UNIVERSITÀ DEGLI STUDI
DI MODENA E REGGIO EMILIA

**Research Doctorate Programme in
"Food and agricultural science, technology and biotechnology"**

XXXVIII cycle

Chemometrics and *green* technologies for a low environmental impact agri-food system

Candidate: Veronica Ferrari

Tutor: Dr. Rosalba Calvini

Co-Tutor: Prof. Alessandro Ulrici

PhD Programme Coordinator: Prof. Fabio Licciardello

Riassunto

La spettroscopia nel vicino infrarosso (NIR), l'imaging digitale e spettrale si stanno sempre più affermando come metodi rapidi, non distruttivi e dal basso impatto ambientale per la valutazione della sicurezza e della qualità degli alimenti. Grazie alla facilità d'uso e alla mancanza di reagenti, tali metodi risultano più sostenibili ed economici rispetto alle tradizionali tecniche analitiche. Inoltre, la spettroscopia NIR permette un'analisi non mirata, fornendo una valutazione completa del profilo chimico distintivo di un campione senza la selezione a monte di composti di interesse.

L'imaging iperspettrale unisce i benefici della spettroscopia con quelli delle tecniche di imaging, grazie alla visualizzazione della distribuzione spaziale dei composti chimici di interesse sulla superficie del campione. Tuttavia, l'acquisizione di immagini iperspettrali prevede l'elaborazione di una grande mole di dati con conseguenti difficoltà computazionali, di gestione, di archiviazione e di analisi. Per superare tali limitazioni vengono spesso applicati metodi di *deep learning*, con ripercussioni negative a livello ambientale. In questo contesto, la chemiometria e l'analisi multivariata delle immagini rappresentano un'alternativa sostenibile.

Questa Tesi di Dottorato propone diverse strategie chemiometriche che prevedono l'utilizzo dell'imaging iperspettrale nel NIR (NIR-HSI) per aumentare la sostenibilità dei sistemi agroalimentari. A tal fine, vengono presentate soluzioni pratiche per il monitoraggio in campo, la cernita in post-raccolta e l'autenticazione attraverso metodi alternativi di riduzione della dimensionalità delle immagini e di classificazione.

Nell'ambito del progetto HALY.ID, sono state proposte diverse strategie per la gestione della cimice asiatica: dal monitoraggio in campo per prevenirne l'attività, all'impianto di cernita per valutare la qualità delle pere nel post-raccolta. Per il monitoraggio in campo il NIR-HSI è stato utilizzato per acquisire immagini di esemplari di cimice su diversi sfondi vegetali, utili a costruire una libreria spettrale per modelli di classificazione allo scopo di migliorare l'identificazione automatizzata in campo dell'infestante.

Per il controllo qualità in post-raccolta, il NIR-HSI è stato vagliato come metodo di cernita per rilevare punture di cimice asiatica su pere, che causano difetti interni alla polpa non rilevabili ad occhio nudo. Le pere biologiche (*cv.* Abate Fétel, *cv.* Williams), raccolte durante due estati consecutive, sono state suddivise in frutti esposti all'insetto e in frutti di controllo. Attraverso un approccio innovativo basato sulla conversione delle immagini in iperspettrogrammi e l'applicazione di modelli di classificazione abbinati alla selezione di *features* spaziali, è stato possibile annotare automaticamente le aree riconducibili alle punture di cimice. Grazie a questa fase, è stato possibile creare un dataset di spettri di riferimento per lo sviluppo di modelli di classificazione.

In tema di autenticità nel settore agroalimentare, le erbe aromatiche e le spezie sono tra i prodotti più soggetti ad adulterazione per fini economici. La ricerca condotta in questo ambito ha dimostrato che NIR-HSI può essere una valida tecnica di screening per differenziare l'origano autentico dall'origano sospettato di adulterazione, proponendo un metodo di classificazione alternativo per migliorare l'efficacia dell'autenticazione di classi fortemente sovrapposte.

Un'altra parte della ricerca esplora i vantaggi nell'applicare tecniche *sparse* nella dimensione spaziale, al fine di selezionare i pixel più rappresentativi di immagini iperspettrali. Come per le variabili spettrali, ciò può migliorare significativamente la compressione dei dati riducendo le esigenze di archiviazione e di calcolo.

Keywords: Alimenti, Chemiometria, Imaging, Spettroscopia NIR, Analisi multivariata

Abstract

Near-infrared (NIR) spectroscopy, digital imaging and spectral imaging have emerged as rapid, non-destructive, and environmentally friendly tools for assessing food safety and quality. Compared to traditional wet analytical techniques, these methods are more sustainable and cost-effective, as they don't require skilled personnel nor chemical reagents. Opposed to destructive targeted approaches, NIR spectroscopy enables untargeted analysis, providing a comprehensive assessment of the distinctive chemical profile of a sample without prior selection of specific compounds of interest.

Hyperspectral imaging combines the strengths of spectroscopy and imaging, allowing the visualization of spatial distribution of the chemical features of interest across a sample's surface. However, its data-richness requires longer computational times thus complicating data handling, storage and analysis. These limitations are typically overcome by applying deep learning methods resulting in enormous environmental implications. In this frame, Chemometrics and Multivariate Image Analysis are a sustainable alternative for addressing the curse of dimensionality and extracting useful information.

The present Doctoral Thesis proposes several strategies based on chemometrics and NIR hyperspectral imaging to increase the sustainability of the agri-food systems. To this aim, practical solutions for pest monitoring, post-harvest sorting, and for food authentication involving image dimensionality reduction and original classification methods are presented.

Within the HALY.ID project, several strategies were explored for Brown Marmorated Stink Bug (BMSB) management, *from field* pest monitoring *until post-harvest* fruit sorting. For pest monitoring, NIR hyperspectral images of BMSB specimens on different plant backgrounds were used to build a spectral library for pixel-level classification models aimed at enhancing automated BSMB in field detection.

For the detection of internal damages to the fruit pulp invisible to the naked eye, NIR hyperspectral imaging was evaluated as a viable method to implement a post-harvest sorting system for the early detection of BMSB punctures on pears. Organic pears (*cv.* Williams and *cv.* Abate Fétel) were collected in the summers of 2022 and 2023, with half of the fruits exposed to BMSB and the remainder used as controls. Approximately 2000 hyperspectral images were acquired for each variety. An innovative approach based on hyperspectrograms and image-level classification coupled with spatial features selection was adopted to automatically annotate BMSB punctures on the collected images. Thanks to this annotation step, a library of representative spectra belonging to both punctures and sound fruits was then built for the development of classification models.

Concerning authenticity issues in agri-food, herbs and spices are frequently subjected to Economically Motivated Adulterations, which are favoured by the complexity of the supply chain. Within this framework, NIR hyperspectral imaging was evaluated as a possible screening technique to differentiate authentic oregano from oregano suspected of adulteration with leaves of plants of lower commercial value. An alternative classification method is proposed to properly address authentication issues involving strongly overlapping classes.

Another part of the research investigated the benefits of applying sparse-based methods in the spatial direction for the selection of representative pixels of hyperspectral data. As with spectral variables, this method may significantly improve data compression by focusing only on the most informative spatial regions, thus reducing storage and computational demands.

Keywords: Food, Chemometrics, Spectral Imaging, NIR Spectroscopy, Image Analysis

Table of Contents

Chapter 1
Introduction

1.1. Background and Aim

Food analysis is fundamental to the food industry and to all sectors linked to food production and control. It mainly focuses on evaluating key product attributes, including quality, safety, nutritional value, sensory characteristics, and stability. While food safety is a straightforward concept, food quality encompasses a broader set of objective and subjective characteristics related to sensory, stability and health attributes, meeting both consumer expectations and regulatory requirements. Consequently, analytical evaluation of food is an essential component of quality management, as products must be monitored during formulation, manufacturing, and even after market release [1].

Food analysis comprises the development, application, and evaluation of analytical techniques used to characterize food and agricultural products. These techniques provide information on structural attributes, composition, physicochemical properties, and sensory characteristics, indispensable for product development, quality control, troubleshooting, and responding to consumer complaints.

However, the rapid growth of the global population, combined with increasing consumer awareness regarding food safety, origin, and authenticity, has intensified the demands placed on food quality management systems. Traditional analytical methods employed for assessing food quality, such as gas chromatography, high performance liquid chromatography and mass spectrometry, while highly accurate, are destructive, require skilled operators, rely on chemical reagents, and involve labour-intensive procedures. As a result, they are time-consuming, costly from both economic and environmental perspectives, and generally incompatible with real-time or in-line monitoring.

As an alternative, near-infrared (NIR) spectroscopy has emerged as a rapid, non-destructive, and environmentally sustainable technique for assessing food safety and quality. NIR spectroscopy is now a well-established analytical method which is widely implemented as process analytical technology (PAT) at multiple critical points in the production chain, given its cost effectiveness, speed, reproducibility, and accuracy [2–5]. Nevertheless, NIR spectrometers provide limited spatial resolution since measuring a small area of the analysed samples' surface, offering no information regarding the location of the constituent or contaminant investigated. This aspect is crucial for many food inspection applications when characterizing the composition of heterogeneous matrices, detecting adulterations or localized defects across the samples' surface [6]. To address these limits, Hyperspectral imaging (HSI) emerged as a powerful tool for non-destructive inspection of food

matrices, thanks to the possibility of combining the advantages of spectroscopic measurements with those of imaging techniques [7–9]. HSI enables the acquisition of spatial information across numerous contiguous spectral bands, producing images in which each pixel corresponds to a full spectrum and it is associated with a precise location on the sample surface [10,11]. Despite its potential, the high costs and sensitivity of optical components, together with considerable computational demands required to manage the large volume of data generated, still limit its widespread deployment as PAT [12]. Consequently, many food-related studies aim to identify a reduced set of informative wavelengths that can be used to develop multispectral imaging systems [13–15]. Multispectral imaging (MSI) relies on the acquisition of images including only fewer wavebands relevant to the specific problem-at-hand, thus resulting in faster and cheaper solutions [15,16].

In general, spectral imaging systems record enormous amount of data, posing issues related to data handling, storage and computation. In this framework, Chemometrics and Multivariate Image Analysis are mandatory to address the *curse of dimensionality* and to efficiently extract meaningful information from such high dimensional data [11,17].

Chemometrics is defined as the use of mathematical and statistical techniques for extracting relevant information from analytical data by identifying patterns and associating them with the chemical and physical properties under investigation [18]. Chemometric methods enable the exploration of image structure and the development of classification models for predicting qualitative attributes, as well as calibration models for quantifying parameters of interest. These tasks heavily rely on dimensionality reduction and feature selection strategies, which ease the extraction of the spectral and/or spatial features most relevant to the problem under investigation.

