Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Materials Science and Technology (37th Cycle)

Facing the challenges of Industry 4.0: advanced techniques of near-infrared spectroscopy and chemometrics to drive innovation in materials science and food analysis

Elena Cazzaniga

Supervisor: Prof. Francesco Savorani

Co-supervisor: Prof. Francesco Geobaldo

To my mum,

whose efforts and sacrifices made it possible

for me to get this far.

Table of Contents

Overview of the contents

The work described in this Thesis concerns the application of Near-Infrared (NIR) spectroscopy coupled with chemometrics in the analysis of food matrices, to face some of the criticalities that characterize the agri-food field nowadays. The main focus regarded the development of rapid and sustainable analytical methods that can save time, avoid the use of solvents and reduce wastes, and also make them easy to use even by untrained personnel. The guidelines of Process Analytical Technology (PAT), circular economy and industry 4.0 were followed in the realization of this purpose.

Traditional analytical methods are often expensive in terms of use of solvents, waste production, and need a significant amount of time to carry out the analyses. Moreover, they often require an initial step of sample preparation, and the sample cannot generally be recovered after the analysis. With this work, it was tried to overcome these limitations by recommending rapid and non-destructive analytical methods and approaches that combine the use of NIR spectroscopy and chemometrics, the latter playing an important role in developing such methods, as multivariate data analysis is fundamental to handle the complex and overwhelming amount of data coming from spectroscopic analyses.

The first project, addressed during this PhD, is named "NEWPOW" and was developed in collaboration with the Department of Agricultural, Forest and Food Science of the University of Turin. This project was aimed at evaluating the use of NIR spectroscopy to perform the quantification of total lipids in hazelnut samples, potentially replacing the conventional analyses. The details about this study are presented in Chapter 3 of this Thesis, and a paper concerning this topic was published [1].

Another project presented in this Thesis is a research line of a larger project, named "Help2Grow", and it involves the use of both portable spectrometer and NIR-HSI in the analysis of grapevine leaves, in collaboration with the Department of Agricultural, Forest and Food Science of the University of Turin, and the University of Modena and Reggio Emilia. The aim was to inspect the potential of NIR spectroscopy in the detection of two diseases affecting grape leaves: powdery and downy mildew. More details of this study are reported in Chapter 4 of this Thesis, while a description of the collected samples can be found in Section 2.3.3.

The last project approached in this Thesis is based on the application of NIR-Hyperspectral Imaging (NIR-HSI) [2,3] and chemometrics in the characterization of 47 rice varieties, with the aim of providing powerful classification models able to identify the different rice types. This project was realized in collaboration with

the University of Barcelona and Ente Nazionale Risi. Chapter 5 of this Thesis describes this project in detail, and a manuscript treating the obtained results is under development.

Thesis structure and organization

This thesis is organised in seven chapters, and it is structured as follows:

- **Chapter 1**: introduction of the theoretical framework and the case studies of this Thesis.
- **Chapter 2**: spectroscopic and chemometric techniques applied, with an overview of the analytical instruments used in this work.
- **Chapter 3**: the "NEWPOW" project (in collaboration with the University of Turin).
- **Chapter 4**: the "Help2Grow" project (in collaboration with the University of Turin and University of Modena and Reggio Emilia).
- **Chapter 5**: the Rice HSI project (in collaboration with the University of Barcelona and Ente Nazionale Risi).
- **Chapter 6**: general conclusions and future perspectives.

References

[1] E. Cazzaniga, N. Cavallini, A. Giraudo, G. Gavoci, F. Geobaldo, M. Pariani, D. Ghirardello, G. Zeppa, F. Savorani, Lipids in a Nutshell: Quick Determination of Lipid Content in Hazelnuts with NIR Spectroscopy, Foods. 12 (2023) 34. https://doi.org/10.3390/foods12010034.

[2] M.J. Khan, H.S. Khan, A. Yousaf, K. Khurshid, A. Abbas, Modern Trends in Hyperspectral Image Analysis: A Review, IEEE Access. 6 (2018) 14118–14129. https://doi.org/10.1109/ACCESS.2018.2812999.

[3] K. Sendin, P.J. Williams, M. Manley, Near infrared hyperspectral imaging in quality and safety evaluation of cereals, Crit. Rev. Food Sci. Nutr. 58 (2018) 575–590. https://doi.org/10.1080/10408398.2016.1205548.

[4] M. Sliwi´nskasliwi´nska-Bartel, D. Thorburn Burns, C. Elliott, Rice fraud a global problem: A review of analytical tools to detect species, country of origin and adulterations, Trends Food Sci. Technol. (2021). https://doi.org/10.1016/j.tifs.2021.06.042.

Chapter 1

Introduction

In recent years, the field of materials science has undergone a deep transformation, linked to the increasing integration of advanced digital technologies, sustainable manufacturing practices, and real-time monitoring systems. This evolution is influenced by three key factors: Process Analytical Technology (PAT, [1,2]), circular economy [3], and Industry 4.0 [4]. These interconnected concepts play a crucial role in the optimisation of industrial production, minimizing wastes, and enhancing process efficiency, contributing to a more sustainable and technologically advanced manufacturing landscape. PAT represents a fundamental shift in the manufacturing methodologies by enabling real-time monitoring and control of production processes. Through the implementation of advanced sensors, data analytics, and machine learning techniques, PAT simplifies the evaluation of materials properties, ensuring a high product quality. This proactive approach enhances the efficiency of the industrial processes and promotes data-driven decision-making, which is essential for optimizing resource utilization and minimizing wastes. Circular economy is a sustainability-driven framework that aims at switching from a traditional linear production model to a more resource-efficient approach, indeed circular. In materials science, this involves designing products and processes that prioritize recyclability, reusability, and waste valorisation. By adopting circular economy principles, the industry can significantly reduce its environmental impact while creating economic value from by-products that would otherwise be discarded. Industry 4.0, the fourth industrial revolution, is characterized by the merging of digital technologies, automation, and artificial intelligence (AI) to create smart production systems. The incorporation of interconnected devices, Internet of Things (IoT) technologies, and cyber-physical systems in materials science allows enhanced process optimization, predictive maintenance, and adaptive manufacturing. With Industry 4.0 technologies, producers can achieve high levels of efficiency, customization, and sustainability in the production processes.

The merging of PAT, circular economy, and Industry 4.0 represents a unique opportunity to revolutionize the field of food science and technology.

This Thesis aims at exploring the integration of these three factors in the context of sustainable food processing and advanced manufacturing. By investigating the role

of rapid, cost-effective and waste-preventive analytical technologies coupled with advanced multivariate data treatment and modelling, this research is intended to develop innovative solutions that align with both environmental and industrial objectives.

The aim of this chapter is to describe the theoretical principles on which this Thesis was based, and to introduce the case studies in which those criteria were applied.

1.1. Chemometrics

Chemometrics is a polyhedral discipline, which merges chemistry, mathematics, statistics, and informatic skills, with deep roots in the field of computer science. In a nutshell, it consists in the application of mathematical and statistical operations on complex chemical data, to make them easier to handle and to understand. The birth of chemometrics can be dated back to the 1960s, although electronic computing machines had already been developed in the 1940s due to the necessity of decrypting the enemy's coded messages during the World War II [5]. During the '60s, the accessibility of computers expanded beyond the exclusive use by engineers and mathematicians, including scientists, and allowing the birth of several new disciplines. The use of computers to help chemists in the synthesis [6], and to better understand the structure [7] of molecules laid the foundation for the development of computational chemistry [8], which is nowadays largely used in molecular modelling [9], especially in the field of biology and pharmaceutical chemistry [10,11].

The first time the term "chemometrics" was used dates back to 1972, in a paper by Svante Wold, who is considered one of the two fathers of this discipline, together with Bruce Kowalski. As Wold himself said, chemometrics can be defined as a way "to get chemically relevant information out of measured data, to represent and display this information, and to get such information into data" [12]. In the field of analytical chemistry, chemometrics is frequently combined with techniques such as UV-Vis, Infrared (IR) and Raman spectroscopies, Nuclear Magnetic Resonance (NMR) and mass spectrometry, and gas or liquid chromatography. Other applications include electrophoresis and potentiometry. What is in common among these analytical techniques is the plentiful amount of data coming from the experimental measurements, which are generally characterized by a huge number of measured variables, for which the application of chemometrics in the data analysis is required.

The multivariate approach was enthusiastically welcomed, as traditional approaches do not work well with many variables, especially if they are correlated or noisy. Moreover, multivariate regression showed to perform better than the bivariate curve, and also multivariate classification was found to work better than the cases with few selected variables [13], encouraging analytical chemists to enthusiastically adopt these new methodologies. Nowadays, chemometrics is widely used in academic research, as well as in industry, and it can be seen as a

right-hand man who helps facing the challenges that characterise the life of a scientist. The main applications of chemometrics involve the treatment of complex and copious information that scientists need to handle [14–16], the choice of the most appropriate sets of conditions of a reaction among multiple possibilities through Design of Experiment (DoE, [17–19]), the inspection of data to catch the presence of trends through exploratory analysis [20,21], and the extraction of quantitative and qualitative information about samples by means of multivariate regression [22–24] and classification [25–27].

Chemometrics is extremely versatile, and this allows its application in any field that requires the treatment of challenging chemical data: pharmaceutical applications involve optimization of formulations, process and dissolution issues understanding, homogenization process monitoring, prediction of drug content in tablets [28–30]. In the agri-food industry, chemometrics is used in the analysis of chemical composition [31,32], quality and authenticity assessment [33–38], anti-food fraud purposes [39–42]. There are also environmental and agricultural applications for pollution analysis and monitoring [43–45] and soil analysis [46–49]. Eventually, another field that sees chemometrics spreading is forensic science [50–52]. This above only represents a non-exhaustive list of current applications in which chemometrics is becoming more and more essential.

1.2. Near-Infrared (NIR) spectroscopy

As the name itself suggests, the Infrared (IR) radiation lies close to the red wavelength of the visible interval and is found at a lower energy level. The discovery of this radiation is dated 1800, thanks to the observation of Sir William Herschel [53] who found that it was possible to measure different temperatures of the scattered solar light, based on the colour of the radiation. In particular, the "hottest" region in this experiment was found beyond the red, where there was apparently no light at all: Herschel named these radiations "calorific rays", which were further named "infra-red rays". Infrared spectroscopy exploits the ability of the molecules to interact with IR radiation by absorbing specific wavelengths. This interaction happens because the chemical bonds present in the molecules resonate with radiations of specific energy, and this makes it possible to study the chemical structure of a material: because of this, IR spectroscopy is a powerful non-destructive technique widely used in analytical chemistry. In particular, two sub-regions in the IR interval are of foremost interest in this field: the near infrared (NIR) and the middle infrared (MIR) regions. The far infrared (FIR) region is mainly used for therapeutic purposes [54]. The NIR range covers the interval between 14000 and 4000 cm^{-1}, while MIR and FIR ranges are 4000–400 and 400–50 cm^{-1} respectively [55].

The main difference in the use of NIR and MIR spectroscopies is linked to the vibrational modes that can be detected in the two wavelength ranges. A molecule can be represented as an anharmonic oscillator, in which the atoms move

constrained by the chemical bonds. These movements are called “vibrations”, each one represented by different levels of potential energy: the type of vibration that can occur depends on the energy to which the molecule is exposed. To visually explain this theory, **Figure 1.1** displays the vibrational levels of an anharmonic oscillator composed by two atoms, moving in between their chemical bond.

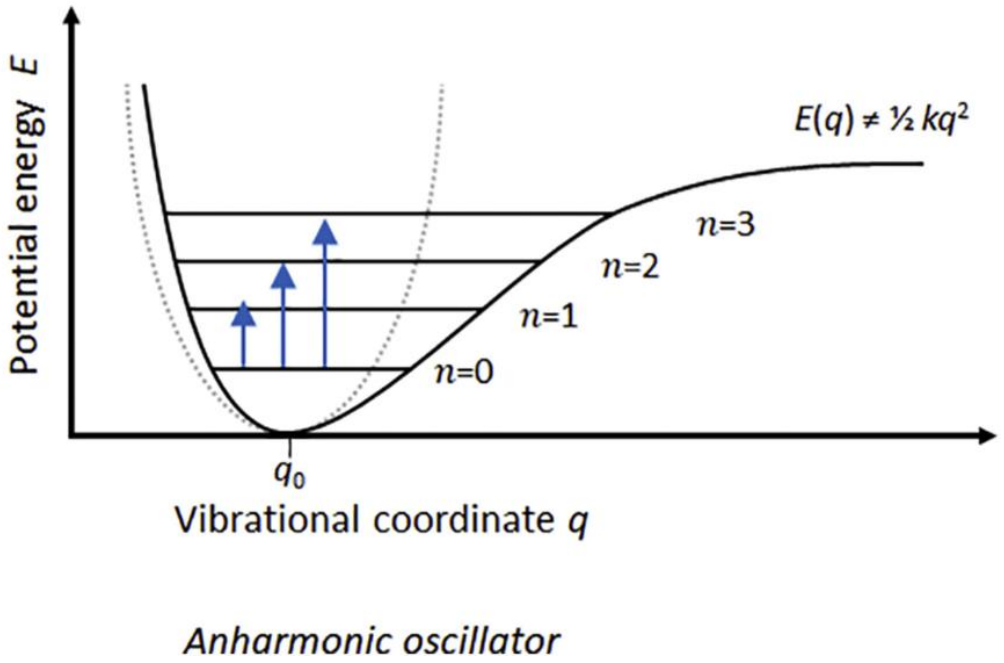

Figure 1.1. illustration of an anharmonic oscillator composed by two atoms and its vibrational levels. The vibrational transitions are represented as blue arrows. (from Comprehensive Analytical Chemistry, 2022 [56])

The MIR radiation stimulates the fundamental vibrations of molecular bonds, i.e., the transitions from the ground state to the first vibrational level (n = 0 → n = 1).

NIR radiation carries more energy than MIR, allowing transitions from the ground state to integer multiples of fundamental vibrations (n = 0 → n (n > 1)); the resulting resonant frequency is called “overtone”. Another signal detected with NIR spectroscopy occur when two or more fundamental vibrations are excited simultaneously, resulting in the so-called “combination band”. The signals detectable in the NIR range refer mostly to the oscillations of hydrogen-containing groups, such as C–H, O–H, and N–H, which are typical of most organic compounds. Because these vibrations in the NIR range are less probable than the fundamental ones, their intensity are consequently weaker if compared to the signals in the MIR range.

1.2.1. NIR instruments and applications

A common NIR spectrometer is composed by a light source (generally Mercury-Argon or Mercury-Neon lamps, but also a diode laser coupled with an optic fibre can be used), a monochromator to split the light beam into selected wavelengths, a

detector (usually Indium-Gallium-Arsenide detectors) and optical components such as lenses, integrating spheres etc. Benchtop spectrometers are routinely used for laboratory analyses and ensure high accuracy and reliability, but in the last decades a strong trend in the development and use of field spectrometers can be seen. Field spectrometers can be classified into transportable (i.e., carried to the field on a vehicle), in suitcase format (>4 kg) and handheld (<1 kg) [57], with recent progresses in miniaturization that led to very small devices weighing less than 50 g, equipped with a built-in battery and operated by smartphone applications, connected via Bluetooth. This significant change in the dimensions of the devices leads to a consequent change in the working characteristics, such as the spectral range and resolution, and the sensitivity of the instrument, which become generally lower. Comparing the performances of benchtop and portable devices, it has been shown that benchtop spectrometers usually provide better outcomes for prediction or classification models parameters [58–60]. This certainly depends on the fact that benchtop instruments consist in structured devices which benefit of years of technological development and improvements and are used by expert personnel in a controlled environment. On the other hand, the field spectrometers overcome the lower sensitivity and resolution with other advantages, such as the portability and possibility to perform analyses directly where necessary, avoiding transporting the sample to a laboratory and consequently saving time and preserving the quality and original characteristics of the sample itself during the analysis. Other benefits involve their accessibility, as costs are significantly lower, and their ease of use, extending their application also in contexts where operators are not used to this type of analysis [61]. In any case, despite some technological limitations of portable spectrometers in comparison to the traditional benchtop instruments, several recent studies proved the effectiveness of these new analytical devices [62–65], giving an optimistic outlook in the affirmation of portable instruments for the analyses with NIR spectroscopy.

NIR spectroscopy represents a widespread analytical technique in the world of research, as well as in many industrial fields. Its principal applications include medical studies [66,67], and pharmaceutical [68] and food [69,70] analyses. Concerning the latter, which is of particular interest in the work presented in this Thesis, more detailed information, concerning the application of NIR spectroscopy to the different investigations characterising the present work, will be provided when introducing the specific studies.

1.3. Hyperspectral Imaging (HSI)

The term “Hyperspectral Imaging” (HSI) was first used in the 1980s, in a study on remote sensing [71], the field in which HSI is rooted. The word “hyperspectral” underlines the multidimensionality of the produced data: the first approach developed from remote sensing, called Multispectral Imaging (MSI, [72]), consisted in the detection of just few single spectral bands, while with HSI each single pixel in the image contains a complete spectrum within a specified range.

The information obtained with HSI is represented by a 3D image, also called "hypercube", consisting in two dimensions which describe the spatial location of each pixel in the image (i.e., x and y coordinates), and one spectroscopic dimension (λ) carrying the information of the chosen wavelength range of analysis [73]. Thus, it is possible not only to acquire the spectra of a specimen, but also to get additional details, i.e., the spatial information of the sample.

1.3.1. HSI instrument and applications

Starting from the 1990s, HSI became very popular in the agri-food field, where it is often coupled with NIR spectroscopy [74]. Lots of studies have been conducted with HSI-NIR to inspect food qualitative properties [75–77], build classification models [78,79], ensure food quality [80–82], and prevent food fraud through products authentication [42]. Thanks to its versatility, HSI can be also coupled with many other spectroscopic techniques, for example Raman [83–85], Vis-NIR [86–88], and fluorescence [89,90] spectroscopies. In these cases, the fields of application are diverse, and mainly include food quality and safety assessment [91–94], medical diagnosis [95,96], artwork authentication [97] and forensic analyses [98–100]. Because of the overwhelming amount of data that this technique provides, chemometrics is always required when dealing with this type of analysis [38,39]. In this way, it becomes easier to extract and exploit information also from such complex and abundant datasets.

The experimental setup of HSI makes this technique advantageous if compared to the classic ones, such as liquid or gas chromatography, or mass spectrometry. It consists in a light source, a wavelength dispersion device, and an area detector. There are four ways to acquire the 3D hyperspectral image, namely point scanning, line scanning, area scanning, and single shot method [101]. In the first case, a single point is scanned along the two spatial dimensions, by moving either the sample or the detector. With line scanning, a row of pixels is acquired, obtaining the total image by moving the sample support along the vertical spatial dimension. Area scanning is not a spatial-scanning method as the previous one: it works by acquiring a single-band 2D grayscale image of the sample, obtaining in this way the full spatial information in just one acquisition. Then, a selected number of wavelengths is used to perform a scanning in the spectral domain. In this way, the image is recorded quickly, as no time for spatial scanning is required. Finally, the single shot method involves the application of a detector with a large area that allows to capture the whole image of the sample, complete with all wavelengths, thus acquiring all the data at one time. **Figure 1.2** displays how the four acquisition methods work.

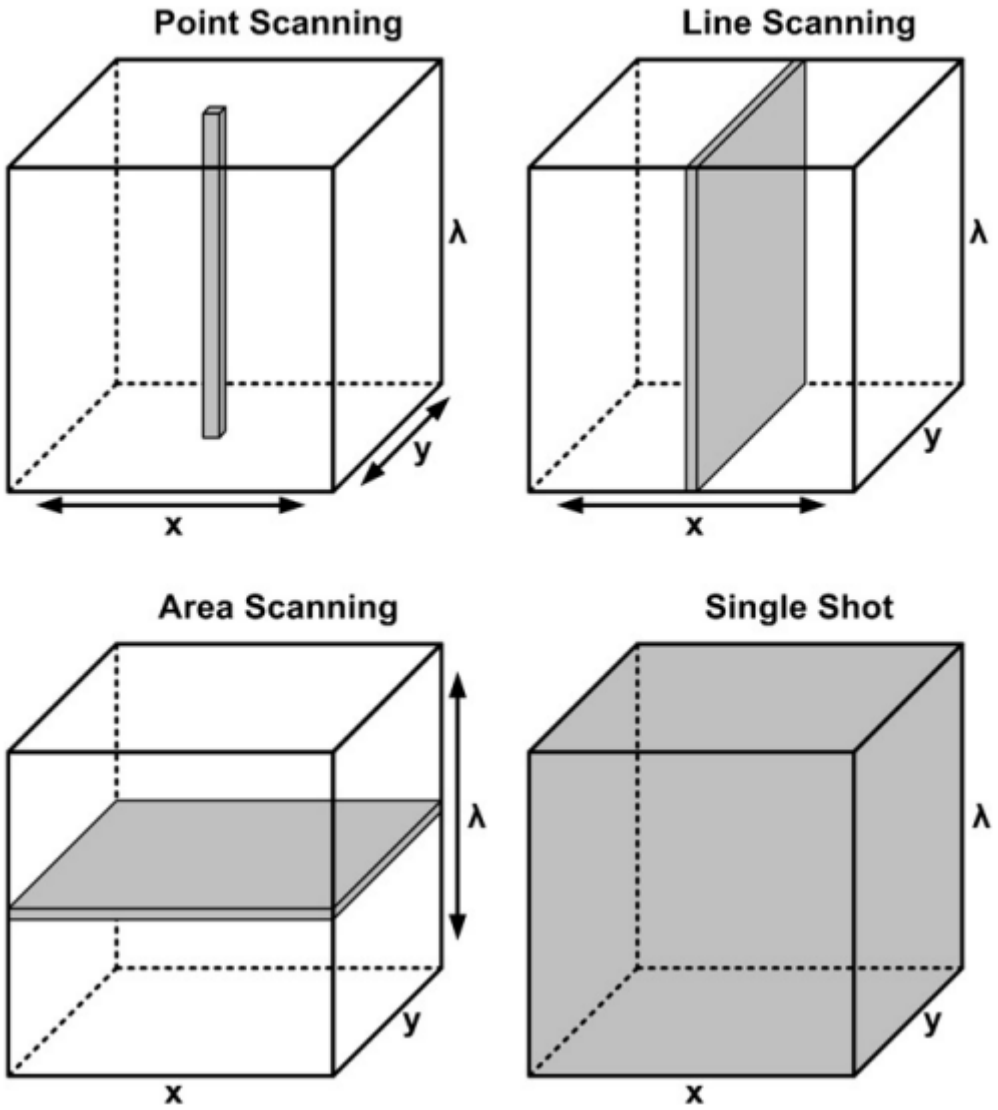

Figure 1.2. the four methods used for the acquisition of the hyperspectral image. The arrows represent the scanning direction, while the grey areas in the cubes indicate the data registered at each single acquisition (adapted from Hyperspectral Imaging for Food Quality Analysis and Control, 2010 [101]).

Using this analytical technique, samples can be analysed without any pre-treatment, avoiding the use of solvents, and saving time. It is also possible to prevent wastes and to re-use the samples, as after the analysis they can be recovered without any alteration. HSI also allows to analyse many samples at the same time within the same image, giving the opportunity to obtain many samples fingerprint, as well as to perform surface analysis, making it possible to obtain a chemical map of the analysed samples.

1.4. Case studies

The research studies performed during the PhD period refers to some specific Projects that were conducted at the Department of Applied Science and Technology of Politecnico di Torino and which are presented in brief in the following Sections.

1.4.1. NEWPOW

This project was aimed at evaluating the use of NIR spectroscopy in the analysis of the lipid content of hazelnuts from different origins, as an alternative to the traditional analytical approaches. Traditionally, the analysis of lipids involves three steps: solvent extraction from the samples, separation through chromatographic or electrophoretic methods, and final identification and quantification [102]. This methodology is generally expensive, time consuming, and it often involves hazardous and non-ecofriendly solvents. Using NIR spectroscopy, it was possible not only to avoid the use of solvents, but also to directly analyse the samples without any pre-treatment, which in turn allowed to recover them intact after the analysis. The NIR data were firstly inspected through chemometrics tools to catch the most relevant information and to detect the presence of trends among samples, and then regression models were developed to be further used in analyses of new samples with the aim of predicting the lipid content.

More information about the samples and the lipid extraction performed by the colleagues of the University of Turin can be found in Section 2.3.1. The details of this project and the results are presented in Chapter 3.

1.4.2. Help2Grow

The first aim of this project deals with the necessity of rapid and cost-effective analytical methods that allow in-field analyses to detect and monitor the growth of the grapevine, allowing to prevent the spread of diseases caused by microorganisms. In particular, the development of downy and powdery mildew was studied. The presence of these diseases can be normally noticed to the naked eye, when the plant illness reaches a rather advanced state: downy mildew develops a white dots pattern on the back side of the grapevine leaves, while powdery mildew consists in a whitish and dusty patina on the surface of the frontal side of the leaf. A first set of measurements were conducted *in-situ* at the Department of Agricultural, Forest and Food Science of the University of Turin during the summer and autumn 2022, with a portable NIR spectrometer. During the last acquisition session, made on October 26th, 2022, some leaves were collected and frozen, to be further analysed on May 6th, 2024, with HSI-NIR at the University of Modena and Reggio Emilia. A detailed discussion of this study can be found in Chapter 4 of this Thesis.

1.4.2. The Rice-HSI project

The aim of this project was to face one of the main issues of the food industry, in particular the world of rice production: food fraud [103]. In the rice industry, it is recurring that low-quality varieties are sold as the most expensive ones because of their structural similarity, but their nutritional values and organoleptic features can be different. Consequently, it is important to find a way to detect the differences among the rice types and to exploit this information to avoid scams. This project is described in detail in Chapter 5 of this Thesis.

References | Chapter 1

[1] L.L. Simon, H. Pataki, G. Marosi, F. Meemken, K. Hungerbühler, A. Baiker, S. Tummala, B. Glennon, M. Kuentz, G. Steele, H.J.M. Kramer, J.W. Rydzak, Z. Chen, J. Morris, F. Kjell, R. Singh, R. Gani, K. V. Gernaey, M. Louhi-Kultanen, J. Oreilly, N. Sandler, O. Antikainen, J. Yliruusi, P. Frohberg, J. Ulrich, R.D. Braatz, T. Leyssens, M. Von Stosch, R. Oliveira, R.B.H. Tan, H. Wu, M. Khan, D. Ogrady, A. Pandey, R. Westra, E. Delle-Case, D. Pape, D. Angelosante, Y. Maret, O. Steiger, M. Lenner, K. Abbou-Oucherif, Z.K. Nagy, J.D. Litster, V.K. Kamaraju, M. Sen Chiu, Assessment of recent process analytical technology (PAT) trends: A multiauthor review, Org. Process Res. Dev. 19 (2015) 3–62. https://doi.org/10.1021/OP500261Y.

[2] A.L. Pomerantsev, O.Y. Rodionova, Process analytical technology: a critical view of the chemometricians, J. Chemom. 26 (2012) 299–310. https://doi.org/10.1002/CEM.2445.

[3] S. Geisendorf, F. Pietrulla, The circular economy and circular economic concepts—a literature analysis and redefinition, Thunderbird Int. Bus. Rev. 60 (2018) 771–782. https://doi.org/10.1002/TIE.21924.

[4] M. Ghobakhloo, Industry 4.0, digitization, and opportunities for sustainability, J. Clean. Prod. 252 (2020) 119869. https://doi.org/10.1016/J.JCLEPRO.2019.119869.

[5] P.M. Sosa, W. Academia, V. Caracas, THE INFLUENCE OF WWII IN THE CREATION OF COMPUTERS, (2010).

[6] E.J. Corey, W. Todd Wipke, Computer-assisted design of complex organic syntheses, Science (80-.). 166 (1969) 178–192. https://doi.org/10.1126/science.166.3902.178.

[7] J. Lederberg, G.L. Sutherland, B.G. Buchanan, E.A. Feigenbaum, A. V. Robertson, A.M. Duffield, C. Djerassi, Applications of Artificial Intelligence for Chemical Inference. I. The Number of Possible Organic Compounds. Acyclic Structures Containing C, H, O, and N, J. Am. Chem. Soc. 91 (1969) 2973–2976. https://doi.org/10.1021/JA01039A025.

[8] C.J. Cramer, Essentials of Computational Chemistry - Theories and Models, John Wiley Sons. (2004).

[9] G.D.K.N. K. I. Ramachandran, Computational Chemistry and Molecular Modeling: Principles and Applications, (2008). https://doi.org/10.1007/978-3-540-77304-7.

[10] J.P. Bowen, P.S. Charifson, P.C. Fox, M. Kontoyianni, A.B. Miller, D. Schnur, E.L. Stewart, C. Van Dyke, Computer-Assisted Molecular Modeling: Indispensable Tools for Molecular Pharmacology, J. Clin. Pharmacol. 33 (1993) 1149–1164. https://doi.org/10.1002/J.1552-4604.1993.TB03915.X.

[11] E.F. Meyer, S.M. Swanson, J.A. Williams, Molecular modelling and drug design, Pharmacol. Ther. 85 (2000) 113–121. https://doi.org/10.1016/S0163-7258(99)00069-8.

[12] S. Wold, Chemometrics and Intelligent Laboratory Systems, 30 (1995) 109–115. https://doi.org/doi.org/10.1016/0169-7439(95)00042-9.

[13] S. Wold, M. Sjöström, Chemometrics, present and future success, Chemom. Intell. Lab. Syst. 44 (1998) 3–14. https://doi.org/10.1016/S0169-7439(98)00075-6.

[14] H. Martens, Quantitative Big Data: where chemometrics can contribute, (2015).

https://doi.org/10.1002/cem.2740.

[15] H. Parastar, R. Tauler, Big (Bio)Chemical Data Mining Using Chemometric Methods: A Need for Chemists, Angew. Chemie Int. Ed. 61 (2022) e201801134. https://doi.org/10.1002/ANIE.201801134.

[16] E. Szymańska, Modern data science for analytical chemical data – A comprehensive review, Anal. Chim. Acta. 1028 (2018) 1–10. https://doi.org/10.1016/J.ACA.2018.05.038.

[17] S.N. Deming; S.L. Morgan, Experimental Design: A Chemometric Approach, 2nd ed., 1993.

[18] F.A.N. Fernandes, Experimental design for chemometrics: best practices, Chemometrics. (2024) 39–59. https://doi.org/10.1016/B978-0-443-21493-6.00003-4.

[19] Q.S. Xu, Y. Da Xu, L. Li, K.T. Fang, Uniform experimental design in chemometrics, J. Chemom. 32 (2018) e3020. https://doi.org/10.1002/CEM.3020.

[20] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. 2 (2010) 433–459. https://doi.org/10.1002/wics.101.

[21] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. https://doi.org/10.1039/c3ay41907j.

[22] L.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools, Technometrics. 35 (1993) 109–135. https://doi.org/10.1080/00401706.1993.10485033.

[23] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

[24] L.E. Eberly, Multiple Linear Regression, Methods Mol. Biol. 404 (2007) 165–187. https://doi.org/10.1007/978-1-59745-530-5_9.

[25] F. Marini, Classification Methods in Chemometrics, Curr. Anal. Chem. 6 (2009) 72–79. https://doi.org/10.2174/157341110790069592.

[26] M. Cocchi, A. Biancolillo, F. Marini, Chemometric Methods for Classification and Feature Selection, Compr. Anal. Chem. 82 (2018) 265–299. https://doi.org/10.1016/BS.COAC.2018.08.006.

[27] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods. 5 (2013) 3790–3798. https://doi.org/10.1039/C3AY40582F.

[28] Y. Roggo, K. Degardin, P. Margot, Identification of pharmaceutical tablets by Raman spectroscopy and chemometrics, Talanta. 81 (2010) 988–995. https://doi.org/10.1016/J.TALANTA.2010.01.046.

[29] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies, J. Pharm. Biomed. Anal. 44 (2007) 683–700. https://doi.org/10.1016/J.JPBA.2007.03.023.

[30] I. Singh, P. Juneja, B. Kaur, P. Kumar, M. Haji Shabani, E. Klodzinska, A. Tsantili-Kakoulidou, Pharmaceutical Applications of Chemometric Techniques, Int. Sch. Res. Not. 2013 (2013) 795178. https://doi.org/10.1155/2013/795178.

[31] J.M. Amigo, I. Martí, A. Gowen, Hyperspectral Imaging and Chemometrics: A Perfect Combination for the Analysis of Food Structure, Composition and Quality, Data Handl. Sci. Technol. 28 (2013) 343–370. https://doi.org/10.1016/B978-0-444-59528-7.00009-0.

[32] A.A.F. Zielinski, C.W.I. Haminiuk, C.A. Nunes, E. Schnitzler, S.M. van Ruth, D. Granato, Chemical Composition, Sensory Properties, Provenance, and Bioactivity of Fruit Juices as Assessed by Chemometrics: A Critical Review and Guideline, Compr. Rev. Food Sci. Food Saf. 13 (2014) 300–316. https://doi.org/10.1111/1541-4337.12060.

[33] C.M. Andre, C. Soukoulis, Food Quality Assessed by Chemometrics, Foods 2020, Vol. 9, Page 897. 9 (2020) 897. https://doi.org/10.3390/FOODS9070897.

[34] A. González-Casado, A. María, J. Carvelo, R. González-Domínguez, A. Sayago, Á. Fernández-Recamales, An Overview on the Application of Chemometrics Tools in Food Authenticity and Traceability, Foods 2022, Vol. 11, Page 3940. 11 (2022) 3940. https://doi.org/10.3390/FOODS11233940.

[35] M. Efenberger-Szmechtyk, A. Nowak, D. Kregiel, Implementation of chemometrics in quality evaluation of food and beverages, Crit. Rev. Food Sci. Nutr. 58 (2018) 1747–1766. https://doi.org/10.1080/10408398.2016.1276883.

