



UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

Data Fusion of Laser-Induced Breakdown Spectroscopy
and Spectral Reflectance Techniques for Estimating the
Mineralogical Composition of Copper Concentrates

POR DANNY ALBERTO LUARTE CANTO

Tesis presentada a la Dirección de Postgrado de la Universidad de Concepción para optar al grado académico de Doctor en Ciencias de la Ingeniería con mención en Ingeniería Eléctrica

Profesor Guía: PhD. Daniel Gerónimo Sbarbaro Hoffer
Profesor Co-guía: PhD. Jorge Carlos Yáñez Solorza

Octubre 2021
Concepción, Chile

Abstract

The pyrometallurgical copper industry faces some challenges in terms of the instrumentation for its processes. In this work, Laser-Induced Breakdown Spectroscopy (LIBS) data will be studied and combined with Diffuse Reflectance Spectroscopy (DRS) data and also with Hyperspectral Imaging (HSI) data to characterize the elemental and mineralogical composition in copper concentrates. This knowledge can be used to develop a sensor that replaces the current procedure used, which is risky, slow, and generates toxic waste and gaseous emissions.

LIBS spectra are used for elemental characterization of samples, whereas DRS spectra can be used for molecular or mineral determination. HSI sensors provide a wider range of data for the sample material. The information from these sources can be fused to obtain a more reliable characterization. These spectroscopic techniques are high dimensional in terms of features or wavelengths. In order to process these datasets, it is essential to reduce their dimensionality, which can be done by using variable selection techniques. In LIBS, the expert selection is frequently used since there are peaks that are known to be associated with certain elemental species. For DRS and HSI data, it is less direct how to choose some wavelengths. Thus some automatic variable selection algorithms can be applied for this task. In this work, two variable selection methods are proposed for LIBS data. Both methods combine the use of expert knowledge to select the best wavelengths.

Before fusing LIBS and HSI datasets, DRS is fused with LIBS data using a small dataset. LIBS and HSI data are finally fused using low-level and mid-level data fusion techniques.

For each regression analysis, artificial neural networks (ANN) were used, which have gained attention for regression studies due to the flexibility in dealing with large amounts of nonlinear correlated data. The results show that by using mid-level data fusion, it is possible to outperform the performance of the individual sources, with root mean squared errors of prediction reductions ranging from 4% to 70% in the case of LIBS-DRS data fusion, and from 1% to 74% in the case of LIBS-HSI data fusion.

Resumen

La industria pirometalúrgica del cobre se enfrenta a algunos desafíos en cuanto a la instrumentación de sus procesos. En este trabajo, los datos de Espectroscopía de Descomposición Inducida por Láser (LIBS) serán estudiados y combinados con los datos de Espectroscopía de Reflectancia Difusa (DRS) y también con los datos de Imágenes Hiperespectrales (HSI) para caracterizar la composición elemental y mineralógica en concentrados de cobre. Este conocimiento puede utilizarse para desarrollar un sensor que sustituya al procedimiento actual, que es arriesgado, lento y genera residuos tóxicos y emisiones gaseosas.

Los espectros LIBS se utilizan para la caracterización elemental de las muestras, mientras que los espectros DRS pueden utilizarse para la determinación molecular o mineral. Los sensores HSI proporcionan una gama más amplia de datos para el material de la muestra. La información de estas fuentes puede fusionarse para obtener una caracterización más fiable. Estas técnicas espectroscópicas son altamente dimensionales en términos de variables o longitudes de onda. Para procesar estos conjuntos de datos, es esencial reducir su dimensionalidad, lo que puede hacerse utilizando técnicas de selección de variables. En LIBS, la selección experta se utiliza con frecuencia, ya que hay intensidades que se sabe que están asociadas a determinadas especies elementales. En el caso de los datos DRS y HSI, es menos directo cómo elegir algunas longitudes de onda. Por lo que se pueden aplicar algunos algoritmos de selección automática de variables para esta tarea. En este trabajo se proponen dos métodos de selección de variables para datos LIBS. Ambos métodos combinan el uso de conocimiento experto para seleccionar las mejores longitudes de onda.

Antes de fusionar los conjuntos de datos LIBS y HSI, los datos DRS se fusionan con los datos LIBS utilizando un pequeño conjunto de datos. Los datos LIBS y HSI se fusionan finalmente utilizando técnicas de fusión de datos de nivel bajo y medio.

Para cada análisis de regresión se utilizaron redes neuronales artificiales (RNA), que han ganado atención en los estudios de regresión debido a la flexibilidad para tratar grandes cantidades de datos correlacionados de manera no lineal. Los resultados muestran que utilizando la fusión de datos de nivel medio es posible superar el rendimiento de las fuentes individuales, con reducciones de los errores cuadráticos medios de predicción que van del 4% al 70% en el caso de la fusión de datos LIBS-DRS, y del 1% al 74% en el caso de la fusión de datos LIBS-HSI.



In memory of Juan Luarte

Acknowledgments

I would like to start by thanking the financial support of Conicyt, which funded my PhD. studies from 2017 to 2021.

My most heartfelt thanks to Professor Daniel Sbarbaro, who once again trusted me and for all his guidance throughout these years. I would also like to thank Professor Jorge Yáñez for letting me be part of LABTRES with incredible people such as Claudio, Eimmy, Jonnathan, Marizú, Martín. Special thanks to Yerko, who went with me to the Chemometrics Winter School in Porto Alegre, where we had the chance to learn a lot about chemometrics and the Brazilian culture as well.

I would like to express my sincere gratitude to Ashwin, whose help and guidance were fundamental to fulfill the requirements of my PhD. studies.

Many thanks to the people from CEFOP, Benjamín, José, Pablo, and Rodrigo, with whom I spent the early stage of my PhD. studies. Also, thanks to Pamela Campos for her outstanding project management skills.

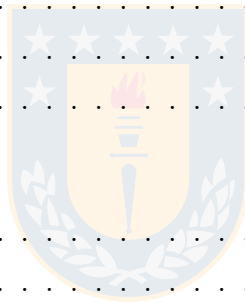
Special thanks to Mr. Patricio Orellana, who has always supported me since my undergraduate studies.

To my friends from my undergraduate studies, Fabián, Guillermo, Israel, Marco, and Nicolás. Also, to my friends from my eSports team, NIUPI, especially to Alejandro, Darío, Germán and Miguel.

Lastly, thanks to my parents and sister for their unconditional support throughout my life. This dissertation is dedicated to my late father, who left us last year and is now watching over the three of us.

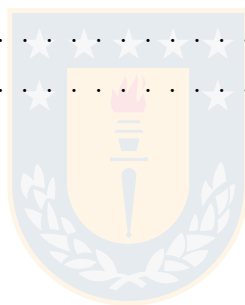
Table of Contents

Acknowledgments	v
Figures Index	ix
Tables Index	xii
1 Proposed Research	1
1.1 Introduction	1
1.2 General Formulation	1
1.3 Work Hypothesis	2
1.4 General Objective	2
1.5 Specific Objectives	2
1.6 Main Contributions	3
1.7 Thesis Organization	3
2 Bibliographic Discussion	4
2.1 Introduction	4
2.2 Previous Work	4
3 LIBS, DRS and HSI Techniques	11
3.1 Introduction	11
3.2 Laser-Induced Breakdown Spectroscopy	11
3.3 Diffuse Reflectance	14
3.4 Hyperspectral Imaging	14
4 Data Fusion Fundamentals	16
4.1 Introduction	16
4.2 Joint Directors of Laboratories Model	16
4.3 Data Fusion Levels	18
5 Combining prior knowledge with input selection algorithms	20
5.1 Introduction	20
5.2 LIBS Setup	23
5.3 Sample Preparation	24



5.4	Sample Treatment	25
5.5	Wavelength and Model Selection Method	25
5.5.1	Spectroscopic Prior Knowledge	25
5.5.2	Proposed Methodology	26
5.5.3	KBest Algorithm	29
5.5.4	LASSO Regularization	29
5.5.5	Principal Component Analysis	30
5.5.6	CARS Algorithm	30
5.5.7	Figures of Merit	31
5.6	Experimental Results	31
5.6.1	Software and Computing	32
5.6.2	ANN Training	32
5.7	Results and Discussion	32
5.8	Conclusions	41
6	An optimization approach to combine prior knowledge and LASSO regularization	42
6.1	Introduction	42
6.2	Theory	45
6.2.1	Wavelength Selection Method and Algorithm	45
6.2.2	Metrics	48
6.3	Application	49
6.3.1	Material and Methods	49
6.3.2	LIBS Setup	49
6.3.3	Software and Computing	50
6.3.4	ANN Training	50
6.4	Results and Discussion	50
6.5	Conclusions	57
7	Data fusion of LIBS-DRS and LIBS-HSI for improved analysis of mineral species in copper concentrates	58
7.1	Introduction	58
7.2	LIBS-DRS Setup	60
7.3	Sample Preparation and Treatment	61
7.4	General Methodology	65
7.4.1	Preprocessing	66

7.4.2	Data Fusion Results	67
7.5	Data Fusion of LIBS and HSI Data	70
7.5.1	LIBS-HSI Setup	70
7.5.2	Sample Preparation and Treatment	71
7.5.3	Preprocessing	71
7.5.4	Data Fusion Results	72
7.6	Conclusions	74
8	General Discussion	75
8.1	Introduction	75
8.2	Conclusions	75
8.3	Future Work	76
9	Publications	77
9.1	Introduction	77
9.2	Journals	77
9.3	Conferences	77
	Bibliography	79



Figures Index

3.1	LIBS setup [1]	11
3.2	Hyperspectral imaging with spectral scanning technique [2]	15
4.1	JDL process model	16
5.1	Experimental arrangement for the LIBS acquisition of copper concentrate pellets	24
5.2	Overview of the procedure for the selection of variables/wavelengths with different algorithms like KBest, LASSO regularization, CARS and PCA	27
5.3	Scores obtained using the KBest algorithm. The selection of a particular wavelength depends on the higher number of scores and correlated with the LIBS library data (enlarged picture)	34
5.4	Cumulative variance explained using number of principal components. A total of 8 principal components explain 99% of the variance cumulatively	35
5.5	Loadings plots of principal component analysis	35
5.6	Sum of weights obtained using CARS method	36
5.7	Sum of weights obtained using the LASSO regularization method	37
5.8	RMSEP for the test dataset and the ANN with optimized number of hidden units and different number of inputs	37
5.9	Copper wavelengths selected by the methodology and the different variable selection algorithms and prior knowledge. The vertical red lines are used to show separate parts of the spectrum within the same figure	38

5.10	Regression curves of PK (prior knowledge), LASSO, KBest, CARS and PCA algorithms for copper analysis fitted using a linear model $y=ax+b$. The corresponding figures of merits are provided in Table 5.2. The scatter plot of the measured values are the same values i.e., $y=x$ (linear), which are provided not only to depict all algorithms have a similar trend of the slope with the hypothetical line $y=x$ but also to visualize the stratified division of test data across the regression curve	39
6.1	AIC with respect to different λ_{PK} values, with $\lambda_{Lu} = 1$	51
6.2	RMSEP for the test dataset and the ANN with an optimized number of hidden units for copper	54
6.3	RMSEP for the test dataset and the ANN with an optimized number of hidden units for iron	54
6.4	RMSEP for the test dataset and the ANN with an optimized number of hidden units for arsenic	55
6.5	Copper wavelengths selected by the method and by prior knowledge	55
6.6	Iron wavelengths selected by the method and by prior knowledge	56
6.7	Arsenic wavelengths selected by the method and by prior knowledge	56
7.1	LIBS-DRS setup	62
7.2	LIBS-DRS setup: Fiber array	63
7.3	Pellet sample	63
7.4	Sample LIBS spectrum	64
7.5	Sample DRS spectrum	64
7.6	Overview of the data fusion strategies. (a) Low-level data fusion. (b) Mid-level data fusion	67
7.7	Regression plot for mid-level data fusion strategy of LIBS-DRS data	70

7.8 HSI spectra of copper concentrate samples 72

7.9 Regression plot for mid-level data fusion strategy of LIBS-HSI data 73



Tables Index

3.1	Typical laboratory laser parameters	12
5.1	Copper, iron and arsenic elemental emission lines observed in the LIBS spectra of copper concentrate [3–12]	28
5.2	Different model fit statistics in terms of parameters like AIC, RMSEP, R^2 , selected variables, number of neurons and linear model fitting parameters for different algorithms KBest, LASSO, PCA, CARS and prior knowledge for copper	40
5.3	Different model fit statistics in terms of parameters like AIC, RMSEP, R^2 , selected variables and number of neurons for different algorithms KBest, LASSO, PCA, CARS and prior knowledge for iron	40
5.4	Different model fit statistics in terms of parameters like AIC, RMSEP, R^2 , selected variables and number of neurons for different algorithms KBest, LASSO, PCA, CARS and prior knowledge for arsenic	40
6.1	Performance metrics for the newly proposed method, the previously proposed method, and prior knowledge for copper	53
6.2	Performance metrics for the newly proposed method, the previously proposed method, and prior knowledge for iron	53
6.3	Performance metrics for the newly proposed method, the previously proposed method, and prior knowledge for arsenic	53
7.1	Analytical figure of merits of LIBS, DRS, LLDF, HLDF	69
7.2	Analytical figure of merits of LIBS, HSI, LLDF, HLDF	73

Nomenclature

γ_0	multiplying factors threshold
\hat{y}	predicted concentrations
λ	regularization parameter
λ_i	LIBS wavelength
λ_j	DRS wavelength
λ_k	HSI wavelength
λ_{Lu}	multiplying factors regularization parameter
λ_{PK}	prior knowledge regularization parameter
L	loading matrix of X
T	score matrix of X
X	independent data matrix
$\mathbf{x}(i)$	input vector
ρ	correlation
f	cost function
k	number of estimated parameters in ANN model
m	number of neurons in hidden layer
n	number of model inputs
n_c	number of principal components
$o(i)$	estimated concentration
$o_L(i)$	estimated concentration for LASSO-based model
$o_{Lu}(i)$	estimated concentration for LASSO-based model with uncoupled effect



S	set of selected wavelengths
S_{min}	minimum number of selected wavelengths
w	ANN weights
x_0	spatial coordinate on x-axis
x_i	model input
y	measured concentrations
y_0	spatial coordinate on y-axis
\mathcal{D}	LIBS spectra dataset
\mathcal{S}	LIBS spectra test set
\mathcal{T}	LIBS spectra training set
\mathcal{V}	LIBS spectra validation set
$\dot{\sigma}$	derivative of the activation function
Γ	multiplying factors
\mathbf{b}	bias
$\mathbf{W}^{(1)}$	input weight vector
$\mathbf{W}^{(2)}$	output weight vector
σ	activation function
n_D	number of samples
n_S	number of samples of LIBS spectra test set
n_T	number of samples of LIBS spectra training set
n_V	number of samples of LIBS spectra validation set



Abbreviations

AAS atomic absorption spectroscopy

AES atomic emission spectroscopy

AIA accuracy influence analysis

AIC Akaike information criterion

ALH adaptive local hyperplane

ANN artificial neural networks

ANNR artificial neural network regression

ANOVA analysis of variance

ARS adaptive reweighted sampling

ASD atomic spectra database

BIC Bayesian information criterion

CARS competitive adaptive reweighted sampling

CC calibration curve

CCD charge-coupled device

CF-LIBS calibration-free LIBS

DR diffuse reflectance

DRS diffuse reflectance spectroscopy

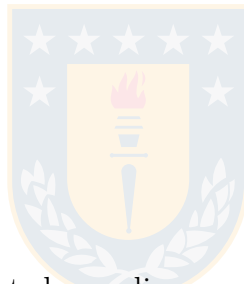
EDF exponentially decreasing function

ENR elastic net regression

FSC full-spectrum correction

FWHM full width at half maximum

GA genetic algorithm



GPR Gaussian process regression

HSI hyperspectral imaging

IR infrared

ISODATA iterative self-organizing data analysis

JDL joint directors of laboratories

KNN K-nearest neighbors

kNNR k-nearest neighbors regression

LASSO least absolute shrinkage and selection operator

LDA linear discriminant analysis

LGR logistic regression

LIBS laser-induced breakdown spectroscopy

LIW layered interval wrapper

LLDF low-level data fusion

MIR mid-infrared

MLDF mid-level data fusion

MLP multi-layer perceptron

MRCE multivariate regression with covariance estimation

MSE mean squared error

Nd:YAG neodymium-doped yttrium aluminium garnet

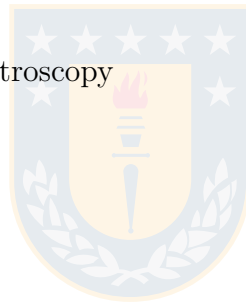
NIR near-infrared

NIST national institute of standards and technology

NMR nuclear magnetic resonance

NNR neural network regression

PC principal component



PCA principal component analysis

PCR principal component regression

PK prior knowledge

PLS partial least squares

PLSDA partial least squares discriminant analysis

PLSR partial least squares regression

QEMSCAN quantitative evaluation of minerals by scanning electron microscopy

RBF radial basis function

RF random forest

RMSE root mean squared error

RMSEP root mean squared error of prediction

ROC receiver operating characteristic

RR ridge regression

RSS residual sum of squares

SFFS sequential forward floating search

SIMCA soft independent modeling of class analogy

SPA successive projection algorithm

SVM support vector machines

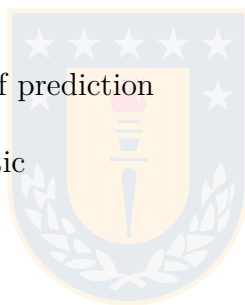
SVR support vector regression

SWIR short-wave infrared

UV ultraviolet

VCA vertex component analysis

Vis visible



1. Proposed Research

1.1 Introduction

As a mining country, Chile needs to improve the performance of its metallurgical processes. One way to do this is by developing spectral sensors for important operational variables.

This work aims to use LIBS, DRS and HSI data to obtain a more reliable chemical or mineral characterization of the input stream of a Cu smelter, which implies one kind of sample: copper concentrate.

The fusion of different data sources has proven to obtain better results than using the data sources individually [13]. Hence, for the first time, in this work, sensor and data fusion techniques will be used to characterize the elemental and mineral composition in copper concentrate samples using LIBS, DRS and HSI data.

1.2 General Formulation

The problem consists of improving the characterization of the input stream of a Cu smelter. For this, data fusion of LIBS and DRS data as well as data fusion of LIBS and HSI data will be used. The latter fusion will be made on a single point fixed by LIBS. Further details are given in Section 7.5.

LIBS and HSI technologies have been used together to classify some materials, but no sensor or data fusion techniques have been developed using these technologies [14]. This finding appears as a problem to be solved and to be applied to the pyrometallurgical industry.

Some advantages of the LIBS system are the accurate results obtained from its data or the possibility to identify any element. In contrast, the drawback of this system is the small spatial resolution, limited to just a single point per measure. On the other hand, one advantage of the HSI system is obtaining a wider range of data per measure. At the same time, some drawbacks are the less accurate results obtained from training a model using HSI data only and the lack

of full labeling for the data. Further advantages and limitations for the techniques are given in Chapter 3.

Considering the advantages and drawbacks of LIBS and HSI systems, the fusion of these seems like a good approach for getting the best out of the two techniques since the small spatial resolution of LIBS is compensated by the large spatial resolution of the HSI system, and the less accurate results obtained from HSI are compensated by the accurate results obtained from LIBS.

This fusion will enable the characterization of the copper concentrate samples better than using the individual sources and will be used in the future to determine specific regions where to use the LIBS technique to calibrate the HSI system online.

On the other hand, the fusion of LIBS and DRS data includes two complementary spectral sources, elemental information from LIBS and molecular information from DRS, which can be used for the determination and quantification of mineral compositions.

1.3 Work Hypothesis

It is possible to obtain improved results on the characterization of the elemental composition of copper, iron and arsenic, and mineral composition of bornite, chalcopyrite, covellite, enargite and pyrite in copper concentrate samples by fusing the data of LIBS and DRS technologies, and by fusing the data of LIBS and HSI technologies since these are complementary data sources.

1.4 General Objective

To improve the quantitative characterization in the input stream of a Cu smelter by fusing the data from LIBS and DRS techniques, and by fusing the data from LIBS and HSI techniques.

1.5 Specific Objectives

- To develop variable selection methods for reducing the dimensionality of LIBS spectra.

- To quantitatively analyze LIBS, DRS and HSI data through artificial neural networks.
- To fuse LIBS and DRS data through low and mid-level data fusion techniques to estimate the mineralogical composition of copper concentrate samples.
- To fuse LIBS and HSI data through low and mid-level data fusion techniques to estimate the mineralogical composition of copper concentrate samples.

1.6 Main Contributions

- Chapter 5 was submitted, accepted and published in *Analytical Methods* in 2021 and early versions of the proposed method in this Chapter were presented at the IV EIQ, Porto Alegre and at PITTCON, Philadelphia, in 2019.
- Chapter 6 was submitted, accepted and published in *Minerals Engineering* in 2021.
- Chapter 7 will be submitted to *Chemometrics and Intelligent Laboratory Systems* in 2021.

1.7 Thesis Organization

This thesis deals with the study of data fusion techniques for the quantitative analysis of copper concentrates. In Chapter 2, a bibliographic discussion related to the main topics covered in this work is given. Chapter 3 and Chapter 4 present the fundamental concepts of the main spectroscopic techniques and the data fusion methodologies used in this work, respectively. In Chapter 5, a first variable selection method is proposed for LIBS spectra. Then, in Chapter 6, a second variable selection method for LIBS spectra is proposed, which is an improvement with respect to the previously proposed method in terms of required steps to achieve the variable selection. In Chapter 7, the data fusion techniques are implemented to fuse LIBS-DRS and LIBS-HSI data. Chapter 8 outlines the conclusions and further work. Finally, in Chapter 9, the contributions from this work are presented.

2. Bibliographic Discussion

2.1 Introduction

In this Chapter, the literature associated with the research is revised. It starts by reviewing the main topics covered in this thesis, then a discussion about the literature is made regarding what it will be done in this work.

2.2 Previous Work

The work of Anabitarte et al. reviews the theory of the LIBS system, starting with an explanation of the physics involved in plasma induction and the features of this plasma in LIBS. It is then followed by a description of the primary devices which compose a LIBS setup. The main algorithms for quantitative chemical analysis or sample classification are described, which include artificial neural networks (ANN), support vector machines (SVM), and K-nearest neighbors (KNN). In addition, it mentions some algorithms to discard redundant information from the spectral data. It ends with future challenges and applications for the LIBS technique [15]. In a different work from the previous group, they propose a sensor system based on a laser-induced breakdown spectroscopy setup to detect and discriminate the sample before the welding process. A spectral algorithm based on support vector machines is used as a classifier to identify areas with aluminum presence automatically. The algorithm uses a polynomial kernel, and the addition of a previous data reduction step is then used to improve the performance both in the execution time and classification error [16]. In the work of Gething et al., the laser-induced breakdown spectroscopy technique is used to recycle chromated copper arsenate-treated materials. For the classification process, a series of regression-fitted calibration lines were used. The most appropriate regression analysis with data reduction procedures was determined and subsequently used to comparatively predict the level of residual preservative relative to reclaimed decking lumber [17]. Godoi et al. use the laser-induced breakdown spectroscopy technique to directly analyze plastic toys and Cd, Cr, and Pb determination. The classification models used were Partial Least Squares - Discriminant Analysis (PLSDA), Soft Independent Modeling of Class Analogy (SIMCA), and K-Nearest Neighbors. The latter obtained the best results since

there was no clear separation between the classes. KNN performed better because this is a deterministic model, unlike SIMCA and PLS-DA models, which are probabilistic methods and require classes to be well characterized [18]. The work of Harmon et al. describe recent applications of the laser-induced breakdown spectroscopy technique to the analysis of geological and environmental materials. Following a summary of the fundamentals of the LIBS analytical technique and its potential for chemical analysis in real-time, the history of LIBS application to the analysis of natural fluids, minerals, rocks, soils, sediments, and other natural materials is also described. Statistical signal processing is included as well, reviewing works that have used the partial least squares and artificial neural network techniques, and others, such as the wavelet approach, principal component regression (PCR), soft independent modeling of class analogy, linear discriminant analysis (LDA), among others [19]. In the work of Cong et al., the traditional calibration curve (CC) method is compared with the partial least square method. It finds that the PLS method overcomes the CC method in the LIBS quantitative analysis since the PLS method establishes a more stable quantitative calibration model due to the reduction of the dimensionality of raw data with multiple correlations between variables [20]. In the work of Pokrajac et al., several classification algorithms are applied and compared for the multi-class classification of laser-induced breakdown spectroscopy data of four commercial samples of proteins diluted in phosphate-buffered saline solution at different concentrations. This task is achieved by using principal component analysis (PCA) as a method for dimensionality reduction. The algorithms applied were K-nearest neighbors, support vector machines, adaptive local hyperplane (ALH) and linear discriminant analysis [21]. The work of Coelho et al. proposes a novel analytical methodology for soil classification based on laser-induced breakdown spectroscopy and chemometric techniques. Linear discriminant analysis is employed to build a classification model based on a reduced subset of spectral variables. For variable selection, three techniques are considered, namely the successive projection algorithm (SPA), the genetic algorithm (GA), and a stepwise formulation. The use of a data compression procedure in the wavelet domain is also proposed to reduce the computational workload involved in the variable selection process [22]. In the paper of Sheng et al., LIBS based on chemometrics was performed to identify and classify iron ore samples. Random Forest (RF) algorithm was used for the classification, and its results were compared to those of the support vector machines algorithm. Although results show that the prediction accuracies of SVM and RF models were acceptable, RF exhibited the best classification predictions [23]. In the work of Sirven et al., laser-induced breakdown spectroscopy was applied to the analysis of three chromium-doped soils. Two chemometric techniques, principal component analysis and neural network analysis, were used to discriminate the soils based on their LIBS spectra. Neural networks proved to be

more efficient than PCA, with a correct identification rate of 100% [24]. In the work of Tan et al., the feasibility of laser-induced breakdown spectroscopy is investigated for sea salts classification. The principal component analysis of the LIBS spectra provided the score plot with quite a high degree of clustering. Classification models were developed by partial least squares discriminant analysis and evaluated. Cross-validation used to classify provenance resulted in 85% accuracy with five PCs [25]. An automated method for classifying four types of proteins from laser-induced breakdown spectroscopy data is proposed in the work of Vance et al. The high dimensionality spectroscopy data is preprocessed using the linear dimensionality reduction technique of principal component analysis. Then classification is performed using support vector machines and adaptive local hyperplane. The wrapper method is used with various principal components to determine the optimal number of extracted features. The classification accuracy is estimated using four-fold cross-validation [26].