The research activities carried out in this doctoral Thesis focused on applying chemometric strategies and NIR-Hyperspectral Imaging (NIR-HSI) to enhance the sustainability of the agri-food systems. Specifically, practical solutions were developed for diverse objectives, including pest monitoring, post-harvest fruit sorting, and product authentication. The issues addressed in this Thesis were faced by means of three Multivariate Image Analysis approaches:

- *Feature selection*, allowing the identification and selection of spectral (wavelengths) or spatial (pixels) features relevant to the problem at hand. In particular, sparse-based methods were considered not only as a spectral variable selection strategy [19–21], but they were also applied row-wise to investigate their potential for the selection of spatial features [22].
- *Feature extraction*, involving the conversion of each spectral image into a one-dimensional signal, thereby performing dimensionality reduction of large image datasets enabling rapid chemometric analysis. These signals, referred to as hyperspectrograms, encode quantities derived

from Principal Component Analysis (PCA) [23,24]. The resulting matrix of hyperspectrograms can then be used for exploratory analysis and for the development of calibration or classification models.

- *Classification by Soft PLS-DA*, a soft discriminant variant of PLS-DA that incorporates additional constraints to perform class assignment of new observations [25]. This configuration combines the advantages of Class Modelling and Discriminant Analysis methods, allowing at the same time both outlier detection and the maximization of differences between the modelled classes. The robustness and flexibility of Soft PLS-DA were assessed across several applications, including authentication issues which were by far the most challenging.

1.2. Thesis Overview

The present thesis is organized as follows:

Chapter 2 presents a state-of-the-art review of NIR spectroscopy and spectral imaging applications within the agri-food context, along with the key principles underlying spectral data acquisition. It further describes Chemometric and Multivariate Image Analysis methods, with a particular focus on *feature selection* and *feature extraction* strategies for high-dimensional spectral image datasets.

Chapter 3 discusses the potential of spectral imaging combined with several chemometric approaches for the management of the Brown Marmorated Stink Bug (BMSB) invasive pest. From field, NIR spectral imaging was assessed as a cutting-edge technology enabling automated pest monitoring while, to preserve pears quality during post-harvest, the damages caused by BMSB feeding were detected using the same technology.

Chapter 4 shows the advantages of soft discriminant methods to handle food authentication problems when the classes of interest strongly overlap. Oregano authentication was considered as a case study, as this herb is frequently subjected to adulteration with leaves of plants with lower commercial values, hardly detectable due to their slight chemical differences.

Chapter 5 illustrates the benefits of applying sparse based selection in the spatial domain, enabling the extraction of a reduced subset of pixels, i.e., Regions of Interest (ROIs), within an image.

Chapter 6 draws general conclusions derived from the results of this research work.

The majority of the results reported in the thesis are presented in preprint format of the following articles:

- Ferrari, V., Calvini, R., Boom, B., Menozzi, C., Rangarajan, A.K., Maistrello, L., Offermans, P., Ulrici, A. (2023). Evaluation of the potential of near infrared hyperspectral imaging for monitoring the invasive brown marmorated stink bug, *Chemometrics and Intelligent Laboratory Systems*, 234, 104751;
- Ferrari, V., Calvini, R., Menozzi, C., Costi, E., Giannetti, D., Offermans, P., Maistrello, L., Ulrici, A. (2025). NIR hyperspectral imaging to identify damage caused by *Halyomorpha halys* on pears: Automated identification of Regions of Interest related to punctured areas, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 343, 126543;
- Ferrari, V., Calvini, R., Menozzi, C., Costi, E., Offermans, P., Maistrello, L., Ulrici, A. NIR hyperspectral imaging to identify damage caused by *Halyomorpha halys* on pears: development of classification models, submitted for publication;
- Ferrari, V., Calvini, R., Menozzi, C., Ulrici, A., Bragolusi, M., Piro, R., Tata, A., Suman, M., Foca, G. (2024). Addressing adulteration challenges of dried oregano leaves by NIR HyperSpectral Imaging, *Chemometrics and Intelligent Laboratory Systems*, 249, 105133.

References

[1] S.S. Nielsen, Introduction to Food Analysis, in: B.P. Ismail, S.S. Nielsen (Eds.), Nielsen's Food Analysis, Springer International Publishing (2024), pp. 3–14. https://doi.org/10.1007/978-3-031-50643-7_1.

[2] S. Grassi, C. Alamprese, Advances in NIR spectroscopy applied to process analytical technology in food industries, *Current Opinion in Food Science* 22 (2018) 17–21. https://doi.org/10.1016/j.cofs.2017.12.008.

[3] E. Skibsted, S.B. Engelsen, Spectroscopy for Process Analytical Technology (PAT), *Encyclopedia of Spectroscopy and Spectrometry* (2010) 2651–2661. https://doi.org/10.1016/B978-0-12-374413-5.00026-9.

[4] B.M. Nicolaï, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K.I. Theron, J. Lammertyn, Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review, *Postharvest Biology and Technology* 46 (2007) 99–118. https://doi.org/10.1016/j.postharvbio.2007.06.024.

[5] M. Manley, Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials, *Chemical Society Reviewes* 43 (2014) 8200–8214. https://doi.org/10.1039/C4CS00062E.

[6] X. Yang, P. Berzaghi, Near-Infrared Spectroscopy, in: A.M. Jiménez-Carvelo, A. Arroyo-Cerezo, L. Cuadros-Rodríguez (Eds.), Non-Invasive and Non-Destructive Methods for Food Integrity, Springer Nature Switzerland, Cham, 2024: pp. 41–59. https://doi.org/10.1007/978-3-031-76465-3_3.

[7] N.C. Basantia, L.M.L. Nollet, M. Kamruzzaman, eds., Hyperspectral Imaging Analysis and Applications for Food Quality, 1st ed., CRC Press, 2018. https://doi.org/10.1201/9781315209203.

[8] D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O.L. García-Navarrete, J. Blasco, Recent Advances and Applications of Hyperspectral Imaging for Fruit and Vegetable Quality Assessment, *Food and Bioprocess Technology* 5 (2012) 1121–1142. https://doi.org/10.1007/s11947-011-0725-1.

[9] D. Wu, D.-W. Sun, Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part I: Fundamentals, *Innovative Food Science & Emerging Technologies* 19 (2013) 1–14. https://doi.org/10.1016/j.ifset.2013.04.014.

[10] J. Burger, P. Geladi, Hyperspectral NIR image regression part II: dataset preprocessing diagnostics, *Journal of Chemometrics* 20 (2006) 106–119. https://doi.org/10.1002/cem.986.

[11] J. Burger, A. Gowen, Data handling in hyperspectral image analysis, *Chemometrics and Intelligent Laboratory Systems* 108 (2011) 13–22. https://doi.org/10.1016/j.chemolab.2011.04.001.

[12] D. Tanzilli, M. Cocchi, J.M. Amigo, A. D'Alessandro, L. Strani, Does hyperspectral always matter? A critical assessment of near infrared versus hyperspectral near infrared in the study of heterogeneous samples, *Current Research in Food Science* 9 (2024) 100813. https://doi.org/10.1016/j.crfs.2024.100813.

[13] R. Calvini, J.M. Amigo, A. Ulrici, Transferring results from NIR-hyperspectral to NIR-multispectral imaging systems: A filter-based simulation applied to the classification of Arabica and Robusta green coffee, *Analytica Chimica Acta* 967 (2017) 33–41. https://doi.org/10.1016/j.aca.2017.03.011.

[14] R. Calvini, A. Ulrici, J.M. Amigo, Growing applications of hyperspectral and multispectral imaging, in: *Data Handling in Science and Technology*, Elsevier, 2019: pp. 605–629. https://doi.org/10.1016/B978-0-444-63977-6.00024-9.

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

[16] J.M. Amigo, S. Grassi, Configuration of hyperspectral and multispectral imaging systems, in: *Data Handling in Science and Technology*, Elsevier, 2019: pp. 17–34. https://doi.org/10.1016/B978-0-444-63977-6.00002-X.

[17] A. Gowen, C. Odonnell, P. Cullen, G. Downey, J. Frias, Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, *Trends in Food Science & Technology* 18 (2007) 590–598. https://doi.org/10.1016/j.tifs.2007.06.001.

[18] P. Geladi, H. Isaksson, L. Lindqvist, S. Wold, K. Esbensen, Principal component analysis of multivariate images, *Chemometrics and Intelligent Laboratory Systems* 5 (1989) 209–220. https://doi.org/10.1016/0169-7439(89)80049-8.

[19] M.A. Rasmussen, R. Bro, A tutorial on the Lasso approach to sparse modeling, *Chemometrics and Intelligent Laboratory Systems* 119 (2012) 21–31. https://doi.org/10.1016/j.chemolab.2012.10.003.

[20] R. Calvini, A. Ulrici, J.M. Amigo, Sparse-Based Modeling of Hyperspectral Data, in: *Data Handling in Science and Technology*, Elsevier, 2016: pp. 613–634. https://doi.org/10.1016/B978-0-444-63638-6.00019-X.

[21] P. Filzmoser, M. Gschwandtner, V. Todorov, Review of sparse methods in regression and classification with application to chemometrics, *Journal of Chemometrics* 26 (2012) 42–51. https://doi.org/10.1002/cem.1418.

[22] R. Bro, E.E. Papalexakis, E. Acar, N.D. Sidiropoulos, Coclustering—a useful tool for chemometrics, *Journal of Chemometrics* 26 (2012) 256–263. https://doi.org/10.1002/cem.1424.

[23] R. Calvini, G. Foca, A. Ulrici, Data dimensionality reduction and data fusion for fast characterization of green coffee samples using hyperspectral sensors, *Analytical and Bioanalytical Chemistry* 408 (2016) 7351–7366. https://doi.org/10.1007/s00216-016-9713-7.

[24] C. Ferrari, G. Foca, A. Ulrici, Handling large datasets of hyperspectral images: Reducing data size without loss of useful information, *Analytica Chimica Acta* 802 (2013) 29–39. https://doi.org/10.1016/j.aca.2013.10.009.

[25] R. Calvini, G. Orlandi, G. Foca, A. Ulrici, Development of a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging, *Journal of Spectral Imaging* (2018) a13. https://doi.org/10.1255/jsi.2018.a13.

Chapter 2
Spectral Imaging in agri-food systems

2.1. Spectral Imaging in food analysis

2.1.1. NIR spectroscopy

NIR spectroscopy emerged as a rapid, non-destructive and versatile optical technique evaluating heterogeneous organic matrices, with numerous applications in the feed, food, pharmaceutical and many other product sectors. NIR spectroscopy is a well-established technique for multi-constituent analysis of various food matrices [1,2], widely used for quality control and assurance.