[36] I.S. Arvanitoyannis, M.N. Katsota, E.P. Psarra, E.H. Soufleros, S. Kallithraka, Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics), Trends Food Sci. Technol. 10 (1999) 321–336. https://doi.org/10.1016/S0924-2244(99)00053-9.

[37] M. Mohd Ali, N. Hashim, S.A. Aziz, O. Lasekan, Emerging non-destructive thermal imaging technique coupled with chemometrics on quality and safety inspection in food and agriculture, Trends Food Sci. Technol. 105 (2020) 176–185. https://doi.org/10.1016/J.TIFS.2020.09.003.

[38] C. Kumaravelu, A. Gopal, A review on the applications of Near-Infrared spectrometer and Chemometrics for the agro-food processing industries, Proc. - 2015 IEEE Int. Conf. Technol. Innov. ICT Agric. Rural Dev. TIAR 2015. (2015) 8–12. https://doi.org/10.1109/TIAR.2015.7358523.

[39] H. Nobari Moghaddam, Z. Tamiji, M. Akbari Lakeh, M.R. Khoshayand, M. Haji Mahmoodi, Multivariate analysis of food fraud: A review of NIR based instruments in tandem with chemometrics, J. Food Compos. Anal. 107 (2022) 104343. https://doi.org/10.1016/J.JFCA.2021.104343.

[40] F. Pennisi, A. Giraudo, N. Cavallini, G. Esposito, G. Merlo, F. Geobaldo, P.L. Acutis, M. Pezzolato, F. Savorani, E. Bozzetta, Differentiation between Fresh and Thawed Cephalopods Using NIR Spectroscopy and Multivariate Data Analysis, Foods 2021, Vol. 10, Page 528. 10 (2021) 528. https://doi.org/10.3390/FOODS10030528.

[41] J. Van De Steene, J. Ruyssinck, J.-A. Fernandez-Pierna, L. Vandermeersch, A. Maes, H. Van Langenhove, C. Walgraeve, K. Demeestere, B. De Meulenaer, L. Jacxsens, B. Miserez, Fingerprinting methods for origin and variety assessment of rice: development, validation and data fusion experiments, Food Control. 151 (2023) 109780. https://doi.org/10.1016/j.foodcont.2023.109780.

[42] J. Mendez, L. Mendoza, J.P. Cruz-Tirado, R. Quevedo, R. Siche, Trends in application of NIR and hyperspectral imaging for food authentication, Sci. Agropecu. 10 (2019) 143–161. https://doi.org/10.17268/SCI.AGROPECU.2019.01.16.

[43] M.M.R. Mostert, G.A. Ayoko, S. Kokot, Application of chemometrics to analysis of soil pollutants, TrAC Trends Anal. Chem. 29 (2010) 430–445. https://doi.org/10.1016/J.TRAC.2010.02.009.

[44] A. Inobeme, V. Nayak, T.J. Mathew, S. Okonkwo, L. Ekwoba, A.I. Ajai, E. Bernard, J. Inobeme, M. Mariam Agbugui, K.R. Singh, Chemometric approach in environmental pollution analysis: A critical review, J. Environ. Manage. 309

(2022) 114653. https://doi.org/10.1016/J.JENVMAN.2022.114653.

[45] S. Mas, A. de Juan, R. Tauler, A.C. Olivieri, G.M. Escandar, Application of chemometric methods to environmental analysis of organic pollutants: A review, Talanta. 80 (2010) 1052–1067. https://doi.org/10.1016/J.TALANTA.2009.09.044.

[46] J.A.M. Demattê, L. Ramirez-Lopez, K.P.P. Marques, A.A. Rodella, Chemometric soil analysis on the determination of specific bands for the detection of magnesium and potassium by spectroscopy, Geoderma. 288 (2017) 8–22. https://doi.org/10.1016/J.GEODERMA.2016.11.013.

[47] D. Cozzolino, A. Morón, Potential of near-infrared reflectance spectroscopy and chemometrics to predict soil organic carbon fractions, Soil Tillage Res. 85 (2006) 78–85. https://doi.org/10.1016/J.STILL.2004.12.006.

[48] J.B. Carra, M. Fabris, J. Dieckow, O.R. Brito, P.R.S. Vendrame, L. Macedo Dos Santos Tonial, Near-Infrared Spectroscopy Coupled with Chemometrics Tools: A Rapid and Non-Destructive Alternative on Soil Evaluation, Commun. Soil Sci. Plant Anal. 50 (2019) 421–434. https://doi.org/10.1080/00103624.2019.1566465.

[49] I. Barra, S.M. Haefele, R. Sakrabani, F. Kebede, Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–A review, TrAC Trends Anal. Chem. 135 (2021) 116166. https://doi.org/10.1016/J.TRAC.2020.116166.

[50] C. Malegori, E. Alladio, P. Oliveri, C. Manis, M. Vincenti, P. Garofano, F. Barni, A. Berti, Identification of invisible biological traces in forensic evidences by hyperspectral NIR imaging combined with chemometrics, Talanta. 215 (2020) 120911. https://doi.org/10.1016/j.talanta.2020.120911.

[51] C.S. Silva, A. Braz, M.F. Pimentel, Vibrational Spectroscopy and Chemometrics in Forensic Chemistry: Critical Review, Current Trends and Challenges, J. Braz. Chem. Soc. 30 (2019) 2259–2290. https://doi.org/10.21577/0103-5053.20190140.

[52] V. Sharma, R. Kumar, Trends of chemometrics in bloodstain investigations, TrAC Trends Anal. Chem. 107 (2018) 181–195. https://doi.org/10.1016/J.TRAC.2018.08.006.

[53] E.F.J. Ring, The discovery of infrared radiation in 1800, Imaging Sci. J. 48 (2000) 1–8. https://doi.org/10.1080/13682199.2000.11784339.

[54] F. Vatansever, M.R. Hamblin, Far infrared radiation (FIR): Its biological effects and medical applications, Photonics Lasers Med. 1 (2012) 255–266. https://doi.org/10.1515/PLM-2012-0034.

[55] Y. Ozaki, Near-infrared spectroscopy-its versatility in analytical chemistry, Anal. Sci. 28 (2012) 545–563. https://doi.org/10.2116/ANALSCI.28.545.

[56] K.B. Beć, J. Grabska, C.W. Huck, Physical principles of infrared spectroscopy, Compr. Anal. Chem. 98 (2022) 1–43. https://doi.org/10.1016/BS.COAC.2020.08.001.

[57] K.B. Beć, J. Grabska, H.W. Siesler, C.W. Huck, Handheld near-infrared spectrometers: Where are we heading?, NIR News, Vol. 31, Issue 3-4. 31 (2020) 28–35. https://doi.org/10.1177/0960336020916815.

[58] H. Yu, H. Liu, Q. Wang, S. van Ruth, Evaluation of portable and benchtop NIR for classification of high oleic acid peanuts and fatty acid quantitation, LWT. 128 (2020) 109398. https://doi.org/10.1016/J.LWT.2020.109398.

[59] S. Mayr, K.B. Beć, J. Grabska, E. Schneckenreiter, C.W. Huck, Near-infrared spectroscopy in quality control of Piper nigrum: A comparison of performance of benchtop and handheld spectrometers, Talanta. 223 (2021) 121809. https://doi.org/10.1016/J.TALANTA.2020.121809.

[60] J.J. Acosta, M.S. Castillo, G.R. Hodge, Comparison of benchtop and handheld near-infrared spectroscopy devices to determine forage nutritive value, Crop Sci. 60 (2020) 3410–3422. https://doi.org/10.1002/CSC2.20264.

[61] N. Cavallini, E. Cavallini, F. Savorani, Monitoring the homemade fermentation of readymade malt extract using the SCiO NIR sensor: A convergence of technology and tradition, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 325 (2025) 125126. https://doi.org/10.1016/J.SAA.2024.125126.

[62] V. Wiedemair, C.W. Huck, Evaluation of the performance of three hand-held near-infrared spectrometer through investigation of total antioxidant capacity in gluten-free grains, Talanta. 189 (2018) 233–240. https://doi.org/10.1016/J.TALANTA.2018.06.056.

[63] V. Wiedemair, D. Langore, R. Garsleitner, K. Dillinger, C. Huck, Investigations into the Performance of a Novel Pocket-Sized Near-Infrared Spectrometer for Cheese Analysis, Molecules. 24 (2019). https://doi.org/10.3390/molecules24030428.

[64] H. Yan, H.W. Siesler, Identification Performance of Different Types of Handheld Near-Infrared (NIR) Spectrometers for the Recycling of Polymer Commodities, Appl. Spectrosc. Vol. 72, Issue 9, 1362-1370. 72 (2018) 1362–1370. https://doi.org/10.1177/0003702818777260.

[65] H. Yan, H.W. Siesler, Quantitative analysis of a pharmaceutical formulation: Performance comparison of different handheld near-infrared spectrometers, J. Pharm. Biomed. Anal. 160 (2018) 179–186. https://doi.org/10.1016/J.JPBA.2018.07.048.

[66] V.R. Kondepati, H.M. Heise, J. Backhaus, Recent applications of near-infrared spectroscopy in cancer diagnosis and therapy, Anal. Bioanal. Chem. 390 (2008) 125–139. https://doi.org/10.1007/S00216-007-1651-Y.

[67] H.M. Heise, Medical Applications of NIR Spectroscopy, Near-Infrared Spectrosc. Theory, Spectr. Anal. Instrumentation, Appl. (2021) 437–473. https://doi.org/10.1007/978-981-15-8648-4_20.

[68] J. Luypaert, D.L. Massart, Y. Vander Heyden, Near-infrared spectroscopy applications in pharmaceutical analysis, Talanta. 72 (2007) 865–883. https://doi.org/10.1016/J.TALANTA.2006.12.023.

[69] L.M. Reid, C.P. O'Donnell, G. Downey, Recent technological advances for the determination of food authenticity, Trends Food Sci. Technol. 17 (2006) 344–353. https://doi.org/10.1016/j.tifs.2006.01.006.

[70] H. Huang, H. Yu, H. Xu, Y. Ying, Near infrared spectroscopy for on/in-line monitoring of quality in foods and beverages: A review, J. Food Eng. 87 (2008) 303–313. https://doi.org/10.1016/j.jfoodeng.2007.12.022.

[71] A.F.H. Goetz, G. Vane, J.E. Solomon, B.N. Rock, Imaging Spectrometry for Earth Remote Sensing, Science (80-.). 228 (1985) 1147–1153. https://doi.org/10.1126/SCIENCE.228.4704.1147.

[72] J. Qin, K. Chao, M.S. Kim, R. Lu, T.F. Burks, Hyperspectral and multispectral imaging for evaluating food safety and quality, J. Food Eng. 118 (2013) 157–171. https://doi.org/10.1016/J.JFOODENG.2013.04.001.

[73] S. Selci, The Future of Hyperspectral Imaging, J. Imaging 2019, Vol. 5, Page 84. 5 (2019) 84. https://doi.org/10.3390/JIMAGING5110084.

[74] M. Manley, Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials, Chem. Soc. Rev. 43 (2014) 8200–8214. https://doi.org/10.1039/C4CS00062E.

[75] J.A. Fernández Pierna, P. Vermeulen, O. Amand, A. Tossens, P. Dardenne, V. Baeten, NIR hyperspectral imaging spectroscopy and chemometrics for the detection of undesirable substances in food and feed, Chemom. Intell. Lab. Syst. 117 (2012) 233–239. https://doi.org/10.1016/J.CHEMOLAB.2012.02.004.

[76] C. Mo, G. Kim, M.S. Kim, J. Lim, S.H. Lee, H.S. Lee, B.K. Cho, Discrimination methods for biological contaminants in fresh-cut lettuce based on VNIR and NIR hyperspectral imaging, Infrared Phys. Technol. 85 (2017) 1–12. https://doi.org/10.1016/J.INFRARED.2017.05.003.

[77] U. Siripatrawan, Y. Makino, Monitoring fungal growth on brown rice grains using rapid and non-destructive hyperspectral imaging, Int. J. Food Microbiol. 199 (2015) 93–100. https://doi.org/10.1016/J.IJFOODMICRO.2015.01.001.

[78] T. Lei, X.H. Lin, D.W. Sun, Rapid classification of commercial Cheddar cheeses from different brands using PLSDA, LDA and SPA–LDA models built by hyperspectral data, J. Food Meas. Charact. 13 (2019) 3119–3129. https://doi.org/10.1007/s11694-019-00234-0.

[79] P.J. Williams, S. Kucheryavskiy, Classification of maize kernels using NIR hyperspectral imaging, Food Chem. 209 (2016) 131–138. https://doi.org/10.1016/J.FOODCHEM.2016.04.044.

[80] B.W. Mulvey, Determination of Fat Content in Foods Using a Near-Infrared Spectroscopy Sensor, Proc. IEEE Sensors. 2020-October (2020). https://doi.org/10.1109/SENSORS47125.2020.9278647.

[81] R. Calvini, S. Michelini, V. Pizzamiglio, G. Foca, A. Ulrici, Exploring the potential of NIR hyperspectral imaging for automated quantification of rind amount in grated Parmigiano Reggiano cheese, Food Control. 112 (2020) 107111. https://doi.org/10.1016/J.FOODCONT.2020.107111.

[82] J.Y. Barnaby, T.D. Huggins, H. Lee, A.M. McClung, S.R.M. Pinson, M. Oh, G.R. Bauchan, L. Tarpley, K. Lee, M.S. Kim, J.D. Edwards, Vis/NIR hyperspectral imaging distinguishes sub-population, production environment, and physicochemical grain properties in rice, Sci. Reports 2020 101. 10 (2020) 1–13. https://doi.org/10.1038/s41598-020-65999-7.

[83] J. Qin, K. Chao, M.S. Kim, Raman Chemical Imaging System for Food Safety and Quality Inspection, Trans. ASABE. 53 (2010) 1873–1882. https://doi.org/10.13031/2013.35796.

[84] J. Qin, K. Chao, M.S. Kim, Simultaneous detection of multiple adulterants in dry milk using macro-scale Raman chemical imaging, Food Chem. 138 (2013) 998–1007. https://doi.org/10.1016/J.FOODCHEM.2012.10.115.

[85] H. Lee, M.S. Kim, J. Qin, E. Park, Y.R. Song, C.S. Oh, B.K. Cho, Raman Hyperspectral Imaging for Detection of Watermelon Seeds Infected with Acidovorax citrulli, Sensors 2017, Vol. 17, Page 2188. 17 (2017) 2188. https://doi.org/10.3390/S17102188.

[86] J. Li, W. Huang, X. Tian, C. Wang, S. Fan, C. Zhao, Fast detection and visualization of early decay in citrus using Vis-NIR hyperspectral imaging, Comput. Electron. Agric. 127 (2016) 582–592. https://doi.org/10.1016/J.COMPAG.2016.07.016.

[87] Y. Shao, Y. Shi, Y. Qin, G. Xuan, J. Li, Q. Li, F. Yang, Z. Hu, A new quantitative index for the assessment of tomato quality using Vis-NIR hyperspectral imaging, Food Chem. 386 (2022) 132864. https://doi.org/10.1016/J.FOODCHEM.2022.132864.

[88] J. Li, L. Chen, W. Huang, Q. Wang, B. Zhang, X. Tian, S. Fan, B. Li, Multispectral detection of skin defects of bi-colored peaches based on vis–NIR

hyperspectral imaging, Postharvest Biol. Technol. 112 (2016) 121–133. https://doi.org/10.1016/J.POSTHARVBIO.2015.10.007.

[89] H. Yao, Z. Hruska, R. Kincaid, R.L. Brown, D. Bhatnagar, T.E. Cleveland, Detecting maize inoculated with toxigenic and atoxigenic fungal strains with fluorescence hyperspectral imagery, Biosyst. Eng. 115 (2013) 125–135. https://doi.org/10.1016/J.BIOSYSTEMSENG.2013.03.006.

[90] G. Zavattini, S. Vecchi, R.M. Leahy, D.J. Smith, S.R. Cherry, A hyperspectral fluorescence imaging system for biological applications, IEEE Nucl. Sci. Symp. Conf. Rec. 2 (2003) 942–946. https://doi.org/10.1109/NSSMIC.2003.1351850.

[91] N. Prieto, R. Roehe, P. Lavín, G. Batten, S. Andrés, Application of near infrared reflectance spectroscopy to predict meat and meat products quality: A review, Meat Sci. 83 (2009) 175–186. https://doi.org/10.1016/J.MEATSCI.2009.04.016.

[92] K. Sendin, P.J. Williams, M. Manley, Near infrared hyperspectral imaging in quality and safety evaluation of cereals, Crit. Rev. Food Sci. Nutr. 58 (2018) 575–590. https://doi.org/10.1080/10408398.2016.1205548.

[93] J.H. Cheng, D.W. Sun, Hyperspectral imaging as an effective tool for quality analysis and control of fish and other seafoods: Current research and potential applications, Trends Food Sci. Technol. 37 (2014) 78–91. https://doi.org/10.1016/J.TIFS.2014.03.006.

[94] G. Elmasry, D.F. Barbin, D.W. Sun, P. Allen, Meat Quality Evaluation by Hyperspectral Imaging Technique: An Overview, Crit. Rev. Food Sci. Nutr. 52 (2012) 689–711. https://doi.org/10.1080/10408398.2010.507908.

[95] B. Fei, Hyperspectral imaging in medical applications, Data Handl. Sci. Technol. 32 (2019) 523–565. https://doi.org/10.1016/B978-0-444-63977-6.00021-3.

[96] S. V. Panasyuk, S. Yang, D. V. Faller, D. Ngo, R.A. Lew, J.E. Freeman, A.E. Rogers, Medical hyperspectral imaging to facilitate residual tumor identification during surgery, Cancer Biol. Ther. 6 (2007) 439–446. https://doi.org/10.4161/CBT.6.3.4018.

[97] F. Daniel, A. Mounier, J. Pérez-Arantegui, C. Pardos, N. Prieto-Taboada, S. Fdez-Ortiz de Vallejuelo, K. Castro, Hyperspectral imaging applied to the analysis of Goya paintings in the Museum of Zaragoza (Spain), Microchem. J. 126 (2016) 113–120. https://doi.org/10.1016/J.MICROC.2015.11.044.

[98] M.Á. Fernández de la Ossa, J.M. Amigo, C. García-Ruiz, Detection of residues from explosive manipulation by near infrared hyperspectral imaging: A promising forensic tool, Forensic Sci. Int. 242 (2014) 228–235. https://doi.org/10.1016/J.FORSCIINT.2014.06.023.

[99] C.S. Silva, M.F. Pimentel, R.S. Honorato, C. Pasquini, J.M.P. Montalbán, A. Ferrer, Near infrared hyperspectral imaging for forensic analysis of document forgery, Analyst. 139 (2014) 5176–5184. https://doi.org/10.1039/C4AN00961D.

[100] G.J. Edelman, E. Gaston, T.G. van Leeuwen, P.J. Cullen, M.C.G. Aalders, Hyperspectral imaging for non-contact analysis of forensic traces, Forensic Sci. Int. 223 (2012) 28–39. https://doi.org/10.1016/J.FORSCIINT.2012.09.012.

[101] J. Qin, Hyperspectral Imaging Instruments, Hyperspectral Imaging Food Qual. Anal. Control. (2010) 129–172. https://doi.org/10.1016/B978-0-12-374753-2.10005-X.

[102] A. Carrasco-Pancorbo, N. Navas-Iglesias, L. Cuadros-Rodríguez, From lipid analysis towards lipidomics, a new challenge for the analytical chemistry of the 21st century. Part I: Modern lipid analysis, TrAC Trends Anal. Chem. 28 (2009) 263–278. https://doi.org/10.1016/J.TRAC.2008.12.005.

[103] M. Sliwi´nskasliwi´nska-Bartel, D. Thorburn Burns, C. Elliott, Rice fraud a global problem: A review of analytical tools to detect species, country of origin and adulterations, Trends Food Sci. Technol. (2021). https://doi.org/10.1016/j.tifs.2021.06.042.

Chapter 2

Materials and Methods

This section of the Thesis is aimed at introducing the main analytical approaches used to face some issues linked to the agri-food field. A concise description of the utilized chemometric tools and the experimental setups is reported. More details about the theoretical principles behind the multivariate data analysis techniques will be discussed in the chapters describing the projects in which these techniques have been used.

2.1. Chemometrics

The following sections introduce the main chemometric tools used in the research projects presented in this Thesis.

2.1.1 Pre-processing

After instrumental acquisition, raw data are generally noisy and can be affected by undesired effects, such as baseline drifts or, in the case of NIR measurements, light scattering. These effects significantly impact the quality of the data and consequently the reliability and effectiveness of the subsequent multivariate data analysis. Therefore, data pre-processing [1,2] is necessary to transform raw data into a "cleaner" version ready for further analysis. This process potentially consists in several steps, which can be resumed in:

- *Data cleaning*: where the presence of noise, outliers, missing data and duplicates is detected and tackled.
- *Data transformation*: raw data are converted into formats or units that are more suitable for the multivariate methods that will be used for the analysis.
- *Data reduction*: data are treated to remove redundancy and can be reorganised to efficiently perform the data analysis steps.

Depending on the data type, different approaches can be used. A list of the pre-processing techniques employed in this Thesis is described below.

Pre-processing of NIR data [3]:

- **Savizky-Golay derivative**: sample-wise method used to smooth and differentiate spectral data while preserving important features like peaks. It calculates a polynomial fit of a chosen order in each filter window of selected width as the filter is moved across the spectrum, calculating a derivative of order n in each filter window. In the case of this Thesis' work, it was chosen a first-order derivative, second-order polynomial and a 11pt window.
- **Standard Normal Variate (SNV,** [4]**)**: sample-wise method that corrects scatter effects in spectral data. It works by performing a normalization of the spectra consisting in subtracting from each spectrum its own mean and then dividing it by its own standard deviation.
- **2-Norm**: also known as "Euclidean normalisation", allows to have all the data points in a consistent scale, reducing bias due to intensity differences. It is a sample-wise method that involves the division of the spectrum's values by the sum of the squared value of all variables for the given sample.
- **Mean centering**: column-wise method that helps emphasizing the relative differences among variables, making the data suitable for techniques like PCA which rely on variance for features extraction. It involves the calculation of the mean value of the whole dataset and its subtraction from each spectrum.

Pre-processing of morphological features (HSI rice project):

- **Autoscale**: column-wise method used to scale the variables with different unit or magnitude, so that they contribute equally to multivariate analysis. It can be seen as an extension of mean centering, as it starts by mean-centering the values of a variable, and then it divides the result by the variable's standard deviation.

2.1.2. Exploratory data analysis

Exploratory data analysis (EDA) can be considered as the very first step in the process of data evaluation. In this phase, the aim is to look at the data to inspect for the information contained in the data, to see if there are correlations among samples or variables, and to discern between information and noise. Generally, EDA relies on graphical visualisations to display the results of the analyses, and sometimes it operates with the aid of a toolbox.

EDA was born in the 1970s and the credit of its development is to be ascribed to John Tukey, who published a book [5] in 1977 which is considered as the foundation of the modern exploratory analysis. Traditionally, data analysis was executed as a hypothesis-driven approach, and scientists used statistics to confirm or reject a starting research question. The revolution carried by Tukey involved the consideration of how the importance of exploring the data before the formulation of a hypothesis, and to consequently develop a research direction based on the results of this investigation. In this concern, EDA becomes an unsupervised

methodology, which means that it is released from any *a priori* postulations and it is used by the analyst as a tool to extract the naturally present information from the data, which is then interpreted according to the knowledge of the analyst.

2.1.2.1. Principal Component Analysis (PCA)

Within the world of exploratory data analysis, Principal Component Analysis (PCA, [6,7]) is the most known and used methodology for data inspection. The popularity of PCA is linked to the possibility of treating complex and plentiful data matrices to obtain a clear and rapid representation of the data: this makes PCA suitable for its application in the scientific field, allowing to treat large datasets and to reduce the time of data analysis.

This method consists in a bilinear decomposition technique that allows to switch from the high-dimensional space of the original data to a new one, characterised by lower dimensions, so that the complexity of the data is reduced, while preserving the information and structure of the data. This new space is described by new summarizing variables, called Principal Components (PCs), which are created by a linear combination of the starting variables. The Equation 2.1 shows the decomposition performed by PCA:

$$X = TP^T + E = \hat{X} + E \tag{2.1}$$

where X is the original matrix of the data, T is the vector of the scores that represents the projection of the samples in the new space, and P is the vector of the loadings, which are the coefficients (a sort of "weights") of the original variables in the space of the PCs. $\hat{X}$ corresponds to the matrix of the modelled data, while E is the matrix of the unmodelled data, also called the residual matrix, which is useful to identify anomalies in the modelled data, such as the presence of outliers, and to assess if the chosen number of PCs is correct or not (for example, large residuals mean that a significant amount of information remains unexplained, suggesting that more PCs might be included in the model).

The role of the PCs is to explain the variability contained in the original data. The first PC describes the direction of maximum variance and is followed by the second PC that explains the next maximum possible variance in the orthogonal direction with respect to the first PC, and so on for all the following PCs. In this new space, the samples distributions present in the original data can be visualized with the scores plot, with the possibility of inspecting possible similarities, groups and/or trends among the samples. The original variables are represented by the loadings plot, with which it is possible to inspect the patterns of correlation among the variables. Another important source of information is represented by the residuals plots, which is very useful for the detection of samples not well-described by the model (provided that the correct number of PCs is defined for the model), usually reflecting substantial differences between them and the majority of the samples;

these samples are generally identified with the term “outliers”, and their possible removal from the dataset is subject to careful considerations.

This technique was used to inspect the data in all projects contained in this Thesis.

2.1.2.2. Cluster analysis

One of the areas of interest in the field of multivariate data analysis is the so-called pattern recognition, a term which stands for the process of identification of patterns within the data. One of the most used techniques in this context is Cluster Analysis [8]. Similarly to PCA, Cluster Analysis is another unsupervised analytical method, as it does not require specific instructions about how to look for groups, thus it can be applied also to datasets where data are not labelled. A difference with PCA consists in the need to give some instructions about how to perform the clustering, by defining specific criteria prior to the analysis: Cluster Analysis involves the calculation of the distance among samples to obtain the clusters, so a parameter that is preliminary chosen is the type of distance, or the clustering method (for example the Ward’s method [9]).

The algorithms used for this multivariate analysis are various and differ according to the approach used to group the data. The most famous are partitioning, hierarchical and density-based algorithms. The first type refers to all the algorithms that divide the data according to a defined number of clusters. The most popular algorithm in this respect is K-means [10,11]. Density-based algorithms, as the name itself suggests, focus the attention on discerning high data density areas from the lower ones. A very used algorithm in this sense is DBSCAN (Density-Based Spatial Clustering of Applications with Noise, [12]).

When using hierarchical clustering, two approaches can be applied: the agglomerative and the divisive one. The first one is the most common and used, and it starts considering each sample as a single cluster. Then the algorithm iteratively identifies the samples that are most similar to each other, by means of calculations of a distance metric (for example, the Euclidean, Mahalanobis or Manhattan distance). These samples are then merged, originating new clusters, and the distance metric is recalculated. The algorithm repeats these steps until all samples are assembled in an all-embracing cluster. Visually, the result is a nested structure generally known with the name of “dendrogram”. An example is shown in **Figure 2.1**.

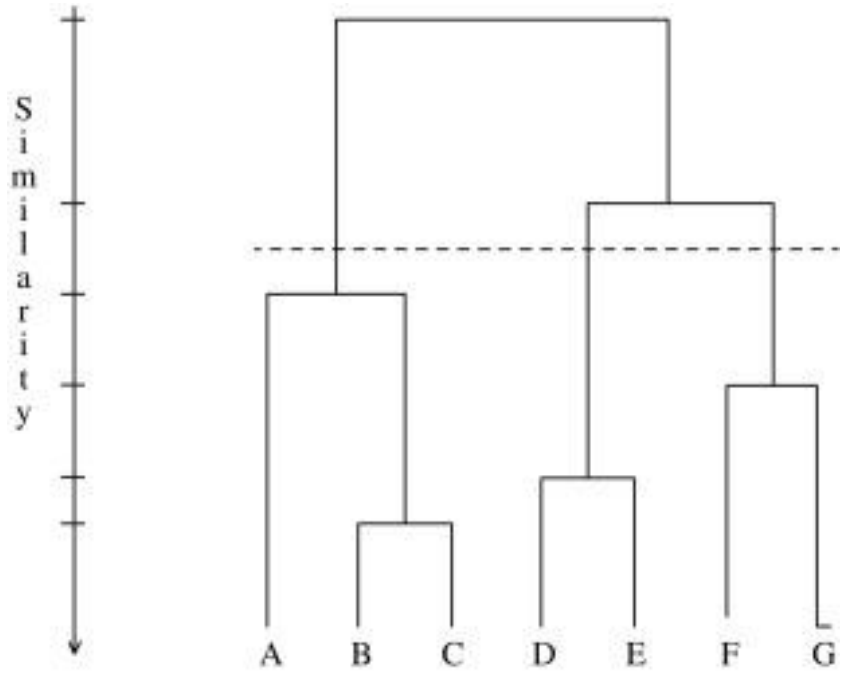

Figure 2.1. Dendrogram representing a hierarchical clustering structure (from Algebraic and Combinatorial Computational Biology, 2019).

The divisive approach works in the opposite way: initially, all the samples represent a single cluster, which is iteratively divided into sub-clusters by calculating the distance metric and finding samples that are most dissimilar to each other. The original cluster is divided into two sub-clusters, the distance metric is recalculated, and the process continues until a desired number of clusters is reached. Depending on the aim of the research, one of these two approaches can be selected. For this work, the agglomerative approach was chosen, and Ward's method was chosen as criterion to build the dendrogram structure.

As clustering is an unsupervised method, the analysed data are unlabelled. Consequently, the results obtained with cluster analysis require an expert knowledge to be interpreted, and this is one of the main criticalities of this approach. Nevertheless, cluster analysis is a powerful analytical technique and allows to treat unlabelled data, a point that makes this method a viable solution when dealing without defined groups among samples.

Hierarchical clustering was employed in this Thesis for the research conducted in the rice project. For a more detailed explanation of this methodology and its application in this work, please refer to Section 5.3.

2.1.3. Multivariate regression

One of the first thing learnt by chemistry students in their Bachelor's laboratory lessons is how to build a calibration line that will be used to find the concentration of an unknown sample, often exploiting UV-Vis measurements. Here, the relation

between the concentration of the samples and the response signal can be expressed as:

$$y = a + bx \tag{2.2}$$

where y is the response signal, in this case the UV-Vis absorbance of the samples, and x is the concentration of the sample. The variables a and b represent the unknown values of the equation, i.e. the intercept and the slope of the line, and are experimentally calculated.

The same idea can be extended in the context of multivariate data. In this case, Y and X become matrices, and Equation 2.2 can be written as follows:

$$Y = \beta X + \varepsilon \tag{2.3}$$

where β stands for the set of regression coefficients, represented as a column vector, and ε is the "error" term, which is also a column vector. X is the matrix of the independent variables, while Y is a column vector of the dependent variables. This equation represents a system of linear equations, where each row of Y is explained by the corresponding row in X and the coefficients in β.

The aim of this approach is to model the relationships between a set of response variables Y and a set of predictor variables X, so that it will be possible to predict the response variables of samples for whom just the values of the predictor variables are known (i.e., new/unknown samples). The main multivariate regression techniques used in chemometrics are Principal Component Regression (PCR, [13]) and Partial Least Squares (PLS) regression [14].

2.1.3.1. Partial Least Squares (PLS) Regression

While PCR is a more basic and traditional form of multivariate regression, not so much applied in modern modelling, PLS Regression is a very common and powerful multivariate regression method: its success and widespread application is due to the possibility of dealing with the complex and plentiful datasets derived from the modern analytical techniques, such as spectroscopy and chromatography. These data, besides providing a huge number of predictor variables, are also generally noisy and correlated. For this reason, traditional regression methods such as PCR and Multiple Linear Regression (MLR, [15]) are being substituted with other multivariate regression approaches, and PLS Regression proved to be one of the most performing in this sense. More information about this technique and its application in the present work is available in Section 3.3.

2.1.4. Classification

Classification represents a supervised learning approach that is used to create categories of samples depending on their properties. The starting point of this methodology involves the training of a mathematical model on a dataset of samples with labelled characteristics. The information contained in the data is recognised and used by the model to differentiate samples into categories, and once these patterns are learnt and tested, they are used to predict in which group a new unknown sample would be assigned to.

Multivariate classification methods started to be used in the 20th century, when machine learning algorithms were developed specifically for classification purposes. In these years, techniques like Linear Discriminant Analysis (LDA, [16]) and Support Vector Machines (SVM, [17]) were born. The development of machine learning techniques and their application in chemometrics during the 21st century led to increasingly robust and powerful classification models, allowing new methods to bloom. Nowadays, the applications of multivariate classification are mainly addressed to sample identification [18,19], quality control [20,21] and process monitoring [22].

2.1.4.1. Partial Least Squares - Discriminant Analysis (PLS-DA)

Partial Least Squares - Discriminant Analysis (PLS-DA, [12]) is one of the most known and consolidated classification methods, thanks to its versatility and robustness. It is a widespread classification technique, particularly appropriate to handle complex and abundant datasets with rapid and reliable responses. The key, in this sense, is the compression of the data in a space of lower dimensionality, that simplifies their visualisation and interpretation, and also lightens the computational operations and the operating time. Another relevant advantage of this method is represented by the possibility of reducing the effects of noise and missing data, avoiding influences and random fluctuations that can mislead the model. More details about PLS-DA and its application in this Thesis are reported in Section 5.4.1.

2.1.4.2. Hierarchical classification

Hierarchical modelling [23,24] represents a classification technique which is gaining popularity thanks to its efficiency in dealing with complex matrices: this methodology is used when a hierarchical connection among the classes is present, i.e., subclasses are included in bigger ones. For this reason, hierarchical models are generally represented by tree-structured diagrams, whose main structures are shown in **Figure 2.2**. These diagrams are also known by the name of “dendrograms”, from ancient Greek “dendron”, meaning “tree”.