The work of Gray et al. summarizes the reflection characteristics of about 200 opaque mineral ore minerals, aiming to provide an aim to mineral identification. It finds broad relationships between spectral profiles and conoscopic features of reflection rotation and apparent angle of rotation, which are of paramount importance to the theoretical reasoning on questions of mineral sequence and fabric, and are vital factors in the establishment of precise determinative procedures [27]. In the work of Pirard et al., a multispectral system at a microscopic scale for ore analysis as well as the procedures for proper calibration to obtain precise reflectance measurements are described. This system is compared to colour imaging systems, and the advantages of the former is discussed in quantitative terms. It ends with discussing the potential for automatic identification of ore minerals [28]. The work of Catalina et al. presents the CAMEVA system, a multispectral reflectance microscopy system specially designed to facilitate the identification and characterization of mineral phases present in polished preparations of metallic ores, as well as to automate the performance of different types of quantitative analysis on them. The tests carried out show that the system allows the automated and reliable identification of ores of industrial interest, based on multispectral information in the VNIR range collected in a specific database; this database, which includes the 70 minerals of greatest interest, is easily expandable [29]. In the work of López-Benito et al., the performance of the system relying on the measurement of multispectral specular reflectance, supported by a multispectral reflectance database covering the VNIR range built with the AMCO System, is analysed, comparing the reliability of different classification methods to achieve ore identification, based respectively on spectral angle mapper, euclidean distance, Mahalanobis distance and linear discriminant analysis. The tests carried out reveal that the last two techniques are powerful tools to determine to which mineral corresponds a pixel based on its reflectance spectrum [30]. The work of Tessier et al. describes a general

machine vision approach for on-line estimation of rock mixture composition, and it is illustrated on a very challenging nickel mineral system: very heterogeneous minerals, similar coloration, and rock fragments can be dry or wet. Through a pilot plant conveyor belt application, very good results were obtained for dry minerals [31].

In the work of Picón et al., a way to classify non-ferrous waste material is developed using hyperspectral data. For this purpose, an algorithm merges the spectral and spatial features. The dimensionality of the hyperspectral data is reduced by constructing a bio-inspired spectral fuzzy set to achieve an efficient implementation. This technique is compared to other decorrelation techniques such as Principal Component Analysis, Linear Discriminant Analysis, and Wavelet decomposition, finding that it overcomes all these techniques. The multivariate Gaussian classifier is proposed to perform the material classification since the Gaussian distributions approximate the dispersion of the spectral-spatial feature vectors within each class of non-ferrous materials [32]. In the work of Melessanaki et al., the laser-induced breakdown spectroscopy along with hyperspectral imaging analysis were used in the identification of pigments in an illuminated manuscript dating from the 12th-13th century AD. LIBS analysis led to the identification of most pigments. Combining LIBS with hyperspectral imaging provides improved information for the discrimination of pigments having similar color appearance but a different chemical structure since imaging provides mapping information regarding pigments' spatial distribution [14]. In the work of Luo et al., spectral and geometrical features are integrated for the classification of hyperspectral images. Vertex Component Analysis (VCA), a linear unmixing algorithm, is used to reduce the hyperspectral data's high dimensionality and extract the endmembers and their abundance maps. Support Vector Machines with Gaussian kernel is used as a classifier. The optimal scale parameter of the Gaussian kernel is selected by 5-fold cross-validation. The results are similar to those obtained using methods based on Principal Component Analysis and Extended Morphological Profiles [33]. In the work of Bris et al., a super spectral sensor dedicated to urban materials classification is designed. The selection of the optimal spectral band subset used for this sensor is described, which is attained by the incremental algorithm Sequential Forward Floating Search (SFFS) and the Genetic algorithm. The score used to evaluate the relevance of band subsets is a wrapper score that relies on a Random Forests classifier and takes into account classification confidence. The data was taken from different libraries available, and some synthetic data was generated. The selected band subsets were evaluated considering the classification quality reached using a radial basis function (RBF) kernel support vector machine classifier. It was confirmed that a limited band subset was sufficient to classify common urban materials [34]. The work of Melgani et al. addresses the problem of the classification of hyperspectral remote sensing images by support vector machines. The

performances of two nonparametric classifiers are compared, concluding that SVMs are much more efficient in terms of classification accuracy, computational time, and stability to parameter setting. Another objective of this work was to solve multiclass problems in hyperspectral data using ensembles of binary SVMs, finding that the multiclass strategy should be selected according to a proper tradeoff between classification accuracy and computational time [35]. In the work of El Rahman, unsupervised hyperspectral image classification algorithms are used to classify an image of Washington DC. Principal Component Analysis and then K-means and Iterative Self-Organizing Data Analysis (ISODATA) algorithms are applied. It is found that the ISODATA overall accuracy is better than that of the K-means classifier [36]. In the work of Candiani et al., a hyperspectral image system is used to characterize five essential metal particles derived from the fine-shredded waste of electrical and electronic equipment. Several combinations of different methods for illumination compensation and different classification algorithms were tested and compared, including spectral angle mapper, minimum distance, Mahalanobis distance and maximum likelihood, being the Mahalanobis distance the one with the best results and the minimum distance the one with the worst [37].

In the work of Ramos et al., low-level and mid-level data fusion are used to classify ochre pigments. Raman and X-ray fluorescence data are fused. Partial least squares-discriminant analysis is used as a classifier. The best results are obtained with the mid-level data fusion technique, i.e., the combination of signal features. The benefits and drawbacks of each technique are discussed [38]. In the work of Di Anibal et al., mid-level and high-level data fusion are used to classify Sudan dyes. Two spectroscopic techniques are fused, UV-visible and H NMR. Partial least squares-discriminant analysis is used as a classifier. Fuzzy aggregation connective operators are used in the decision-level fusion technique. Both data fusion techniques performed better than the classification with individual sources, demonstrating that the two spectroscopic techniques' information has a synergistic effect [39]. In the work of Bevilacqua et al., low-level and mid-level data fusion were used for food authentication and traceability. NIR and MIR spectral data were fused. Partial least squares-discriminant analysis is used as a classifier. The best results are obtained with the mid-level data fusion technique since this does not suffer from the increase in the number of irrelevant predictors as in the case of low-level data fusion [40]. In the work of Biancolillo et al., low-level and mid-level data fusion are used to characterize a high-quality beer. Five instrumental techniques are fused: thermogravimetry, mid-infrared, near-infrared, ultra-violet, and visible spectroscopy. Partial least squares-discriminant analysis and soft independent modeling of class analogies are used as classifiers. The best results are obtained with the mid-level data fusion technique, classifying all the training and validation samples correctly [41]. In the work of Jiang et al., low-level and mid-level data fusion are used for

simultaneous determination of six ginsenosides and four saccharides in Chinese herbal injection. NIR and UV spectral data were fused. Partial least squares regression and uninformative variable elimination by PLS are used as regressors. The best results are obtained with the mid-level data fusion technique, and both performed better than the regression with individual sources [42]. In the work of Makarau et al., mid-level data fusion is used for urban area classification. Multispectral image and laser Digital Surface Model data are fused. Feature extraction, dimensionality reduction, and supervised classification are used. A neural network is used as a classifier. The results are improved by using this data fusion technique [43]. In the work of Marhoubi et al., low-level data fusion is used for activity recognition in mobile device space. Data from a mobile accelerometer and gyroscope are fused. Four algorithms are used as a classifier, namely multi-layer perceptron, random forests, bagging with MLP, and bagging with RF. A moving average filter was used to minimize the noise in both sensors, improving the classification rate for the MLP [44]. In the work of Li et al., low-level, mid-level, and high-level data fusion are used for geographical traceability of *Panax notoginseng*. Data from Fourier transform mid-IR spectrum and near-IR spectrum were fused. Random forest was used as a classifier in all cases. For low-level data fusion, the data was directly concatenated into a new matrix. For mid-level data fusion, some variables are extracted from the spectral data using principal component analysis. For high-level data fusion, the results of the RF classifiers were concatenated, and then an SVM was used to fuse the decision result. The best results were obtained for high-level data fusion [45]. In the work of Obisesan et al., mid-level and high-level data fusion are used to authenticate the geographical origin of palm oil. Data from liquid chromatography coupled to two detectors (ultraviolet and charged aerosol) were fused. Partial least squares-discriminant analysis was used as a classifier. For mid-level data fusion, principal component analysis and interval partial least squares were used to obtain the variables. For high-level data fusion, fuzzy aggregation connective operators were used. The best results were obtained with high-level data fusion, and both techniques were better than the individual techniques [46].

Both qualitative and quantitative analyses for the LIBS technique have been reviewed. Different models have been tested on spectral data to identify materials or finding the material composition. In [15], the fundamentals of LIBS are explained. [16], [17], [18], [21], [22], [23], [25] and [26] use classification for LIBS qualitative analysis. Many of these use dimensionality reduction techniques to handle the spectral data. On the other hand, in [20] and [24], regression is used for LIBS quantitative analysis. Whereas in [19], some applications of the LIBS technique are reviewed.

ANNs have proven to obtain better results than some classic algorithms because ANNs are more complex models, which can address the nonlinearities of the data. These models work for both classification and regression problems. Considering this, ANNs will be used in this work to deal with LIBS, DRS and HSI data.

The reviewed works show that reflectance data can be used for determining relevant information about mineral ores, obtaining good results in identification and determination of ore compositions. Regarding the HSI technique, all the reviewed articles use classification in order to analyze the data. All these works will be used as an initial step for analyzing the actual data in this project. Moreover, some of the mentioned algorithms will be used in the spectral data.

In [14], no data fusion techniques were used, despite having different data sources (LIBS and HSI). This finding appears as a problem to explore and apply to this project. One of the benefits this could bring is taking advantage of the HSI data, which provides a vast amount of information, and the LIBS data, which provides more accurate outputs from a model. Thus, combining these two sources will get the best out of the two techniques. The other benefit is to allow the possibility of an online characterization of the process, with a more accurate result than with the individual sources.

Three levels of data fusion have been reviewed, low-level, mid-level, and high-level data fusion. All of these obtain better results than those with individual sources, as shown in the reviewed works. LIBS and HSI data have been successfully analyzed with chemometrics techniques individually, and the use of data fusion, which has not been applied to these, will only increase the performance of the models, because data fusion uses the best from the individual models, and guarantees at least equal results than with individual models.

3. LIBS, DRS and HSI Techniques

3.1 Introduction

Laser-induced breakdown spectroscopy and Diffuse Reflectance Spectroscopy are techniques that provide an accurate in situ quantitative chemical analysis. On the other hand, the Hyperspectral Imaging technique provides a wider range of data per measure.

In this Chapter, the fundamental theories of LIBS, DRS and HSI techniques will be reviewed, as well as the advantages, considerations, and some applications.

3.2 Laser-Induced Breakdown Spectroscopy

LIBS is an emerging technique for determining elemental composition along with chemometrics techniques, which allows qualitative and quantitative analysis from the spectral data obtained. Different types of classifiers have been used along with the LIBS technique. The spectral data obtained is a fingerprint of the measured sample, which allows the chance to classify and characterize the sample's elemental composition.

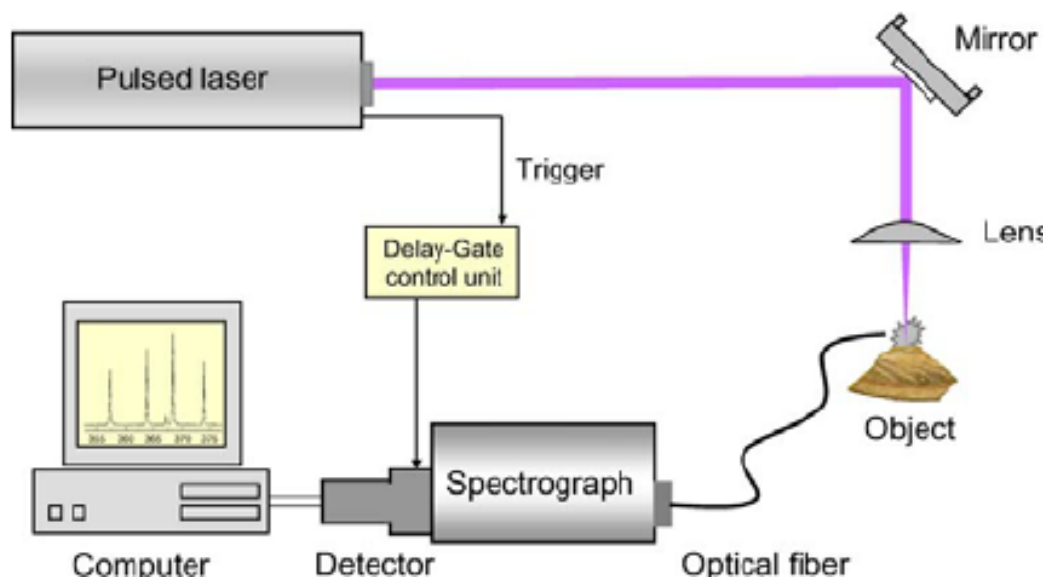


Fig. 3.1: LIBS setup [1]

In Fig. 3.1, there is a typical setup for LIBS. The setup is mainly composed of a laser, optical systems, and detection system. A computer is also used to process the obtained data.

The main component of LIBS is the laser, and the primary laser used in LIBS technique is the Nd:YAG laser. In this, a laser-shot or flashlamp is fired to produce excitation on the material. If this excitation is sufficiently strong, ablation is produced, and a temporary microplasma is formed. An optical system can acquire the emitted light to obtain a spectrum, which can be further processed to characterize the elemental composition.

Important parameters for the laser specification include pulse energy, pulse repetition rate, beam mode quality, size/weight, and cooling and electrical power requirements. The wavelength of the laser beam is also considered. The fundamental wavelength is 1064 [nm] and can be converted to others. Typical values for laboratory lasers are shown in Table 3.1 [47].

Table 3.1: Typical laboratory laser parameters

Parameter	Value	Unit
Pulse energy	450	mJ
Pulse width	5-8	ns
Repetition rate	10	Hz
Mass	71	kg
Cooling	Water	-
Power requirements	220, 12	V, A

The optical system is composed of lenses and mirrors to focus the laser pulses onto the sample. These can also be used to collect the plasma light and focus it into a spectrograph or onto fiber optic cables. Relevant parameters of the lens include the focal length, diameter, and material.

An optical arrangement coupled to a optical fiber are used to collect plasma light when the detection system is far from the target sample.

The detection system is composed of a spectrograph and a detector. The spectrograph is used to separate the plasma light into a wavelength spectrum and to record the data, whereas the detector returns a current in proportion to the intensity of the incident light. By integrating the detector current on a capacitor for a specific time, a voltage will result, which can be digitalized to produce a signal in the form of counts.

As an atomic emission spectroscopy (AES) method, LIBS has the advantage of detecting all elements, and the capability of simultaneous multi-element detection, compared with some

non-AES-based methods.

Additionally, LIBS has some advantages compared with other AES-based methods. These are listed below [48]:

- simplicity,
- rapid or real-time analysis,
- no sample preparation,
- allows in situ analysis requiring only optical access to the sample,
- ability to sample gases, liquids, and solids equally well,
- good sensitivity for some elements, difficult to monitor with conventional AES methods,
- adaptability to a variety of different measurements scenarios,
- robust plasma that can be formed under conditions not possible with conventional plasmas.

One thing to be considered is the homogeneity of the sample, especially in solids. The same happens with bulk non-uniformity or non-representative surface composition, where one sample does not represent the entire material, considering the laser only shots at the sample's surface.

Another consideration to take into account is the matrix effect, that is, the physical properties (specific heat, latent heat of vaporization, thermal conductivity, absorption) and composition of the sample affect the element signal, or when the presence of one element affects the emission characteristics of another element.

The way the laser is focused or lens-to-sample distance may also affect the element emission signals. Lastly, there are safety considerations when using LIBS, which require the operations procedures to be standardized. Some of these considerations may affect the characterization of the samples.

A wide range of applications exists for the LIBS technique, including metals, alloys, carbon and steel; explosive, biological and chemical threat detection; soil analysis, carbon in soils, contamination; minerals, oil, crystals and gems, ceramics; water and aqueous solutions, among many others.

3.3 Diffuse Reflectance

In contrast to LIBS, DRS does not utilize a high-power laser source. It uses simple broadband light sources to illuminate the sample. The diffuse reflectance spectra is a result of a combination of different processes, like scattering and absorption, and it provides rich spectral information. The back reflected, diffusely scattered light (some of which is absorbed by the sample) is then collected by the accessory and directed to the detector optics.

Some factors related to a high spectral quality for diffuse reflectance sampling are; reducing particle size, sample dilution, homogeneity and packing of the sample.

Applications in which DRS can be applied ranges from color measurements of textiles, pharmaceuticals, building materials, paper and pulp materials etc., to adsorption studies and other basic investigations in physical, inorganic and organic chemistry [49].

DRS has several advantages over other types of spectroscopic techniques, including enhanced scattering phenomenon in powder materials. Moreover, the effects of light scattering in the absorption spectra of powder samples dispersed in liquid media can be avoided using DRS [50]. Finally, it reduces time in sample preparation.

3.4 Hyperspectral Imaging

As well as the previous techniques, the HSI technique provides spectral data, which can be analyzed to obtain information about the elemental composition of a sample. A hyperspectral image is characterized for having many wavelengths or other variable bands and by the possibility to express a pixel as a spectrum with spectral interpretation, spectral transformation, or spectral data analysis. [51].

The way the spectral data is obtained differs from LIBS or DRS, although. There exist four main ways of obtaining spectral data. Two of them will be explained.

The first way to obtain spectral data is called spatial scanning. In this, each sensor output represents a full slit spectrum. The spatial dimension is collected through platform movement or scanning, and the image is then reconstructed.

Another way to obtain spectral data is called spectral scanning. In this, each 2-D sensor

output represents a monochromatic spatial map of the scene. HSI devices for spectral scanning are typically based on optical band-pass filters. The scene is spectrally scanned by exchanging one filter after another while the platform must be stationary.

Fig. 3.2 shows the hypercube from an HSI measure and the spectral data obtained from a couple of pixels of the image, using spectral scanning.

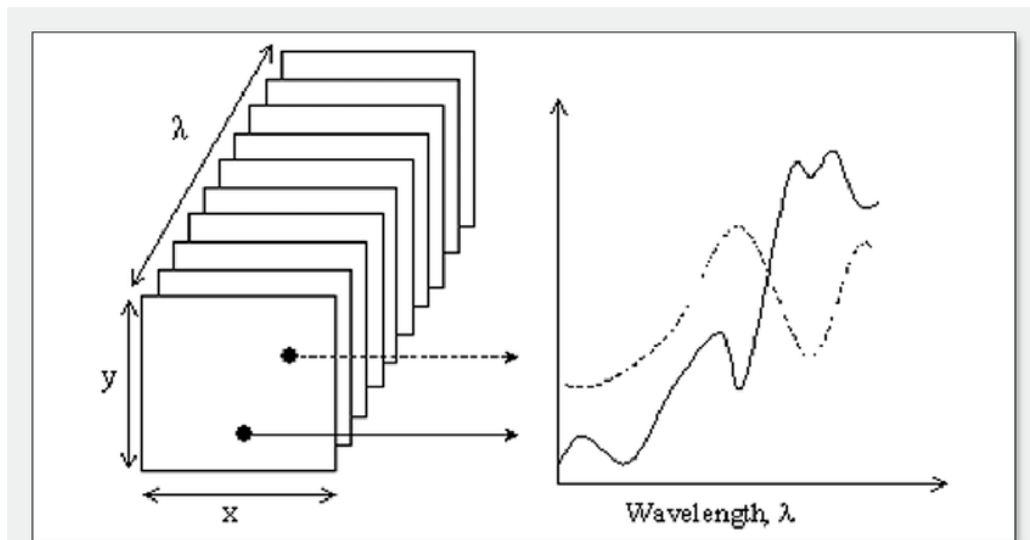


Fig. 3.2: Hyperspectral imaging with spectral scanning technique [2]

The main advantage of HSI is that it takes a spectrum for every pixel or point, where all the information is available to process.

The main disadvantages of HSI are cost and complexity, which are due to the ample data storage capacities, fast computers, and sensitive detectors that are needed.

Among the applications of HSI, we find several fields, including agriculture, eye care, food processing, mineralogy, surveillance, physics, astronomy, chemical imaging, environment, among others.

4. Data Fusion Fundamentals

4.1 Introduction

Data fusion is defined as the combining of sensory data or data derived from sensory data such that the resulting information is in some sense better than would be possible when these sources were used individually [13]. In this Chapter, the fundamentals of data fusion will be introduced, and some models to use data fusion will be proposed.

4.2 Joint Directors of Laboratories Model

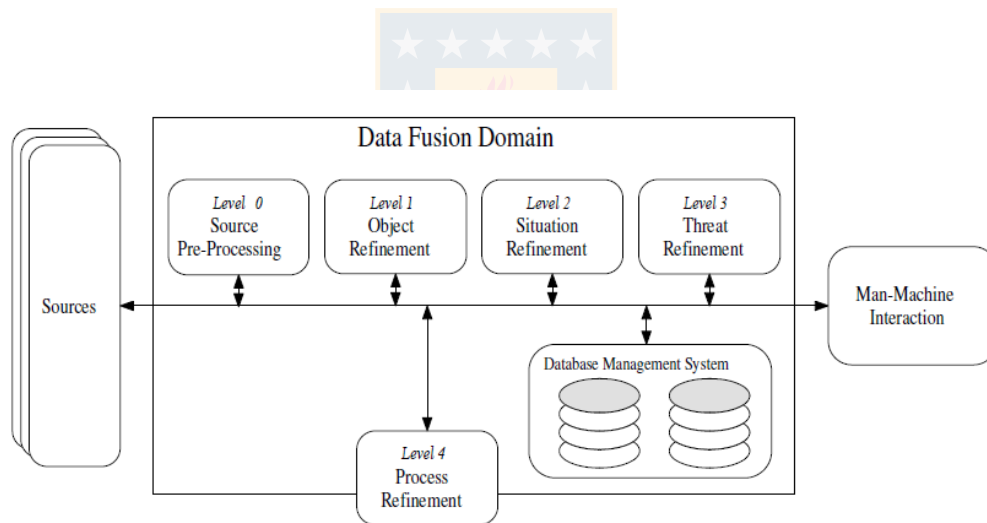


Fig. 4.1: JDL process model

Fig. 4.1 shows the JDL model, one of the main process models for data fusion. The elements of this model are described below [13].

Sources: The sources provide information from a variety of data sources, including sensors, a priori information, databases, human input.

Source preprocessing (Level 0): The task of this element is to reduce the processing load of the fusion processes by prescreening and allocating data to appropriate processes.

Object refinement (Level 1): This level performs data alignment (transformation of data to a consistent reference frame and units), association (using correlation methods), tracking actual and future positions of objects, and identification using classification methods.

Situation refinement (Level 2): The situation refinement attempts to find a contextual description of the relationship between objects and observed events.

Threat refinement (Level 3): Based on a priori knowledge and predictions about the future situation, this processing level tries to draw inferences about vulnerabilities and opportunities for operation.

Process refinement (Level 4): Level 4 is a meta-process that monitors system performance (e. g., real-time constraints) and reallocates sensors and sources to achieve particular mission goals.

Database management system: The task of the database management system is to monitor, evaluate, add, update, and provide information for the fusion processes.

Man-machine interaction: This part provides an interface for human input and communication of fusion results to operators and users.

Data fusion has several advantages over individual techniques. One of them is increased robustness and reliability by having data from multiple sources. Reduced ambiguity and uncertainty are other advantages of data fusion. Finally, it helps improve a model's results by including more data, adding more features, or changing the individual model.

Data fusion has its initial steps on military applications, and it then included some non-military applications.

Among the military applications, we find:

- Automated target recognition.
- Guidance for autonomous vehicles.
- Battlefield surveillance.
- Automated threat recognition systems.

Non-military applications include:

- Monitoring of manufacturing processes.
- Maintenance of complex machinery.
- Robotics.
- Medical diagnosis.

4.3 Data Fusion Levels

Data fusion typically has two goals: increasing the prediction accuracy and obtaining better understanding of the studied phenomena. Three data fusion levels are found: low, mid, and high-level data fusion levels. In low-level data fusion, the raw data is concatenated and used as input for a model. In mid-level data fusion, some variables or features are obtained from the sources and then used as inputs for a model. In high-level data fusion, two or more models are fitted with the individual sources. Then a final model outputs a response, with techniques including voting, fuzzy logic, and Bayesian inference.

Low-level data fusion has the advantage of being the most straightforward way to fuse data sources and also the possibility to interpret the results directly in terms of the original variables collected in each data block. However, it may yield overfitted results as increases the number of inputs for the predictive models. One advantage in adopting mid-level data fusion approach is that the noninformative variance can be removed in the features reduction step, and thus the final models may show better performance (e.g., in prediction). High-level data fusion has the possibility of further improving the results, however it is considerably more used for classification purposes.

Data blocks can be concatenated sample wise when the same set of samples is analyzed by different techniques, as in the cases under study, where the same samples are analyzed using different spectroscopic techniques.

One of the main issues for data fusion is how to deal with the heterogeneity of data variables (data entries in general) with respect to different physical units, measurement scales, etc. This issue can be addressed by preprocessing such as scaling and normalization. A further issue is how to ensure a fair contribution by each data block in the fusion process considering their different size, variance, and information content. In simple concatenation, in most of the cases this is dealt with blockscaling, e.g., to equal sum of squares, whereas in multiblock methods it

is usually dealt with by weighting the data blocks, and the weighting scheme choice is guided by the objective that is pursued, e.g., having an equal contribution from each data block or weighting a data block more if it has a lower degree of redundancy with the others, etc. Other aspects, which can be addressed by preprocessing, are misalignment among data blocks, e.g., in time or space, outliers, and spurious data. A critical issue is also represented by the possible difference in the type/level of noise present in different data blocks, which may affect the data fusion process [52].

Another aspect to consider, especially in mid-level data fusion, relies on how the variables are selected, which can be achieved by directly discarding the least relevant variables or by using a feature transformation algorithm such as PCA. The former is preferred since the results can be more easily interpreted.

Regarding the input data to be used for the models, LIBS data is composed of n , DRS data is composed of m reflectance values at coordinate (x_0, y_0) . Whereas for HSI data, this is composed of l reflectance values at coordinate (x_0, y_0) , where (x_0, y_0) are the coordinates of the data fusion point:

LIBS Data(λ_i, x_0, y_0), $i=1, \dots, n$

DRS Data(λ_j, x_0, y_0), $j=1, \dots, m$

HSI Data(λ_k, x_0, y_0), $k=1, \dots, l$



For the models, the following assumptions are necessary:

- The spectral data are supposed to match for both LIBS and HSI techniques.
- The fusion is made on the spatial points fixed by the LIBS technique, which implies that the spatial information from LIBS will be more relevant.

5. Combining prior knowledge with input selection algorithms

Laser-induced breakdown spectroscopy (LIBS) is an emerging technique for the analysis of rocks and mineral samples. Artificial neural networks (ANN) have been used to estimate the concentration of minerals in samples from the LIBS spectra. These spectra are very high dimensional data, and it is known that only specific wavelengths have information on atomic and molecular features of the sample under investigation. This Chapter presents a systematic methodology based on the Akaike information criterion (AIC) for selecting the wavelengths of LIBS spectra as well as the ANN model complexity, by combining prior knowledge and variable selection algorithms. Several variable selection algorithms are compared within the proposed methodology, namely KBest, a least absolute shrinkage and selection operator (LASSO) regularization, principal component analysis (PCA), and competitive adaptive reweighted sampling (CARS). As an illustrative example, the estimation of copper, iron and arsenic concentrations in pelletized mineral samples is performed. A dataset of LIBS emission spectra with 12287 wavelengths in the range of 185-1049 nm obtained from 131 samples of copper concentrates is used for regression analysis. An ANN is then trained considering the selected reduced wavelength data. The results are satisfactory using LASSO and CARS algorithms along with prior knowledge, showing that the proposed methodology is very effective for selecting wavelengths and model complexity in quantitative analyses based on ANN and LIBS.