Although its discovery is attributable to William Herschel at the beginning of the nineteenth century, NIR spectroscopy has long lying idle due to the complex interpretation of the resulting spectra, characterized by broad, highly overlapped and unresolved bands. NIR spectroscopy breakthrough for the analysis of agricultural products dates back to the 1960s, mainly thanks to Karl Norris and co-workers, who demonstrated that proper statistical tools could help the interpretation of relevant NIR spectral regions for the analysis of grains [3]. Afterwards, the rapid improvements in terms of instrumentation and computing power have created the perfect environment for the development of many applications and a widespread use by the agrifood industry.

The NIR region of the electromagnetic spectrum covers the wavelength range between 780 and 2500 nm (12,821-4000 cm^{-1}). As the name suggests, NIR spectroscopy is based on the physical interaction between light in the infrared region and molecular bonds, mainly involving “hydrogen containing groups” like O-H, N-H and C-H. When infrared light interacts with a sample, part of this light is absorbed by the molecules, while the remaining light is either reflected or transmitted. The absorption bands falling in the NIR region are mainly related to combinations of fundamental vibrations and to overtones of molecular bonds involving pairs of atoms with dissimilar mass and a sufficiently large dipole moment. As the absorption is related to the vibrational status of each dipole molecule, NIR spectroscopy is then classified as a vibrational spectroscopy.

Although the NIR spectrum primarily reflects the chemical and physical characteristics of a sample, it is also strongly influenced by environmental factors. Nonetheless, alongside the following key characteristics, the versatility of NIR spectroscopy remains a key factor enabling its wide applicability across diverse analytical contexts [4–6]:

- *non-destructive*: NIR spectroscopy is based on matter-light interaction, without alteration and changes of the original sample. Conversely to traditional wet analytical methods which require the

destruction of the sample, NIR spectroscopy enables the analysis of food products without compromising their integrity. This is particularly important both for on-line monitoring and when dealing with valuable or limited quantities of samples;

- *rapid analysis*: point-wise NIR spectroscopy instruments provide quick results, making it ideal for real-time quality control in food production processes. Regardless of the type of instrument (i.e., handheld, benchtop, etc.), the time required for analysis is in the order of few seconds or even milliseconds;
- *minimal or no sample preparation required*: intact samples can be analysed without prior dilution or particular preparation; this is particularly valuable in food processing environments where products need to be analysed very rapidly as they are;
- *untargeted analysis*: NIR spectroscopy entails a comprehensive examination of a sample's distinctive chemical profile, effectively providing its chemical "fingerprint" without requiring prior knowledge of specific compounds of interest. On the other hand, due to the sensitivity of the technique, it is recommended for the determination of major components;
- *untrained personnel*: in contrast to wet analytical techniques, highly skilled personnel is not needed for sample preparation or spectra acquisition. On the other hand, proper statistical methods are needed for data analysis in order to obtain interpretable results;
- *multiparameter analysis*: as NIR spectroscopy simultaneously captures the absorption across different wavelengths, advanced statistical techniques can be employed to predict multiple parameters, including moisture, fat, protein content, among others. Beyond determining compositional parameters in food, NIR spectroscopy can also be used for the determination of complex quality properties such as texture and sensory attributes.

Its rapid, non-destructive nature, coupled with the ability to simultaneously quantify multiple chemical constituents, makes NIR spectroscopy particularly suitable for on-line or at-line quality control. Thanks to these characteristics, NIR spectroscopy has progressively evolved from a laboratory analytical tool into a reliable process analytical technology (PAT) suitable for routine use in the agri-food sector. In this context, it is crucial to follow a Quality by Design approach, whose main purpose is to ensure the quality of the product by employing statistical, analytical and risk-management strategies during the whole supply-chain [7–9].

From the industrial point-of-view, the quality assessment of the composition of cereals, milk and vegetable oils using NIR spectrometers at-line or on-line has been applied successfully, replacing wet chemistry [10–13]. However, despite its recognized potential, standardized procedures for sampling, control, measurement, data acquisition and development of predictive models for the determination

of analytical parameters of foodstuff and related materials are still lacking. To address this gap, UNI—the Italian Standards Organization—has recently begun filling this regulatory vacuum in the agri-food sector by establishing technical standards for both quantitative and qualitative NIR applications [14,15].

The relevance of point-wise NIR spectroscopy applications in the agri-food sector is also reflected in the growing volume of research conducted in the last twenty-five years. **Figure 2.1** shows the trend of the production of publications (i.e., articles, conference papers, reviews, book chapters, books, conference reviews and editorials) related to NIR Spectroscopy alone (NIRS) and to its application as PAT in the agri-food sector. The published documents related only to NIRS include the terms "NIR", "Near Infrared spectroscopy", "Near-infrared spectroscopy", "NIRS" or "NIR spectroscopy", alongside "food" or "beverage" in the article title, abstract and keywords. For the research of documents related specifically to NIRS application as PAT, the terms "industrial scale", "process monitoring", "process analytical technology", "in-line", "on-line" or "PAT" were added. To evaluate only the applications involving point-wise NIRS, the terms "imaging" and "image*" were excluded.

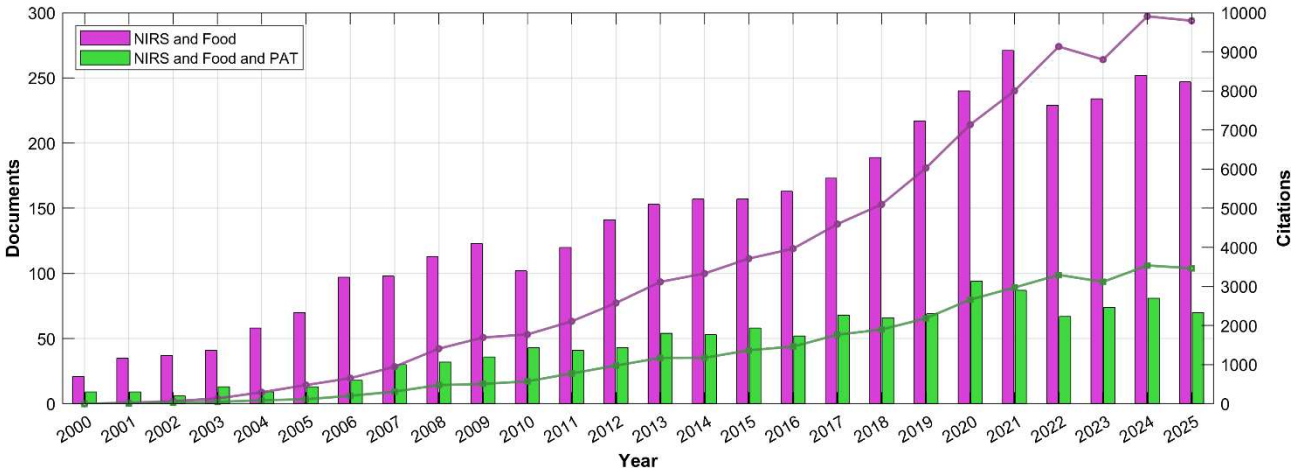


***Figure 2.1** Production of documents related to "**NIRS and Food**" and "**NIRS and Food and PAT**" in the last twenty-five years along with the citations trend. The bars correspond to the number of published documents while the solid lines correspond to the citations registered over the years.*

The plot shows the number of total publications registered for each year, along with the corresponding number of citations. In total, 3738 documents have been found for "NIRS and Food", of which 1195 are specifically related to its applications at the industrial scale or process monitoring ("NIRS and Food and "PAT"). In particular, for "NIRS and Food" both production and citation of publications exhibit a constant upward trend through 2025, reaching a maximum peak of 271 publications and

9909 citations in Scopus. The production and citation of publications related to "NIRS and Food and "PAT" show a moderate ascending trend, which may be due to the lack of common standardized guidelines and/or trade secrets.

For a global assessment of the topics of main co-occurrence related to the "NIRS and Food" bibliographic research on Scopus, the most frequently chosen keywords were visualized using VOSviewer (**Figure 2.2**). As shown in **Figure 2.2 a**, NIRS covers a broad range of different applications in the agri-food industry, from the determination of chemical composition (e.g., moisture, protein, fat, starch) enclosed in the green cluster, the monitoring of textural or functional properties (e.g., freshness and ripeness or maturity indices like soluble solids and sugars) relegated in the red cluster along with fruit quality, to safety assessment and authentication issues (e.g., shelf life evaluation, detection of adulterants, contaminants, fraud and spoilage-related changes) [16–20].

Another crucial aspect is related to data analysis, whose related keywords are mainly clustered in blue. In **Figure 2.2 b**, the overlay visualization displays a glaring increase of the terms "chemometrics" and "machine learning" in the last five years, especially opposed to the keywords related to chemical composition. Indeed, the application of proper chemometric strategies and machine learning arose with the usage of advanced instrumentation, enhancing the extraction of meaningful information from NIR spectra regarding molecular and physical structure of samples. In this frame, multivariate data analysis seem reliable to address both quantitative (i.e., determination of a specific analyte in a given matrix by regression analysis) and qualitative (i.e., classification of samples into different categories based on the specific qualitative chemical characteristics) matters [21–24].

Another keyword summarizing NIRS application in the agri-food sector are "quality" and "quality control". The global objectives for quality assessment are quite broad as they can be related to authentication based on species, variety or geographical origin, monitoring of on-going fermentation or ripening parameters (e.g., lactic acid fermentation in dairy products, acetic acid fermentation in vinegars), freshness and shelf life of liable to spoilage products such as fish, meat or fresh-cut fruits and vegetables, in multiple critical points of the production process. In addition to quality, food safety is essential to perform a reliable assessment of human health risks. In this category, the primary objectives include the detection of contaminants, adulteration with prohibited materials, or defects and spoilage caused by pathogens or pests.