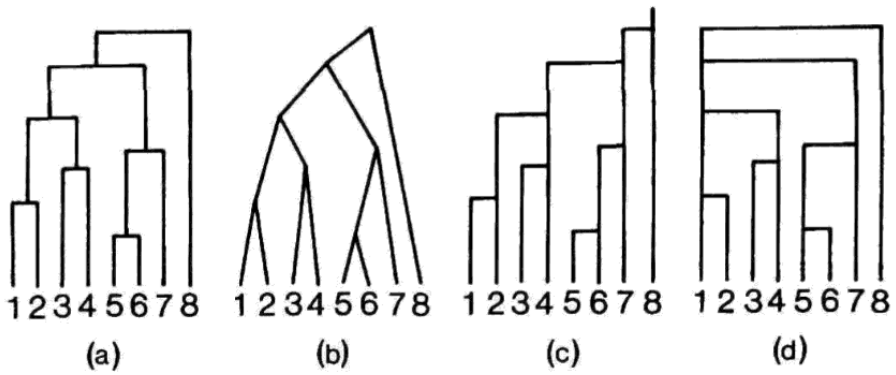

Figure 2.2. main ways for visual representation of dendrograms. (a) and (b) are the most used. (from Clustering and Classification, Arabie Hubert and De Soete, 1999, [25])

Although there are many ways to find groups of samples by means of this methodology, the most used hierarchical algorithms in chemometrics are agglomerative and divisive. In the first case, each object in the dataset represents a class itself, and at each iteration of the algorithm the number of classes progressively decreases, as the most similar ones are agglomerate (merged). Depending on the definition of similarity/dissimilarity between classes, different clustering strategies can be chosen. Divisive algorithms exploit the same criteria for the identification of groups among samples, but work in the opposite direction: initially, all the objects belong to a general and comprehensive class; then, any existing class is divided into two, until each object of the dataset is separated and represents an individual class.

A deeper description of the algorithms behind this methodology, the types of hierarchical classification and how this approach was used for the classification purposes of this Thesis can be found in Section 5.4.2.

2.2. Near-Infrared (NIR) spectroscopy

In the following sections, the instruments used for the spectroscopic analyses in the projects of this Thesis are presented.

2.2.1. NIR experimental setup

The NIR spectrometers consist in a source of radiation (the most used is a tungsten-halogen lamp, but also LED and diode lasers are employed), an interferometer that allows to measure all the wavelength of the interval of interest at once, and a detector. The information is collected and represented by an interferogram, which needs to be mathematically transformed in order to make the information interpretable: to do that, the Fourier transform is applied, converting the interferogram into the infrared spectrum.

The NIR spectrometers exploit two main different operating modes: transmittance and reflectance. Transmittance is usually applied in the analyses of liquid samples, with the light beam going through the specimen and being detected on the opposite side. On the other hand, reflectance is used to analyse solid samples, and in this case the light beam is reflected by the surface of the material and then collected by the detector, which lies on the same side of the light source. The physical principle behind the operation mode of the spectrometers implies that the light beam coming from the source is depleted of certain wavelengths after the interaction with samples, due to the absorption of light by the molecules. The spectrometers measure the difference between the starting signal and the one recorded after the interaction and translate this information into a transmittance or reflectance spectrum, which can be eventually turned into an absorbance spectrum.

Lots of NIR spectrometers are benchtop laboratory instruments, but in the last decades portable handheld NIR devices became increasingly popular [26], thanks to their suitability and to the possibility of making *in-situ* analyses. In this thesis, both benchtop spectrometers and the SCiO Pocket molecular sensor (v1.2, Consumer Physics Inc., Tel Aviv, Israel) were used to collect the NIR spectra in reflectance mode.

2.2.1.1. Benchtop instrument: Bruker MPA

A benchtop Fourier transform-NIR (FT-NIR) spectrometer (Multi-Purpose Analyser–MPA, Bruker Optics, Ettlingen, Germany) equipped with an integrating sphere and an optical fibre reflectance probe was used for samples acquisitions. This spectrometer is represented in **Figure 2.3**. The instrumental settings for the MPA operated in sphere mode were: 800–2780 nm (12500–3600 cm^{-1}) spectral range, 8 cm^{-1} optical resolution, and 10 kHz scanner velocity. A sample holder of 9 cm of diameter, equipped with a quartz window on the bottom, was used to sample acquisition. Regarding the MPA operated in optical fibre probe mode the instrumental settings were: 800–2500 nm (12500–4000 cm^{-1}) spectral range, 16 cm^{-1} optical resolution and 20 kHz (probe) scanner velocity. In both MPA acquisition modes, 64 scans for both sample and background acquisition were collected, which were eventually averaged resulting in one individual spectrum for each sample. Background scans were performed using the instrument's internal reference standard. The Opus software (v6.5, Bruker Optics, Ettlingen, Germany) was used for instrumental control and for spectra acquisition.

Figure 2.3. Multi-Purpose Analyser—MPA II by Bruker.

2.2.1.2. Portable instrument: the SCiO

A representation of the SCiO portable spectrometer is displayed in **Figure 2.4**. The instrumental settings for the SCiO device are non-customizable and are fixed as follows: 740–1070 nm (13510–9340 cm^{-1}) spectral range, 10 cm^{-1} resolution, and time scan of approximately 5s. Spectral scans were managed through the SCiO smartphone app (The Lab, version 2.5.3), which allows for controlling the sensor via Bluetooth connection and also uploads all acquired data on the Consumer Physics Cloud database, which is accessible via browser to inspect and download the raw data.

Figure 2.4. the portable SCiO spectrometer.

2.2.1.3. HSI instrumentation

The HSI device used in the rice project is shown in **Figure 2.5**. It involved a NIR-HSI camera (Specim FX17, Finland), connected with a motorized scanning bed (LabScanner Setup 40×20 cm) where samples were placed. The instrumental settings for this setup were: 935–1720 nm spectral range, 2.25 mm/s bed scanning speed, 3.5 ms exposure time, 32 Hz frame rate. To calibrate the NIR camera before each acquisition, an internal black reference standard and an external white reference standard surface were used, as the black standard represents the complete absence of light reflectance, while the white one provides the total reflection of light. The white surface lies at the beginning of the scanning bed, just before the samples, and the image acquisition starts once the white calibration is finished.

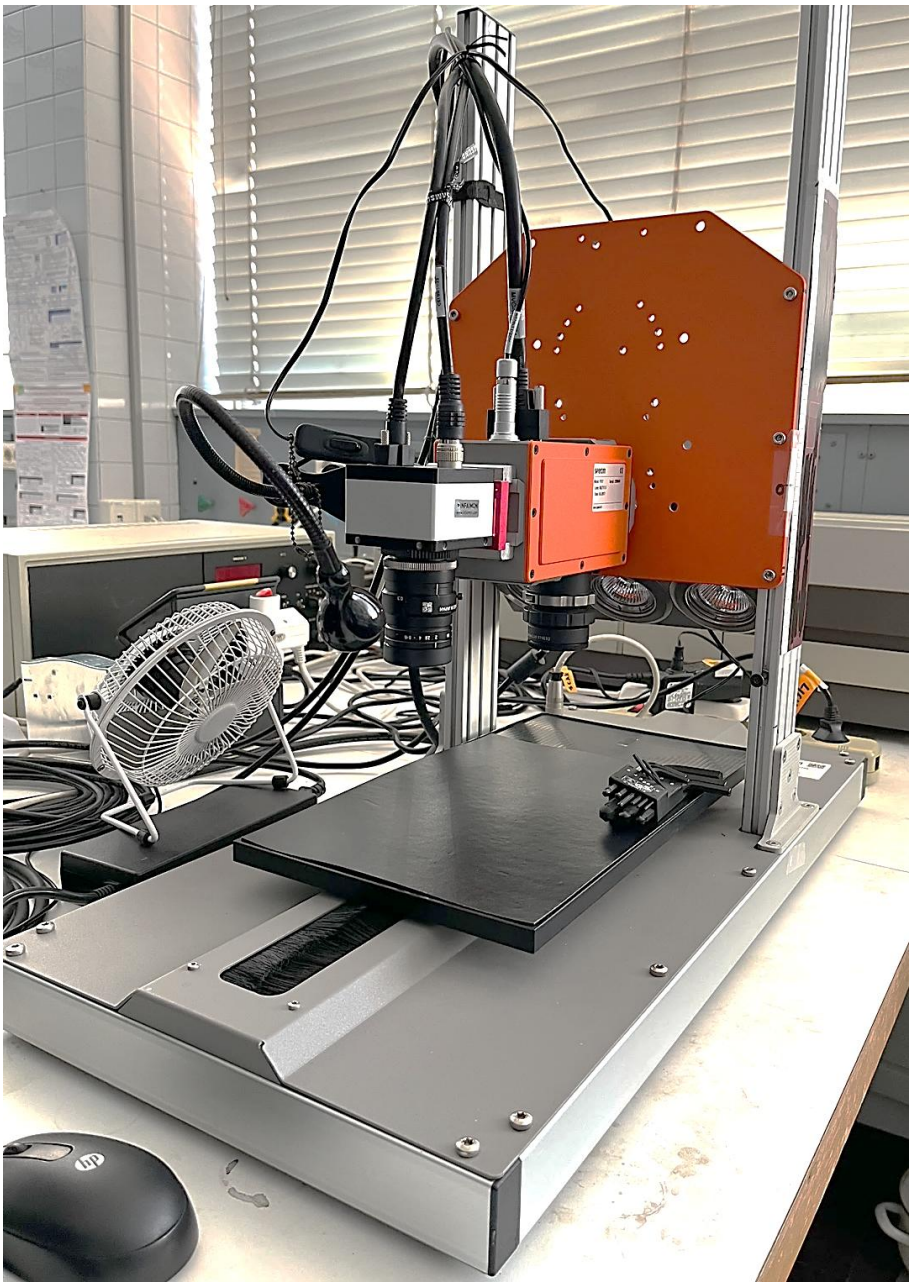

Figure 2.5. NIR-HSI setup of the University of Barcelona, used in the rice project.

Concerning the Help2Grow project, the hyperspectral images were acquired using a line scanning system (NIR Spectral Scanner, DV Optic) equipped with a Specim ImSpector N17E imaging spectrometer coupled to a Xenics Xeva-1.7-320 camera (320 × 256 pixels) embedding Specim Oles 31 f/2.0 optical lens. The spectral range covered by the hyperspectral system was between 900 a 1700 nm, although only the wavelengths in the interval 980–1660 nm were considered, to remove the noisy regions in the extremities of the spectra. The spectral resolution was 5 nm. The background of the hyperspectral images was represented by a black silicon carbide sandpaper sheet. A ceramic tile with a 99% reflectance standard reference and two ceramic tiles with intermediate reflectance values, respectively corresponding to 89% and 46%, were included in the images.

2.3. Case studies: samples and datasets

In this section, the samples analysed in this Thesis and the data pre-processing are described. The samples of the NEWPOW project are firstly presented, followed by those of the Rice-HSI and Help2Grow projects.

2.3.1. NEWPOW project: description of the hazelnut samples and lipid extraction

A total of 56 samples of raw hazelnuts from different countries (Italy, Turkey, Azerbaijan, Georgia, and Chile) were analysed. Each sample was ground with a Retsch ZM200 grinder (Retsch Gmbh, Haan, Germany) and sieved using a vibratory sieve shaker BA 200N (CISA Sieving Technologies, Lliçà de Vall, Barcelona, Spain). The particles of sizes in the range 250–500 μm were selected, and finally all samples were stored in hermetically sealed polyethylene bags at about –20 °C. Prior to analysis, samples were allowed to warm up to room temperature (~20 °C).

Colleagues of the University of Turin handled the lipids extraction processes, detailed hereafter. To reduce the volume of organic solvent used and the time extraction [27], the crude fat content of the samples was determined according to the Randall/Soxtec modification of the standard Soxhlet extraction method (AOAC 948.22)[28]. The hot solvent extraction process was carried out with a SER 148 Solvent Extractor (Velp Scientifica Srl, Usmate Velate (MB), Italy) equipped with three Soxhlet posts. In each Soxhlet post, 5.000 ± 0.001 g of hazelnuts powder was extracted with 99% n-hexane, analytical grade (Sigma-Aldrich, Milan, Italy), for a total of 120 min. The extraction process involved three semi-automated steps. During the first step, the thimbles containing the sample were immersed in the boiling solvent (60 mL, 130 °C, 60 min); then, the level of the solvent was lowered below the extraction thimbles. The second step (washing step, 60 min) allowed the continuous flow of condensed solvent over the sample and through the thimble, completing the solvent extraction. During the last step (30 min), as much solvent as

possible was distilled and recovered from extraction cups until apparent dryness. Finally, the extraction cups, including the extracts, were dried at 105 °C for 1 h, cooled in a desiccator to room temperature, and weighed to calculate the extract percentage.

The results were expressed as total grams of extracted lipids, with their respective percentages related to the initial weight of ground hazelnuts. Three replicates were extracted simultaneously for each hazelnut sample.

2.3.1.1. NIR spectra acquisitions and pre-processing

All NIR analyses were performed using the benchtop MPA spectrometer and the SCiO portable device described in Section 2.2.1, where also the instrumental settings for the two instruments are highlighted. The measurements were performed in reflectance mode on the ground samples obtained as described in Section 3.3.1. For each specimen, three spectra were collected as replicates with the benchtop MPA instrument, while six replicates were acquired with the SCiO portable device. An average spectrum was then calculated from the replicates after proper replicate quality evaluation. One average spectrum was therefore obtained for each individual sample, and all further data analysis steps were done on the averaged spectra.

The raw NIR spectra were analysed under MATLAB environment. The MPA spectra needed to be cut at the extremities, to remove areas that were particularly noisy. The obtained range was 1117–2561 nm for the MPA operating in sphere mode and 1121–2254 nm for the MPA operating in probe mode. Such correction was not necessary for the SCiO spectra. Furthermore, all datasets were pre-processed with SNV, with the purpose of eliminating artefacts and correcting nonlinear behaviours, due to potential scattering effects originating from the granular nature of the samples. Mean-centring was also applied prior to any multivariate data analysis.

Figure 2.6 provides a visual representation of the data preprocessing steps, depicted according to the three datasets. The spectra with no pre-treatment (**Figure 2.6a–c**) are reported in the first row; the spectra pre-processed with SNV (**Figure 2.6d–f**) lie in the central row, and the spectra with SNV followed by mean centering preprocessing (**Figure 2.6g–i**) in the bottom row. The samples are coloured according to their lipid content: blue corresponds to lower lipid content determined by Randall/Soxtec extraction, while yellow corresponds to higher lipid content.

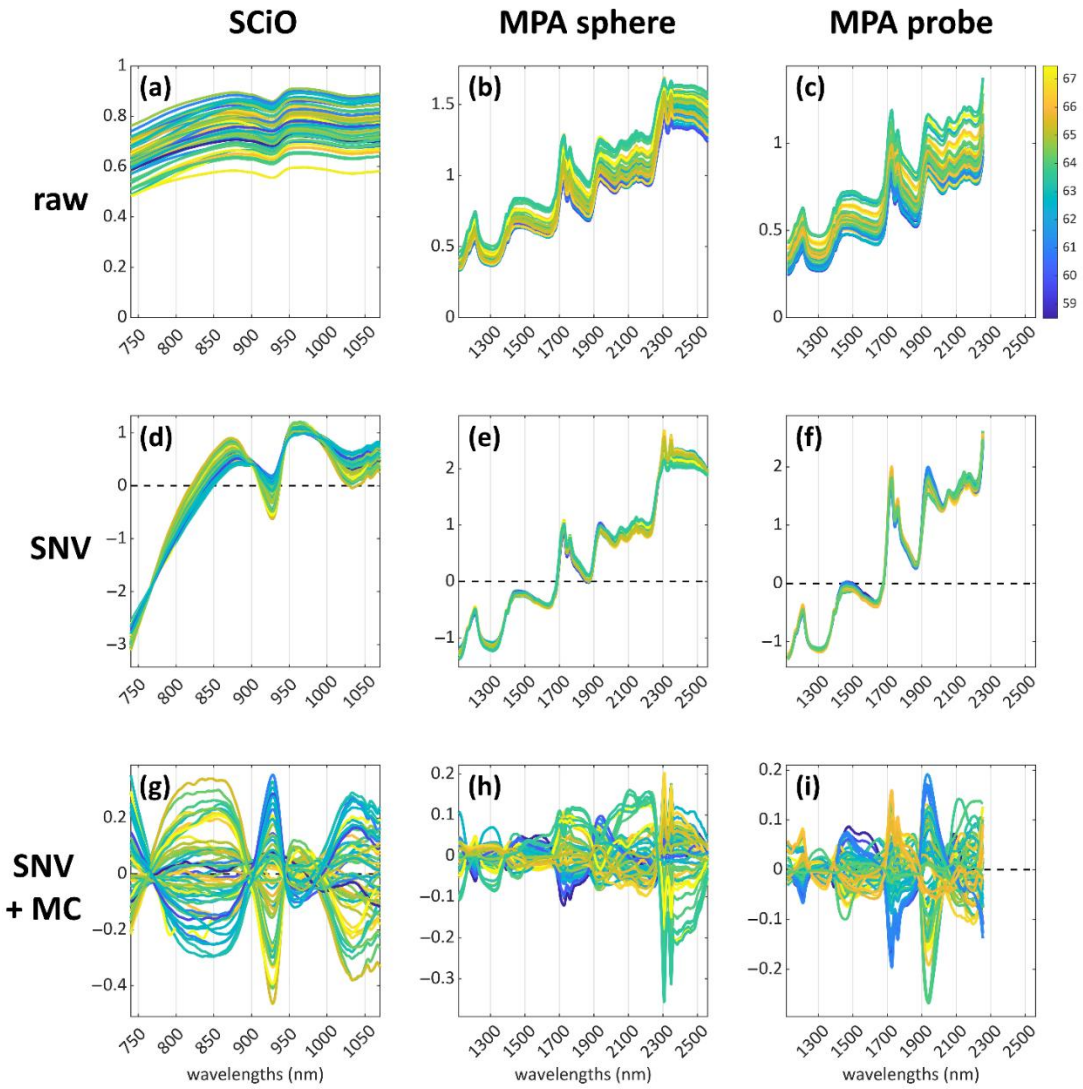

Figure 2.6. Visual representation of the data-preprocessing pipeline (from top to bottom along each column): raw NIR spectra of the hazelnut samples (a–c); spectra preprocessed with Standard Normal Variate (SNV) (d–f); spectra preprocessed with SNV + mean centering (MC) (g–i).

2.3.2. Help2Grow project: description of the vineyard and the leaves samples

A scheme of the vineyard where the acquisitions were performed is displayed in **Figure 2.7**, while **Figure 2.8** represents a picture of the vineyard taken during an investigation day in June. The vineyard consisted in 16 grapevine lines, divided into 8 lines where the fungi responsible of the development of downy mildew were inoculated, and 8 lines for the inoculation of the fungi responsible of the development of powdery mildew. Thirteen different treatments were applied, each one replicated four times at different vineyard heights (indicated in **Figure 2.7** with the letters a, b, c, d). Among the 13 total treatments applied to the plants, only seven were considered of interest for the purposes of the Help2Grow project, and

consequently just the grapevines that underwent the chosen treatments were analysed. The description of the selected treatments with the corresponding numbers in the grapevine lines are summarised in **Table 2.1**.

Table 2.1: description of the treatments applied to the grapevine plants and the respective number in the vineyard.

Treatment number on the vineyard	Description
0	No treatment applied to the plants
1	Plants treated just with water
2	Plants treated with fungicide
6	Plants treated with activated water "T-sonik"
7	Plants treated with fungicide + activated water "T-sonik"
9	Plants treated with half-dosage fungicide
13	Plants treated with half-dosage fungicide + activated water "T-sonik"

The treatments number six, seven and thirteen involved the use of the "T-sonik" water: this name refers to a technology that involves the use of ultrasonic devices to improve the properties of water that will be used in agriculture. These ultrasonic activators increase the oxigenation of water and ease the solubilisation of nutrients, improving the crop quality and yield with resulting in less maintenance.

The *in-situ* analyses were performed from the 6th of June until the 1st of August, one time per week. After the summer break, three more sets of analyses were carried out in October (12th, 19th and 26th of October 2022). The plants were analysed *in-situ* with the SCiO portable spectrometer, collecting six spectra for each treatment, also analysing the four replicates for both the diseases affecting the plants. The amount of spectra collected after each investigation day was 336, which was eventually reduced to 14 spectra after averaging firstly the six acquisitions per treatment, and then the four replicates.

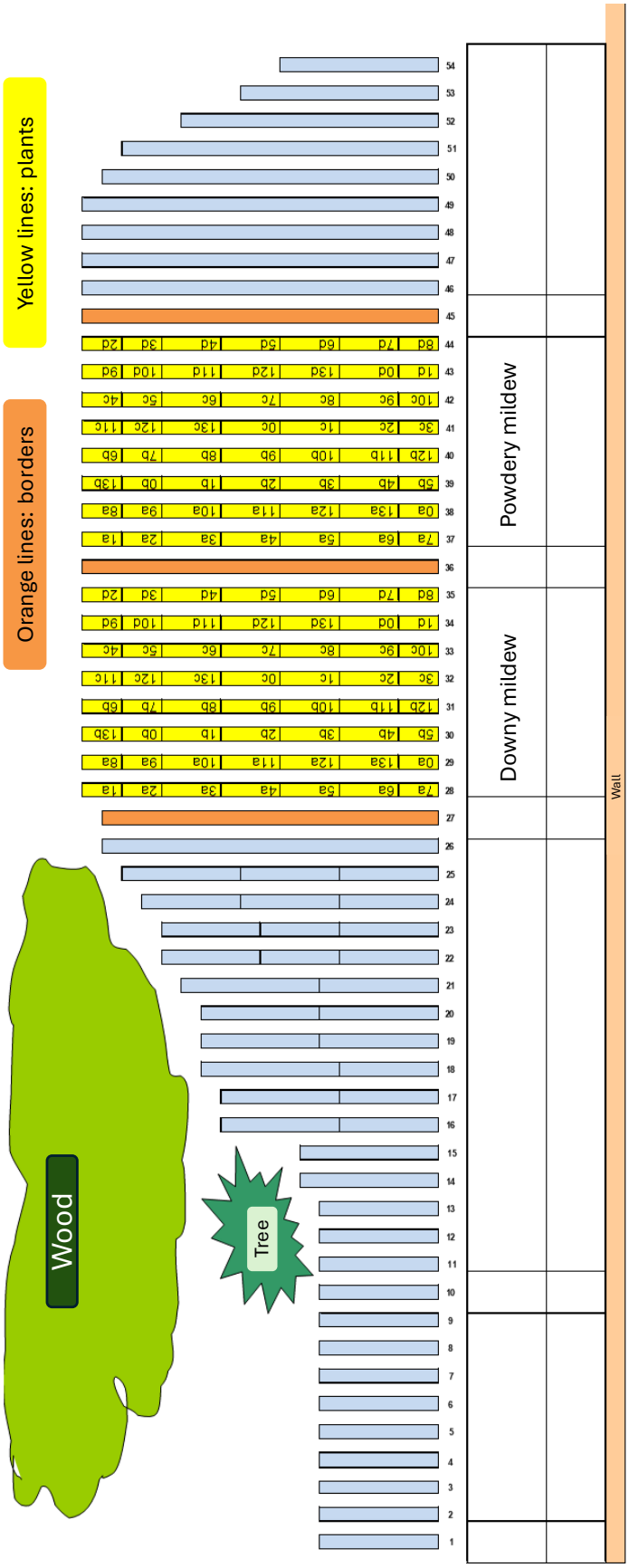

Figure 2.7. scheme of the vineyard in the Department of Agricultural, Forest and Food Science of the University of Turin, where the acquisitions were performed.

Figure 2.8. photo of the vineyard of the University of Turin, taken during one investigation day in June.

Unfortunately, some criticalities occurred in the achievement of the Help2Grow targets: the Summer of 2022 was particularly hot and dry, and these extreme conditions limited the normal and expected development of the downy and powdery mildew. In fact, during the last summer investigation on August 1st, the grapevine leaves still looked healthy, and no typical signs of the diseases were observed. During the investigation days of October, it was noticed that some plants started developing the diseases, particularly those affected by the downy mildew. Concerning the other diseases, it was too difficult to state the presence of the powdery mildew with the naked eye. In the attempt of obtaining some more information about the development of these diseases, in particular for the powdery mildew, it was chosen to collect some leaves during the last investigation on October 26th, which were frozen at -20 °C the same day in the laboratory of Polytechnic of Turin. Later, they were brought to the laboratory of the University of Modena and Reggio Emilia to be analysed through hyperspectral imaging.

The leaves were collected following this criterion: both for the lines of downy and powdery mildew diseases, the plants were inspected to collect a set of healthy and a set of ill leaves, or those that at least appeared to be so to the naked eye. This was particularly difficult considering the powdery mildew, because even expert personnel were not able to certainly state which leaves were ill or healthy. On the contrary, the plants affected by downy mildew clearly showed the typical white dots profile on the back of the leaves, thus in this case it was easier to discern between healthy and ill leaves and to correctly sample them.

Another challenge arisen during the hyperspectral analyses concerned the leaves dryness and yellowing, due to the advance of the fall season, and that made impossible for many leaves to be analysed. For this reason, it was not possible to analyse the leaves of plants belonging to all the treatments considered in this project, and the analysis was restricted to the leaves of treated plants showing better physical conditions. For each selected treatment, 4 leaves were collected, reaching a final amount of samples equal to 40. **Table 2.2** summarizes the sampling done during the last investigation day of October.

Table 2.2. health condition and type of treatment of the leaves collected in the investigation day of the 26th of October 2022.

Health condition of the leaf (to the naked eye)	Type of treatment considered
Healthy	• No treatment • Fungicide • Activated H_2O "T-sonik"
Ill	• No treatment • Activated H_2O "T-sonik"

The sampled leaves were kept frozen until the 6th of May 2024, when they were transported to Reggio Emilia and analysed with HSI. This part proved to be critical, as the first problem concerned the long conservation time of the leaves, which could have caused their lyophilization and a possible consequent adulteration of the samples. Secondly, the leaves were transported to the University of Modena and Reggio Emilia with a thermostatic box equipped with ice-cold tiles.

2.3.2.1. Data acquisition and pre-processing

Once the laboratory located in Reggio Emilia was reached, the leaves were firstly analysed with the SCiO instrument, and then, immediately after, with the HSI equipment. For each leaf, three acquisitions were performed with SCiO on the front and the back side, collecting six spectra per leaf which were averaged, obtaining one spectrum for the front and one for the back side. The spectra were pre-processed with SNV prior to any multivariate data analysis.

Concerning the hyperspectral acquisitions, two leaves of the same treatment at a time were put on the black sandpaper sheet, and an image was collected for the front and one for the back side. The final amount of collected images was 80. After the acquisition, the images were processed with an in-house MATLAB routine. The first step involved the selection of the Region of Interest (ROI, [29]) ich was used to cut the images as close as possible to the leaves' profiles. Then, the background was removed by selecting a threshold value of reflectance at the wavelength of 1000

nm: all the pixels having reflectance lower than 0.42 were not considered in the further steps.

Prior to any multivariate data analysis, the hyperspectral images were then converted to one-dimensional signals, called Common Space Hyperspectrograms (CSH, [30]): this procedure starts with an initial unfolding into 2D matrices, with each row corresponding to one image pixel and each column corresponding to each wavelength in the analysed spectral range. These matrices are then pre-processed with SNV and scaled using the mean spectrum obtained by calculating the average spectrum of all the pixels of the collected images. After this step, a PCA was calculated considering the first four PCs, and the CSH of each image was obtained by plotting the frequency distribution curves of the four PCs, the Q residuals and the Hotelling's T^2 vectors, considering 137 bins. The representation of the global CSH for the collected images is represented in **Figure 2.9**.

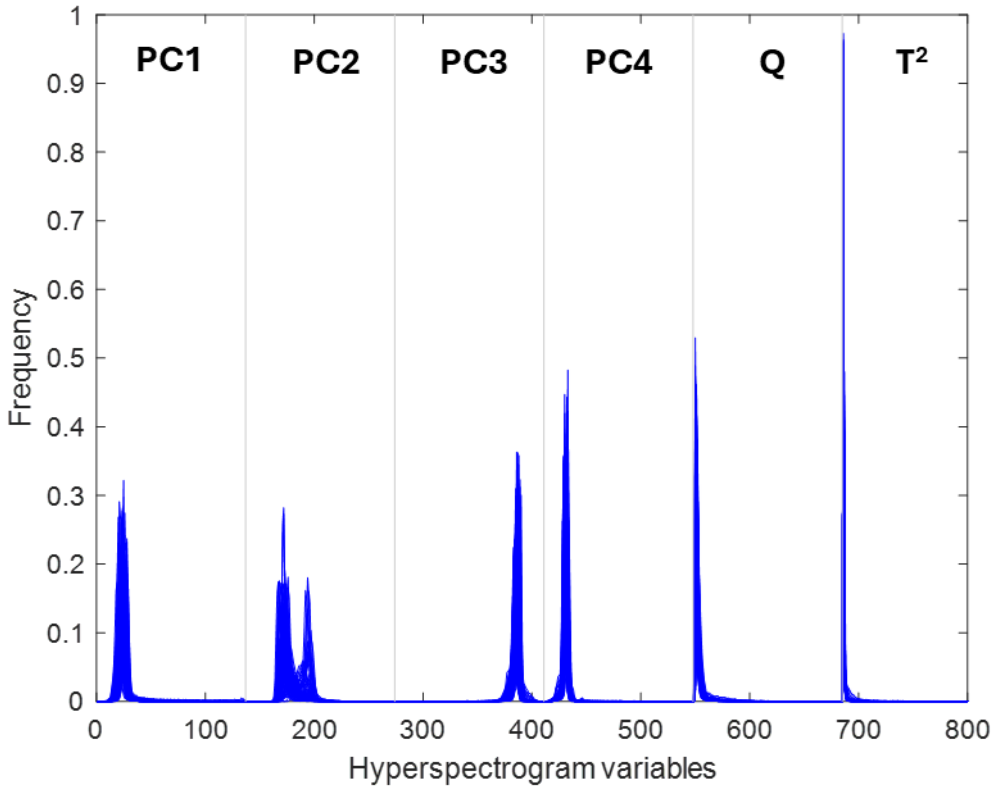

Figure 2.9. visual representation of the CSH for the 40 leaves analysed with HSI.

Once calculated, CSH data were analysed through PCA to inspect the information about the grape leaves, as described in Section 4.3.

2.3.3. HSI project: description of the rice samples

Ente Nazionale Risi, a public economic body supervised by the Italian Ministry of Agriculture, Food Sovereignty and Forestry, provided rice grains samples of 47 different rice varieties: 36 Italian, 18 from all over the world, in particular Sri Lanka, Malesia, Bangladesh, Philippines, Indonesia, Taiwan, Australia, Brazil, India, and Louisiana. All descriptive information about the samples is summarised in **Table 2.3**.

Table 2.3. Information about the rice varieties provided by Ente Nazionale Risi, with specifications concerning taxonomy, origin, shape of the rice grain and amylose content.

Variety	Taxonomy	Origin	Shape of the rice grain	Amylose content
Arborio	Oryza sativa (japonica)	Italy	Long A (Internal Market)	Low
Argo	Oryza sativa (japonica)	Italy	Medium	High
Baldo	Oryza sativa (japonica)	Italy	Long A (Internal Market)	Low
Carnaroli	Oryza sativa (japonica)	Italy	Long A (Internal Market)	High
Castelmochi	Oryza sativa (japonica)	Italy	Round	Waxy
CL12	Oryza sativa (japonica)	Italy	Round	Low
CL15	Oryza sativa (japonica)	Italy	Round	Low
CL18	Oryza sativa (japonica)	Italy	Round	Low
CL26	Oryza sativa (japonica)	Italy	Long B	High
CL28	Oryza sativa (japonica)	Italy	Long B	High
CL31	Oryza sativa (japonica)	Italy	Long A for parboil	Low
CL33	Oryza sativa (japonica)	Italy	Long A for parboil	Low
CL35	Oryza sativa (japonica)	Italy	Long A for parboil	Low
CL71	Oryza sativa (japonica)	Italy	Long B	High
CL80	Oryza sativa (japonica)	Italy	Long B	High
CL338	Oryza sativa (japonica)	Italy	Long A (Internal Market)	Low
CL510	Oryza sativa (japonica)	Italy	Long A (Internal Market)	Low
Cripto	Oryza sativa (japonica)	Italy	Round	High
CRLB1	Oryza sativa (japonica)	Italy	Long B	High
CRW3	Oryza sativa (japonica)	Italy	Round	Waxy
Dedalo	Oryza sativa (japonica)	Italy	Long B	High
Drago	Oryza sativa (japonica)	Italy	Long A for parboil	Low
Duilio	Oryza sativa (japonica)	Italy	Medium	Low
Elio	Oryza sativa (japonica)	Italy	Round	High
Europa	Oryza sativa (japonica)	Italy	Long A for parboil	Low
Iarim	Oryza sativa (japonica)	Italy	Long B	High
Italmochi	Oryza sativa (japonica)	Italy	Round	Waxy
Lince	Oryza sativa (japonica)	Italy	Long A for parboil	Low
Padano	Oryza sativa (japonica)	Italy	Medium	Low
Pegaso	Oryza sativa (japonica)	Italy	Long B	High
Prometeo	Oryza sativa (japonica)	Italy	Round	High
Puma	Oryza sativa (japonica)	Italy	Long A for parboil	Low
S. Andrea	Oryza sativa (japonica)	Italy	Long A (Internal Market)	Low
Selenio	Oryza sativa (japonica)	Italy	Round	Low
Tiberio	Oryza sativa (japonica)	Italy	Long A for parboil	High
Valente	Oryza sativa (japonica)	Italy	Long A for parboil	Low

Kaluheenati	Oryza sativa	Sri Lanka	Medium	Not defined
Hetadawee	Oryza sativa	Sri Lanka	Medium	Not defined
IR 6	Oryza sativa	Philippines	Long B	High
IR 50	Oryza sativa	Philippines	Long B	Not defined
IR 64	Oryza sativa	Philippines	Long B	High
Cisokan	Oryza sativa	Indonesia	Medium	High
Taichung sen 17	Oryza sativa	Taiwan	Medium	Not defined
Doongara	Oryza sativa	Australia	Long B	High
IAC 165	Oryza sativa	Brazil	Long A for parboil	High
Fedearroz 50	Oryza sativa	South America	Long B	High
Cypress	Oryza sativa	Louisiana	Long B	High

A total amount of 115 rice grains per variety was analysed. Prior to the acquisition of the hyperspectral images, it was necessary to remove the husk that covered the rice grains, a step which was done manually. Both before and after the images acquisition, the samples were stored in plastic bags at room temperature (~20 °C).