5.1 Introduction

LIBS is an emerging spectroscopic technique based on the spectral emission of a plasma generated by a pulsed laser. Chemometric based publications are increasing in the field of LIBS, see for instance recent survey papers [53, 54]. The spectral lines from the LIBS spectra are related to atomic, ionic and molecular features. Owing to these rich features/variables several fields of studies have been explored using this spectroscopy [47, 48]. Highly heterogeneous matrices like soil and rock samples are also explored for qualitative and quantitative purposes [55, 56].

The application of a multivariate regression approach is important for the quantitative anal-

ysis, while univariate analysis is not desirable because of the limitation in addressing nonlinearities arisen due to plasma instability and matrix effects. Among the multivariate methods, the most applied techniques for regression study are principal component regression (PCR) and partial least square regression (PLSR) as linear regression methods. Other nonlinear regression methods like K-nearest neighbour regression (kNNR), support vector regression (SVR) and artificial neural network regression (ANNR) are also vividly utilized. Another method called calibration free LIBS (CF-LIBS) is alternate to all these regression methods. This method is generally used to overcome the matrix effect, it relies upon deducing plasma parameters and is still yet not fully adopted by researchers due to its complex procedures and assumptions [57].

Due to the flexibility in dealing with large amounts of nonlinear correlated data, ANN has gained attention for regression studies. The processes which disrupt linearity between the concentrations and spectral intensities depend on many factors such as: continuum emission, background noise, self-absorption, matrix effects, and spectral interferences. These factors can be taken into account if their effects are repeatable in all stages of measurements including training, validation, and testing. Some drawbacks of ANNs are that they take a lot of effort to train, may be trapped in local minima, and require domain knowledge and expertise. However, these effects are diminished when working with smaller datasets, features and model architectures.

Nowadays, due to developments in detection schemes such as broadband detectors like echelle spectrographs or multiple channel CCD spectrographs, the spectral data size became larger. The direct use of the entire data set poses a serious limitation by producing overfitted results. However, it is known that only specific characteristic wavelengths, which are signature of the elements present in the sample, are more informative for quantification. These wavelengths, we termed as prior knowledge wavelengths. Thus, finding a suitable subset of prior knowledge wavelengths is an important step towards building an effective ANN model for quantitative analysis.

The input of data should be assessed by models that consider peaks owning spectroscopic relevance, avoiding peaks affected by different phenomena like self-absorption and saturation. A spectral window approach can be utilized where the elemental peaks exist [58]. On the other hand, selecting too many wavelengths increases the number of input variables, model complexity and may also increase the chance of overfitting.

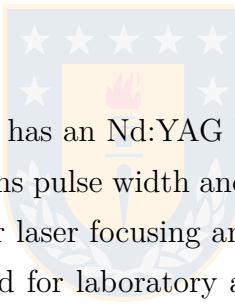
Variable selection methods can be used with spectral data to select the best input variables for the models. These less sized inputs are useful for testing real-time concentration analysis, where computation time is a crucial factor.

In the LIBS literature, there are several feature/variable selection approaches employed for the purpose of classification and regression. The feature selection methods can be divided into two categories: (i) Preexisting spectral knowledge based feature selection (ii) algorithm based feature selection. The first feature selection category considers atomic, molecular lines, peak attributes (such as area, width and height), sub spectra, combination of sub spectra [58], ratios related to peak attributes [59] and full spectra, which are routine in the spectroscopic research. The second category relates to the mathematical approaches like: (a) dimensionality reduction techniques, PCA and PLS (b) model-based techniques, wavelet compression methods, filter, wrapper, and embedded methods. These methods are assessed based on different evaluation parameters like R^2 , mean squared error (MSE), root mean squared error (RMSE), precision, accuracy, sensitivity, specificity, receiver operating characteristics (ROC), Akaike information criterion (AIC) and Bayesian information criterion (BIC). A conceptual review of various variable selection methods for applying ANN models can be found in the reference [60] and can be extended to any other model. When compared to other spectroscopic methods like IR spectroscopy and Raman, the use of algorithm-based feature selection in LIBS is found to be lesser. Two feature selection approaches, the least absolute shrinkage and selection operator (LASSO) and sparse multivariate regression with covariance estimation (MRCE), were used for identifying wavelengths associated with carbon and its prediction in soil [61]. Xu et al proposed a new method, accuracy influence analysis (AIA), for opting informative features of steel LIBS spectra for the SVM classification model [62]. Besides conventional variables like sub-spectrum selection, Myakalwar et al proposed the use of GA for variable reduction, where using only 10% of the data, they were able to classify better than with the full spectra and the sub-spectral data which depends on prior knowledge [58]. Gonzaga et al determined Mn content in steel samples with regression analysis by proposing a new method, forward variable selection [63]. In the work of Duan et al, an automatic variable selection method for LIBS is proposed, which is based on full-spectrum correction (FSC) and modified iterative predictor weighting-partial least squares [64]. This method was compared with GA and SPA, showing similar results but faster computation times. In the work of Lu et al, a new layered interval wrapper (LIW) feature selection method was proposed for LIBS data, showing better accuracy results in the classification of LIBS data in comparison to ANOVA and LGR filter methods [65].

Several researchers have proposed the use of ANN for quantitative analysis [24, 66–68], but they did not address the optimal selection of both the optimum number of neurons and wavelengths which contribute most to decrease the estimation error with simpler models. Here the need for parsimonious models with good predictive capabilities and few parameters, which can be achieved by using model selection criteria.

The main goal is to propose a methodology that can address the integration of expert knowledge and data driven algorithms for selecting the most informative wavelengths leading to low complexity models. It combines the prediction results based on manually chosen wavelengths resulting from spectroscopic expert often termed in this article as LIBS library data, and the ones obtained by frequently used methods such as: filter based techniques, penalizing regularization methods, dimensionality reduction methods and adaptive reweighted sampling methods. The optimization of both model and different variable selection methods is carried out using a penalized model selection criteria. For this reason, we selected each type of feature selection methods: KBest, LASSO, PCA and finally with CARS respectively as an illustration, and Akaike information criterion (AIC) as model selection criterion [69]. The AIC criterion has demonstrated to be more suitable than other information criteria for ANN regression, such as Bayesian information criterion, since it penalizes less the model complexity [70].

5.2 LIBS Setup



The LIBS set up, as seen in Fig. 5.1, has an Nd:YAG UV laser head (Ultra Quantel), 266 nm, with an energy of 25 mJ per pulse, 7 ns pulse width and a repetition rate of up to 10 Hz. It has an optical confocal coaxial system for laser focusing and collecting optical emission emanating from the plasma which was developed for laboratory analysis, as previously described in [71]. The focal lens distance is of 100 mm and reception of the emitted radiation was collected on different places on the surface of the sample with an optical fiber connected to the Aurora's spectrometer (Applied Spectra, CA, US). The spectrometer is a six channel spectrometer from 186-1049 nm, with spectral ranges covering UV-Vis-NIR. The spectral ranges for the six channels are 186-309, 309-460, 460-568, 568-672, 672-964, 964-1049 nm, with spectral resolutions of 0.1 to 0.12 nm (FWHM). A pulse/delay generator is used and the external Q-switch trigger system was activated and synchronized with the computer-controlled data acquisition system. Copper concentrate pellets are placed on XY translation stage to receive the plasma emission in Z-direction as shown in the Fig. 5.1.

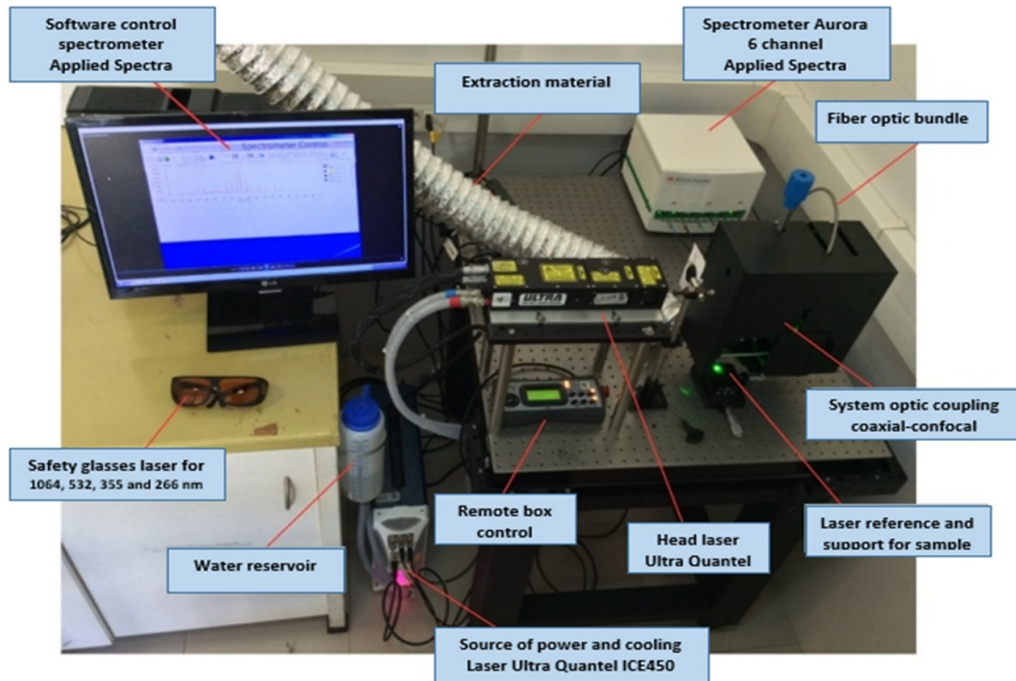


Fig. 5.1: Experimental arrangement for the LIBS acquisition of copper concentrate pellets

5.3 Sample Preparation

For LIBS measurements, homogeneous copper concentrate powder samples are collected from the processing plants, and a portion of 1000 mg is taken and pressed in form of pellet with 13 mm of diameter using a hydraulic pressing machine (CARVER) by applying 4 tons of pressure for about 2 min. This procedure allows a better efficiency of ablation for compact samples than powders, since the pressing produces dense granules with a flat and uniform surface to be exposed to the laser [72]. The collected LIBS data was obtained choosing 5 different locations of the surface and for each point 50 spectra were taken, accumulating a total of 250 shots per sample. As a criterion the first 5 spectra are eliminated, since these shots are considered as part of the surface cleaning, the remaining 45 spectra are averaged and are used for the calibration models. Delay and exposure time were adjusted to optimize the signal/background ratio, being these parameters set to 634.4 us and 1.05 ms, respectively.

5.4 Sample Treatment

One hundred and fifty three copper concentrate samples are obtained from different copper processing plants throughout Chile. The acid digestion treatment ($\text{HNO}_3/\text{HClO}_4/\text{HCl}$) is applied to the samples and the copper concentrations (%) are determined by atomic absorption spectroscopy (AAS) with a flame atomizer Analytikjena, novAA 400p (Jena, Germany). The methodologies and methods used for the characterization of these elements are validated with certified reference materials (OREAS 990 and OREAS 99b) provided by OREAS (North Vic, Australia). A subset of 131 fully labeled samples is used for this study. The elemental composition of copper, iron and arsenic are analyzed.

5.5 Wavelength and Model Selection Method

5.5.1 Spectroscopic Prior Knowledge

The quantitative estimation based on LIBS relies on detecting the emission lines amplitude of the elements present in the sample. Since these emission lines are known, they can be used to guide the selection of informative wavelengths for quantitative analysis.

For instance, copper concentrate samples contain mainly copper (Cu), iron (Fe), silicon (Si) and sulfur (S) as principal constituents. Arsenic (As) is a minor element, but very important for the further processing of the concentrates. With the help of atomic spectra database interface (ASD) for LIBS provided by the National Institute of Standards and Technology (NIST) several peaks associated with the emission lines are identified [71,73]. The present study is a regression case study of two main elements (Cu, Fe) and one minor element (As); however, it can be extended to all elements. The identified lines of copper, iron and arsenic are shown in Table 5.1.

In this work, the following definitions are used, and explained using copper wavelengths as an example:

- (i) the spectroscopic prior knowledge wavelengths from Table 1 of copper comprises 12 wavelengths are labelled as base LIBS library,

- (ii) the base LIBS library with a window of 3 wavelengths that constitutes 36 wavelengths in total are termed as extended LIBS library,
- (iii) If the base LIBS library is made to increase its window to span 11 wavelengths, which leads to 132 wavelengths in total are considered as full LIBS library data.

5.5.2 Proposed Methodology

The problem to be addressed is the concentration estimation of a known element in a given sample by means of LIBS and an artificial neural network. A three-layer ANN is enough to approximate any continuous function [74]. Thus the structure of the ANN model is:

$$o(i) = \mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{x}(i) + \mathbf{b}^{(1)}) + b^{(2)} \quad (5.1)$$

where $o(i) \in R$ is the estimated concentration associated to the input vector $\mathbf{x}(i) \in R^n$ representing the measured intensities of the emission at n wavelengths. The weight vectors are $\mathbf{W}^{(1)} \in R^{m \times n}$ and $\mathbf{W}^{(2)} \in R^{1 \times m}$, $\mathbf{b}^{(1)} \in R^m$ and $b^{(2)} \in R$ are biases, and σ is a tansig activation function, where n represents the number of wavelengths and m is the number of units in the hidden layer.

Thus, given a set of LIBS spectra collected from samples with known concentrations $\mathcal{D} = \{(y(i), \mathbf{x}(i)) \in R^{1+n} | i = 1, \dots, n_D\}$, where n_D is the number of samples, and a set of wavelengths corresponding to the element emission lines $\mathcal{E} = \{w_i \in R | i = 1, \dots, l\}$. The problem is to find the best number of hidden units and a subset of wavelengths from LIBS spectra in order to estimate the elemental concentrations, while avoiding an increase in model complexity and overfitting problems. The data set \mathcal{D} is divided into three data sets: training data set \mathcal{T} , validation \mathcal{V} , and test \mathcal{S} , where N_T, N_V, N_S are the number of elements of each set respectively. In order to tackle this problem, the methodology depicted in Fig. 5.2 is proposed along with the data set used in each step. This methodology combines the prior knowledge concerning the main emission lines and variable selection methods.

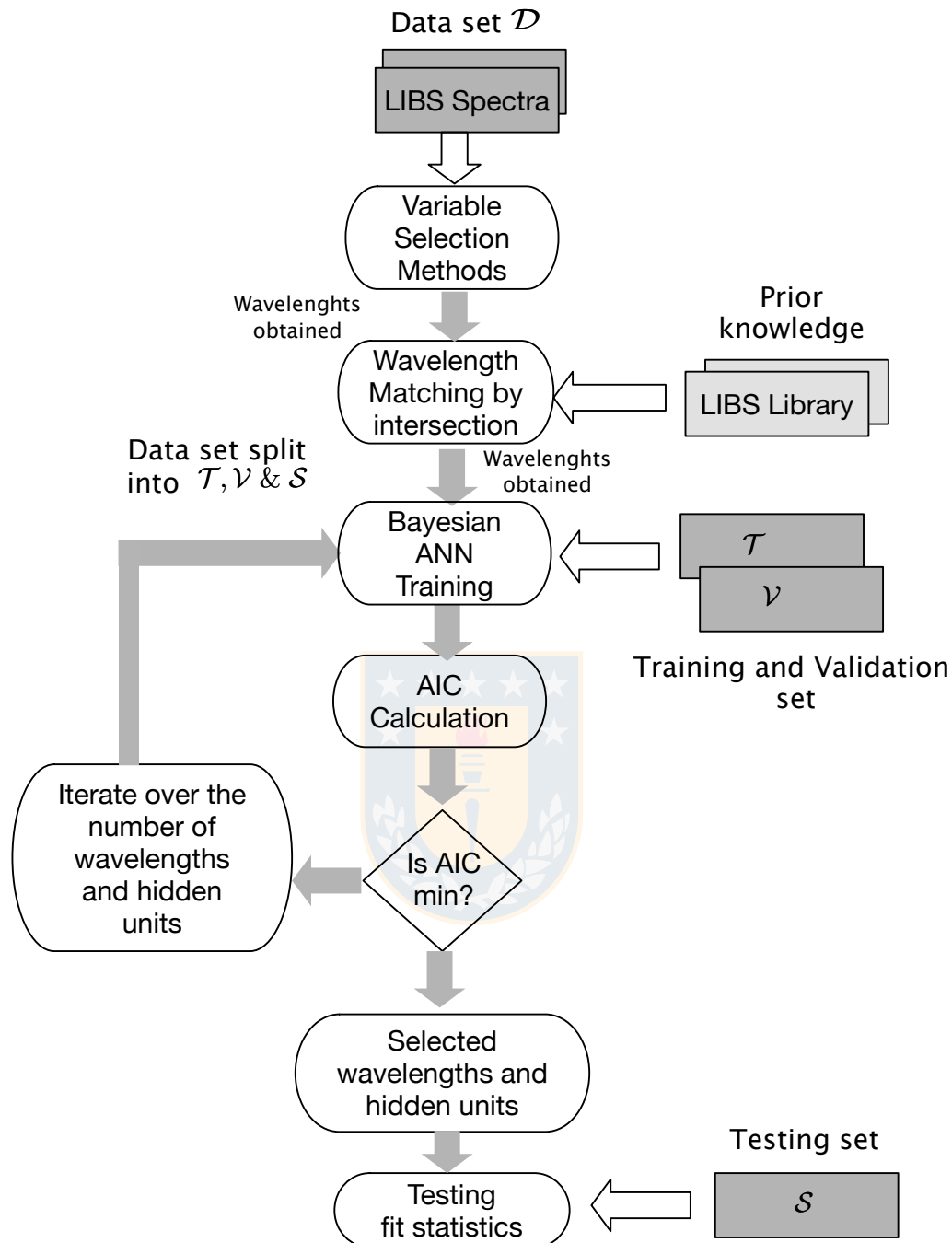


Fig. 5.2: Overview of the procedure for the selection of variables/wavelengths with different algorithms like KBest, LASSO regularization, CARS and PCA

The proposed methodology considers three steps:

First, from the full spectra, a variable selection algorithm is used to select the best features or wavelengths. These wavelengths are then ranked according to the respective algorithm's criterion.

Element	Line [nm]
Cu	324.76/327.41/458.70/465.10
	470.45/510.57/515.32/521.87
	529.22/569.98/578.24/809.56
Fe	234.30/238.20/239.50/240.40
	259.90/263.10/273.90/274.90
	275.50/358.10/364.70/373.70
	374.50/374.90/375.80/489.19
	492.21/495.73
As	228.82/278.02/286.05

Table 5.1: Copper, iron and arsenic elemental emission lines observed in the LIBS spectra of copper concentrate [3–12]

Secondly, the selected wavelengths from the first step pass through a wavelength matching process, where only those wavelengths inside the full LIBS library for the element are then selected.

Finally, a range from 2 wavelengths to the number of wavelengths at the extended LIBS library of the element, is used in the ANN regression model. Also, a range from 2 to 10 neurons is used in the ANN. The performance of the models is measured with the average AIC criterion from a 5-fold cross validation process, using a grid search, where the optimum number of wavelengths and neurons is finally chosen. AIC is defined in Eq. 5.2,

$$AIC = 2k + n_T \ln \left(\frac{RSS}{n_T} \right) \quad (5.2)$$

where $k = mn + m + n + 1$ is the number of estimated parameters in the ANN model, n is the number of model inputs; i.e. number of selected wavelengths, m is the number of neurons in the hidden layer, and RSS is the residual sum of squares, which is defined as in Eq. 5.3

$$RSS = \sum_{i=1}^{n_T} (y(i) - \hat{y}(i))^2 \quad (5.3)$$

where y is the measured value and \hat{y} is the predicted value.

As mentioned, different variable selection algorithms are considered: KBest, LASSO, PCA and CARS. The details of these methods are given in the following subsections.

5.5.3 KBest Algorithm

KBest belongs to the filter based variable selection algorithms [75], it is a univariate algorithm that selects the K best variables, according to a scoring function, for instance: ANOVA F-value or chi-square for classification problems, and F-value for regression problems.

For every input x_i , the correlation ρ_i with the output y is computed, this is:

$$\rho_i = \frac{(x_i - \bar{x}_i)(y - \bar{y})}{std(x_i)std(y)} \quad (5.4)$$

where

$$\bar{x}_i = \frac{1}{n_T} \sum_{j=1}^{n_T} x_i(j) \quad (5.5)$$

and

$$std(x_i) = \sqrt{\frac{1}{n_T} \sum_{j=1}^{n_T} (x_i(j) - \bar{x}_i)^2} \quad (5.6)$$

The F-statistic is computed:

$$F_i = \frac{\rho_i^2}{1 - \rho_i^2} (n_T - 2) \quad (5.7)$$

Finally, the highest F-values are selected by the KBest algorithm.

KBest algorithm is fast and has simple calculations. However, since it is an univariate linear method it can provide misleading estimates for multivariate and nonlinear problems.

5.5.4 LASSO Regularization

Regularization methods penalize input variables with less information, making the associated parameters tend to zero. Hence, if we apply regularization to an ANN model, we can select those variables which are not penalized. In LASSO (least absolute shrinkage and selection operator) regularization the estimated parameters (weights and biases) are those that minimize the following cost function:

$$\sum_{i=1}^{n_T} (y(i) - \hat{y}(i))^2 + \lambda \sum_{i=1}^m \sum_{j=1}^n |w_{ij}^{(1)}| \quad (5.8)$$

where λ is a regularization parameter. Once solved the optimization problem, the inputs x_i are ranked according to the values of $\sum_{j=1}^m |w_{ji}^{(1)}|$; i.e. the sum of the absolute values of weights $w_{ij}^{(1)}$ associated to input i [61].

LASSO algorithm is a nonlinear algorithm considering at the same time both the minimization of the approximation error and the selection of the input variable. The main disadvantages are the need of tuning the regularization parameters, may be slow depending of the network structure and it can lead to different results for different runs.

5.5.5 Principal Component Analysis

PCA is a data reduction technique that can be used to reduce the dimension of the input space by performing a data transformation of the original spectral data. However, it can also give insight about the informative wavelengths of a spectrum by inspecting the loading vectors [76, 77].

In PCA the original data matrix $\mathbf{X} \in R^{n_T \times n}$ is approximated by:

$$\mathbf{X} = \mathbf{T}\mathbf{L}^T + \mathbf{E} \quad (5.9)$$

where $\mathbf{T} \in R^{n_T \times n_C}$ is the score matrix, and $\mathbf{L} \in R^{n \times n_C}$ is the loading matrix, with n_C the principal components, and $\mathbf{E} \in R^{n_T \times n}$ is the residuals matrix. Most of the cumulative variance of the data is accumulated in the starting leading principal components (PCs) and very little variance in the trailing PCs. As a general practice, the first three loading vectors are carefully examined and wavelengths are selected based on the maximum values of the loading variables.

In this work, PCA is based on the covariance matrix, i.e., the input data is centered.

The main advantages of the PCA algorithm are its speed and simple calculation. The disadvantage stems from the fact that it is a linear method.

5.5.6 CARS Algorithm

The competitive adaptive reweighted sampling (CARS) is proposed in 2009 and related details can be found in [78]. Briefly, the working scheme of CARS method is as follows: primarily, create a PLS model and estimate the absolute values of coefficients of regression. Subsequently

in a competitive and iterative manner, N subsets of wavelengths are chosen via implementing Monte Carlo sampling established on the importance of each wavelength. Then, two algorithms, exponentially decreasing function (EDF) and adaptive reweighted sampling (ARS), are applied respectively to select the important wavelengths according to the coefficients of regression. At the end, using lowest RMSE, cross validation is employed to select the subset.

CARS algorithm is nonlinear method easy to setup for calculations; however, it may be slow and it can provide different results for different runs.

5.5.7 Figures of Merit

Four metrics are used in this work, namely the coefficient of determination (R^2), the root mean squared error of prediction (RMSEP), the residual sum of squares (RSS) and the Akaike information criterion (AIC). The two latter are previously defined, whereas R^2 and RMSEP are defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n_T} (y(i) - \hat{y}(i))^2}{\sum_{i=1}^{n_T} (y(i) - \bar{y})^2} \quad (5.10)$$

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_T} (y(i) - \hat{y}(i))^2}{n_T}} \quad (5.11)$$

where \hat{y} represents the predicted values, y the measured values and n_T the number of samples in the test set.

5.6 Experimental Results

In order to illustrate step by step the methodology, the estimation of copper concentration in pellets samples of copper concentrates varying with Cu in the range of 5 - 40% is considered. The dataset considers emission spectra in 185-1049 nm wavelength range (12287 wavelengths) obtained from 131 pellet samples of copper concentrates. In addition, the final results for iron and arsenic are also presented to demonstrate the validity of the methodology. Considering the non-ideal characteristics in the pellet samples, the main prior knowledge emission lines may be not sufficient to directly quantify copper, iron and arsenic concentrations, motivating the

proposed method for selecting other suitable sets of wavelengths. The experimental procedures for gathering the LIBS dataset is described in the following sections.

5.6.1 Software and Computing

KBest and PCA algorithms were implemented with the scikit-learn library of python platform, LASSO was implemented with the tensorflow library of python platform, whereas CARS algorithm was implemented with the MATLAB code provided in [79]. Finally, the ANN was implemented with MATLAB's neural network toolbox.

5.6.2 ANN Training

Mean squared error is used as the cost function, and Bayesian regularization is used in the training step [80]. The data is divided into three sets: 60% for training, 20% for validation and 20% for test. The training set is used to fit the model, the validation set is used to stop the training of the model to avoid overfitting and the test set is used to evaluate the model trained on the training set. A 5-fold cross-validation was used to split the data sets and obtain the network performances. Early stopping is used to prevent overfitting the training set along with regularization. The input and target data are normalized.

5.7 Results and Discussion

In this present quantitative study, the number of samples utilized was 131, which are considered to be limited for the division into two sample sets for the feature and regression studies; because it reduces the advantages of employing multivariate analysis. For instance, in the case of regression using ANN, the remaining samples would be the half. The splitting of these samples further into three sets for the purpose of training, validation and test sets, restricts approximately to 21 samples for each, where stratified data splits may not be possible i.e., the representation of concentrations across the regression curve may not be uniform. Since the algorithms and regression study, rely on the statistical number and true representation of samples, the division of the data has not been opted at this stage. However, in future practical field tests in industry the availability of a large number of samples may yield a better prediction. The work presented

here is to demonstrate the proof of concept through the prediction of elements such as copper, iron and arsenic.

The prior knowledge dataset considered a set of wavelengths for copper, iron and arsenic as summarized in Table 5.1.

We applied all the methods and optimized results using AIC for several elements Cu, Fe and As. However, we have limited the discussion mainly to copper element to understand the process. At the end, different related results for iron and arsenic were tabulated and commented.

For copper 132 wavelengths were used as prior knowledge to perform wavelength matching. This set is defined as the full LIBS library.

The variable selection algorithms were applied on the experimental data of LIBS to select multiple elements such as copper, iron and arsenic. No prior preprocessing have been performed. The obtained variables (wavelengths) are matched with the full LIBS library of data and the selected intersection of wavelengths from both of them. This step generates a ranking according to the matching criterion of multiple elements separately.

