BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF PHYSICAL ENGINEERING

FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV FYZIKÁLNÍHO INŽENÝRSTVÍ

CLASSIFICATION OF METALS BY MEANS OF LASER-INDUCED BREAKDOWN SPECTROSCOPY AND CHEMOMETRIC METHODS

KLASIFIKACE KOVŮ POMOCÍ SPEKTROSKOPIE LASEREM BUZENÉHO PLAZMATU A CHEMOMETRICKÝCH METOD

DIPLOMA THESIS
DIPLOMOVÁ PRÁCE

AUTHOR AUTOR PRÁCE	Bc. ERIK KÉPEŠ
SUPERVISOR VEDOUCÍ PRÁCE	Ing. JAN NOVOTNÝ, Ph.D.

BRNO 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání diplomové práce

Ústav:	Ústav fyzikálního inženýrství
Student:	**Bc. Erik Képeš**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Fyzikální inženýrství a nanotechnologie
Vedoucí práce:	**Ing. Jan Novotný, Ph.D.**
Akademický rok:	2016/17

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Klasifikace kovů pomocí spektroskopie laserem buzeného plazmatu a chemometrických metod

Stručná charakteristika problematiky úkolu:

Spektrometrie laserem buzeného plazmatu (LIBS) je metoda vhodná pro rychlou prvkovou analýzu vzorků. Jako jedna z mála metod je relativně dobře aplikovatelná v mimolaboratorních podmínkách a v podobě specializovaného zařízení je možné ji implementovat přímo do průmyslových provozů. Její přednosti, jako je rychlost analýzy a nenáročnost na přípravu vzorků, ji dělají ideálním kandidátem například do linek třídících materiál dle chemického složení.

Spolehlivý provoz těchto přístrojů je závislý na vhodném způsobu zpracování signálu pomocí chemometrických metod (PCA, PLS, PLS-DA). Nezbytné je nalézt takovou, která by s dostatečnou přesností a opakovatelností byla schopna vzájemně rozlišovat definované typy ocelí a slitin hliníku.

Cíle diplomové práce:

1) Rešerše dosavadních způsobů řešení dané problematiky a zmapování využívaných chemometrických metod.

2) Identifikovat vhodnou chemometrickou metodu (PCA, PLS, PLS-DA,...) pro třídění kovových materiálů pomocí techniky LIBS.

3) Experimentální porovnání vybraných metod z hlediska přesnosti a opakovatelnosti na sériích definovaných kovových vzorků.

Seznam literatury:

MIZIOLEK, A. W., V. PALLESCHI and I. SCHECHTER. Laser-induced breakdown spectroscopy (LIBS): Fundamentals and applications. 1st edition. Cambridge: Cambridge University Press, 2006, 620 pages. ISBN 0-521-85274-9.

ERIKSSON L., T. BYRNE, E. JOHANSSON, J. TRYGG, and C. VIKSTRÖM. Multi- and Megavariate Data Analysis: Basic Principles and Applications. 3rd editiont. Umetrics Academy, 2013, 503 pages. ISBN - 13: 978-91-973730-5-0.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2016/17

V Brně, dne

L. S.

prof. RNDr. Tomáš Šikola, CSc.
ředitel ústavu

doc. Ing. Jaroslav Katolický, Ph.D.
děkan fakulty

Summary

This thesis deals with the classification of metals by means of laser-induced breakdown spectroscopy (LIBS) and chemometric methods. The work gives a review of the studies reported on the subject. Three widely used chemometric classification methods are selected: Soft Independent Modelling of Class Analogy (SIMCA), Partial Least Squares Discriminant Analysis (PLS-DA) and a variation of Artificial Neural Networks (ANN), the Feedforward Multilayer Perceptron. Several approaches to exploratory data analysis are also considered. The methods are described, briefly stating their working principle. Subsequently, the performance of the classifiers is experimentally assessed, using several figures of merit.

Abstrakt

Táto diplomová práca sa zaoberá klasifikáciou kovov pomocou spektroskopie laserom indukovanej plazmy (LIBS) a chemometrických metód. Práca poskytuje prehľad o štúdiách na danú tému. Sú vybrané tri široko používané chemometrické klasifikačné metódy: "Soft Independent Modeling of Class Analogy" (SIMCA), "Partial Least Squares Discriminant Analysis" (PLS-DA) a variácia umelých neurónových sietí (ANN), "Feedforward Multilayer Perceptron". Rôzne prístupy k prieskumovej analýze su tiež preskúmané. Metódy sú stručne opísané. Následne sú klasifikátory experimentálne porovnané.

Keywords

laser-induced breakdown spectroscopy, LIBS, chemometrics, classification, SIMCA, PLS-DA, ANN

Klíčová slova

spektroskopia laserom indukovanej plazmy, LIBS, chemometria, klasifikácia, SIMCA, PLS-DA, ANN

KÉPEŠ, E. *Classification of metals by means of Laser-induced Breakdown Spectroscopy and chemometric methods.* Brno: Brno University of Technology, Faculty of Mechanical Engineering, 2017. 77 p. Supervised by: Ing. Jan Novotný, Ph.D.

I declare that I have written this master's thesis on the subject of *Classification of Metals by Means of Laser-Induced Breakdown Spectroscopy and Chemometric Methods* independently, under the guidance of the work's supervisor and using the literature listed in the bibliography.

Bc. Erik Képeš

First and foremost, I would like to thank my supervisor, Ing. Jan Novotný, PhD. for his help and guidance during the preparation of this work. In addition, I'm grateful for the help and useful discussions provided by Ing. Pavel Pořízka, Ph.D. and Ing. Jakub Klus.

V neposladnom rade ďakujem rodičom za podporu počas môjho štúdia.

Bc. Erik Képeš

Contents

Introduction

Laser Induced Breakdown Spectroscopy (LIBS) is a versatile atomic emission spectroscopic method with a very robust instrumentation [1]. This enables LIBS to be used in a great variety of applications, ranging from industrial material analysis [2], through archaeological sample identification [3], to pharmaceutical forgery detection [4]. In addition, further applications of LIBS are being constantly developed. Classification of chemically similar metals being one of such new applications.

The correct identification and classification of metals is a crucial step in applications such as metal recycling and on-line quality control. By the nature of these tasks, a quick paced chemical analysis tool, operating in industrial environment is a necessity. LIBS, requiring little or no sample treatment [1], is a great candidate and has already attracted some attention [5].

Two approaches exist for using LIBS spectra for classification purposes. A more complicated and labour intensive way to classify samples is to correctly assess the concentration of the constituents [6]. This approach requires elaborate calibration procedures, using standardised samples with known composition and matrix.

A more robust and more straightforward method is chosen in this work. By comparing the LIBS spectra and assessing their correlation, it's possible to discriminate between two different materials [7]. To achieve this, several well established chemometric techniques are applied on the LIBS spectra.

The present work has three goals to fulfil. First it tries to give an exhausting overview of the possibilities of using LIBS to classify metals. Secondly, it aims to identify the optimal chemometric classification techniques usable for the task. And lastly, it presents an experimental comparison of the chosen methods.

The work gives a short introduction to the LIBS method in the first chapter, describing its instrumentation, applications and limitations.

Subsequently, in chapter 2 it summarises the studies done on the subject of metal identification and classification by LIBS. This is further expanded by an overview of the combination of LIBS with chemometric methods aimed at material identification and classification in general.

The third chapter deals with the theoretical background of the data processing aspect of classification process. As such, attention is given to exploratory data analysis (EDA), consisting of outlier filtering, various normalisation techniques and data reduction.

Subsequently, the working principles of the three chosen classification techniques is described. These are Soft Independent Modelling of Class Analogy (SIMCA) [8], Partial Least Squares Discriminant Analysis (PLS-DA) [9] and Artificial Neural Networks (ANN) [10].

In the last, fourth chapter, the experimental comparison of the chosen classification methods is presented, considering the influence of the applied EDA.

1 Laser-Induced Breakdown Spectroscopy

Laser Induced Breakdown Spectroscopy is a young [11] atomic emission spectroscopic method. Its development began by the construction of the first laser in 1960 by Maiman [16]. It has since grown into a promising material analysis tool [1], with a wide range of applications:

- Product control of metal production [5]
- Geological sample analysis [12]
- Monitoring of soil contamination [13]
- Identification of chemical and biological warfare agents [14]
- Identification of explosives [15]
- etc.

Principles of LIBS

The principles of the method are schematically shown in 1.1. LIBS is based on the spectroscopic analysis of light emitted by a plasma. As the name of the method suggests, the source of the plasma is a short (in the orders of a few ns, down to fs) laser pulse, focused into a tight spot on the surface of the analysed material. This results in a very high irradiance (several $\mathrm{GW} \cdot \mathrm{cm}^{-2}$).

Due to the high irradiance, multiphoton absorption takes place. Depending on the pulse length several processes can occur. In the case of ns[1] pulses the multiphoton absorption is followed by thermal excitation, melting, ionisation and the creation of seed electrons [1], the latter being a crucial step in creating laser induced plasmas (LIP).

The free (seed) electrons present in the material amplify its ability to absorb energy from the laser pulse [1]. This leads to further ionisation and the consequent evaporation of the sample material. The evaporated ions and electrons form a micro-plasma. A few µs after the laser pulse the plasma begins to cool down, the ions and electrons recombine, the excited atoms are neutralised [1],[2].

During the recombination and neutralisation processes, energy is released by the ions and atoms in the form of photons. The energy of these photons is given by the energy levels of the source ions and atoms. Since the energy levels of each element are unique, spectroscopic analysis of the emitted light enables the identification of the atomic composition of the sample material.

[1] Which is the case for the utilised laser, and the focus of this work. In the case of fs pulses, the pulse is shorter than the thermal relaxation processes. Therefore, no thermal relaxation and no melting takes place and the material is ionised directly after the multiphoton absorption. [85]

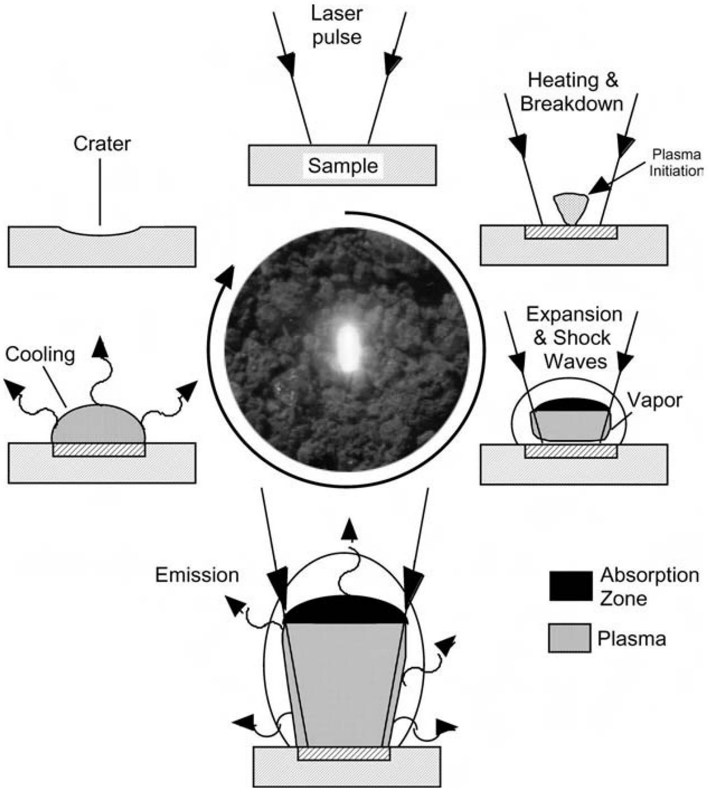

Figure 1.1: The principles of LIBS. From top clockwise: the laser pulse is focused onto the sample surface. This results in the heating and breakdown of the sample, followed by the creation of the laser induced plasma (LIP). The LIP amplifies the absorbtion of the laser light and expands rapidly, creating a shockwave. Brehmsstrahlung radiation is emitted by the expanding LIP still heated by the laser pulse. After the end of the laser pulse the plasma begins to cool down, the ions and electrons recombine, the excited atoms are neutralised, resulting in the characteristic emission spectra of the elements ablated from the sample. After the plasma is extinguished a small crater remains on the sample's surface. Taken from [1].

Advantages and Limitations of LIBS

The straightforward approach of LIBS offers:

- detection of most of the elements [1]
- a very robust instrumentation with the possibility of on-line and even remote analysis [12]
- little, or no sample preparation (cleaning of the sample surface, if needed, can be achieved by the laser with no additional adjustments) [2]
- chemical mapping of the surface (by acquiring spectra from spots all over the surface) [1]
- depth profiling (repeated laser shots into the same spot can drill into the sample) [1]

While LIBS excels at fast and cheap qualitative chemical analysis of any type of sample (solid, liquid and gaseous), its quantification capabilities are considered to be one the method's main limitations. Due to the complex nature of the laser-matter interaction creating the plasma, i.e. the emission source, calculating the elemental composition of the sample from the spectra is rather challenging [1].

For quantitative analyses, therefore, so called calibration curves are constructed. These curves show the relationship between the elements' concentration and its emission lines intensity in the acquired spectra. However, the influence of the sample's chemical matrix on the ablation characteristics complicates the construction.

Consequently, the quantitative analysis requires a set of calibration samples with varying elemental composition but similar matrices. These strict requirements are usually not met, leading to the limited quantitative capabilities of LIBS [2].

2 Material Identification by LIBS

With the aim of selecting the right classification methods, this chapter gives an overview of the studies done on the subject of identifying different materials by using LIBS spectra. The first section is aimed at the identification and classification of metallic samples.

As most of these studies does not involve chemometric methods, or deals with rather small sample variation, the subsequent review is expanded to material identification by LIBS in general.

2.1 Identification and Classification of Metallic Samples by LIBS

The following section reviews the uses of LIBS with the purpose of classifying metallic samples. The aim is to show the different approaches, and to sum up the achievements. It's important to note, that most of the studies did not focus on the comparison of the performance of different classification methods, nor the influence of different Exploratory Data Analysis (EDA - see 3.2). With a few exceptions, they only report the possibility to use LIBS for the given task.

Gurrel *et al.* reported the use of LIBS spectra from 18 certified reference metals to construct calibration curves Fe and Cr [6]. Subsequently, they used the calibration curve to classify 8 other steel samples based on their Fe and Cr content. They investigated the use of several cleaning shots. They achieved a 100 % correct classification rate after 7-8 shots in the same spot, due to the surface layer of sample being removed. However, using the spectra from the first shot only the classification rate was significantly lower, reaching 70 %. The multivariate analysis (MVA) method used for constructing the calibration curve, nor any data treatment methods were reported.

Koujelev *et al.* [72] used artificial neural networks (ANN) with hand-selected emission lines as input to identify 16 different Al alloys. The investigated alloys were similar in composition, hence the identification relied on the signal of trace elements. The selected lines were normalised by a specific Al emission line. They used cleaning shots to remove the surface layer of the metals and subsequently recorded the spectra by averaging the signal from 50 shots to build train the ANN model. The test set consisted of 100 single-shot spectra from each sample. The final total accuracy of the classification was 96.5 %.

Anabitarte *et al.* [74] performed binary classification of steel samples by the presence (or absence) of a boronic anti-oxidant layer. Data reduction by kernel-PCA (k-PCA) and normalisation was done prior to the classification by Support Vector Machines (SVM). The details of the normalisation were not reported. The performance of the classification was not quantified.

Noharet [75] reported a 94 % correct classification rate of 4 different steel alloys (certified reference materials) on-line in single shot mode, an approach which can be implemented in industrial environments. The classification was done by analysing the correlation between intensity ratios of different emission lines. Details about the data processing were, however, not reported.

Jurado-Lopez and Luque de Castre [76] used LIBS to correctly classify 7 metallic alloys into 2 classes, depending on their use in jewelry. A different set of 25 alloys was used to construct a library (12 and 13 alloys in the 2 classes, respectively). Subsequently, the correlation coefficients of the test samples and the spectra in the library were calculated. The samples were assigned to the class containing the spectrum with the highest correlation coefficient.

Merk *et al.* [77] utilised PCA and PLS-DA to classify 10 metallic alloys, from which 9 contained a high percentage of different metal (more than 80 %). They studied the use of LIBS in industrial environment, and thus the samples were covered in a layer of dust. Prior to classification the spectra were truncated to a selected spectral range, normalised and smoothed. The smoothing consisted of rejecting spectral lines with intensity lower than a selected level. PLS-DA outperformed PCA, with the two methods reaching total accuracy of 87 % and 59 %, respectively.

Goode *et al.* [78] reported the classification of 39 metal alloys into 4 classes with the total accuracy of 100 %. The classes were stainless steel, Zr and Ni based alloys, Cu and brass alloys, and mild steels. They used Multiple Discriminant Analysis, with no details about the data pre-treatment.

Werheit *et al.* [79] 60 Al cast and 168 Al wrought samples into two classes (cast and wrought) with total accuracy of 96 %. Subsequently 8 different Al wrought alloys were identified with 95 % total accuracy. The classification was carried out based on the ratio of the intensity of different emission lines, no chemometric method was used. The LIBS spectra were acquired from the samples moving on a conveyor belt, proving the possibility of LIBS being used for industrial metal discrimination.

Kong *et al.* [80] reported the classification of 27 different steel alloys into 4 classes (carbon, low alloy, high alloy and stainless steels). The authors used ANN for classification and compared feature selection (manual line selection) and PCA for data reduction, with the former achieving 100 % total accuracy, and the latter 97 %. No details about the data pre-treatment were included in the study.

Grzegorzek *et al.* [81] reported on the comparison 3 classification methods. 14 different Al alloys were classified by Bayes Classifier, Nearest Neighbour classifier and Support Vector Machine method. The latter outperformed the other two, achieving a 91 % total accuracy.

Moncayo *et al.* [82] investigated the use of PCA and SIMCA for the classification of 6 copper based samples, often found in archaeology. They worked with an unusually small dataset, consisting of 10 spectra for each sample. 7 spectra per each sample were used to build the SIMCA model, the remaining 3 for testing. From the 18 testing observations, 1 was not classified.

Noll *et al.* [5] reported on the use of an industrial LIBS system, capable of distinguishing 35 different material grades. The details of the classification method, however, were not included in the work.

Gornushkin *et al.* [83] used correlation analysis to classify 8 cast iron samples. The authors built a library consisting of spectra from the 8 samples. The classification of unknown observations was done by calculating the correlation between each and every spectrum, finally assigning the spectrum to the class with the highest correlation coefficient. The only data pre-treatment carried out was background correction.

Porizka *et al.* [7] studied the influence of different normalisation approaches on the resulting classification accuracy of SIMCA. In total a set of 13 Al alloys and 38 steel and cast-iron samples were classified (the Al alloys and Fe based samples were studied separately). From the several normalisation techniques, normalisation to $\langle 0, 1 \rangle$ yielded the highest total classification accuracy of 95 % for both the steel and Al datasets.

Porizka *et al.* [22] reported on the use of SIMCA for classifying 10 steel samples. They studied the use of different outlier detection methods. Best results were achieved by selecting outliers by the total energy of the spectra, achieving an improvement of 10% increase in total classification accuracy, reaching a total of 95 %.

Summary of Metal Identification Studies

The ability of LIBS to correctly identify and discriminate metallic samples has been successfully demonstrated by several authors. However, only limited investigation of the use of chemometric classification methods eliminating the need of calibration has been reported. In addition, the number of classes considered in the works is often rather small.

Hence, this work focuses on approaches not relying on the use of calibration sets and quantitative analysis of the samples' composition and incorporates the analysis of an increased number of samples.

2.2 Identification and Classification of Materials by LIBS

To identify the feasible chemometric classification methods, the review of material identification studies has been further expanded to non-metallic materials as well. The most widespread classifiers used in combination with LIBS have been found to be Soft Independent Modelling of Class Analogy (SIMCA - 3.3.1), Partial Least Squares Discriminant Analysis (PLS-DA - 3.3.2), and Artificial Neural Networks (ANN - 3.3.3). The following subsections give an overview of their uses combined with LIBS data.

2.2.1 Soft Independent Modelling of Class Analogy Studies

SIMCA, being one of the most widely adopted classifiers coupled with LIBS data, has been for a wide range of applications. Coalo *et al.* [25] utilised SIMCA for the classification of geological samples. Similar studies were reported by De Lucia *et al.* [15], Clegg *et* *al.* [45], and Anderson *et al.* [24].

A similar application was found by Pontes *et al.* [62], who used SIMCA to differentiate between contaminated and clean soil samples.

Another frequent application of SIMCA on LIBS data is the classification of pharmaceutics, as reported by Myakalwar *et al.* [4], [55], and Dingari *et al.* [54]. From the biological applications of LIBS, SIMCA proved to be useful for the identification of human bone samples by Moncayo *et al.* [39] and infected plant tissue was studied by Pereira *et al.*[52].

SIMCA found possible applications in archeology as well, specifically for the identification of dyes, inks [59] and painting pigments [3].

2.2.2 Partial Least Squares Discriminant Analysis Studies

Delucia *et al.* [15] studied the use of PLS-DA for the classification of geomaterials. The authors advocated the use of PLS-DA, emphasizing its ability to distinguish between the variance coming from shot-to-shot variance of the signal and the variance originating from difference in sample.

Gottfired *et al.* [44] used PLS-DA to classify geological materials, and compared the use of dual-pulse LIBS with single-pulse. They didn't achieve a significant improvement by using a dual-pulse excitation. Further classification of geological materials by PLS-DA was reported by Alvey *et al.* [47], Tion *et al.* [67], Remus *et al.* [36], Hark *et al.* [68], and Zhu *et al.* [69].

Kim *et al.* [13] reported the binary classification of soil samples by PLS-DA, according to the presence of heavy metal contamination.

Larsson *et al.* [31] studied the use of different data preparation techniques for the subsequent classification of biological samples by PLS-DA. They achieved similar results by normalisation of the spectra to unit total energy and feature selection, both outperforming the use of raw spectra.