2.3.3.1. Images acquisition and pre-processing

All the images are collected in reflectance mode and pre-processed through an in-house written MATLAB routine. The first step after the acquisition is the conversion of the signal from reflectance to absorbance. Subsequently, the pixels were binned to obtain a square shape, as initially the vertical dimension was double than the horizontal one: sometimes, it is possible to obtain images where pixels are not squared, and one axis (generally the vertical one) is longer or shorter than the other. The reason is that the shape of the pixel is affected by the rate of the image acquisition: while the horizontal axis is fixed, the vertical one follows the movement of the scanning bed, so the lower the rate of acquisition, the longer the vertical dimension of the pixel, and vice versa. The binning can solve this criticality, shaping the pixels squared. Subsequently, the images were treated to remove the background: each hypercube was unfolded multiplying the two spatial dimensions, resulting in a 2D matrix where each row corresponds to one pixel and each column to the wavelength in the selected spectral interval. Then, a threshold of absorbance was chosen, below which all the values were turned to zeros. In this way, the noisy signals due to the absorbance of the background were removed from the images.

Afterwards, another in-house MATLAB function was developed to correct the optical distortion provoked by the round shape of the camera's lenses, which caused an elongation in the pixels at the extremities of the image, consequently altering the actual shape of the rice grains. This optical effect affects the computation of the morphological parameters and, if not corrected, unreal shapes could be processed and then fed to the multivariate analysis steps. As a consequence, the measured morphological parameters would not reflect the reality of the scanned samples introducing a bias. Once these steps were finished, the 2D matrix was folded back

to a 3D hypercube. This final matrix was composed of 115 rows, one representative of each rice grain.

2.3.3.2. Morphological parameters extraction

In-house MATLAB functions and scripts were exploited for the extraction of all the morphological information from the hypercubes. Initially, it was necessary to define which morphological parameters to compute and analyse; the chosen parameters and their description are resumed in **Table 2.4**. Before any multivariate analysis, the morphological data were pre-processed with Autoscaling.

Table 2.4. Morphological parameters used for the structural analysis of the samples.

Parameter	Description
Area	Measured as the number of pixels in a selected region of the image, returned as a scalar.
Circularity	A quantification of the roundness or similarity of a shape to a perfect circle. $\frac{4 \cdot \pi \cdot Area}{Perimeter^2} \cdot \left(1 - \frac{0.5}{r}\right)^2$, where $r = \frac{Perimeter}{2 \cdot \pi} + 0.5$
Eccentricity	A measurement of how elongated or stretched a shape is compared to a perfect circle. $\frac{Distance\ between\ the\ foci\ of\ the\ ellipse}{Major\ axis\ length}$
Major axis length	Length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the region, returned as a scalar.
Minor axis length	Length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the region, returned as a scalar.
Perimeter	Length of the boundary of the identified sample's region, returned as a scalar.

The morphograms

During this study, it was chosen to try to explore an original way to handle the morphological data, with the so-called "morphograms", which allow to inspect several morphological parameters of interest in datasets with a high number of samples. The morphogram of a rice variety is elaborated concatenating the histograms of the different morphological features, which are obtained using the 115 values of the rice grains analysed (otherwise, no distributions could be drawn from an averaged dataset). The first step for this elaboration involved writing an in-house MATLAB function to preliminary define the proper number of bins for this analysis, which was chosen to be 30. Further, as the values of the morphological features for each rice grain per image fall into their respective bins based on defined intervals, they are counted accordingly, resulting in a histogram in the end. As an example of this methodology, in **Figure 2.10** the morphogram of Dedalo (in red) and Elio (in blue) rice varieties are displayed. The comparison between these two varieties was chosen because their shape is poles apart: Dedalo belongs to the "Long B" shape category, while Elio is a round-shaped rice type.

The figure shows six concatenated histograms (from a to f), each one corresponding to one of the morphological parameters defined in **Table 2.4**.

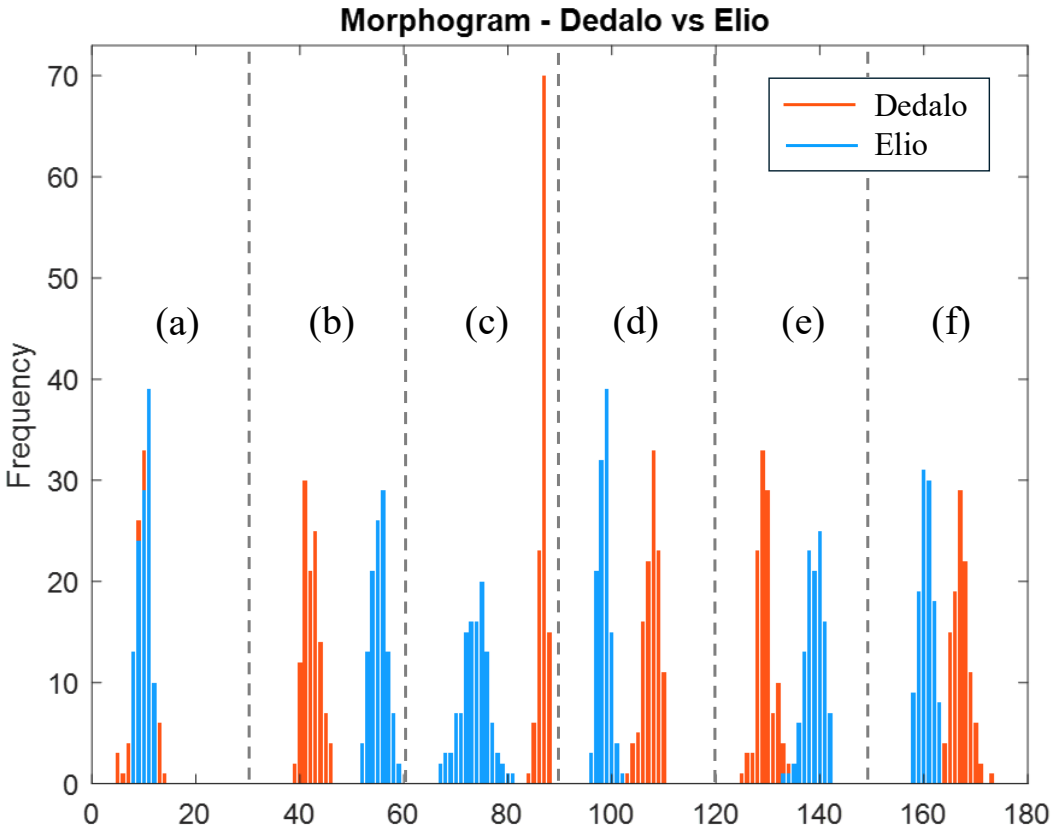

Figure 2.10. A morphogram example comparing the Dedalo (red) and Elio (blue) varieties. The distributions of the considered morphological parameters are sorted from the left to the right in the following order: (a) Area, (b) Circularity, (c) Eccentricity, (d) Major axis length, (e) Minor axis length, (f) Perimeter.

In this visual representation of the morphological parameters of the samples, the structural differences among them are highlighted: despite the values of the area of the rice grains are distributed very similarly for the two varieties, the other parameters underline significant differences in the shape of the kernels (i.e., for all the other parameters). The most different parameter between the two varieties under examination is Circularity (**Figure 2.10b**), for which the distributions of both varieties lie in completely separated intervals, and the Elio rice shows clearly higher values. Eccentricity (**Figure 2.10c**) is antithetical to circularity, and in fact Dedalo shows higher values as the shape of these rice type is more elongated. Similar considerations can be done for the other parameters.

Other than a visual representation of the morphological parameters, morphograms can represent an original way to handle the morphological data since, in principle, they carry more information than the simple set of averaged values: the latter consists of the mean value alone, while the morphogram describes the distributions, which encompass the mean value, the dispersion around the mean, but also possible deviations from the distribution around the mean; thus, a normal or at least centered distribution is not assumed, while with the simple mean value that number is taken as representative for that sample. This type of data was exploited to look for

structural similarities/differences among the inspected rice varieties, and the results for this inspection are shown in Section 5.2.1.1.

2.3.3.3. NIR spectra extraction and pre-processing

To extract the NIR data from the hypercubes, the MIA_Toolbox was exploited. In each image, each rice grain is described by several pixels, each one containing an NIR spectrum. In the present case, an average of the spectra (and thus, of the pixels) belonging to each individual grain was computed, so that for each grain a single representative spectrum was obtained. As an output, each rice variety was therefore represented by 115 NIR spectra (one for each grain), which were then pre-processed with Savitzky-Golay derivative (1^{st} derivative, 2^{nd} polynomial order and 11 pt window), normalization (2-Norm) and mean centering. The spectral range was cut to remove noisy areas at the extremities, obtaining a new range of 1100–1634 nm.

2.4. Software, toolboxes and in-house MATLAB routines

All data analyses reported in this Thesis were carried out under the MATLAB environment (2020a, The Mathworks Inc., Natick, USA).

All **multivariate data analyses** were performed using the PLS_Toolbox (v9.1, Eigenvector Research Inc. WA, USA).

For instrumental control and for spectra acquisition with the **benchtop Bruker MPA spectrometer**, the Opus software (v6.5, Bruker Optics, Ettlingen, Germany) was used.

For the **HSI instrumental settings and acquisition**, the LUMO (Specim, Spectral Imaging LTD, Oulu, Finland) software was used for the acquisitions of rice images. To acquire the images of the grapevine leaves, the SpectralScanner (DV S.r.l., Padua, Italy) software was employed.

NIR spectra extraction from the HSI images was performed using the MIA_Toolbox (v3.1, Eigenvector Research Inc. WA, USA).

Morphological parameters extraction from the HSI images was computed using an in-house MATLAB routine written by Rodrigo Rocha de Oliveira (University of Barcelona).

To create the **Common Space Hyperspectrograms (CSH)**, Hyper-Tools (v3.0, freely available at https://www.hypertools.org/, last visited in Jan 2025) was used.

References | Chapter 2

[1] Å. Rinnan, L. Nørgaard, F. Van Den Berg, J. Thygesen, R. Bro, S.B. Engelsen, Data Pre-processing, in: Elsevier (Ed.), 2009. https://doi.org/10.1016/B978-0-12-374136-3.00002-X.

[2] Å. Rinnan, Pre-processing in vibrational spectroscopy-when, why and how, Anal. Methods 6 (2014) 7124–7129. https://doi.org/10.1039/C3AY42270D.

[3] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC Trends Anal. Chem. 28 (2009) 1201–1222. https://doi.org/10.1016/J.TRAC.2009.07.007.

[4] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra:, Http://Dx.Doi.Org/10.1366/0003702894202201 43 (2016) 772–777. https://doi.org/10.1366/0003702894202201.

[5] A.S.C. Ehrenberg, J.W. Tukey, Exploratory Data Analysis., Appl. Stat. 28 (1979) 79. https://doi.org/10.2307/2346818.

[6] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831. https://doi.org/10.1039/c3ay41907j.

[7] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. 2 (2010) 433–459. https://doi.org/10.1002/wics.101.

[8] R. Garcia-Dias, S. Vieira, W.H. Lopez Pinaya, A. Mechelli, Clustering analysis, Mach. Learn. Methods Appl. to Brain Disord. (2020) 227–247. https://doi.org/10.1016/B978-0-12-815739-8.00013-4.

[9] R.K. Blashfield, The Growth Of Cluster Analysis: Tryon, Ward, And Johnson, Multivariate Behav. Res. 15 (1980) 439–458. https://doi.org/10.1207/S15327906MBR1504_4.

[10] H.-H. Bock, Clustering Methods: A History of k-Means Algorithms, (2007) 161–172. https://doi.org/10.1007/978-3-540-73560-1_15.

[11] M. Ahmed, R. Seraj, S.M.S. Islam, The k-means Algorithm: A Comprehensive Survey and Performance Evaluation, Electron. 2020, Vol. 9, Page 1295 9 (2020) 1295. https://doi.org/10.3390/ELECTRONICS9081295.

[12] K. Khan, S.U. Rehman, K. Aziz, S. Fong, S. Sarasvady, A. Vishwa, DBSCAN: Past, present and future, 5th Int. Conf. Appl. Digit. Inf. Web Technol. ICADIWT 2014 (2014) 232–238. https://doi.org/10.1109/ICADIWT.2014.6814687.

[13] L.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools, Technometrics 35 (1993) 109–135. https://doi.org/10.1080/00401706.1993.10485033.

[14] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

[15] L.E. Eberly, Multiple Linear Regression, Methods Mol. Biol. 404 (2007) 165–

187. https://doi.org/10.1007/978-1-59745-530-5_9.

[16] P. Xanthopoulos, P.M. Pardalos, T.B. Trafalis, Linear Discriminant Analysis, (2013) 27–33. https://doi.org/10.1007/978-1-4419-9878-1_4.

[17] H. Bhavsar, M.H. Panchal, A Review on Support Vector Machine for Data Classification, Int. J. Adv. Res. Comput. Eng. Technol. 1 (2012) 2278–1323.

[18] J. Sun, X. Lu, H. Mao, X. Jin, X. Wu, A Method for Rapid Identification of Rice Origin by Hyperspectral Imaging Technology, J. Food Process Eng. 40 (2017) e12297. https://doi.org/10.1111/JFPE.12297.

[19] M. Manfredi, E. Robotti, F. Quasso, E. Mazzucco, G. Calabrese, E. Marengo, Fast classification of hazelnut cultivars through portable infrared spectroscopy and chemometrics, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 189 (2018) 427–435. https://doi.org/10.1016/J.SAA.2017.08.050.

[20] L. Wang, D.W. Sun, H. Pu, J.H. Cheng, Quality analysis, classification, and authentication of liquid foods by near-infrared spectroscopy: A review of recent research developments, Crit. Rev. Food Sci. Nutr. 57 (2017) 1524–1538. https://doi.org/10.1080/10408398.2015.1115954.

[21] I.S. Arvanitoyannis, M.N. Katsota, E.P. Psarra, E.H. Soufleros, S. Kallithraka, Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics), Trends Food Sci. Technol. 10 (1999) 321–336. https://doi.org/10.1016/S0924-2244(99)00053-9.

[22] Y. Zhou, J. Hahn, M.S. Mannan, Process monitoring based on classification tree and discriminant analysis, Reliab. Eng. Syst. Saf. 91 (2006) 546–555. https://doi.org/10.1016/J.RESS.2005.03.019.

[23] A.D. Gordon, A Review of Hierarchical Classification, J. R. Stat. Soc. Ser. A 150 (1987) 119. https://doi.org/10.2307/2981629.

[24] C.N. Silla, A.A. Freitas, A survey of hierarchical classification across different application domains, Data Min. Knowl. Discov. 22 (2011) 31–72. https://doi.org/10.1007/S10618-010-0175-9.

[25] P. Arabie, L.J. Hubert, G. De Soete, Clustering and Classification, Clust. Classif. (1996). https://doi.org/10.1142/1930.

[26] K.B. Beć, J. Grabska, H.W. Siesler, C.W. Huck, Handheld near-infrared spectrometers: Where are we heading?, NIR News, Vol. 31, Issue 3-4 31 (2020) 28–35. https://doi.org/10.1177/0960336020916815.

[27] C.T. Srigley, M.M. Mossoba, Current Analytical Techniques for Food Lipids, Food Saf. Innov. Anal. Tools Saf. Assess. (2016) 33–64. https://doi.org/10.1002/9781119160588.CH3.

[28] AOAC, Official Method 948.22. Fat (crude) in nuts and nut products. Gravimetric methods, in: Official Methods of Analysis of AOAC International, 19th ed., AOAC International, Gaithersburg, MD, USA, 2012., (n.d.). http://www.aoacofficialmethod.org/index.php?main_page=product_info&products_id=588 (accessed October 7, 2022).

[29] M.M. Jan, N. Zainal, S. Jamaludin, Region of interest-based image retrieval techniques: A review, IAES Int. J. Artif. Intell. 9 (2020) 520–528.

https://doi.org/10.11591/IJAI.V9.I3.PP520-528.

[30] R. Calvini, G. Foca, A. Ulrici, Data dimensionality reduction and data fusion for fast characterization of green coffee samples using hyperspectral sensors, Anal. Bioanal. Chem. 408 (2016) 7351–7366. https://doi.org/10.1007/S00216-016-9713-7.

Chapter 3

NEWPOW: determination of lipid content in hazelnuts

In this chapter, the research conducted within the NEWPOW project is presented.

3.1 The project

Hazelnuts (*Corylus avellana L.*) are broadly cultivated and exported worldwide, with a production volume exceeding one billion tons in 2020 [1]. Turkey stands as the main producer, contributing approximately 70% of the world's production, followed by Italy at nearly 20% [2]. The success of hazelnuts as food, beyond the taste, is due to their high nutritional and nutraceutical properties, mainly coming from their fat content (approximately 60% in weight), mostly consisting in unsaturated fatty acids, particularly oleic, linoleic and linolenic acids) [3]. Moreover, hazelnuts are also rich in proteins, which provide near the 25% of their total energy content [4], as well as carbohydrates and dietary fibres, and are rich in essential micronutrients and non-nutritive valuable components, including vitamins (e.g., tocopherols, valuable antioxidant vitamins), healthy minerals (such as calcium, magnesium, and potassium), phytosterols, and polyphenols [5].

Quantitatively, the lipid fraction is the most abundant constituent of hazelnuts, and therefore, the most significant in terms of energy intake, organoleptic quality, and storability [4]. In some cases, a high lipid content is beneficial, for example in the field of hazelnut oil production. On the other hand, other circumstances like the excessive calorie intake, or the increase in susceptibility to lipid oxidation, could lead to an opposite choice. Therefore, total fat content becomes a relevant quality parameter for processors and manufacturers. In this context, the most used analytical techniques to determine and quantify the fatty acid composition, at laboratory scale, are generally time-consuming and involve expensive instrumentation, demanding also high scientific expertise. With respect to these observations, the application of NIR spectroscopy to the analysis of hazelnut samples was explored with the aim of quantifying the total lipids, as a potential alternative to the conventional Soxhlet extraction methods, which are expensive, time consuming, and involve hazardous and non-ecofriendly solvents.

3.2. PCA results and discussion

The results of the exploratory analysis on the NIR data obtained from the hazelnut samples (Section 2.3.1.1.) are shown in **Figure 3.1**: for each dataset, the most informative combinations of principal components are reported (**Figure 3.1a–c**) together with the respective loadings (**Figure 3.1d–f**), and the samples are coloured according to their lipid content. It is possible to observe a trend in the distribution of the samples: the scores follow the directions described by the red arrows, from lower to higher contents of lipids.

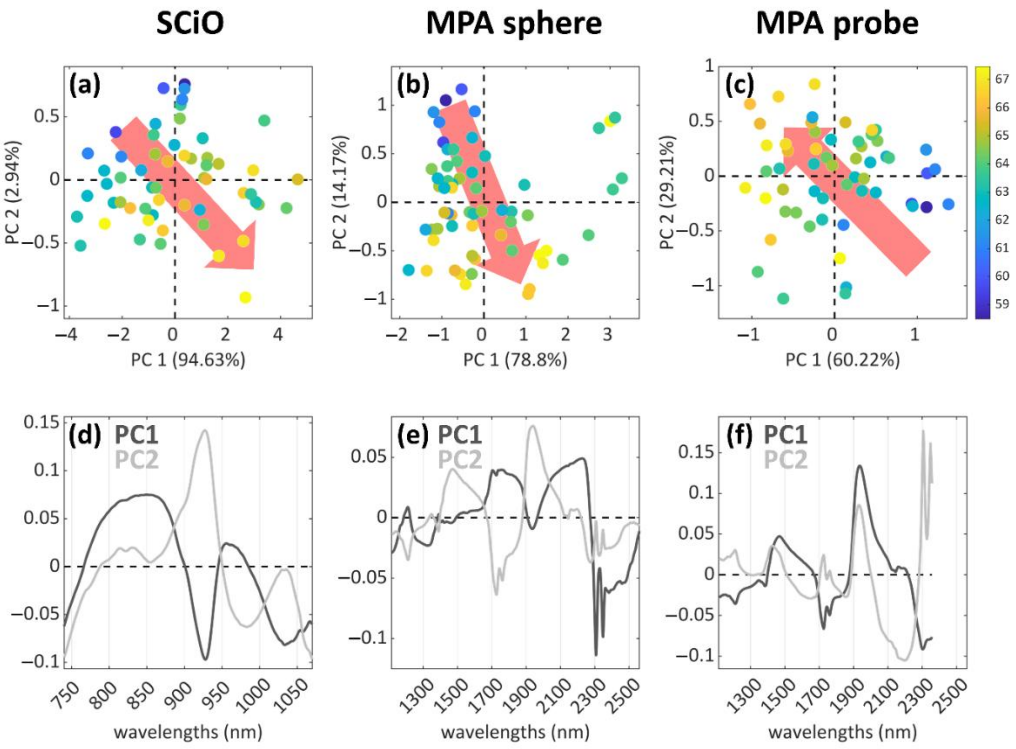

Figure 3.1. PCA results for the three analytical techniques of the study. The score plots are shown in the first row (a–c), with the samples coloured according to the lipid content and the content trends highlighted by the red arrows (from low to high). In the second row (d–f) the corresponding loadings plot are reported, showing the variables that are important for each PC reported in a–c.

These preliminary results suggested the presence of a trend in the spectral data related to the lipid content of hazelnuts, according to their chemical profile. Starting from these observations, it was decided to continue with a further modelling step, with the aim of predicting the fatty acids content directly from the spectral information, using Partial Least Square (PLS) regression.

3.3 Partial Least Square (PLS) Regression

PLS-R and PLS-DA are both children of the PLS family, and both work with LVs to lower the data dimensionality and to model *X* and *Y*, which is conceptually similar to the decomposition made by PCA. These decomposition methods aim at modelling relationship among variables, and this information can be used to further describe samples with unknown response value. The main difference between PCA and PLS is that the latter is a supervised method, so during the modelling phase, the bilinear decomposition is performed so that only the information contained in *X* that relates to the response variable *Y* is actually modelled. In the case of PLS-R, the *Y* response is quantitative: for example, it is possible to relate spectral data to the chemical features of the samples, measured with other/traditional reference techniques.

When a model is built and validated, it is possible to exploit its predictive power to obtain the chemical composition of a sample simply based on its spectrum. This is one of the main advantages of this technique with respect to the traditional analytical methods in chemistry, as it allows rapid and cost-effective analyses, avoiding wastes and use of solvents.

In this project, to build the PLS model, the spectral dataset was initially split into a training and a test set, with the test set consisting of 33% of the total number of samples. In this way, a division into 38 samples in calibration and 18 samples in the test set was obtained. This step was pursued using the Duplex algorithm [6], which allows homogeneous sampling of the initial and complete dataset. Subsequently, PLS was used for building the regression models to predict the lipid content of hazelnuts, expressed in percentage, as obtained with the Randall/Soxtec extraction method (Section 2.3.1). The regression performances were evaluated in terms of the coefficient of determination (R^2), the root mean squared error (RMSE), and the Ratio of Prediction to Deviation (RPD, [7,8]).

3.3.1. PLS regression results and discussion

Starting from the preprocessed spectral dataset obtained from the three spectroscopic techniques (i.e., MPA sphere, MPA probe, SCiO), one PLS regression model was built for each one of them, with model-specific training and test sets for model building and validation. The regression parameters are represented in **Table 3.1**, reporting the models' complexities (number of latent variables, LVs), the coefficients of determination (R^2), the root mean squared errors (RMSEs), and the RPD value. Both calibration and cross-validation (CV) values are reported in this table: the parameters referring to CV are generally of major interest, as the CV procedure better elucidates the robustness of the information contained in the data, and consequently, the evaluation of the model performances is more reliable. A visual representation of the predicted lipid content values, expressed in percentage, plotted against the measured ones can be found in **Figures 3.2a–c**, while in **Figure 3.2d–f** the corresponding regression vectors are reported.

According to these results, the regression model built with the MPA sphere data showed the best performances for both calibration and validation, with a value of R^2_{CV} = 0.903, which was higher than those obtained with MPA probe and SCiO. Concerning the model's error (the RMSEs), the MPA sphere in validation also showed the lowest value, with $RMSE_{CV}$ = 0.645, confirming that MPA sphere models seemed to be more robust than the others. Despite this, even better results were achieved with the MPA probe when considering the prediction step, with R^2_{PRED} = 0.897 and $RMSE_{PRED}$ = 0.712: these results suggest that this technique could be performing better in correctly predicting the lipid content of hazelnuts when compared to the other NIR techniques used for the same purpose.

Table 3.1. Regression parameters from PLS models for the three NIR techniques. The three models have different complexities (number of latent variables, LVs), and the parameters used for describing the performances are the coefficient of determination (R2) and the root mean squared errors (RMSEs). The subscripts stand for CAL = calibration, CV = cross validation, and PRED = prediction (test set).

	SCiO	**MPA Sphere**	**MPA Probe**
LVs	7	4	3
R^2_{CAL}	0.717	0.925	0.846
R^2_{CV}	0.461	0.903	0.793
R^2_{PRED}	0.550	0.713	0.897
$RMSE_{CAL}$	0.966	0.566	0.870
$RMSE_{CV}$	1.332	0.645	1.008
$RMSE_{PRED}$	1.217	1.109	0.712
RPD_{CAL}	1.904	3.709	2.583
RPD_{CV}	1.380	3.254	2.229
RPD_{PRED}	1.909	1.773	2.167

Globally, the results achieved with the SCiO proved to be worse than those obtained with the benchtop MPA instrument, especially with respect to the values of the coefficients of determination, with R^2_{CV} = 0.461 and R^2_{PRED} = 0.550, which resulted in much worse values than those obtained with the MPA instrument. The reason of this behaviour can be ascribed to the restricted spectral range of acquisition of this portable instrument, which can represent a limit if compared to a benchtop spectrometer. In fact, the SCiO spectral absorption range only contains the last overtones and combination bands of the NIR interval, while the larger acquisition range of MPA can detect more signals than SCiO. Moreover, the reduced number of samples could also represent a limitation to the performances of this instrument. Despite this, the results obtained with the SCiO were interesting in the perspective of in situ analyses: even if this technique is less accurate, it could lead to reliable models if the number of samples employed to build the calibration model is increased.

From the point of view of the RPD values, the best model, even if failing in prediction, proved to be the MPA sphere acquisition mode, with values above 3 for

the calibration and CV models. These results put the MPA sphere model in the "screening/quality control" RPD bracket according to Williams [8], even if the prediction RPD value of this model is very poor according to the same RPD scheme. The MPA probe model can be placed into the "rough screening" RPD bracket, and this is consistent with the interpretation of the other regression parameters, as previously discussed: if compared to the MPA sphere acquisition model, the model shows lower performances; therefore, the suggested applicability by Williams is also less trustworthy (screening/quality control vs. rough screening). The SCiO model once again proved to be less reliable, showing very poor performances from the RPD point of view, especially when considering the CV figure, which places this model into the Williams bracket of "not recommended" to be used. It is important to consider that the aim of the study was to assess the performances of the different instruments and acquisition modes, and the rather limited number of samples already showed some limitations on the performances of the models when inspected from the point of view of the traditional chemometric regression figures, i.e., coefficient of determination and RMSE.

To identify and interpret the most influent NIR signals, the regression vectors of each model, depicted in **Figure 3.2d–f**, were inspected. Lipids in hazelnuts represent about 60% of the total amount of components, so it is consistent to think that the signals of the regression coefficients can be mostly related to the lipids absorption. The information found in literature concerning the determination of fatty acids in hazelnuts shows that the main fatty acids analysed in this substrate are palmitic, stearic, oleic, and linoleic acids, with a percentage of unsaturated acids above 90% [9].

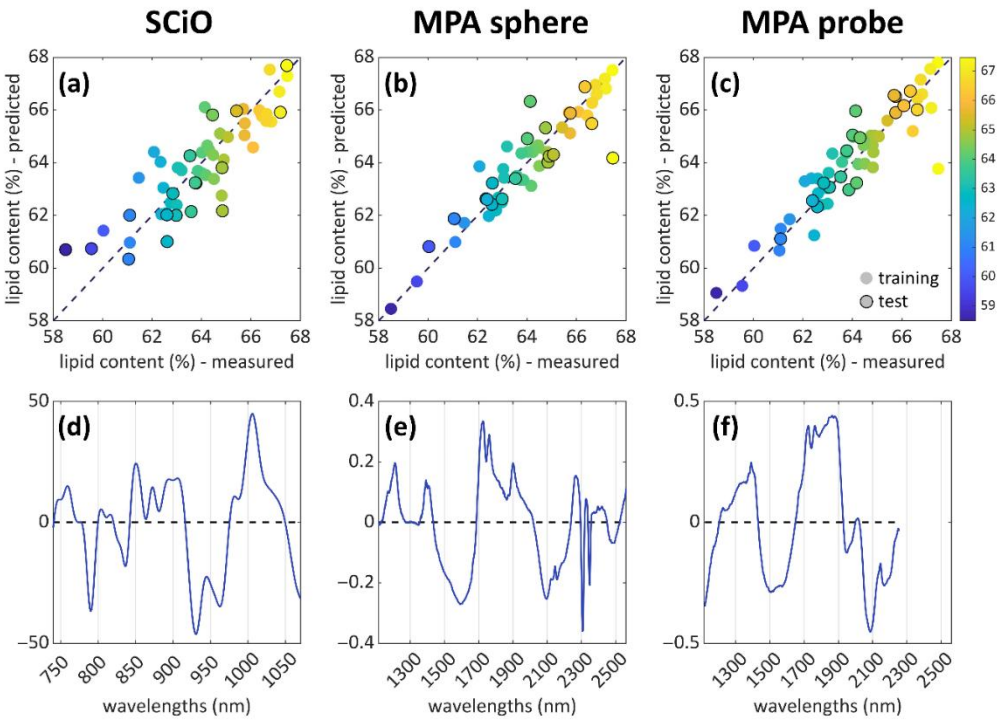

Figure 3.2. PLS results for the three analytical techniques of the study. The prediction plots are shown in the first row (a–c), with the samples coloured according to the lipid content. The black-bordered markers correspond to the test set (predicted samples). In the second row (d–f), the corresponding regression vectors are reported, showing the variables that are important for each PLS model.

The interpretation of the regression coefficients is reported in **Table 3.2**. Aliphatic peaks are located in the region between 1700–1900 nm, and they are related to the first overtone, C–H, stretching [10,11]. These peaks fall within the MPA spectral range, as the SCiO only covers the wavelengths between 740 and 1070 nm, so it is not possible to detect these vibrations with this portable device. In particular, the signals referred to C–H, stretching are the two peaks in the range of 1730–1760 nm, which appear well defined for the MPA sphere, while the MPA probe shows a partial overlap with the band at ~1900 nm, which is ascribable to O–H deformation and stretching combination bands. Another strong signal which is possible to detect is two narrow double bands at 2300–2340 nm, which could be related to C–H stretching and deformation combinations of CH_2 of lipids. As the largest part of fatty acids in hazelnuts is composed by unsaturated fatty acids, a strong band at 2100 nm, related to the C=C stretching, can be noticed. Other intense signals are the C–H second overtone at 1214 nm of CH_2, and CH_2 stretching at 1390 nm.

Table 3.2. Interpretation of spectroscopic signals related to the fatty acids, with literature references.

	SCiO	MPA (Sphere and Probe)
C–H [9,10]	920 nm, stretching third overtones	1214 nm, second overtone CH_2 1390 nm, stretching CH_2 1730–1760 nm, first overtone 2300–2340 nm, stretching and deformation combinations CH_2
C=C [9,10]	/	2100 nm, stretching
O–H [11]	960–1050 nm, stretching second overtones	1900 nm, combination band
Unassigned	780 nm, 830–850 nm	1600 nm

Within the SCiO spectral range, the vibration associated with the third overtone of the C–H stretch of lipids at ~920 nm [10] is observed. Other strong signals detected by this instrument can be referred to with O–H stretching second overtones, in the range between 960–1050 nm [11]. This device can detect the signals related to the absorption of fatty acids, in accordance with the aim of this study. This information was used in the construction of the model, even if the regression parameters underline that the obtained model is not robust enough to be used for reliable predictions. As already stated above, this issue might be solved by considering a larger number of samples to be included in the model.

References | Chapter 3

[1] I. Oliveira, A. Sousa, J.S. Morais, I.C.F.R. Ferreira, A. Bento, L. Estevinho, J.A. Pereira, Chemical composition, and antioxidant and antimicrobial activities of three hazelnut (Corylus avellana L.) cultivars, Food Chem. Toxicol. 46 (2008) 1801–1807. https://doi.org/10.1016/J.FCT.2008.01.026.

[2] F. Tüfekci, Ş. Karataş, Determination of geographical origin Turkish hazelnuts according to fatty acid composition, Food Sci. Nutr. 6 (2018) 557–562. https://doi.org/10.1002/FSN3.595.