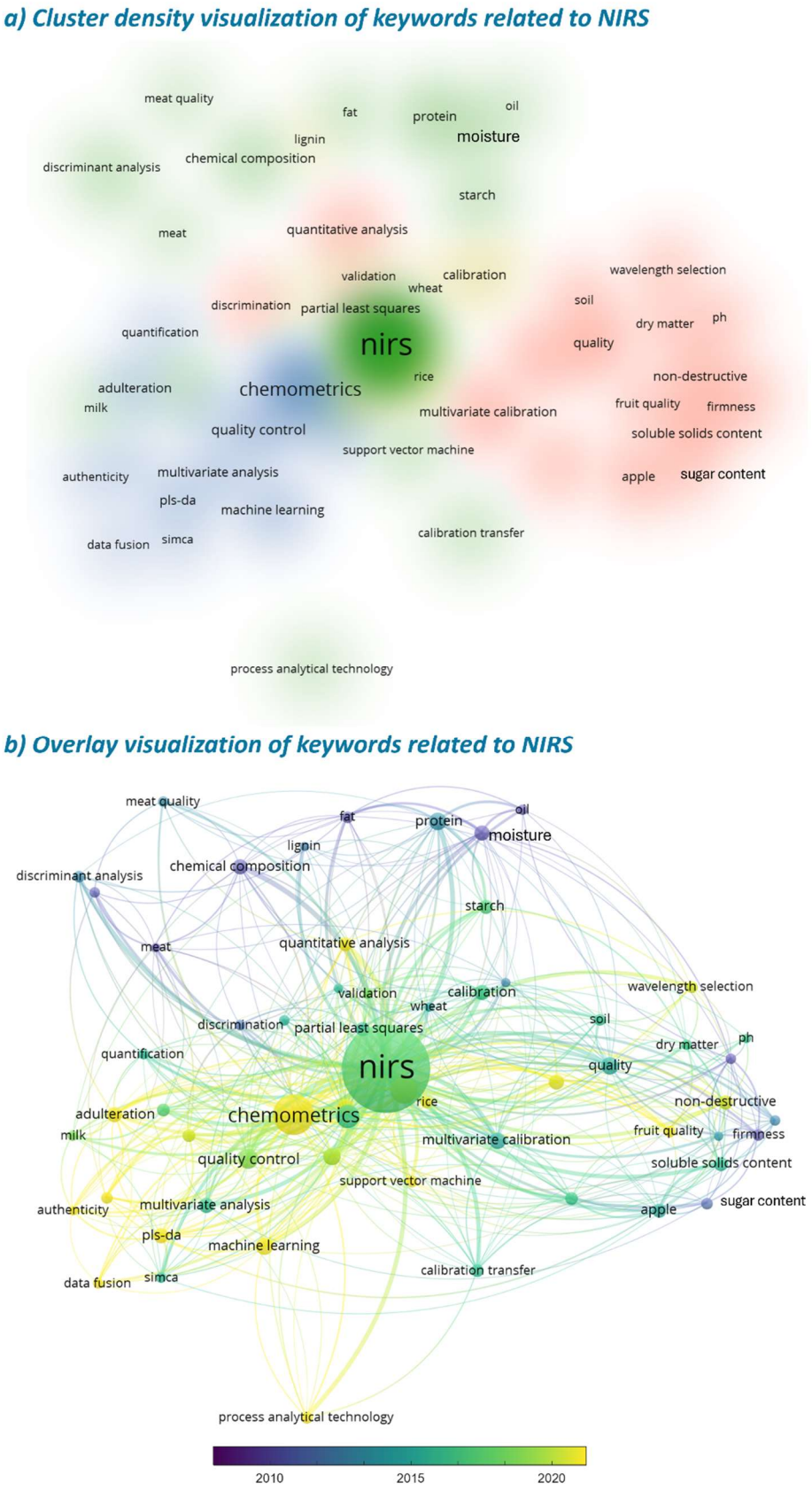


Figure 2.2 *Overview of the most cited keywords related to the "NIRS and Food" bibliographic research. In a) cluster density visualization, the keywords are coloured by topic and the colour intensity depends on frequency of citing, while in b) overlay visualization, the keywords are coloured by publication year.*

2.1.2. Spectral Imaging

Despite its numerous advantages, NIR spectroscopy instruments are based on point-wise measurements, meaning that these devices have a fixed-size measuring window. Therefore, the point-wise configuration is able to provide a single measurement of the spot or area within the measuring window [4]. This severely limits the application on extremely heterogeneous samples, where the spatial distribution of chemicals across the sample surface could help in the identification of different materials or substances.

In contrast, conventional imaging and computer vision systems are widely used in on-line processing to perform an objective evaluation of visible attributes of heterogeneous samples. However, as they operate in the visible range, conventional imaging systems are only able to grasp chemical differences related to colour, i.e., visible to the naked eye.

In this scenario, *spectral imaging* systems, which combine the advantages of both spectroscopy techniques and imaging systems, are gaining an increasing relevance as they are able to retrieve spatial and spectral information from the analysed products. These methods are emerging as one of the key elements of process control within the concept of Industry 4.0 [25,26].

Spectral images are three-dimensional (3-D) matrices composed of one spectral ($\boldsymbol{\lambda}$) and two spatial ($\mathbf{x, y}$) dimensions, obtained by stacking together grey-scale images acquired at different wavelengths. Therefore, this technique combines the strengths of spectroscopy and imaging, allowing the visualization of chemical distribution across a sample's surface. As reported by [27], "if conventional imaging tries to answer the question ***where?*** and point-wise spectroscopy tries to answer the question ***what?***, then spectral imaging tries to answer the question ***where is what?***".

Depending on the characteristics and resolution of the spectral dimension, spectral imaging can be divided into two main techniques: *multispectral imaging* (MSI) and *hyperspectral imaging* (HSI).

M*ultispectral images* are 3-D matrices composed of grey-scale images acquired at discrete and specific wavebands (**Figure 2.3 a**). The wavebands implemented on a multispectral imaging system are sensitive to the features of interest for a defined application [28,29], generally determined from a tailored solution of a previous variable selection step performed throughout hyperspectral imaging systems.

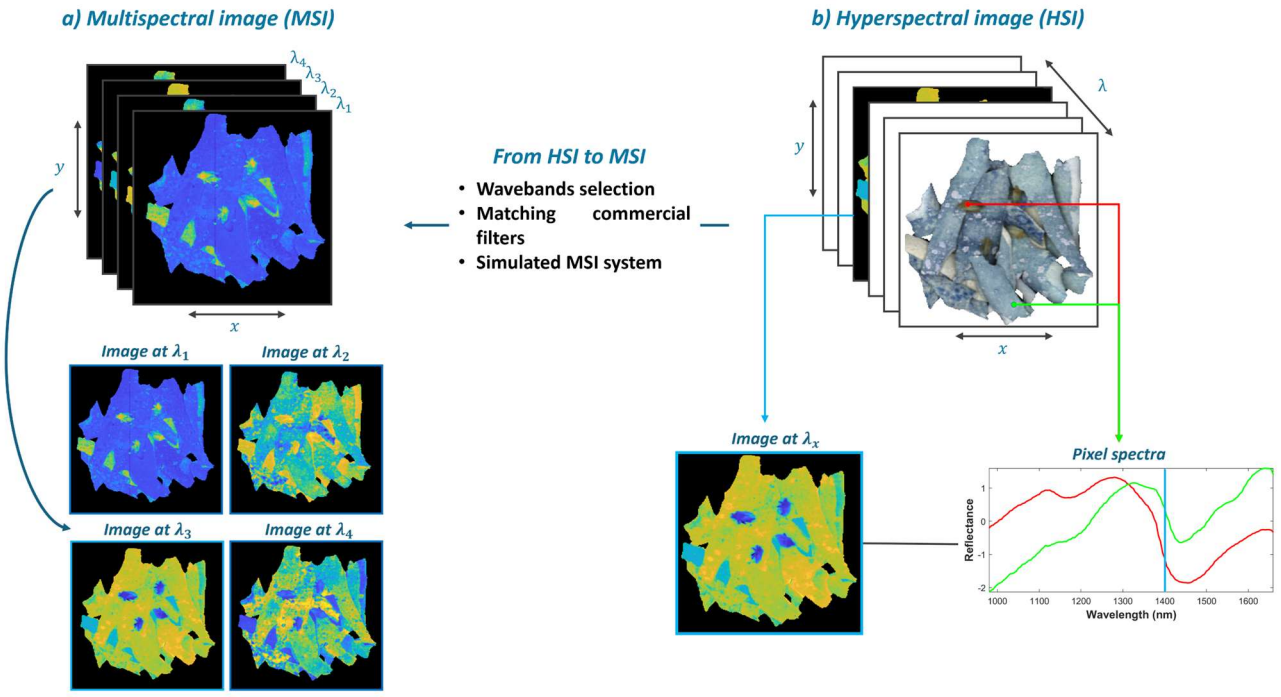


Figure 2.3 *General structure of multispectral (a) and hyperspectral (b) images. For multispectral images, the λ values are referred to the central wavelength.*

Usually, multispectral imaging systems are composed by a detector and optical band-pass filters assembled on a wheel, which its placed between the detector and the optical lenses [25,29]. These systems operate by rotating a filter wheel, thus enabling only a narrow portion of the frequency of light to pass through each band-pass filter. For the implementation of multispectral systems, the following filters' features have to be considered (**Figure 2.4**):

- central wavelength (CWL): the wavelength at the centre of the pass band;
- filter width at half maximum (FWHM): the bandwidth at 50% of the maximum;
- peak transmission (PT): the percentage of light transmission at the CWL.

Nowadays, sensors are composed by many light-sensitive materials like silicon (Si), germanium (Ge), indium gallium arsenide (InGaAs), indium antimonite (InSb), or mercury cadmium telluride (HgCdTe). Silicon (Si) works in the 300-1100 nm range (UV-Vis_NIR), and is the most widely used material due to its lower cost, simple processing, and wide temperature range. With other materials, it is possible to obtain sensors that cover different spectral ranges such as 1000-5000 nm (InSb, HgCdTe, InGaAs). These sensors enable the acquisition of high-quality images when there is sufficient light reaching the sensor [28,30].