PLS-DA was also used in biological applications. Moncayo *et al.* [39] classified human bones, while Gottfried *et al.* [70] was able to differentiate between biological nerve agents and other biomaterials. Putnam *et al.* [41] also reported a possible biological application of LIBS coupled with PLS-DA, being able to classify 13 bacterial species.

Similarly to SIMCA, PLS-DA was also used for the classification of pharmaceutical materials by Myakalar *et al.* [55], and Dingari *et al.* [54], energetic materials (explosives) by QianQian *et al.* [60], De Lucia *et al.* [26], [15] and painting pigments by Duchene *et al.* [3].

2.2.3 Artificial Neural Networks Studies

Artificial Neural Networks coupled with LIBS data has found a wide range of interest. Makalwar *et al.* [55] reported a comparison of SIMCA, PLS-DA and ANN for the classification of pharmaceutical samples. PLS-DA outperformed SIMCA, while ANN achieved a 100 % total accuracy, pointing out the possibility of using LIBS to detect fake drugs.

Pokrajac *et al.* [48], [49] investigated the use of LIBS and ANN to classify proteins. They reported an enhanced convergence of the ANN algorithm while using PCA for data reduction. The authors also emphasized the importance of data normalisation to unity prior to utilising the ANN.

Koujelev *et al.* [72] demonstrated the classification capabilities of ANN on a wide range of materials, including metal alloys, marbles, granites, soils, and clay samples. However, they focused only on the identification of the type of the sample.

Sirven *et al.*. [28] used ANN for binary classification of chromium doped and ”clean” soil samples. Elhaddad *et al.* [73] applied ANN for the classification of soil samples as well.

ANN coupled with LIBS data has found several applications in biology. Yueh *et al.* [53] was able to correctly identify organs, while Monzoor *et al.* [27] used ANN to differentiate between different bacterial strains. Lastly Snyder *et al.* [14] successfully classified biological warfare agents and surrogate biological samples.

Summary of the Material Identification Studies

The three selected classifiers, SIMCA, PLS-DA and ANN (will be described in the following chapter) were reported to successfully classify a wide range of materials. Therefore all three of them are good candidates for fulfilling the goals of the present work.

3 Chemometrics

The aim of the following chapter is to outline the chemometric techniques used for the data analysis. A quick introduction to chemometrics is followed by the introduction to the exploratory data analysis (EDA) approaches.

Subsequently, a description of the chosen classification techniques is given. These being Soft Independent Modelling of Class Analogy (SIMCA), Partial Least Squares Discriminant Analysis (PLS-DA) and Artificial Neural Networks (ANN).

3.1 Basic Ideas of Chemometrics

Chemometrics is a set of statistical and mathematical tools used to solve problems in analytical chemistry [17]. As such, it was first introduced by in the 1970s [18]. Its rise in popularity is closely linked to the improvement of modern analytical tools [19]. The vast amount of data recorded in modern day experiments calls for the development and use of highly efficient statistical analysis methods.

To maximise the information extracted from the available data, looking at each variable separately is no longer sufficient, the multivariate approach is required. Considering every variable at once, hidden relationships between predictors and dependent variables can be discovered, which could remain hidden during the univariate approach.

The ever-increasing cost efficiency in computing power of present-day computers was a key to the implementation of the often computationally demanding chemometric methods. A present-day personal computer is able to handle LIBS datasets, consisting of tens of thousands of variables, in real time.

This has led to the recent rise of application of multivariate data analysis tools on LIBS spectra [21]. The use of chemometric tools has found a wide range of applications, ranging from multivariate calibration [6], through classification of different materials [7], to a calibration-free approach to quantitative analysis of LIBS spectra [20].

3.2 Exploratory Data Analysis (EDA)

The aim of Exploratory Data Analysis (EDA) is to prepare the acquired raw data for analysis. This is achieved by the following steps:

- Organisation of the dataset into the proper form ($m \times n$ matrix, mean centering[1], scaling[2], etc.).
- Identification and removal of extreme values (outlier filtering)
- Reduction of the effect of shot-to-shot variance by data normalisation

[1] Mean centering means the subtraction of the arithmetic mean of a variable from each observed value of the variable. This changes the arithmetic mean of the variable to 0.

[2] Scaling is the division of each observed value of a variable by the standard deviation of the variable.

- Dimensional reduction of the data to simplify further analysis

3.2.1 Outlier Filtering

Outliers are aberrated observations (illustrated in fig. 3.1), which do not correctly represent the corresponding sample [19]. The definition of outliers depends on the expectations laid on the data and can vary between applications. In LIBS applications, an observation is considered to be an outlier depending on different criteria:

- It doesn't fit a particular distribution of a chosen feature [22].
- Its distance[3] from the mean or arithmetic average is bigger than a chosen value [22].
- Its variance from the rest of the observations is bigger than a chosen value [22].

Outliers, by definition, can skew the classification models, leading to decreased classification accuracy. It is, therefore, recommended to find and remove outliers before proceeding with further analysis [22].

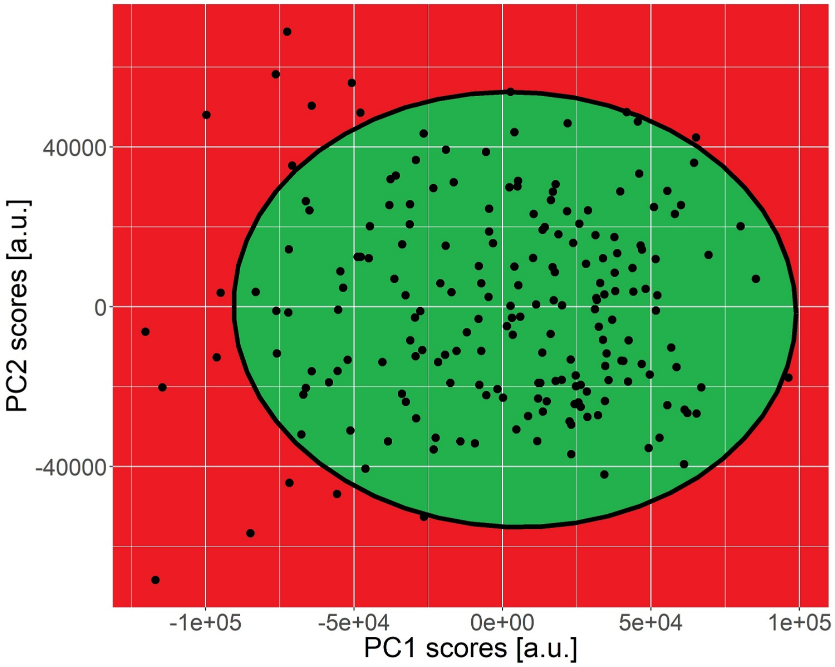

Figure 3.1: Illustration of outliers. The dots show individual spectra projected onto the scores space by PCA (see 3.2.4). The spectra in the red area are considered to be outliers, since their distance from the center is greater than a chosen value.

Review of the Reported Outlier Detection Methods

The most comprehensive study of outlier filtering of LIBS spectra with the aim of classification was carried out by Porizka *et al.*. [22]. The authors utilised outlier detection approaches based on the total energy of the spectra, distance of the spectra in the PCA scores space (for further information see 3.3), and linear correlation of the spectra. The

[3] Euclidian, Mahalanobis (see Appendix B: Statistics), etc.

outlier detection methods were evaluated by the subsequent classification of 10 steel samples by SIMCA (see 3.3.1), showing an improvement in total accuracy compared to the use of unfiltered data.

The distance of spectra in the PCA scores space was also utilised by Devangad *et al.* [23] to detect outliers. However, they did not study other approaches. Outlier detection based on the statistical distribution of the total energy of the spectra was also reported in the work of Anderson *et al.* [24].

An approach based on the intensity distribution of a few selected lines was presented by Colao *et al.*. [25]. They used the distribution of the intensity of hand selected emission lines to detect outliers. No comparison with other methods was reported in their work.

3.2.2 Normalisation

As mentioned earlier, the complex nature of the plasma formation, the matrix effects, the stability of the laser source, and the laser-matter coupling linked to surface inhomogeneity of the sample, results in shot-to-shot variation of the spectra, even when analysing the same sample.

By normalising (schematically shown in 3.2) an attribute describing the laser induced plasma to unity, these effects can be moderated [26]. Some of these attributes are the following:

- Intensity of a chosen emission line [27] [13] [28] [29]
- Total energy of the plasma [30] [31] [32]
- Standard deviation of every variable in the spectrum (also called scaling) [7]

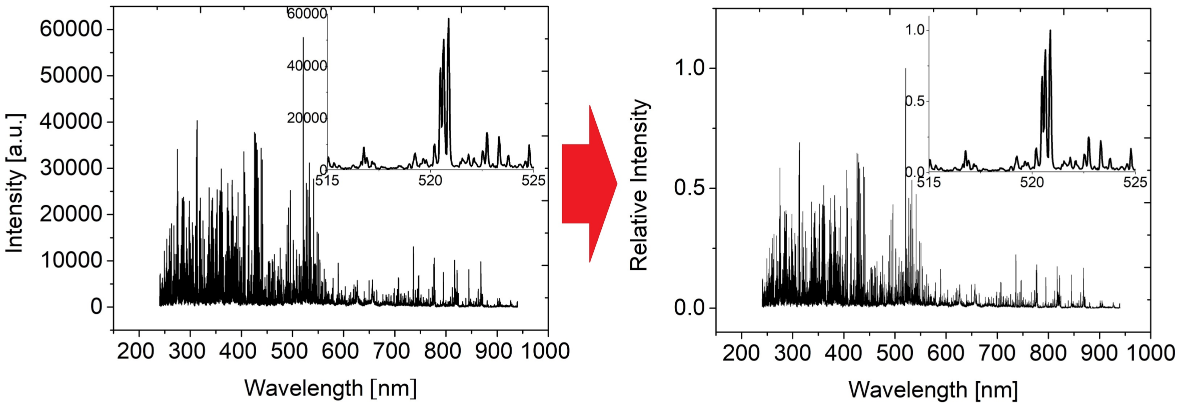

Figure 3.2: Illustration of the influence of a normalisation of a spectrum. The original (raw) spectrum is shown on the left. The left spectrum was normalised to unit maximum intensity, by dividing each point by the highest intensity value in the spectrum.

Review of the Reported Normalisation Methods

Porizka *et al.* [7] carried out an extensive study, comparing the use of several normalisation techniques. The authors compared normalisation to unit total energy, unit intensity of selected emission lines, normalisation of the spectral intensities to $\langle 0, 1\rangle$ and unit standard deviation. The evaluation was done by comparing the classification results of metallic samples done by SIMCA.

Lee *et al.* [33] normalised their spectra to unit total energy. However, they excluded the emission lines with the highest intensity, arguing they are self-absorbed. No comparison with other methods was reported.

3.2.3 Data Reduction

While modern analytical tools produce data with vast number of variables (in the order of tens of thousands), not every variable carries valuable information for the task. Some variables can be highly correlated, others can be noise.

To reduce the required computing power and to ease visualisation, implementing a data reduction technique is recommended [21]. The aim of data reduction is to transform the original data. In the process, the variables carrying the most information are identified. The highly correlated and noisy variables are, on the other hand, discarded.

The data consisting of n variables and m measurements (an $m \times n$ matrix $\boldsymbol{X}$) can be plotted as m points in an n dimensional space. The transformation of $\boldsymbol{X}$ can be regarded as a projection of the original data onto a lower dimensional space. The basis vectors of the new space are the variables carrying the most information.

Data reduction is also possible without transforming the original data. An approach, called feature selection, relies on the selection of emission lines by the operator. This, however, requires prior knowledge about the sample composition [34].

Review of the Reported Data Reduction Methods

The studies utilising PCA for dimensional reduction are listed in 3.3. Therefore, the following summary focuses on other approaches.

Colao *et al.* [25] used hand selected spectral regions and emission lines for their analysis, thus reducing the number of redundant variables.

Amadorhernandez *et al.* [35] utilised both feature selection, and subsequent PCA as well.

Devangad *et al.* [23] used 3 narrow spectral regions for analysis, discarding the rest of the spectra from further analysis.

Hybl *et al.* [30] and Kim *et al.* [13] reported the use of intensity ratio of selected emission lines to reduce the number of redundant variables. In addition, the relative intensity of different emission lines can provide further information about the sample.

Lee *et al.* [33] calculated the correlation coefficient of the intensity of emission lines, following a hand selection of the non-correlated ones, i.e. the emission lines with low correlation coefficient.

Remus *et al.* reported a 10 % better total accuracy of classification using feature selection, compared to the use of the full spectra [36]. In a later study [37] they compared the use of the full spectra and the use of selected spectral ranges, with no significant difference in the resulting classification accuracy of geological samples.

Vars *et al.* [38] utilised a unique data reduction approach, selecting only emission lines with intensity higher than $\frac{T}{n}$, where T is the intensity of the line, and n is the number of principal components used by SIMCA.

Lastly, feature selection was reported by several authors, like Moncayo *et al.* [39], Moros *et al.* [40], De Lucia *et al.* [15], Neiva *et al.* [32], Putnam *et al.* [41], and El Haddad *et al.* [21].

In their work, Myakalwar *et al.* [34] reported better classification results of high energy materials achieved by using the full spectra, compared to feature selection. This leads to uncertainty regarding the use of feature selection.

Cluster Analysis

The aim of cluster analysis (also clustering) is the identification of similarities (or differences) in the data, based on which it can be divided into groups (an example is shown in 3.3). This is done without explicit knowledge of the classes to which the spectra belong to. In other words, clustering would divide spectra with high variance, while grouping together ones with high correlation.

In this work, Principal Component Analysis (see below) has been implemented for data reduction, cluster analysis and the subsequent data visualisation.

3.2.4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a commonly used approach to data reduction, data clustering and data visualisation [17]. It was first introduced by Karl Pearson in 1901 [42], based on the principal axis theorem [43]. Thanks to its non-parametric use and simple implementation, PCA has become one of the most widely utilised tools to tackle both data reduction and data visualisation problems [17].

The Principles of PCA

The data reduction by PCA is based on two principles:

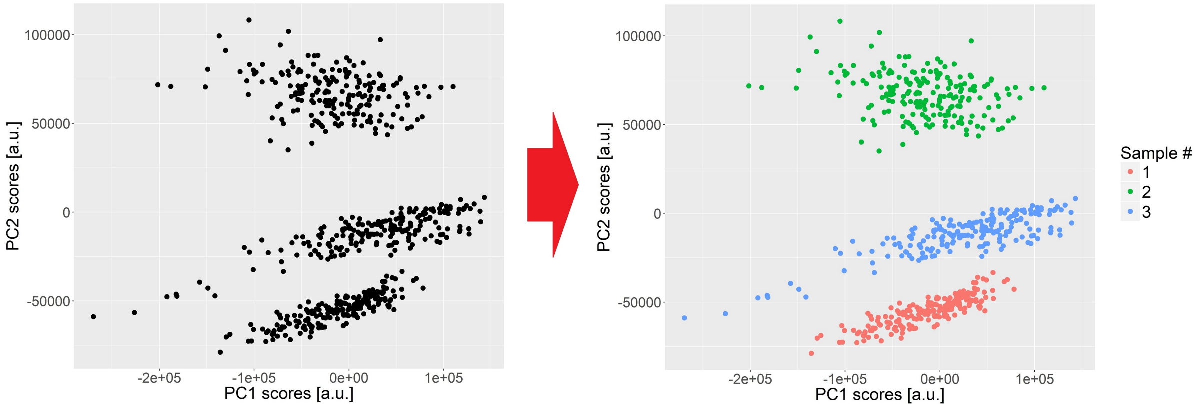

Figure 3.3: Illustration of the cluster analysis done by PCA (see 3.2.4). Each dot represents an individual spectrum projected onto a lower dimensional (2-D) PCA scores space. The left figure was created without explicit knowledge about the source of the spectra, the separation was carried out by identifying the differences in the spectral data. By colouring the dots according to source, the division of the spectra by source sample is evident.

- Highly correlated variables carry the same information, which means a high level of redundancy.

- Variables with high variance are signs of differences between samples, i.e. they carry valuable information about the samples and may help discriminate between different samples.

Following these two principles, PCA aims to find new variables, which explain most of the difference between different samples. These new variables are called principal components (PC). Meanwhile, PCA identifies and discards the linearly dependent, i.e. the highly-correlated variables. This process can be expressed by the following transformation of the original data $\boldsymbol{X}$:

$$\boldsymbol{RX} = \boldsymbol{Y} \tag{3.1}$$

where $\boldsymbol{Y}$ is the new, transformed data and $\boldsymbol{R}$ is the transformation matrix containing the basis vectors of the new, transformed data[4] . The problem PCA is solving is hence reduced to finding the optimal set of basis vectors, arranged as rows of the transformation matrix $\boldsymbol{R}$.

To accomplish this goal, the linear dependency between variables must be expressed. By definition, covariance is the measure of linear relationship between variables. Hence, assuming the $\boldsymbol{X}$ data matrix is mean centered, we can calculate the covariance matrix $\boldsymbol{\Sigma}_X$:

$$\boldsymbol{\Sigma}_X = \frac{1}{m-1}\boldsymbol{X}\boldsymbol{X}^T \tag{3.2}$$

where $\frac{1}{m-1}$ is a constant of normalization [17], an m is the number of observations. In addition to the covariance expressed by the off-diagonal terms, $\boldsymbol{\Sigma}_X$ also contains information about the variance of each original variable, expressed by the diagonal terms.

[4] By looking at the way the columns of $\boldsymbol{Y}$ are calculated, the dot product between the rows of $\boldsymbol{R}$ and columns of $\boldsymbol{X}$ can be recognised. This expresses the projection of $\boldsymbol{X}$ onto the new basis vectors $\boldsymbol{r_i}$.

The projection $\boldsymbol{R}$ should, therefore, maximise the diagonal terms of $\boldsymbol{\Sigma}_Y$ (the covariance matrix of the new, transformed data), and at the same time minimise the off-diagonal terms. Considering the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ given by 3.1, $\boldsymbol{R}$ can be found by following these steps:

$$\begin{aligned}
\boldsymbol{\Sigma}_Y &= \frac{1}{m-1}\boldsymbol{Y}\boldsymbol{Y}^T \\
&= \frac{1}{m-1}\boldsymbol{R}\boldsymbol{X}(\boldsymbol{R}\boldsymbol{X})^T \\
&= \frac{1}{m-1}\boldsymbol{R}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R})^T \\
&= \frac{1}{m-1}\boldsymbol{R}(\boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{R}^T \\
\boldsymbol{\Sigma}_Y &= \frac{1}{m-1}\boldsymbol{R}\boldsymbol{A}\boldsymbol{R}^T
\end{aligned}$$

where $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{X}^T$ is, by theorem 2 of Appendix A, symmetric.

By invoking theorems 3 and 4 of Appendix A, we have

$$\boldsymbol{A} = \boldsymbol{E}_{eig}\boldsymbol{D}\boldsymbol{E}_{eig}^T$$

where $\boldsymbol{D}$ is a diagonal matrix and $\boldsymbol{E}_{eig}$ is a matrix of eigenvectors[5]. By selecting $\boldsymbol{R} = \boldsymbol{E}_{eig}^T$, i.e. a matrix, where the rows are the eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$, $\boldsymbol{\Sigma}_Y$ can be written:

$$\begin{aligned}
\boldsymbol{\Sigma}_Y &= \frac{1}{m-1}\boldsymbol{R}\boldsymbol{A}\boldsymbol{R}^T \\
&= \frac{1}{m-1}\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{D}\boldsymbol{R})\boldsymbol{R}^T \\
&= \frac{1}{m-1}(\boldsymbol{R}\boldsymbol{R}^T)\boldsymbol{D}(\boldsymbol{R}\boldsymbol{R}^T) \\
&= \frac{1}{m-1}(\boldsymbol{R}\boldsymbol{R}^{-1})\boldsymbol{D}(\boldsymbol{R}\boldsymbol{R}^{-1}) \\
\boldsymbol{\Sigma}_Y &= \frac{1}{m-1}\boldsymbol{D}
\end{aligned}$$

This means, that by choosing $\boldsymbol{R}$ to be the matrix containing the eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$, it's possible to diagonalize $\boldsymbol{\Sigma}_Y$. Therefore, the base vectors of the new data space are the eigenvectors of the covariance matrix of the original data $\boldsymbol{\Sigma}_X$.

Data Reduction by PCA

In other words, the resulting PCs are linear combinations of the original variables, with high variance variables having a higher weight. The weights are the elements of the eigenvectors and are called loadings of the principal component. The actual value of the new variables, called the score of the principal component, is attained by calculating the related linear combination. This can be written:

$$S_j = c_{1j}x_1 + c_{2j}x_2 + \ldots + c_{nj}x_n$$

[5] In other words, the columns of $\boldsymbol{E}_{eig}$ are the eigenvectors of $\boldsymbol{A}$.

where S_j is the j^{th} score of a spectrum, c_{ij} is the i^{th} term of the j^{th} eigenvector, x_k is the k^{th} original variable.

The last important term in PCA is the variance explained by the PC. Since the aim of PCA is to find the most valuable variables, i.e. the variables responsible for most of the variance in the data, it's important to quantify how much variance is explained by each PC. This is done by comparing the eigenvalues. The variance explained by i^{th} PC, V_i is estimated by:

$$V_i = \frac{\lambda_i}{\sum_l \lambda_l}$$

where, λ_i is the eigenvalue belonging to the i^{th} eigenvector.

Data Visualisation by PCA

By finding the most valuable variables and combining them into a new set of variables, the data is greatly simplified. This process can be understood as a projection from an $\mathbb{R}^n$ space into $\mathbb{R}^p$, where n is the number of variables in the raw data and $p < n$ is the number of PCs, i.e. the number of new variables. The decreased number of variables eases the possibility of visualising the data (as shown in 3.4), which is limited to a 3-dimensional space.