[3] A.I. Köksal, N. Artik, A. Şimşek, N. Güneş, Nutrient composition of hazelnut (Corylus avellana L.) varieties cultivated in Turkey, Food Chem. 99 (2006) 509–515. https://doi.org/10.1016/J.FOODCHEM.2005.08.013.

[4] E. Ros, Health benefits of nut consumption, Nutrients 2 (2010) 652–682. https://doi.org/10.3390/NU2070652.

[5] H.-G. Jang, B.-G. Heo, Y.S. Park, J. Namiesnik, D. Barasch, E. Katrich, K. Vearasilp, S. Trakhtenberg, S. Gorinstein, Chemical Composition, Antioxidant and Anticancer Effects of the Seeds and Leaves of Indigo (Polygonum tinctorium Ait.) Plant, Appl. Biochem. Biotechnol. 2012 1677 167 (2012) 1986–2004. https://doi.org/10.1007/S12010-012-9723-7.

[6] R.D. Snee, Validation of Regression Models: Methods and Examples, Technometrics 19 (1977) 415–428. https://doi.org/10.1080/00401706.1977.10489581.

[7] T. Fearn, Assessing Calibrations: SEP, RPD, RER and R2, Http://Dx.Doi.Org/10.1255/Nirn.689 13 (2017) 12–13. https://doi.org/10.1255/NIRN.689.

[8] P. Williams, The RPD Statistic: A Tutorial Note, Http://Dx.Doi.Org/10.1255/Nirn.1419 25 (2010) 22–26. https://doi.org/10.1255/NIRN.1419.

[9] F. Davrieux, F.-O. Allal, G. Piombo, B. Kelly, J.B. Okulo, M. Thiam, O.B. Diallo, J.-M. Bouvet, Near Infrared Spectroscopy for High-Throughput Characterization of Shea Tree (Vitellaria paradoxa) Nut Fat Profiles, J. Agric. Food Chem 58 (2010) 7811. https://doi.org/10.1021/jf100409v.

[10] B.W. Mulvey, Determination of Fat Content in Foods Using a Near-Infrared Spectroscopy Sensor, Proc. IEEE Sensors 2020-October (2020). https://doi.org/10.1109/SENSORS47125.2020.9278647.

[11] V. Wiedemair, D. Langore, R. Garsleitner, K. Dillinger, C. Huck, Investigations into the Performance of a Novel Pocket-Sized Near-Infrared Spectrometer for Cheese Analysis, Molecules 24 (2019). https://doi.org/10.3390/molecules24030428.

Chapter 4

Help2Grow: analysis of grapevine leaves

In the following sections, the analysis of grape leaves carried out in collaboration with the University of Turin and the University of Modena and Reggio Emilia will be discussed.

4.1 The project

Vitis vinifera, also known as common grapevine, is the most famous flowering plant in the family of Vitaceae. The story that involves the cultivation and domestication of this plant dates back to antiquity, as it is supposed that this practice was born some millennia before Christ, in a geographical area situated between present-day Iran and the Black Sea [1–3]. Nowadays, the main interest towards the cultivation of this plant is linked to the production of wine [4,5], but due to the particular sensitivity of the grapevine to the environmental conditions, only specific areas of the planet are suitable for this practice. These territories lie in the so-called "wine belt", and are Europe, Asia, New Zealand, and some parts of Africa, America and Australia. The main European wine producers are Italy, France and Spain: in 2022, the European union produced 16.1 billion litres of wine, with Italy and Spain contributing with 5 bn litres each, and France with 3.4 bn litres (EUROSTAT, [6]). Although this amount seems to be considerable, the wine production is facing a gradual decline year by year, due, among other reasons such as economic and sociologic matter, to the troubling phenomenon of climate change: the extreme climatic conditions impact significantly the global wine production [7–9], requiring continuous adaptation of cultivation practices and wine-making techniques to ensure the survival and competitiveness of this sector.

Another criticality linked to the cultivation of grapevine and the wine production involves the development of diseases caused by pathogens. These microorganisms are generally viruses, bacteria and fungi and can infect the plant causing diseases development both in pre- and post-harvest phases [10]. Regarding the fungal diseases, grapevine is particularly vulnerable to downy mildew, grey mold, and powdery mildew [11]. The issue of pathogen assault leads to the necessity of relying

on pesticides, with harmful consequences for the environment and, potentially, also the human health [12–14]. Due to the substantial risks associated with the use of pesticides, many alternative strategies have been under development in the agricultural field, with the aim of reaching a higher level of safety and sustainability [15–17].

Nowadays, several methods are applied to monitor the growth of the vineyard, mainly involving the use of remote sensing [18,19], or the application of sensors on the field [20]. The main disadvantages of these technologies are the high costs and the necessity of periodical maintenance to ensure the efficiency through time, as well as the need of trained personnel for its use and the interpretation of collected data. Moreover, these methods deeply depend on the atmospheric conditions for the acquisition of data.

In the attempt to overcome these criticalities, the Help2Grow project proposed an alternative analytical method based on acquisitions with NIR spectroscopy coupled with chemometrics. The aim was to evaluate this method as a rapid, cost-effective alternative in the monitoring of the growth of grapevine plants located at the Department of Agricultural, Forest and Food Science of the University of Turin. These plants were processed with different treatments to preserve their integrity from pathogens, and with this study it was tried to test the efficacy of these treatments and their combination in the obstruction of fungal growth.

The acquisitions were performed firstly with the SCiO portable spectrometer (more details about this instruments in Section 2.2.1.2), which allowed to collect the spectra *in-situ* during the investigation days throughout the plant productive season; then, during the last investigation, some leaves were taken, brought to the laboratories of Polytechnic of Turin and frozen, to be subsequently transported to the University of Modena and Reggio Emilia for the analyses with HSI.

4.2 PCA results and discussion

A first PCA model was calculated considering all the acquisitions carried out on the field, from the 6th of June to the 26th of October 2022, and the results are shown in **Figure 4.1**. In this figure, samples are coloured in blue or in red, according to the disease developed on the leaf, with a darkening shade representing the evolution of the acquisition time. The shape of the scores represents the type of treatment the plants underwent, which involved treatment with water, fungicide, activated water, half-dosage fungicide and no treatment.

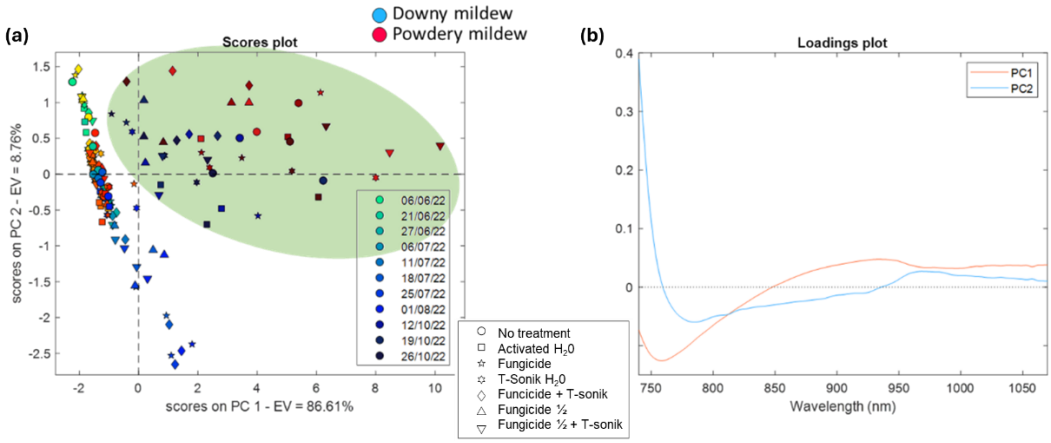

Figure 4.1: PCA results for the analysis of grape leaves with SCiO, from the 6th of June to the 26th of October. The blue scores refer to the samples affected by downy mildew, while the red ones refer to the samples affected by powdery mildew.

What emerged from this analysis was that the samples acquired in October (in the green area) are well distinguished from the others. The leaves analysed in October appeared dry and ruined, so these results proved that this analytical method can detect structural differences in the samples linked to the seasonal deterioration of the leaves. No trends in the samples were observed concerning the type of treatment carried out.

Subsequently, it was chosen to separately inspect the data acquired in the summer surveys, as the leaves were not deteriorated by the progress of the Fall season. **Figure 4.2** displays the results of PCA with the data collected from the 6th of June

to the 1st of August.

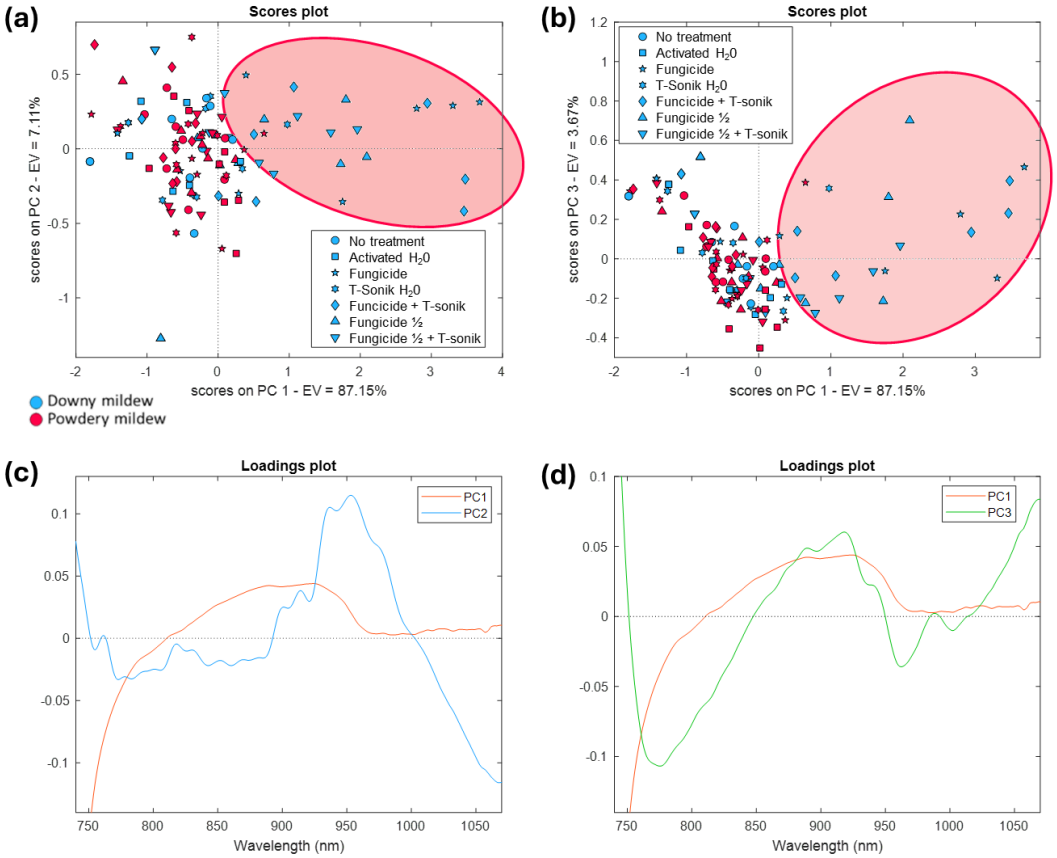

Figure 4.2. PCA scores (a-b) and loadings (c-d) for the analysis of grape leaves with SCiO, from the 6th of June to the 1st of August. The blue objects in the Scores plots refer to the samples affected by downy mildew, while the red ones refer to the samples affected by powdery mildew.

The scores plot in **Figure 4.2a** and **4.2b** highlight a peculiar behaviour for the sample of grape leaves affected by downy mildew: on the right side of PC1 only lie the samples exposed to this disease and treated with fungicide (in the red circle). A viable assumption is that these treatments had a positive impact in contrasting the development of the disease, generating changes detectable with spectroscopy in the leaves which are different from the other samples submitted to treatments with just water, or that were not treated at all, and that could consequently have started to develop the disease. Concerning the loadings, the positive peak around 950nm in PC2, which becomes negative in PC3, (**Figure 4.2c** and **d**) can be ascribable to the absorption of water [21], in particular the third overtone of the O−H stretching. The PC3 shows also a negative peak between 760 and 790 nm, probably due to the absorption of chlorophyll [22]. This observation might suggest that the differentiation among the leaves under examination is influenced by the content of these two chemical components, which is consistent with the fact that the leaves analysed in October were visibly less green if compared to those analysed during Summer, and their texture was drier.

A deeper investigation of this trend was conducted, to see whether there was any correlation with the evolution of the acquisition time or not. As no trends were detected in samples affected by powdery mildew, a PCA model was calculated considering only the samples exposed to downy mildew. The results of this analysis are displayed in **Figure 4.3**.

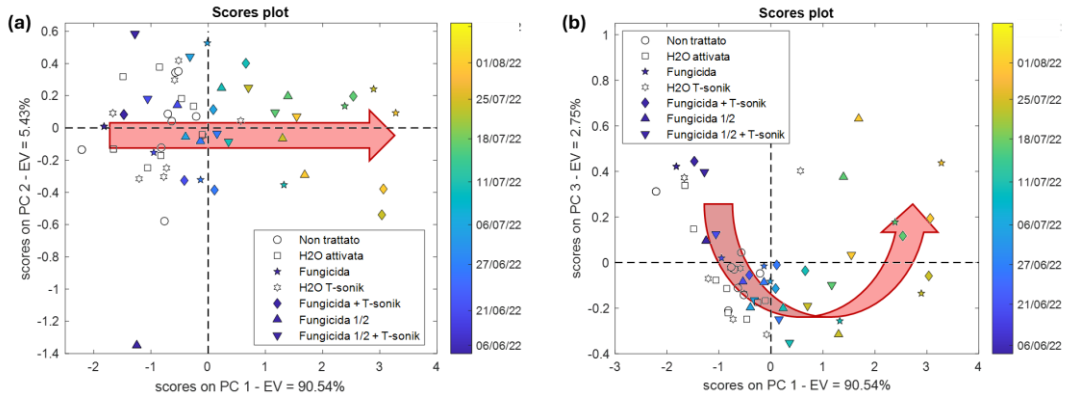

Figure 4.3. scores plot of PC1 vs PC2 (a) and PC1 vs PC3 (b) of samples exposed to downy mildew. Samples are coloured according to the time development of the *in-situ* acquisitions.

As a result, it was noticed that the acquisition time had an actual impact on this trend: the samples distributed along the PCs from the first acquisition date to the last, following the directions indicated by the arrows. If the sample distribution is linked to the effectiveness of the treatment on the plant, as hypothesized, this disposition can suggest that the fungicide can help protecting the leaves from the attack of the pathogen, maintaining a healthy status, and consequently distancing the samples from the others that, on the contrary, are probably developing the disease.

As mentioned above, no trends were found in samples of leaves exposed to powdery mildew. This is probably a consequence of the particularly hot and dry Summer of that year, which hampered the growth of the fungus responsible for the development of this disease.

Analyses on non-averaged data were conducted at a later stage: firstly, spectra collected in different positions on the same leaf were examined to inspect intra-leaf variability, but no relevant information were observed. Subsequently, the spectra of leaves from different plants subjected to the same treatment were analysed. Although the replicates were located at difference heights within the vineyard, no significant insights were found in this analysis, indicating that the spectral response of these samples is not influenced by the position of the plants, and that the variability among these samples is not significant.

4.3. Results of the acquisition on defrosted leaves

The description concerning the setup for the analyses at the University of Modena and Reggio Emilia are described in detail in Section 2.3.2. **Figure 4.4** below shows an example of the analysed samples, with the white dots on the back side of the leaves ascribable to the development of downy mildew.

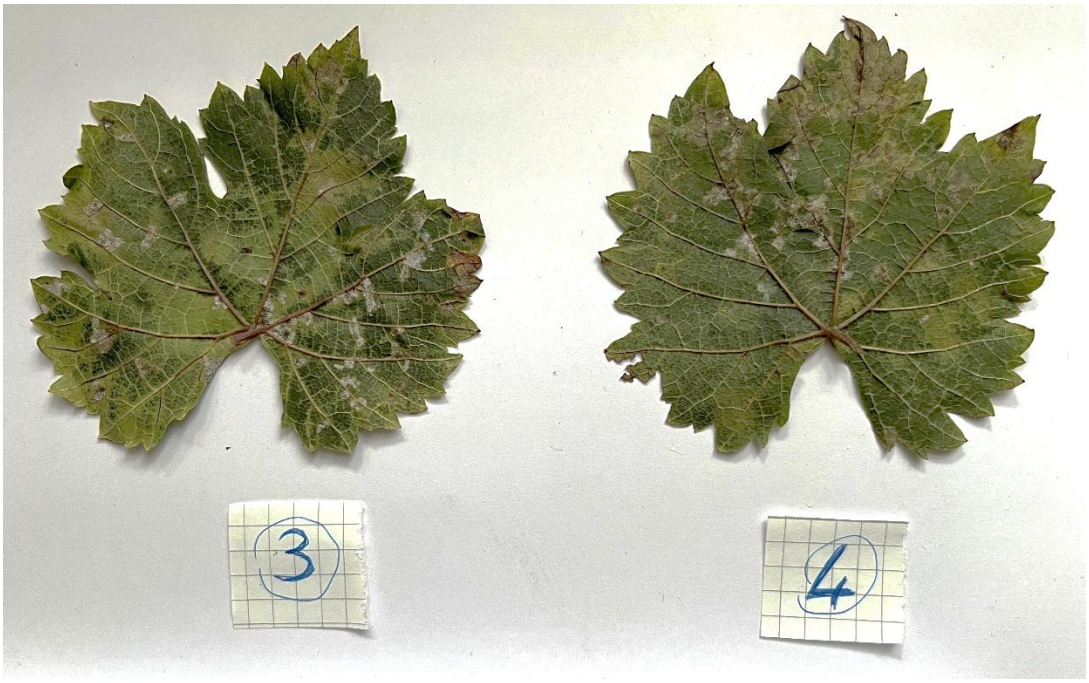

Figure 4.4. this image shows the third and fourth leaves representing the sample affected by downy mildew and treated with activated water.

The data obtained from SCiO and HSI were separately inspected through PCA to see if it was possible to extract some information from the data, concerning the type of treatment on the leaves or the development of the inoculated diseases. Unfortunately, neither for the SCiO nor for the HSI, interesting results were obtained, as shown in **Figure 4.5** and **4.6** respectively, where the results concerning the treatment are displayed. In these images it is possible to see that no significant trends within the samples are considerable, and leaves belonging to the same treatment were not even correlated among them, lying apart one from the others.

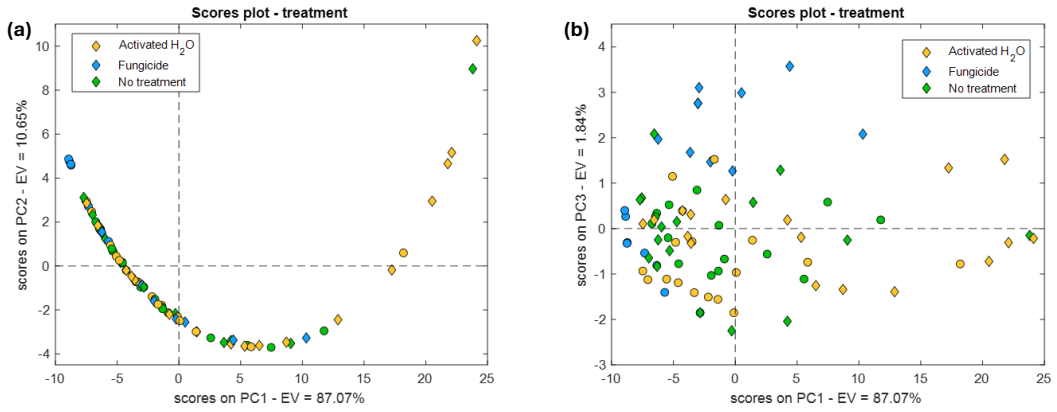

Figure 4.5. PCA results from SCiO analysis on the grapevine leaves. The circles refer to samples affected by powdery mildew, while the diamonds to the samples affected by downy mildew. The curvilinear behaviour in the scores plot of PC1 vs PC2 suggests a possible correlation among the variables of the NIR data.

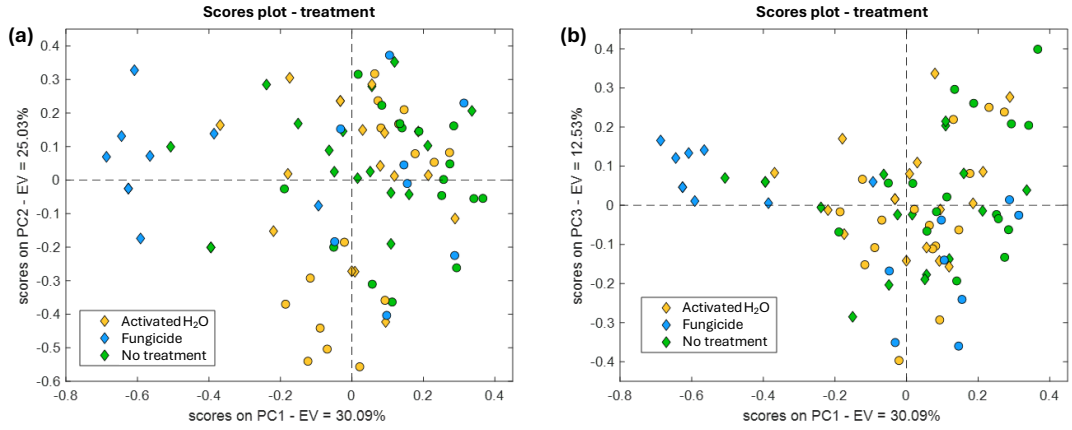

Figure 4.6. PCA results from HSI analysis on the grapevine leaves. The circles refer to samples affected by powdery mildew, while the diamonds to the samples affected by downy mildew.

The few clusters visible in **Figure 4.6** (for example, the six diamond-shaped blue scores on the left side of the plot, and the six yellow circles on the central bottom part of the plot) refer to three of the four leaves analysed per treatment, i.e., they belong to the same sample under examination. For this reason, these clusters cannot be considered as informative for the purposes of this analysis.

Given the aim of the present study, the results obtained with PCA were considered appropriate for gaining insight into the data, providing results that confirmed what observed during the *in situ* acquisitions concerning the development of the plants diseases. Therefore, the application of more advanced analytical methods in addition to PCA was not pursued. Particularly, it was assessed that employing more complex multilinear models such as PARAFAC would not have offered substantial benefits in terms of interpretability or modelling.

Many criticalities affected the fulfilment of this part of the project, and possibly the long storage time caused the deterioration of the samples and the loss of information that was thought to be possible to inspect. Another attempt of analysis in this context should be considered, hoping first in favourable seasonal conditions that promote the growth of the fungi responsible for the development of downy and powdery mildew. Regrettably, it was not possible to perform the analyses in faster times, and this issue clearly hampered the success of the analysis of the frozen leaves. An appropriate weather and immediate analyses with HSI would therefore surely be beneficial in providing better results for the development of this study in the future.

References | Chapter 4

[1] P.E. Mc Govern, D.L. Glusker, L.J. Exner, P.E. Mc Govern, Neolithic resinated wine, Nat. 1996 3816582 381 (1996) 480–481. https://doi.org/10.1038/381480a0.

[2] E. Patrick, S. James, H. Solomon, The Origin and Domestication of the Vinifera Grape, (2003) 29–43. https://doi.org/10.4324/9780203392836-5.

[3] J.F. Terral, E. Tabard, L. Bouby, S. Ivorra, T. Pastor, I. Figueiral, S. Picq, J.B. Chevance, C. Jung, L. Fabre, C. Tardy, M. Compan, R. Bacilieri, T. Lacombe, P. This, Evolution and history of grapevine (Vitis vinifera) under domestication: new morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient European cultivars, Ann. Bot. 105 (2010) 443–455. https://doi.org/10.1093/AOB/MCP298.

[4] E. Vaudour, A.B. Shaw, A Worldwide Perspective on Viticultural Zoning, South African J. Enol. Vitic. 26 (2005) 106–115. https://doi.org/10.21548/26-2-2125.

[5] L.F. Bisson, A.L. Waterhouse, S.E. Ebeler, M.A. Walker, J.T. Lapsley, The present and future of the international wine industry, Nat. 2002 4186898 418 (2002) 696–699. https://doi.org/10.1038/nature01018.

[6] EUROSTAT, Wine production reached 16.1 billion litres in 2022, (2023). https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20231116-1 (accessed August 9, 2024).

[7] C. van Leeuwen, G. Sgubin, B. Bois, N. Ollat, D. Swingedouw, S. Zito, G.A. Gambetta, Climate change impacts and adaptations of wine production, Nat. Rev. Earth Environ. 2024 54 5 (2024) 258–275. https://doi.org/10.1038/s43017-024-00521-5.

[8] R. Mira de Orduña, Climate change associated effects on grape and wine quality and production, Food Res. Int. 43 (2010) 1844–1855. https://doi.org/10.1016/J.FOODRES.2010.05.001.

[9] M.R. Mozell, L. Thachn, The impact of climate change on the global wine industry: Challenges & solutions, Wine Econ. Policy 3 (2014) 81–89. https://doi.org/10.1016/J.WEP.2014.08.001.

[10] G. Armijo, R. Schlechter, M. Agurto, D. Muñoz, C. Nuñez, P. Arce-Johnson, Grapevine pathogenic microorganisms: Understanding infection strategies and host response scenarios, Front. Plant Sci. 7 (2016) 183585. https://doi.org/10.3389/FPLS.2016.00382.

[11] C. Pirrello, C. Mizzotti, T.C. Tomazetti, M. Colombo, P. Bettinelli, D. Prodorutti, E. Peressotti, L. Zulini, M. Stefanini, G. Angeli, S. Masiero, L.J. Welter, L. Hausmann, S. Vezzulli, Emergent Ascomycetes in Viticulture: An Interdisciplinary Overview, Front. Plant Sci. 10 (2019) 478340. https://doi.org/10.3389/FPLS.2019.01394.

[12] K.H. Kim, E. Kabir, S.A. Jahan, Exposure to pesticides and the associated human health effects, Sci. Total Environ. 575 (2017) 525–535. https://doi.org/10.1016/J.SCITOTENV.2016.09.009.

[13] E.T. Rodrigues, M.F. Alpendurada, F. Ramos, M.Â. Pardal, Environmental and human health risk indicators for agricultural pesticides in estuaries, Ecotoxicol. Environ. Saf. 150 (2018) 224–231. https://doi.org/10.1016/J.ECOENV.2017.12.047.

[14] P. Nicolopoulou-Stamati, S. Maipas, C. Kotampasi, P. Stamatis, L. Hens, Chemical Pesticides and Human Health: The Urgent Need for a New Concept in Agriculture, Front. Public Heal. 4 (2016) 178764. https://doi.org/10.3389/FPUBH.2016.00148.

[15] E. Borsato, M. Zucchinelli, D. D'Ammaro, E. Giubilato, A. Zabeo, P. Criscione, L. Pizzol, Y. Cohen, P. Tarolli, L. Lamastra, F. Marinello, Use of multiple indicators to compare sustainability performance of organic vs conventional vineyard management, Sci. Total Environ. 711 (2020) 135081. https://doi.org/10.1016/J.SCITOTENV.2019.135081.

[16] C. Pirrello, G. Magon, F. Palumbo, S. Farinati, M. Lucchin, G. Barcaccia, A. Vannozzi, Past, present, and future of genetic strategies to control tolerance to the main fungal and oomycete pathogens of grapevine, J. Exp. Bot. 74 (2023) 1309–1330. https://doi.org/10.1093/JXB/ERAC487.

[17] M. Volanti, C. Cubillas Martínez, D. Cespi, E. Lopez-Baeza, I. Vassura, F. Passarini, Environmental sustainability assessment of organic vineyard practices from a life cycle perspective, Int. J. Environ. Sci. Technol. 19 (2022) 4645–4658. https://doi.org/10.1007/S13762-021-03688-2.

[18] L. Pádua, P. Marques, J. Hruška, T. Adão, E. Peres, R. Morais, J.J. Sousa, Multi-Temporal Vineyard Monitoring through UAV-Based RGB Imagery, Remote Sens. 2018, Vol. 10, Page 1907 10 (2018) 1907. https://doi.org/10.3390/RS10121907.

[19] Y. Barnard, The use of technology to improve current precision viticulture practices: predicting vineyard performance, (2018). http://hdl.handle.net/10019.1/104893 (accessed October 9, 2024).

[20] J.M. Terrón, J. Blanco, F.J. Moral, L.A. Mancha, D. Uriarte, J.R. Marques Da Silva, Evaluation of vineyard growth under four irrigation regimes using vegetation and soil on-the-go sensors, SOIL 1 (2015) 459–473. https://doi.org/10.5194/SOIL-1-459-2015.

[21] Q. Zhang, Q. Li, G. Zhang, Rapid determination of leaf water content using VIS/NIR spectroscopy analysis with wavelength selection, Spectrosc. (New York) 27 (2012) 93–105. https://doi.org/10.1155/2012/276795.

[22] A.A. Gitelson, O.B. Chivkunova, M.N. Merzlyak, Nondestructive estimation of anthocyanins and chlorophylls in anthocyanic leaves, Am. J. Bot. 96 (2009) 1861–1868. https://doi.org/10.3732/AJB.0800395.

Chapter 5

The Rice HSI project

The following sections describe the project carried out in collaboration with the Italian "Ente Nazionale Risi" and the University of Barcelona in the analysis of the rice samples through Hyperspectral Imaging.

5.1. The project

Rice stands as the most consumed cereal on a global level, within the big family of cereal products suitable for human sustenance. Statistics from the Food and Agricultural Organisation (FAO) reveal that the 90% of the global production of rice comes from Asia, with China, India, and Indonesia as main producers. Concerning the situation in Europe, Italy is the most representative producer (53% of the European production, [1]), with Piedmont and Lombardy regions which collectively contribute to the 93.2% of the national production [2].

The high nutritional value of rice makes it a really appreciated food, thanks to the considerable content of carbohydrates (around 80% [3]), which represents a significant source of energy for the human body. Moreover, rice is also rich in vitamins and minerals, as well as antioxidants such as flavonoids, anthocyanins, tocopherols, etc. [4]. Rice exhibits very rich genetic diversity: thousands of different rice varieties exist [5], although the most famous and farmed are the *Oryza sativa* and the *Oryza glaberrima*, mostly grown respectively in Asia and Africa. The first way to distinguish different rice varieties consists in the evaluation of their structural parameters, such as the grain size, which can be short, medium, long and round. Size and shape, alongside colour and chalkiness, are other qualitative attributes that describe the morphology of the rice kernel, while chemical indicators of rice quality are moisture, and the content of proteins, sugars and lipids [6].

The main issue concerning the rice industry involves the problem of food fraud. As it is normally difficult to distinguish among rice types which show similar structural features but are different from a qualitative point of view (for example, different amylose content that leads to different cooking modalities and culinary applications), it is easy to sell counterfeit rice, which is a fraudulent practice that happens when a rice variety of lower quality is sold as a premium product. For example, Basmati rice, which is famous for its peculiar flavour and aroma, is often

adulterated with cheaper varieties that do not present any aromatic characteristic [7], but still look very similar.

Several strategies have been undertaken in the agri-food field with the aim of improving the analytical methods and the detection of food frauds, and image analysis has proven to be particularly effective in this context [8], with many applications in the analysis of meat [9,10], fish [11–13], fruits and vegetables [14,15], cereals [16–18], and bakery goods [19]. The main advantages of this technique lie in the possibility to have rapid and cost-effective analyses, as well as high accuracy, allowing to overcome the limits of human visual inspection characterized by high variability and labour costs. Hyperspectral imaging is extremely useful: it allows to both obtain the chemical fingerprint of the sample, as each pixel in the collected image contains a complete spectrum within a specific wavelength range [20], and to record the shape of the grains, from which the morphological traits can be derived.

In this project the application of NIR-HSI in the analysis of 47 different rice varieties is developed. The choice of using this technique, besides the rapidity of the analysis and the ease of use, is based on the possibility of exploiting two sources of information, i.e., the morphological parameters and the NIR spectrum of the samples. The aim of this project was to test a technique able to easily discern among different rice types, providing a useful tool against the widespread issue of the food fraud in the rice industry.

5.2. Exploratory analysis: results and discussion

For information concerning the theorical principles of the PCA, please refer to Section 2.1.2.1.

To perform exploratory analysis on the rice samples, 115-rows matrices of morphology and NIR spectra obtained from the HSI images' treatment were initially averaged to obtain one mean value representative of each rice variety. The resulting single-variety averaged vectors were merged to obtain two matrices of 47 rows each, on a matrix with the morphology and the other with the spectroscopic data. The purpose of working with averaged information is to obtain clearer visualisations in the space of the PCs (as one single score per variety would be present), but also to make the results more robust.

The external information provided by Ente Nazionale Risi and reported in **Table 2.3** of Section 2.3.3 was used as metadata to further describe the samples and interpret the results. To model the morphological parameters, the information concerning the shape of the rice kernels was considered: this information is illustrated in **Figure 5.1**, with a visual explanation of the five different possible shapes of the rice grain, namely, "Round", "Medium", "Long A from internal market", "Long A for parboil" and "Long B".

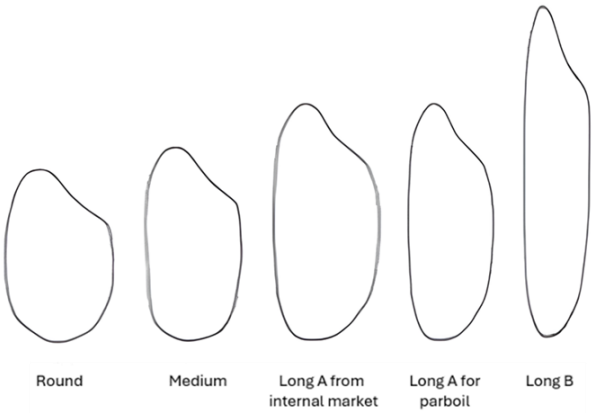

Figure 5.1. different shapes of the rice grains. Information provided by Ente Nazionale Risi.