It is important to understand that the algorithms i.e., KBest, LASSO, PCA and CARS, cannot identify automatically copper elemental peaks, but rather calculate weights, scores or loadings based on the algorithm and supplied datasets. The scores of Kbest method show the relationship between the wavelengths and the concentrations. The higher the scores, the closer the dependence of the variables with that of the concentration. However it is likely to select wrong peaks other than copper element; to overcome this false data selection, the LIBS library data was used for sorting unrelated wavelengths as discussed in the flowchart in Figure 5.2 explicitly. From the remaining wavelengths based on the higher value of scores different wavelengths were chosen. It was observed that sometimes the algorithms may choose different wavelengths in the vicinity of elemental peak central wavelength, which also have spectroscopic significance as it is well known that a peak contains several wavelengths spread due to different broadening mechanisms in the plasma and as well as instrumental resolving power (spectral interferences).

Fig. 5.3 shows the scores obtained using the KBest algorithm. For a better visualization, the inserted diagram is drawn for explaining how the peak 521.89 peak is selected utilizing the score.

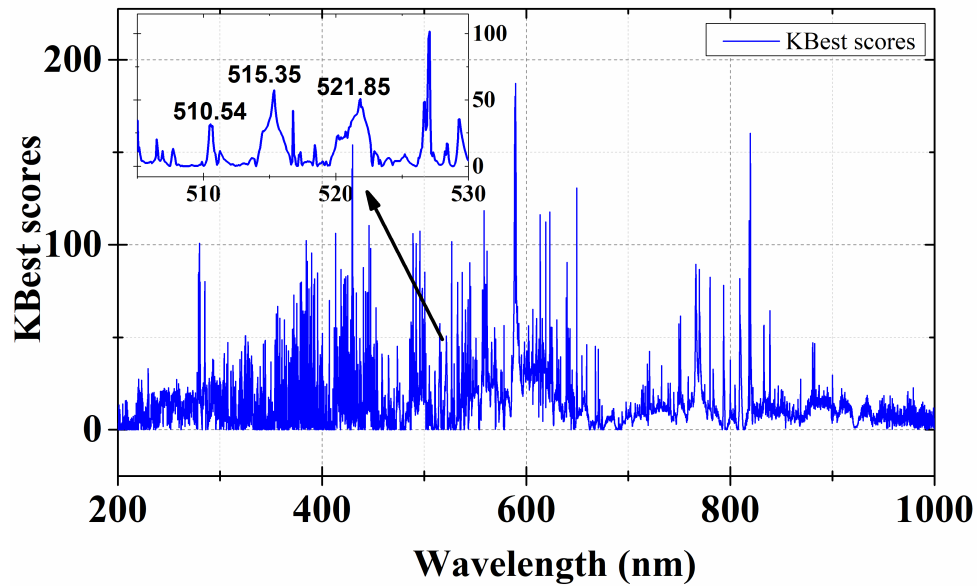


Fig. 5.3: Scores obtained using the KBest algorithm. The selection of a particular wavelength depends on the higher number of scores and correlated with the LIBS library data (enlarged picture)

In the case of PCA, the first three loadings features were studied for the associated elemental emissions of copper as shown in Fig. 5.5. As seen in the figure, the first loading has all the necessary peaks/wavelengths related to copper. For this reason, only the first loading is considered for wavelength selection; while the other loadings have information of Fe, Si and S. The cumulative explained variance achieved for the first eight PCs explained 99% variance as shown in Fig. 5.4. The first principal component accounted 69% variance. Like in the other methods. For PCA, from the first loading, the loading values are sorted with the wavelengths associated and copper elemental pixels were chosen with the help of the full LIBS library.

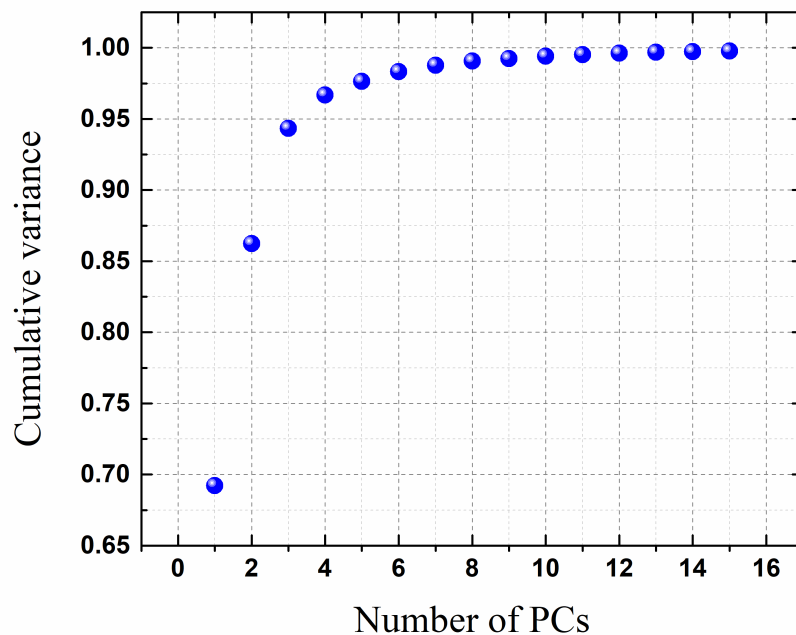


Fig. 5.4: Cumulative variance explained using number of principal components. A total of 8 principal components explain 99% of the variance cumulatively

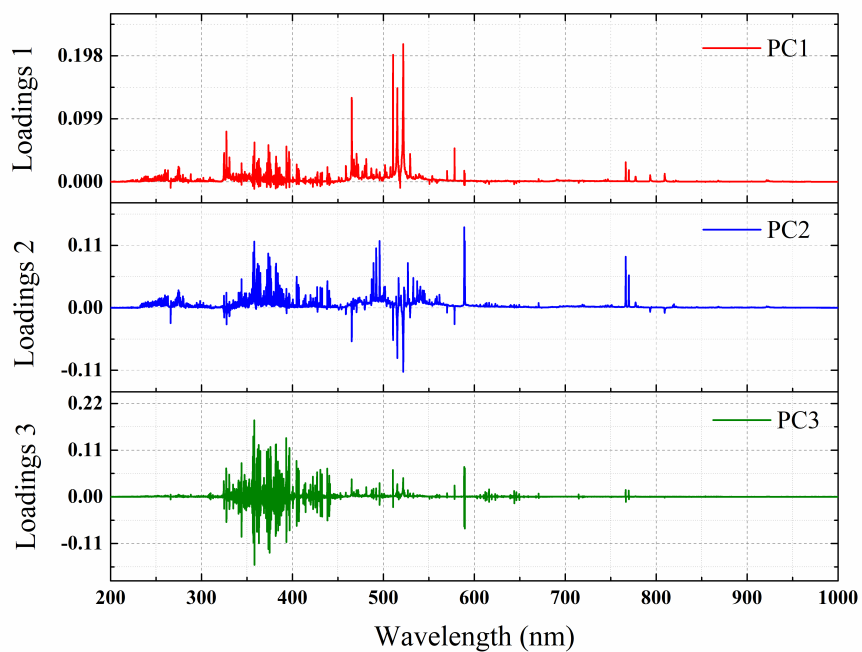


Fig. 5.5: Loadings plots of principal component analysis

For CARS, 14 latent variables were obtained for the initial Monte Carlo 5-fold cross val-

idation step with an optimum reached at 33 runs. CARS is designed to obtain an optimum number of variables without a cut-off number. In order to use it for this work, we obtained fixed numbers of selected variables from CARS weights. The 36 highest sums of the obtained weights in absolute value are plotted against the corresponding wavelengths in Fig. 5.6. The obtained output is similar to the KBest weights, but considers non-zero values only with the selected wavelengths.

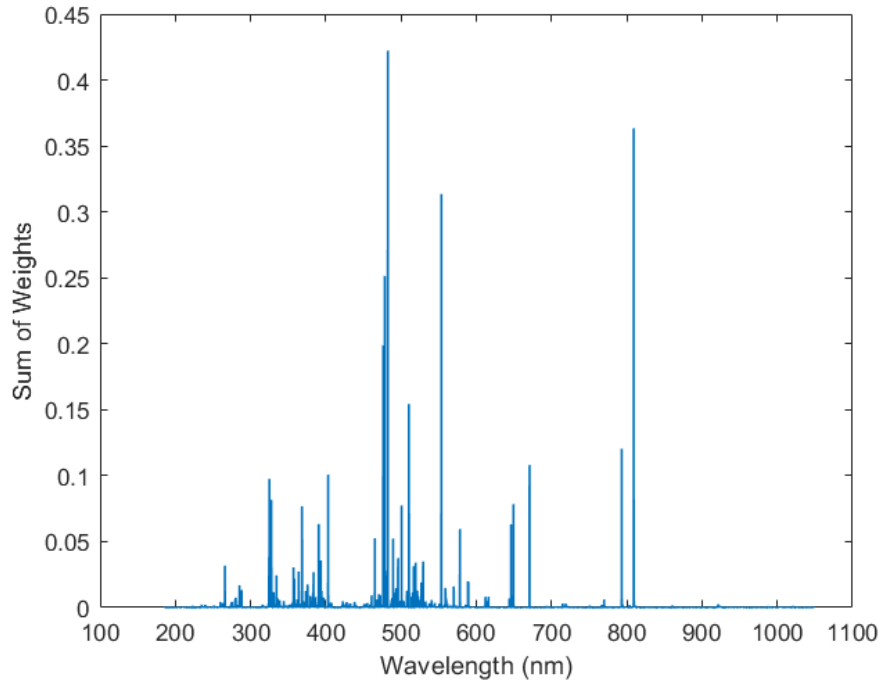


Fig. 5.6: Sum of weights obtained using CARS method

For LASSO, a fully-connected one hidden layer neural network with 10 neurons in the hidden layer was considered. In this case, there are 10 weights associated to every input. For every input, the sum of these 10 values (in absolute value) is calculated. These sums of weights are plotted against the corresponding wavelengths as seen in Fig. 5.7. Combining the LIBS library data and sorting, different number of variables are chosen similarly to the KBest method. The regularization parameter was set to 0.01.

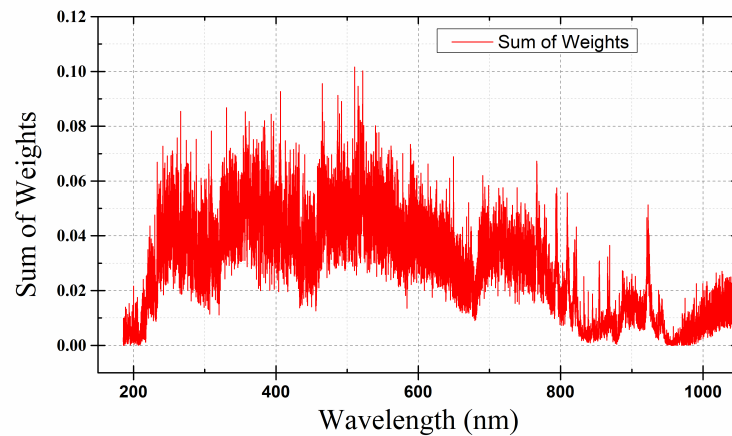


Fig. 5.7: Sum of weights obtained using the LASSO regularization method

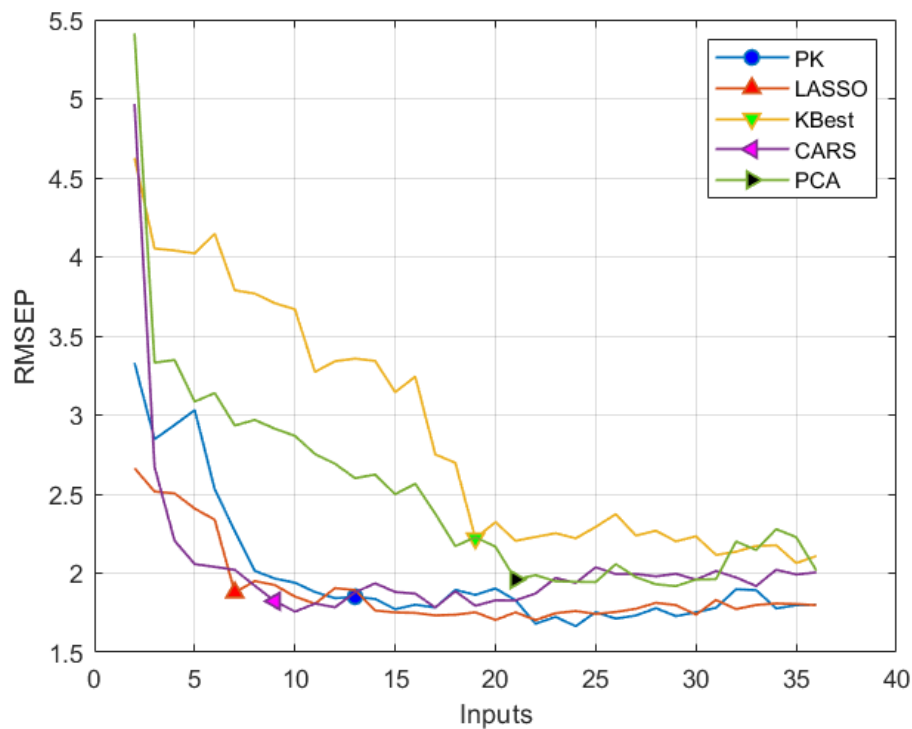


Fig. 5.8: RMSEP for the test dataset and the ANN with optimized number of hidden units and different number of inputs

To minimize the AIC an exhaustive grid search over the number of inputs and hidden units is carried out. The minimum AIC scores achieved for different number of hidden units and input variables are summarized in Table 5.2. As seen in the table, all the algorithms reach the minimum for two or three units in the hidden layer. The root mean squared error prediction calculated over the test dataset for the ANN with the optimal values found using AIC criterion

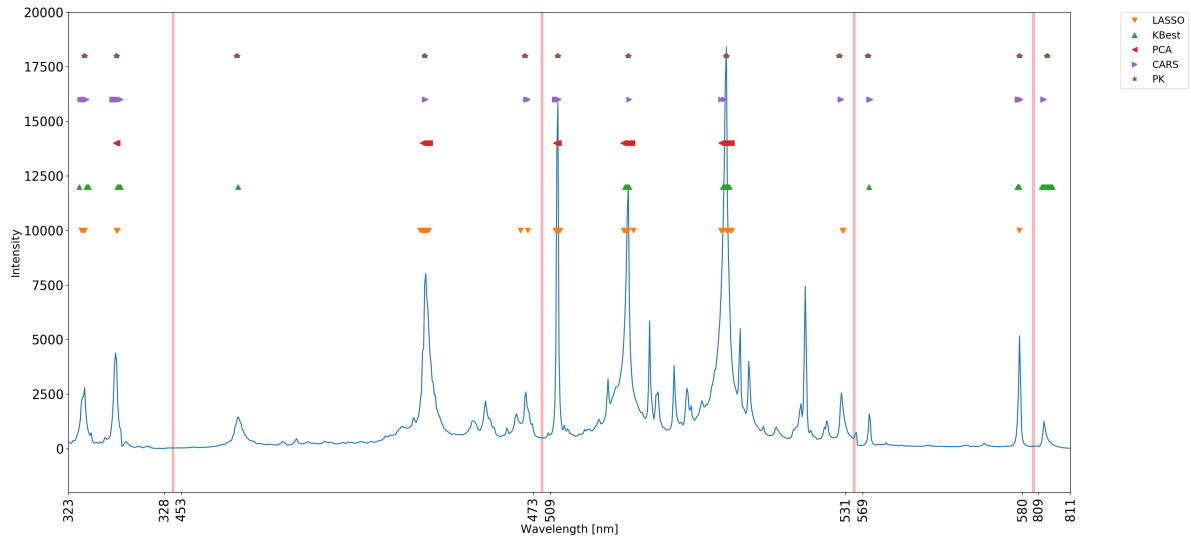


Fig. 5.9: Copper wavelengths selected by the methodology and the different variable selection algorithms and prior knowledge. The vertical red lines are used to show separate parts of the spectrum within the same figure

and different number of inputs are depicted in Fig. 5.8. These figures show that the RMSEP follow the minimum AIC criterion; i.e. the lowest RMSEP values are reached within the optimal values obtained by AIC.

The RMSEP as also seen in Fig. 5.8, shows that an increase in model complexity does not improve the RMSEP. The prior knowledge wavelengths were chosen according to the highest intensity peak values from the extended LIBS library. In the case of the prior knowledge approach the same methodology was applied to obtain the number of selected variables, meaning that a ranking was made for the wavelengths from the extended LIBS library. The wavelengths were ranked according first to the highest intensity peaks values and secondly according to the highest intensity values of the wavelengths immediately next to the peaks. A comparison among the different curves shows that the best performing algorithms are the non-linear ones; i.e. LASSO and CARS.

The best 36 selected wavelengths for the algorithms along with the full LIBS library wavelengths are shown in Fig. 5.9, in the case of copper. It can be observed that the algorithms select different wavelengths. For instance, KBest selects wavelengths clustered mostly in the region around 325 and 809 nm., which are considered less relevant wavelengths for copper, while using PCA, CARS or LASSO, the selected wavelengths are mostly on the highest peak values.

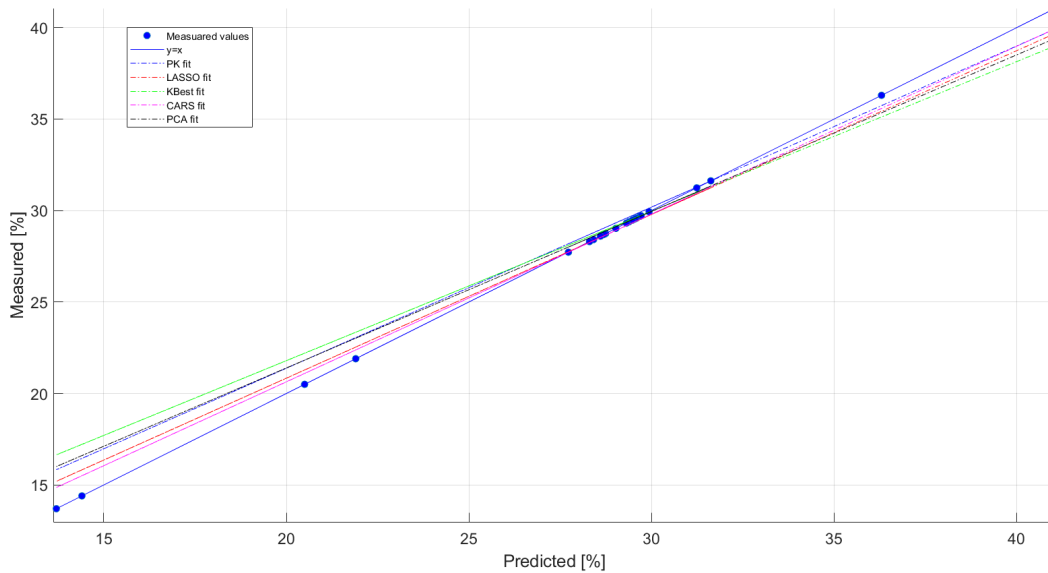


Fig. 5.10: Regression curves of PK (prior knowledge), LASSO, KBest, CARS and PCA algorithms for copper analysis fitted using a linear model $y=ax+b$. The corresponding figures of merits are provided in Table 5.2. The scatter plot of the measured values are the same values i.e., $y=x$ (linear), which are provided not only to depict all algorithms have a similar trend of the slope with the hypothetical line $y=x$ but also to visualize the stratified division of test data across the regression curve

Fig. 5.10 shows the regression plot for a particular fold for each method, for copper analysis, showing just the linear fits for each method and the measured values along the measured range for the test set. It can be seen that the best fits are obtained for the non-linear methods. The details of the linear models parameters for each algorithm are shown in Table 5.2. Comparing biases and weights (in linear model parameters slope and intercept) LASSO and CARS possess closely related values and also good agreement with AIC and other figures of merit.

From Table 5.2, it can be perceived that LASSO algorithm has an AIC, RMSEP, R2 and number of input wavelengths of 93.73, 1.72, 0.92 and 7 respectively, which are the best parameters for Cu elemental regression among all the other algorithms. However, the number of hidden neurons for all algorithms lies in between 2 to 3. We have tested our methodology by extending to other elements such as Fe and As. The regression fit statistics corresponding to Fe and As, using all wavelength selection algorithms and prior knowledge wavelengths are tabulated in Table 5.3 and Table 5.4. From Table 5.3, Fe regression fit statistics using all algorithms, it is evident that LASSO preserves the best method in terms of AIC, number of hidden neurons and number of input wavelengths. The other parameters, RMSEP and R2 are just lagging in the second decimal places compared to KBest method. The parameters such as RMSEP and

R2 seem to be very good when compared to other elemental regression results.

Table 5.2: Different model fit statistics in terms of parameters like AIC, RMSEP, R^2 , selected variables, number of neurons and linear model fitting parameters for different algorithms KBest, LASSO, PCA, CARS and prior knowledge for copper

	KBest	LASSO	PCA	CARS	P.K.
AIC	170.38	93.73	165.47	107.07	110.52
RMSEP (test)	2.37	1.72	2.17	1.90	1.80
R^2 (test)	0.85	0.92	0.88	0.91	0.91
Input variables	19	7	21	9	13
Number of hidden units	2	3	2	2	2
Slope (a)	0.8310	0.8951	0.8557	0.9170	0.8807
Intercept (b)	4.9983	2.9335	4.2802	2.2980	3.7667

Table 5.3: Different model fit statistics in terms of parameters like AIC, RMSEP, R^2 , selected variables and number of neurons for different algorithms KBest, LASSO, PCA, CARS and prior knowledge for iron

	KBest	LASSO	PCA	CARS	P.K.
AIC	135.30	128.22	142.86	137.70	141.87
RMSEP (test)	1.84	1.89	2.09	1.90	1.94
R^2 (test)	0.75	0.74	0.66	0.73	0.73
Input variables	12	6	8	8	10
Number of hidden units	2	2	2	2	2

The selection of As in this study serves as a good example for illustrating the estimation of a minor element with a concentration range of 0.09 – 4.2% and weak signals as compared with copper and iron. Under this less favorable experimental conditions the most sensible choice is to use prior knowledge. By analyzing Table 5.4 the methodology has indeed suggested this alternative. It is also worth pointing out that the RMSEP and R2 metrics are very good for all the variable selection algorithm compared to the other elemental regression results.

Table 5.4: Different model fit statistics in terms of parameters like AIC, RMSEP, R^2 , selected variables and number of neurons for different algorithms KBest, LASSO, PCA, CARS and prior knowledge for arsenic

	KBest	LASSO	PCA	CARS	P.K.
AIC	-329.91	-389.03	-354.80	-335.02	-408.90
RMSEP (test)	0.15	0.18	0.18	0.16	0.19
R^2 (test)	0.98	0.97	0.96	0.97	0.97
Input variables	9	4	7	9	7
Number of hidden units	3	4	4	2	4

Hence, we conclude that the combined optimization in terms of an information criterion of input variables and model complexity provides good results in terms of low complexity models and approximation accuracy.

5.8 Conclusions

This work has presented a systematic method to select the number of LIBS wavelengths and hidden units of an ANN model for regression analysis using LIBS spectra. The methodology combines prior knowledge and data driven variable selection algorithms to optimize model performance and complexity expressed in terms of the Akaike information criterion (AIC).

As illustrative example, the problem of copper, iron and arsenic elemental composition estimation by LIBS in pellets of copper concentrates is addressed. Four input selection methods; i.e. KBest, LASSO, PCA and CARS along with prior knowledge wavelengths, were tested within the methodology. The experimental results show that the proposed methodology is able to select a set of wavelengths to improve the accuracy, without increasing the model complexity. For this application under favorable experimental conditions; i.e. strong signals as in Cu and Fe, the combination of prior knowledge and LASSO, outperformed CARS, KBest, PCA and prior knowledge in terms of combined performance metrics and model complexity. On the other hand, under less favorable experimental conditions; i.e. weak signal, the most sensible option is to rely on prior knowledge, as was suggested by the methodology for As. In any case, the performance metrics are very similar among all the algorithms.

As far as the variable selection methods are concerned LASSO and CARS provided the least complex models for the three analyzed elements.

Finally, we conclude that the utilization of methodologies for obtaining parsimonious models is essential to obtain good predictive capabilities for the estimation of the elemental composition in copper concentrates with high dimensional data possessing large number of acquisitions.

6. An optimization approach to combine prior knowledge and LASSO regularization

Features (wavelengths) selection from high dimensional laser-induced breakdown spectroscopic data is highly gaining importance due to supervised and unsupervised tasks. Unlike conventional regression methods, neural network regressions (NNR) are much more flexible and adaptive to the variabilities produced due to experimental data fluctuations and sample properties like sample inhomogeneity in pellets. Even with the NNR method, proper selection of input parameters and selecting an appropriate choice of figures of merit is a critical task. We have proposed Akaike information criterion (AIC) as a model selection criterion, which was a promising approach for selecting an appropriate model along with different feature selection methods. As similar to all other wavelength selection LIBS articles, the wavelengths selected from the feature selection approach were compared with the LIBS library data, which is sometimes laborious, time consuming and may likely suffer from human error. On the contrary, in this chapter, we developed such as an algorithm within the framework of LASSO, which can automatically select wavelengths without human intervention of matching wavelengths. We modified the cost function by adding penalization over the LIBS libraries of spectral lines in the least absolute shrinkage and selection operator (LASSO) regularization with NNR. This approach avoids manual selection from LIBS library spectral lines and produces a model which is effectively robust, interpretable, and with better prediction capability.

6.1 Introduction

Laser-induced breakdown spectroscopy is essentially an optical emission spectroscopy. Due to the ease in system design, real-time analysis capability, portability, and standoff measurement ability, it has been used to determine the concentration of constituents of complex heterogeneous matrices like soils, minerals, and for molten metal analysis. The spectra generated using LIBS contain more variabilities due to excitation source fluctuations (laser), plasma fluctuations,

micro-level inhomogeneity in pellets, and self-absorption of lines. The total data generated from wide spectral detector channels are sometimes creating some spurious data correlations, noise from some specific line emissions, which may not contribute significantly to the analysis. Hence, it is important to select the features that can effectively bring some meaningful relationships with respect to the predictor variables like concentration. In general, for most cases of complex rock and soil LIBS data, the employment of least square models along with prior knowledge data sometimes may not yield satisfactory results in the prediction of constituent elements because the data generated by LIBS does not strictly follow linearity. The usage of variable selection and shrinkage methods can significantly enhance the prediction accuracies, and the models can effectively explain the data variability. The shrinkage method implies the inclusion of penalty terms in the least square regression methods which limit the coefficients towards zero. There are different kinds of methods which induce sparsity, such as ridge regression (RR), the least absolute shrinkage and selection operator (LASSO) regression, and elastic net regressions (ENR). In the ridge regression, the penalty term L2 includes squared terms, while in LASSO the penalty term L1 is in absolute modulus, and in elastic net is the combination of these two methods. The most important aspect of these methods is to shrink the coefficients towards zero, and ENR is more advantageous where it can retain all possible features. LASSO can be considered as a special case of ENR because ENR contains both lasso and ridge terms intrinsically.