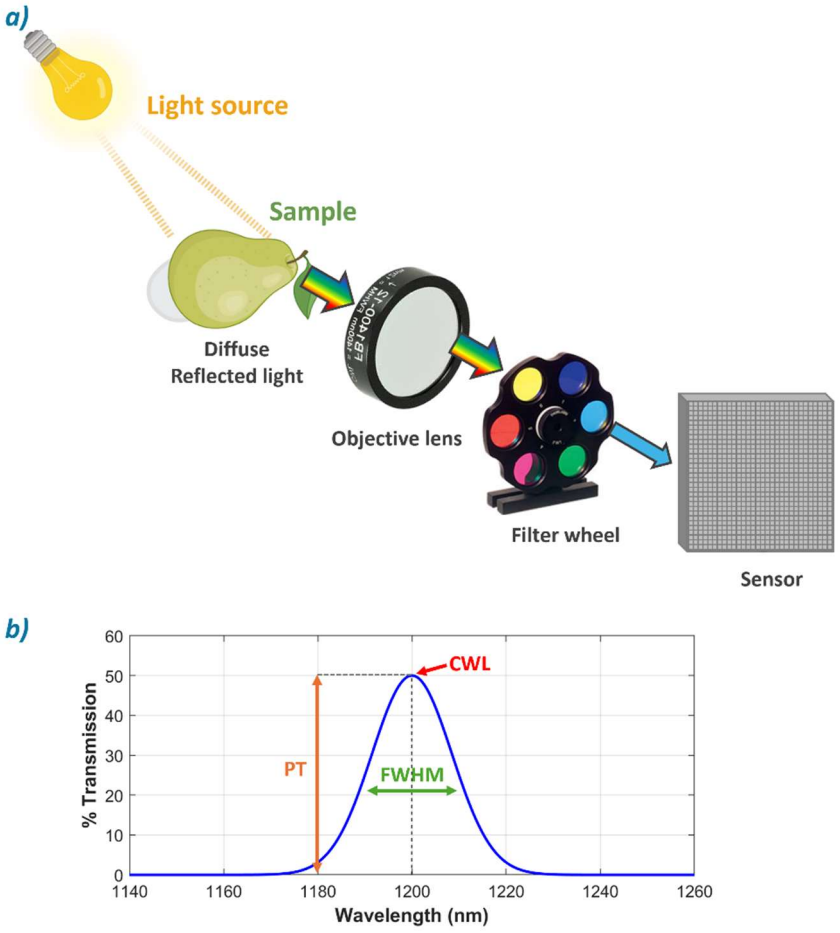


Figure 2.4 *Fundamental aspects of multispectral imaging system: a) schematic representation of the system and b) filters' features visualized onto a Gaussian-shaped transmission profile of a hypothetical bandpass filter.*

As reported in **Figure 2.3 b**, *Hyperspectral images* are also 3-D matrices composed of one spectral dimension and two spatial dimensions, but they are obtained by stacking together hundreds of (**x, y**) grey-scale images acquired at different continuous wavelengths (**λ**). Thanks to this configuration, each pixel, defined by the spatial coordinates **x** and **y**, contains a whole spectrum of **λ** wavelengths. As for point-wise NIR spectroscopy, each pixel-spectrum can be seen as a chemical fingerprint of the sample composition, having specific coordinates onto the image domain.

At the same time, depending on the distribution of the chemical components, it is possible to visualize the spatial variation of spectral intensity of the imaged sample at a single wavelength, also known as *slab* [26].

Generally, thanks to its high spectral resolution and high sensitivity, hyperspectral imaging permits a more detailed representation of the analysed samples. Despite its great potential, hyperspectral imaging systems require costly instrumentation and time-consuming data acquisition process, which limits the speed of online processing [31]. More importantly, hyperspectral imaging data-richness is a double-edge sword which leads to data handling, storage and analysis related issues. These aspects

limit hyperspectral imaging applications mainly at the laboratory scale, eventually adopting proper statistical tools to perform variable selection for the implementation of *ad hoc* multispectral imaging systems [28,32].

Indeed, multispectral imaging systems are more suitable for applications at the industrial scale. Thanks to the reduced data amount, MSI requires shorter computational times and cheaper and more robust optical components than HSI, making the technology more cost effective [28]. As it will be further discussed in **Section 2.3**, in this scenario it is crucial to adopt statistical tools able to analyse images as a whole, to select or extract relevant variables or pixels from spatially resolved Regions of Interests (ROIs), improving spectral imaging industrial applicability.

Many applications of MSI and HSI for the assessment of food quality and safety have been reported in the last two decades; a list of representative publications is reported in **Table 2.1**.

	Main objectives	Product	Spectral range	
			MSI	**HSI**
Quality	**Freshness:** -chemical composition -pH -TVB-N -Tenderness	Beef	SWIR: 900-1700 nm [33]	SWIR: 900-1700 nm [34,35]
		Lamb	SWIR: 900-1700 nm [36]	VNIR: 500-1000 nm [37], SWIR: 900-1700 nm [37,38]
		Pork	VNIR: 400-1000 nm [39] SWIR: 900-1700 nm [40]	SWIR: 910–1700 nm [41,42]
		Fish fillet	VNIR: 400-1000 nm [43,44]	VNIR: 400-1000 nm [45,46]
	Maturity: -TSS -TA -SSC -Firmness	Apple	VNIR: 680-940 nm [47]	
		Banana		VNIR: 400-1000 nm [48]
		Blueberry	VNIR: 400-1000 nm [49]	VNIR: 400-1000 nm [50]
		Nectarine		VNIR: 450-1040 nm [51,52]
		Pear		VNIR: 400-1000 nm [53]
		Bell pepper	VNIR: 400-1000 nm [54]	VNIR: 400-1000 nm [55]
	Mechanical damages (e.g., bruises, chilling injuries, mealiness)	Apple	VNIR: 450-1000 nm [56,57]	VNIR: 500-1000 nm [57,58] NIR: 1000-2500 nm [58]
		Kiwifruit	SWIR: 900–1700 nm [59]	VNIR: 400-1100 nm [60] SWIR: 900–1700 nm [59]
		Mushroom	SWIR: 900–1500 nm [61]	VNIR: 400-1000 nm [62]
		Pear		SWIR: 950–1650 nm [63] VNIR: 400-1000 nm [64]
Safety	**Adulteration and frauds**	Black pepper		SWIR: 950–1700 nm [65]
		Beef	VNIR: 400-970 nm [66,67]	VNIR: 400-1000 nm [68] SWIR: 900-1700 nm [69]
		Coffee	SWIR: 960-1700 nm [70]	SWIR: 960-1700 nm [70]
		Honey		VNIR: 400-1000 nm [71]
		Pork	VNIR: 400-970 nm [66]	
	Biological damage -Molds -Aflatoxin detection -TVC of bacteria	Citrus	VNIR: 400-1000 nm [51,72]	VNIR: 320-1030 nm [73,74]
		Fish	VNIR: 360-1000 nm [43,44]	VNIR: 360-1000 nm [43,44]
		Maize kernel		VNIR: 400-1000 nm [75] SWIR: 1000–2500 nm [76]
		Meat	VNIR: 400-1000 nm [77]	VNIR: 430–960 nm [78] SWIR: 910–1700 nm [42]
		Pomes		VNIR: 400-1000 nm [57,79]
		Strawberry	SWIR: 1000-2500 nm [80]	VNIR: 400-1000 nm [81] SWIR: 1000-2500 nm [80]

Table 2.1 *List of some spectral imaging applications concerning diverse objectives for the assessment of food quality and safety.*

2.1.3. Acquisition methods

The instrumentation is the key point of any reliable measurement system which inherently depends on the main objective of research. Generally, the basic setup of a spectral imaging system consists of a light source (i.e., a lighting system), adequate objective lenses, a set of pass-band filters or a wavelength dispersion device, and a camera with a proper detector. Other than these major components shared by every system, spectral imaging cameras are classified according to the mechanical procedure used to acquire an image [82]. As summarized in **Figure 2.5**, spectral cameras can be grouped into four configurations: point scanning, line scanning, area scanning, and single shot [25,27,83].

- *Point scanning*, also known as *whisker-broom method*, consisting of the acquisition of one spectrum at a single spatial location (i.e., pixel) each time, therefore capturing a spectrum in every measurement. To obtain a 3-D image, the whole spatial scene is acquired by moving either the detector or the sample continuously in the two spatial dimensions [25,82,83]. Since it basically works as a normal spectrometer, the acquisitions are quite time-consuming. However, this configuration reduces the side effect of the sample illumination, as it guarantees the retention of a constant lighting path between the optical system and the sample. Compared to other acquisition methods, it guarantees a higher spatial resolution, which could be more important than speed in some fields (e.g., microscopic imaging). On the other hand, point scanning is not ideal to evaluate food quality or safety at the industrial scale.
- *Line scanning*, also known as *push-broom method*, captures a line of spatial information and the full spectrum for each corresponding pixel. Each line gives a 2D spatial-spectral information and, thanks to the line by line augmentation, the complete volume of data is obtained with only one direction movement between the sample and the detector [25,82,83]. This aspect makes this acquisition mode well suited for conveyor belt systems, advantageous for industrial implementations. Moreover, by recording a line at a time, line scanning systems are much faster compared to point scanning systems, resulting compatible with online processing and production.
- *Area scanning*, or *stare-down configuration*, is a spectral scanning method, where the image field-of-view is fixed and the corresponding 2-D grayscale image is acquired sequentially over a full spectral range [25,82]. This approach is suitable for applications where the sample must be kept stationary or where its movement is not necessary. As the *push-broom* method, this method could be easily adapted to multispectral image set-ups enabling rapid acquisitions [29].
- *Single shot*, also known as *snapshot*, records both spatial and spectral information of the sample with one single shot. The approach entails a large area detector to acquire a 3-D image in a single

integration time [25,29]. This acquisition method has been recently emerging as viable for rapid acquisition of spectral images, achieving up to 20 multispectral images per second [27]. However, the methodology is still evolving: for now, it is characterized by limited spatial resolutions and narrow spectral ranges. The potentialities of this configuration make it worth to further explore it for video recording or real-time applications involving multispectral devices [82].

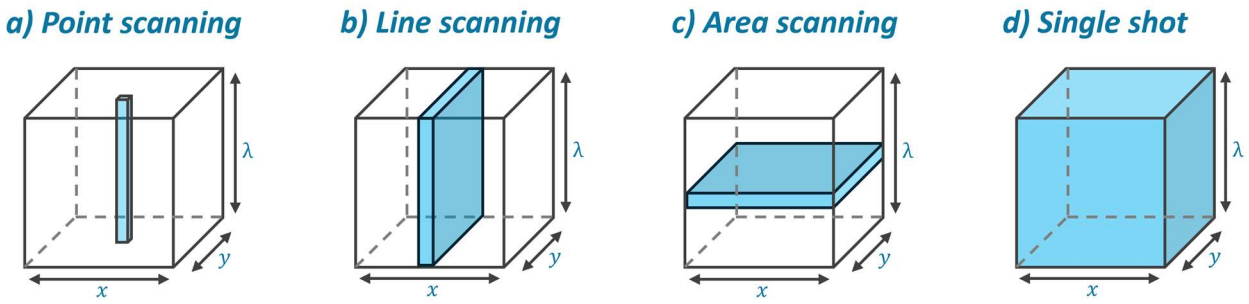


Figure 2.5 *Acquisition configurations in spectral imaging devices: a) point scanning, b) line scanning, c) area scanning, and d) single shot. Scheme adapted from [27].*

After defining ***how the spectral images are acquired***, it is crucial to define ***which sensing mode*** is more suitable for the objective in exam.

When analysing solid and heterogeneous products, the commonly used sensing modes are reflectance, transmittance and interactance, which differ mainly in the position of the light source and the detector (**Figure 2.6**).

- *Reflectance mode*: both the light source and the detector are placed above the sample, hence the measurement enables to capture the diffuse reflected light of the illuminated area of the sample. This sensing mode is widely used to detect the quality attributes of solid food products [50,57]. On the other hand, diffuse reflectance light may be not ideal for internal defects detection of fruits and vegetables due to its limited penetration.
- *Transmittance mode*: the light source and the detector are positioned in the opposite side of the sample. The measurement of transmitted light retrieves information about internal features of the product, thus being suitable for detecting internal defects [84,85]. On the other hand, this sensing mode requires high-intensity light source and a high-sensitivity detector, therefore complicating its practical applications. Moreover, transmittance measurements are influenced by product dimensions (i.e., size, shape) and light pathlength in the product, which further complicate quality assessment [86].