The NIR data were modelled exploiting the information concerning the amylose content of the rice samples. Three types of amylose content were defined: “Low”, “High”, and “Waxy”. The latter refers to a particular rice type, whose starch is mainly composed by amylopectin, while amylose is present just in a low percentage [21,22]. This peculiarity is responsible of the sticky consistency of this rice type during cooking, which is not normally seen in the non-waxy rice varieties were the amylose content is predominant.

5.2.1. PCA of morphology

Before each exploratory analysis, the morphological data were pre-processed with Autoscaling, to make the different parameters comparable.

The PCA results of morphology are shown in **Figure 5.2**. Each marker in the scores plot of **Figure 5.2a** represents one variety, as all 115 grains of each variety were averaged. The visual coding of the scores is related to the information concerning the shape of the rice grain, as reported in **Table 2.3** of Section 2.3.3.

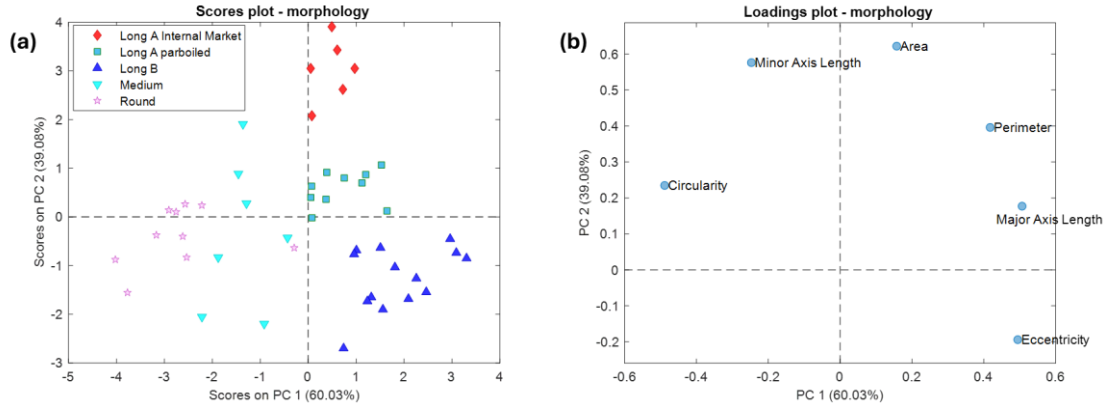

Figure 5.2. PCA scores (a) and loadings (b) plots of the morphological data for the first two PCs, with colouring based on the information about the shape of the rice kernel. Each point in the scores plot represents one individual variety.

In the scores plot of **Figure 5.2a** it is possible to notice the presence of different groups of samples, clearly associated with the shape of the rice grains. Indeed, five well-defined groups can be seen, with just one small exception with the variety Italmochi, belonging to the "Round" variety, located among the samples of class "Medium". The "Medium" varieties lie in a middle position with respect to all other groups: considering its name and related characteristics, its placement seems to be consistent with the nature of this rice type. "Round" and "Long B" rice varieties are located on the opposite sides of PC1, suggesting that their morphological features are counterposed. The loadings plot of **Figure 5.2b** confirms this hypothesis, because the features at negative PC1 loadings are Circularity and Minor Axis Length, which are coherent with a round shape and correspond to the samples of "Round" type in the negative direction of the scores of PC1, while in the positive loadings direction there is Major Axis Length, a parameter with high value for the "Long B" varieties, due to their elongated shape.

5.2.1.1. PCA of morphograms

As seen for the morphological parameters, also the morphograms were used to inspect the data through exploratory analysis. **Figure 5.3** shows the results of the PCA performed using the morphograms of each rice variety. The data were pre-processed with mean-centering.

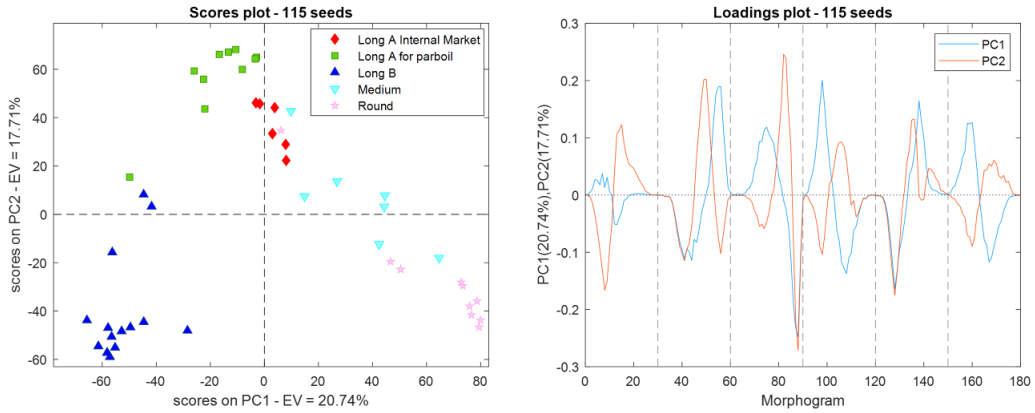

Figure 5.3. PCA scores (a) and loadings (b) plots of the morphograms for the first two PCs, with colouring based on the information about the shape of the rice kernel. Each point in the scores plot represents one individual variety.

Also in this case, the scores plot (**Figure 5.3a**) reveals distinct groups of samples based on the rice grain shape. In particular, samples are gradually separated along PC1 according to their roundness, starting from the most elongated rice type ("Long B") on the left of PC1, and ending with the "Round" type on the right. The "Medium" type lies in a middle position between the "Long A" scores and the "Round" ones, while the "Long B" type represents the cluster that separates the most from the others, except for two samples that lie close to one "Long A for parboil" score. Analysing the loadings plot, it can be stated that the variables which

mostly influence the model are Circularity and Eccentricity (second and third block of peaks). The strong negative peak in PC1 and PC2 for Eccentricity is consistent with the position of the "Long B" scores. The "Medium" and "Round" rice types lie in the positive part of PC1, where also the loadings for the Circularity are positive, and the "Long A" rice types lie in an average position along PC1.

5.2.2. PCA of NIR data

The information carried by the NIR spectra was explored as seen in the previous sections for morphology: after the pre-treatment discussed in Section 3.3.2.4, a single representative spectrum for each variety was calculated averaging the spectra of the 115 rice grains. Then, PCA was performed exploiting the information concerning the amylose content of the rice types (**Table 2.3**). These results are shown in **Figure 5.4**.

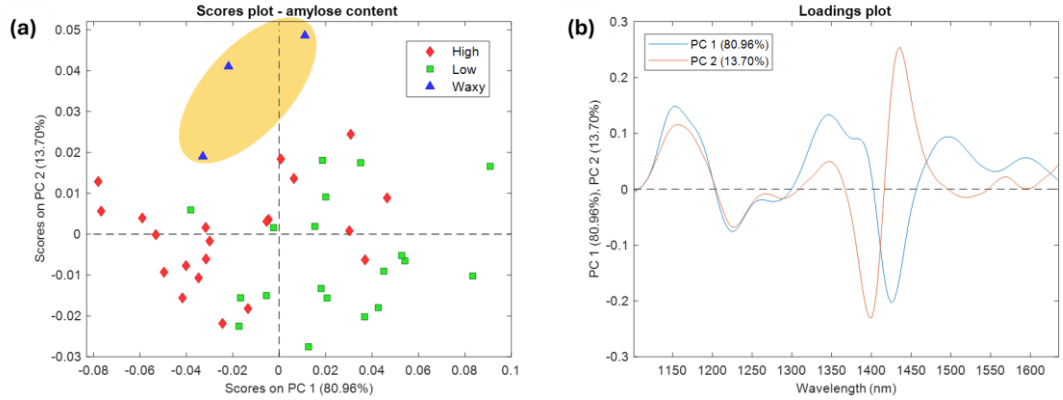

Figure 5.4. PCA scores (a) and loadings (b) plots of the NIR data for the first two PCs, coloured and shaped based on the information about the amylose content. Each point in the scores plot represents one individual variety.

The first observation about the results of this PCA is that the Waxy varieties are quite well distinguished from the other two amylose types, which appear slightly overlapped between them, as it can be clearly seen in **Figure 5.4a**. Despite this, there is a separation tendency along PC1: the high-amylose rice varieties occupy the left side of PC1, while the low-amylose scores lie mainly on the right side.

Concerning the loadings plot, a strong downward peak around 1450 nm for PC1 can be noticed, which can be ascribable to the combination of first overtone of O–H anti-symmetric and symmetric stretching of amylose [23,24]. This is coherent with the observation on the scores plot, where the samples with high content of amylose lie mostly on the negative side of PC1. Another strong peak can be detected along PC2, around 1400 nm, and can be attributed to the first overtone of O–H stretching of water [25], suggesting that the Waxy rice types could be characterized by lower content of water.

5.2.3. PCA of fused data

The results concerning the study of morphology and NIR spectra provided in both cases interesting insights about the possible differentiation among different rice varieties. While the analysis of NIR data provided insights into rice distinction, these models proved to be more limited in robustness and reliability if compared to the same kind of analysis developed using the morphological features.

Starting from these premises, it was decided to merge the two sources of information obtaining a fused dataset to be studied. To do that, a low-level data-fusion approach was implemented [26,27]. The "low-level" approach consists in a first pre-processing of the morphology and NIR data separately, as detailed in Section 2.3.3.2 and Section 2.3.3.3. Then, each matrix was divided by its own norm, and were finally merged, obtaining one final matrix with both these information together.

The results of the PCA on data fusion are reported in **Figure 5.5**, where the scores are coloured based on the amylose content of each variety, while the markers describe the rice grain shape, as seen for the PCA of NIR spectra and morphology.

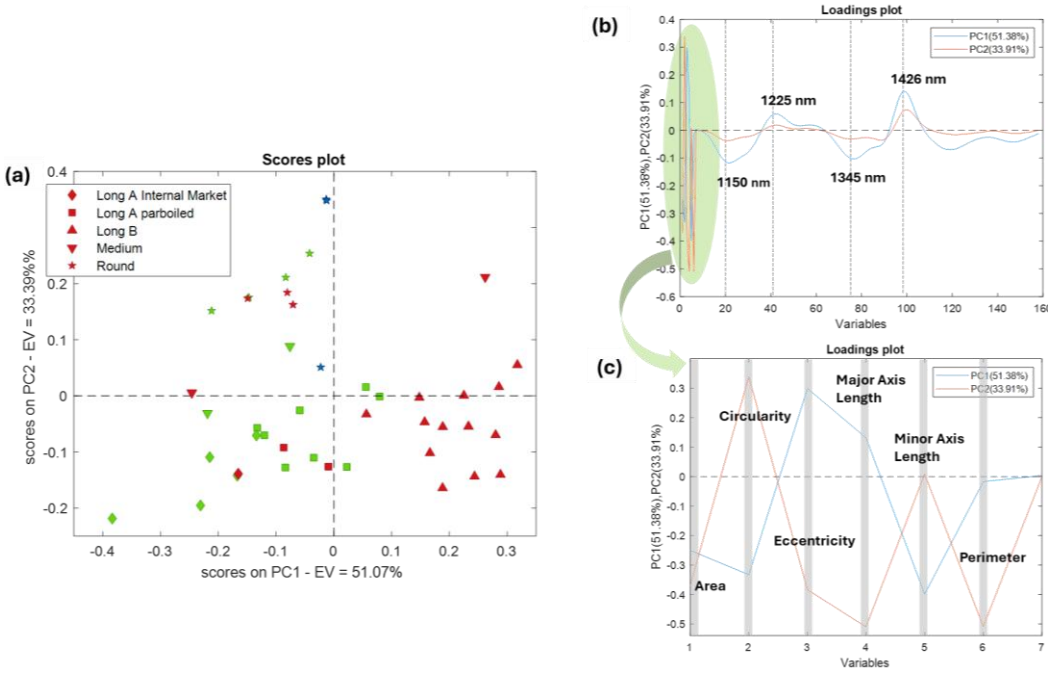

Figure 5.5. PCA scores (a) and loadings (b) plot of the fused matrices of NIR and morphology. Samples are represented by the shape of the rice grain and coloured according to the amylose content: green = low amylose, red = high amylose, blue = waxy. Figure 5.6b is zoomed (Figure 5.6c) to highlight the contributions of the morphological parameters.

The loadings plot in **Figure 5.5b** shows that the morphological parameters (at the left side of the plot) strongly influence the PCA model. A correlation between morphological and spectroscopic features can be deduced by observing the loadings plot: Eccentricity and Major Axis Length lie in the positive direction of PC1, as

well as two NIR features, namely the absorption bands at 1225 nm and 1426 nm, ascribable to the second overtone symmetric stretching of methyl group CH_3 of carbohydrates and fat [28,29], and the second overtone stretching of the O–H group of amylose, respectively. Circularity and Minor Axis Length lie in the opposite direction of PC1, together with the absorption bands at 1150 nm and 1345 nm, the first one referring to the absorption of carbohydrates and fat, and the second one to the first overtone of O–H stretching of water. This correlation between structural and spectroscopic features suggests that the rice varieties with an elongated shape contain higher levels of amylose, while the circular shape is more linked to higher levels of water in the grain.

Indeed, this observation is confirmed in the scores plot (**Figure 5.5a**): all the "Long B" rice types (positive direction of PC1) show high levels of amylose, while almost all "Long A from Internal Market" (negative and central parts of PC1) show low levels of this chemical component. The three Waxy varieties share the same round shape and lie in the positive side of PC2, whose Circularity is high in the loadings plot; their position is slightly moved to the left side of PC1, which is mostly occupied by the low-amylose rice varieties, coherently with the nature of this particular type of rice. So, it is possible to assert that the data fusion approach proved a useful strategy to correlate two different sources of information and to obtain a deeper understanding of the data under examination.

5.3. Hierarchical clustering results and discussion

Another step in the exploratory analysis of rice included the application of hierarchical clustering [30,31] to define natural groups of rice varieties. The intent was to observe how the different varieties group, and to exploit the information obtained from this methodology to build classification models, making the classification task simpler since modelling 47 varieties individually can be rather difficult to achieve and, most importantly, to interpret. The first step in this sense involved the application of hierarchical clustering to the averaged matrices of all varieties, using the information of amylose content for the NIR data, and the shape of the rice grain for the matrix concerning the morphological parameters. Secondly, the same method was used to further classify the varieties by considering them as individual classes, without any external class information.

5.3.1. Clustering of morphology

The first hierarchical clustering performed in this work involved the morphological data, which were initially explored using the shape of the rice grain as class information. The resulting dendrogram obtained from this approach is shown in **Figure 5.6**. Observing this figure, it is possible to understand how the model captures the main differences/similarities among the samples and how it uses this information to efficiently partition the five (expected) groups of rice shapes: more

specifically, the two groups that seem to be the most similar are those containing the “Medium” and “Round” varieties, since their linking distance is the shortest compared to the other groups, and consequently appear as well distinguished from the other rice types. These long linking distances describe a very well-defined clustering structure, mostly corresponding to the expected groups (i.e., the five rice shapes), with the exception of the “Medium” shape, which is consistent with the nature of this rice type.

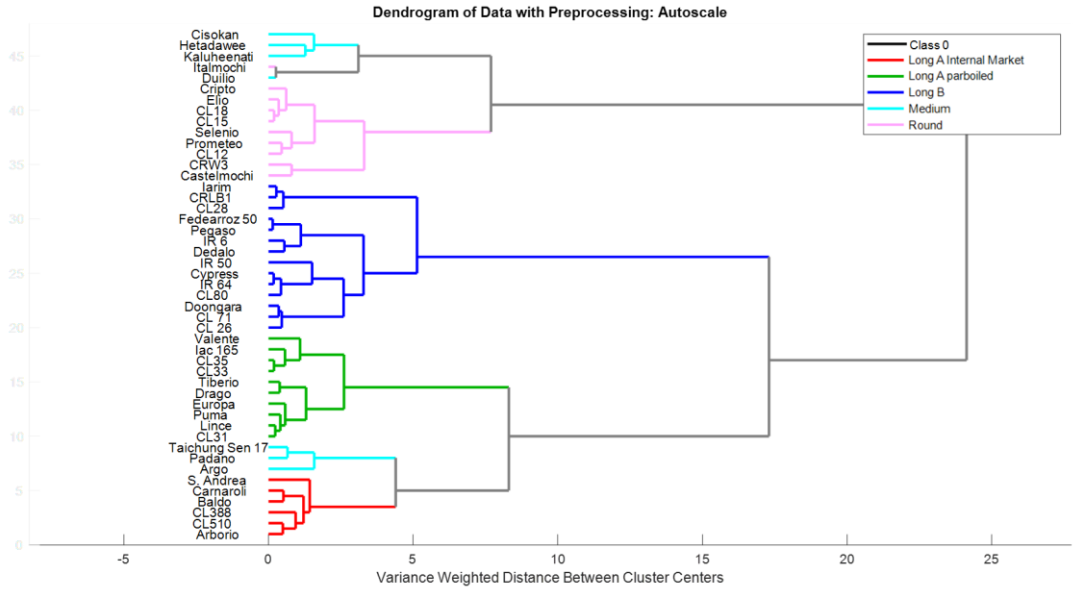

Figure 5.6. Dendrogram obtained from the hierarchical clustering for the morphological parameters using Ward’s method. Samples are coloured according to the shape of the rice grain.

These results provide a useful basis for defining few groups of varieties for building further classification models, as needed for the PLS-DA models that will arrange the structure of the manual hierarchical classification analysis discussed in Section 5.4.2., both for the classification according to the kernel shape, and the one considering a variety as a class itself.

5.3.2. Clustering of NIR data

The next step involved the application of hierarchical clustering on the NIR data. As among the samples there were four rice varieties for which the amylose content was not provided by Ente Risi, it was chosen to treat the data in the following way: those samples were removed from the averaged matrix, obtaining a final amount of 43 rice varieties instead of 47, and the “Very High” amylose varieties were merged with the “High” ones, resulting in just three classes, namely “Waxy”, “Low”, and “High” amylose content. This was done also to work with more balanced classes. The resulting matrix was analysed through hierarchical clustering, resulting in the dendrogram shown in **Figure 5.7**.

If compared to the previous results concerning morphology, in the case of NIR spectra the separation among the classes is more confused, and it is not possible to see the samples divided into three well-defined groups. For this reason, as the number of classes is very low, it was decided to perform also a PLS-DA on this matrix and to refer to its results to obtain a better explanation about the class division for this type of data. These results will be discussed later in Section 5.4.2.

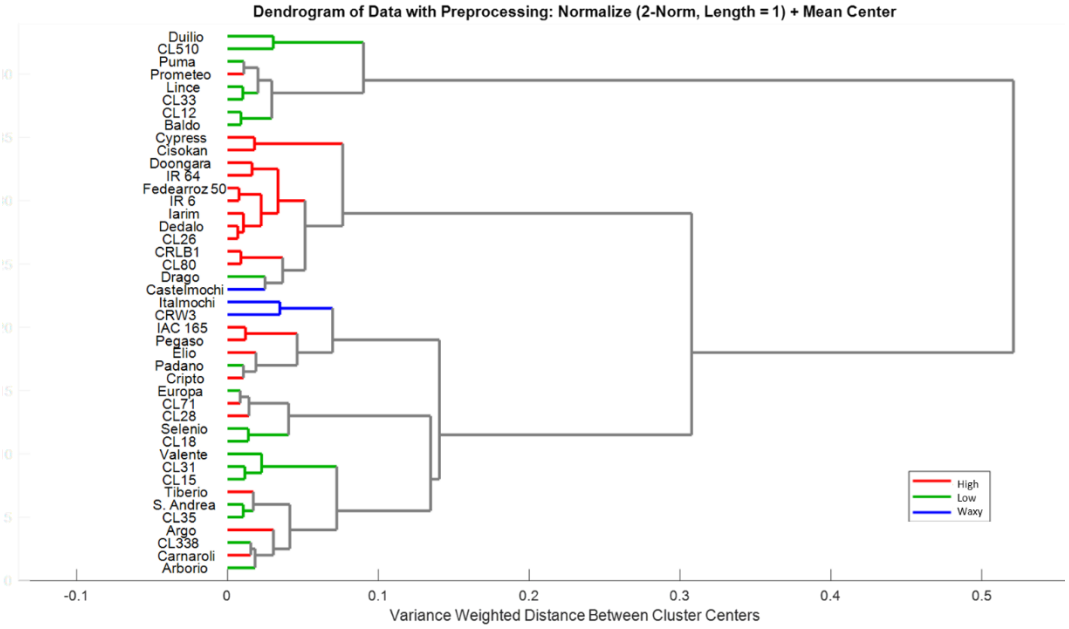

Figure 5.7. Dendrogram obtained from the hierarchical clustering for the NIR data preprocessed with first derivative, normalization and mean centering. Samples are colored according to the amylose content.

Furthermore, another hierarchical clustering analysis was performed to inspect the presence of groups of samples among the varieties themselves. To do that, the averaged matrix containing all the 47 rice varieties included in this study was used. The results of this analysis are shown in **Figure 5.8**.

A clarification about the pre-processing needs to be done: both the matrices used in the construction of these dendrograms were separately pre-processed with a Savizky-Golay (first-order derivative, second-order polynomial and a 11pt window) smoothing before any cluster analysis. Then, the other pre-processing steps were done directly in the PLS_Toolbox when performing the hierarchical clustering. For this reason, in **Figure 5.7** and **Figure 5.8** the pre-processing displayed in the upper part of the figures are just normalization and mean centering.

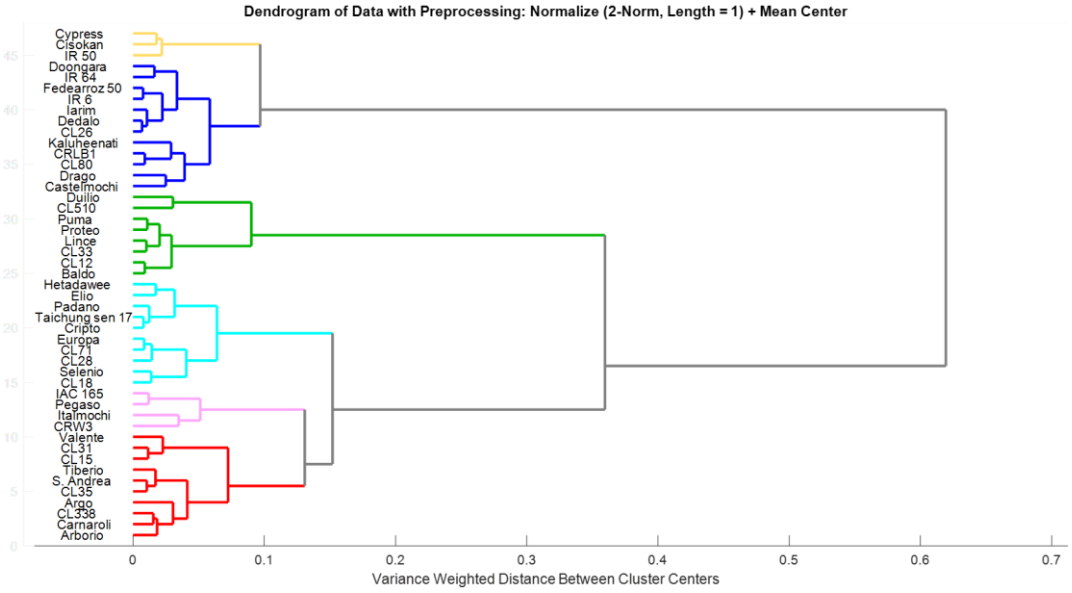

Figure 5.8. Dendrogram obtained from the hierarchical clustering for the NIR data, preprocessed with first derivative, normalization and mean centering. The colouring does not correspond to any additional information, but it is used for better visualisation of the identified clusters.

5.3.3. Clustering of fused data

As the results obtained with the exploratory analysis of fused data proved to be informative, it was chosen to deepen their exploration by building hierarchical models with the same criteria illustrated in the previous sections. A first hierarchical clustering was done by using the shape of the rice grains to colour the dendrogram, to visually inspect the presence of groups of samples according to this parameter. **Figure 5.9** displays the results.

The "Medium" class proved again to be quite tricky to isolate, with a larger dispersion within the other categories than what was previously seen in **Figure 5.6**. Besides the "Medium" group, all the others maintain their entirety and few misplacements can be found, suggesting that the morphological traits bring predominant information in the definition of the dendrogram when dealing with data fusion, as already proved with the PCA results on this type of data.

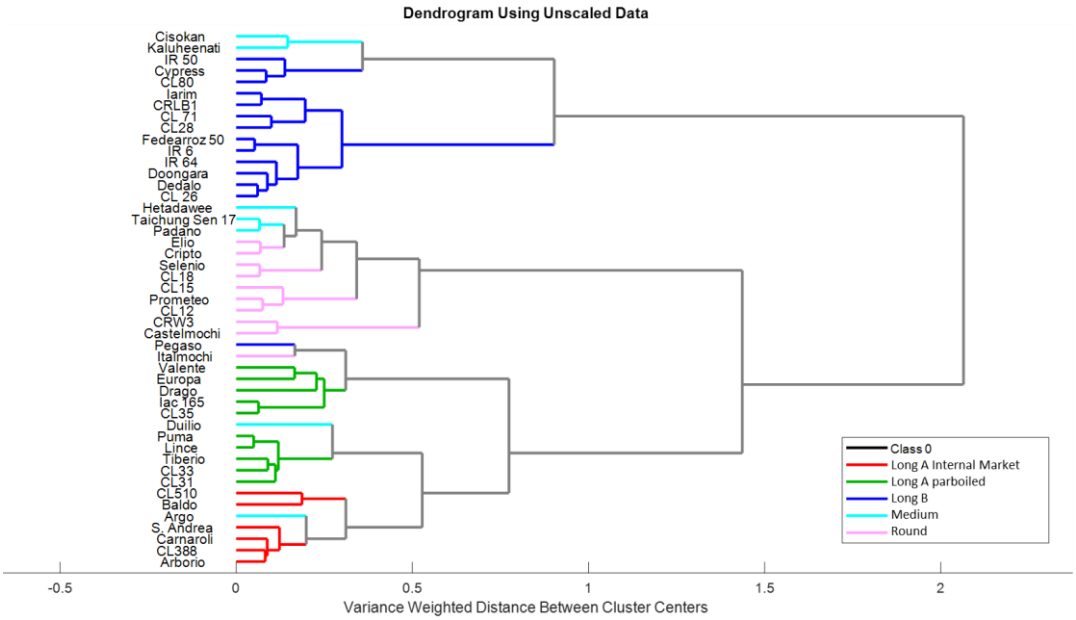

Figure 5.9. Dendrogram obtained from the hierarchical clustering for the fused data of morphology and NIR spectra. Samples are coloured according to the shape of the rice grain.

Then, the same method was applied on the data colour coded by means of the amylose content, with the results displayed in **Figure 5.10**.

The two Waxy varieties Castelmochi and CRW3 end up in the same final cluster, and they represent the two varieties that are overlapping in the PCA scores plot in **Figure 5.4**. Italmochi, which is the third Waxy variety, lies in a completely different group, very close to a high amylose variety. If considering the PCA of the amylose content of NIR data in **Figure 5.3**, the three Waxy rice types are well-distinguished and grouped from the others, while in the PCA of morphological features in **Figure 5.2**, Italmochi is the only "Round" variety which lies far from the "Round" group. This seems to confirm that there is a strong influence of the morphological parameters in the results of the analysis of data fusion.

Finally, nearly all the varieties with high levels of amylose group in the same cluster, improving the results of PCA of NIR spectra (**Figure 5.3**) where a distinction among high- and low-amylose varieties was not very clear. This can be evidence that the data fusion approach enriches and strengthens the information carried by the samples, helping to solve the criticalities seen with the analysis of the NIR spectra themselves. Moreover, the information represented by the PCA of fused data in **Figure 5.4** is confirmed by this dendrogram: almost all the high amylose varieties are represented by the "Long B" shape of the rice grain.

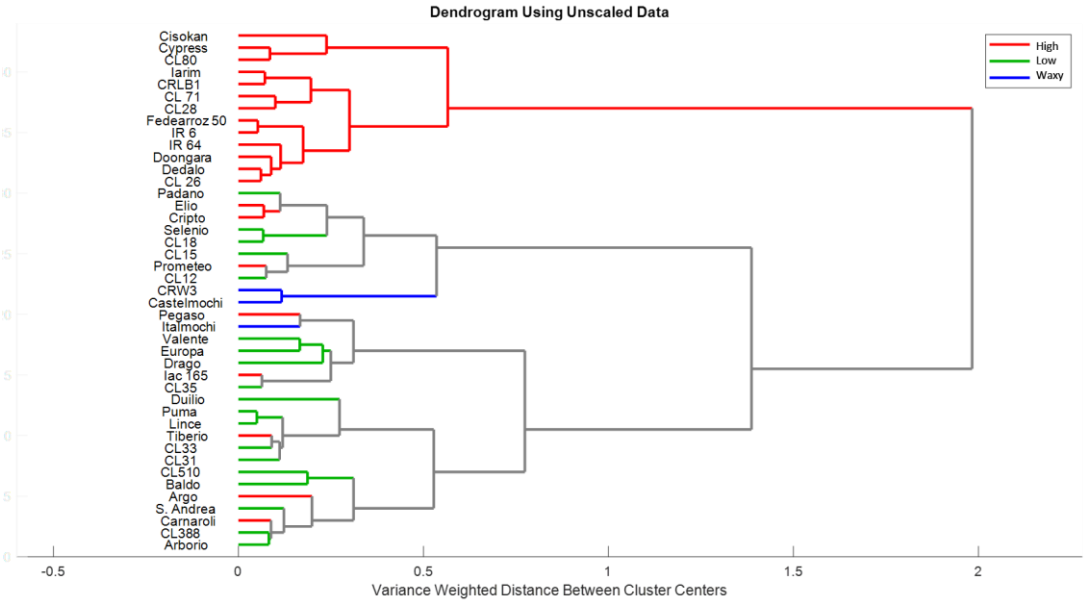

Figure 5.10. Dendrogram obtained from the hierarchical clustering for the fused data of morphology + NIR data. Samples are coloured according to the amylose content.

5.4 Classification models: results and discussion

The information obtained from all the exploratory steps became a useful starting point in the development of classification models aimed at distinguishing among varieties in the anti-food fraud optic. To do that, two types of models were developed: Partial Least Squares-Discriminant Analysis (PLS-DA) and hierarchical classification models.

5.4.1. Partial Least Squares-Discriminant Analysis (PLS-DA)

The core of this technique lies in the exploitation of the PLS regression algorithm to discern and classify samples: the working principle of PLS-DA is to look for the maximum covariance between the original variables and the class labels, the latter defined as the Y-dependent variables. Similarly to PCA, PLS-DA works in a transformed space, where the original variables are converted into the so-called Latent Variables (LVs). These new variables describe the variability of the original data and allow to reduce their dimensionality, in addition to enable their graphical visualisation. The decomposition carried by PLS-DA can be resumed in the following Equations 5.1 and 5.2:

(5.1) $$X = TP^T + E = \hat{X} + E$$

(5.2) $$Y = TQ^T + F = \hat{Y} + F$$

Undeniably, Equations 5.1 and 5.2 are very similar to each other. The main difference here is that the decomposition is performed also on the response vector Y, which is divided into the T scores matrix and the Q loadings matrix, and F is the error matrix.

Starting from here, it is mandatory to define the weights matrix W, which highlights the importance of the original variables from X in the definition of the LVs. This matrix is calculated iteratively by the PLS-DA algorithm and is used to create the scores matrix based on how much every single variable in X contributes to the LVs. The scores matrix can be rewritten as in Equation 5.3:

$$T = XW(P^TW)^{-1} \tag{5.3}$$

and consequently Equation 5.2 becomes:

$$Y = XW(P^TW)^{-1}Q^T + F = \hat{Y} + F \tag{5.4}$$

Where $\hat{Y}$ is the predicted value of the response Y, and the error term F defines the difference between the real and the predicted value of Y. In the present work, PLS-DA was used to predict to which class the rice varieties belonged, by means of morphological and spectroscopic features. Differently from the exploratory analysis, in the case of the classification models no averaged matrices were used, considering all 115 rice grains per variety in the construction of the models. Initially, the pre-processed matrix was split into training and test set, with the aid of Kennard-Stone [11] algorithm. The test set consisted in the 33% of the total amount of samples. The training set matrix was used to create and calibrate the model, whose performances were eventually evaluated through the test set. The cross-validation of all models involved the venetian blinds selection scheme with 10 splits and blind thickness equal to 1.

To evaluate the model performances, four quality indexes were considered:

1. **Specificity (Spec)**: the ability to avoid false positives in classification. It is calculated as shown in Equation 5.5.
2. **Sensitivity (Sens)**: the ability to avoid false negatives in classification. It is calculated as shown in Equation 5.6.
3. **Non-error rate (NER)**: obtained calculating the mean of the sensitivities and it represents the ability of the model to correctly classify the samples. It is calculated as shown in Equation 5.7.
4. **Accuracy (Acc)**: it represents an estimation of the model's error, and it is calculated as shown in Equation 5.8.

(5.5) $$\text{Spec} = \frac{\text{TN}}{\text{TN+FP}}$$

(5.6) $$\text{Sens} = \frac{\text{TP}}{\text{TP+FN}}$$

(5.7) $$\text{NER} = \frac{1}{n}\sum_{i=1}^{n} \text{Sensitivity}_{\text{i}}$$

(5.8) $$\text{Acc} = \frac{\text{TP} + \text{TN}}{n}$$

where TP = true positives, TN = true negatives, n = n° of samples

5.4.1.1 PLS-DA of morphology

The PLS-DA model for the morphological features was built following the criteria of the PCA model: all the 47 rice varieties, represented by their 115 rice kernels, were classified according to the shape of the rice grain. The performances of this PLS-DA model are shown in **Table 5.1**.