LASSO is a nonlinear regularization algorithm, and it can be used for quantification and variable selection tasks. There is only some handful literature available using LASSO regression for LIBS data. The first quantitative report of comparison of LASSO and ENR was carried out with the data of rocks for the application of Mars planetary surface analysis. ENR was found to be pretty better than LASSO in terms of the root mean squared error predictions (RMSEP). The information from these findings not only provides elemental determination but also delivers channels selection for each atomic emission [81]. Further, in another article from the same research unit, partial least square regression (PLSR) and LASSO were compared with the LIBS data of 100 igneous rocks. Two variants of PLS viz., PLS1 and PLS2 were applied. The interpretation of PLS components with the LASSO(β) coefficients was interesting. They reported that the negative β coefficients are correlated with PLS anti-correlated values and vice versa. The information of overlap of lines and matrix effects explained by different wavelengths was well reported by LASSO. Thus, LASSO explained most of the underlying physical processes by selecting appropriate features that can deduce best intrinsic relation for predicting the concentration [82]. Brickley et al., reported the utilization of LASSO and sparse multivariate regression with covariance estimation (MRCE) for elemental emission wavelengths selection for

prediction of carbon [61]. Boucher et al. compared nine different regression models, including LASSO in search of interpretable regression methods considering igneous rocks LIBS data. From a broader perspective of results, linear models predicted major elemental concentrations with better prediction ability [83]. Ytsma et al. studied in different investigations utilizing LASSO and PLS for the prediction of H, Li, B, and S from rock standards [84, 85]. Chen et al. utilized PLSR and SVR methods for determining Cd content in lettuce, and LASSO was utilized for wavelengths associated with Cd elemental emission [86]. Erler et al. used PLS, LASSO, and GPR for soil nutrients detection. Out of them, LASSO and GPR yielded the best results [87]. Zhang et al. utilized LASSO for variable selection along with NNR and predicted Carbon concentration in steel [53]. Bertsimas et al. developed a regression algorithm based on mixed-integer optimization, compared the results with LASSO and sparse PLS [88]. Duarte et al. recently compared different variable selection methods, including LASSO where they selected the data from the prior knowledge LIBS library. They showed that the LASSO-based method has better variable selection capabilities, producing better prediction results in conjunction with neural network regression (NNR) [89].

In most of the articles discussed above, the authors chose the intersection of the selected wavelengths from the variable selection method and LIBS library data. To fill this gap of manual comparison process, we built such an algorithm where LASSO is utilized to select wavelengths automatically by penalizing LIBS library data inside the cost function, which has not been reported earlier. The selected wavelengths using the proposed method are then used as input data for an artificial neural network, where the performances are obtained based on KFold cross-validation. The optimum number of wavelengths and hidden units for the model are obtained by using the AIC criterion. The most critical aspect of this work is to implement a hard feature selection method that is fully focused or biased towards the selection of appropriate variables from the broad range of wavelengths of LIBS library data by reducing or ruling out matrix effect causing wavelengths. This approach reduces the manual search timing, overlapped wavelengths choices, and human errors.

6.2 Theory

6.2.1 Wavelength Selection Method and Algorithm

The use of LASSO regularization for wavelength selection considers the penalization of the weights associated to each input; i.e.

$$\sum_{i=1}^{n_T} (y(i) - o(i))^2 + \lambda \sum_{i=1}^m \sum_{j=1}^n |w_{ij}^{(1)}| \quad (6.1)$$

where $o(i) \in R$ is the estimated concentration modelled as an ANN with the following structure

$$o(i) = \mathbf{W}^{(2)} \sigma(\mathbf{W}^{(1)} \mathbf{x}(i) + \mathbf{b}^{(1)}) + b^{(2)} \quad (6.2)$$

In this model, the input vector $\mathbf{x}(i) \in R^n$ represents the measured intensities of the emission at n wavelengths. The weight vectors are $\mathbf{W}^{(1)} \in R^{m \times n}$ and $\mathbf{W}^{(2)} \in R^1 \times m$, $\mathbf{b}^{(1)} \in R^n$ and $b^{(2)} \in R$ are biases, and σ is an activation function, where n represents the number of wavelengths and m is the number of units in the hidden layer. The parameter λ is a regularization parameter.

An alternative approach proposed in [90] adds an extra parameter multiplying each input to the neural network model, as follows:

$$o_L(i) = \mathbf{W}^{(2)} \sigma(\mathbf{W}^{(1)} \mathbf{\Gamma} \mathbf{x}(i) + \mathbf{b}^{(1)}) + b^{(2)} \quad (6.3)$$

where $\mathbf{\Gamma} = \text{diag}[\gamma_1 \dots \gamma_n]$ are the multiplying factors. The cost function to be minimized is given by

$$\sum_{i=1}^{n_T} (y(i) - o_L(i))^2 + \lambda_L \sum_{j=1}^n |\gamma_j| \quad (6.4)$$

where λ_L is the regularization parameter. This cost function must be minimized with respect to the weights and multiplying factors.

The advantage of using Eq. 6.4 instead of Eq. 6.1 is that each variable is directly linked to only one parameter, which is particularly useful for the case of an ANN with several hidden units.

In order to take into account previous knowledge in terms of LIBS emission lines, this work proposes to add an additional penalization term over the parameter γ . Thus the following cost

function is proposed:

$$f = \sum_{i=1}^{n_T} (y(i) - o_{Lu}(i))^2 + \lambda_{Lu} \sum_{j=1}^n |\gamma_j| + \lambda_{PK} \sum_{k \in PK} (\gamma_k - 1)^2 \quad (6.5)$$

where PK is the set of wavelengths that fall inside the prior knowledge windows, λ_{PK} and λ_{Lu} are regularization parameters. In order to limit the effect of the multiplying parameters, a sigmoidal function of these parameters is considered in the model

$$o_{Lu}(i) = \mathbf{W}^{(2)} \sigma(\mathbf{W}^{(1)} \sigma(\mathbf{\Gamma}) \mathbf{x}(i) + \mathbf{b}^{(1)}) + b^{(2)} \quad (6.6)$$

It can be observed that Eq. 6.5 has a similar structure to that of the elastic net, where both LASSO and Ridge regularization methods are combined, however in this case, the squared term is restricted to a specific domain.

The proposed method aims to reduce the dimensionality of LIBS spectra considering prior knowledge data by adding an extra penalization on the cost function, discarding less relevant



wavelengths. The proposed method is described in Algorithm 1.

Algorithm 1: Proposed method for variable selection in LIBS

Step 1: Use wide spectral windows around the prior knowledge intensity peaks and discard overlapping wavelengths.

Step 2: Make S the set of wavelengths from Step 1.

Step 3:

while $length(S)$ is greater than S_{min} **do**

 Obtain Γ from solving the optimization problem associated to Eq. 6.5 using

$n = length(S)$

foreach $\gamma \in S$ **do**

if $\gamma < \gamma_0$ **then**

 discard from S the wavelength associated to γ

end

end

end

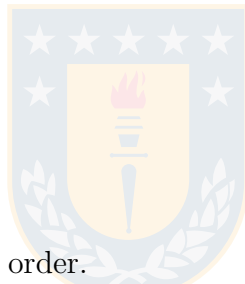
Step 4:

foreach $\gamma \in S$ **do**

$\delta = \gamma - 1$

end

Step 5: Sort $abs(\delta)$ in ascending order.



γ_0 is a threshold to discard less relevant wavelengths, which should be close to 0. λ_{Lu} and λ_{PK} are the regularization parameters considered in this work, and they can be selected using random search, grid search, or Bayesian optimization. In random search, some random hyperparameter values from a statistical distribution are chosen, and the model's performance is obtained for each combination of values. The best combination of values are finally returned [91]. In the case of the grid search, all possible values from a grid are chosen, and the best combination of these values are returned based on the model's performance [92]. These two hyperparameter tuning methods use individual experiments for each combination of values, meaning that the results are independent of each other. In Bayesian optimization, the results from previous runs of the optimization algorithm can be used as prior knowledge to obtain a posterior that attempts to approximate the model's objective function by using a surrogate model and an acquisition function, which proposes sampling points in the search space [93].

The use of wide spectral windows is for helping the algorithm to speed up its convergence.

In order to optimize Eq 6.5 subject to Eq 6.6, a projected scaled sub-gradient (Gafni-

Bertsekas variant) proposed in [94] is used. This algorithm requires the cost function and its derivatives with respect to the parameters:

$$\frac{\partial f}{\partial W^{(1)}} = 2(o_{Lu} - y)\dot{\sigma}(W^{(1)}\sigma(\mathbf{\Gamma})X)W^{(2)T}\sigma(\beta)X \quad (6.7)$$

$$\frac{\partial f}{\partial W^{(2)}} = 2(o_{Lu} - y)\sigma(W^{(1)}\sigma(\mathbf{\Gamma})X) \quad (6.8)$$

$$\frac{\partial f}{\partial \mathbf{\Gamma}} = 2(o_{Lu} - y)\dot{\sigma}(W^{(1)}\sigma(\mathbf{\Gamma})X)W^{(2)T}W^{(1)}\sigma(\mathbf{\Gamma}).(\mathbf{1} - \sigma(\mathbf{\Gamma}))X + 2\lambda_{PK}(\mathbf{\Gamma} - \mathbf{1}) \quad (6.9)$$

where $\mathbf{1}$ is a vector with ones.

Once the best wavelengths in S are obtained using Algorithm 1, the second and last step is to use these wavelengths as input for an ANN. The final result is an optimized set of wavelengths with an optimized number of hidden units in the ANN regression model. The average AIC criterion is used to assess the models' performance along with a grid search 5-fold cross-validation process, where the optimal number of wavelengths and hidden units are found.

6.2.2 Metrics

The Akaike Information Criterion (AIC) is proposed as a performance metric for this work. It typically penalizes less the model complexity, which tends to be high on ANN models. [70] AIC is defined in Eq. 6.10,

$$AIC = 2k + n_T \ln \left(\frac{RSS}{n_T} \right) \quad (6.10)$$

where $k = mn + m + n + 1$ is the total number of estimated parameters in the ANN, n is the number of inputs; i.e., number of selected wavelengths, m is the number of hidden units, and RSS is the residual sum of squares, defined in Eq. 6.11

$$RSS = \sum_{i=1}^{n_T} (y(i) - \hat{y}(i))^2 \quad (6.11)$$

where y is the measured concentration and \hat{y} is the predicted concentration.

Other metrics used in this work, namely the coefficient of determination (R^2) and the root mean squared error of prediction (RMSEP) are defined as:

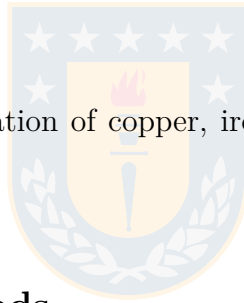
$$R^2 = 1 - \frac{\sum_{i=1}^{n_T} (y(i) - \hat{y}(i))^2}{\sum_{i=1}^{n_T} (y(i) - \bar{y})^2} \quad (6.12)$$

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_T} (y(i) - \hat{y}(i))^2}{n_T}} \quad (6.13)$$

where \hat{y} depicts the predicted concentration values, y the measured concentration values and n_T the number of samples that are in the test set.

6.3 Application

The application considers the estimation of copper, iron, and arsenic concentration in pellets samples of copper concentrates.



6.3.1 Material and Methods

Pellets samples of copper concentrates varying with Cu in the range of 5 - 40%, Fe in the range of 12 - 35%, and As in the range of 0.1 - 4.2% are considered, which were determined by atomic absorption spectroscopy and inductively coupled plasma - optical emission spectrometry. The dataset considers emission spectra in 185-1049 nm wavelength range (12287 wavelengths) obtained from 131 pellet samples of copper concentrates from different copper processing plants throughout Chile.

6.3.2 LIBS Setup

The LIBS setup used in this work is described in [89]. The same procedures and criteria for sample treatment were used in this case.

6.3.3 Software and Computing

The variable selection algorithm was implemented based on the work in [94], and the ANN was implemented with MATLAB's Neural Network Toolbox.

6.3.4 ANN Training

Bayesian regularization is utilized as an optimizer in the training step [80]. As in our previous work, the data is divided into three sets: 60% for training, 20% for validation, and 20% for test. 5-fold cross-validation is used to split the data sets and also to obtain the performances. Early stopping and regularization are utilized to prevent over-fitting of the training set. Both the input and target data are normalized.

6.4 Results and Discussion

Copper, iron, and arsenic concentrations are used for predictions in this work. For this study, the nearest ten wavelengths around the emission lines were used as prior knowledge windows.

In this work, 100 wavelengths-width windows were used as wide spectral windows. The spectral resolutions in this work range from 0.1 to 0.12 nm. S_{min} was set to 36 in the case of copper, 54 in the case of iron, and 9 in the case of arsenic, whereas β_0 was set to 0.25, which helps discard less relevant wavelengths.

For implementing the variable selection algorithm, a multi-layer perceptron neural network with 5 neurons in the hidden layer was used. The values used for the regularization parameters were obtained through an exhaustive grid search with values $\lambda_{Lu} = \{1, 5, 10\}$ and $\lambda_{PK} = \{0, 5, 10, 30, 50, 80, 100, 150\}$. In all cases, the optimum values were $\lambda_{Lu} = 1$ and $\lambda_{PK} = 50$. Fig. 6.1 shows the variation of AIC with respect to λ_{PK} for the three elements, where λ_{Lu} was fixed to 1.

The results from the proposed method (referred as $LASSO_2$) are compared with those of the direct use of prior knowledge wavelengths and those of our previous work using LASSO (referred as $LASSO_1$). In order to make this comparison, the 36 best copper wavelengths for all methods were ranked according to in the case of the newly proposed method to the lowest values

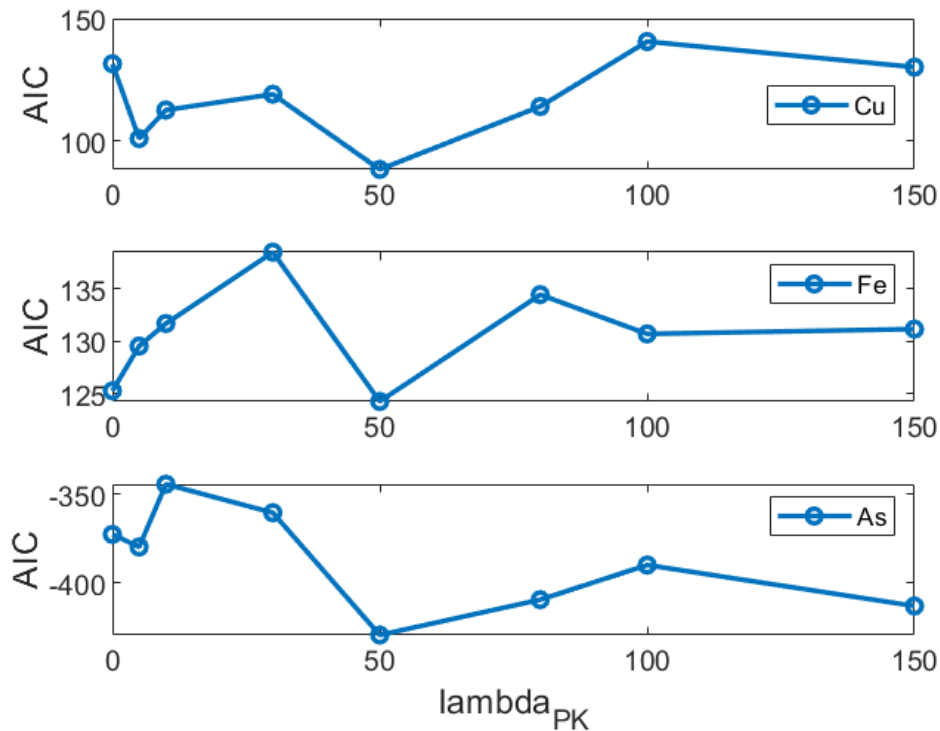


Fig. 6.1: AIC with respect to different λ_{PK} values, with $\lambda_{Lu} = 1$

of $\text{abs}(\delta)$, as stated in Step 5 of Algorithm 1. In the case of prior knowledge, the wavelengths were ranked according first to the highest intensity peaks values and secondly according to the highest intensity values of the wavelengths immediately next to the peaks. In all cases, the data sets used exactly the same data splits using 5-fold cross-validation, and for the ANN, the same random seeds were used. We utilized a range from 2 to S_{min} wavelengths as inputs for the ANN model. In addition, a range from 2 to 10 neurons is used.

The metrics for the three methods in the case of copper are summarized in Table 6.1. As seen in this table, both $LASSO_2$ and prior knowledge methods reached a minimum AIC in 2 neurons and 13 input variables. In terms of AIC, the newly proposed method outperforms the prior knowledge approach, and in terms of the fit, the results are slightly better for the newly proposed method with LASSO.

Fig. 6.2 shows the RMSEP over different numbers of inputs for the newly proposed method and the prior knowledge-based method. The markers show the RMSEP for the optimized number of inputs based on AIC. The two trends are similar over the number of inputs, obtaining both methods a minimum RMSEP at 13 inputs.

Fig. 6.5 shows the best 36 copper wavelengths selected by the new method with LASSO and

by prior knowledge. It can be observed that the wavelengths are picked from different regions, however many are clustered in the region around 325 and 529 nm, which are considered less relevant wavelengths for copper, and explain the higher RMSEP of the proposed method when using larger numbers of input variables in comparison to the prior knowledge approach. In the end, when using the optimized number of inputs by AIC, the proposed method outperforms the prior knowledge approach.

The same procedure is repeated in the case of iron and arsenic. Table 6.2 shows the metrics for all methods in the case of iron. Both $LASSO_2$ and prior knowledge methods reached a minimum AIC in 2 neurons, however the new method with LASSO uses fewer input variables and obtains a similar performance in terms of the goodness of fit, implying that the AIC is lower in comparison to the prior knowledge method.

In the case of arsenic, Table 6.3 shows the metrics for all methods. The prior knowledge method obtains its best results at 7 input variables and 4 neurons, whereas the newly proposed method uses 8 input variables and 3 neurons. In terms of the goodness of fit, the newly proposed method obtains slightly better results, whereas, for the AIC, the newly proposed method with LASSO outperforms the prior knowledge method.

Fig. 6.3 and Fig. 6.4 show the RMSEP over different numbers of inputs for the new proposed and prior knowledge methods for iron and arsenic, respectively. The markers indicate the RMSEP for the optimized number of inputs based on AIC. In the case of iron, the newly proposed method starts with lower RMSEP values, which are increased as the number of inputs is higher. High values of RMSEP can be found for 25 or more inputs, especially on the newly proposed method, which obtains its minimum RMSEP at 4 inputs, whereas the prior knowledge-based method has its minimum RMSEP at 10 inputs. In the case of arsenic, the RMSEP values for the newly proposed method are always lower than those of the prior knowledge-based method, obtaining its minimum RMSEP value at 8 inputs, whereas for the prior knowledge-based method is obtained at 7 inputs.

Fig. 6.6 shows the best 54 iron wavelengths selected by the new method with LASSO and by prior knowledge, where it can be observed that many wavelengths are selected from regions with low values of intensity, which may explain the high RMSEP values as seen in Fig 6.3.

Finally, Fig. 6.7 shows the best 9 arsenic wavelengths selected by the new and prior knowledge methods. It can be observed that the selected wavelengths are more spread out throughout the peaks, which is the most noticeable difference between the two methods.

Table 6.1: Performance metrics for the newly proposed method, the previously proposed method, and prior knowledge for copper

	LASSO ₂	LASSO ₁	P.K.
AIC	88.50	93.73	110.52
RMSEP (test)	1.60	1.72	1.80
R^2 (test)	0.93	0.92	0.91
Input variables	13	7	13
Number of hidden units	2	3	2

Table 6.2: Performance metrics for the newly proposed method, the previously proposed method, and prior knowledge for iron

	LASSO ₂	LASSO ₁	P.K.
AIC	124.29	128.22	141.87
RMSEP (test)	1.91	1.89	1.94
R^2 (test)	0.73	0.74	0.73
Input variables	4	6	10
Number of hidden units	2	2	2

Table 6.3: Performance metrics for the newly proposed method, the previously proposed method, and prior knowledge for arsenic

	LASSO ₂	LASSO ₁	P.K.
AIC	-429.29	-389.03	-408.90
RMSEP (test)	0.13	0.18	0.19
R^2 (test)	0.98	0.97	0.97
Input variables	8	4	7
Number of hidden units	3	4	4

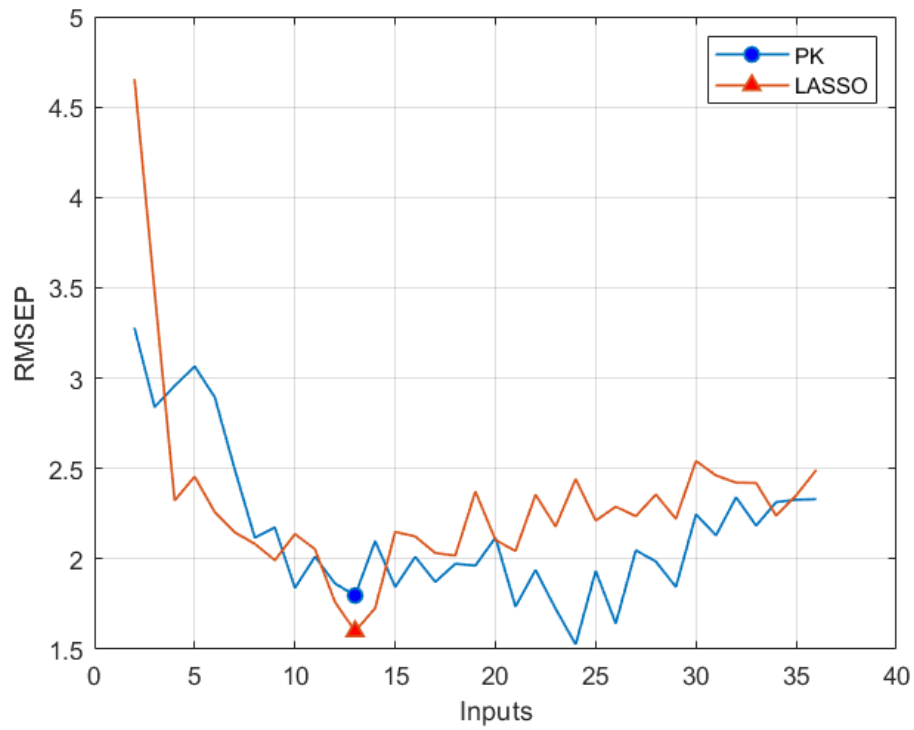


Fig. 6.2: RMSEP for the test dataset and the ANN with an optimized number of hidden units for copper

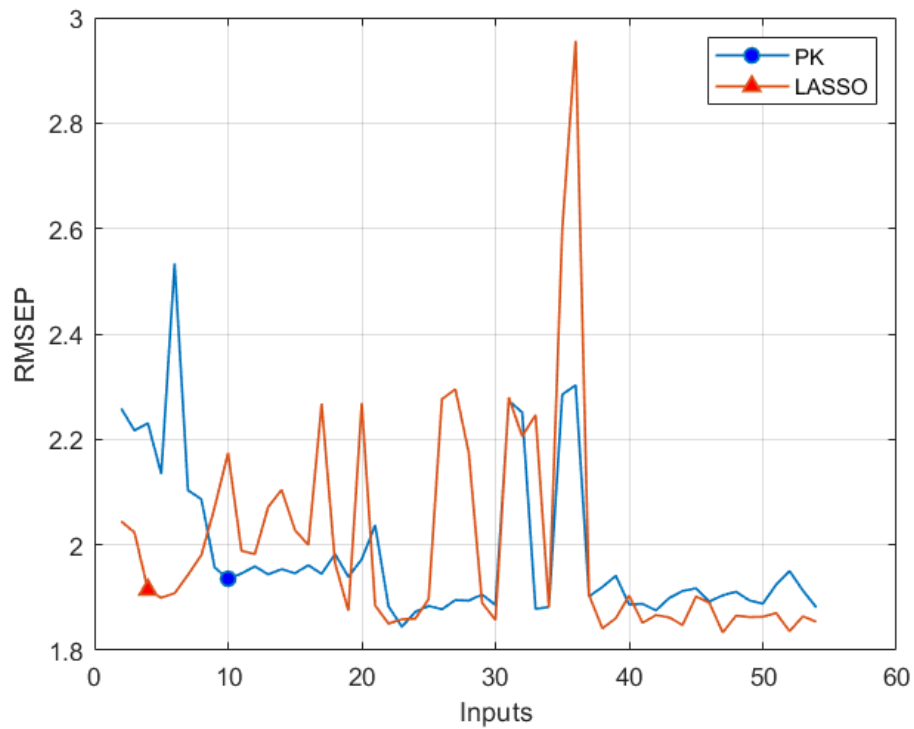


Fig. 6.3: RMSEP for the test dataset and the ANN with an optimized number of hidden units for iron

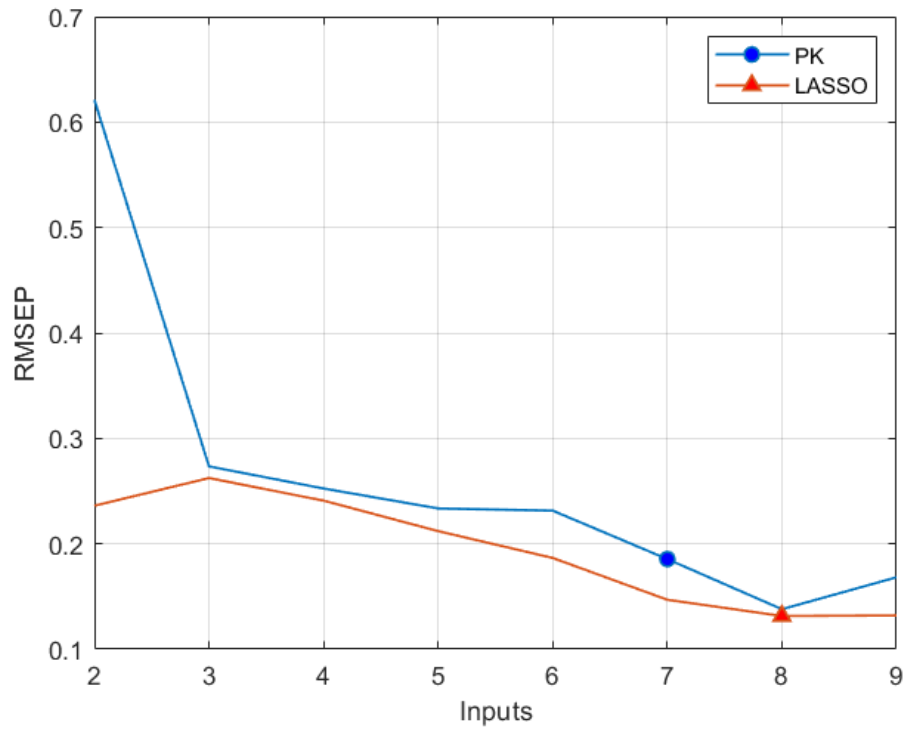


Fig. 6.4: RMSEP for the test dataset and the ANN with an optimized number of hidden units for arsenic