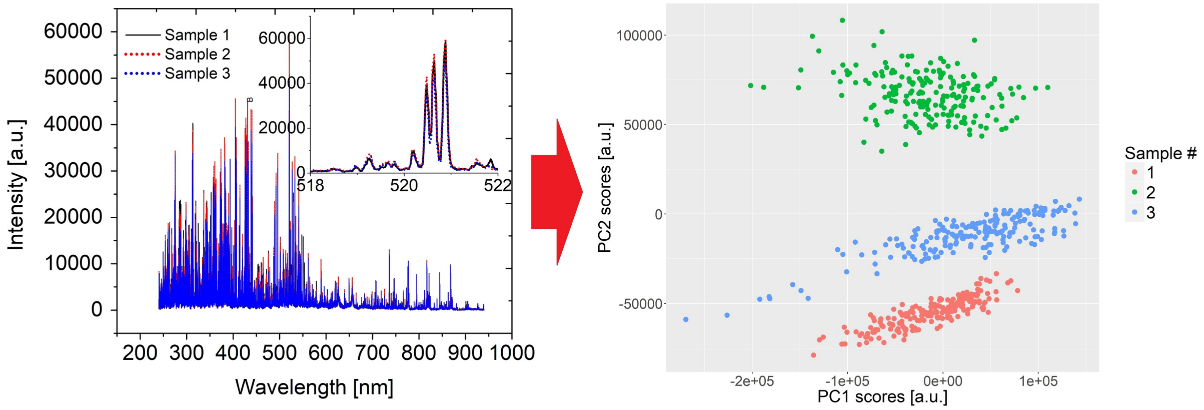

Figure 3.4: Illustration of the visualisation capabilities of PCA. Three spectra from 3 different samples are shown on the left. The difference between the spectra are hardly noticeable, their separation is challenging. On the right side, 200 spectra are shown for each sample. Due to the projection of the spectra onto the lower dimensional scores space, the separation of the spectra is clear.

Usual Assumptions for the Use of PCA

It's important to note, that while deriving the form of the transformation matrix $\boldsymbol{P}$, the $\boldsymbol{\Sigma}_X$ covariance matrix was calculated by eq. 3.2. To use this form, the data matrix X has to be mean centered, i.e. the mean value of each variable has to subtracted from each observed value of the corresponding variable. Additionally, a usual assumption while using PCA is a relatively high signal to noise ratio (SNR) [17], defined by:

$$\mathrm{SNR} = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

where σ_{signal} is the standard deviation of the signal, and σ_{noise} is the standard deviation of the noise. Scaling the dataset, i.e. dividing each observed value of a variable by its standard deviation, reduces the SNR. Therefore, scaling is not recommended for LIBS applications.

Review of the LIBS Studies Utilising PCA

PCA is often used for data reduction and visualisation. As such, it has found applications on the analysis of geological samples by De Lucia *et al.* [15], Gottfried *et al.* [44], Clegg *et al.* [45], Lassue *et al.* [46] [47], biological samples by Pokrajac *et al.*[48], [49], Hybl *et al.* [30], Labb *et al.* [50], Larsson *et al.* [31], Moncayo *et al.* [39], Munson *et al.* [51], Pereira *et al.* [52], Yueh *et al.* [53], and pharmaceutical materials by Myakalwar *et al.* [4], [55] and Dingari *et al.* [54].

Devangad *et al.* [23] also utilised the clustering capabilities of PCA. Another study including PCA clustering was done by Samuels *et al.* [56]. Laudstrm *et al.* [57] used clustering by PCA to differentiate between bioaerosols and paint materials. A similar approach was taken by Unnikrishnan *et al.* [58], who used the Mahalanobis distance of spectra in the scores space to differentiate between different plastics.

PCA found further applications in identification of dyes and inks, as reported by Hoehse *et al.* [59], metal contamination detection in soils studied by Kim *et al.* [13], and Sirven *et al.* [28], analysis of sea salts' origin reported by Lee *et al.* [33], and the identification of energetic materials carried out by QianQian *et al.* [60].

3.3 Classification

The goal of classification is the following. The data usually consists of observations from several different samples. The observations (whole spectra in the case of LIBS) can be grouped together by origin, i.e. by sample. These groups are called classes. The classification methods' aim is to assign new observations with unknown origin to these classes.

This is achieved by building statistical models of each class (sample) from a so-called training data. The training data consists of observations with known origin. Each classification method uses a different approach to build the statistical model (see below) in a way, that allows the assignment of observations with unknown origin to the existing classes.

By building a statistical model for each class, classification methods eliminate the need of comparing newly attained spectra with the existing set of measurements. Instead, the observations with unknown class are simply compared to the existing models. Subsequently they are classified by a pre-defined rule, varying by classification technique.

Classification vs. Clustering

While the final goal of classification might be similar to clustering (grouping data together), there is a substantial difference between the two approaches. Cluster analysis uses no information about the existing classes. On the other hand, classification methods are built to find the inter-class differences and intra-class regularities. This not only allows the grouping of observation from the same sample together, but also helps to distinguish between observations from different classes.

Under- and Overfitting

Classification models face two common issues, namely under- and overfitting (schematically shown in 3.5). Underfitting results in ambiguity. This means, that the mathematical models are not precise enough, they possibly overlap, causing possible misclassification. This is often caused by insufficient size of the training data or extreme values (outliers), which do not represent the class precisely enough.

On the other hand, tight mathematical models result in over-fitting. Over-fitted models allow for only a small variance in data, resulting in a high rate of unclassified samples.

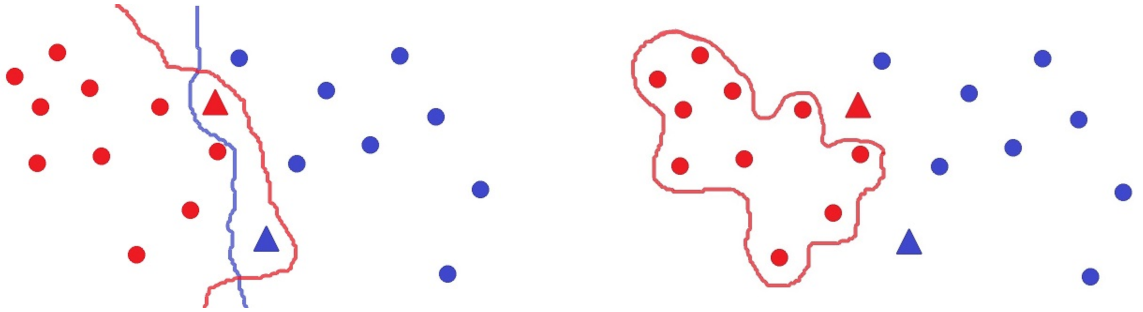

Figure 3.5: Illustration of under- and overfitting. The dots represent the observations from which the models are built (the training data), the different colours meaning different samples. The triangles are the observations to be classified (testing data). The colour of the triangles shows to which class they should be classified. The model on the left side is underfitted. Therefore, the classification is unclear, both testing observations could belong to each class. The (red) model on the right side, on the other hand, is overfitted, and is not able correctly classify the red triangle.

Figures of Merit

To compare the performance of different classification techniques, a set of figures of merit (FoM) needs to be defined. Unfortunately, the studies done on the subject do not uniformly use the same FoMs. The encountered ones include the following:

- Total Accuracy:

$$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

- Sensitivity:
$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$
- Specificity:
$$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

where TP is the number true positives, TN true negatives, FP false negatives, and FN false negatives, which are well-known terms from statistics.

3.3.1 Soft Independent Modelling of Class Analogy (SIMCA)

Soft Independent Modelling of Class Analogy (SIMCA) was first introduced by S. Wold and M. Sjöström in 1977 [8] as a classification method closely related to PCA (see below). Being built on PCA, SIMCA is a linear classification method, with relatively simple implementation.

Model Building

The aim of SIMCA is to build a PCA model of each class separately (hence the independent in its name), based on the training data:

$$\boldsymbol{Y}_j = \boldsymbol{X}_j \boldsymbol{R}_j$$

where $\boldsymbol{Y}_j$ is the score matrix of the j^{th} class, $\boldsymbol{X}_j$ is the matrix of observations in the training data belonging to the j^{th} class and $\boldsymbol{R}_j$ is the loadings matrix of the j^{th} class, and $j = 1, \ldots, k$ for k different classes.

Classification Rule

The new observations are then classified by calculating its orthogonal distance (OD) from the center of the j^{th} group's model, defined by [61]:

$$\mathrm{OD}_j = \|\boldsymbol{y}_{new} - \bar{\boldsymbol{y}}_j\|$$

where OD_j is the orthogonal distance of the observation to be classified from the centre of the j^{th} class, $\boldsymbol{y}_{new} = \boldsymbol{R}^T \boldsymbol{x}_{new}$ is the scores vector of the observation to be classified $\boldsymbol{x}_{new}$, and $\bar{\boldsymbol{y}}_j$ is the arithmetic mean of scores of the observations in j^{th} class.

The observation is assigned to the class with the smallest distance (depicted in 3.6), or to a class (or to several classes) based probability, while assuming a random distribution of the score values in the classes, i.e. by performing the t-student test. Should the test fail for each class, the object is not classified.

The possibility of assigning an observation to more than one class or not assigning it to any is called soft modelling.

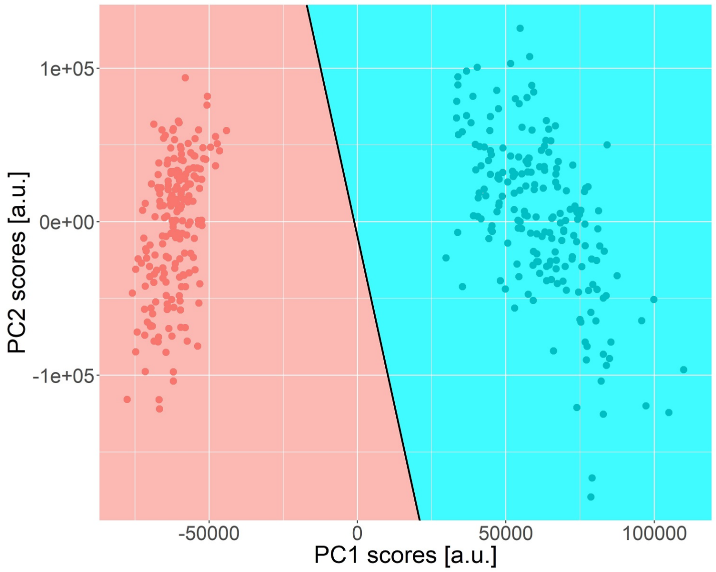

Figure 3.6: Simple visualisation of the working principle of SIMCA. by building a PCA model of the two (red and blue) samples, the scores space is divided by a plane, according to the distance from the center of the two models. Observations in the red (blue) area will be classified as members of the red (blue) class. For multiclass classification the models in the higher dimensional space are separated by hyperplanes.

3.3.2 Partial Least Squares Discriminant Analysis (PLS-DA)

Partial Least Squares Discriminant Analysis (PLS-DA) is a linear classification technique [17], being one of the most widely used classification tools [21]. PLS originates from the work of Herman *et al.* [9]. However, it was not constructed as a classification method. Rather, it was devised to solve regression problems [63].

An extension to PLS, namely the PLS-DA was originally devised by Gottfires J. *et al.* [64] to utilise PLS for classification. Since then, PLS-DA has become a widely-spread technique for solving classification problems [63].

Model Building

Similarly to PCA, PLS-DA also performs a dimensional reduction by finding a new set of variables, namely a set of latent variables (LVs). The goal of this approach, again, is to minimise the redundancy among the predictors. However, PLS-DA differs from PCA in that it uses information about the classes to find the LVs. The difference of the scores of the two methods are shown in 3.7, although Brereton *et al.* warns users about interpreting PLS-DA scores plots the same way as PCA scores plots are interpreted [66].

For PLS-DA, the objects' classes in the training set are organised into a class matrix $\boldsymbol{C}$. For m objects belonging to k different classes, $\boldsymbol{C}$ is an $m \times k$ dimensional matrix. For the p^{th} object belonging to q^{th} class $(p \leq n, q \leq k)$, the pq^{th} term of $\boldsymbol{C}$ is 1 and the rest of the p^{th} row's values are 0.

Therefore, in addition to maximising the covariance between the LVs, PLS-DA also maximises the correlation between the LVs and the response variables, i.e. the matrix

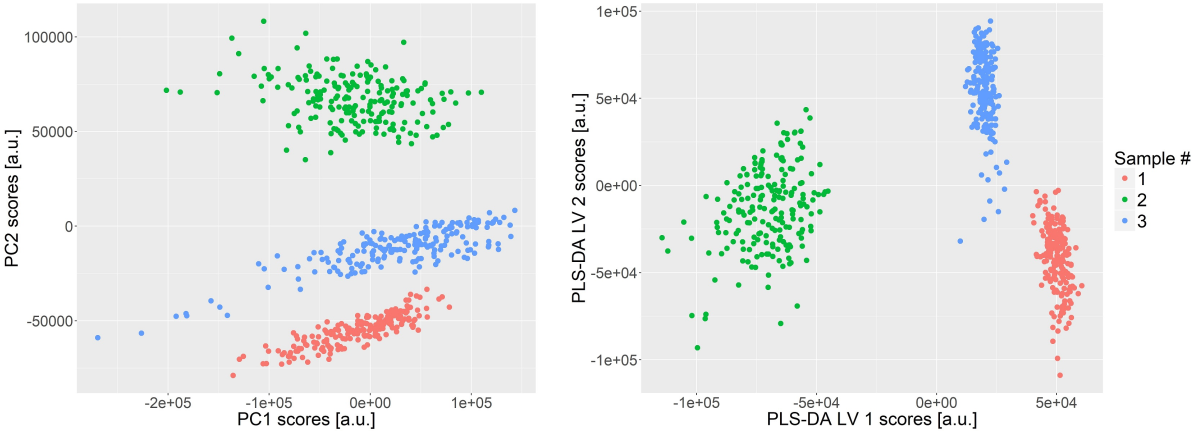

Figure 3.7: Illustration of the difference in PCA (on the left) and PLS-DA (on the right) scores.

$\boldsymbol{C}$. This is achieved by projecting both the independent variables $\boldsymbol{X}$, and the response variables $\boldsymbol{C}$ onto the latent variables by:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P} + \boldsymbol{E} \tag{3.3}$$

$$\boldsymbol{C} = \boldsymbol{T}\boldsymbol{Q} + \boldsymbol{F} \tag{3.4}$$

where $\boldsymbol{T}$ is the common scores matrix, $\boldsymbol{P}$ and $\boldsymbol{Q}$ are the predictor and response loading matrices, respectively. Lastly, $\boldsymbol{E}$ and $\boldsymbol{F}$ are the error matrices. An important feature of PLS-DA is the way of relating the predictor and response variables indirectly, through the common scores. [19].

In other words, while PCA is solving the eigenvalue problem of the correlation matrix, PLS-DA is considering the combined variance-covariance matrix $\boldsymbol{X}\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{X}^T$ and the matrix $\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C}\boldsymbol{C}^T$ [65]. This is done by carrying out in the following steps [66]:

1. A vector $\boldsymbol{u}$ is chosen as one of the column vectors of $\boldsymbol{C}$

2. The PLS weight vector $\boldsymbol{w}$ is calculated: $\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{u}$

3. The t scores are estimated:

$$\boldsymbol{t} = \frac{\boldsymbol{X}\boldsymbol{w}}{\sqrt{\sum w_i^2}}$$

   where w_i are the elements of $\boldsymbol{w}$

4. The $\boldsymbol{x}$ loadings are estimated (loadings of only 1 LV):

$$\boldsymbol{p} = \frac{\boldsymbol{t}^T\boldsymbol{X}}{\sum t_i^2}$$

   where t_i are the elements of $\boldsymbol{t}$

5. The c loadings are estimated (again, the loadings of only 1 LV):

$$\boldsymbol{q} = \frac{\boldsymbol{C}^T\boldsymbol{t}}{\sum t_i^2}$$

   where t_i are the elements of $\boldsymbol{t}$

6. Calculate a new vector
$$\boldsymbol{u} = \frac{\boldsymbol{Cq}}{\sum q_i^2}$$
where q_i are the elements of $\boldsymbol{q}$, and return to step 2.

7. Compare $\boldsymbol{t}_{k-1}$ and $\boldsymbol{t}_k$:
$$\sum (t_{k-1,i} - t_{k,i})^2$$
where $\boldsymbol{t}_k$ is the scores vector from the k^{th} iteration and $t_{k,i}$ are its elements. If the sum is small, the model is sufficiently precise and the calculation can proceed to the next step. Otherwise a new $\boldsymbol{u}$ is calculated and the process is repeated from the 2nd step.

8. The effect of the calculated LV (consisting of the loadings $\boldsymbol{p}$ and $\boldsymbol{q}$, and scores $\boldsymbol{t}$) is subtracted from the original data:
$$^{resid}\boldsymbol{X} = \boldsymbol{X} – \boldsymbol{t} \cdot \boldsymbol{p}$$

9. The new estimation of $\boldsymbol{C}$ is calculated
$$\boldsymbol{C}_k = \boldsymbol{C}_{k-1} + \boldsymbol{t} \cdot \boldsymbol{q}$$
where $\boldsymbol{C}_k$ is the class matrix from the k^{th} iteration. Summing the contribution of every component the $\boldsymbol{C}_{sum}$ is calculated, which is then used to get the residual of $\boldsymbol{C}$:
$$^{resid}\boldsymbol{C} = \boldsymbol{C} - \boldsymbol{C}_{sum}.$$

10. For the estimation of further latent variables the $\boldsymbol{X}$ and $\boldsymbol{C}$ matrices are replaced by their residuals and the calculation is repeated from step 1.

Classification Rule

After building the model (calculating the desired number of LVs), the subsequent classification is carried out by following several decision rules. The simplest rule is to assign the observation to a class with the highest value of predicted c [66].

3.3.3 Artificial Neural Networks (ANN)

In general, artificial neural networks (ANNs) are a computational model for supervised learning inspired by the brain's biological structure. ANNs mimic the connections between neurons to complete the given computational tasks. In their current form, they were first devised by Werbos *et al.*. in 1974 [10]. They have quickly earned a wide range of interest due to their ability to model non-linear problems without any prior knowledge about the form of the involved function.

The term ANN covers a wide range of computational methods. Therefore, only the one used in the present work will be described. This being the multilayer perceptron (MLP) model with backpropagation [71].

Multilayer Perceptron (MLP)

The working principle of the MLP is the following. It consists of several layers and neurons, as shown in 3.8. The first (left-most) layer is called the input layer, and the last (right-most) layer is the output layer. An arbitrary number of layers can be placed between these two layers, with no clear rule to determine their number [71].

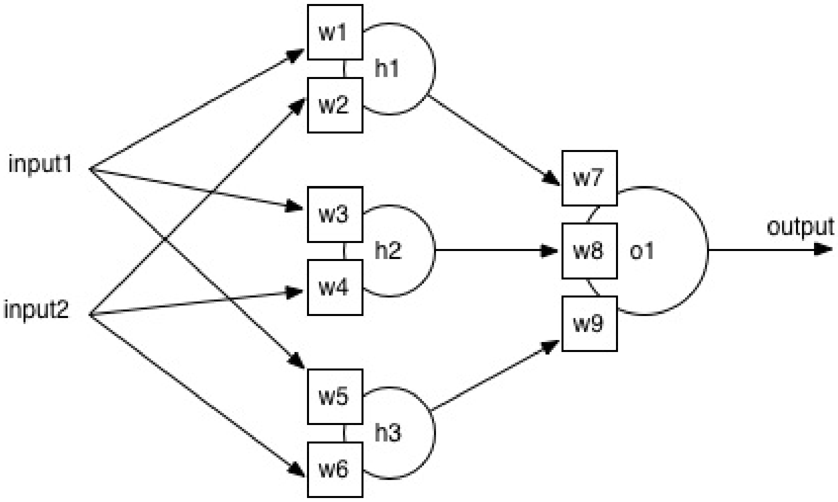

Figure 3.8: Simple depiction of an artificial neural network with no hidden layers, 3 input neurons and 1 output neuron: w_i represent the weights of each connection, h_i the output of the function in the neurons, $o1$ the output of the function in the output neuron. Taken from http://cowlet.org.

The number of neurons in the first layer is given by the number of predictor variables, for each variable there is one neuron in the input layer. The number of neurons in the output layers is equal to the number of expected results (the number of classes). The number of neurons in the middle layers is not clearly defined, similarly to the number of layers [71].

The neurons

Every neuron is identical [71]. Each of them receives a set of weighted inputs (one from each neuron in the previous layer), sums them and calculates an output value. Each connection between neurons is assigned a separate weight value, which can be adjusted independently.

The output value is in turn relayed to the next layer of neurons, i.e. to each neuron in the next layer. Therefore, the data flows only in one direction, towards the output layer. This attribute is referred to as feedforward model (illustrated in 3.9).

The output is calculated by a chosen function. A usual choice is the sigmoid function, also called the logistic function (shown in figure 3.10):

$$f(x) = \frac{1}{1+e^{-x}}$$

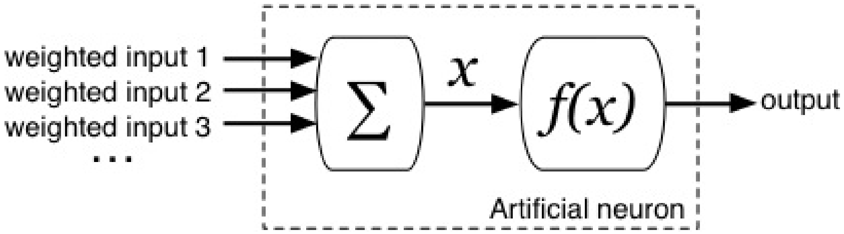

Figure 3.9: Illustration of the working principle of an artificial neuron. The weighted inputs are summed, used as input for the function, and the output of the function is sent towards each connected neuron in the following layer. Taken from http://cowlet.org.