Table 5.1: Classification results for PLS-DA model of morphology, with the information about the shape of the rice kernel. All values refer to the quality indexes in prediction and are expressed in percentage.

	LVs	Sens	Spec	NER	Accuracy
Long A from internal market	4	97.9	95.3	94.1	96.0
Long A for parboil		97.9	63.4		87.3
Long B		97.1	96.3		95.0
Medium		87.9	67.2		87.2
Round		89.7	90.1		92.6

Based on the classification performances, it is possible to deduce that the model can globally discern well the five different shapes of the rice grains. The “Medium” class is still a bit problematic, showing the lowest value of Specificity, followed by the “Long A for parboil” class. This means that for these two classes the number of

false positives can be rather high, as the model mistakes rice types of a different shape as "Medium" or "Long A for parboil" and consequently leads to classification mistakes for these two classes in particular. This observation confirms what was already seen with PCA in Section 4.3.1: the morphology of these types of varieties, which both lie in the central part of the scores plot, can generate a sort of confusion during the classification process, and rice grains of a different type of shape can be misclassified as "Medium" or "Long A for parboil". The "Long B" class is the best classified one, probably due to its peculiar shape that can more easily be recognised from the others.

5.4.1.2. PLS-DA of NIR data

The pre-processed 115 rice grains spectral matrices were used to build the PLS-DA model, using the amylose content as class information. The results are shown in **Table 5.2**.

Table 5.2. Classification results for PLS-DA model of NIR spectra, with the information about the amylose content. All the values refer to quality indexes in prediction and are expressed in percentage.

	LVs	Sens	Spec	NER	Accuracy
High		74.2	82.6		91
Low	7	87.7	78.5	86.7	81.6
Waxy		98.3	97.6		98

This PLS-DA model shows high values for Sensitivity and Specificity, in particular for the Waxy varieties: this observation is in accordance with the PCA results, where the Waxy rice types were separated from the others in the space of the PCs. More criticalities are found in the distinction among samples with low and high levels of amylose: for these varieties, the performances of the model are lower, but still acceptable when looking for a preforming classification model. Based on these results, it is possible to say that this model can distinguish well the Waxy rice type from the others.

5.4.1.3. PLS-DA of fused data

A trial of classification using the fused data matrix was done, even if the results in the exploratory analysis proved a bit tricky if compared to those obtained by morphology and NIR. The classification followed the criteria of the PCA, starting

with a first PLS-DA model with the shape of the rice kernel as class information. These results are resumed in **Table 5.3**.

Table 5.3: Classification results for PLS-DA model of data fusion, with the information about the shape of the rice kernel. All values refer to the quality indexes in prediction and are expressed in percentage.

	LVs	Sens	Spec	NER	Accuracy
Long A from internal market	3	96.6	96	94.9	96.9
Long A for parboil		86.6	52.4		91.2
Long B		96.7	95		96.5
Medium		88.6	67.7		86.3
Round		91.3	86.4		91.9

The performances of this model are good, with high values of Sensitivity and consequently a high Non-error rate, and a good level of Accuracy, suggesting that this model is sufficiently robust for the classification of samples into these five rice types. A criticality is found with the "Long A for parboil" class, which shows the lowest performances, especially for the Specificity. If compared to the PLS-DA model obtained with the morphology itself, it can be seen that the performances of this model with data fusion worsen for this class. Consequently, even if this model classification is globally robust, it would be probably better to consider a classification with only the morphological features of the samples, avoiding merging them with NIR data, when classifying the samples according to their shape.

The second PLS-DA model for this dataset was built using the amylose content to classify the 47 rice varieties, with the results shown in **Table 5.4**.

Table 5.4. Classification results for PLS-DA model of fused data, with the information about the amylose content. All the values refer to quality indexes in prediction and are expressed in percentage.

	LVs	Sens	Spec	NER	Accuracy
High	4	71.3	86.5	89	83.7
Low		95.7	75.4		84.2
Waxy		100	97.9		99.2

These results are comparable with those obtained from the PLS-DA on NIR data, even if some improvements are found with the data fusion approach: for example, both the Waxy and the Low amylose classes improved in Sensitivity, and this leads to a higher Non-error rate if compared to NIR classification. A high Sensitivity means a lower presence of false negatives, i.e., a better ability of the model to correctly classify samples in the belonging class. The data fusion approach can consequently be preferred with respect to the NIR data itself to obtain a more robust classification in this sense.

5.4.2. Hierarchical classification

Trying to build classification models with datasets characterised by a large number of classes may be a challenge for the traditional chemometrics approaches. In this respect, hierarchical modelling can be considered as a valuable alternative to traditional classification techniques, thanks to the possibility to efficiently analyse datasets with considerable complexity. Typical examples of hierarchical classification involve decision trees, k-nearest neighbour (KNN, [32]) and support vector machines (SVM, [33]), although in chemometrics the most common way to classify in this respect is to use hierarchical structures based on PLS-DA models. The troublesome part is the construction of these structures, as it demands the knowledge of a skilled person, and it generally takes time. A solution to this concern can be the employment of a newly developed automatized approach, the Automatic Hierarchical Classification Model Builder (AHIMBU, [34]). This method can help saving time and efforts, allowing rapid modelling for classification purposes even in the case of a huge number of samples that makes it impossible to manually build a model structure.

In this thesis, both automatic and manual hierarchical models were used to classify the morphological and spectroscopic data. This type of models exploits a series of PLS-DA models, one for each classification step that occurs during the analysis, enabling to start from an all-inclusive initial class consisting of all the samples and subclasses together, and finishing with the partitioning of each object of the dataset

as a class itself. The individual PLS-DA models act as “rules” of the hierarchical model, indicationg how to classify the samples and giving the typical tree structure to the hierarchical model. Consequently, at the end of the classification the number of branches in the hierarchical structure will be twice the number of the PLS-DA models applied in the construction, as each model operates the distinction between two classes.

In the case of manual hierarchical models, their building was done under the human supervision, starting with the calculations of the PLS-DA models constituting the nodes of the structure. The information given by the clustering was used to divide the samples into groups. Then, the information obtained was used to calculate the PLS-DA models, which were further used to build the structure of the hierarchical manual model. On the contrary, AHIMBU works calculating automatically the PLS-DA nodes from the training set, after a preliminary definition of the appropriate number of latent variables for the analysis. The training set and test set for the hierarchical models were obtained from the complete dataset as described in Section 5.4.1.

5.4.2.1. Hierarchical models of morphology

The same dataset used in the construction of the PLS-DA model for the morphological data was used to build the hierarchical models. The first application involved the automatic hierarchical model: to build the AHIMBU structure, PLS-DA models with 6 LVs were automatically calculated, as previously explained. An image showing the structure of AHIMBU model is shown in **Figure 5.11**, while the results for this classification are shown in **Table 5.5.**

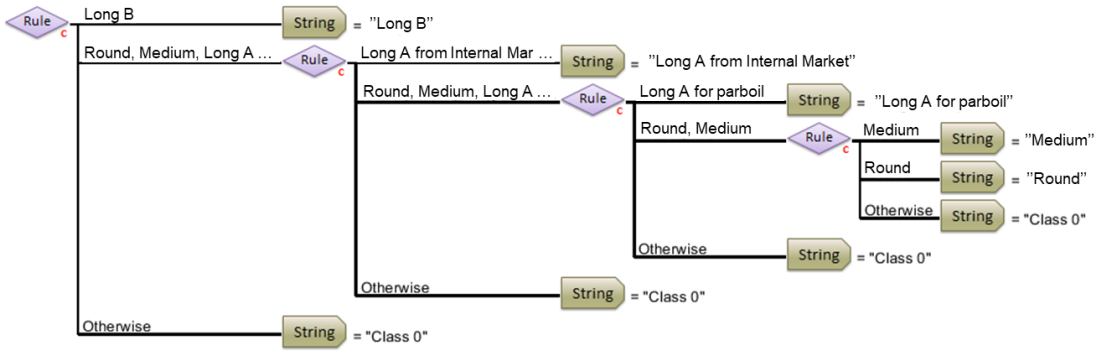

Figure 5.11. structure of the AHIMBU for the morphological data. Samples are classified according to the shape of the rice grain.

Globally, the performances of the automatic hierarchical model are good and comparable with those obtained with the global PLS-DA model. In the case of the hierarchical model, Specificity proved to be generally better than the case of PLS-

DA, while, on the other hand, PLS-DA shows better performances concerning the Sensitivity. This suggests that the hierarchical model better avoids wrong inclusions from different classes, while PLS-DA better recognizes the correct class. Globally, these methods can be considered strong and reliable, as well as complementary, as both work very well concerning the aim of this work. The choice of using of one method instead of the other can be driven by the type of desired classification, i.e., if it is more important to obtain high Specificity or high Sensitivity. Also, the work that must be done to obtain the models, in terms of effort and time can be taken into account.

Table 5.5. Classification results for AHIMBU of morphology, with the information about the shape of the rice kernel. All values refer to the quality indexes in prediction and are expressed in percentage.

	LVs	Sens	Spec	NER	Accuracy
Long A from internal market	6	97.4	97.1	87.1	97.2
Long A for parboil		81.5	95.6		92.6
Long B		98.0	96.4		96.9
Medium		71.4	98.1		94.1
Round		87.2	97.9		95.6

Starting from these observations, it was chosen to try another way to classify the samples with hierarchical models: the idea was to use each variety as an individual class, to see which groups of varieties the models were able to correctly identify, based on the morphological information extracted from the images. The structure of this model is shown in **Figure 5.12**. With this approach, the model classifies all the samples starting from a first division into two big groups, and then the division goes on until each individual variety is classified separately from the others. In this work we chose to stop the class divisions when the node reached the threshold value of Specificity in calibration equal or greater than 80%, assuming that this value could be used as a limit threshold to keep the reliability of the model high enough.

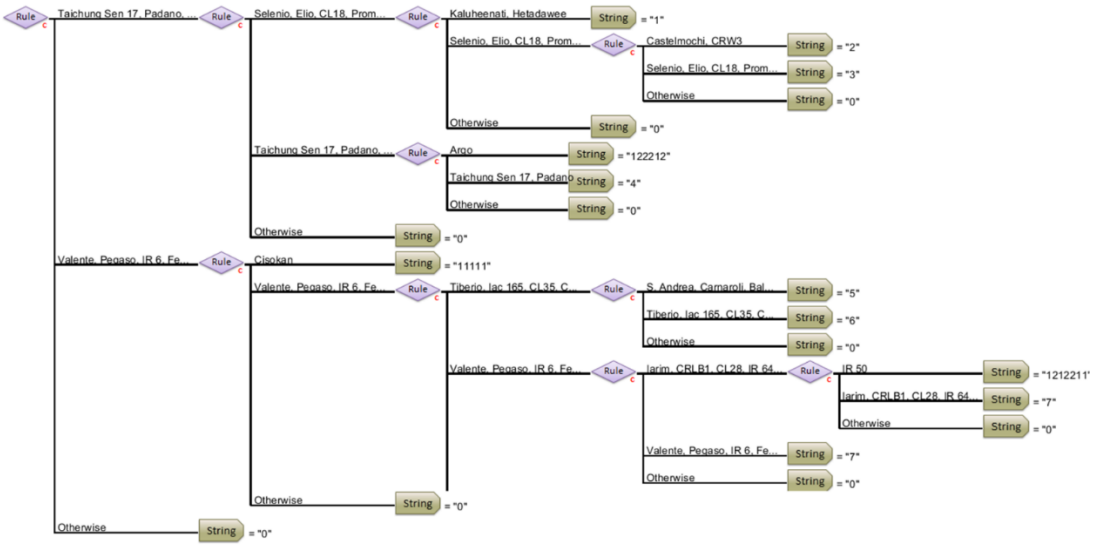

Figure 5.12. Structure of the AHIMBU for the morphological data, at the threshold of Specificity(Cal) ≥ 80%. Each variety represents a class itself.

Once reached the selected threshold, AHIMBU stopped, generating 11 groups of varieties. In **Figure 5.12**, these groups are coded with numeric strings, to facilitate the calculation of the model quality indexes in MATLAB. The actual groups of varieties and the classification performances of this model are resumed in **Table 5.6**.

Table 5.6: Classification results for AHIMBU of morphology at the threshold of Specificity(Cal) ≥ 80%. All values refer to the quality indexes in prediction and are expressed in percentage.

	Varieties	LVs	Sens	Spec	NER	Accuracy
Group 1	CL31, CL33, CL35, Drago, Duilio, Europa, IAC165, Italmochi, Lince, Puma, Tiberio	6	89.7	98.7	83.3	96.6
Group 2	CL26, CL28, CL71, CL80, CRLB1, Cypress, Dedalo, Doongara, Iarim, IR64		90.8	98.6		97.1
Group 3	Fedearroz 50, IR6, Pegaso, Valente		80.8	98.6		97.2
Group 4	CL12, CL15, CL18, Cripto, Elio, Prometeo, Selenio		84.6	99.0		96.8
Group 5	Castelmochi, CRW3		93.6	99.3		99.1
Group 6	Cisokan		82.1	98.2		97.9
Group 7	Hetadawee, Kaluheenati		84.6	98.1		97.5
Group 8	Arborio, Baldo, Carnaroli, CL388, CL510, S. Andrea		97.4	99.6		99.2
Group 9	Padano, Taichung Sen 17		84.6	98.1		97.5
Group 10	Argo		76.9	99.7		99.2
Group 11	IR 50		51.3	99.3		98.3

Observing the groups defined by the model, it was chosen to inspect whether there was or not a correlation among the varieties included in these eleven groups, and the types of rice grains shapes used to classify samples so far. Interestingly, the varieties in each group share the same shape: a clearer visualisation of this result is displayed in **Figure 5.13**.

This observation underlines the ability of this approach to detect information from the samples and to correctly classify them according to their morphological characteristics. Moreover, this classification method is suitable for consistent datasets with lots of classes, allowing to overcome the limits of the PLS-DA models, which generally work better with a low number of classes involved.

Hetadawee	Hetadawee	Hetadawee	Hetadawee	Medium		
Kaluheenati	Kaluheenati	Kaluheenati	Kaluheenati			
CRW3	CRW3	CRW3	CRW3	CRW3	Round	
Castelmochi	Castelmochi	Castelmochi	Castelmochi	Castelmochi		
CL12	CL12	CL12	CL12	CL12	Round	
CL15	CL15	CL15	CL15	CL15		
Cripto	Cripto	Cripto	Cripto	Cripto		
Prometeo	Prometeo	Prometeo	Prometeo	Prometeo		
CL18	CL18	CL18	CL18	CL18		
Elio	Elio	Elio	Elio	Elio		
Selenio	Selenio	Selenio	Selenio	Selenio		
Argo	Argo	Argo	Argo	Medium		
Padano	Padano	Padano	Padano			
Taichung sen 17	Taichung sen 17	Taichung sen 17	Taichung sen 17			
Cisokan	Cisokan	Cisokan				
CL510	CL510	CL510	CL510	CL510	Long A from I.M.	
Arborio	Arborio	Arborio	Arborio	Arborio		
CL338	CL338	CL338	CL338	CL338		
Baldo	Baldo	Baldo	Baldo	Baldo		
Carnaroli	Carnaroli	Carnaroli	Carnaroli	Carnaroli		
S. Andrea	S. Andrea	S. Andrea	S. Andrea	S. Andrea		
Duilio	Duilio	Duilio	Duilio	Duilio		
Italmochi	Italmochi	Italmochi	Italmochi	Italmochi		
Drago	Drago	Drago	Drago	Drago	Long A for parboil	
Europa	Europa	Europa	Europa	Europa		
CL31	CL31	CL31	CL31	CL31		
Lince	Lince	Lince	Lince	Lince		
Puma	Puma	Puma	Puma	Puma		
CL33	CL33	CL33	CL33	CL33		
CL35	CL35	CL35	CL35	CL35		
Iac 165	Iac 165	Iac 165	Iac 165	Iac 165		
Tiberio	Tiberio	Tiberio	Tiberio	Tiberio		
IR50	IR50	IR50	IR50	IR50	IR50	
Dedalo	Dedalo	Dedalo	Dedalo	Dedalo	Dedalo	Long B
CL26	CL26	CL26	CL26	CL26	CL26	
CL71	CL71	CL71	CL71	CL71	CL71	
Doongara	Doongara	Doongara	Doongara	Doongara	Doongara	
CL80	CL80	CL80	CL80	CL80	CL80	
Cypress	Cypress	Cypress	Cypress	Cypress	Cypress	
IR64	IR64	IR64	IR64	IR64	IR64	
CL28	CL28	CL28	CL28	CL28	CL28	
CRLB1	CRLB1	CRLB1	CRLB1	CRLB1	CRLB1	
Iarim	Iarim	Iarim	Iarim	Iarim	Iarim	
Fedearroz 50	Fedearroz 50	Fedearroz 50	Fedearroz 50	Fedearroz 50	Long B	
IR6	IR6	IR6	IR6	IR6		
Pegaso	Pegaso	Pegaso	Pegaso	Pegaso		
Valente	Valente	Valente	Valente	Valente		

Figure 5.13. Visual representation of the groups of varieties defined by AHIMBU, with the information about the rice shape.

Starting from these optimistic results, the following step concerned the application of manual hierarchical models on the same dataset. Again, the first classification was performed on the data classified based on their rice grain shapes. First, it was necessary to define how the model would divide the groups of classes: to do that, the hierarchical clustering built in the exploratory step was considered. Then, individual PLS-DA models were calculated and used to build the structure of the manual hierarchical model, which is displayed in **Figure 5.14**.

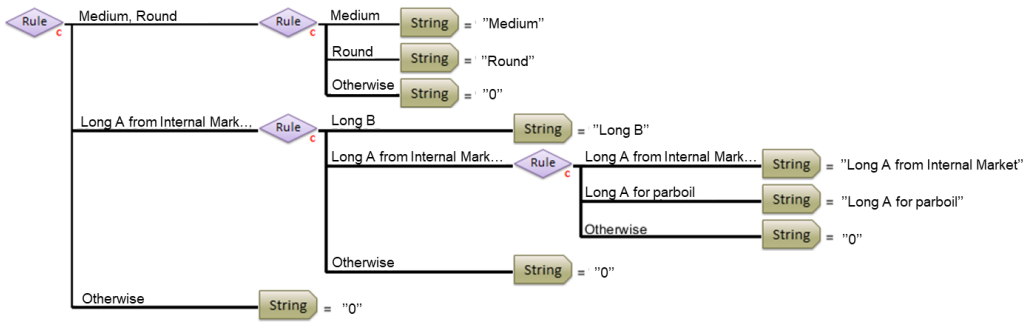

Figure 5.14. Structure of the manual hierarchical model for the morphological data. Samples are classified according to the shape of the rice grain.

The performances of this hierarchical model are resumed in **Table 5.7**. With respect to these results, the manual hierarchical classification approach performs slightly better than the automatic one, as the classification performances globally reach higher values both for what concerns Specificity and Sensitivity. If considering the latter, the same assumptions previously made are still valid: the "Medium" class is the most difficult to correctly classify, due to the inhomogeneous nature of this shape, as already discussed.

Table 5.7: Classification results for manual hierarchical model of morphology, with the information about the shape of the rice kernel. All values refer to the quality indexes in prediction. Values are expressed in percentage.

	LVs	Sens	Spec	NER	Accuracy
Long A from internal market	3	97.0	99.6	90.3	96.6
Long A for parboil	3	89.2	95.8		94.4
Long B	1	97.1	97.8		97.5
Medium	5	81.0	97.6		95.1
Round	5	87.2	97.9		95.6

Subsequently, the manual approach was applied in the construction of a hierarchical model to classify all the 47 varieties as single classes, resulting in the structure shown in **Figure 5.15**. The first step involved again the use hierarchical clustering to divide samples into groups, and then for each division, PLS-DA models were built. The threshold of Specificity(Cal) ≥ 80 % was applied as threshold for the model. Again, the groups are numerically coded for a matter of practicality in the further data treatment. **Table 5.8** shows the actual groups and the classification results for this manual hierarchical model.

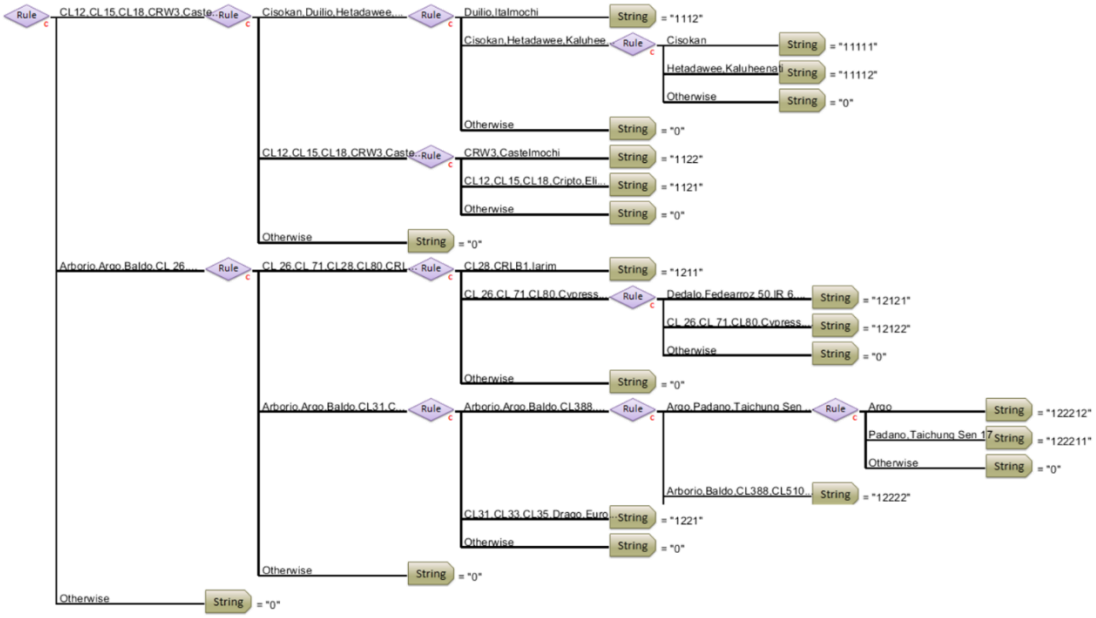

Figure 5.15. Structure of the manual hierarchical model for the morphological data, at the threshold of Specificity(Cal) ≥ 80%. Each variety represents a class itself.

The first difference between the manual and the automatic hierarchical models concerns the number of groups of varieties identified by the two approaches: AHIMBU defined 11 groups, while 12 were found by the manual approach. As a second difference, the manual model has two groups with very low values of Sensitivity, namely Group 1 and Group 11. For this reason, also the quality index of Non-error rate is lower than the one of AHIMBU. This observation suggests that in the case of a high number of classes, the automatic approach could represent a better choice with respect to the manual one, as it shows better performances and allows for rapid classification.

	Varieties	LVs	Sens	Spec	NER	Accuracy
Group 1	Duilio, Italmochi	4	34.6	97.8		95.1
Group 2	CL12, CL15, CL18, Cripto, Elio, Prometeo, Selenio	2	86.5	95.5		94.2
Group 3	Castelmochi, CRW3	2	96.2	98.8		98.7
Group 4	CL28, CRLB1, Iarim	1	94	95.3		95.3
Group 5	CL31, CL33, CL35, Drago, Europa, IAC165, Lince, Puma, Tiberio, Valente	2	88.7	95.3		93.9
Group 6	Cisokan	2	84.6	99.9	71.6	99.6
Group 7	Hetadawee, Kaluheenati	2	74.4	99		97.9
Group 8	Dedalo, Fedearroz 50, IR6, Pegaso	4	69.9	96.8		94.5
Group 9	CL26, CL71, CL80, Cypress, Doongara, IR50, IR64	4	65.9	98		93.2
Group 10	Arborio, Baldo, Carnaroli, CL388, CL510, S. Andrea	4	94.9	99.6		98.9
Group 11	Padano, Taichung Sen 17	4	7.7	99.7		95.8
Group 12	Argo	4	61.5	99.7		98.9

Table 5.8: Classification results for manual hierarchical model of morphology at the threshold of Specificity(Cal) ≥ 80 %. All values refer to the quality indexes in prediction and are expressed in percentage.

The groups obtained with the manual hierarchical model were inspected as it was done for the AHIMBU case: it was proven that also with the manual approach the varieties included in each group share the same type of grain shape, as displayed in **Figure 5.16**. This highlights that also the manual model can be successfully used to exploit the information contained in the data to classify the samples thanks to their structural features, with reliable results that make it useful in the industrial applications.

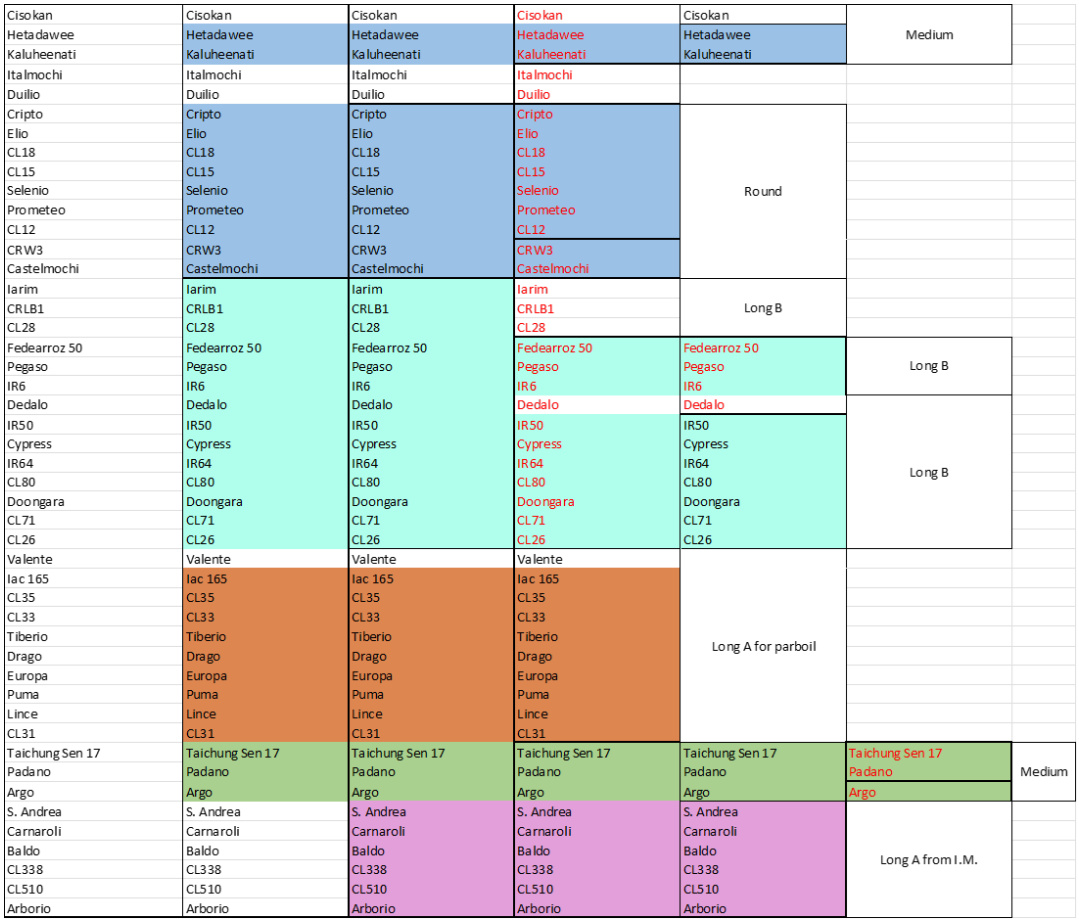

Figure 5.16. Visual representation of the groups of varieties defined by the manual hierarchical model, with the information about the rice shape.

Finally, the fact that both approaches (automatic and manual) provide very similar results both from the point of view of the classification performances and actual composition of the groups of variety indicates that the morphological information that can be extracted from the data seems to be rather robust, as reproducibility seems to hold for both approaches. A choice between the two approaches can be driven by the rapidity of the AHIMBU, which allows to skip the part of the calculation of the singular PLS-DA models and the supervised building of the hierarchical structure, saving a considerable amount of time and avoiding random accidental mistakes.

5.4.2.2. Hierarchical models of NIR data

As seen for the morphology, the first hierarchical model built to classify the NIR data was based on the AHIMBU approach. Samples were classified according to the amylose content, as seen with PLS-DA. The AHIMBU structure for this type of data is shown in **Figure 5.17**.

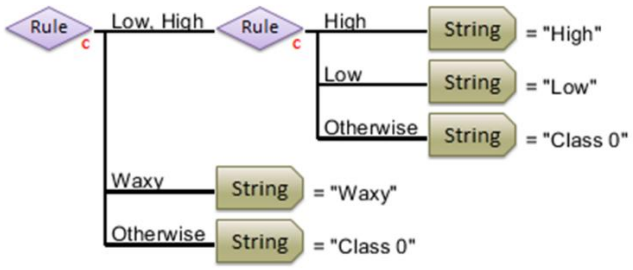

Figure 5.17. Structure of the AHIMBU for the spectroscopic data. Samples are classified according to the amylose content.

From the tree structure of the AHIMBU model it is possible to see that, also in this case, the first class to be separated is the Waxy one, confirming that also the hierarchical model detects more differences for this type of rice. The performances of AHIMBU are resumed in **Table 5.9**.

Globally, the performances of this model are rather good and comparable to the prediction quality indexes obtained with the PLS-DA model, suggesting that the two approaches have the same relevance in terms of reliability and robustness. The distinction between varieties with low and high amylose levels is again challenging, but the prediction quality indexes are still acceptable for considering this approach robust enough for the purpose of this classification.

Table 5.9. Classification results for AHIMBU of NIR spectra, with the information about the amylose content. All the values refer to the quality indexes in prediction and are expressed in percentage.

	LVs	Sens	Spec	NER	Accuracy
High		74.6	88.2		81.6
Low	5	84.5	80.2	85.5	82.1
Waxy		97.4	97.5		97.5

The second step with AHIMBU implicated the classification of all the 47 varieties as singular classes themselves. Also in the case of NIR data, the threshold of Specificity(Cal) ≥ 80% was chosen, so that the model stopped once this rule was not respected. The final structure of this modelling has the shape displayed in **Figure 5.18**.

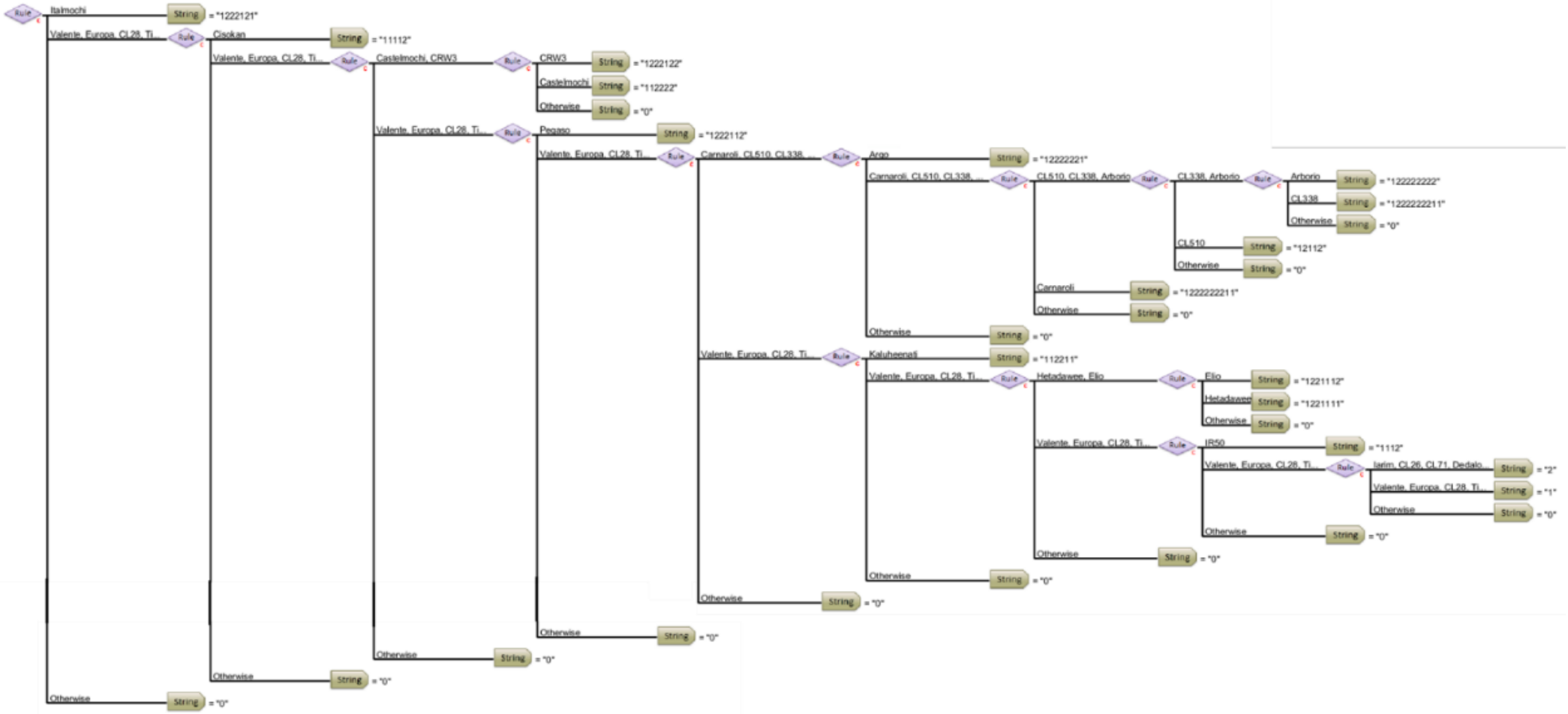

Figure 5.18. Structure of the AHIMBU for the NIR data, at the threshold of Specificity(Cal) ≥ 80%. Each sample represents a class itself.

