov, V. V. Kiris, P. A. Naumenkov, and S. N. Raikov. Calibration-Free Laser Spectral Analysis of Glasses and Copper Alloys. *J. Appl. Spectrosc.*, 71(5):740–746, Sep 2004.

[21] F. Colao, R. Fantoni, V. Lazic, A. Paolini, and A. Baliva. Investigation of LIBS feasibility for in situ planetary exploration: An analysis on Martian rock analogues. *Planet. Space Sci.*, 52(1):117, Mar 2004.

[22] B. Sallé, J.-L. Lacour, P. Mauchien, P. Fichet, and G. Manhès. Comparative study of different methodologies for quantitative rock analysis by Laser-Induced Breakdown Spectroscopy in a simulated Martian atmosphere. *Spectrochim. Acta, Part B*, 61(3):301–313, Mar 2006.

[23] Li Wang, Chijian Zhang, and Yuan Feng. Controlled calibration method for laser induced breakdown spectroscopy. *Chin. Opt. Lett., COL*, 6(1):5–8, Jan 2008.

[24] A. De Giacomo, M. Dell'Aglio, O. De Pascale, S. Longo, and M. Capitelli. Laser induced breakdown spectroscopy on meteorites. *Spectrochim. Acta, Part B*, 62(12):1606–1611, Dec 2007.

[25] Jean-Michel Mermet. Quality of calibration in inductively coupled plasma atomic emission spectrometry. *Spectrochim. Acta, Part B*, 49(12):1313–1324, Oct 1994.

[26] Jean-Michel Mermet. Limit of quantitation in atomic spectrometry: An unambiguous concept? *Spectrochim. Acta, Part B*, 63(2):166–182, Feb 2008.

[27] A. Ciucci, M. Corsi, V. Palleschi, S. Rastelli, A. Salvetti, and E. Tognoni. New Procedure for Quantitative Elemental Analysis by Laser-Induced Plasma Spectroscopy. *Appl. Spectrosc.*, 53(8):960–964, Aug 1999.

[28] E. Tognoni, G. Cristoforetti, S. Legnaloli, V. Palleschi, and I. Gomushkin. A numerical study of expected accuracy and precision in Calibration-Free Laser-Induced Breakdown Spectroscopy in the assumption of ideal analytical plasma. *SPECTROCHIMICA ACTA PART B-ATOMIC SPECTROSCOPY*, 62(12):1287–1302, Dec 2007.

[29] Mike J. Adams and Neil W. Barnett. *Chemometrics in Analytical Spectroscopy*. The Royal Society of Chemistry, Feb 2004.

[30] Analytica Chimica Acta | Vol 391, Issue 2, Pages N5-N8, 101-252 (31 May 1999) | ScienceDirect.com by Elsevier, Jan 2021.

[31] David M. Haaland and Edward V. Thomas. Partial least-squares methods for spectral analyses. 2. Application to simulated and glass spectral data. *Anal. Chem.*, 60(11):1202–1208, Jun 1988.

[32] G. C. Lalor and C. Zhang. Multivariate outlier detection and remediation in geochemical databases. *Sci. Total Environ.*, 281(1-3):99–109, Dec 2001.

[33] M. Blanco and I. Villarroya. NIR spectroscopy: a rapid-response analytical tool. Trends Anal. Chem. 21, 240-250. *TrAC, Trends Anal. Chem.*, 21(4):240–250, Apr 2002.

[34] Herbert Fink, Ulrich Panne, and Reinhard Niessner. Process Analysis of Recycled Thermoplasts from Consumer Electronics by Laser-Induced Plasma Spectroscopy. *Anal. Chem.*, 74(17):4334–4342, Sep 2002.

[35] Madhavi Z. Martin, Nicole Labbé, Timothy G. Rials, and Stan D. Wullschleger. Analysis of preservative-treated wood by multivariate analysis of laser-induced breakdown spectroscopy spectra. *Spectrochim. Acta, Part B*, 60(7):1179–1185, Aug 2005.

[36] Ling Gao and Shouxin Ren. Simultaneous Multicomponent Spectrophotometric Kinetic Determination by a Hybrid Intelligent Computing Method. In *Artificial Intelligence and Computational Intelligence*, pages 348–355. Springer, Berlin, Germany, Sep 2011.