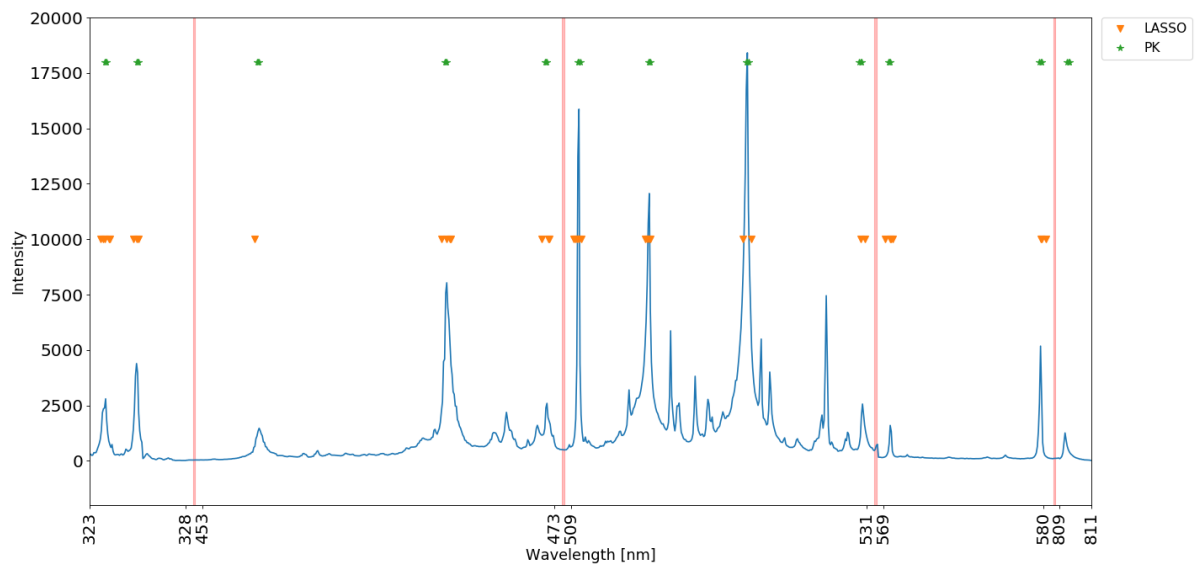


Fig. 6.5: Copper wavelengths selected by the method and by prior knowledge

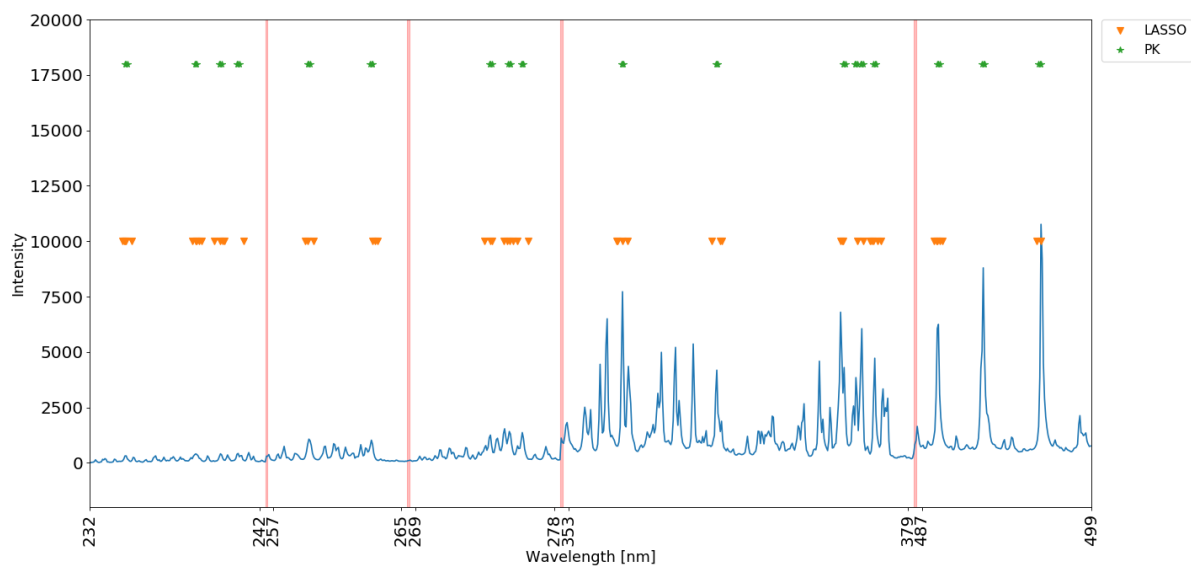


Fig. 6.6: Iron wavelengths selected by the method and by prior knowledge

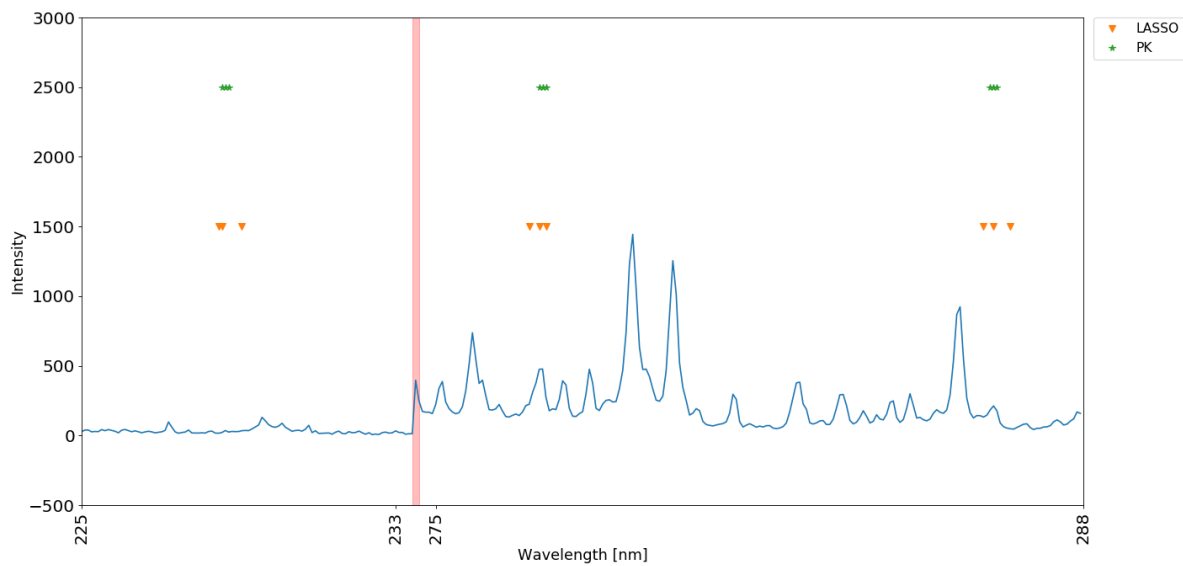


Fig. 6.7: Arsenic wavelengths selected by the method and by prior knowledge

6.5 Conclusions

This Chapter has proposed a new algorithm with the basis of spectroscopic significance for deselection of spurious wavelengths in LIBS data, which are not related to elements of interest. Using an iterative process, the proposed method can reduce the dimensionality of LIBS spectra by discarding wavelengths that are less relevant according to prior knowledge data. The challenge of estimating copper, iron, and arsenic concentrations in pellets of copper concentrates using LIBS was addressed as an example in this study. Based on the Akaike information criterion, the experimental results suggest that the proposed strategy can improve results above those achieved using merely a prior knowledge approach.



7. Data fusion of LIBS-DRS and LIBS-HSI for improved analysis of mineral species in copper concentrates

Copper concentrates are an intermediate product of the copper production process, which can be sold or further processed by pyrometallurgical and electrolytic processes to obtain copper with a grade of 99.99%. The mineralogical analysis of these concentrates is very important for quality control and monitoring the feed of smelters. The use of spectroscopy provides a means to carry out this analysis by measuring the reflected or emitted spectra when the samples are illuminated by different light sources. Diffuse Reflectance (DRS), Hyperspectral Imaging (HSI) and Laser Induced Breakdown Spectroscopy (LIBS) provide key complementary information for identifying and quantifying mineralogical components. Combining the information of these spectra is crucial to perform the measurements at the same spatial domain. This contribution presents the analysis of copper concentrates using a multi-purpose set-up, measuring spectra of LIBS-DRS as well as LIBS-HSI, at the same spatial position, and the use of data fusion techniques for blending the information. Low and mid-level data fusion strategies are compared in terms of their effectiveness. In this application, the experimental results show that mid-level data fusion provides the best result outperforming the predictions done by the separated information sources. These results indicate that the spectral measurements obtained by LIBS-DRS and LIBS-HSI from the same measuring point with high spatial resolution can be combined for enhancing the mineralogical analysis of copper concentrates, which is tested using the following mineral species: bornite (Cu_5FeS_4), chalcopyrite ($CuFeS_2$), covellite (CuS), enargite (Cu_3AsS_4) and pyrite (FeS_2).

7.1 Introduction

The analysis of copper concentrates is very important from an operational and economical point of view. If the copper concentrates are sold, the quality control of the product is an integral part of the purchase agreement, where off control parameters are heavily penalized.

On the other hand, if the concentrate is further processed by smelters, its analysis will provide key information for controlling the process. The traditional analysis is done by dissolving the samples using acids and then the chemical analysis is carried out over the solution. These procedures require time consuming sample preparation and analysis. In recent years, Laser Induced Breakdown Spectroscopy (LIBS) has been proposed to analyze different mineral samples to measure their elemental composition. A review of the recent advances concerning the use of LIBS for identifying and analyzing rocks and minerals can be found in [95]. This technique can be used in-situ for qualitative and quantitative analysis of mineral ore [96] [97] and remotely [98]. The online analysis of bulk materials has been demonstrated in [99] and [100]. Recently in [101], the method has extended to estimate the mineral composition of rock samples.

LIBS is a plasma optical emission spectroscopic technique, where a high-power laser is focused to create plasma on the sample. The optical electromagnetic radiation emitted from plasma provides atomic and molecular information of the sample. In contrast, DRS does not utilize a high-power laser source. It uses simple broadband light sources to illuminate the sample. The diffuse reflectance spectra is a result of a combination of different processes, like scattering and absorption, and it provides rich spectral information.

Early work on the reflectance of ore minerals can be found in [27], where the reflection characteristics of about 200 ore mineral species were summarized. Criddle and Standley [102] provide a compilation of reflectance standards and the measurement of reflectance properties of all known ore minerals. Based on this knowledge several authors have proposed systems for analyzing ore samples based on reflectance. Pirard describes a multispectral system at a microscopic scale for ore analysis [28]. Catalina and Castroviejo describe a system for the automated microscopic characterization of ores with 20 reflectance bands [29]. This system is used to classify ore species based on reflectance measurements and different classification algorithms [30]. The use of reflectance has been taken a step further by using hyperspectral cameras for online estimation of run-of-mine ore composition on conveyor belts [31]. These works show that reflectance spectrum can provide mineralogical information, but in some situations, the discrimination between mineralogical species having similar spectral signatures can be challenging.

Data fusion is a technique to enhance the effectiveness of predictive models by combining different sources of information about the sample. The fusion of information can be carried out at three different levels. The low-level considers the use of concatenated data set, while the mid-level strategy concatenates features and high-level strategy merges the results of individual predictive models [103]. In [38] low-level and mid-level data fusion are used to classify ochre

pigments using Raman and X-ray fluorescence data. In [39] mid-level and high-level data fusion are used to classify Sudan dyes using UV-Vis and HNMR data. The low-level fusion of Raman and LIBS spectra to discriminate geological specimens is described in [104]. Low-level data fusion of LIBS and Raman data can improve the identification of sulfates and salts in a basaltic matrix [105]. Low and mid-level data fusion strategies considering mid-wave infrared (MWIR) and long-wave infrared (LWIR) spectra have been proposed to improve the prediction accuracy of SiO_2 , Al_2O_3 , and Fe_2O_3 concentrations in a polymetallic sulphide deposit [106]. The classification of 21 edible salts products by fusing DRS and LIBS information is described in [107]. In this application, the optimal mid-level fused model, based on a linear combination of Principal Component scores, performed better than the individual models enabling the accurate classification of edible salt samples. In all these works data fusion was able to provide better results than the individual data sources.

One of the key issues for using the information of LIBS and DRS is the need of acquiring the spectra from the same spatial location of the sample. To address this issue, a set-up has been designed for performing both measurements almost simultaneously and avoiding sample movements. Besides, this work explores data fusion for improving the regression analyses of copper concentrate samples.

7.2 LIBS-DRS Setup

This work proposes a new optical coupling system to perform simultaneously LIBS and DR measurements. The system consists of a coaxial-confocal design to increase the light irradiance for the generation of the plasma, subsequent ablation, and capture the emitted light signal from the plasma and subsequent atomic emission. The system is coaxial because the focusing and light collecting parts share the same optical axis of propagation; and it is confocal since the image plane of the focusing system is the plane object of the collection system. This optical arrangement duplicates images in three images: the close-up image for detection by coupling to the optical fiber of the polychromator; a second image plane for visible inspection camera; and finally a third image plane for a white projection system, useful for aligning and centering the samples. Thereby, all the planes object-image of the system are finite conjugates and of revolution. Therefore, the optical adjustment is identical for the ablation and detection, visualization and alignment, which is fulfilled automatically by the optical design. The DRS/LIBS set up, as seen in Fig. 7.1, has an Nd:YAG UV laser head (Ultra Quantel, France), 266 nm,

with an energy of 25 mJ per pulse, 7 ns pulse width and a repetition rate of up to 10 Hz. A similar optical coaxial system was utilized earlier for LIBS studies explicitly described previously in [101] and [108]. The focal lens distance is 100 mm and the emitted radiation was collected by an optical fiber connected to an Aurora's spectrometer (Applied Spectra, CA, US). The LIBS spectrometer is a six channel spectrometer from 186-1049 nm, with spectral ranges covering UV-Vis-NIR. The spectral ranges for the six channels are 186-309, 309-460, 460-568, 568-672, 672-964, 964-1049 nm, with spectral resolutions of 0.1 to 0.12 nm (FWHM). A pulse/delay generator is used and the external Q-switch trigger system was activated and synchronized with the computer-controlled data acquisition system.

Copper concentrate pellets are placed on XY translation stage to receive the plasma emission in Z-direction as shown in Fig. 7.1. A fiber array, as depicted in Fig. 7.2, was implemented to collect the light coming from the sample and to illuminate the sample with a Tungsten Halogen light source.

To perform the DRS measurements, firstly the sample is illuminated by a lamp and the reflectance is measured using a single channel spectrometer (Ocean Optic HR-2000, 190-1100 nm, 0.9 nm FWHM). Then, the laser pulse radiation is concentrated in the image plane, using a bi-convex lens, increasing the irradiance on the sample above 1×10^{13} W/cm², fulfilling the condition of laser ablation formation and atomic emission; if the sample was located within the ablation length, a characteristic sound of the ablation will be produced, whose intensity and duration will be recorded by the microphone, which will confirm the success or failure of the detection of the laser pulse. Furthermore, as reported in [109], the acoustic signals can provide additional information since they are correlated with the sample hardness/density and the volume of the crater.

Immediately the detection occurs because the light was guided by the optical fiber towards the polychromator. Subsequently, visual confirmation can be obtained, through an image camera.

7.3 Sample Preparation and Treatment

Twenty homogeneous copper concentrate powder samples are collected from different processing plants. These samples have natural variations of the minerals that are present in the copper concentrate as pyrite, chalcopyrite, bornite, as major minerals. The particle size distribution

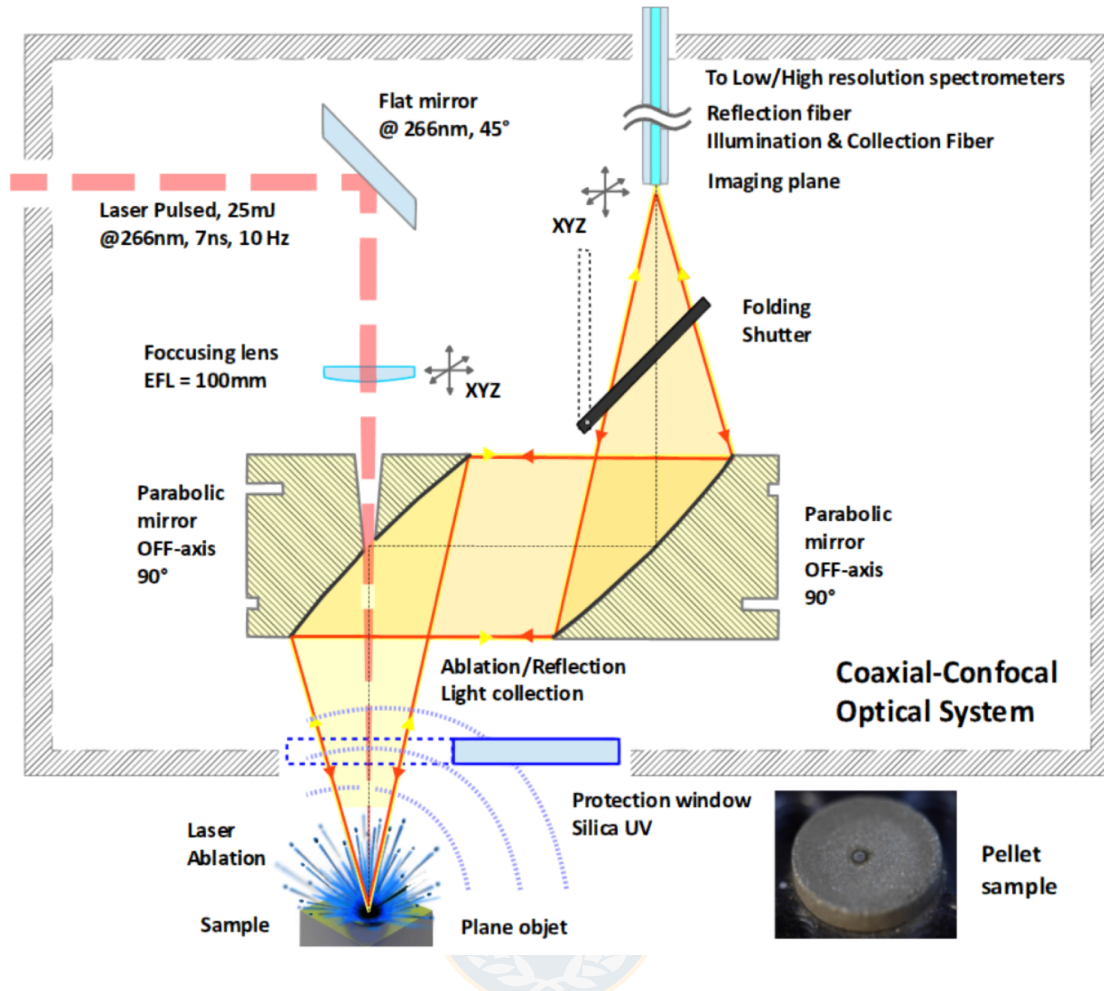


Fig. 7.1: LIBS-DRS setup

of samples was measured by light scattering analysis in Analysette 22 MicroTec plus (Fritsch, USA). The particle size distribution of the copper concentrate shows that the 96.4% of the particles have a size of less than $74 \mu\text{m}$. The samples were characterized mineralogically to identify the phases that are present in the copper concentrate by Quantitative Evaluation of Minerals by Scanning Electron Microscopy (QEMSCAN) (ThermoFisher, FEG Quanta 650, USA). The particle size was measured in Analyzer Frisch Analysette 22 (Idar-Oberstein, Germany). The values obtained by QEMSCAN were used as a gold standard method for the multivariate methods.

Portions of 1000 mg are taken and pressed in the form of a pellet with 13 mm of diameter using a hydraulic pressing machine, as shown in Fig. 7.3. The collected LIBS and DR data were obtained from a single location of the surface. LIBS spectral data is depicted in Fig. 7.4, and Fig. 7.5 shows the DR spectra. Both have very distinctive features; LIBS being mostly a discontinuous spectrum while DR is continuous. For LIBS 50 spectra were taken, accumulating

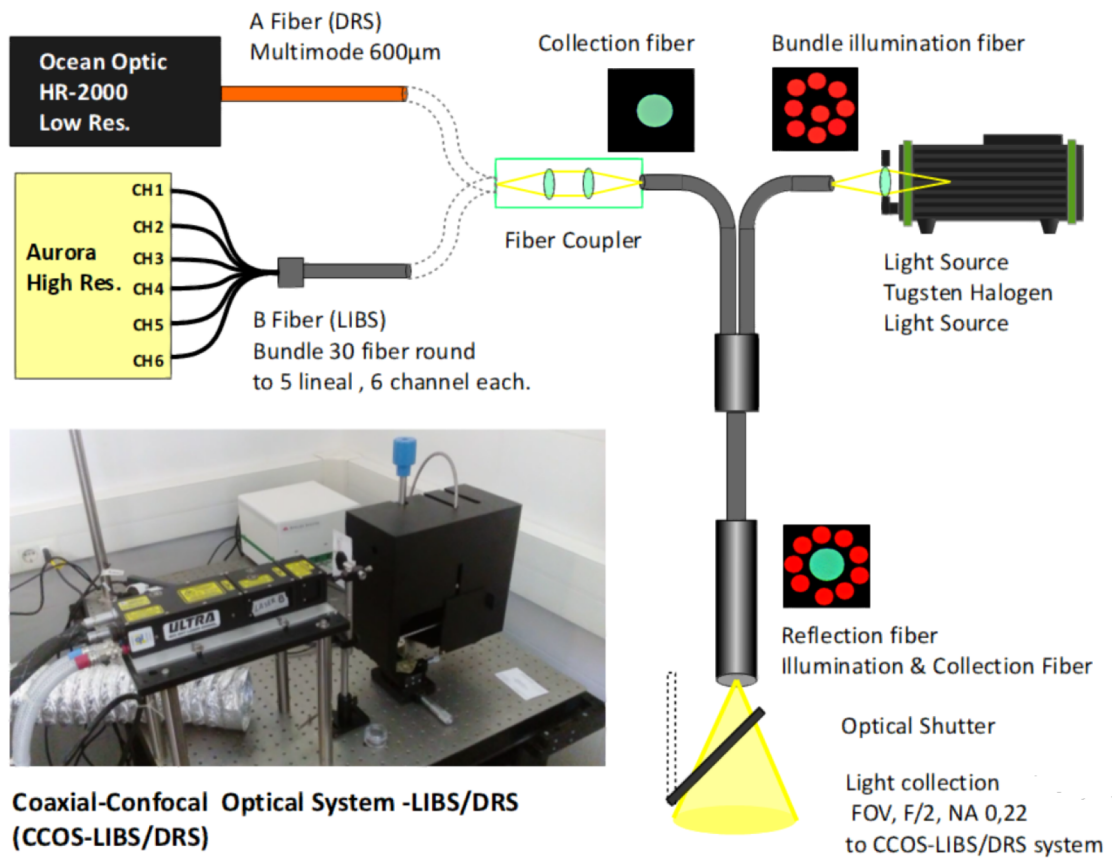


Fig. 7.2: LIBS-DRS setup: Fiber array

a total of 250 shots per sample. As a criterion the first 5 spectra are eliminated, since these shots are considered as part of the surface cleaning, the remaining 45 spectra are averaged and are used for the calibration models.

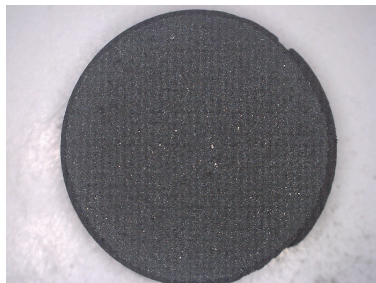
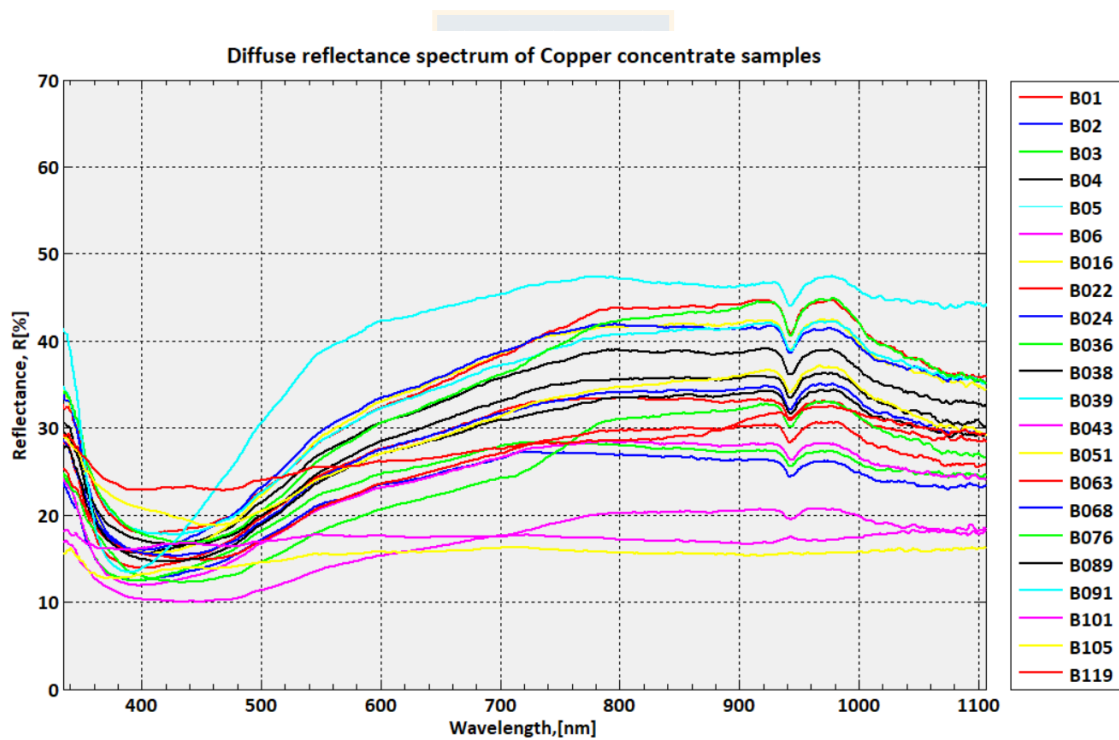
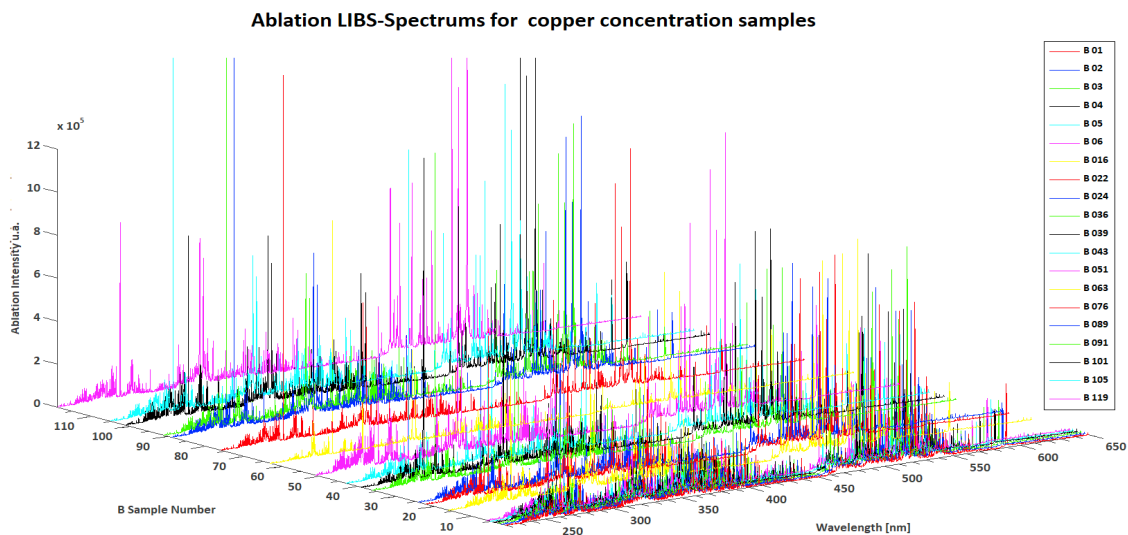


Fig. 7.3: Pellet sample



7.4 General Methodology

The general methodology for analyzing copper concentrates by the fusion of LIBS and DRS considers the following steps:

- Sample preparation. Copper concentrate is a powder where 96.5% of the particles have a size of less than $74 \mu m$. Thus, to perform LIBS analysis copper concentrate must be prepared as a pellet to stand the laser energy.
- Sample characterization by a gold standard method. All the spectroscopy methods require data for calibrating a model. Hence the prepared samples are analyzed by a gold standard method. These results are then taken as a reference value for calibrating the spectroscopic sensors.
- Spectral data acquisition. The spectral data acquisition requires the measurements of a spatial point by both LIBS and DRS. Since LIBS changes the surface of the sample, DRS must be performed first. To this end, it is very important to have a suitable setup for doing these measurements with minimum interference.
- Data preprocessing. Basic spectral data processing is required to enhance features, filter noise, and perform baseline corrections before proceeding to model calibration.
- Data fusion and model calibration. To transform the spectral data in an estimate of the sample composition a model is required. Furthermore, if several sources of information are available it is necessary to consider strategies for fusing the spectral information.
- Model validation. The model is validated using a set of data not used in model calibration.