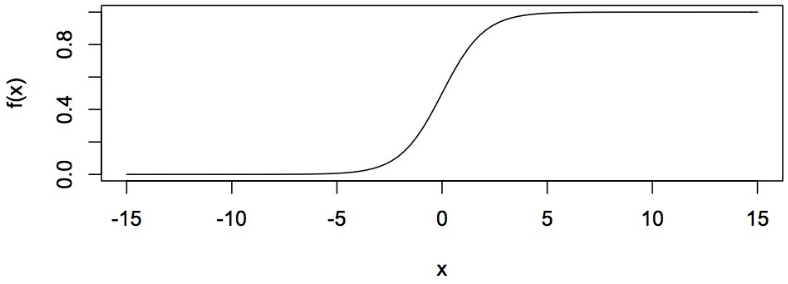

Figure 3.10: Shape of the logistic function. Taken from http://cowlet.org.

Since the logistic function's output is limited to values between $\langle 0, 1\rangle$, it is advised to normalise the input values to unity for better results [71].

Model Building and Backpropagation

For building an ANN classification model a matrix containing the response variables is created, similarly to the one for PLS-DA. For m observations and k classes this means an $m \times k$ matrix, where for the r^{th} observation belonging to s^{th} class ($r \leq m, s \leq k$), the rs^{th} term is 1 and the rest of the r^{th} row's values are 0.

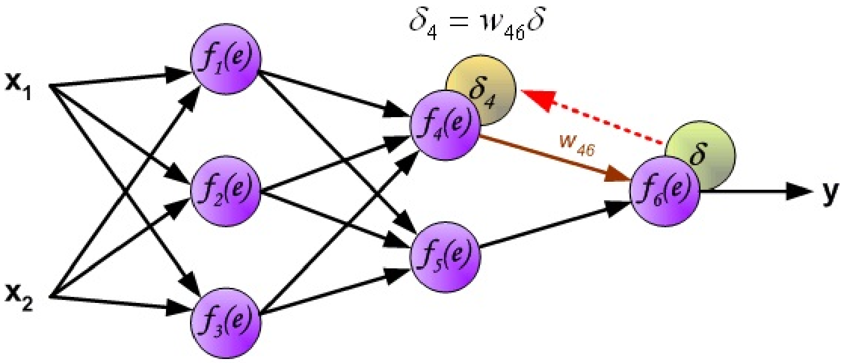

Figure 3.11: First step of the backpropagation algorithm. The error function is passed to the previous layer of neurons. Taken from http://galaxy.agh.edu.pl.

The model building is done by calculating the output y for each observation and comparing it to the expected value y_{true}, contained in the response matrix[6]. If the difference between the two $\delta = y - y_{true}$ is greater than a chosen number, the weights of the neurons' connections are adjusted. The adjustment is calculated by the backpropagation algorithm, shown in 3.11-3.13.

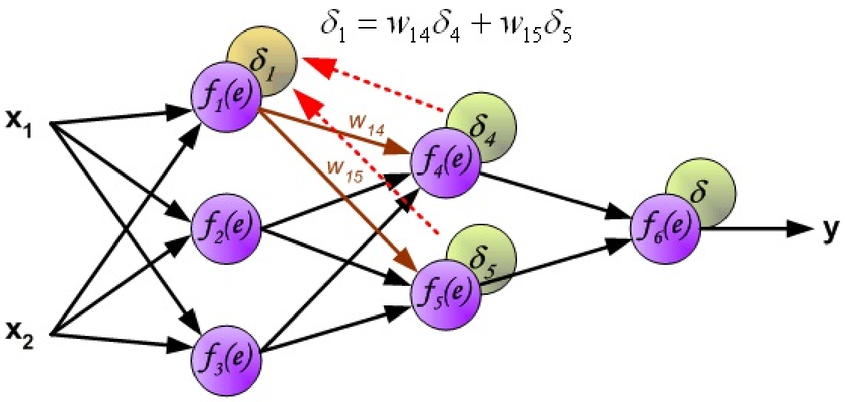

Figure 3.12: Calculation of the error signal at each neuron. Taken from http://galaxy.agh.edu.pl.

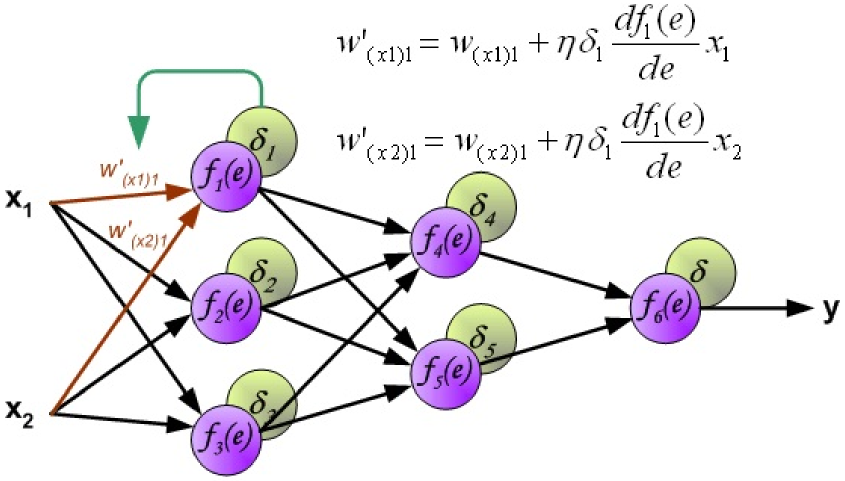

Figure 3.13: Depiction of the weights' adjustment. Taken from http://galaxy.agh.edu.pl

The adjustment of the weights is in principle a backwards propagation of the resulting error δ, called error signal, through the network, hence the name backwards propagation or backpropagation.

The value of the error signal at each neuron δ_i is calculated using the existing weights:

$$\delta_i = \sum w_{ij}\delta_j$$

[6] For simplicity only a single output is considered in the description of the working principles of MLP. For multiclass classification problems there is an output neuron for each class.

where w_{ij} is the weight between neurons i and j, δ_j is the value of the error signal at neuron j. This step is shown in 3.12.

Subsequently the new weights w'_{ij} are calculated:

$$w'_{ij} = w_{ij} + \eta \delta_j \frac{df_j(x)}{dx} y_i$$

where w_{ij} is, again, the (existing) weight between neurons i and j, δ_j is the value of the error signal at neuron j, $\frac{df_j(x)}{dx}$ is the derivative of the used function and y_i is the output of the i^{th} neuron. The coefficient η is chosen by the operator, or it's adjusted by the algorithm automatically. For the neurons in the first layer the y_i is replaced by the input x_i and w_{ij} is the weight of the i^{th} input variable at the j^{th} neuron. This step is depicted in 3.13, and it's repeated until the final error δ is sufficiently small.

Classification Rule

After building the classification model, i.e. finding the appropriate weights, the observations with unknown origin (observations in the test set) are used as input for the existing model. For each observation several outputs are calculated, one from each output neuron. Each output neuron corresponds to exactly 1 class. Hence, the observation with unknown origin is assigned to the class for which the corresponding output neurons output is the highest (closest to 1).

4 Experimental Comparison of the Chosen Methods

The following chapter deals with the experimental comparison of the selected classification methods, while taking into account several EDA steps. In addition, a statistical analysis of the distribution of LIBS data is carried out.

The experimental setup is described in the first section, followed by the description of the chosen samples in the second section.

Every step of the data analysis was carried out using the R statistical language [88]. The utilised libraries are briefly described in Appendix C: R Programming Language.

4.1 Experimental Setup and Parameters

The measurements were carried out by the Sci-Trace LIBS system (AtomTrace, CZ) [84], including the LIBS interaction chamber shown in 4.1. The excitation source is an Nd:YAG laser with its second harmonic wavelength of 532 nm, and pulse length of 10 ns (CFR 400, Quantel, FR).

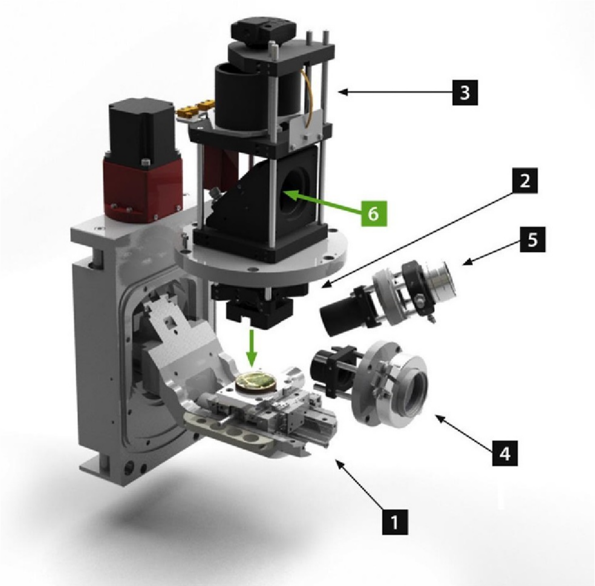

Figure 4.1: Inner view of the experimental chamber used in for the measurements: 1. motorized manipulator, 2. focusing optics of the primary laser beam, 3. optical system providing an overview of the sample, 4. focusing optics of the secondary laser beam[2], 5. collector optics for the plasma emission, 6. entrnce of the primary laser pulse into the chamber. Taken from [84].

The laser pulse is guided into the interaction chamber accommodating the sample by a series of mirrors (NB1-K13, Thorlabs, US). The sample itself is placed on a motorised manipulator enabling 3-axial movement.

Subsequently, the pulse is focused perpendicularly onto the surface of the sample by a triplet lens with a focal length of 25.4 mm (Sill Optics, GE).

[2] Used for double-pulse LIBS applications, not relevant for the present work.

The plasma emission is collected by an objective positioned under a 60° angle to the surface normal. The objective consists of an 100 mm focal length CaF2 and a 75 mm focal length UVFS lens (Thorlabs, US).

The collected light is delivered to the entrance slit of an echelle spectrometer with a 200-900 nm wavelength range (EMU 65, Catalina Scientific, US) through an optical fiber with a 400 μm core diameter.

The spectrally resolved light is finally recorded by an EMCCD camera (Falcon Blue, Raptor Photonic, IR) and saved on a personal computer.

The system and the measurement settings (see below) are controlled by the AtomChamber software (AtomTrace, CZ).

Experimental Parameters

The movement of charged particles (ions and electrons) inside the plasma results in the emission of a continuous radiation. This carries no valuable information for the analysis, and is regarded unwanted. Therefore, to minimise its presence in the recorded spectra, the EMCCD is gated. This means, that the camera starts recording the light emitted by the plasma only after a chosen amount of time (Gate Delay) and for a chosen time window (Gate Width).

These parameters influence the signal to noise ratio of the recorded spectra and have to be optimised. Nevertheless, no optimisation was carried out during the present work and the parameters were taken from a previous work done on the same system by Pořízka *et al.* [7]. The reason for this decision was the possibility of using the data from the present and past measurements for a future study, aimed at the system's stability.

The parameters used are, therefore, the following: the energy of the laser pulses were set to 50 mJ, and it was focused into a spot of 100 μm spot, resulting in a 64 GW.cm^{-2} irradiance. The gate delay was 1000 ns and the gate width 50 μs. One measurement was taken from each spot with 200 μm distance between each spot. 200 measurements were carried out on each sample (see below), resulting in 5200 spectra, 200 for each sample.

4.2 Samples

To assess and compare the capabilities of the three chosen classification methods, 26 steel samples have been selected, with similar chemical composition, which is shown in Appendix D: Samples.

No elaborate sample preparation has been carried out, as LIBS does not require any. However, the surface of the samples has been ground by a 80-grit sandpaper (Bosch, GE). Care has been taken not to contaminate the samples with the remains of the other ones and for each sample a new, clean piece of sandpaper has been used. Subsequently the samples were cleaned with isopropyl alcohol.

The aim of this sample preparation step was to minimise the effects of the differences in surface finish on the classification results. Different samples having different surfaces might result in the classification not reflecting the difference in chemical composition, but rather the laser-matter coupling properties.

4.3 Exploratory Data Analysis

The EDA process began by organising the dataset into an $m \times n$ matrix, with $m = 5200$ being the number of spectra collected during the measurement (200 spectra for each sample) and $n = 35000$ being the number of variables, the wavelength in the $\langle 240, 940 \rangle$ nm range with 0.02 nm resolution. The data was then mean centered, as a mathematical requirement for PCA.

Subsequently the dataset was divided into training and testing sets. For each sample 133 spectre were selected for training and 67 for testing. Any further analysis was carried out on the training set, leaving the testing set only for the final assessment of the classifiers.

Several outlier filtering and data normalisation techniques, and their influence on the resulting classification capabilities of the classifiers were assessed.

The utilised outlier filtering methods were:

- Outlier filtering by total energy: the spectra were individually integrated to get the total energy. The gap $\mathrm{E_{Gap}}$ between the first and third quartiles (see Appendix C: Statistics) were calculated. Spectra with total energy E where considered to be outliers if
$$|\mathrm{E} - \bar{\mathrm{E}}| > 3 \cdot \mathrm{E_{Gap}}$$
where $\bar{\mathrm{E}}$ is the mean total energy for the sample.

- Outlier filtering based on the Mahalanobis distance (see Appendix C: Statistics) of the spectrum from the center of the PCA model in the scores space: If the spectrum's distance $\mathrm{D_{Mah}}$ is greater than the standard deviation, the spectrum is removed as outlier.

- Outlier filtering based on the correlation of the spectra: The correlation between every spectra of the sample were calculated and the spectra with the lowest correlation were removed.

- Data reduction, if possible: SIMCA and PLS-DA carry out data reduction during the model building. The former builds a PCA model for each sample individually, while the latter performs a similar projection onto the latent variables. Hence, data reduction by PCA was only used for ANN, which, subsequently, used the PCs as input. SIMCA and PLS-DA, on the other hand, used the whole spectra as input.

The used data normalisation methods were:

- Normalisation to unit total energy: The spectra were divided by their numerical integral.

- Normalisation to unit maximum intensity: The spectra were divided by the intensity of the highest emission line.

- Scaling the intensity to $\langle 0, 1 \rangle$ range: The lowest value of the spectrum was subtracted from the spectrum, which was subsequently divided by the highest value of the spectrum.

The influence of the the different EDA approaches on the actual classification is presented in the next subsection. In this subsection, however, their effect on the statistical distribution of the data will be investigated.

Data Visualisation and Distribution Analysis

The outlier filtering and data normalisation steps were followed by initial observations of the data. For this purpose PCA was used to visualise the data. Some of the results of PCA are shown in figures 4.2 and 4.3 gained from the training set normalised to unit total energy and unit maximum intensity, respectively.

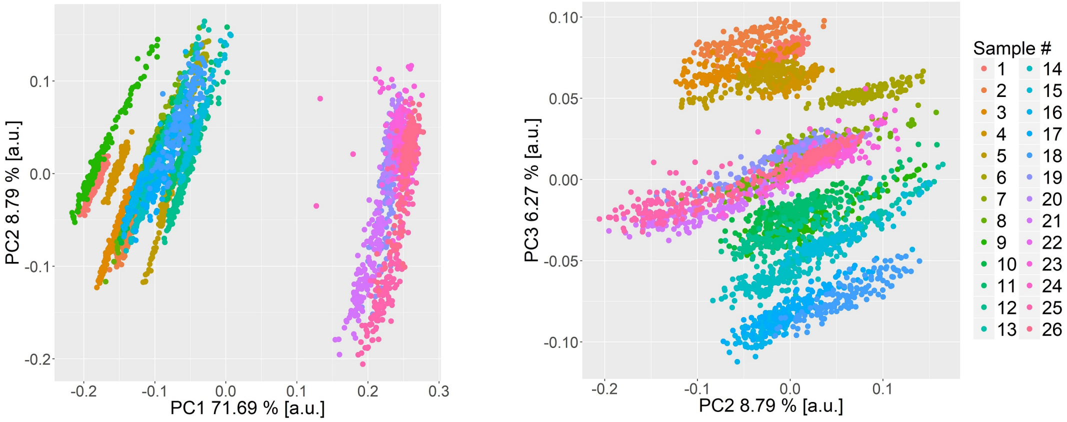

Figure 4.2: PCA scores plot of the training data normalised to unit total energy without outlier filtering. The figure on the left side shows the first two PCs plotted against each other. The percentage given in the axes labels gives the total variance explained by the respective principal component. The left figure shows the second and third principal components' scores. By plotting the data projected onto the scores space a clear separation of the spectra can be observed

Comparing figures 4.2 and 4.3, it can be observed that the higher the variance explained by the PC, the better separation is achieved (as is the case for PC1 in the shown cases).

The separation of the spectra by PCA shows, that the differentiation between samples is possible and the classification is achievable.

Following the visual inspection of the PCA scores plots, the data's distribution in the scores space was analysed by utilising both the Euclidean (fig. 4.4) and Mahalanobis distance (4.5) [3]. In addition, the distribution of the total energy was also investigated (shown in fig. 4.6).

The figures 4.4-4.6 clearly show, that the distribution of the spectra in the PCA scores space, and the distribution of the total energy of the spectra follow the extreme value distribution (EVD, described in Appendix B: Statistics).

[3] Both are described in Appendix B: Statistics

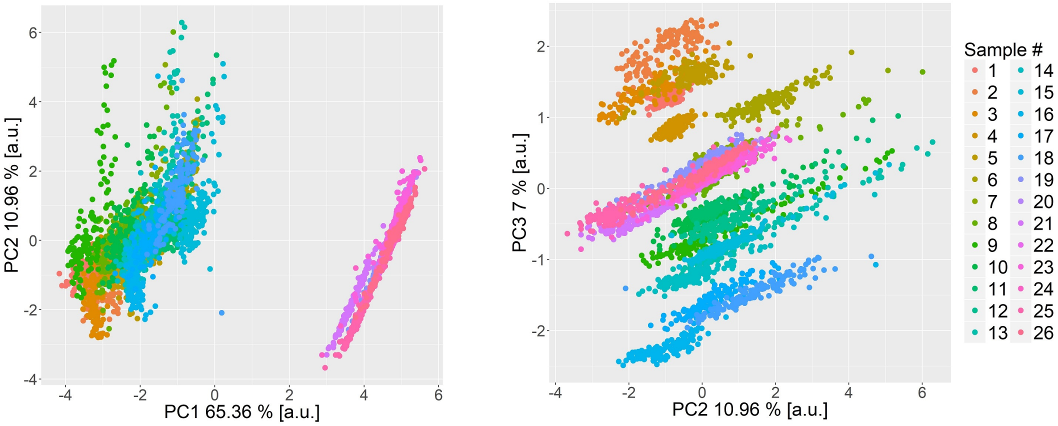

Figure 4.3: PCA scores plot of the training data normalised to unit maximum intensity. The meaning of the axes is the same as in fig. 4.2

.

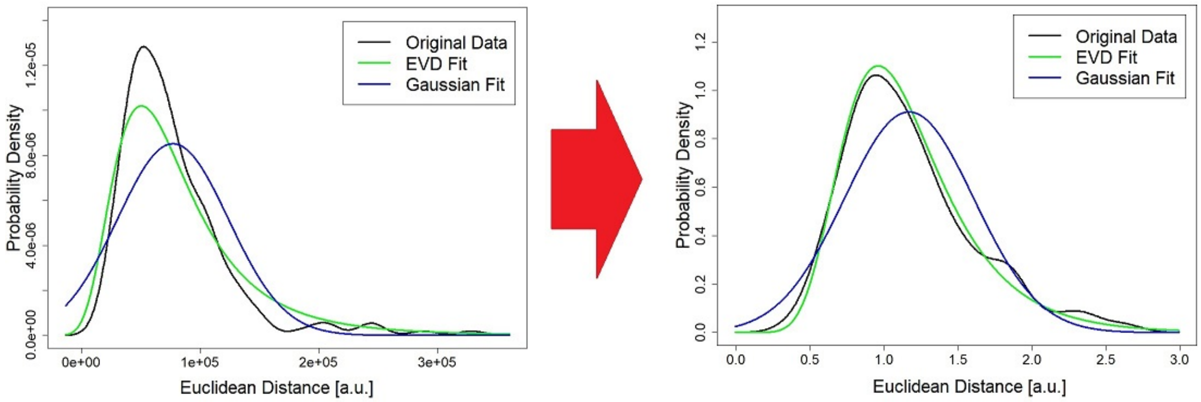

Figure 4.4: Statistical distribution of the Euclidean distance of the spectra in the PCA scores space from the center of the sample's model. Data shown for sample 3, without outlier filtering; without data normalisation on the left, and normalised to unit maximum intensity on the right.

The comparison of the distribution calculated from raw and normalised datasets shows, that the distribution of distances does not change, i.e. it follows the EVD after the normalisation as well, while the distribution of the total energy after normalisation seems to be closer to Gaussian, than for raw data. To quantify these claims, the Kolmogorov-Smirnov test (KS test, described in Appendix B: Statistics) was used as a goodness-of-fit test. The results are shown in table 4.1, along with the curves' parameters.

The results shown in table 4.1 agree with the ones reported by Klus *et al.* [86], who reported that during repeated measurements of the same sample, the distribution of the intensity of an emission line does not follow the normal distribution, but rather the EVD.

After the normalisation of the dataset, the distribution of the total energy is equally well described by both the normal distribution and the EVD. For the two distances, the normalisation results in a better fit for both distributions (lower D_{GEVD} and D_{Gauss}), but the EVD still fits the distribution better.