[37] J.-B. Sirven, B. Bousquet, L. Canioni, and L. Sarger. Laser-Induced Breakdown Spectroscopy of Composite Samples: Comparison of Advanced Chemometrics Methods. *Anal. Chem.*, 78(5):1462–9, Apr 2006.

[38] Christopher R. Dockery and Scott R. Goode. Laser-induced breakdown spectroscopy for the detection of gunshot residues on the hands of a shooter. *Appl. Opt.*, 42(30):6153–6158, Oct 2003.

[39] Ota Samek, Helmut H. Telle, and David C. S. Beddows. Laser-induced breakdown spectroscopy: a tool for real-time, in vitro and in vivo identification of carious teeth. *BMC Oral Health*, 1(1):1–9, Dec 2001.

[40] Morgan S. Schmidt and Amy J. Ray Bauer. Preliminary correlations of feature strength in spark-induced breakdown spectroscopy of bioaerosols with concentrations measured in laboratory analyses. *Appl. Opt.*, 49(13):C101–C109, May 2010.

[41] Frank C. De Lucia, Jr., Jennifer L. Gottfried, Chase A. Munson, and Andrzej W. Miziolek. Multivariate analysis of standoff laser-induced breakdown spectroscopy spectra for classification of explosive-containing residues. *Appl. Opt.*, 47(31):G112–G121, Nov 2008.

[42] Svante Wold, Henrik Antti, Fredrik Lindgren, and Jerker Öhman. Orthogonal signal correction of near-infrared spectra. *Chemom. Intell. Lab. Syst.*, 44(1):175–185, Dec 1998.

[43] R. Barbini, F. Colao, R. Fantoni, A. Palucci, S. Ribezzo, H. J. L. van der Steen, and M. Angelone. Semi-quantitative time resolved LIBS measurements. *Appl. Phys. B*, 65(1):101–107, Jul 1997.

[44] G. Galbács, I. B. Gornushkin, B. W. Smith, and J. D. Winefordner. Semi-quantitative analysis of binary alloys using laser-induced breakdown spectroscopy and a new calibration approach based on linear correlation. *Spectrochim. Acta, Part B*, 56(7):1159–1173, Jul 2001.

[45] P. A. Mosier-Boss, S. H. Lieberman, and G. A. Theriault. Field demonstrations of a direct push FO-LIBS metal sensor. *Environ. Sci. Technol.*, 36(18):3968–3976, Sep 2002. `arXiv:12269750`.

[46] Robert L. Green, Gert Thurau, Nathan C. Pixley, Arthur Mateos, Robert A. Reed, and John P. Higgins. In-Line Monitoring of Moisture Content in Fluid Bed Dryers Using Near-IR Spectroscopy with Consideration of Sampling Effects on Method Accuracy. *Anal. Chem.*, 77(14):4515–4522, Jul 2005.

[47] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *J. Sens.*, 2015(2):1–12, Jul 2015.

[48] Lu-Ning Li, Xiang-Feng Liu, Wei-Ming Xu, Jian-Yu Wang, and Rong Shu. A laser-induced breakdown spectroscopy multi-component quantitative analytical method based on a deep convolutional neural network. *Spectrochim. Acta, Part B*, 169:105850, May 2020.

[49] R. C. Wiens, S. Maurice, J. Lasue, O. Forni, R. B. Anderson, S. Clegg, S. Bender, D. Blaney, B. L. Barraclough, A. Cousin, L. Deflores, D. Delapp, M. D. Dyar, C. Fabre, O. Gasnault, N. Lanza, J. Mazoyer, N. Melikechi, P.-Y. Meslin, H. Newsom, A. Ollila, R. Perez, R. L. Tokar, and D. Vaniman. Pre-flight calibration and initial data processing for the ChemCam laser-induced breakdown spectroscopy instrument on the Mars Science Laboratory rover. *Spectrochim. Acta, Part B*, 82:1–27, Apr 2013.

[50] Madhavi Z Martin, Nicole Labbé, Timothy G Rials, and Stan D Wullschleger. Analysis of preservative-treated wood by multivariate analysis of laser-induced breakdown spectroscopy spectra. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 60(7-8):1179–1185, 2005.