- *Interactance mode*: the detector and light source are positioned on the same side, but a light barrier is placed between them to ensure that the detected light has gone through a minimum distance [86]. It is applied to high moisture products, like fruits, vegetables or pieces of meat or fish, illuminating a limited spot with high intensity light. This sensing mode represents a compromise between reflectance and transmittance measurements [4,87], enabling the evaluation of sublayers of the sample.

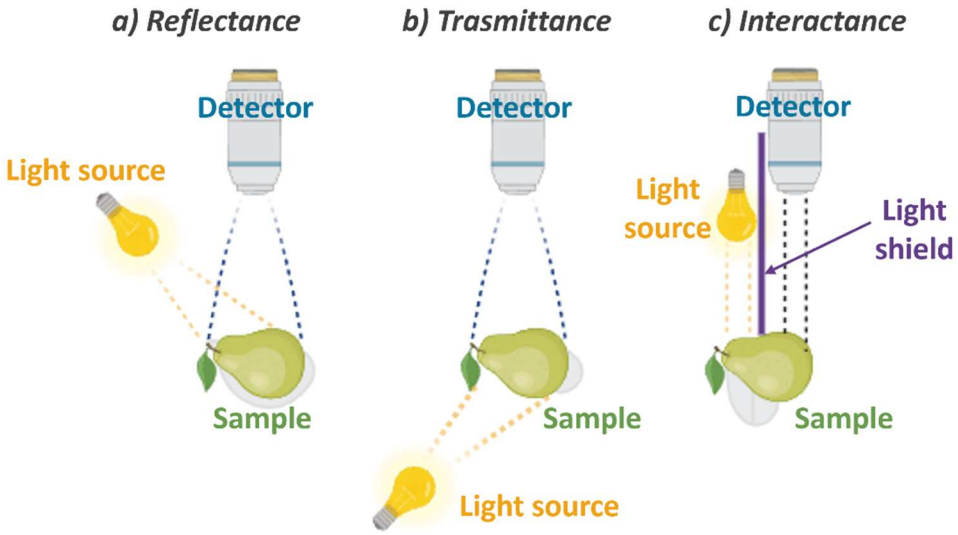


Figure 2.6 *Sensing modes in spectral imaging devices for the analysis of solid agri-food products: a) Reflectance, b) Trasmittance, and c) Interactance. Scheme adapted from [25].*

Generally, the acquired images are saved in *raw* format, where each pixel reports the intensity counts for each measured wavelength (for HSI) or waveband (for MSI) from the detector. Usually, the acquisition software automatically performs a calibration procedure allowing the conversion into the intensity values of the desired sensing mode. Coherently with the case studies that will be presented in this thesis, the conversion into *reflectance* values is essential for analysis.

The images are converted into reflectance throughout the measurements of a standard reference characterised by high reflectance value, or white reference (e.g., spectralon®, ceramic or teflon tiles), and the dark current, i.e., instrumental noise of the detector registered when the camera lens has been shut (**Equation 2.1**).

$$R = \frac{(I-D)}{(W-D)} \qquad (2.1)$$

R corresponds to the reflectance values, I is the intensity measured from the instrument, D is the intensity of the dark current, W is the intensity measured for the standard white reference.

2.2. Multivariate Image Analysis

Multivariate Image Analysis (MIA) mainly consists in the application of chemometric methods for the analysis of images having more than one measurement per pixel (i.e., more than one spectral channel). This step is crucial as it allows the extraction of relevant information to address the problem at hand, partially solving the curse of dimensionality. Following the MIA approach, each pixel of an image is considered as an individual object, characterised by a defined number of variables, i.e., $\boldsymbol{\lambda}$ wavebands for spectral images [88,89].

The first step of MIA usually consists in the *unfolding* of a three-dimensional spectral image into a two-dimensional matrix, containing as many rows as the number of image pixels ($\mathbf{x, y}$) and as many columns as the number of spectral channels ($\boldsymbol{\lambda}$). The unfolding procedure implies the temporary loss of the spatial information of the image, which can be recovered by refolding the two-dimensional matrix into the original image domain. The same procedure is valid for the output quantities obtained by multivariate models, e.g., a PCA score vector can be refolded into the corresponding score image, as it will be further described in **Section 2.2.2**.

As shown in **Figure 2.7**, the main objectives of MIA can be summarized in:

- image exploration: application of unsupervised methods enabling the exploration of images' structure, like e.g., PCA;
- image classification: application of supervised algorithms allowing the classification based on qualitative characteristics, like e.g., Soft Independent Modelling of Class Analogies (SIMCA) and Partial Least Squares Discriminant Analysis (PLS-DA);
- calibration: application of supervised algorithms allowing quantitative prediction of a property of interest, like e.g., Partial Least Squares Regression (PLS-R).

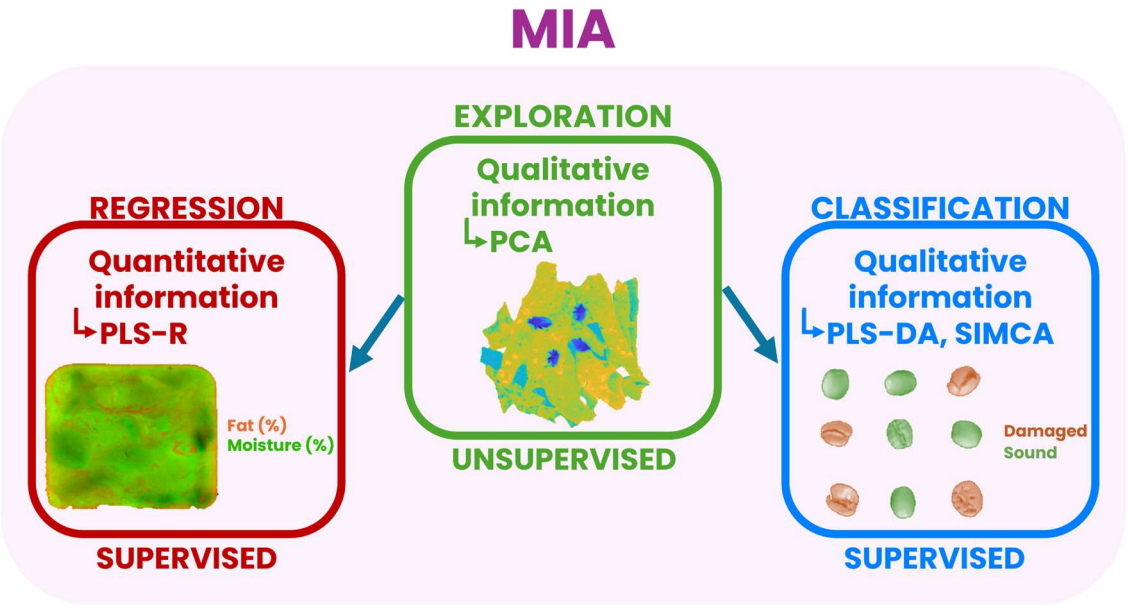


Figure 2.7 *Schematic representation of the different aims of Multivariate Image Analysis (MIA).*

The case studies presented in this thesis are focused on exploring the image structure and on the development of classification models, therefore calibration approaches won't be further addressed.

As it will be further discussed in **Section 2.3**, traditional MIA approaches usually calculate unsupervised or supervised models considering each pixel as an individual sample; however, according to the investigated issue, it may be more convenient to perform image analysis by considering each object within an image or even each image as an individual sample. This latter approach is particularly useful when a large number of images has to be simultaneously analysed.

2.2.1. Data preprocessing

As previously mentioned, spectral imaging techniques analyse the chemical and physical behaviour of light reflected or scattered from a surface. Each element involved in the measurement (i.e., light source, sensor, sample) contributes to a response that includes relevant information as well as noise and artefacts. Data preprocessing is therefore essential to correct the measured signal, reduce unwanted variability, and enhance the extraction of meaningful information.

Spectral distortions are a common issue related to data recording instruments, fluctuations of the light source and the nature of the analysed sample. Despite today' sensors are able to measure the information with a high signal-to-noise ratio, the final response will be always affected by spectral noise.

Light scattering and detector saturation are additional concerns. In heterogeneous samples, variations in texture or chemical composition alter the incident light angle, producing multiplicative or additive

effects that cause baseline drift or detector saturation. Considering spectral imaging system based on reflectance sensing mode and NIR radiation, light scattering is a major concern.

Data preprocessing methods applied on the spectral direction, known as *row-wise* preprocessing, mitigate or erase systematic differences, such as offset or baseline drifts, improving the extraction of useful information [90]. The row-wise preprocessing methods most commonly applied to NIR spectra are the following:

- Detrend, which permits to correct baseline drifts when linear or non-linear trends are present [91]. Following a least squares regression, detrend is based on fitting a polynomial of a given order to the whole spectrum and, then, the calculated values are subtracted from the original spectrum.
- Derivatives can be used to correct baseline vertical shifts and drifts. The derivatives are usually calculated using the Savitzky-Golay (Sav-Gol) algorithm [92,93]. In this frame, a chosen sub-window of points is fitted with a polynomial function of a given order calculated using least squares regression. Then, the derivative value of the polynomial function at the wavelength at the centre of the moving window is calculated. These steps are repeated until all the derivative values are calculated. The most utilised in NIR spectroscopy are first and second derivatives: the former removes baseline vertical shifts while the latter eliminates both baseline shifts and linear drifts. Savitzky-Golay algorithm is also used for smoothing by filtering out the high frequency noise in the signals.
- Multiplicative Scatter Correction (MSC) handles multiplicative scatter effects. Each spectrum is regressed against a reference (typically the mean or median spectrum), and the resulting offset and slope are used for correction [93]. Its performance strongly depends on the selection of an appropriate reference spectrum, which can be challenging in spectral images containing mixtures of different compounds [93].
- Standard Normal Variate (SNV) consists in subtracting the mean value from each signal, then the values are divided by the standard deviation of the signal, enabling normalization. This row-wise preprocessing minimizes scattering and baseline shifts by eliminating additive effects [90].

MSC and SNV are usually classified as scatter-correction methods as they are primarily focused on removing systematic variations due scattering by standardising and normalising the spectra. On the other hand, both detrend, derivatives and smoothing are considered as filtering methods, which are designed to modify or smooth the spectra to enhance signal features or remove noise (**Figure 2.8**) [94,95].