Before applying data fusion, LIBS and DR data sets are processed to take into account measurement conditions enhancing the important features and reducing noise. Some variable selection approaches are also used to reduce the problem dimension. Only the spectroscopic meaningful features from the LIBS spectra or wavelengths are considered. Considering the full wavelength range may cause serious over-fitting problems.

For LIBS data, a set of preprocessing approaches has been proposed to increase the effectiveness of the regression algorithms. The most common preprocessing are background subtraction,

internal standardization, centering and scaling [97]. To reduce the number of variables feature selection algorithms and the use of heuristic consideration are used to define the number of wavelengths to be considered [110].

In the case of DRS, for a given wavelength range there are several preprocessing approaches to deal with different experimental conditions. Smoothing by Savitzky-Golay algorithm, first derivative, $\log(1/R)$, mean centering, standard variate and multiplicative scatter correction [111]

Before using the preprocessed data for building a fusion model, this should be normalized or standardized. This is an important step since features differ in intensity values, and the model should not emphasize more the data from one source than the other.

Two fusion strategies are considered: low-level and mid-level data fusion, as depicted in Fig. 7.6.a and Fig. 7.6.b respectively.

For Low-Level Data Fusion (LLDF), the full preprocessed spectra from LIBS and DRS data sets are concatenated and used as input for the regression model. At this level, the raw data are directly provided as input to the data fusion process, which provides more accurate data (a lower signal-to-noise ratio) than the individual source [112].

In the case of Mid-Level Data Fusion (MLDF), the full preprocessed spectra from both data sets are concatenated as well, but a variable selection algorithm is used to select the best features before the concatenation. The variable selection algorithm used in this case is CARS (competitive adaptive reweighted sampling) [79], which uses a PLS model, Monte Carlo sampling, and the coefficients of regression from the exponentially decreasing function and adaptive reweighted sampling algorithms. The final selected variables are chosen by means of cross-validation.

The regression model for data fusion was implemented with MATLAB's neural network toolbox. The variable selection algorithm used in MLDF was CARS, implemented with the code for MATLAB provided in [113].

7.4.1 Preprocessing

Before applying data fusion, LIBS and DRS datasets were preprocessed. A total of 107 wavelengths were chosen representing the main constituent elements Cu, Fe, S and As using the information from NIST LIBS database [73], which is frequently termed as spectroscopic prior

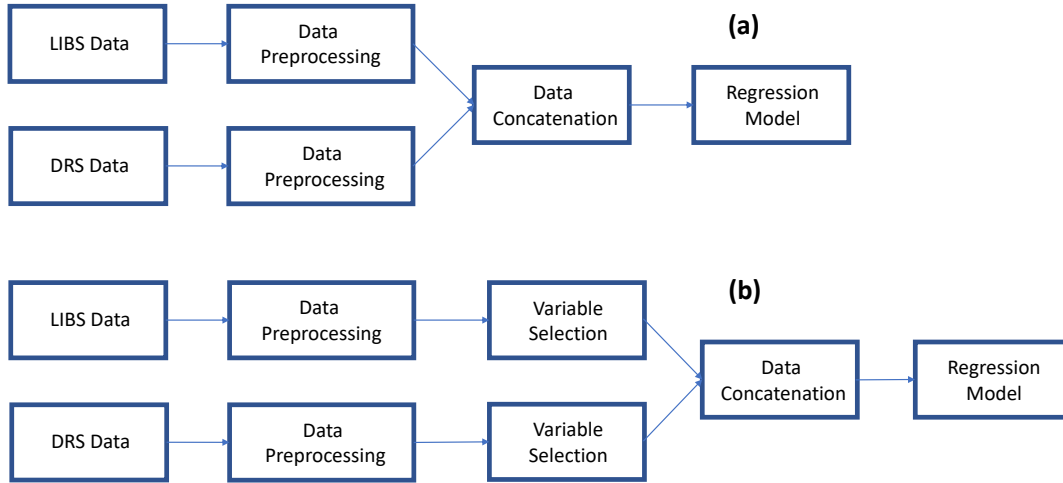


Fig. 7.6: Overview of the data fusion strategies. (a) Low-level data fusion. (b) Mid-level data fusion

knowledge data. In the case of DRS, a total of 559 wavelengths between 400nm to 652nm were considered.

7.4.2 Data Fusion Results

A three-layer neural network with one output was used to estimate the concentrations of each mineral; i.e.

$$\hat{\mathbf{y}}(i) = \mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{x}(i) + \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}; \quad (7.1)$$

where $\hat{\mathbf{y}}(i) \in R$ are the estimated compositions of a given mineralogical component associated to the input vector $\mathbf{x}(i) \in R^n$ representing the network inputs. The weight vectors are $\mathbf{W}^{(1)} \in R^{m \times n}$ and $\mathbf{W}^{(2)} \in R^{1 \times m}$, $\mathbf{b}^{(1)} \in R^n$ and $\mathbf{b}^{(2)} \in R^5$ are biases, and σ is a tansig activation function, where n represents the number inputs and m is the number of units in the hidden layer. To keep a low model complexity 5 units in the hidden layer were selected.

Thus, given a set of LIBS spectra collected from samples with known concentrations $\mathcal{I} = \{(y(i), \mathbf{x}_L(i), \mathbf{x}_D(i)) \in R^{1+n_L+n_D}, |i = 1, \dots, n_S\}$, where n_S is the number of samples. The problem is to find the best fusion strategy to estimate the mineralogical concentrations, while avoiding an increase in model complexity and overfitting problems. The data set \mathcal{I} is divided

into three data sets: training data set \mathcal{T} , validation \mathcal{V} , and test \mathcal{S} , where N_T , N_V , N_S are the number of elements of each set respectively.

The network parameters were obtained by minimizing the mean squared error defined as

$$RSS = \sum_{i=1}^{n_T} (y_j(i) - \hat{y}_j(i))^2 \quad (7.2)$$

by means of a gradient descent algorithm with adaptive learning rate [114]. The data was divided into three sets: 60% for training, 20% for validation and 20% for test.

The training set is used for computing the gradient and updating the network weights and biases. The validation set is used during the training process for monitoring the mean square error and detect the point where overfitting occurs; i.e. if validation mean squared error does not decrease for six iterations, the training is stopped [115]. To quantify the quality of the results, the test set is used for computing the R^2 and RMSEP defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n_T} (y(i) - \hat{y}(i))^2}{\sum_{i=1}^{n_T} (y(i) - \bar{y})^2} \quad (7.3)$$

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_T} (y(i) - \hat{y}(i))^2}{n_T}} \quad (7.4)$$

where \hat{y} represents the predicted values, y the measured values and n_T the number of samples in the test set.

R^2 and $RMSEP$ are both metrics that are used to assess the performance of models for regression tasks. R^2 indicates the proportion of total variance that can be explained by a model [116]. It can be scaled between 0 and 1, if R^2 tends to 1 then more variance is explained by the model, which means that the goodness of fit is increasing. Otherwise, less variance is explained by the models, which means a poor fit. In the case of $RMSEP$, this is a measure of the standard deviation of the unexplained variance, and its value is in the same units as the measured values [117]. If $RMSEP$ tends to 0 then there is less unexplained variance, which means that the goodness of fit is increasing. Otherwise there is more unexplained variance, meaning a poor fit. It is important to notice that while these metrics may show excellent results, it does not necessarily mean that the model will perform well with unseen data, as it may be suffering from overfitting problems, which should be tackled using other techniques, such as regularization, model selection, or others more specific to certain models, including early

Table 7.1: Analytical figure of merits of LIBS, DRS, LLDF, HLDF

Sample	Individual LIBS Figures of merit/ metrics		Individual DRS Figures of merit/ metrics		Low level data fusion (LLDF) Figures of merit/ metrics		Mid-level data fusion (MLDF) Figures of merit/ metrics	
	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2
Bor	0.401	0.615	0.3529	0.7018	0.4632	0.4864	0.2897	0.7991
Ccp	4.509	0.6858	5.052	0.6056	4.0981	0.7404	1.1123	0.9809
Cov	1.0196	0.9444	3.2205	0.4428	2.2869	0.719	0.8377	0.9623
Enr	0.9979	0.9182	0.7042	0.9577	1.3569	0.8429	0.1982	0.9966
Pyr	2.8474	0.8649	6.2571	0.3476	5.1809	0.5527	2.5825	0.8889

stopping, dropout or batch normalization for neural networks.

LIBS and DR data sets were used independently to build regression models and then together using data fusion strategies.

Individual LIBS spectroscopic prior knowledge data analysis results in good R^2 metrics for Covellite, Enargite and Pyrite 0.94, 0.92 and 0.87 respectively, while poor performance in the case of Bornite and Chalcopyrite. Individual DRS data analysis is only able to predict Enargite with good R^2 while the others with less than 0.75. DRS data alone poses serious drawbacks in acquiring prediction ability. Table 7.1 summarizes the performance of the regression model for each strategy used and for each of the mineral species. The results show that by using mid-level data fusion, it is possible to outperform the performance of the individual sources, with root mean squared errors of prediction reductions ranging from 4% to 70%.

To improve prediction ability Low Level Data Fusion (LLDF) has been opted. LLDF strategy comprises the concatenation of the spectroscopic prior knowledge data and the DRS data; i.e a total of 666 input variables. As seen in Table 7.1, this strategy just improves the prediction results for Chalcopyrite.

The MLDF strategy uses a total of 50 features; i.e. 25 variables selected from LIBS spectra and 25 variables from DRS spectra. In both cases, the best 25 variables according to CARS ranking were selected; i.e. the variables associated with the highest sums of CARS weights in absolute values were chosen by the algorithm. The results summarized in Table 7.1 show that MLDF has the best performance. In terms of prediction, as can be seen from the regression plot in Fig. 7.7, all of the 5 minerals were well predicted.

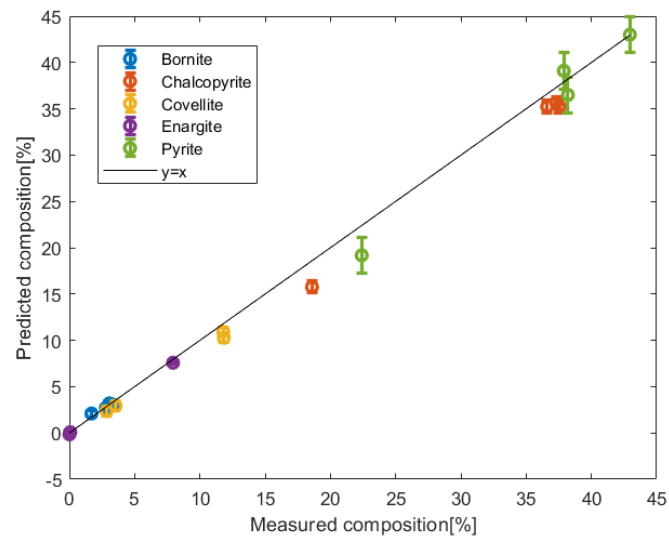


Fig. 7.7: Regression plot for mid-level data fusion strategy of LIBS-DRS data

7.5 Data Fusion of LIBS and HSI Data

In this Chapter, two spectroscopy techniques are evaluated and combined to quantify different mineral species sample compositions in pellet samples, namely Laser-Induced Breakdown Spectroscopy (LIBS) and Hyperspectral Imaging (HSI). Low-level and mid-level data fusion strategies are used. A dataset of LIBS and HSI obtained from 38 samples of copper concentrates is used for regression tests. For both techniques, variable selection is performed, and Artificial Neural Networks (ANN) are used with the selected wavelengths. The results are satisfactory using especially mid-level data fusion, in comparison to using the individual sources.

7.5.1 LIBS-HSI Setup

The setup for this experiment combines LIBS and HSI equipment. The LIBS part is essentially the same used in Chapter 5, Chapter 6 and Chapter 7. There are two hyperspectral cameras for the HSI system, one for the VIS region with a spectral range of 400-1000 nm and a sensor size of 10.85x10 mm - 1024x1024 px. The other camera is for the SWIR region, with a spectral range of 900-1700 nm, and a sensor size of 8.8x6.6 mm - 320x256 px. In this work, the SWIR camera is used for data fusion studies.

Other basic components of the HSI part of the setup include a light source to provide

illumination, an optical fiber; a detector which obtains both spectral and spatial information simultaneously; a hyper-spectrograph to disperse the wavelengths of the reflected, transmitted, or scattered light and deliver signals to the photosensitive surface of the detector; an objective lens to adjust the range of light acquisition; and finally a computer to compose and store the three-dimensional hypercube [118].

7.5.2 Sample Preparation and Treatment

The samples are a subset of the copper concentrates used in Chapter 5 and Chapter 6. The samples are firstly subject to HSI measurements and then subject to LIBS measurements since the last procedures would affect the samples' surface due to the ablation process causing a crater at the locations of the laser shots.

In the case of LIBS, a 10x10 grid is used, which gives a total of 100 spatial points with LIBS measurements. For HSI measurements, the whole samples are scanned using a resolution of 140x160 pixels, and with 256 spectral bands in the SWIR region.

38 samples of copper concentrates compose the dataset, and the mineralogical compositions of bornite, chalcopyrite, covellite, enargite, and pyrite are analyzed through regression tests. The HSI spectra from these samples are plotted in Fig. 7.8.

7.5.3 Preprocessing

Before applying data fusion, LIBS and HSI datasets were preprocessed. A total of 107 wavelengths were chosen representing the main constituent elements Cu, Fe, S, and As using the information from NIST LIBS database [73]. In the case of HSI, a total of 217 wavelength bands were considered.

Because of the lack of labels for every single point at the LIBS grid, the average spectra from these 100 locations was used as the LIBS dataset. In the case of HSI, the exact points where the sample was shot were used as input where once again, the 100 locations were averaged.

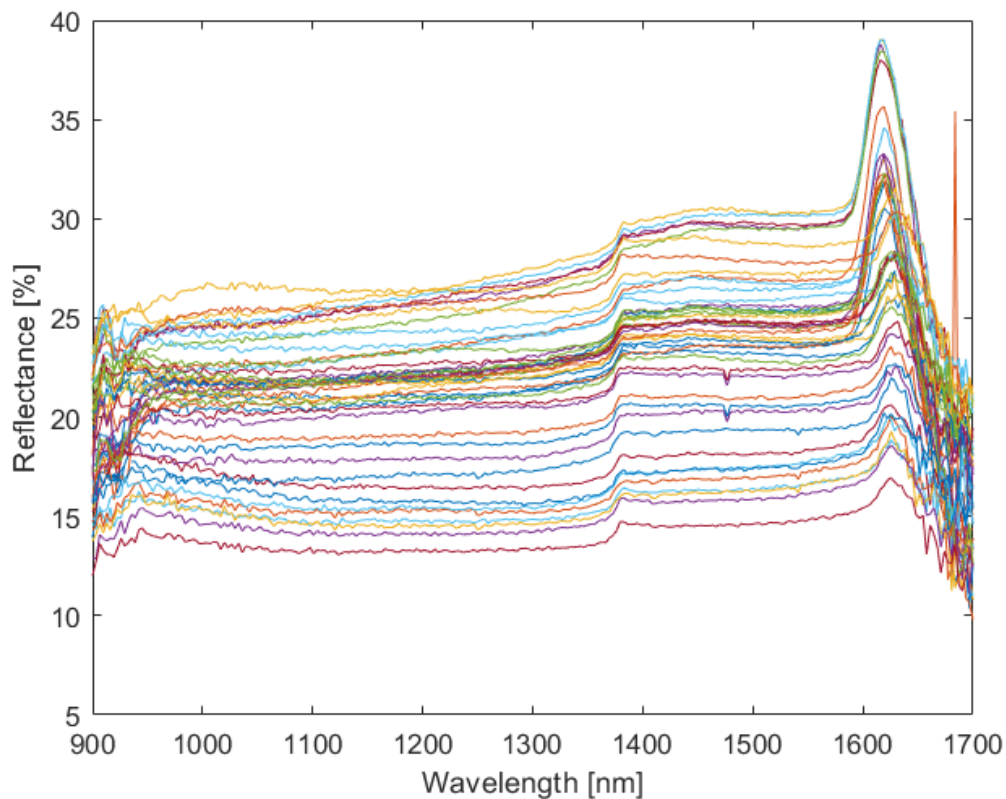


Fig. 7.8: HSI spectra of copper concentrate samples

7.5.4 Data Fusion Results

The same procedures used in the fusion of LIBS and DRS were replicated in this case, in terms of the model architecture and training processes. LIBS and HSI data sets were used independently to build regression models and then together using data fusion strategies.

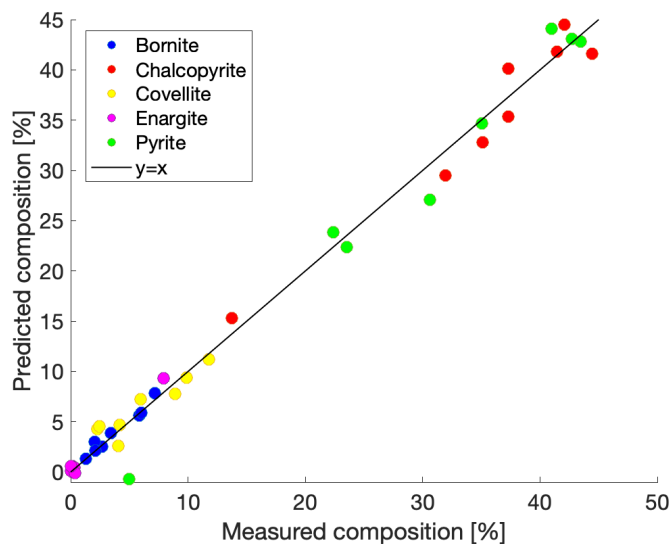
Individual LIBS spectroscopic prior knowledge data analysis results in good R^2 values for Enargite, acceptable R^2 values for Bornite and Pyrite, while poor performance in the case of Covellite and Chalcopyrite. Individual HSI data analysis is only able to predict Chalcopyrite with a good R^2 value, while the others with less than 0.70 in R^2 values. Table 7.2 summarizes the performance of the regression model for each strategy used and for each of the mineral species. For HSI data, various sources of error explain the poor performance, including the spatial and spectral resolutions, the presence of dead pixels, and other experimental errors. The results show that by using mid-level data fusion, it is possible to outperform the performance of the individual sources, with root mean squared errors of prediction reductions ranging from 1% to 74%.

Table 7.2: Analytical figure of merits of LIBS, HSI, LLDF, HLDF

Metric	LIBS	HSI	LLDF	MLDF
Bornite RMSEP (test)	1.3363	1.5715	0.9143	0.4490
Bornite R^2 (test)	0.5806	0.4199	0.8037	0.9527
Chalcopyrite RMSEP (test)	8.5343	6.1019	7.2561	2.2452
Chalcopyrite R^2 (test)	0.4139	0.7073	0.5861	0.9380
Covellite RMSEP (test)	1.3611	1.3893	1.1704	1.3435
Covellite R^2 (test)	0.3309	0.3029	0.5053	0.8399
Enargite RMSEP (test)	0.8259	2.2620	1.7740	0.6121
Enargite R^2 (test)	0.8970	0.2275	0.5245	0.9434
Pyrite RMSEP (test)	8.2389	9.9015	6.4433	2.7077
Pyrite R^2 (test)	0.6362	0.4745	0.7775	0.9517

To improve prediction ability, Low-Level Data Fusion (LLDF) has been opted. LLDF strategy comprises the concatenation of the spectroscopic prior knowledge data and the HSI data, i.e., a total of 324 input variables. As seen in Table 7.2, this strategy improves the prediction results for Bornite, Covellite, and Pyrite.

The MLDF strategy uses a total of 50 features, i.e., 25 variables selected from LIBS spectra and 25 variables from HSI spectra. In both cases, the best 25 variables according to CARS ranking were selected; i.e., the variables associated with the highest sums of CARS weights in absolute values were chosen by the algorithm. The results summarized in Table 7.2 show that MLDF has the best performance. In terms of prediction, as can be seen from the regression plot in Fig. 7.9, all of the five minerals were well predicted.

**Fig. 7.9:** Regression plot for mid-level data fusion strategy of LIBS-HSI data

7.6 Conclusions

This Chapter has illustrated the combined use of LIBS and DRS, and LIBS and HSI to estimate the mineralogical composition of copper concentrates. To perform the measurements a multi-purpose set-up for acquiring LIBS and DR spectra as well as for acquiring LIBS and HSI data at the same spatial position was developed. The spectral information was combined by data fusion techniques. Thus, LIBS elemental information combined with molecular information provided by DRS or HSI can be used to enhance the mineralogical analysis of copper concentrates. The analysis of copper concentrates from different mining operations demonstrate that, for this application, the use of mid-level data fusion outperforms the prediction done by low-level data fusion technique and the separated information sources. Based on the prediction errors from the two data fusion scenarios studied, the fusion between LIBS and HSI showed slightly better results. These encouraging results open the possibility of building more precise measuring systems using a simple optical set-up and suitable data fusion techniques.



8. General Discussion

8.1 Introduction

In this Chapter, a summary of the conclusions derived from this work is presented. Future work is also outlined.

8.2 Conclusions

In this thesis, data fusion techniques were successfully exploited for the quantitative analysis of copper concentrates. The complete data analysis process was executed from preprocessing of the data, model training with optimized hyperparameters, model validation, and model testing. For dealing with the high dimensionality of LIBS spectra, which contain thousands of wavelengths, two variable selection methods were proposed. Both methods are based on prior knowledge or expert selection of intensity peaks and also consider the selection of an optimized ANN architecture in terms of model complexity. The first method considers using an automatic variable selection algorithm to preselect variables that are then filtered based on expert knowledge. The second method proposes a modified cost function to minimize in a regression problem of an ANN, which includes a term considering the prior knowledge information, obtaining improved results compared to those of the first method. Based on the Akaike information criterion, the experimental results show that the proposed methodologies can select a set of wavelengths to improve the accuracy, compared to manual selection, without increasing the model complexity. It was confirmed that the use of nonlinear variable selection methods provided improved results in quantification terms. We conclude that the utilization of methodologies for obtaining parsimonious models is essential to getting good predictive capabilities with high dimensional data possessing a large number of acquisitions. This work has illustrated the combined use of LIBS and DRS, and LIBS and HSI to estimate the mineralogical composition of copper concentrates. It was shown that LIBS elemental information combined with molecular information can be used to enhance the mineralogical analysis of copper concentrates. The analysis of copper concentrates from different mining operations demonstrate that, for this application, the use of mid-level data fusion outperforms the prediction done by low-level data fusion technique and

the separated information sources. The results show that by using mid-level data fusion, it is possible to outperform the performance of the individual sources, with root mean squared errors of prediction reductions ranging from 4% to 70% in the case of LIBS-DRS data fusion, and from 1% to 74% in the case of LIBS-HSI data fusion. These encouraging results open the possibility of building more precise measuring systems using a simple optical setup and suitable data fusion techniques.

8.3 Future Work

Future work considers but is not limited to the following:

- To develop a sensor for online calibration of HSI cameras based on LIBS measurements.
- To use data fusion with combined Vis-SWIR HSI data.
- To extend the analyses to any other chemical element/composition.
- To continue optimizing the second proposed variable selection method considering, for instance, the use of Bayesian optimization for hyperparameter tuning.

9. Publications

9.1 Introduction

In this Chapter, the contributions of this work in terms of publications in Journals and Conferences are presented.

9.2 Journals

- D. Luarte, A. Kumar, M. Velásquez, J. Álvarez, C. Sandoval, R. Fuentes, J. Yáñez, D. Sbarbaro, “Combining prior knowledge with input selection algorithms for quantitative analysis by neural network in laser induced breakdown spectroscopy”, accepted and published in *Analytical Methods*, 2021.
- R. Fuentes, D. Luarte, C. Sandoval, A. Kumar, J. Yáñez, D. Sbarbaro, “Data fusion of Laser Induced Breakdown Spectroscopy and Diffuse Reflectance for improved analysis of mineral species in copper concentrates”, accepted and published in *Minerals Engineering*, 2021.
- D. Luarte, A. Kumar, M. Velásquez, J. Álvarez, C. Sandoval, R. Fuentes, J. Yáñez, D. Sbarbaro, “An optimization approach to combine prior knowledge and LASSO regularization for quantitative analysis by LIBS and ANN”, to be submitted to *Chemometrics and Intelligent Laboratory Systems*, 2021.

9.3 Conferences

- D. Luarte, A. Kumar, J. Yáñez, D. Sbarbaro, “On the Selection of Variables for Quantitative Laser-Induced Breakdown Spectroscopy using Principal Component Analysis and Artificial Neural Networks”, IV EIQ, Brasil, Porto Alegre, 2019.

- D. Luarte, J. Yáñez, D. Sbarbaro, “On the Selection of Variables for Quantitative Multi-Elemental LIBS Using Artificial Neural Networks”, PITTCON, USA, Philadelphia, 2019.