[51] JM Tucker, MD Dyar, MW Schaefer, SM Clegg, and RC Wiens. Optimization of laser-induced breakdown spectroscopy for rapid geochemical analysis. *Chemical Geology*, 277(1-2):137–148, 2010.

[52] MD Dyar, ML Carmosino, EA Breves, MV Ozanne, SM Clegg, and RC Wiens. Comparison of partial least squares and lasso regression techniques as applied to laser-induced breakdown spectroscopy of geological samples. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 70:51–67, 2012.

[53] Yun Zhao, Mahamed Lamine Guindo, Xing Xu, Miao Sun, Jiyu Peng, Fei Liu, and Yong He. Deep Learning Associated with Laser-Induced Breakdown Spectroscopy (LIBS) for the Prediction of Lead in Soil. *Appl. Spectrosc.*, 73(5):565–573, Jan 2019.

[54] Ying Zhang, Ying Li, Wendong Li, Zigang Sun, and Yunfeng Bi. Classification of Geological Samples Based on Soft Independent Modeling of Class Analogy Using Laser-Induced Breakdown Spectroscopy. *J. Spectro.*, 2018, Jun 2018.

[55] A. Metzinger, D. J. Palásti, É. Kovács-Széles, T. Ajtai, Z. Bozóki, Z. Kónya, and G. Galbács. Qualitative Discrimination Analysis of Coals Based on Their Laser-Induced Breakdown Spectra. *Energy Fuels*, 30(12):10306–10313, Dec 2016.

[56] TG Barroso, L Ribeiro, H Gregório, F Santos, and RC Martins. Self-learning artificial intelligence for spectroscopy point-of-care for veterinary hemogram analysis:erytrocytes, hemoglobin and hematocrit. Biosensors, 2020.

[57] Rosalba Gaudiuso, Ebo Ewusi-Annan, Weiming Xia, and Noureddine Melikechi. Diagnosis of Alzheimer's disease using laser-induced breakdown spectroscopy and machine learning. *Spectrochim. Acta, Part B*, 171:105931, Sep 2020.

[58] Yanwu Chu, Tong Chen, Feng Chen, Yun Tang, Shisong Tang, Honglin Jin, Lianbo Guo, Yong feng Lu, and Xiaoyan Zeng. Discrimination of nasopharyngeal carcinoma serum using laser-induced breakdown spectroscopy combined with an extreme learning machine and random forest method. *J. Anal. At. Spectrom.*, 33(12):2083–2088, Nov 2018.

[59] Explainable Artificial Intelligence, Nov 2020. [Online; accessed 3. Feb. 2021]. URL: `https://www.darpa.mil/program/explainable-artificial-intelligence`.

[60] Erhan Gokcay and Jose C. Principe. Information Theoretic Clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(2):158–171, Feb 2002.

[61] Sudhir Rao, Allan de Medeiros Martins, and José C. Príncipe. Mean shift: An information theoretic perspective. *Pattern Recognit. Lett.*, 30(3):222–230, Feb 2009.

[62] R. Jenssen, K.E. Hild, D. Erdogmus, J.C. Principe, and T. Eltoft. Clustering using renyi's entropy. 1:523–528 vol.1, 2003.

[63] J. Sumaili, H. Keko, Vladimiro Miranda, and G. Chicco. On the use of information theoretic mean shift for electricity load patterns clustering. *2011 IEEE Trondheim PowerTech*, pages 1–6, 2011.

[64] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory*, 21(1):32–40, Jan 1975.

[65] Vapnik, V. (1998) Statistical Learning Theory. Wiley, New York. - References - Scientific Research Publishing, Jun 2021. [Online; accessed 3. Jun. 2021].

[66] K. Amal, S. H. Elnaby, V. Palleschi, A. Salvetti, and M. A. Harith. Comparison between single- and double-pulse LIBS at different air pressures on silicon target. *Appl. Phys. B*, 83(4):651–657, May 2006.

[67] Andrew J. Effenberger and Jill R. Scott. Effect of Atmospheric Conditions on LIBS Spectra. *Sensors*, 10(5):4907–4925, May 2010.