Looking at this structure, it is possible to notice that the first four varieties that the model separates are the Italmochi, Cisokan, Castelmochi and CRW3. Apart from Cisokan, the other rice types are the three Waxy varieties present in this study. Italmochi is the first variety that is isolated from the totality of the samples as a single class itself, while Castelmochi and CRW3 are firstly grouped together and then separated one from another. This observation confirms once again what was previously seen: these rice types are detected also by the AHIMBU as deeply different from the others, as seen with AHIMBU on amylose content, but also with PLS-DA and also PCA. The results of this classification are resumed in **Table 5.10**.

Table 5.10. Classification results for AHIMBU of NIR spectra at the threshold of Specificity(Cal) ≥ 80%. All the values refer to the quality indexes in prediction and are expressed in percentage.

	Varieties	LVs	Sens	Spec	NER	Accuracy
Class 1	Baldo, CL12, CL15, CL18, CL28, CL31, CL33, CL35, CL80, Cripto, CRLB1, Cypress, Drago, Duilio, Europa, Fedearroz50, IAC165, IR6, Lince, Padano, Prometeo, Puma, S. Andrea, Selenio, Taichung Sen 17, Tiberio, Valente	4	61.7	96.8	54.3	76.7
Class 2	CL26, CL,71, Dedalo, Doongara, Iarim, IR64	4	42.3	97.1		90.1
Class 3	IR50	2	92.3	91.8		91.8
Class 4	Cisokan	3	100	96.1		96.2
Class 5	CL510	4	48.7	98.8		97.8
Class 6	Kaluheenati	2	0	99.2		97.1
Class 7	Castelmochi	6	84.6	98.2		97.9
Class 8	Hetadawee	4	66.7	98		97.3
Class 9	Elio	4	25.6	97.9		96.4
Class 10	Pegaso	2	74.4	88.8		88.5
Class 11	Italmochi	3	97.4	96.3		96.3
Class 12	CRW3	6	2.5	99.8		97.8
Class 13	Argo	6	82.1	97.2		96.8
Class 14	Arborio	2	20.5	99.9		98.2

Class 15	CL388	2	28.2	99.1		97.7
Class 16	Carnaroli	3	41	99.3		98.1

Observing the performances of this model, it seems that AHIMBU finds some criticalities when classifying the 47 varieties using the NIR data. This is particularly true for the Specificity, with globally low values, and consequently the Non-error rate is very low. On the other hand, the values of Specificity are high and all above 90%. This might suggest that this AHIMBU model can be used to predict very well which rice types do not belong to one class under examination: in the optic of fighting food fraud, this model might be suitable for analyses aimed at excluding rice types that do not belong to a selected variety.

The last step in the classification analyses based on NIR data involved the application of manual hierarchical models. Two PLS-DA models were calculated and used as rules of the hierarchical structure, which resulted to be the same shown in **Figure 5.19**. In **Table 5.11**, the results of this classification method are displayed.

Table 5.11. Classification results for manual hierarchical model of NIR spectra, with the information about the amylose content. All the values refer to the quality indexes in prediction. Values are expressed in percentage.

	LVs	**Sens**	**Spec**	**NER**	**Accuracy**
High		74.6	88.2		81.6
Low	5	84.5	80.2	85.5	82.1
Waxy		97.4	97.5		97.5

The first observation that can be highlighted from these results is that the prediction performances are exactly the same of those of AHIMBU shown in **Table 5.9**. This happens since these two approaches follow the same path in the construction of the model structure, using PLS-DA models to divide samples into groups, and the number of chosen LVs is the same. This outcome underlines that, in the case of a classification with a low number of classes, the choice in the use of one of these two approaches can be led by the rapidity of the automatic models that allow to obtain trustworthy results without the need for a preliminary construction of PLS-DA models to insert in the hierarchical structure, a step which is time-consuming and can lead to random accidental mistakes.

Finally, the rice samples data were also classified on the basis of the NIR data, considering each variety as a class, as previously seen with the AHIMBU approach. The dendrogram built in the exploratory step was exploited to choose how to group the varieties. Then, according to this division, PLS-DA models were built continuing to divide samples into classes until the threshold of Specificity ≥ 80% was reached. **Figure 5.19** shows the structure of the manual hierarchical model.

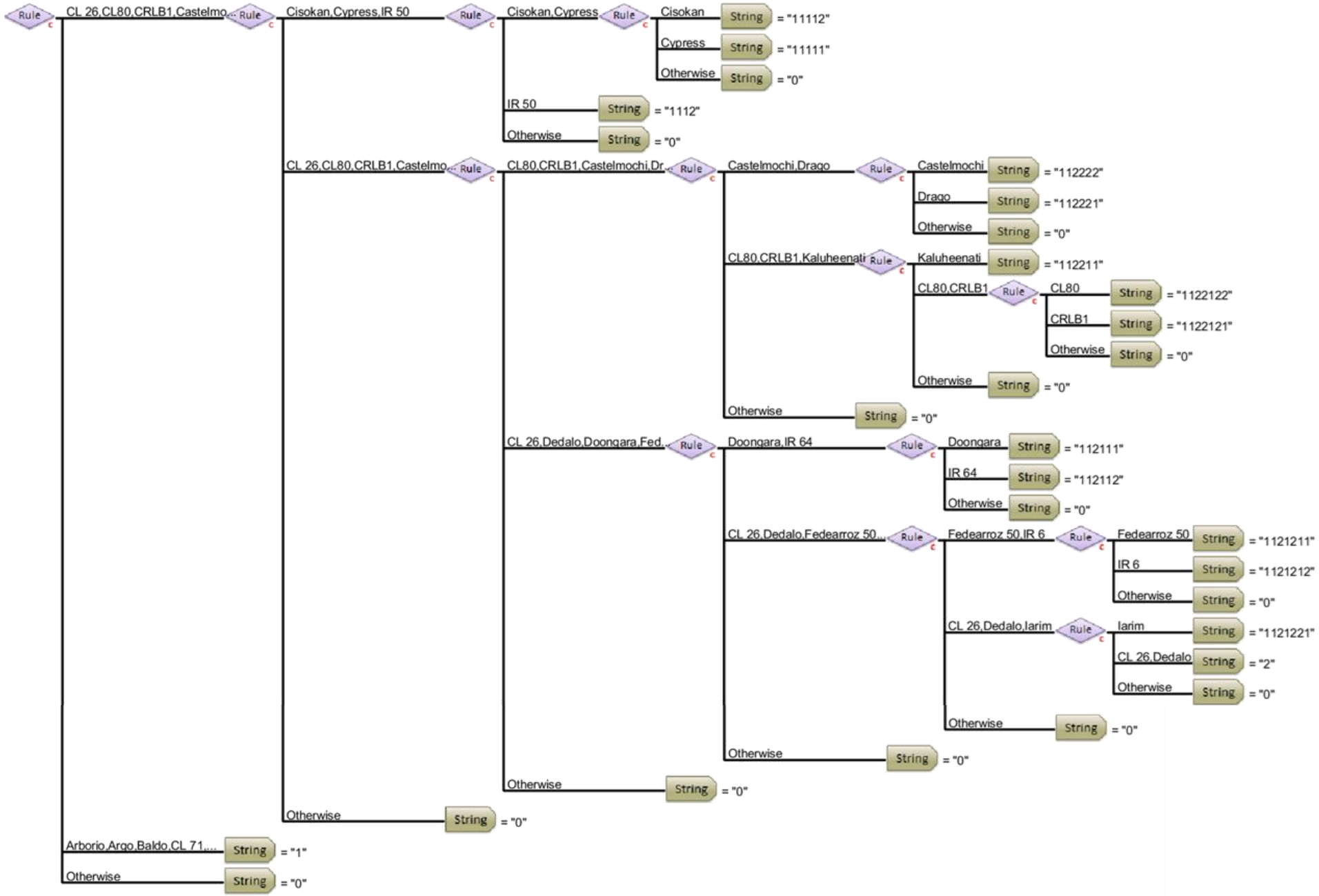

Figure 5.19. Structure of the manual hierarchical model for the NIR data, at the threshold of Specificity(Cal) ≥ 80%. Each sample represents a class itself.

This structure is slightly different from that obtained for the AHIMBU approach shown in **Figure 5.18**: here, the Waxy varieties are not the first to be separated: actually, the variety Italmochi, which proved to be the most particular in the exploratory and classification analyses, still lies within a group of varieties in Class 1, together with CRW3, while Castelmochi appears close to Drago variety. Other differences can be found observing the results of the manual hierarchical classification in **Table 5.12** below.

The values of Sensitivity are low also in this case, even if differently from AHIMBU, there are no zero values, and the Non-error rate is comparable for the two approaches. The Specificity is again high, but not all the values are above 90%, as seen with the automatic modelling. Globally, the performances of the two methodologies are comparable, even if AHIMBU seems to identify and classify the Waxy varieties consistently with the results obtain with the other analyses. It can be assumed that, in the view of these results, the automatic approach can be considered more advisable than the manual one as it saves a considerable amount of time and also works better in recognising the different classes under examination.

Table 5.12. Classification results for manual hierarchical model of NIR spectra at the threshold of Specificity(Cal) ≥ 80%. All the values refer to the quality indexes in prediction and are expressed in percentage.

	Varieties	LVs	Sens	Spec	NER	Accuracy
Class 1	Arborio, Argo, Baldo, Carnaroli, CL12, CL15, CL18, CL28, CL71, CL31, CL33, CL35, CL388, CL510, Cripto, CRW3, Duilio, Elio, Europa, Hetadawee, Italmochi, Lince, Padano, Pegaso, Prometeo, Puma, IAC165, S. Andrea, Selenio, Tiberio, Taichung Sen 17, Valente	5	85.6	82.9		84.7
Class 2	CL26, Dedalo	4	26.9	98.1		95
Class 3	IR50	5	87.2	98.1		97.8
Class 4	Cypress	3	84.6	97.5		97.2
Class 5	Cisokan	3	89.7	98.8		98.6
Class 6	Doongara	5	87.2	97.7	56.6	97.4
Class 7	IR64	5	82.1	97.4		97.1
Class 8	Kaluheenati	6	84.6	99.9		99.6
Class 9	Drago	6	10.7	99.3		97.4
Class 10	Castelmochi	6	61.5	98.2		97.4
Class 11	Fedearroz 50	5	10.3	99.6		97.7
Class 12	IR6	5	23.1	99.2		97.5
Class 13	Iarim	4	59	99.1		98.3
Class 14	CRLB1	5	35.9	98.4		97.1
Class 15	CL80	5	20.5	99.1		97.4

5.4.2.3. Hierarchical models for fused data

The first step in the classification of the data fusion matrix with the hierarchical approach involved the application of the AHIMBU approach exploiting the information about the shape of the rice grain. The structure of this model is the same shown in **Figure 5.12** for the same classification with the morphological features. In **Table 5.13**, the classification quality indexes for this model are displayed.

Table 5.13. Classification results for AHIMBU of fused data, with the information about the shape of the rice kernel. All values refer to the quality indexes in prediction and are expressed in percentage.

	LVs	Sens	Spec	NER	Accuracy
Long A from internal market	3	98.3	95.7	87.5	96.1
Long A for parboil		80.3	97.2		93.7
Long B	4	98.4	96.8		97.2
Medium	3	70	97.8		93.3
Round		90.3	97.8		96

Some criticalities are found for the "Medium" and "Long A for parboil" classes concerning the Sensitivity values, which proved to be the lowest for the five classes under examination. Even for the fused data, the classification of these two classes is a bit tricky for the same reasons explained previously. These results are comparable to those obtained from the same model using the morphological features only, with just a slight improvement in the classification of the "Round" varieties for the data fusion. These two approaches, i.e. the data fusion and the morphology, are both successful with respect to the rice classification according to the shape of the grain, therefore adding the NIR spectral information does not lead to an improvement in the classification of this type of data when using the automatic hierarchical classification.

Following the previous systematic way to proceed with data modelling, the manual hierarchical model should have also been built. Nevertheless, it was chosen not to proceed, as a consequence of the results obtained from the dendrograms in the exploratory step: the groups found by clustering were not well-defined, with a

particular misclassification of the “Medium” rice type, which made it impossible to obtain a reliable division of the samples according to the classes under examination. For this reason, the further step was to directly switch to the classification based on the data fusion matrix with respect to the amylose content information.

Even in this case, the structure of the AHIMBU was the same previously obtained when classifying the NIR data with amylose content as class information illustrated in **Figure 5.17**. The classification quality indexes of this AHIMBU approach are shown in **Table 5.14**.

Table 5.14. Classification results for AHIMBU of fused data, with the information about the amylose content. All the values refer to the quality indexes in prediction and are expressed in percentage.

	LVs	Sens	Spec	NER	Accuracy
High		77.2	92.9		85.2
Low	4	91.6	80.8	89.6	85.6
Waxy		100	99.5		99.5

The performances of the model proved to be good, with both high Sensitivity and Specificity for all the three classes. Moreover, if compared to the AHIMBU of NIR data on the amylose content of rice, the data fusion results proved to be better, with an incrementation in Specificity for the High class, in Sensitivity for the Low class, and in both the quality indexes for the Waxy class. This can suggest that the classification with the data fusion matrix can lead to stronger and more reliable models with respect to the use of NIR data only.

Subsequently, a manual hierarchical model was built. This manual model shares the same structure of the AHIMBU approach, and a visual representation of the manual hierarchical model can again be seen in **Figure 5.17**. The manual model performances are resumed in **Table 5.15**.

Table 5.15. Classification results for the manual hierarchical model of data fusion, with the information about the amylose content. All the values refer to the quality indexes in prediction and are expressed in percentage.

	LVs	Sens	Spec	NER	Accuracy
High	6	76.7	97.1	90.4	87.1
Low		95.3	81.2		87.4
Waxy	5	99.2	98.4		98.4

The performances of the two hierarchical models are comparable. There is an improvement in the performances of the Low class, which proved to be higher both in Sensitivity and Specificity for the manual model, while these quality indexes are slightly lower for the Waxy rice type. However, both the approaches work well and proved reliable, so the choice of using one approach instead of the other can be led by the rapidity of the AHIMBU approach.

References | Chapter 5

[1] V. Giuliana, M. Lucia, R. Marco, V. Simone, Environmental life cycle assessment of rice production in northern Italy: a case study from Vercelli, Int. J. Life Cycle Assess. 1 (2022) 1–18. https://doi.org/10.1007/S11367-022-02109-X.

[2] M. Arcieri, G. Ghinassi, Rice cultivation in Italy under the threat of climatic change: Trends, technologies and research gaps, Irrig. Drain. 69 (2020) 517–530. https://doi.org/10.1002/ird.2472.

[3] P.R. Chaudhari, N. Tamrakar, L. Singh, A. Tandon, D. Sharma, C.R. Prabha Chaudhari, Rice nutritional and medicinal properties: A review article, J. Pharmacogn. Phytochem. 7 (2018) 150–156. https://www.phytojournal.com/archives/2018.v7.i2.3233/rice-nutritional-and-medicinal-properties-a-review-article (accessed March 22, 2024).

[4] A. Moongngarm, N. Daomukda, S. Khumpika, Chemical Compositions, Phytochemicals, and Antioxidant Capacity of Rice Bran, Rice Bran Layer, and Rice Germ, APCBEE Procedia 2 (2012) 73–79. https://doi.org/10.1016/J.APCBEE.2012.06.014.

[5] N.K. Fukagawa, L.H. Ziska, Rice: Importance for Global Nutrition, J. Nutr. Sci. Vitaminol. (Tokyo). 65 (2019) S2–S3. https://doi.org/10.3177/JNSV.65.S2.

[6] L. Wang, D. Liu, H. Pu, D.-W. Sun, W. Gao, Z. Xiong, Use of Hyperspectral Imaging to Discriminate the Variety and Quality of Rice, (n.d.). https://doi.org/10.1007/s12161-014-9916-5.

[7] M. Sliwi´nskasliwi´nska-Bartel, D. Thorburn Burns, C. Elliott, Rice fraud a global problem: A review of analytical tools to detect species, country of origin and adulterations, Trends Food Sci. Technol. (2021). https://doi.org/10.1016/j.tifs.2021.06.042.

[8] R.D. Tille'vr, Image Analysis for Agricultural Processes: a Review of Potential Opportunities, J. Agric. Engng Res 50 (1991) 247–258.

[9] G. Elmasry, D.F. Barbin, D.W. Sun, P. Allen, Meat Quality Evaluation by Hyperspectral Imaging Technique: An Overview, Crit. Rev. Food Sci. Nutr. 52 (2012) 689–711. https://doi.org/10.1080/10408398.2010.507908.

[10] M. Kamruzzaman, G. ElMasry, D.W. Sun, P. Allen, Prediction of some quality attributes of lamb meat using near-infrared hyperspectral imaging and multivariate analysis, Anal. Chim. Acta 714 (2012) 57–67. https://doi.org/10.1016/J.ACA.2011.11.037.

[11] J. Qin, F. Vasefi, R.S. Hellberg, A. Akhbardeh, R.B. Isaacs, A.G. Yilmaz, C. Hwang, I. Baek, W.F. Schmidt, M.S. Kim, Detection of fish fillet substitution and mislabeling using multimode hyperspectral imaging techniques, Food Control 114 (2020) 107234. https://doi.org/10.1016/J.FOODCONT.2020.107234.

[12] H.J. He, D. Wu, D.W. Sun, Nondestructive Spectroscopic and Imaging Techniques for Quality Evaluation and Assessment of Fish and Fish Products, Crit. Rev. Food Sci. Nutr. 55 (2015) 864–886. https://doi.org/10.1080/10408398.2012.746638.

[13] J.H. Cheng, D.W. Sun, Hyperspectral imaging as an effective tool for quality analysis and control of fish and other seafoods: Current research and potential applications, Trends Food Sci. Technol. 37 (2014) 78–91. https://doi.org/10.1016/J.TIFS.2014.03.006.

[14] Y. Lu, Y. Huang, R. Lu, Innovative Hyperspectral Imaging-Based Techniques for Quality Evaluation of Fruits and Vegetables: A Review, Appl. Sci. 2017, Vol. 7, Page 189 7 (2017) 189. https://doi.org/10.3390/APP7020189.

[15] Y.Y. Pu, Y.Z. Feng, D.W. Sun, Recent Progress of Hyperspectral Imaging on Quality and Safety Inspection of Fruits and Vegetables: A Review, Compr. Rev. Food Sci. Food Saf. 14 (2015) 176–188. https://doi.org/10.1111/1541-4337.12123.

[16] N. Caporaso, M.B. Whitworth, I.D. Fisk, Near-Infrared spectroscopy and hyperspectral imaging for non-destructive quality assessment of cereal grains, Appl. Spectrosc. Rev. 53 (2018) 667–687. https://doi.org/10.1080/05704928.2018.1425214.

[17] K. Sendin, P.J. Williams, M. Manley, Near infrared hyperspectral imaging in quality and safety evaluation of cereals, Crit. Rev. Food Sci. Nutr. 58 (2018) 575–590. https://doi.org/10.1080/10408398.2016.1205548.

[18] G. Fox, M. Manley, Applications of single kernel conventional and hyperspectral imaging near infrared spectroscopy in cereals, J. Sci. Food Agric. 94 (2014) 174–179. https://doi.org/10.1002/JSFA.6367.

[19] J.M. Amigo, B.M. Jespersen, F. van den Berg, J.J. Jensen, S.B. Engelsen, Batch-wise versus continuous dough mixing of Danish butter cookies. A near infrared hyperspectral imaging study, Food Chem. 414 (2023) 135731. https://doi.org/10.1016/J.FOODCHEM.2023.135731.

[20] S. Selci, The Future of Hyperspectral Imaging, J. Imaging 2019, Vol. 5, Page 84 5 (2019) 84. https://doi.org/10.3390/JIMAGING5110084.

[21] L.H. Xie, S.Q. Tang, X.J. Wei, Z.H. Sheng, G.N. Shao, G.A. Jiao, S.K. Hu, Wang-Lin, P.S. Hu, Simultaneous determination of apparent amylose, amylose and amylopectin content and classification of waxy rice using near-infrared spectroscopy (NIRS), Food Chem. 388 (2022) 132944. https://doi.org/10.1016/J.FOODCHEM.2022.132944.

[22] S.R. Delwiche, R.A. Graybosch, Identification of Waxy Wheat by Near-infrared Reflectance Spectroscopy, J. Cereal Sci. 35 (2002) 29–38. https://doi.org/10.1006/JCRS.2001.0400.

[23] H. Zhang, Y. Miao, H. Takahashi, J.Y. Chen, Amylose Analysis of Rice Flour Using Near-Infrared Spectroscopy with Particle Size Compensation, Food Sci. Technol. Res 17 (2011) 361–367.

[24] P. Mishra, E.J. Woltering, Identifying key wavenumbers that improve prediction of amylose in rice samples utilizing advanced wavenumber selection techniques, Talanta 224 (2021) 121908. https://doi.org/10.1016/J.TALANTA.2020.121908.

[25] H. Bu¨ningbu¨ning-Pfaue, Analysis of water in food by near infrared spectroscopy, (n.d.). https://doi.org/10.1016/S0308-8146(02)00583-6.

[26] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion

methodologies for food and beverage authentication and quality assessment - a review, Anal. Chim. Acta 891 (2015) 1–14. https://doi.org/10.1016/J.ACA.2015.04.042.

[27] R.R. de Oliveira, C. Avila, R. Bourne, F. Muller, A. de Juan, Data fusion strategies to combine sensor and multivariate model outputs for multivariate statistical process control, Anal. Bioanal. Chem. 412 (2020) 2151–2163. https://doi.org/10.1007/S00216-020-02404-2/FIGURES/9.

[28] L. Yang, H. Gao, L. Meng, X. Fu, X. Du, D. Wu, L. Huang, Nondestructive measurement of pectin polysaccharides using hyperspectral imaging in mulberry fruit, Food Chem. 334 (2021) 127614. https://doi.org/10.1016/J.FOODCHEM.2020.127614.

[29] H. Zhang, S. Zhang, Y. Chen, W. Luo, Y. Huang, D. Tao, B. Zhan, X. Liu, Non-destructive determination of fat and moisture contents in Salmon (Salmo salar) fillets using near-infrared hyperspectral imaging coupled with spectral and textural features, J. Food Compos. Anal. 92 (2020) 103567. https://doi.org/10.1016/J.JFCA.2020.103567.

[30] C.N. Silla, A.A. Freitas, A survey of hierarchical classification across different application domains, Data Min. Knowl. Discov. 22 (2011) 31–72. https://doi.org/10.1007/S10618-010-0175-9.

[31] A.D. Gordon, A Review of Hierarchical Classification, J. R. Stat. Soc. Ser. A 150 (1987) 119. https://doi.org/10.2307/2981629.

[32] O. Kramer, K-Nearest Neighbors, (2013) 13–23. https://doi.org/10.1007/978-3-642-38652-7_2.

[33] H. Bhavsar, M.H. Panchal, A Review on Support Vector Machine for Data Classification, Int. J. Adv. Res. Comput. Eng. Technol. 1 (2012) 2278–1323.

[34] L. Marchi, I. Krylov, R.T. Roginski, B. Wise, F. Di Donato, S. Nieto-Ortega, J.F.Q. Pereira, R. Bro, Automatic hierarchical model builder, J. Chemom. 36 (2022). https://doi.org/10.1002/CEM.3455.

Chapter 6

General conclusions and future perspectives

The PhD path has for me represented a fundamental phase of both scientific and personal growth. On a personal level, I never expected this PhD journey to change me as much as it did. I've learnt how to collaborate and communicate effectively with colleagues, which often came from different backgrounds, how to stay resilient and face challenges with a problem-solving mindset, and how to keep going even in moments of high stress and uncertainty; I've also gained the confidence to stand in front of hundreds of people and present my work. I had the chance to live in a foreign city I had always dreamed of, stepping out of my comfort zone and embracing new experiences, I discovered the exciting and inspiring atmosphere of international conferences, where I had the chance to connect with brilliant minds. But most importantly, I met incredible people along the way, some of whom I can now call friends. This PhD has been more than just an academic challenge: it has been an evolutive path that strengthened my confidence and prepared me to tackle future challenges with determination and enthusiasm.

Concerning the scientific side, throughout my PhD I have gained a solid understanding of spectroscopic and chemometric techniques, learning how to combine their potential to improve the quality and the efficiency of analytical processes in the agri-food field. I have developed a strong interdisciplinary mindset, integrating knowledge from chemistry, physics, and data science to tackle challenges related to the world of food analysis. One of the most significant contributions of my work has been the implementation of Process Analytical Technology (PAT) in food analysis, demonstrating how real-time, non-destructive techniques can enhance process control, reduce waste, and improve product consistency. The adoption of these tools not only reduces the use of chemical reagents and analysis time but also opens new perspectives for real-time monitoring in the agri-food supply chain, improving traceability and food safety, in line with sustainability and circular economy principles. Moreover, by integrating NIR spectroscopy and chemometrics into industrial workflows, it was shown how advanced analytical technologies can facilitate data-driven decision-making, aligning with the core concepts of Industry 4.0.

In conclusion, with the present work, it was demonstrated the effectiveness of NIR spectroscopy coupled with chemometrics as a key tool for developing rapid, sustainable and cost-effective analytical methodology in the agri-food sector, to tackle the main criticalities that characterize this field. To face this challenge, three projects were introduced, and their results were discussed.

Based on the outcomes of the **NEWPOW project**, it can be stated that the methodology proposed in this Thesis demonstrated its usefulness in the achievement of robust regression models, which proved to be powerful in the prediction of the lipid content of hazelnuts. The best results were obtained using the benchtop MPA spectrometer in probe acquisition mode, as the regression figures in prediction were generally better with respect to the other techniques. The MPA sphere showed slightly worse performances in prediction, while the cross-validation step suggested that this model could be the most robust, with R^2_{CV} = 0.903, $RMSE_{CV}$ = 0.645, and RPD_{CV} = 3.254. In general, the regression parameters of the SCiO portable molecular sensor proved to be lower than the MPA benchtop instrument, suggesting that the robustness of the model obtained through this data was not enough to ensure reliable results when applied to analyse new samples.

Additional analyses on hazelnut samples would surely be very beneficial in the attempt to improve the performances of the SCiO portable spectrometer, as well as to further improve, test, and validate the performances of the MPA instrument in both acquisition modes, hopefully leading to more robust models (especially regarding the obtained RPD values). These implementations would significantly help the advancement of this study to a more developed stage of rapid and cost-effective in situ analyses, to achieve a robust classification as quality control. The concrete result of this project was a research paper [1 in Overview] published on Foods.

Concerning the **Help2Grow project**, the PCA analysis performed on the SCiO data provided important insights into the spectral variations of grape leaves affected by downy and powdery mildew. A first PCA model was built considering all the acquisitions from June 6th to October 26th, 2022. The results highlighted a clear distinction between samples collected in October and those from earlier months, suggesting that the analytical method effectively captured structural differences in the leaves due to seasonal deterioration. However, no discernible trends were observed regarding the type of treatment applied to the plants.

To further investigate the data, a separate analysis was conducted focusing exclusively on summer acquisitions (June 6th to August 1st, 2022), when leaves were not yet deteriorated by fall conditions. This PCA model revealed a significant trend among grape leaves affected by downy mildew: samples treated with fungicide clustered distinctly on the right side of PC1, suggesting that the treatment affected the spectral response of the leaves. This observation led to further analysis isolating only the downy mildew-affected samples, which confirmed that acquisition time influenced the PCA distribution. The progressive separation of treated samples over time suggests that the fungicide may have contributed to healing the plant,

distinguishing treated leaves from those left untreated or exposed to other treatments. On the other hand, no specific trends were identified among samples affected by powdery mildew. This is probably due to the particularly hot and dry summer conditions, which may have hindered the growth of the fungus responsible for the disease.

A further investigation using hyperspectral imaging (HSI) data was also carried out to evaluate whether this technique could provide additional insights into disease development or treatment effects. Unfortunately, no meaningful trends were observed: leaves subjected to the same treatment did not cluster together, indicating that the expected spectral correlations were not present. This lack of clear results can be attributed to several critical factors, including extended storage times and transport conditions, which probably contributed to sample deterioration and information loss.

Future research should consider optimizing storage and transportation methods to preserve sample integrity and allow for more reliable analyses. Moreover, conducting the study under normal seasonal conditions that promote fungal growth may enhance the ability to detect significant spectral variations associated with disease progression and treatment efficacy. Adjusting the experimental timeline to minimize delays in analysis could also improve the success of similar studies in the future.

The results of the **Rice HSI project** proved the efficacy of the NIR-Hyperspectral Imaging in the analysis and classification of 47 different rice varieties. The information contained in the images, extracted and explored through multivariate analysis tools, allowed to build classification models aimed at distinguishing samples both through their morphological and spectroscopic features, facing the critical issue of food fraud in the rice industry. To this end, three classification approaches were implemented: a first PLS-DA classification, followed by two hierarchical classification models, both manual and automatic. The latter proved to be able to handle a high number of classes, enhancing the classification process. All these models were initially used to categorize samples based on the shape grain morphology and the amylose content. Additionally, for the hierarchical models, another classification was developed using the type of variety as class information.

The PLS-DA models proved to be robust and effective, with both high Sensitivity and Specificity. The best classification performances were obtained using the data fusion approach, while the worst performances were found for the classification with the NIR data alone, indicating that integrating multiple data sources enhances the models' reliability. The hierarchical models also proved to classify well, with results comparable to the respective PLS-DA models. When analysing the morphology, the hierarchical models showed better performances when comparing Specificity; on the other hand, PLS-DA models showed higher Sensitivity, suggesting that this approach might be more suitable for targeted recognition of specific rice types, as it works very well in detecting true positives. In the case of NIR data, some criticalities were found concerning Sensitivity, suggesting that this

model can be exploited with the intention of recognising samples that do not correspond to a given reference variety, taking advance of the higher results reached with Specificity. The morphological data proved more effective than the NIR data for hierarchical classification of rice varieties, achieving higher performances both for Sensitivity and Specificity. Among the hierarchical models, the AHIMBU approach led to better results compared to the manual method, both in the cases of morphology and NIR data.

Concerning the fused data, combining rice grain morphology and amylose content showed to improve classification results. Although no hierarchical classification models of the rice varieties were built for data fusion due to time constraints, the promising results obtained suggest that implementing this model in future studies could enhance classification efficiency, ultimately improving counterfeit detection in the rice supply chain. The study on the fused data revealed a correlation between rice grain shape and chemical features, in particular for the “Long B” and “Long A” rice types, where the morphological characteristics were associated with specific amylose levels. This observation suggests that the rice shape can become an indicator of the chemical composition of some rice varieties, strengthening the potential of combining spectral and structural information to improve classification results. This correlation highlights the advantages of an integrated approach in rice authentication, enhancing the ability to distinguish rice varieties based on both structural and compositional attributes.

This study provides valuable insights into the application of NIR-Hyperspectral Imaging for food authentication, highlighting the potential of advanced multivariate approaches in agri-food analysis. The results demonstrate that manual and hierarchical models lead to comparable results, suggesting that automatic classification offers a time-efficient solution without compromising the model performances. Hierarchical models showed higher Specificity if compared to PLS-DA models: this can be due to their stepwise classification approach, which progressively reduces misclassification by refining decisions progressively. This classification process enhances the model ability to correctly exclude non-matching samples, consequently minimizing false positives. On the contrary, PLS-DA maximizes class separation globally, showing higher Sensitivity as it detects subtle differences among samples and assigns them more easily to a given class. However, this increases the risk of misclassification, leading to lower Specificity.

Future research can further optimize classification performance, making food authentication processes more robust and reliable. Given the impact of rice fraud on global markets, the ability to accurately classify and authenticate rice varieties using rapid and non-destructive techniques represents a significant step forward in food safety and quality control. A research paper detailing this study is currently in its final writing phase.

In conclusion, the contributions of this Thesis underline the growing importance of advanced analytical methods in the agri-food sector. By offering more sustainable, efficient, and cost-effective solutions to food analysis, the integration of NIR

spectroscopy and chemometrics offers new opportunities for industrial improvements. The present research provides a solid foundation for future studies, which will continue to increase these findings to further enhance food quality, in line with global sustainability goals.

Acknowledgements

First and foremost, I want to thank my supervisor Francesco, for having believed in my potential without knowing me in advance. Without your trust in me, this path would have never been possible.

I want to thank my Spanish team, Anna, Rodrigo and Adrian, for having made me feel warmly welcome in an unfamiliar place, and for having been friends beyond just being great teammates.

I want to thank Rosalba and Veronica from UNIMORE, for the assistance in the analyses with HSI, and for having enriched our collaboration with pleasant moments of shared wine and tigelle.

Thanks to Giusy and Federico, for having been my family from the moment I arrived in Turin, for the countless happy memories, and for remaining one of the biggest parts of my heart.

Thanks to Stefania, for being one of the strongest and smartest people I ever met in my life, for always standing by my side in my toughest moments, for being a steady presence in a world of uncertainty.

Thanks to Ilaria, for being with me for half of our lives, for tolerating my endless voice messages, for always finding time and energy for supporting me. You are an example of those special people that make life feel lighter.

Thanks to Giuliana, for having turned our house into our home, for your remarkable strength and empathy, for being a sister more than a flatmate.

Thanks to Guglielmo, Maria Laura and Tegitu, for having been my family in Barcelona, my reference point in an unknown place, and for still being a constant presence in my life despite the distance. Without you, Barcelona would have never felt as warm and meaningful.

Thanks to Spazio MovArt for having turned my life upside down, shattering the walls of my uncertainties and opening my mind and heart. I feel enriched not only by the work we did together, but by the connection I experienced with such an extraordinary group of people. I will always remember the impact that this journey had on me.