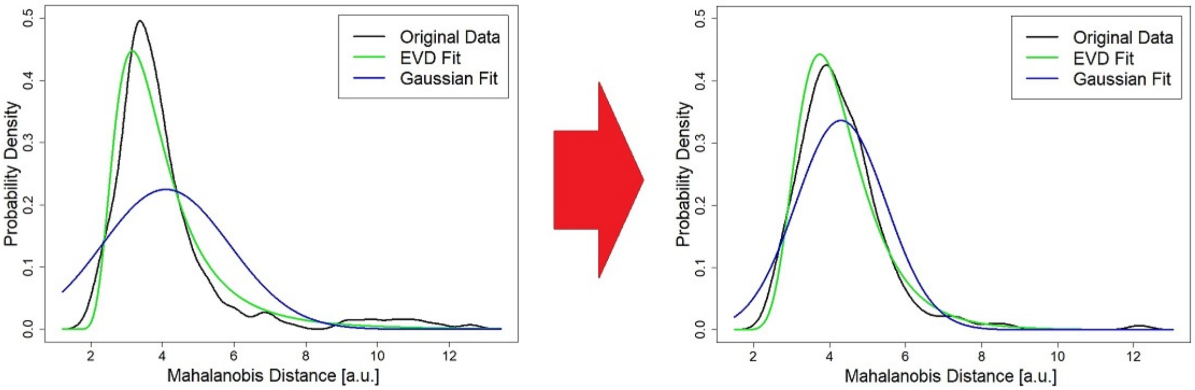

Figure 4.5: Statistical distribution of the Euclidean distance of the spectra in the PCA scores space from the center of the sample's model. Data shown for sample 3, without outlier filtering; without data normalisation on the left, and normalised to unit maximum intensity on the right.

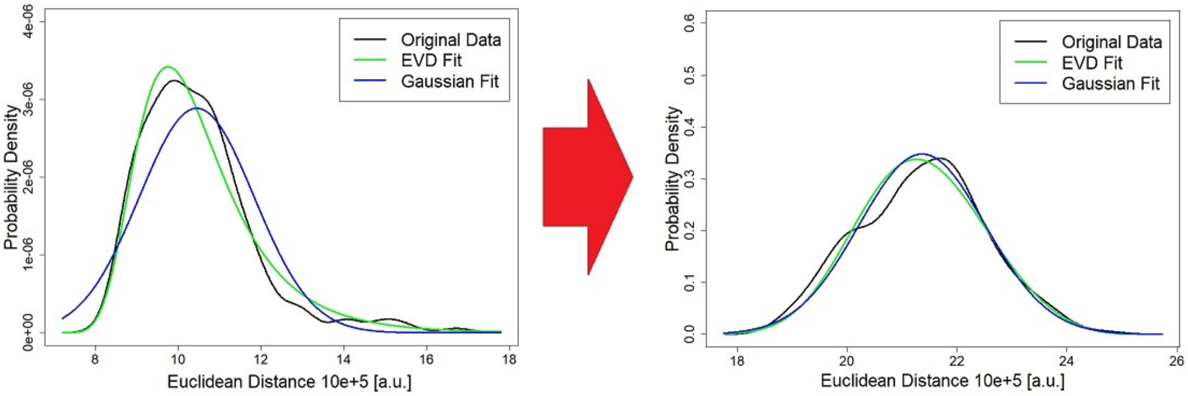

Figure 4.6: Statistical distribution of the total energy of the spectra. Data shown for sample 3, without outlier filtering; without data normalisation on the left, and normalised to unit maximum intensity on the right.

Table 4.1: The results D_{GEVD} is the result of the KS test on the generalised EVD, D_{Gauss} is the result of the KS test on the Gaussian distribution, μ_{GEVD} is the location parameter of the EVD, μ_{Gauss} is the mean of the normal distribution, σ_{GEVD} is the scale parameter of the EVD, σ_{Gauss} is the standard deviation of the normal distribution, and ξ is the shape parameter of the EVD. The values in bold show the values corresponding to the better fit, i.e. the distribution better describing the data.

	Non-normalised			Normalised to unit maximum intensity		
Attribute	$\mathrm{D}_{\mathrm{Mah}}$	$\mathrm{D}_{\mathrm{Euc}}$	TE	$\mathrm{D}_{\mathrm{Mah}}$	$\mathrm{D}_{\mathrm{Euc}}$	TE
D_{GEVD}	**0.180**	**0.170**	**0.089**	**0.135**	**0.089**	**0.062**
D_{Gauss}	0.224	0.298	0.190	0.186	0.121	**0.062**
μ_{GEVD}	3.33	$5.6 \cdot 10^4$	$9.8 \cdot 10^5$	3.78	0.97	20.95
μ_{Gauss}	7.32	$1.7 \cdot 10^5$	$1.25 \cdot 10^6$	7.27	1.49	21.75
σ_{GEVD}	0.84	$3.7 \cdot 10^4$	$1.1 \cdot 10^5$	0.83	0.33	1.13
σ_{Gauss}	3.54	$1.1 \cdot 10^5$	$3.1 \cdot 10^5$	3.34	0.87	2.31
ξ	0.23	0.16	0.06	0.05	0.02	-0.25

4.4 Classification

The following section lays out the method and results of the main goal of the present work, which is the experimental comparison of the chosen classification methods. First, the validation process and its results are described, followed by the testing of the 3 classifiers.

4.4.1 Validation

The validation process is aimed at determining the optimal model parameters for each classifier, which are the following:

- SIMCA: number of principal components.
- PLS-DA: number of latent variables.
- ANN: number of inputs (principal components), number of hidden layers (limited to 1 in the present study), number of neurons in the hidden layer

In addition to the method parameters, the optimal data preparation was also assessed. For each outlier filtering and normalisation this consisted of the following steps:

- Random division of the training set: 2/3 for model building and 1/3 for testing (the latter being the validation set). This resulted in 89 spectra in the training set and 44 in the validation set.
- Choosing an outlier and normalisation method.
- Filtering only(!) the model building set[4] and normalising both.
- Building the model from the filtered and normalised training set (or, in some cases, from the unfiltered and/or unnormalised training set)
- Testing for total accuracy by the validation set.
- Repeat from the first step, for 10 iterations in total.

During the validation process only the total accuracy of the methods was considered. This figure of merit is the percentage of correctly classified spectra:

$$\text{Total Accuracy} = \frac{\text{number of correctly classified spectra}}{\text{number of spectra in the validation set}}$$

The random division of the training set into model building and validation sets is crucial for the correct assessment of the optimal parameters. The reason being, there is a chance, that the model building would consist of highly correlated (or similar) spectra, while the testing set would contain a high number of outliers. In this case the resulting total accuracy would be low, not reflecting the true influence of the parameter.

This is avoided by randomly reorganising the two sets and repeating the calculation.

[4] Without knowledge about the origin of the spectrum, it's impossible to tell, if it's an outlier.

Validation of SIMCA

As mentioned above, the validation of SIMCA was aimed at the assessment of optimal number of PCs, outlier filtering and normalisation for the highest achievable total accuracy.

The influence of the outlier filtering and data normalisation on the resulting total accuracy is shown in figures 4.7 and 4.8, respectively. Finally, the results are summarised in figure 4.9.

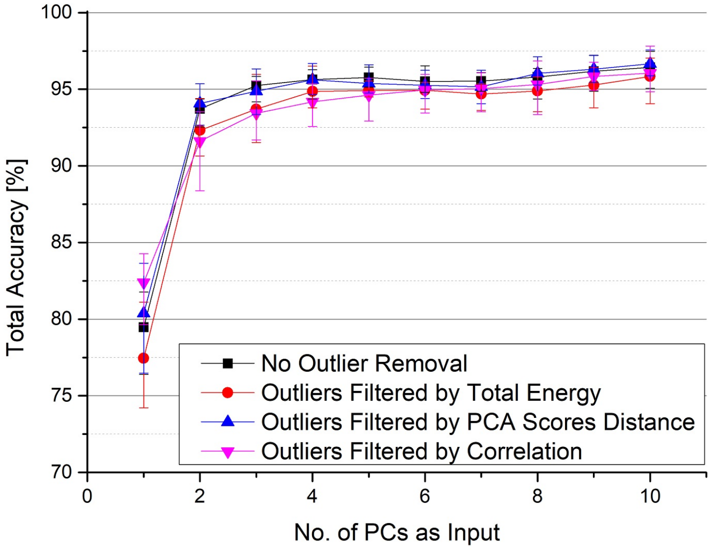

Figure 4.7: Total accuracy's dependency on the number of principal components (PCs) used for building the SIMCA model for the different outlier filtering techniques. The symbols show the mean values achieved during the validation using 10 iterations, while the error bars depict the lowest and highest total accuracy reached. A slight (1-2 %) difference can be seen up to 3 PCs. However, as the method reaches its maximum total accuracy (96 %) around 4 PCs, the outlier filtering no longer has any effect.

From figure 4.7 no significant improvement can be observed related to the outlier filtering. By using 4 or more principal components any effect of the outlier filtering diminishes, resulting in a small spread of the results, related to the randomised spectral selection.

Normalisation appears to significantly improve the classification capabilities of SIMCA, if only 1 PC is used for model building (figure 4.8). For 2 and more PCs, however, the effect of normalisation is vanishes.

As the increase in total accuracy reaches its maximum at 4-5 PCs used for model building (figures 4.8 and 4.7), the summary shown in fig. 4.9 was constructed by using 5 PCs. The highest achieved total accuracy by SIMCA is 97 %, with a mean performance around 95.5 %, regardless of the EDA, with one exception.

The outlier filtering technique based on the correlation of the spectra is slightly outperformed by every other data preparation method, even by the lack of outlier filtering.

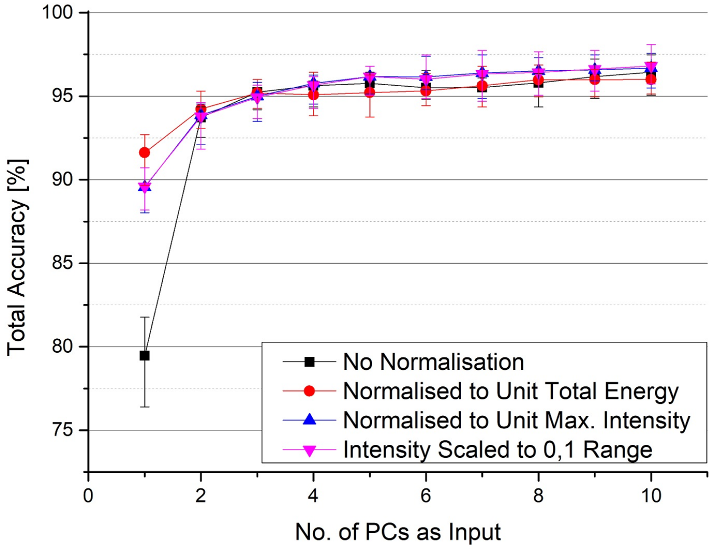

Figure 4.8: Effect of the normalisation on the total accuracy as function of the number of principal components used for model building. The symbols are the mean from 10 iterations, the error bars the minimum and maximum total accuracy values from the validation. For only 1 PC the normalisation of the dataset shows a clear advantage, regardless of the utilised method (from the 3 used in the present study). However, for 2 and more PCs the normalisation no longer leads to increased total accuracy.

This could be related to the issue of overfitting. By removing the spectra with the lowest correlation coefficient, the resulting model might be too tight. Another possibility is related to the amount of spectra removed. The outlier filtering based on correlation removes more spectra, than the other two, leading to an insufficient number to build precise enough models.

To test this hypothesis, the influence of the number of spectra used for model building on the resulting total accuracy was also investigated, and is shown in figure 4.10.

The results shown in fig. 4.10 show a significant decrease in performance of SIMCA if the training set size is limited. However, over a certain size of the training set (30 spectra for each sample in this case) the performance of SIMCA seems to be unaffected by the further increase of the training set's size.

This means, that decreased performance is SIMCA after the outlier filtering step based on correlation is linked to overfitting of the model, rather than the influence of the the training set's size. This outlier filtering approach removes 25 spectra, with the lowest correlation, leaving 64 in the training set. For this training set size the performance is unaffected.

Validation of PLS-DA

The validation of PLS-DA consisted of similar steps as the validation of SIMCA. The total accuracy was assessed as function of number of latent variables used for model building. This relation was evaluated for the different outlier filtering (example shown in fig. 4.11)

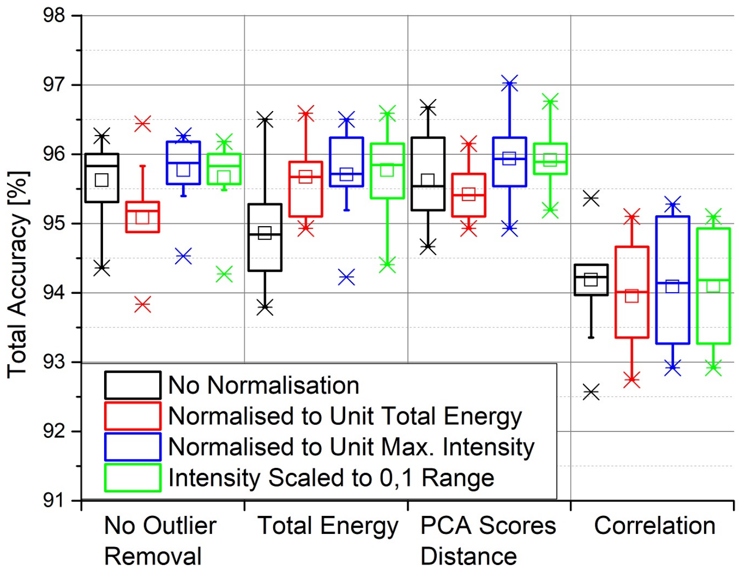

Figure 4.9: A box chart (the meaning of the chart is explained in Appendix B: Statistics) summarising the effect of every combination of outlier filtering and normalisation technique, constructed from the results using 5 principal components from 10 iterations. The colour of the boxes refers to the utilised normalisation, while the position gives the outlier filtering approach. The outlier filtering based on correlation seems to be slightly outperformed, but the distribution of the results overlap and no definite conclusion can be made.

and data normalisation methods (example shown in fig. 4.12). Finally, the influence of every combination of outlier filtering and data normalisation approach is summarised in fig. 4.13.

The range of the number of latent variables extends to 50, compared to the 10 used for SIMCA. This difference is comes from the limitation of the implemented algorithm for SIMCA (which allows the use of up to 10 principal components). Nevertheless, the SIMCA results reach a plateau around 4-5 PCs, and no significant further increase in total accuracy is to be expected. On the other hand, the total accuracy achieved by PLS-DA keeps increasing with each additional latent variable.

The total accuracy achieved as function of number of latent variables used for model building follows the same curve, regardless of the used outlier filtering (fig. 4.11). Over 23 latent variables the curves completely overlap, indicating that the outlier filtering does not lead to improved classification.

Data normalisation shows the same negligible effect (fig. 4.12) as outlier filtering.

The results shown in fig. 4.13 (using 25 latent variables for model building) are consistent with the previously discussed ones. The distribution of the results for each outlier filtering and data normalisation method overlap. Hence, there is no significant im-

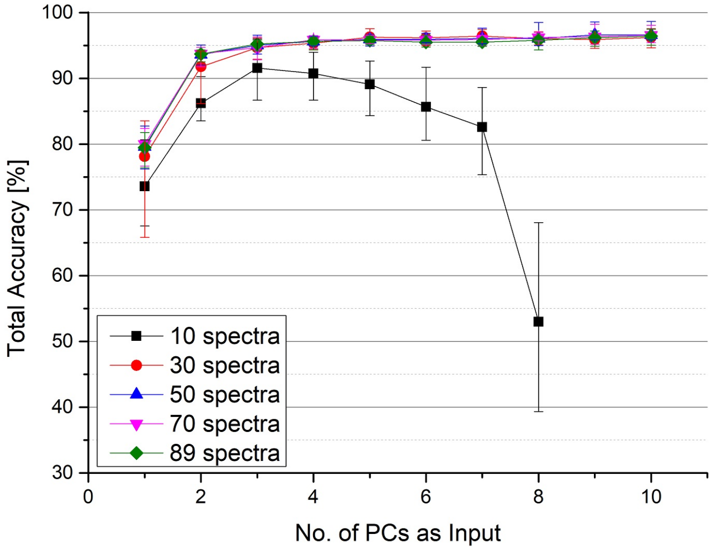

Figure 4.10: Total accuracy as function of principal components used for model building plotted for different training set sizes. The symbols represent the mean from 10 iterations, the error bars give the extreme values. The plot was constructed using raw spectra (without outlier filtering, or normalisation). A significant decrease in performance can be observed while using a training set containing only 10 spectra for each sample. However, the drop occurs only if 4 or more PCs are used.

provement of total accuracy achieved by either outlier filtering, nor data normalisation, achieving a total accuracy in the range of 94-98 % with every data preparation approach.

Although no implication was observed of the number of spectra used for model building influencing the classification capabilities of PLS-DA, it was investigated for the sake of completeness. The results are shown in fig. 4.14.

The results shown in fig. 4.14 suggest, that PLS-DA performs well, even with small training set sizes (10 spectra for each sample). No overfitting of the model can be observed either.

Validation of ANN

The validation of ANN follows follows the same steps as the previous two. The achieved total accuracy is assessed as function of number of principal components used as input[5] after the application of different outlier filtering (fig. 4.15) and data normalisation (fig. 4.16) methods. In addition, the influence of the number of neurons in the hidden layer is also investigated (fig. 4.17).

Unlike for the previous methods, data without normalisation was not considered (as mentioned above). This is due to the working principle of the method, which requires values in the range of $\langle 0, 1\rangle$ to converge properly.

[5] These principal components are the result of the PCA analysis carried out during EDA.

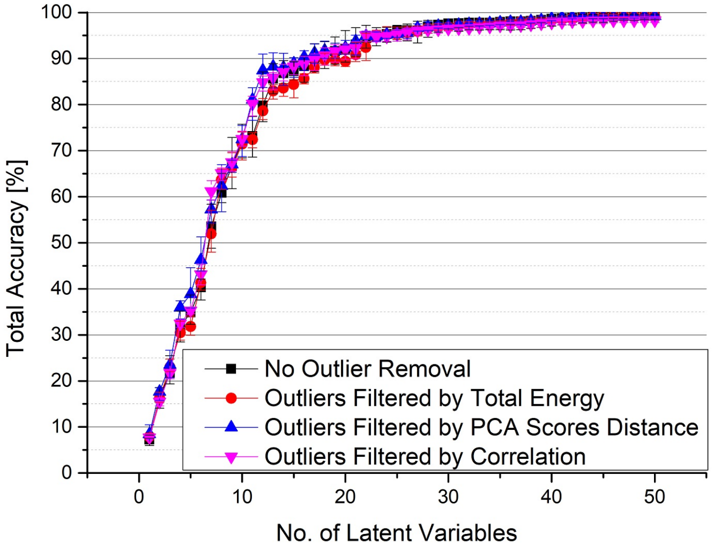

Figure 4.11: Total accuracy as function of the number of latent variables utilised for the building of the PLS-DA model and the influence of the outlier filtering. The symbols are the mean values, while the error bars show the minimum and maximum total accuracy achieved during the validation process, carrying out 10 iterations. Similarly to the results from the SIMCA validation, the influence of the different outlier filtering methods seems to be insignificant, as each dependency follows the same curve.

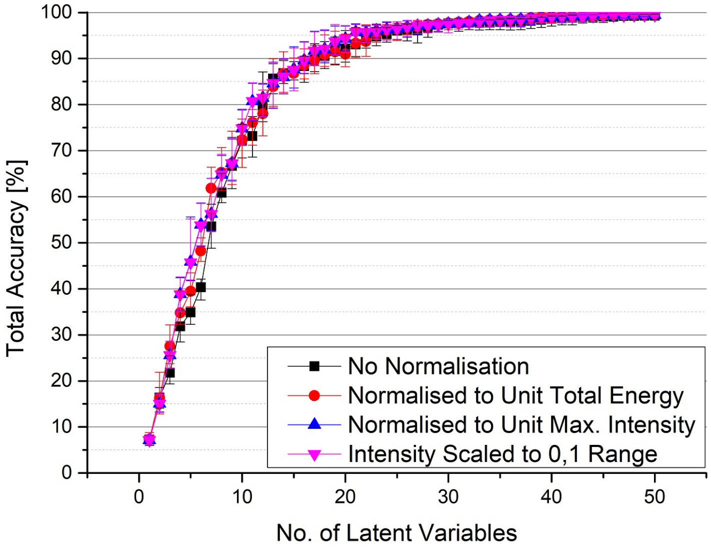

Figure 4.12: Total accuracy as function of the number of latent variables used for building the model with different data normalisation methods used on the data. The symbols are the mean values, while the error bars show the minimum and maximum total accuracy achieved during the validation process. As in the case of outlier filtering, the normalisation of the dataset doesn't affect the performance of PLS-DA either.

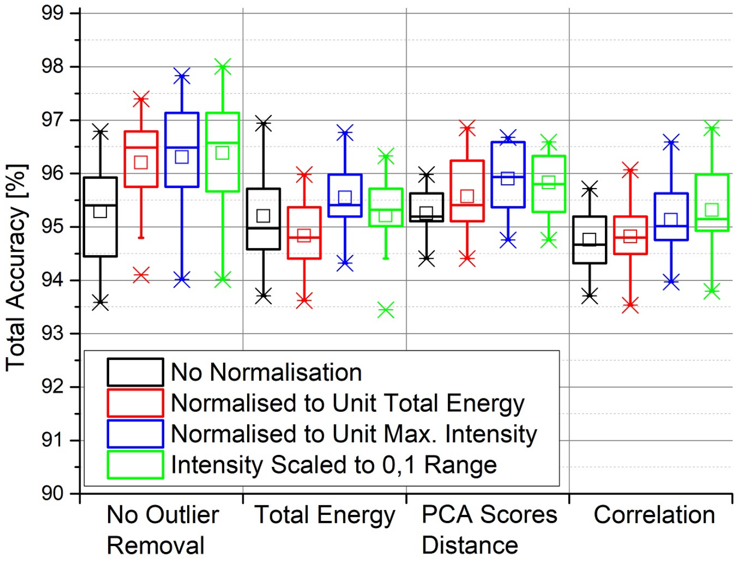

Figure 4.13: Box chart summarising the achieved results, constructed from the results utilising 25 latent variables. The colour of the boxes refers to the utilised normalisation, while the position gives the outlier filtering approach. No statistically significant difference can be observed.