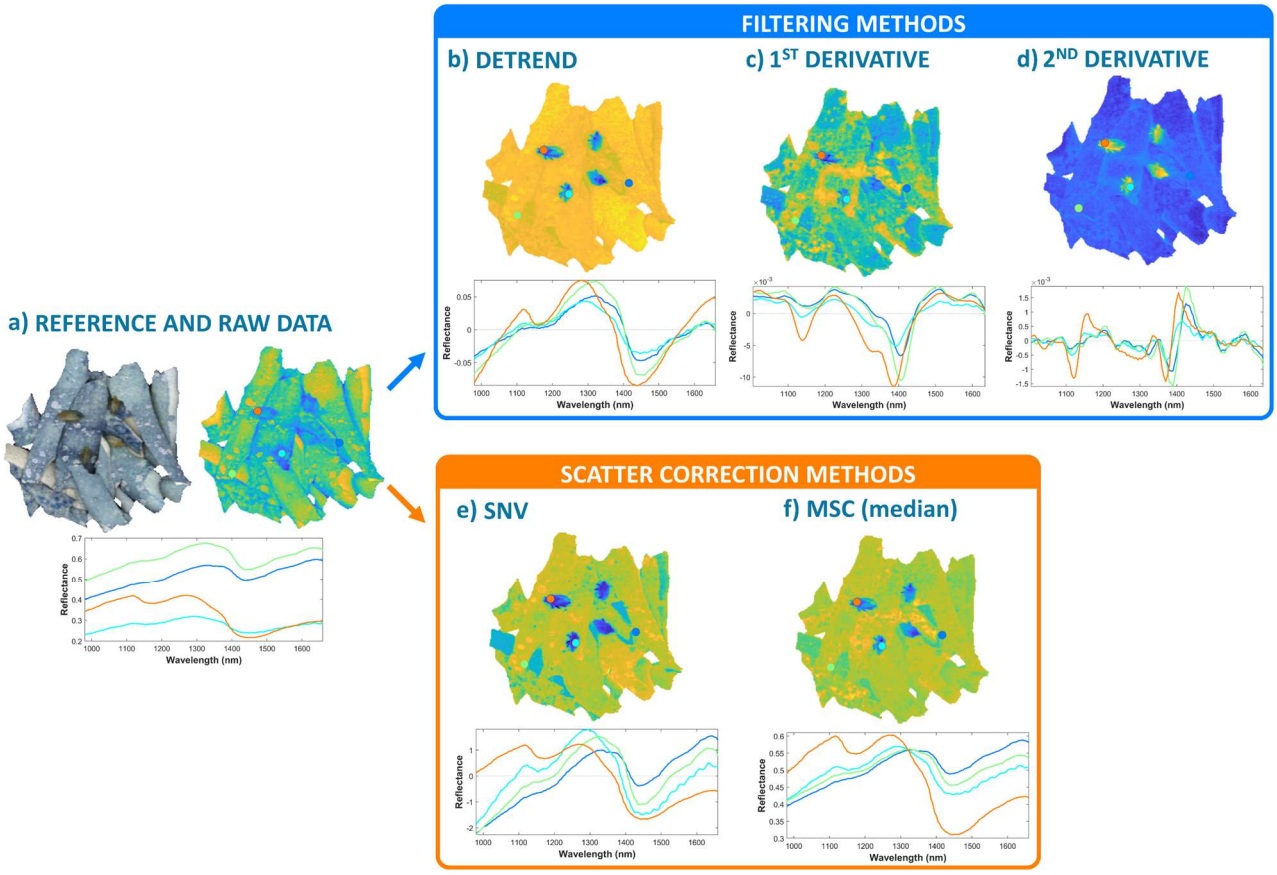


Figure 2.8 *Effect of row-wise preprocessing in removing scattering influence of a reference image of tree bark and bugs a). Both filtering, i.e., linear-detrend b), 1st derivative c) and 2nd derivative d), and scatter correction methods, i.e., SNV and MSC (median), considered in different applications presented in this thesis.*

Row-wise preprocessing methods can be applied to improve the quality of information and interpretability of the results, but they are not mandatory. On the other hand, *column-wise* preprocessing is applied to correct systematic variations among variables [96]. Therefore, these methods must be always applied before the application of chemometric strategies, wisely choosing the most suitable method depending on the type of data [93]. The most widely used column-wise preprocessing methods are mean center and autoscale.

- Mean center (MC) removes the average spectrum contribution, emphasizing variable-specific variation. Fundamentally, mean center translates the axis origin in correspondence of the data centroid. This method is applied to dependent variables having the same nature, such as spectroscopy or chromatography data.

  For each j-th original variable, the mean value $\bar{x}_j$ of the corresponding column vector $\boldsymbol{x}_j$ is subtracted from each vector element x_{ij}:

$$x_{ij,MC} = x_{ij} - \bar{x}_j \tag{2.2}$$

- Autoscale (AUTO) is useful for the pretreatment of independent variables, thus having different nature (i.e., different scales, also depending on different measurements units), allowing them to have the same weight on the data analysis.

  For each *j*-th original variable, the corresponding column vector $\boldsymbol{x}_j$ is mean centered and scaled to unit variance by subtracting the mean value $\bar{x}_j$ from each vector element $\boldsymbol{x}_{ij}$, and then dividing by the standard deviation s_j:

$$x_{ij,AUTO} = \frac{x_{ij} - \bar{x}_{ij}}{s_j} \tag{2.3}$$

2.2.2. Exploring Image structure

Data exploration is a preliminary step of multivariate analysis, which is crucial for gaining information about the structure of image data useful for the development of supervised methods. For this purpose, Principal Component Analysis (PCA) is one of the most widely used multivariate statistical techniques, being able to enhance any existing relationship between objects (i.e., pixel spectra) and variables (i.e., wavelengths), while removing the variability associated with noise.

PCA represents the information within the original dataset in a space of reduced dimensionality defined by *A* principal components (PCs), which are a new set of orthogonal variables calculated as a linear combination of the original ones. In this alternative space the axes are represented by *A* PCs, where PC1 describes the direction of maximum data variance, PC2 is orthogonal to PC1 and accounts for the second direction of maximum variance of the data and so on [96–98].

From the mathematical point of view, PCA decomposes the variance of the unfolded image $\boldsymbol{X}\{n, m\}$, where *n* is the number of pixel spectra (rows) and *m* is the number of spectral variables (columns) of matrix $\boldsymbol{X}$, into a score matrix $\boldsymbol{T}\{n, A\}$, accounting for the systematic variance associated to the samples, a loading matrix $\boldsymbol{P}\{m, A\}$, accounting for the systematic variance associated to the variables, and the residuals matrix $\boldsymbol{E}\{n, m\}$, accounting for the non-systematic (stochastic) variation. Therefore, for the pretreated matrix $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} = \sum_{a=1}^{A} \boldsymbol{t}_a \boldsymbol{p}_a^T + \boldsymbol{E} \tag{2.4}$$

Concerning images, the score matrix represents the pixels position in the PCs space, while the loading matrix describes the weights of the original variables (i.e., wavelengths) in the definition of the PCs. The main goal of PCA is therefore to extract useful information and visualize the data structure through graphical representations. As shown in **Figure 2.9 c**, the values of the score vectors for each PC can be reported in a *score plot*, allowing to visualize the distribution of the objects and thus identifying clusters of pixel spectra having similar characteristics or the presence of possible outliers.

When dealing with images, each PC score vector can be refolded back into the original image domain to visualize the corresponding *score images*.

The interaction between these two types of score vectors representations is often exploited to perform image segmentation by means of *brushing* (**Figure 2.9**), which allows the simultaneous visualization of a selected cluster of pixels in both representations [83].

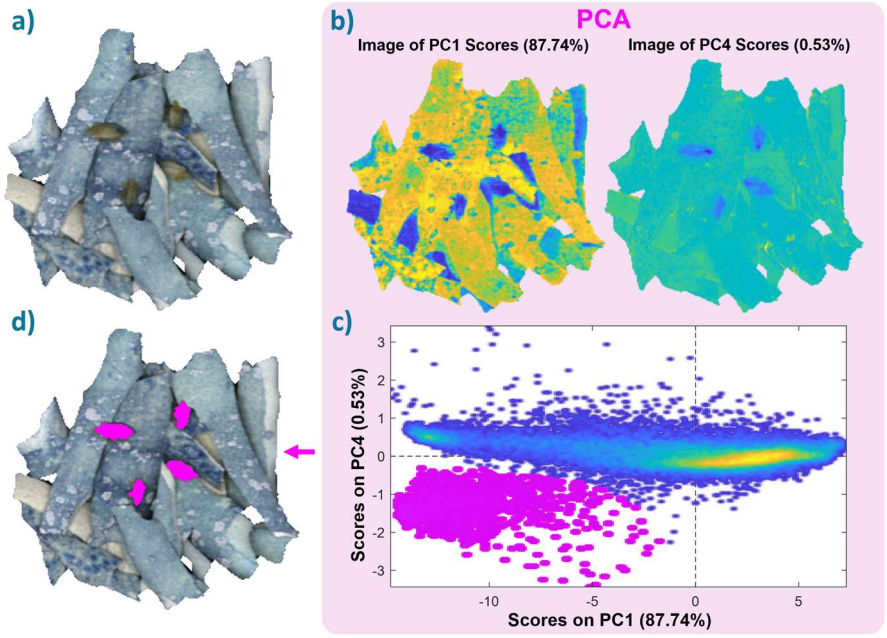


Figure 2.9 *In a) pre-processed hyperspectral image (i.e., visualized in false RGB colours) analysed through PCA. The corresponding score values of PC1 and PC4 can be visualised as b) score images, by re-folding the score values into the original image domain, or as c) score plot. The selection of clustered pixels (i.e., brushing) having negative PC1 and PC4 score values allows the visualization of bugs specimens on bark background.*

Concerning variables, the loading plot can be visualized to identify which original variables are the most influential to determine the direction of a given PC [96]. When dealing with spectral images, where variables are numerous and highly correlated, each loading vector can be represented by a curve that is a function of the domain of the original variables. For interpretation purposes, the regions of the spectrum with loading values that deviate the most from the origin of the y axis correspond to the most influential ones, which have the greatest impact on the direction of the principal component (PC) under consideration.

As previously mentioned, PCA can also detect the presence of possible outlier pixels, which significantly affect the results of the analysis. In this frame, the parameters Q residuals and Hotelling's T^2 are usually considered to verify the nature of the extreme objects [99]:

- Q residual values refer to the distance of each object from its projection on the hyperplane of the PCA model, enabling the visualization of objects following an unusual pattern not described by

the model. For the *i*-th sample (pixel spectrum) defined by the row vector $\boldsymbol{x}_i$, the corresponding Q residual value Q_i is given by:

$$Q_i = \boldsymbol{e}_i \boldsymbol{e}_i^T = \boldsymbol{x}_i(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{P}^T)\boldsymbol{x}_i^T \quad (2.5)$$

where $\boldsymbol{e}_i$ corresponds to the *i*-th row of the residuals matrix $\boldsymbol{E}$, $\boldsymbol{P}$ is the loading matrix and $\boldsymbol{I}$ is the identity matrix.