Bibliography

- [1] A. Elhassan, "Short review of laser-induced breakdown spectroscopy for corrosion diagnostic," *AIP Conference Proceedings*, vol. 1380, pp. 65–69, 09 2011.
- [2] M. Shahin and S. Symons, "Detection of hard vitreous and starchy kernels in amber durum wheat samples using hyperspectral imaging (grl number m306)," *Nir News*, vol. 19, 08 2008.
- [3] N. Khajehzadeh and T. Kauppinen, "Fast mineral identification using elemental libs technique," *IFAC-PapersOnLine*, vol. 48, pp. 119–124, 12 2015.
- [4] E. Ferreira, D. Milori, E. Ferreira, L. Santos, L. Neto, and A. R. Nogueira, "Evaluation of laser induced breakdown spectroscopy for multielemental determination in soils under sewage sludge application," *Talanta*, vol. 85, pp. 435–40, 07 2011.
- [5] P. Ilhardt, J. Nuñez, E. Denis, J. Rosnow, E. Krogstad, R. Renslow, and J. Moran, "High-resolution elemental mapping of the root-rhizosphere-soil continuum using laser-induced breakdown spectroscopy (libs)," *Soil Biology and Biochemistry*, vol. 131, 04 2019.
- [6] D. Paules, S. Hamida, R. Lasheras, M. Escudero, D. Benouali, J. Caceres, and J. Anzano, "Characterization of natural and treated diatomite by laser-induced breakdown spectroscopy (libs)," *Microchemical Journal*, vol. 137, 09 2017.
- [7] P. Singh, E. Mal, A. Khare, and S. Sharma, "A study of archaeological pottery of northeast india using laser induced breakdown spectroscopy (libs)," *Journal of Cultural Heritage*, vol. 33, 04 2018.
- [8] A. Rai, J. Pati, and R. Kumar, "Spectro-chemical study of moldavites from ries impact structure (germany) using libs," *Optics and Laser Technology*, vol. 114, 02 2019.
- [9] M. Elfaham, A. Alnozahy, and A. Ashmawy, "Comparative study of libs and mechanically evaluated hardness of graphite/ rubber composites," *Materials Chemistry and Physics*, vol. 27, pp. 30–35, 12 2017.
- [10] N. Khajehzadeh, O. Haavisto, and L. Koresaar, "On-stream and quantitative mineral identification of tailing slurries using libs technique," *Minerals Engineering*, no. 101-109, 2016.

- [11] A. Haider, M. Ullah, Z. Khan, F. Kabir, and K. M. Abedin, "Detection of trace amount of arsenic in groundwater by laser-induced breakdown spectroscopy and adsorption," *Optics Laser Technology*, vol. 56, pp. 299–303, 03 2014.
- [12] J.-H. Kwak, C. Lenth, C. Salb, E.-J. Ko, and K.-W. Kim, "Quantitative analysis of arsenic in mine tailing soils using double pulse-laser induced breakdown spectroscopy," *Spectrochimica Acta Part B-atomic Spectroscopy - SPECTROCHIM ACTA PT B-AT SPEC*, vol. 64, pp. 1105–1110, 10 2009.
- [13] W. Elmenreich, "An introduction to sensor fusion," *Research Report 47/2001*, 2001.
- [14] K. Melessanaki, V. Papadakis, C. Balas, and D. Anglos, "Laser induced breakdown spectroscopy and hyper-spectral imaging analysis of pigments on an illuminated manuscript," *Spectrochimica Acta Part B-Atomic Spectroscopy*, vol. 56, pp. 2337–2346, 12 2001.
- [15] F. Anabitarte, A. Cobo, and J. López-Higuera, "Laser-induced breakdown spectroscopy: Fundamentals, applications, and challenges," *ISRN Spectroscopy*, vol. 2012, 10 2012.
- [16] F. Anabitarte, J. Mirapeix, O. M. Conde, J. López-Higuera, and A. Cobo, "Sensor for the detection of protective coating traces on boron steel with aluminium–silicon covering by means of laser-induced breakdown spectroscopy and support vector machines," *IEEE Sensors Journal - IEEE SENS J*, vol. 12, pp. 64–70, 01 2012.
- [17] B. Gething, J. Janowiak, and R. Falk, "Assessment of laser induced breakdown spectroscopy (libs) for classification of preservative in cca-treated lumber," *Forest Products Journal*, vol. 59, 03 2009.
- [18] Q. Godoi, F. Leme, L. Trevizan, E. Filho, I. Rufini, D. Santos Junior, and F. Krug, "Laser-induced breakdown spectroscopy and chemometrics for classification of toys relying on toxic elements," *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 66, pp. 138–143, 02 2011.
- [19] M. A. Harith, R. Harmon, R. Russo, and R. Hark, "Applications of laser-induced breakdown spectroscopy for geochemical and environmental analysis: A comprehensive review," *Spectrochim. Acta Part B*, vol. 87, pp. 11–26, 01 2013.
- [20] Z.-B. Cong, L.-x. Sun, Y. Xin, Y. Li, and L.-f. Qi, "Comparison of calibration curve method and partial least square method in the laser induced breakdown spectroscopy quantitative analysis," *Journal of Computer and Communications*, vol. 01, pp. 14–18, 01 2013.

- [21] D. Pokrajac, A. Lazarevic, V. Kecman, A. Marcano Olaizola, Y. Markushin, T. Vance, N. Reljin, S. McDaniel, and N. Melikechi, “Automatic classification of laser-induced breakdown spectroscopy (libs) data of protein biomarker solutions,” *Applied Spectroscopy*, vol. 68, 09 2014.
- [22] M. J. C. Pontes, J. Cortez, R. K. H. Galvão, C. Pasquini, M. C. U. de Araújo, R. M. Coelho, M. K. Chiba, M. F. de Abreu, and B. E. Madari, “Classification of brazilian soils by using libs and variable selection in the wavelet domain,” *Analytica Chimica Acta*, vol. 642(1), pp. 12–18, 2009.
- [23] L. Sheng, T. Zhang, G. Niu, K. Wang, H. Tang, Y. Duan, and H. Li, “Classification of iron ores by laser-induced breakdown spectroscopy (libs) combined with random forest (rf),” *J. Anal. At. Spectrom.*, vol. 30, 12 2014.
- [24] J.-B. Sirven, B. Bousquet, L. Canioni, and L. Sarger, “Laser-induced breakdown spectroscopy of composite samples: Comparison of advanced chemometrics methods,” *Anal. Chem*, vol. 78(5), pp. 1462–1469, 2006.
- [25] M. T. Man, S. Cui, J. Yoo, S.-H. Han, K.-S. Ham, S.-H. Nam, and Y. Lee, “Feasibility of laser-induced breakdown spectroscopy (libs) for classification of sea salts,” *Applied spectroscopy*, vol. 66, pp. 262–71, 03 2012.
- [26] T. Vance, N. Reljin, A. Lazarevic, D. Pokrajac, V. Kecman, N. Melikechi, A. Marcano Olaizola, Y. Markushin, and S. McDaniel, “Classification of libs protein spectra using support vector machines and adaptive local hyperplanes,” 07 2010, pp. 1–7.
- [27] I. Gray and A. Millman, “Reflection characteristics of ore minerals,” *Economic Geology*, vol. 57, pp. 325–349, 1962.
- [28] E. Pirard, “Multispectral imaging of ore minerals in optical microscopy,” *Mineral Mag.*, vol. 68, pp. 323–333, 2004.
- [29] J. C. Catalina and R. Castroviejo, “Multispectral reflectance microscopy: Application to automated recognition of metallic ores,” *Revista de Metalurgia*, vol. 53, no. 4, 2017.
- [30] A. López-Benito, J. Catalina, D. Alarcón, U. Grunwald, P. Romero, and R. Castroviejo, “Automated ore microscopy based on multispectral measurements of specular reflectance. I – A comparative study of some supervised classification techniques,” *Minerals Engineering*, vol. 146, January 2020.

- [31] J. Tessier, C. Duchesne, and G. Bartolacci, "A machine vision approach to on-line estimation of run-of-mine ore composition on conveyor belts," *Minerals Engineering*, vol. 20, no. 12, pp. 1129–1144, 2011.
- [32] A. Picon, O. Ghita, P. Whelan, and P. Iriondo, "Fuzzy spectral and spatial feature integration for classification of nonferrous materials in hyperspectral data," *Industrial Informatics, IEEE Transactions on*, vol. 5, pp. 483 – 494, 12 2009.
- [33] L. Bin and J. Chanussot, "Supervised hyperspectral image classification based on spectral unmixing and geometrical features," *Signal Processing Systems*, vol. 65, pp. 457–468, 12 2011.
- [34] A. Le Bris, N. Chehata, X. Briottet, and N. Paparoditis, "Spectral band selection for urban material classification using hyperspectral libraries," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. III-7, pp. 33–40, 06 2016.
- [35] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, pp. 1778 – 1790, 09 2004.
- [36] S. A., "Hyperspectral image classification using unsupervised algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 7, 04 2016.
- [37] G. Candiani, N. Picone, L. Pompilio, M. Pepe, and M. Colledani, "Characterization of fine metal particles derived from shredded weee using a hyperspectral image system: Preliminary results," *Sensors*, vol. 17, p. 1117, 05 2017.
- [38] P. Ramos, I. Ruisánchez, and K. Andrikopoulos, "Micro-raman and x-ray fluorescence spectroscopy data fusion for the classification of ochre pigments," *Talanta*, vol. 75, pp. 926–36, 06 2008.
- [39] C. Di Anibal, M. Callao, and I. Ruisánchez, "1h nmr and uv-visible data fusion for determining sudan dyes in culinary spices," *Talanta*, vol. 84, pp. 829–33, 05 2011.
- [40] M. Bevilacqua, R. Bucci, A. Magrì, A. Magrì, and F. Marini, "Data fusion for food authentication. combining near and mid infrared to trace the origin of extra virgin olive oils," *NIR news*, vol. 24, pp. 12–15, 03 2013.
- [41] A. Biancolillo, R. Bucci, A. Magrì, A. Magrì, and F. Marini, "Data-fusion for multi-platform characterization of an italian craft beer aimed at its authentication," *Analytica chimica acta*, vol. 820, pp. 23–31, 04 2014.

- [42] C. Jiang, Y. Liu, and H. Qu, "Data fusion strategy based on near infrared spectra and ultraviolet spectra for simultaneous determination of ginsenosides and saccharides in chinese herbal injection," *Anal. Methods*, vol. 5, pp. 4467–4475, 2013. [Online]. Available: <http://dx.doi.org/10.1039/C3AY26540D>
- [43] A. Makarau, G. Palubinskas, and P. Reinartz, "Multi-sensor data fusion for urban area classification," 05 2011, pp. 21 – 24.
- [44] A. Marhoubi, S. Saravi, and E. Edirisinghe, "The application of machine learning in multi sensor data fusion for activity recognition in mobile device space," 05 2015, p. 94810G.
- [45] Y. Li, J. Zhang, and Y. Wang, "Ft-mir and nir spectral data fusion: a synergetic strategy for the geographical traceability of panax notoginseng," *Analytical and Bioanalytical Chemistry*, vol. 410, pp. 91–103, 11 2017.
- [46] K. Obisesan, A. Jiménez-Carvelo, L. Cuadros-Rodríguez, I. Ruisánchez, and M. Callao, "Hplc-uv and hplc-cad chromatographic data fusion for the authentication of the geographical origin of palm oil," *Talanta*, vol. 170, 04 2017.
- [47] D. A. Cremers and L. J. Radziemski., "Handbook of laser-induced breakdown spectroscopy," *2nd Edition*, 2013.
- [48] A. W. Miziolek, V. Palleschi, and I. Schechter, "Laser-induced breakdown spectroscopy (libs) fundamentals and applications," *1st Edition*, 2006.
- [49] R. Frei, "Diffuse reflectance spectroscopy; applications, standards, and calibration (with special reference to chromatography)," *J Res Natl Bur Stand A Phys Chem*, vol. 80A(4), pp. 551–565, 1976.
- [50] A. Escobedo Morales, E. Sánchez Mora, and U. Pal, "Use of diffuse reflectance spectroscopy for optical characterization of un-supported nanostructures," *67.Bf*, vol. 53, 01 2007.
- [51] H. F. Grahn and P. Geladi., *Techniques and Applications of Hyperspectral Image Analysis, 1st Edition, 2007, ch 1: Multivariate Images, Hyperspectral Imaging: Background and Equipment*, 2007.
- [52] M. Cocchi, *Data Fusion Methodology and Applications*. Elsevier, 2019.
- [53] T. Zhang, H. Tang, and H. Li, "Chemometrics in laser-induced breakdown spectroscopy," *Journal of Chemometrics*, vol. 32(11), p. e2983, 2018.

- [54] H. Fu, J. Jia, H. Wang, Z. Ni, and F. Dong, "Calibration methods of laser-induced breakdown spectroscopy," in *Calibration and Validation of Analytical Methods - A Sampling of Current Approaches*, M. Stauffer, Ed. DOI: 10.5772/intechopen.72888: IntechOpen, 2017, ch. 5, pp. 85–107.
- [55] P. Porizka, A. Demidov, J. Kaiser, J. Keivanian, I. Gornushkin, U. Panne, and J. Riedel, "Laser-induced breakdown spectroscopy for in situ qualitative and quantitative analysis of mineral ores," *Spectrochimica Acta Part B*, vol. 101, pp. 155–163, 2014.
- [56] N. Khajehzadeh and T. Kauppinen, "Fast mineral identification using elemental libs technique," *IFAC-PapersOnLine*, vol. 48-17, pp. 119–124, 2015.
- [57] E. Tognoni, G. Cristoforetti, S. Legnaioli, and V. Palleschi, "Calibration-free laser-induced breakdown spectroscopy: State of the art," *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 65(1), pp. 1 – 14, 2010.
- [58] A. Kumar, "Less is more: Avoiding the libs dimensionality curse through judicious feature selection for explosive detection," *Scientific Reports 5*, vol. 13169, 2015.
- [59] F. D. Lucia and J. Gottfried, "Influence of variable selection on partial least squares discriminant analysis models for explosive residue classification," *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 66(2), pp. 122–128, 2011.
- [60] R. May, G. Dandy, and H. Maier, "Review of input variable selection methods for artificial neural networks," *Chapter 2 in Artificial Neural Networks - Methodological Advances and Biomedical Applications, Prof. Kenji Suzuki (Ed.)*, vol. ISBN: 978-953-307-243-2, 2011.
- [61] R. Brickley, D. Brown, P. Turk, and S. Clegg, "Improved intact soil-core carbon determination applying regression shrinkage and variable selection techniques to complete spectrum laser-induced breakdown spectroscopy (libs)," *Applied Spectroscopy*, vol. 67(10), pp. 1185–1199, 2013.
- [62] L. Xu, L. Liang, T. Zhang, H. Tang, K. Wang, and H. Li, "A method of improving classification precision based on model population analysis of steel material for laser-induced breakdown spectroscopy," *Analytical Methods*, vol. 6(20), pp. 8374–8379, 2014.
- [63] F. Barbieri Gonzaga, L. Braga, A. Pimentel Sampaio, T. Martins, C. Oliveira, and R. Pacheco, "A simple method for forward variable selection and calibration: evaluation for compact and low-cost laser-induced breakdown spectroscopy system," *Analytical and Bioanalytical Chemistry*, vol. 409(11), pp. 3017–3024, 2017.

- [64] F. Duan, X. Fu, J. Jiang, T. Huang, L. Ma, and C. Zhang, “Automatic variable selection method and a comparison for quantitative analysis in laser-induced breakdown spectroscopy,” *Spectrochimica Acta Part B*, vol. 143, pp. 12–17, 2018.
- [65] S. Lu, S. Shen, J. Huang, M. Dong, J. Lu, and W. Li, “Feature selection of laser-induced breakdown spectroscopy data for steel aging estimation,” *Spectrochimica Acta Part B*, vol. 150, pp. 49–58, 2018.
- [66] V. Motto-Ros, A. Koujelev, G. Osinski, and A. Dudelzak, “Quantitative multi-elemental laser-induced breakdown spectroscopy using artificial neural networks,” *Journal of the European Optical Society*, vol. 08011, 2008.
- [67] E. D’Andrea, S. Pagnotta, E. Grifoni, G. Lorenzetti, S. Legnaioli, V. Palleschi, and B. Lazzerini, “An artificial neural network approach to laser-induced breakdown spectroscopy quantitative analysis,” *Spectrochimica Acta Part B*, vol. 99, pp. 52–58, 2014.
- [68] J. Haddad, D. Bruyère, A. Ismaël, G. Gallou, V. Laperche, K. Michel, L. Canioni, and B. Bousquet, “Application of a series of artificial neural networks to on-site quantitative analysis of lead into real soil samples by laser induced breakdown spectroscopy,” *Spectrochimica Acta Part B*, vol. 97, pp. 57–64, 2014.
- [69] D. R. Anderson, *Model Based Inference in the Life Sciences: A Primer on Evidence*. Springer, 2008.
- [70] J. Kuha, “Aic and bic comparisons of assumptions and performance,” *Sociological Methods and Research*, vol. 33, pp. 188–229, 2004.
- [71] J. Alvarez, M. Velasquez, A. Myakalwar, C. Sandoval, R. Fuentes, R. Castillo, D. Sbarbaro, and J. Yanez, “Determination of copper-based mineral species by laser induced breakdown spectroscopy and chemometric methods,” *J. Anal. At. Spectrom.*, vol. 34(12), pp. 2459–2468, 2019.
- [72] F. Wallis, B. Chadwick, and R. Morrison, “Analysis of lignite using laser-induced breakdown spectroscopy,” *Appl. Spectrosc.*, vol. 54, pp. 1231–1235, 2000.
- [73] National institute of standards and technology @ONLINE. [Online]. Available: <https://physics.nist.gov/PhysRefData/ASD/LIBS/lib-form.html>
- [74] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.

- [75] T. Mehmood, K. Liland, L. Snipen, and S. Sæbø, “A review of variable selection methods in partial least squares regression,” *Chemometrics and Intelligent Laboratory Systems*, vol. 118, pp. 62–69, 2012.
- [76] M. Kamruzzaman, G. Elmasry, D.-W. Sun, and P. Allen, “Application of nir hyperspectral imaging for discrimination of lamb muscles,” *Journal of Food Engineering*, vol. 104(3), pp. 332–340, 2011.
- [77] Y. Zhao, S. Zhu, C. Zhang, X. Feng, L. Feng, and L. He, “Application of hyperspectral imaging and chemometrics for variety classification of maize seeds,” *RSC Advances*, vol. 8, pp. 1337–1345, 2018.
- [78] H.-D. Li, Y.-Z. Liang, Q. Xu, and D.-S. Cao, “Key wavelength screening using competitive adaptive reweighted sampling method for multivariate calibration,” *Analytica Chimica Acta*, vol. 648, pp. 77–84, 09 2009.
- [79] H.-D. Li, Q. Xu, and Y.-Z. Liang, “libpls: An integrated library for partial least squares regression and linear discriminant analysis,” *Chemometrics and Intelligent Laboratory Systems*, vol. 176, 03 2018.
- [80] M. T. Hagan, H. B. Demuth, M. H. Beale, and O. D. Jesus, *Neural Network Design*, 2014.
- [81] M. Ozanne, M. Dyar, M. Carmosino, E. Breves, S. Clegg, and R. Wiens, “Comparison of lasso and elastic net regression for major element analysis of rocks using laser-induced breakdown spectroscopy (libs),” p. 2391, 03 2012.
- [82] M. Dyar, M. Carmosino, E. Breves, M. Ozanne, S. Clegg, and R. Wiens, “Comparison of partial least squares and lasso regression techniques for laser-induced breakdown spectroscopy data of geological samples,” *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 70, p. 51–67, 04 2012.
- [83] T. Boucher, M. Ozanne, M. Carmosino, M. Dyar, S. Mahadevan, E. Breves, K. Lepore, and S. Clegg, “A study of machine learning regression methods for major elemental analysis of rocks using laser-induced breakdown spectroscopy,” *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 107, 02 2015.
- [84] C. Ytsma and M. Dyar, “Effects of univariate and multivariate regression on the accuracy of hydrogen quantification with laser-induced breakdown spectroscopy,” *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 139, 11 2017.

- [85] ———, “Accuracies of lithium, boron, carbon, and sulfur quantification in geological samples with libs in mars, earth, and vacuum conditions,” *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 162, p. 105715, 10 2019.
- [86] Z. Chen, T. Shen, J. Yao, W. Wang, F. Liu, X. Li, and Y. He, “Signal enhancement of cadmium in lettuce using laser-induced breakdown spectroscopy combined with pyrolysis process,” *Molecules*, vol. 24, p. 2517, 07 2019.
- [87] A. Erler, D. Riebe, T. Beitz, H.-G. Löhmansröben, and R. Gebbers, “Soil nutrient detection for precision agriculture using handheld laser-induced breakdown spectroscopy (libs) and multivariate regression methods (plsr, lasso and gpr),” *Sensors*, vol. 20, p. 418, 01 2020.
- [88] D. Bertsimas, D. Kitane, N. Azami, and F. Doucet, “Novel mixed integer optimization sparse regression approach in chemometrics,” *Analytica Chimica Acta*, vol. 1137, 09 2020.
- [89] D. Luarte, A. K. Myakalwar, M. Velásquez, J. Álvarez, C. Sandoval, R. Fuentes, J. Yañez, and D. Sbarbaro, “Combining prior knowledge with input selection algorithms for quantitative analysis using neural networks in laser induced breakdown spectroscopy,” *Anal. Methods*, vol. 13, pp. 1181–1190, 2021.
- [90] K. Sun, S.-H. Huang, D. Wong, and S.-S. Jang, “Design and application of a variable selection method for multilayer perceptron neural network with lasso,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, pp. 1–11, 03 2016.
- [91] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *The Journal of Machine Learning Research*, vol. 13, pp. 281–305, 03 2012.
- [92] B. H. Shekar and G. Dagnev, “Grid search-based hyperparameter tuning and classification of microarray cancer data,” 02 2019, pp. 1–8.
- [93] J. Snoek, H. Larochelle, and R. Adams, “Practical bayesian optimization of machine learning algorithms,” *Advances in Neural Information Processing Systems*, vol. 4, 06 2012.
- [94] R. Rosales, M. Schmidt, and G. Fung, “Fast optimization methods for l1 regularization: A comparative study and two new approaches,” 09 2007.
- [95] C. Fabre, “Advances in laser-induced breakdown spectroscopy analysis for geology: A critical review,” *Spectrochimica Acta Part B*, vol. 166, 2020.

- [96] P. Porizka, A. Demidov, J. Kaiser, J. Keivanian, I. Gornushkin, U. Panne, and J. Riedel, “Laser-induced breakdown spectroscopy for in situ qualitative and quantitative analysis of mineral ores,” *Spectrochimica Acta Part B*, vol. 101, pp. 155–163, 2014.
- [97] J. K. P. Porizka, E. Kepes, D. Prochazka, D. W. Hahn, and J. Kaiser, “On the utilization of principal component analysis in laser-induced breakdown spectroscopy data analysis, a review,” *Spectrochimica Acta Part B*, vol. 148, pp. 65–82, 2018.
- [98] X. Wan and P. Wang, “Remote quantitative analysis of minerals based on multispectral line-calibrated laser-induced breakdown spectroscopy (LIBS),” *Applied Spectroscopy*, vol. 68, pp. 1132–1136, 2014.
- [99] M. Gaft, I. Sapir-Sofer, and H. Modiano, “Laser induced breakdown spectroscopy machine for bulk minerals online analyses,” *Spectrochimica Acta Part B*, vol. 63, pp. 1496–1503, 2007.
- [100] M. Gaft, L. Nagli, I. Fasaki, M. Kompitsas, and G. Wilsch, “Laser-induced breakdown spectroscopy for on-line sulfur analyses of minerals in ambient conditions,” *Spectrochimica Acta Part B*, vol. 64, pp. 1098–1104, 2009.
- [101] J. Alvarez, M. Velasquez, A. Myakalwar, C. Sandoval, R. Fuentes, R. Castillo, D. Sbarbaro, and J. Yanez, “Determination of copper-based mineral species by laser induced breakdown spectroscopy and chemometric methods,” *J. Anal. At. Spectrom.*, vol. 34, no. 12, pp. 2459–2468, 2019.
- [102] A. Criddle and C. Stanley, Eds., *Quantitative Data File for Ore Minerals (3rd ed.)*, 3rd ed. London: Chapman & Hall, 1993.
- [103] M. Cocchi, Ed., *Data Fusion Methodology and Applications*, ser. Data Handling in Science and Technology. Elsevier, 2019, vol. 31.
- [104] E. Gibbons, R. Léveillé, and K. Berlo, “Data fusion of laser-induced breakdown and raman spectroscopies: Enhancing clay mineral identification,” *Spectrochimica Acta Part B*, vol. 170, 2020.
- [105] K. Rammelkamp, S. Schroeder, S. Kubitzka, D. S. Vogt, S. Frohmann, P. B. Hansen, U. Boettger, F. Hanke, and H.-W. Huebers, “Low-level LIBS and Raman data fusion in the context of in situ mars exploration,” *J. Raman Spectrosc.*, no. 1-20, 2019.
- [106] F. Desta, M. Buxton, and J. Jansen, “Fusion of mid-wave infrared and long-wave infrared reflectance spectra for quantitative analysis of minerals,” *Sensors*, vol. 20, no. 1472, 2020.

- [107] J. Park, S. Kumar, S.-H. Han, S.-H. Nam, and Y. Lee, “Combination of diffuse optical reflectance spectroscopy and laser-induced breakdown spectroscopy for accurate classification of edible salts,” *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 179, p. 106088, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S058485472100032X>
- [108] D. Luarte, A. K. Myakalwar, M. Velásquez, J. Alvarez, C. Sandoval, R. Fuentes, J. Yañez, and D. G. Sbarbaro, “Combining prior knowledge with input selection algorithms for quantitative analysis by neural network in laser induced breakdown spectroscopy,” *Anal. Methods*, pp. –, 2021. [Online]. Available: <http://dx.doi.org/10.1039/D0AY02300K>
- [109] B. Chide, S. Maurice, N. Murdoch, J. Lasue, B. Bousquet, X. Jacob, A. Cousin, O. Forni, O. Gasnault, P.-Y. Meslin, J.-F. Fronton, M. Bassas-Portas, A. Cadu, A. Sournac, D. Mimoun, and R. C. Wiens, “Listening to laser sparks: a link between laser-induced breakdown spectroscopy, acoustic measurements and crater morphology,” *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 153, pp. 50–60, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0584854718305081>
- [110] E. D’Andrea, S. Pagnotta, E. Grifoni, G. Lorenzetti, S. Legnaioli, V. Palleschi, and B. Lazzerini, “An artificial neural network approach to laser-induced breakdown spectroscopy quantitative analysis,” *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 99, 2014.
- [111] A. Rinna, F. van den Berg, and S. B. Engels, “Review of the most common pre-processing techniques for near-infrared spectra,” *Trends in Analytical Chemistry*, vol. 28, no. 10, 2009.
- [112] F. Castanedo, “A review of data fusion techniques,” *The Scientific World Journal*, vol. 6, no. ID. 704504, 2013.
- [113] H.-D. Li, Q. Xu, and Y.-Z. Liang, “libpls: An integrated library for partial least squares regression and linear discriminant analysis,” *Chemometrics and Intelligent Laboratory Systems*, vol. 176, 03 2018.
- [114] M. Hagan, H. Demuth, M. Beale, and O. De Jesús, *Neural Network Design (2nd Edition)*. M. Hagan, 2014.
- [115] S. Ingrassia and I. Morlini, “Neural network modeling for small datasets,” *Technometrics*, vol. 47, pp. 297–311, 2005.

- [116] R. Steel and J. Torrie, *Principles and procedures of statistics : with special reference to the biological sciences*. McGraw-Hill, 1960.
- [117] R. Hyndman and A. Koehler, “Another look at measures of forecast accuracy,” *International Journal of Forecasting*, vol. 22, pp. 679–688, 02 2006.
- [118] H. Huang, L. Liu, and M. Ngadi, “Recent developments in hyperspectral imaging for assessment of food quality and safety,” *Sensors (Basel, Switzerland)*, vol. 14, pp. 7248–76, 04 2014.