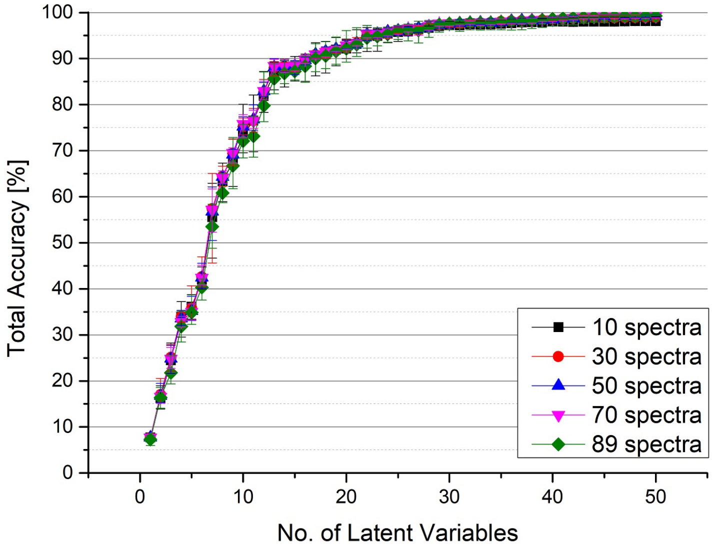

Figure 4.14: Total accuracy as function of number of latent variables plotted for different number of spectra contained in the training set for each sample. The symbols represent the mean from 10 iterations, the error bars show the extreme values. The size of the training set does not seem to affect the performance of PLS-DA in the investigated range.

The effect of the outlier filtering is shown in fig. 4.15. Apart from the method based on correlation, the outlier filtering shows no change in total accuracy. The method based

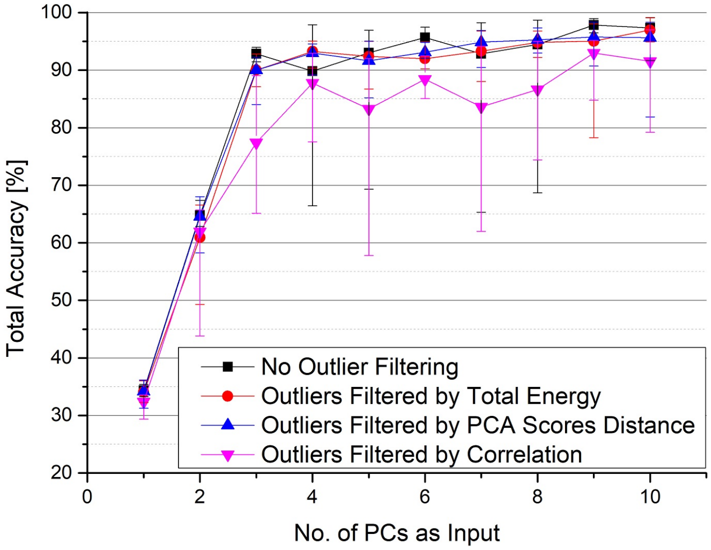

Figure 4.15: Total accuracy as function of the number of principal components used as input for the different outlier filtering approaches. The symbols represent the mean of the results, the error bars show the extreme values of the results. The plot was created by using 10 neurons in the hidden layer from the validation results of the spectra normalised to unit total energy. Similarly to the previous results, the outlier filtering doesn't seem to significantly influence the classification capabilities of the method.

on correlation results in a decreased achieved total accuracy. This, again might be due to overfitting resulting from the removal of weakly correlating spectra, or underfitting as result of leaving insufficient number of spectra for model building. The influence of model building training set size was assessed for further investigation.

The normalisation of the dataset doesn't seem to affect the total accuracy. The low extreme values are probably linked to the randomly chosen initial weights in the algorithm. This could result in very poor results, if the algorithm does not converge properly in the chosen number if iterations.

The influence of the number of neurons in the hidden layer on the resulting total accuracy seems to have no effect in the investigated range (fig. 4.17). With fewer neuron, however, the training of the model does not converge.

Apart from the few very low extreme values (for 6-8 and 13 neurons), there is no significant difference. There is no emerging pattern either. It can be concluded, that after reaching a sufficient number of neurons in the hidden layer, no more improvement can be achieved by further increasing their number.

A summarising comparison of every outlier filtering and data normalisation technique used is shown in fig. 4.18. The chart was constructed from the results of using 5 PCs as input and 12 neurons in the hidden layer. The outlier filtering seems to negatively affect the method's capabilities. The highest total accuracy (98 %) was achieved by normalising the dataset to unit total energy without outlier filtering. The most stable results came

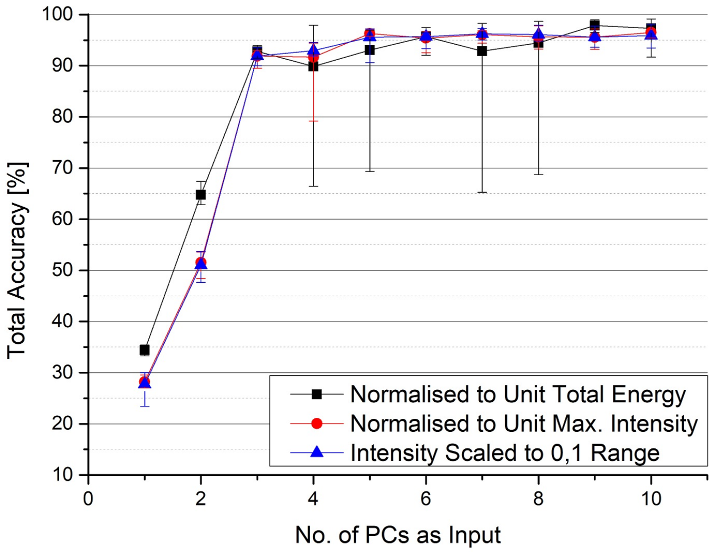

Figure 4.16: Total accuracy as function of the number of principal components used as input for the 3 data normalisation methods (without normalisation to unity the ANN algorithm does not converge). The symbols represent the mean values, the error bars show the extreme results from the validation. The plot was created by using 10 neurons in the hidden layer from the validation results of the spectra without outlier filtering. No significant difference in the results can be seen.

from the data normalised to unit maximum intensity.

As for the previous two methods, the influence of the number of spectra used for model building on the classification capabilities of ANN was investigated, and is shown in fig 4.19.

From fig. 4.19 it can be observed, that an insufficient number of spectra used for training the model will lead to significant loss in performance. After reaching a certain number, however, using additional spectra will not grant any advantages. It is also important to note, that no overfitting of the model is apparent from these results, as the accuracy of the classification does not drop with higher number of spectra.

Summary of the Validation Procedure

The influence of several outlier filtering and data normalisation methods on the classification performance of the chosen classifiers was assessed. The total accuracy was used as figure of merit. No significant improvement in classification capabilities was observed. This can be the sign of a stable experimental system, which minimises the need of outlier filtering.

In addition, the optimal parameters were found for all three methods:

- SIMCA's performance reaches a plateau at 4-5 principal components, with 96 % mean total accuracy.

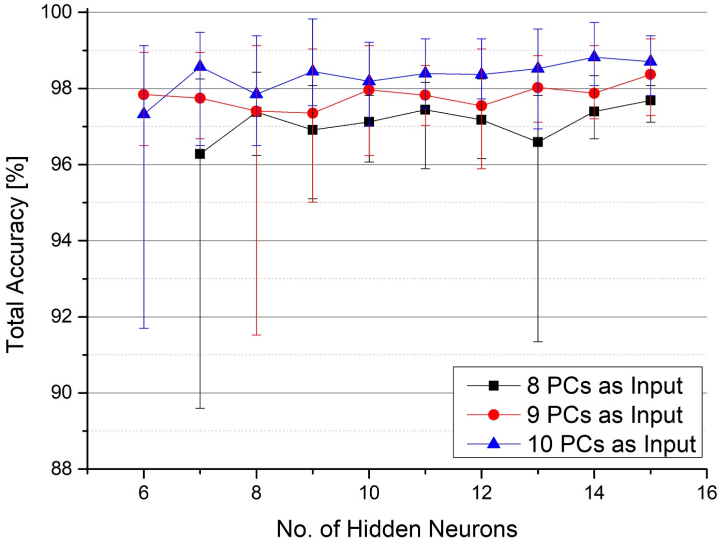

Figure 4.17: Total accuracy as function of number of additional neurons in the hidden layer for different number of principal components used as input. The symbols give the mean value, the error bars the the extreme values of the results. The plot was constructed from the validation results of the spectra normalised to unit total energy without outlier filtering. It's evident, that number of neurons in the hidden layer has no effect on the performance of the ANN.

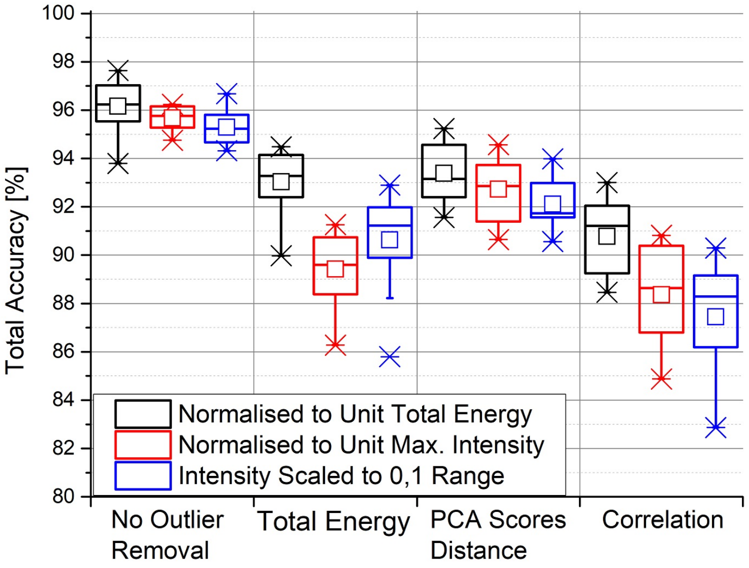

Figure 4.18: Box chart summarising the results of the ANN validation, constructed from the results using 5 principal components as input and 12 neurons in the hidden layer. The colour of the boxes refers to the utilised normalisation, while the position gives the outlier filtering approach. Best results are achieved without outlier filtering.

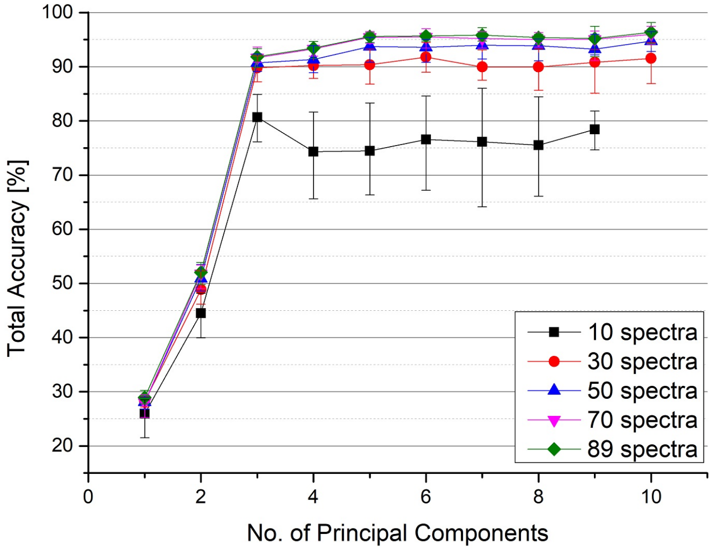

Figure 4.19: Total accuracy as function of number of PCs used as input plotted for different number of spectra utilised for the training of the neural network. The symbols represent the mean values from 10 iterations, the error bars show the extremes. The model performs consistently down to 30 spectra. At 10 spectra a significant drop in resulting total accuracy can be observed.

- PLS-DA's performance keeps increasing, slowly approaching perfect (100 %) total accuracy. For 30 latent variables the method's performance reaches 99 % total accuracy, any additional latent variable granting only a minuscule improvement.

- ANN's performance was found to depend only on the number of principal components used as input, reaching its highest mean performance at 96 % for 5 PCs.

Subsequently, the influence of the training set's size on the classifier's performance was investigated. It was concluded, that the number of spectra available for each sample is sufficient for reaching each model's highest performance.

4.4.2 Testing of the Classification Methods

Using the information about the optimal parameters gained from the validation, the testing of the classifiers was carried out. This consisted of repeating the following steps:

- Randomly removing a sample from the training set (randomly chosen in each iteration).

- Training the models using the truncated training set (both the model building and validation sets, resulting in 133 spectra for each sample).

- Each spectra in the testing set was classified (even the ones taken from the sample, which was removed from the training set).

The aim of the last step was to assess the methods' ability to recognise unknown observations. By removing a sample from the training set, its spectra in the testing set

should be identified as unknowns and not be added to any of the classes.
No outlier filtering was carried out. For the sake of consistent comparison, and due to ANN requiring some kind of data normalisation, the datasets were normalised to unit maximum intensity.

Introducing observations not belonging to any class to the testing process, new figures of merits can be used:

- Accuracy (different from total accuracy used during the validation [6]):

  $$\frac{TP+TN}{TP+FP+TN+FN}$$

  related to the method's ability to correctly classify an observation. TP = the number of true positives, TN = true negatives, FP = false positives, FN = false negatives.

- Sensitivity:

  $$\frac{TP}{TP+FN}$$

  gives the proportion of the correctly classified observations, which belong to a class. In other words, sensitivity ignores observations, which do not belong to any class.

- Specificity:

  $$\frac{TN}{TN+FP}$$

  related to the method's ability to identify observations not belonging to any class.

The SIMCA model was built using 5 PCs, the PLS-DA using 25 latent variables, the ANN using 5 PCs as input and 12 neurons in the hidden layer. The results are shown in figures 4.20-4.23.

As seen in fig. 4.20, PLS-DA achieved the highest mean accuracy (95 %). However, comparing this to the results shown in fig. 4.21, which shows the total accuracy, it can be concluded, that PLS-DA does not handle the classification of classless[7] observations well: the method's total accuracy (which is assessed by not having any classless observations in the test) is 4 % higher. This is further confirmed by the poor (virtually 0 %) specificity of PLS-DA, shown in fig. 4.23.

SIMCA showed little difference between its accuracy and total accuracy - a decrease of 1 %. In addition, it has the highest sensitivity, shown in fig. 4.22. This suggests, that SIMCA is less sensitive to classless observations than PLS-DA. However, while greatly outperforming PLS-DA in specificity with a mean value of 23 %, it is still unsatisfactory in identifying classless observations.

ANN showed very low reproducibility, or in other words, very high spread of results (over a 15 % range in accuracy, 17 % in sensitivity, and 80 % in specificity). In addition, the difference in ANN's accuracy and total accuracy (8 %) suggests that the method is very sensitive for classless observations being introduced to the testing set.

However, the highest achieved sensitivity (99 %) and specificity (81 %) were also the result of ANN. The reason of the low reproducibility of ANN's performance remains to be solved.

[6] If the testing set doesn't contain any observations not belonging to any class, the TN and FN values will be 0, leading to the definition of total accuracy.

[7] Not belonging to any class.

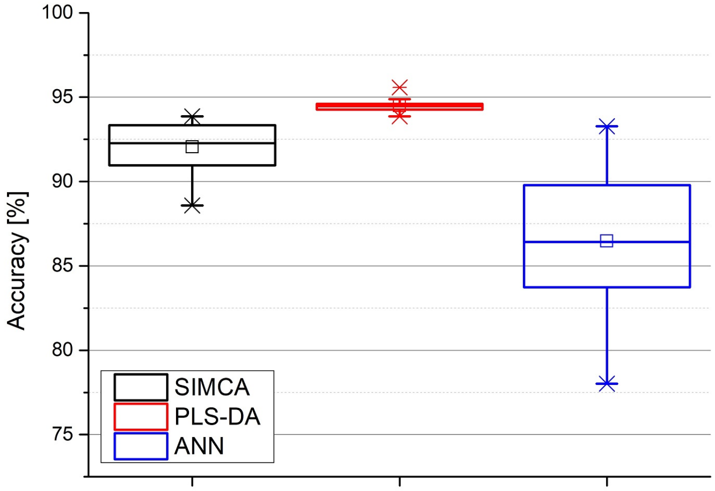

Figure 4.20: Accuracy of the three methods. The models were built by unfiltered data normalised to unit maximum intensity. 5 principal components were used for both the SIMCA and ANN models, 35 latent variables for the PLS-DA. PLS-DA clearly outperforms the other two methods.

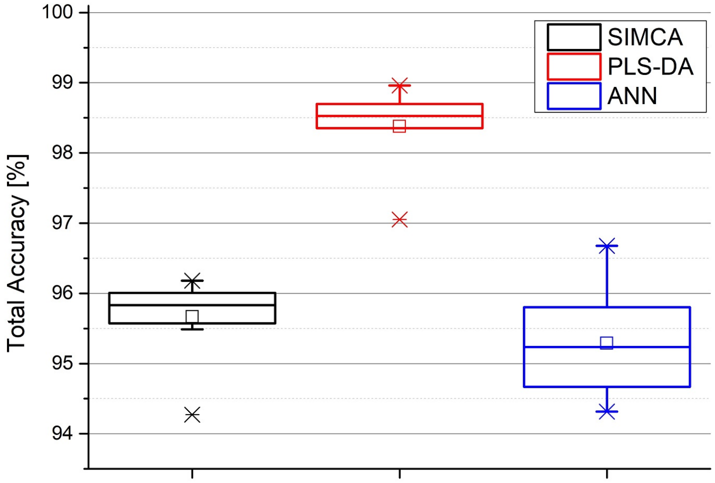

Figure 4.21: Total accuracy of the three methods achieved during the validation process, therefore no observations without class used for testing. 5 PCs used for building the SIMCA and ANN models, 35 latent variables for the PLS-DA. In each case, unfiltered data were used normalised to unit maximum intensity. PLS-DA clearly outperforms the other two methods, which show similar results.

4.4.3 Summary of the Comparison

The optimal EDA and parameters were found for each classifier. Outlier filtering, data normalisation and training set size (over a certain number of spectra) were found to have little effect on the results.

Subsequently the three methods were compared using the optimal parameters, using 26 metallic samples, unfiltered data normalised to unit maximum. The data normalisation was necessary due to the working principle of ANN.

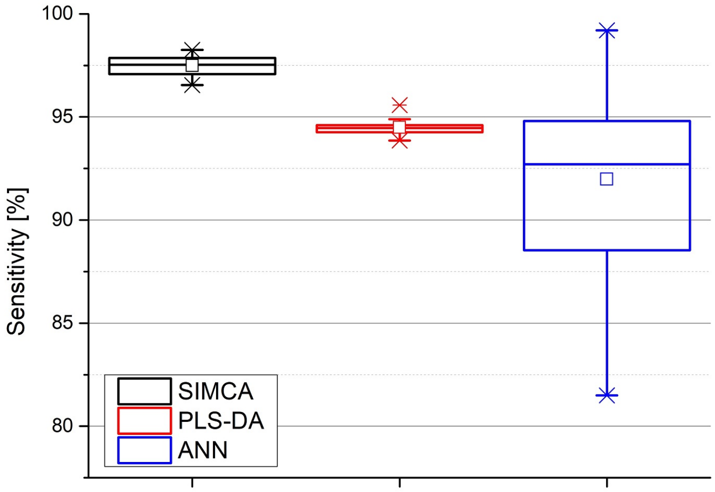

Figure 4.22: Sensitivity of the three methods. The models were built by unfiltered data normalised to unit maximum intensity. 5 principal components were used for both the SIMCA and ANN models, 35 latent variables for the PLS-DA. ANN shows both the best and worst sensitivity from the three, leading to very inconsistent results. SIMCA has stable and high sensitivity, outperforming the other two.

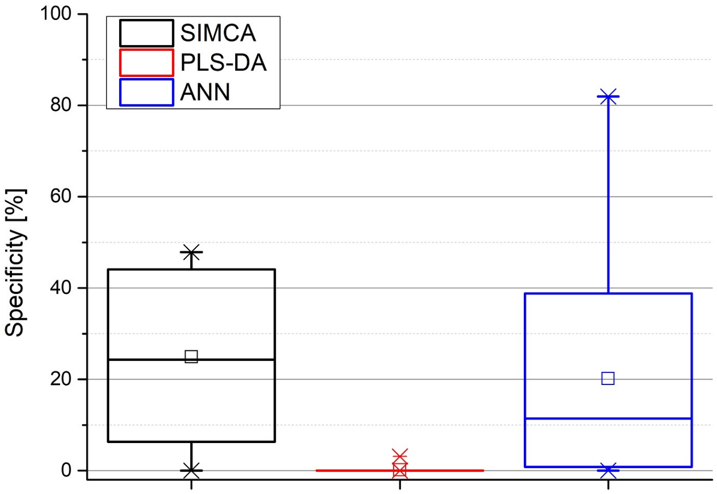

Figure 4.23: Specificity of the three classifiers. The models were built by unfiltered data normalised to unit maximum intensity. 5 principal components were used for both the SIMCA and ANN models, 35 latent variables for the PLS-DA. PLS-DA seems to be unable to identify observations not belonging to any class. ANN reaches 81 % specificity, however, only during one iteration.

PLS-DA showed the best classification accuracy. However, its accuracy dropped by 3 % after introducing observations not belonging to any class for testing. It was also unable to recognise if an unknown observation doesn't belong to any class.

SIMCA showed the second highest accuracy. In addition, its accuracy didn't decrease with the introduction of observations not belonging to any class for testing.

ANN showed lower accuracy than the other two methods and more importantly, lower reproducibility. The latter might be related to the algorithm used, which chooses the models initial weight values randomly at each iteration. This random step might lead significantly affect the stability of the method's learning process. More, in depth investigation is, however, necessary to validate these claims.

Conclusion

This master's thesis deals with the classification of metals by means of laser-induced breakdown spectroscopy and chemometric methods. It had three main goals. It had to map the possibilities of using LIBS for the classification of metals. Subsequently, the optimal classification methods had to be selected. Lastly, the experimental comparison of the selected methods had to be carried out.

Chapter 1 briefly laid out the basic principles of LIBS, listing the method's main advantages and limitations. The attributes of LIBS have been shown, that make LIBS a good candidate for industrial applications, including the classification of metals.

Subsequently, in chapter 2, a comprehensive review of the studies done on the present work's subject has been given. Due to the lack of reports dealing with the use of chemometric classification methods, the review was further extended to material classification in general.

Three widely used chemometric classifiers have been selected, which were described in chapter 3. In addition, various approaches to data preparation have been also briefly given.

The experimental comparison of the methods is described in chapter 4. After describing the experimental setup and the samples, exploratory data analysis was carried out. The observations (spectra) were visualised and their distribution in the PCA space was analysed. It was shown, in agreement with previous studies, that the LIBS spectra follow the extreme value distribution.

The process of the validation and testing of the 3 classification methods was given in the rest of the chapter. The validation was aimed at finding the optimal parameters of the methods, while the testing assessed their capabilities.

From the three investigated classifiers, the highest accuracy was achieved by PLS-DA, followed by SIMCA. ANN showed low reproducibility in the achieved classification.

All three methods showed a decrease in classification accuracy, once a set of classless samples were introduced to the testing set, however, SIMCA seemed to be the least affected one.

In summary, all three investigated classification methods show satisfactory accuracy, over 95 %, once the models are trained with data from every sample.

Appendix A: Linear Algebra

The contents of this appendix were worked out based on [87].
Theorem 1: The inverse of an orthogonal matrix is its transpose, i.e. $\boldsymbol{A}^{-1} = \boldsymbol{A}^T$ for any $m \times m$ orthogonal symmetric matrix $\boldsymbol{A}$. The proof follows from the way the ij^{th} element of matrix $\boldsymbol{A}^T\boldsymbol{A}$ is calculated: $(\boldsymbol{A}^T\boldsymbol{A})_{ij} = \boldsymbol{a}_i^T\boldsymbol{a}_j$, where $\boldsymbol{a}_i$ is the i^{th} column vector of $\boldsymbol{A}$.

Since we defined $\boldsymbol{A}$ as orthogonal matrix, its $\boldsymbol{a}_i$ components are orthogonal (or orthonormal after normalisation to unity). Therefore,

$$(\boldsymbol{A}^T\boldsymbol{A})_{ij} = \boldsymbol{a}_i^T\boldsymbol{a}_j = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

From which we get $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, where $\boldsymbol{I}$ is a unit matrix, and hence $\boldsymbol{A}^T = \boldsymbol{A}^{-1}$.
Theorem 2: For any matrix $\boldsymbol{A}$, the matrices $\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$ are both symmetric. The proof follows from the following two equalities:

$$(\boldsymbol{A}\boldsymbol{A}^T)^T = \boldsymbol{A}^{TT}\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{A}^T$$
$$(\boldsymbol{A}^T\boldsymbol{A})^T = \boldsymbol{A}^T\boldsymbol{A}^{TT} = \boldsymbol{A}^T\boldsymbol{A}$$

Theorem 3: A matrix is symmetrical if and only if it is orthogonally diagonalizable. This is a bi-directional statement, we will focus on proving only the forward case, which is used in the derivation of PCA.

Let $\boldsymbol{A}$ be an orthogonally diagonalizable matrix. Then $\boldsymbol{A}$ is a symmetric matrix. As proof, let's invoke the definition of an orthogonally diagonalizable matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^T$$

where $\boldsymbol{D}$ is a diagonal matrix and $\boldsymbol{E}$ is a transformation that diagonalizes $\boldsymbol{A}$. Now let's calculate $\boldsymbol{A}^T$:

$$\boldsymbol{A}^T = (\boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^T)^T = \boldsymbol{E}^{TT}\boldsymbol{D}^T\boldsymbol{E}^T = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^T$$

where we used the fact, that for a diagonal matrix $\boldsymbol{D}$:

$$\boldsymbol{D}^T = \boldsymbol{D}.$$

Theorem 4: A symmetric matrix is diagonalized by a matrix of its orthogonal eigenvectors. In other words, this states how to calculate the matrix $\boldsymbol{E}$ from the previous theorem:

$$\boldsymbol{E} = [\boldsymbol{e}_1\boldsymbol{e}_2 \dots \boldsymbol{e}_n]$$

where $\boldsymbol{e}_i$ is the i^{th} eigenvector of a symmetric, $n \times n$ matrix $\boldsymbol{A}$ (in a column vector form). Furthermore, let's define a diagonal matrix $\boldsymbol{D}$, with λ_i being the eigenvalues of $\boldsymbol{A}$, in its ii^{th} position.

Comparing the two sides of the equation $\boldsymbol{AE} = \boldsymbol{ED}$ we get $\boldsymbol{Ae}_i = \lambda_i \boldsymbol{e}_i$. This is, by definition, the eigenvalue equation. After rearranging: $\boldsymbol{A} = \boldsymbol{EDE}^{-1}$.

To show, that symmetric matrices have orthogonal eigenvectors we consider λ_1 and λ_2 to be two distinct eigenvalues of a symmetric matrix $\boldsymbol{A}$, and the 2 related eigenvectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$. Then:

$$\begin{aligned}
\lambda_1 \boldsymbol{e}_1 \cdot \boldsymbol{e}_2 &= (\lambda_1 \boldsymbol{e}_1)^T \boldsymbol{e}_2 \\
&= (\boldsymbol{A}\boldsymbol{e}_1)^T \boldsymbol{e}_2 \\
&= \boldsymbol{e}_1^T \boldsymbol{A}^T \boldsymbol{e}_2 \\
&= \boldsymbol{e}_1^T \boldsymbol{A} \boldsymbol{e}_2 \\
&= \boldsymbol{e}_1^T (\lambda_2 \boldsymbol{e}_2) \\
\lambda_1 \boldsymbol{e}_1 \cdot \boldsymbol{e}_2 &= \lambda_2 \boldsymbol{e}_1 \cdot \boldsymbol{e}_2 \\
(\lambda_1 - \lambda_2) \boldsymbol{e}_1 \cdot \boldsymbol{e}_2 &= 0
\end{aligned}$$

and since λ_1 and λ_2 are two distinct values, $\boldsymbol{e}_1 \cdot \boldsymbol{e}_2 = 0$. This proves, that $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are orthogonal, i.e. $\boldsymbol{E}$ is an orthogonal matrix. Recalling theorem 1:

$$\boldsymbol{A} = \boldsymbol{EDE}^T$$

Appendix B: Statistics

The contents of this appendix were worked out based on [19].

Mahalanobis Distance: The Mahalanobis distance $\mathrm{D}_{\mathrm{Mah}}$ is defined as

$$\mathrm{D}_{\mathrm{Mah}} = \sqrt{(\boldsymbol{x} - \bar{\boldsymbol{x}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}})}$$

where $\boldsymbol{x}$ is an observation, $\bar{\boldsymbol{x}}$ is the mean of the observations and $\boldsymbol{\Sigma}$ is the correlation matrix of the observations.

Extreme Value Distribution: The extreme value distribution (EVD), or more precisely the generalised EVD (GEVD) is defined as:

$$G(z) = \exp(-(1 + \xi(\frac{z - \mu}{\sigma}))^{-1/\xi})$$

where μ is called the location parameter, σ the scale parameter and ξ the shape parameter. μ and σ are analogous to the mean and standard deviation of the Gaussian distribution.

Kolmogorov-Smirnov Test: The Kolmogorov-Smirnov test (KS test) uses the comparison of the cumulative distribution function of two distributions to determine the goodness of fit. The KS statistic is defined as:

$$D_n = \sup |F_n(x) - F(x)|$$

where $F_n(x)$ is the empirical distribution function of the variable x and $F(x)$ is the cumulative distribution function of the same variable. The distribution with lower D_n value describes the variable's distribution more precisely.

Box Chart: The meaning of a box chart is explained in fig.

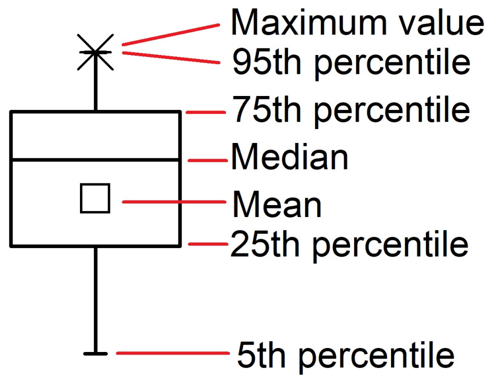

Figure 4.24: Explanation of the meaning of a box chart.

Quartile: The quartiles are 3 points, which divide the distribution of the data into 4 equal parts, each containing 25 % of the total number of observations.

Appendix C: R Programming Language

R is a free open source statistical programming language. The following list contains a brief summary of the libraries used for the analysis:

- ggbiplot: Advanced graphics for R. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

- matrixStats: Statistical functions operating on row and/or columns of matrices. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/matrixStats/matrixStats.pdf

- factoextra: Visualisation of multivariate data. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/factoextra/factoextra.pdf

- labeling: Axis labeling functions for nicer plotting options. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/labeling/labeling.pdf

- rrcovHD: High dimensional multivariate data handling functions, such as PCA. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/rrcovHD/rrcovHD.pdf

- nnet: Package for feedforward backpropagation neural network algorithms. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/nnet/nnet.pdf

- DiscriMiner: Functions for discriminant analysis and classification. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/DiscriMiner/DiscriMiner.pdf

- caret: Classification and regression training functions. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/caret/vignettes/caret.pdf

- car: Regression functions. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/car/car.pdf

- ggfortify: Advanced plotting functions, used, for example, for PCA plots. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/ggfortify/ggfortify.pdf

- data.table: Functions designed to handle manipulations with huge data sets. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/data.table/data.table.pdf

- klaR: Classification and visualisation functions. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/klaR/klaR.pdf

- evd: estimation and plot of extreme value distributions. Manual and detailed description available online at

  https://cran.r-project.org/web/packages/evd/evd.pdf

Appendix D: Samples

The following table contains the chemical composition of the samples used for the experimental comparison of the classifiers. Samples 1-5 reference materials provided by SPL Bohumin. Samples 6-18 are certified reference materials from BAM (Federal Institute for Materials Research and Testing, DE). The composition of samples 19-26 was measured by ICP-OES by Lithea, s.r.o., CZ.

Table 4.2: Table containing the chemical composition of the used samples in wt. %. Content of Fe is the supplement to 100 wt. %.

	Ni	Mn	Cr	C	Si	Mo	Co	Ti	Cu	Mg	S	P
Sample 1	0.52	0.524	9.58	0.142	0.612	1.116	0.0329	0.0236	0.201	–	0.0175	0.031
Sample 2	15.27	0.783	13.12	0.361	1.588	1.023	0.222	0.254	0.986	–	0.0182	0.044
Sample 3	7.93	1.431	18.26	0.044	0.381	0.271	0.151	0	0.389	–	0.0297	0.0284
Sample 4	11.58	1.87	19.7	0.163	1.56	0.625	0.062	0	0.307	–	0.0026	0.029
Sample 5	3.91	0.66	13.08	0.028	0.597	0.443	0.018	0	0.106	–	0.022	0.597
Sample 6	12.55	0.74	12.35	0.092	0.46	−−	−−	−−	−−	−−	−−	−−
Sample 7	9.24	1.38	17.31	0.066	0.405	0.092	0.053	−−	−−	−−	−−	−−
Sample 8	10.72	1.745	16.811	0.0201	0.537	2.111	0.0525	−−	−−	−−	−−	−−
Sample 9	20.5	0.791	25.39	0.086	0.57	−−	0.054	−−	−−	−−	−−	−−
Sample 10	10.2	1.311	17.84	0.0141	0.48	2.776	0.0184	−−	−−	−−	−−	−−
Sample 11	6.124	0.686	14.727	0.0103	0.374	0.0138	−−	−−	−−	−−	−−	−−
Sample 12	12.85	0.722	11.888	0.0345	0.463	0.0304	−−	−−	−−	−−	−−	−−
Sample 13	10.2	1.4	18.46	0.019	0.27	0.265	0.116	−−	−−	−−	−−	−−
Sample 14	8.9	1.7	17.96	0.143	1.41	−−	0.018	−−	−−	−−	−−	−−
Sample 15	5.66	0.89	14.14	0.05	0.21	1.59	0.22	−−	−−	−−	−−	−−
Sample 16	13.39	0.682	10.953	0.016	0.84	0.073	0.136	−−	−−	−−	−−	−−
Sample 17	17.96	1.031	9.89	0.034	0.73	−−	0.083	−−	−−	−−	−−	−−
Sample 18	4.36	0.971	18.73	0.105	1.09	1.71	0.079	−−	−−	−−	−−	−−
Sample 19	0.0234	0.186	0.0359	3.33	2.35	−−	−−	0.0122	0.0261	0.0349	0.0145	0.048
Sample 20	0.058	0.575	0.106	3.27	2.1	0.00125	−−	0.00816	0.0319	0.0433	0.0181	0.0287
Sample 21	0.37	0.225	0.346	2.84	5.34	0.654	0.00257	0.0269	0.107	−−	0.0152	0.0434
Sample 22	0.0527	0.528	0.107	2.93	2.56	0.0039	0.0324	0.0103	0.046	0.0467	0.0209	0.0348
Sample 23	0.0536	0.459	0.214	3.5	2.25	0.0421	0.0038	0.0112	0.107	−−	0.0723	0.208
Sample 24	0.079		0.074	3.18	2.79	0.0027	0.0048	0.0256	0.0183	−−	0.0172	0.067
Sample 25	0.889	0.217	0.0324	2.99	2.47	−−	−−	0.0116	0.99	0.0415	0.00661	0.428
Sample 26	−−	0.61	0.12	3.56	2.06	−−	−−	−−	0.12	−−	0.124	0.36

List of Symbols and Abbreviations

$\boldsymbol{A}^T$	transpose of matrix $\boldsymbol{A}$
$\boldsymbol{C}$	class matrix
c_{ij}	i^{th} term of the j^{th} eigenvector of a spectrum
$\boldsymbol{C}_k$	estimate of the response variables after k iterations
${}^{resid}\boldsymbol{C}$	class matrix with the effect of the calculated latent variables subtracted
δ	error signal
δ_i	error signal at the i^{th} neuron
$\boldsymbol{D}$	a diagonal matrix
D_{Gauss}	the result of the KS test on the Gaussian distribution
D_{GEVD}	the result of the KS test on the generalised EVD
$\mathrm{D_{Mah}}$	Mahalanobis distance
η	coefficient involved in the backpropagation step of ANN, chosen by the operator
$\boldsymbol{E}$	error matrix of the predictor variables
E	total energy of a spectrum
$\bar{\mathrm{E}}$	mean total energy of the spectra from a sample
$\boldsymbol{E}_{eig}$	matrix consisting of the eigenvectors of an other matrix
$\mathrm{E_{Gap}}$	gap between the first and third quartiles of the total energy distribution
$\boldsymbol{F}$	error matrix of the response variables
h_i	the output of the function in the i^{th} neuron
k	number of classes
λ_i	eigenvalue belonging to the i^{th} eigenvector of a spectrum
μ_{Gauss}	the mean of the normal distribution
μ_{GEVD}	the location parameter of the EVD
m	number of observations (spectra)
n	number of variables (wavelengths)

OD_j	orthogonal distance of the j^{th} spectrum
o_i	output from the i^{th} output neuron
p	number of variables in the transformed data
$\boldsymbol{P}$	predictor loadings matrix
$\boldsymbol{p}$	column vector, vector element of matrix $\boldsymbol{P}$
$\boldsymbol{Q}$	response loadings matrix
$\boldsymbol{q}$	column vector, vector element of matrix $\boldsymbol{Q}$
$\boldsymbol{R}$	transformation matrix
$\boldsymbol{r}_i$	i^{th} column vector of matrix $\boldsymbol{R}$
$\boldsymbol{R}_j$	transformation matrix belonging to the j^{th} sample
$\mathbb{R}^n$	n dimensional space
σ_{Gauss}	the standard deviation of the normal distribution
σ_{GEVD}	the scale parameter of the EVD
σ_{noise}	standard deviation of the noise
σ_{signal}	standard deviation of the signal
$\boldsymbol{\Sigma}_X$	covariance matrix of the original data $\boldsymbol{X}$
$\boldsymbol{\Sigma}_Y$	covariance matrix of the transformed data $\boldsymbol{Y}$
S_j	j^{th} score of a spectrum
$\boldsymbol{T}$	scores matrix
$\boldsymbol{t}$	column vector, vector element of matrix $\boldsymbol{T}$
t_i	i^{th} element of $\boldsymbol{t}$
$\boldsymbol{t}_i$	$\boldsymbol{t}$ vector from i^{th} iteration
V_i	variance explained by the i^{th} principal component
$\boldsymbol{w}$	weight vector
w_i	i^{th} element of $\boldsymbol{w}$
w_{ij}	weight of the connection between the i^{th} and j^{th} neuron
w'_{ij}	adjusted weight of the connection between the i^{th} and j^{th} neuron
ξ	the shape parameter of the EVD

$\boldsymbol{X}$	matrix of the observed data
$\boldsymbol{X}_j$	original spectra of the j^{th} sample
x_k	k^{th} variable of a spectrum
${}^{resid}\boldsymbol{X}$	original data with the effect of the calculated latent variables subtracted
y	estimated value of a response variable
$\boldsymbol{Y}$	transformed dataset
$\boldsymbol{Y}_j$	transformed spectra of the j^{th} sample
$\bar{\boldsymbol{y}}_j$	mean of the spectra of the j^{th} sample
$\boldsymbol{y}_{new}$	spectrum to be classified
y_{true}	true value of a response variable

ANN	Artificial Neural Networks
EDA	Exploratory Data Analysis
EMCCD	Electron Multiplying Charge Coupled Device
EVD	extreme value distribution
FN	False negative
FoM	Figures of Merit
FP	False positive
GEVD	generalised EVD
k-PCA	Kernel Principal Component Analysis
KS test	Kolmogorov Smirnov test
LIBS	Laser-Induced Breakdown Spectroscopy
LIP	Laser-Induced Plasma
LV	Latent variable
MLP	Multilayer Perceptron
MVA	Multivariate Analysis
PC	Principal Component
PCA	Principal Component Analysis

PLS-DA	Partial Least Squares Discriminant Analysis
SIMCA	Soft Independent Modelling of Class Analogy
SNR	Signal to Noise Ratio
SVM	Support Vector Machines
TN	True negative
TP	True positive
UVFS	Ultraviolet grade Fused Silica

Bibliography

[1] A. Miziolek, V. Palleschi and I. Schechter: Laser-Induced Breakdown Spectroscopy (LIBS) Fundamentals and Applications, Cambridge, UK: Cambridge University Press, 2006.

[2] D. A. Cremers and L. J. Radziemski: Handbook of Laser-Induced Breakdown Spectroscopy, New York, USA: John Wiley & Sons, Ltd., 2006.

[3] S. Duchene, V. Detalle, R. Bruder and J.B. Sirven: Chemometrics and Laser Induced Breakdown Spectroscopy (LIBS) Analyses for Identification of Wall Paintings Pigments, *Current Analytical Chemistry*, 2016, vol. 6, p. 60-65.

[4] A. K. Myakalwar, S. Sreedhar, I. Barman, N.C. Dingari, S.V. Rao, P.P. Kiran, S.P. Tewari, G.M. Kumar: Laser-Induced Breakdown Spectroscopy-based Investigation and Classification of Pharmaceutical Tablets Using Multivariate Chemometric Analysis, *Talanta*, 2011, vol. 87, p. 53-59.

[5] R. Noll, H. Bette, A. Brysch, M. Kraushaar, I. Monch, L. Peter, V. Sturm: Laser-Induced Breakdown Spectrometry - Applications for Production Control and Quality Assurance in The Steel Industry, *Spectrochimica Acta Part B*, 2001, vol. 56, p. 637-649.

[6] J. Gurell, A. Bengtson, M. Falkenstrom, B.A.M. Hansson: Laser Induced Breakdown Spectroscopy for Fast Elemental Analysis and Sorting of Metallic Scrap Pieces Using Certified Reference Materials, *Spectrochimica Acta Part B*, 2012, vol. 74-75, p. 46-50.

[7] P. Pořízka, J. Klus, A. Hrdlička, J. Vrábel, P. Škarková, D. Prochazka, J. Novotný, K. Novotný, J. Kaiser: Impact of Laser-Induced Breakdown Spectroscopy Data Normalisation on Multivariate Classification Accuracy, *Journal of Analytical Atomic Spectrometry*, 2016, vol. 00, p. 1-3.

[8] S. Wold and M. Sjöström: SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy, *ACS Symposium Series*, 1977, vol. 52.

[9] H. Wold: Multivariate Analysis, New York, USA: Academic Press, 1981.

[10] P.J. Werbos: Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences, PhD thesis, Harward University, 1974.

[11] J.D. Winefordner, I.B. Gornushkin, T. Correll, E. Gibb, B.W. Smith, N. Omenetto: Comparing Several Atomic Emission Spectrometric Methods to the Super Stars: Special Emphasis on Laser Induced Breakdown Spectrometry, LIBS, a Future Super Star, *Journal of Analytical Atomic Spectrometry*, 2004, vol. 19, p. 1061-1083.

[12] R. Brennetot, J.L. Lacour, E. Vors, A. Rivoallan, D. Vailhen, S. Maurice: Mars Analysis by Laser-Induced Breakdown Spectroscopy (MALIS): Influence of Mars Atmosphere on Plasma Emission and Study of Factors Influencing Plasma Emission with the Use of Doehlert Designs, *Applied Spectroscopy*, 2003, vol. 57, p. 744-752.

[13] G. Kim, J. Kwak, K.R. Kim, H. Lee, K.W. Kim, H. Yang, K. Park: Rapid Detection of Soils Contaminated with Heavy Metals and Oils by Laser Induced Breakdown Spectroscopy (LIBS) *Journal of Hazardous Materials*, 2013, vol. 263, p. 754-760.

[14] E.G. Snyder, C.A. Munson, J.L. Gottfried, F.C. De Lucia, B. Gullett, A. Miziolek: Laser-Induced Breakdown Spectroscopy for the Classification of Unknown Powders, *Applied Optics*, 2008, vol. 47, p. 80-87.

[15] F.C. De Lucia and J.L. Gottfried: Rapid Analysis of Energetic and Geo-materials Using LIBS, *Materials Today*, 2011, vol. 14, p. 274-281.

[16] T.H. Maiman: Optical and Microwave-Optical Experiments in Ruby, *Physical Review Letters*, 1960, vol. 4, p. 564.

[17] R.G. Brereton: Chemometrics Data Analysis for the Laboratory and Chemical Plant, New York, USA: John Wiley & Sons, Ltd., 2003.

[18] S. Wold: Chemometrics; What do We Mean with It, and What Do We Want from It? *Chemometrics and Intelligent Laboratory Systems*, 1995, vol. 30, p. 109-115.

[19] T. Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning Data Mining, Inference, and Prediction, Stanford, USA: Springer, 2008

[20] M. Corsi, G. Cristoforetti, M. Hidalgo, S. Legnaioli, V. Palleschi, A. Salvetti, E. Tognoni, C. Vallebona: Double Pulse, Calibration-free Laser-Induced Breakdown Spectroscopy: A New Technique for In Situ Standard-less Analysis of Pulluted Soils, *Applied Geochemistry*, 2006, vol. 21, p. 748-755.

[21] J. El Haddad, L. Canioni, B. Bousquet: Good Practices in LIBS Analysis: Review and Advices, *Spectrochimica Acta Part B*, 2014, vol. 101, p. 171-182.

[22] P. Pořízka, J. Klus, D. Prochazka, E. Képeš, A. Hrdlička, J. Novotný, K. Novotný, J. Kaiser: Laser-Induced Breakdown Spectroscopy Coupled with Chemometrics for the Analysis of Steel; The Issue of Spectral Outliers Filtering, *Spectrochimica Acta Part B*, 2016, vol. 123, p. 114-120.

[23] P. Devangad, V.K. Unnikrishnan, M.M. Tamboli, K.M.M. Shameem, R. Nayak, K.S. Choudhari, C. Santhosh: Quantification of Mn in Glass Matrices Using Laser Induced Breakdown Spectroscopy (LIBS) Combined with Chemometric Approaches, *Analytical Methods*, 2016, vol. 8, p. 7177-7184.

[24] R.B. Anderson, J.F. Bell, R.C. Wiens, R.V. Morris, S.M. Clegg: Clustering and Training Set Selection Methods for Improving the Accuracy of Quantitative Laser Induced Breakdown Spectroscopy, *Spectrochimica Acta Part B*, 2012, vol. 70, p. 24-32.

[25] F. Colao, R. Fantoni, P. Ortiz, M.A. Vazquez, J.M. Martin, R. Ortiz, N. Idris: Quarry Identification of Historical Building Materials by Means of Laser Induced Breakdown Spectroscopy, X-ray Fluorescence and Chemometric Analysis, *Spectrochimica Acta Part B*, 2009, vol. 65, p. 688-694.

[26] F.C. De Lucia, J.L. Gottfried, C.A. Munson, A.W. Miziolek: Multivariate Analysis of Standoff Laser-Induced Breakdown Spectroscopy Spectra for Classification of Explosive-containing Residues, *Applied Optics*, 2008, vol. 47, p. 112-121.

[27] S. Manzoor, S. Moncayo, F. Navarro-Villoslada, J.A. Ayala, R. Izquierdo-Hornillos, F.J. Manuel de Villena, J.O. Caceres: Rapid Identification and Discrimination of Bacterial Strains by Laser Induced Breakdown Spectroscopy and Neural Networks, *Talanta*, 2014, vol. 121, p. 65-70.

[28] J.B. Sirven, B. Bousquet, L. Canioni, L. Sarger, S. Teller, M. Potin-Gautier, I. Le Hecho: Qualitative and Quantitative Investigation of Chromium-polluted Soils by Laser Induced Breakdown Spectroscopy Combined with Neural Networks Analysis, *Analytical and Bioanalytical Chemistry*, 2006, vol. 385, p. 256-262.

[29] J.B Sirven, B. Salle, P. Mauchien, J.L. Lacour, S. Maurice, G. Manhes: Feasibility Study of Rock Identification at the Surface of Mars by Remote Laser Induced Breakdown Spectroscopy and Three Chemometric Methods, *Journal of Analytical Atomic Spectrometry*, 2007, vol. 22, p. 1471-1480.

[30] J.D. Hybl, G.A. Lithgow, S.G. Buckley: Laser Induced Breakdown Spectroscopy Detection and Classification of Biological Aerosols, *Applied Spectroscopy*, 2003, vol. 57, p. 1207-1215.

[31] A. Larsson, H. Andersson, L. Landstrom: Impact of Data Reduction on Multivariate Classification Models Built on Spectral Data From Bio-samples, *Journal of Analytical Atomic Spectrometry*, 2015, vol. 30, p. 1117-1127.

[32] A.M. Neiva, M.A.C. Jacinto, M.M. de Alencar, S.N. Esteves, E.R. Pereira-Filho: Proposition of Classification Models for the Direct Evaluation of the Quality of Cattle and Sheep Leathers Using Laser-Induced Breakdown Spectroscopy (LIBS) Analysis, *Journal of Analytical Atomic Spectrometry*, 2016, vol. 6, p. 104827-104838.

[33] Y. Lee, K.S. Ham, S.H. Han, J. Yoo, S. Jeong: Revealing Discrimination Power of the Elements in Edible Sea Salts: Line-intensity Correlation Analysis from Laser Induced Plasma Emission Spectra, *Spectrochimica Acta Part B*, 2014, vol. 101, p. 57-67.

[34] A.K. Myakalwar, N. Spegazini, C. Zhang, S.K. Anubham, R.R. Dasari, I. Barman, M.K. Gundawar: Less is More: Avoiding the LIBS Dimensionality Curse Through Judicious Feature Selection for Explosive Detection, *Scientific Reports*, 2015, vol. 5, p. 13169.

[35] J. Amador-Hernandez, J.M. Fernandez-Romero, M.D. Luque de Castro: Three-dimensional Analysis of Screen-printed Electrodes by Laser Induced Breakdown Spectrometry and Pattern Recognition, *Analytica Chimica Acta*, 2001, vol. 435, p. 227-238.

[36] J.J. Remus, J.L. Gottfried, R.S. Harmon, A. Draucker, D. Baron, R. Yohe: Archeological Applications of Laser-Induced Breakdown Spectroscopy: an Example from the Coso Volcanic Field, California, Using Advanced Statistical Signal Processing Analysis, *Applied Optics*, 2010, vol. 49, p. 120-131.

[37] J.J. Remus, R.S. Harmon, R.R. Hark, G. Haverstock, D. Baron, I.K. Potter, S.K. Bristol, L.J. East: Advanced Signal Processing Analysis of Laser-Induced Breakdown Spectroscopy Data for the Discrimination of Obsidian Sources, *Applied Optics*, 2012, vol. 51, p. 65-73.

[38] E. Vors, K. Tchepidijian, J.B. Sirven: Evaluation and Optimization of the Robustness of a Multivariate Analysis Methodology for Identification of Alloys by Laser Induced Breakdown Spectroscopy, *Spectrochimica Acta Part B*, 2016, vol. 117, p. 16-22.

[39] S. Moncayo, S. Manzoor, F. Favarro-Villoslada, J.O. Caceres: Evaluation of Supervised Chemometric Methods for Sample Classification by Laser Induced Breakdown Spectroscopy, *Chemometrics and Intelligent Laboratory Systems*, 2015, vol. 146, p. 354-364.

[40] J. Moros, J. Serrano, C. Sanchez, J. Macias, J.J. Laserna: New Chemometrics in Laser-Induced Breakdown Spectroscopy for Recognizing Explosive Residues, *Journal of Analytical Atomic Spectrometry*, 2012, vol. 27, p. 2111-2122.

[41] R.A. Putnam, Q.I. Mohaidat, A. Daabous, S.J. Rehse: A Comparison of Multivariate Analysis Techniques and Variable Selection Strategies in a Laser-Induced Breakdown Spectroscopy Bacterial Classification, *Spectrochimica Acta Part B*, 2013, vol. 87, p. 161-167.

[42] K. Pearson: On Lines and Planes of Closest Fit to Systems of Points in Space, *Philosophical Magazine*, 1901, vol. 2, p. 301.

[43] G. Strang: Introduction to Linear Algebra, Massachusetts, USA: Wellesley-Cambridge Press, 2009.

[44] J.L. Gottfried, R.S. harmon, F.C. De Lucia, A.W. Miziolek: Multivariate Analysis of Laser-Induced Breakdown Spectroscopy Chemical Signatures for Geomaterial Classification, *Spectrochimica Acta Part B*, 2009, vol. 64, p. 1009-1019.

[45] S.M. Clegg, E. Sklute, M.D. Dyar, J.E. Barefield, R.C. Wiens: Multivariate Analysis of Remote Laser-Induced Breakdown Spectroscopy Spectra Using Partial Least Squares, Principal Component Analysis and Related Techniques, *Spectrochimica Acta Part B*, 2009, vol. 64, p. 79-88.

[46] J. Lasue, R.C. Wiens, T.F. Stepinski, O. Forni, S.M. Clegg, S. Maurice: Nonlinear Mapping Technique for Data Visualisation and Clustering Assessment of LIBS data: Application to ChemCam Data, *Analytical and Bioanalytical Chemistry*, 2011, vol. 400, p. 3247-3260.

[47] D.C. Alvey, K. Morton, R.S. Harmon, J.L. Gottfried, J.J. Remus, L.M. Collins, M.A. Wise: Laser-Induced Breakdown Spectroscopy-based Geochemical Fingerprinting for the Rapid Analysis and Discrimination of Minerals: the Example of Garnet, *Applied Optics*, 2010, vol. 49, p. 168-180.

[48] D. Pokrajac, T. Vance, A. Lazarevic, A. Marcano, Y. Markushin, N. Melikeche, N. Reljin: Performance of Multilayer Perceptrons for Classification of LIBS Protein

Spectra, *10th Symposium on Neural Network Applications in Electrical Engineering*, 2010, University of Belgrade, Serbia.

[49] D. Pokrajac, A. Lazarevic, V. Kecman, A. Marcano, Y. Markushin, T. Vance, N. Reljin, S. McDaniel, N. Melikechi: Automatic Classification of Laser-Induced Breakdown Spectroscopy (LIBS) Data of Protein Biomarker Solutions, *Applied Spectroscopy*, 2014, vol. 68, p. 1067-1075.

[50] N. Labbe, I.M. Swamidoss, N. Andre, M.Z. Martin, T.M. Young, T.G. Rials: Extraction of Information from Laser-Induced Breakdown Spectroscopy Spectral Data by Multivariate Analysis, *Applied Optics*, 2008, vol. 47, p. 158-168.

[51] C.A. Munson, F.C. De Lucia, T. Piehler, K.L. McNesby, A.W. Miziolek: Investigation of Statistics Strategies for Improving the Discrimination Power of Laser-Induced Breakdown Spectroscopy for Chemical and Biological Warfare Agent Simulants, *Spectrochimica Acta Part B*, 2005, vol. 60, p. 1217-1224.

[52] F.M.V. Pereira, D.M.B.P. Milori, A.L. Venancio, M.S.T. Russo, P.K. Martins, J. Freitas-Astua: Evaluation of the Effects of Candidatus Liberibacter Asiaticus on Inoculated Citrus Plants Using Laser-Induced Breakdown Spectroscopy (LIBS) and Chemometric Tools, *Talanta*, 2010, vol. 83, p. 351-356.

[53] F.Y. Yueh, H. Zheng, J.P. Singh, S. Burgess: Preliminary Evaluation of Laser-Induced Breakdown Spectroscopy for Tissue Classification, *Spectrochimica Acta Part B*, 2012, vol. 64, p. 1059-1067.

[54] N.C. Dingari, I. Barman, A.K. Myakalwar, S.P. Tewari, M.K. Gundawar: Incorporation of Support Vector Machines in the LIBS Toolbox for Sensitive and Robust Classification Amidst Unexpected Sample and System Variability, *Analytical Chemistry*, 2012, vol. 84, p. 2686-2694.

[55] A.K. Myakalwar, N.C. Dingari, R.R. Dasari, I. Barman, M.K. Gundawar: Non-Gated Laser-Induced Breakdown Spectroscopy Provides a Powerful Segmentation Tool on Concomitant Treatment of Characteristic and Continuum Emission, *Plos One*, 2014.

[56] A.C. Samuels, F.C. De Lucia, K.L. McNesby, A.W. Miziolek: Laser-Induced Breakdown Spectroscopy of Bacterial Spores, Molds, Pollens, and Protein: Initial Studies of Discrimination Potential, *Applied Optics*, 2003, vol. 42, p. 6205-6209.

[57] L. Landstrom, A. Larsson, P.A. Gradmark, L. Orebrand, P.O. Andersson, P. Wasterby, T. Tjarnhage: Detection and Monitoring of CWA and BWA Using LIBS, *Proceedings of SPIE*, 2014, vol. 9073.

[58] V.K. Unnikrishnan, K.S. Choudhari, S.D. Kulkarni, R. Nayak, V.B. Kartha, C. Santhosh: Analytical predictive Capabilities of Laser Induced Breakdown Spectroscopy (LIBS) with Principal Component Analysis (PCA) for Plastic Classification, *RSC Advances*, 2013, vol. 3, p. 25872-25880.

[59] M. Hoehse, A. Paul, I. Gornushkin: Multivariate Classification of Pigments and Inks Using Combined Raman Spectroscopy and LIBS, *Analytical and Bioanalytical Chemistry*, 2011, vol. 402, p. 1443-1450.

[60] Q.Q. Wang, K. Liu, H. Zhao: Multivariate Analysis of Laser-Induced Breakdown Spectroscopy for Discrimination between Explosives and Plastics, *Chinese Physical Letters*, 2011, vol. 29, p. 044206:1-3.

[61] K. Varmuza and P. Filzmoser: Introduction to Multivariate Statistical Analysis in Chemometrics, New York, USA: Taylor & Francis Group, 2008.

[62] M.J.C. Pontes, J. Cortez, R.K.H. Galvao, C. Pasquini, M.C.U. Araujo, R.M. Coelho, M.K. Chiba, M.F. de Abreu, B.E. Madari: Classification of Brazilian Soils by Using LIBS and Variable Selection in the Wavelet Domain, *Analytica Chimica Acta*, 2009, vol. 642, p. 12-18.

[63] M. Barker and W. Rayens: Partial Least Squares for Discrimination, *Journal of Chemometrics*, 2003, vol. 17, p. 166-173.

[64] J. Gottfries, K. Blennow, A. Wallin, C.G. Gottfries: Diagnosis of Dementias Using Partial Least Squares Discriminant Analysis, *Dementia*, 1995, vol. 6, p. 83-88.

[65] S. Wold, L. Eriksson, J. Trygg, N. Kettaneh: The PLS Method - Partial Least Squares Projections to Latent Structures - and Its Applications in Industrial RDP (Research, Development, and Production), Umea, Sweden: Umea University Press, 2004.

[66] R.G. Brereton and G.R. Lloyd: Partial Least Squares Discriminant Analysis: Taking the Magic Away, *Journal of Chemometrics*, , vol. 28, p. 213-225.

[67] Y. Tian, Z. Wang, X. Han, H. Hou, R. Zheng: Comparative Investigation of Partial Least Squares Discriminant Analysis and Support Vector Machines for Geological Cuttings Identification Using Laser-Induced Breakdown Spectroscopy, *Spectrochimica Acta Part B*, 2014, vol. 102, p. 52-57.

[68] R.R. Hark, J.J Remus, L.J. East, R.S. Harmon, M.A. Wise, B.M. Tansi, K.M. Shughrue, K.S. Dunsin, C. Liu: Geographical Analysis of Conflict Minerals Utilising Laser-Induced Breakdown Spectroscopy, *Spectrochimica Acta Part B*, 2012, vol. 74-75, p. 131-136.

[69] X. Zhu, T. Xu, Q. Lin, L. Liang, G. Niu, H. Lai, M. Xu, X. Wang, H. Li, Y. Duan: Advanced Statistical Analysis of Laser-Induced Breakdown Spectroscopy Data to Discriminate Sedimentary Rocks Based on Czerny-Turner and Echelle Spectrometer, *Spectrochimica Acta Part B*, 2014, vol. 93, p. 8-13.

[70] J.L. Gottfried, F.C. De Lucia, C.A. Munson, A.W. Miziolek: Standoff Detection of Chemical and Biological Threats Using Laser-Induced Breakdown Spectroscopy, *Applied Spectroscopy*, 2008, vol. 64, p. 353-363.

[71] E. Alpaydin: Introduction to machine learning, , Massachusetts, USA: MIT Press, 2010

[72] A. Koujelev, M. Sabsabi, V. Motto-Ros, S. Laville, S.L. Lui: Laser-Induced Breakdown Spectroscopy with Artificial Neural Network Processing for Material Identification, *Planetary and Space Science*, 2010, vol. 58, p. 682-690.

[73] J. El Haddad, D. Bruyere, A. Ismael, G. Gallou, V. Laperche, K. Michel, L. Canioni, B. Bousquet: Application of a Series of Artificial Neural Networks to On-site Quantitative LIBS Analysis of Lead Into Real Soil Samples, *Spectrochimica Acta Part B*, 2014, vol. 97, p. 57-64.

[74] F. Anabitarte, J. Mirapeix, O.M.C. Portilla, J.M. Lopez-Higuera, 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*, 2012, vol. 12, p. 64-70.

[75] B. Noharet, E. Zetterlund, O. Tarasenko, M. Lindblom, J. Gurell, A. Bengtson, P. Lundin: Spectroscopy-based Photonic Instrumentation for the Manufacturing Industry: Contactless Measurement of Distance, Temperature and Chemical Composition, *Photonic Instrumentation Engineering*, 2014, vol. 8992, p. 89920R:1-8.

[76] A. Jurado-Lopez, M.D. Luque de Castro: Rank Correlation of Laser-Induced Breakdown Spectroscopic Data for the Identification of Alloys Used in Jewelry Manufacture, *Spectrochimica Acta Part B*, 2003, vol. 58, p. 1291-1299.

[77] S. Merk, C. Scholz, S. Florek, D. Mory: Increased Identification Rate of Scrap Metal Using Laser Induced Breakdown Spectroscopy Echelle Spectra, *Spectrochimica Acta Part B*, 2015, vol. 112, p. 10-15.

[78] S.R. Goode, S.L. Morgan, R. Hoskins, A. Oxsher: Identifying Alloys by Laser-Induced Breakdown Spectroscopy with a Time-resolved High Resolution Echelle Spectrometer, *Journal of Analytical Atomic Spectrometry*, 2000, vol. 15, p. 1133-1138.

[79] P. Werheit, C. Fricke-Begemann, M. Gesing, R. Noll: Fast Single Piece Identification with 3D Scanning LIBS for Aluminium Cast and Wrought Alloys Recycling, *Journal of Analytical Atomic Spectrometry*, 2011, vol. 26, p. 2166-2174.

[80] H.Y. Kong, L.X. Sun, J.T. Hu, Y. Xin, Z.B. Cong: A Comparative Study of Two Data Reduction Methods for Steel Classification Based on LIBS, *Applied Mechanics and Materials*, 2014, vol. 644-650, p. 4722-4725.

[81] M. Grzegorzek, D. Schwerbel, D. Balthasar, D. Paulus: Automatic Sorting of Aluminium Alloys Based on Spectroscopy Measures, *Proceeding of OAGM/AAPR Workshop*, 2011.

[82] S. Moncayo, M. Kocianova, J. Hulik, J. Plavcan, M. Hornackova, M. Suchonova, P. Veis, J.O. Caceres: Discrimination of Copper Alloys with Archaeological Interest Using LIBS and Chemometric Methods, *WDS'14 Proceedings of Contributed Papers*, 2014, p. 131-135.

[83] I.B. Gornushkin, B.W. Smith, H. Nasajpour, J.D. Winefordner: Identification of Solid Materials by Correlation Analysis Using a Microscopic Laser-Induced Plasma Spectrometer, *Analytical Chemistry*, 1999, vol. 71, p. 5157-5164.

[84] J. Novotný, M. Brada, M. Petrilak, D. Prochazka, K. Novotný, A. Hrdlička, J. Kaiser: A Versatile Interaction Chamber for Laser-Based Spectroscopic Applications, with the Emphasis on Laser-Induced Breakdown Spectroscopy, *Spectrochimica Acta Part B*, 2014, vol. 108, p. 1-7.

[85] R. Valizadeh, O.B. Malyshev, S. Wang, S.A. Zolotovskaya, W.A. Gillespie, A. Abdolvand: Novel Low Secondary Electron Yield Engineered Surface for Mitigation of Electron Cloud, *Applied Physics*, 2014, vol. 105, p. 231605:1-4.

[86] J. Klus, P. Pořízka, D. Prochazka, J. Novotný, K. Novotný, J. Kaiser: Effect of Experimental Parameters and Resulting Analytical Signal Statistics in Laser-Induced Breakdown Spectroscopy, *Spectrochimica Acta Part B*, 2016, vol. 126, p. 6-10.

[87] J. Shlens: A Tutorial on Principal Component Analysis, available online at `http://www.cs.cmu.edu/~tom/10701_sp11/slides/pca_schlens.pdf`, 2017.

[88] R Programming Language Documentation, available online at `https://www.r-project.org/,2017`