- Hotelling's T^2 values quantify the distance between the projection of an object on the PCA model and the centre of the model itself. This parameter represents the objects variability within the model and, therefore, it allows to detect the ones having the highest contribution in the definition of the model itself. For the *i*-th sample (pixel spectrum) defined by the row vector $\boldsymbol{x}_i$, the corresponding Hotelling's T^2 value T_i^2 is given by:

$$T_i^2 = \sum_{a=1}^{A} \frac{t_{ia}^2}{s_a^2} \quad (2.6)$$

where t_{ia} corresponds to the value of the *i*-th sample for the *a*-th PC score vector, while s_a^2 is the variance associated with the *a*-th score vector ($\boldsymbol{t}_a$).

2.2.3. Image classification

Multivariate classification methods are supervised techniques designed to create mathematical models able to assign each sample to its respective class based on a set of measurements. From the mathematical point-of-view, classification techniques handle qualitative responses by defining the relationships between a set of descriptors (e.g., chemical measurements, signal intensities) and a categorical variable which determines class membership [100].

Regardless of the classification algorithm used, the original dataset is subdivided into a training set used for model calculation, which defines a separation line (*threshold*) to attribute a sample to a given class, and an external test set for validation, to evaluate the model's predictive ability. When dealing with supervised methods, it is crucial to select the optimal number of *Latent Variables* (LVs) representing the best trade-off between capturing the important variability in the data and an excessive redundancy in the descriptor variables, to avoid *underfitting* or *overfitting*. The model's dimensionality and the preprocessing methods leading to the best results are usually chosen through a cross-validation step of the training set samples.

Cross-validation consists in dividing the training set into a defined number of deletion groups and iteratively calculating the model while excluding one group at a time. This operation is carried out for each number *a* of LVs (with $1 \leq a \leq A$), where *A* is the maximum number of LVs given the size of the training set. Once all samples have been excluded and predicted, several figures of merit are computed on class predictions of left-out samples. Different cross-validation strategies can be

adopted, e.g., contiguous blocks, venetian blinds, random subsets or custom. The selection of a correct strategy is crucial since affecting the predictive ability of the model: one advisable procedure is to consider the structure of the training set and samples arrangement and, if necessary, implement multiple rounds of cross-validation [101,102].

Classification results can be retrieved from the confusion matrix, a square matrix with size {G × G} of *G* modelled classes, where each element n_{gh} represents the number of samples actually belonging to a class *g* assigned to a class *h*. The matrix entries in the main diagonal indicate the correctly classified samples, while the off-diagonal entries correspond to misclassified samples (**Figure 2.10**) [103].

The most common figure of merits for evaluating the performances of classification models are:

- Sensitivity (SENS) corresponds to the percentage of samples of each modelled class correctly assigned to the corresponding class (i.e., ratio between True Positives, TP, and samples belonging to the class);
- Specificity (SPEC) corresponds to the percentage of samples belonging to other classes correctly rejected by the considered class (i.e., ratio between True Negatives, TN, and samples not belonging to the class);
- Efficiency (EFF) is the geometric mean of SENS and SPEC values;
- Non-Error Rate (NER), also known as average sensitivity, corresponds to overall percentage of samples correctly classified.

These parameters can be calculated in calibration, cross-validation and prediction of the external test set samples.

		Real class		
		Class 1	Class 2	Class 3
Predicted class	Class 1	25	4	1
	Class 2	8	22	0
	Class 3	2	1	27

Figure 2.10 *Example of a confusion matrix.*

Supervised classification algorithms can be grouped in Class Modelling (CM) and Discriminant Analysis (DA) methods. These approaches represent two distinct classification philosophies: CM

defines class boundaries by focusing on similarities among samples belonging to the same class, while DA is based on modelling the differences between the different classes of interest.

Among CM techniques, the most widespread is Soft Independent Modelling of Class Analogy (SIMCA), introduced by Svante Wold in 1976 [104]. As accurately summarized by the acronym, in SIMCA each class is *modelled independently* of others, relying on *analogies* among samples belonging to the same class. If more than one class is modelled, samples can be recognised as members of none, one, or multiple modelled classes.

SIMCA creates an individual PCA model for each class, which is characterised by its own dimensionality. After defining the proper number of PCs in cross-validation for each PCA model, class assignment of a new sample is based on Q residuals and Hotelling's T^2 statistics, previously defined in **Section 2.2.2**. Coherently with literature, for each sample Q residual and Hotelling's T^2 values are often referred to as *Orthogonal Distance* (OD) and *Score Distance* (SD), respectively. OD and SD statistics can be combined in different ways to define class boundaries, originating several variants of SIMCA algorithm. In ***Chapter 4***, a specific variant will be presented.

In contrast, DA techniques define precise boundaries that partition the global data domain into as many regions as the number of classes. The class boundaries are calculated by maximizing the discrimination among samples from different classes. Therefore, for each class model each sample is forcibly assigned to one of the classes considered, regardless its class membership status. One of the most widely used discriminant analysis methods is Partial Least Squares Discriminant Analysis (PLS-DA), a variant of Partial Least Squares (PLS) regression algorithm adapted for discriminant analysis scenarios.

Coherently with its regression counterpart, PLS-DA maximizes the covariance between the original data matrix $\boldsymbol{X}$ of measurements and a dummy matrix $\boldsymbol{Y}$, which contains a set of binary column vectors encoding for class membership of each sample [105]. The dummy matrix $\boldsymbol{Y}$ is composed by as many columns as the number of investigated classes and as many rows as the number of investigated samples. Similarly to PCA, the information is summarized in a subset of successive orthogonal directions, i.e. LVs. However, unlike PCA, in PLS the directions of the LVs are determined by simultaneously considering the information in both the $\boldsymbol{X}$ and $\boldsymbol{Y}$ blocks, specifically by maximizing their covariance [100].

More in detail, both **X** and **Y** matrices are separately decomposed as follows:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \tag{2.7}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{F} \tag{2.8}$$

The decomposition of $\boldsymbol{X}$ and $\boldsymbol{Y}$ matrices has to be optimized in a manner that the objects-related variance of the descriptors block $\boldsymbol{X}$, expressed by the score matrix $\boldsymbol{T}$, can be used to describe $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Q}^T + \boldsymbol{G} \quad (2.9)$$

To this aim, for each LV, a weight vector ($\boldsymbol{w}$) is calculated to weight the original variables according to their contribution to explain the $\boldsymbol{Y}$ matrix. Based on these considerations, the estimated $\boldsymbol{Y}$ matrix ($\widehat{\boldsymbol{Y}}$) can be calculated as follows:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{Q}^T = \boldsymbol{X}\boldsymbol{B} \quad (2.10)$$

where $\boldsymbol{W}$ is the weights matrix accounting for the contributions of the original variables in explaining the $\boldsymbol{Y}$ matrix and $\boldsymbol{B}$ is the set of coefficients derived for each class.

Moreover, the relevance of the contribution of the original variables to the PLS-DA model can be evaluated by calculating the Variables Importance in Projection (VIP) scores [106]:

$$VIP_j = \sqrt{\frac{m\sum_{a=1}^{A}\left[\left(q_{ja}^2\boldsymbol{t}_a^T\boldsymbol{t}_a\right)\left(w_{ja}/\|\boldsymbol{w}_j\|^2\right)\right]}{\sum_{a=1}^{A}\left(q_{ja}^2\boldsymbol{t}_a^T\boldsymbol{t}_a\right)}} \quad (2.11)$$

Where m is the number of variables of the descriptor matrix $\boldsymbol{X}$, w_{ja} is the loading weight of the j-th variable at the a-th latent variable, $\|\boldsymbol{w}_j\|^2$ is the variance accounted from variable j, and $\boldsymbol{t}_a$ and $\boldsymbol{w}_a$ are the a-th column vectors of $\boldsymbol{T}$ and $\boldsymbol{W}$ matrices, respectively, while q_{ka}^2 is the element of matrix $\boldsymbol{Q}$ corresponding to the j-th variable and a-th latent variable.

Once the model has been calculated, class assignment of a new observation to a given class is based on the value of $\hat{y}$ estimated by the model, which is compared with the distribution of the $\hat{y}$ values estimated by the model for the training set samples belonging to the considered class. As a result, for each class the corresponding $\hat{y}$ values estimated by the model will never be exactly equal to 0 or to 1, but they are supposed to lie in a range close to these reference values. In PLS-DA, threshold values for class assignment are defined by means of Bayes Theorem: first, it is assumed that the estimated $\hat{y}$ values for each class follow a Gaussian distribution and, secondly, the threshold is set at the intersection point where the two estimated distribution cross, i.e., the point at which the probabilities of belonging or not belonging to the class are equivalent. Therefore, if the $\hat{y}$ estimated value of a new sample is greater than the threshold, it will be assigned to the corresponding class; conversely, it will be rejected [100].

Recently, soft discriminant methods are emerging as a viable alternative to benefit from the advantages of both classification approaches. In this thesis, Soft PLS-DA algorithm was used for the development of classification models in the different case studies considered. Soft PLS-DA is a variant of PLS-DA algorithm allowing, at the same time, both outlier detection and maximization of the differences between the modelled classes [107].

The main difference is based on ***how*** class assignment is made since the following additional criteria have to be met:

- Q residuals values falling within the 99.9 % confidence limit of the model. This limit was set wide enough to consider as much as possible within classes variability, but allowing at the same time to exclude samples with a very poor fit to the model;
- $\hat{y}$ predicted values falling within an acceptability range for the considered class. The lower limit of this range is defined by the PLS-DA threshold value for the considered class ($y_{tsh1,g}$), while the upper limit allows the rejection of objects located at the extremes of the Gaussian probability density function ($y_{tsh2,g}$). The upper limit is calculated as the sum of the mean $m_{\hat{y},g}$ and 5 times the standard deviation $s_{\hat{y},g}$ related to $\hat{y}$ values of class g:

$$y_{tsh2,g} = m_{\hat{y},g} + 5 \times s_{\hat{y},g} \quad (2.12)$$

- the samples must be unambiguously assigned to one class only.

The samples that do not match all the criteria defined by the Soft PLS-DA constraints are not assigned to any class and automatically labelled as *Not Assigned*.

The general robustness and flexibility of Soft PLS-DA algorithm makes it suitable for a wide range of applications. In ***Chapter 3*** and ***Chapter 4***, the advantages in adopting Soft PLS-DA for pest detection, damage detection and authentication will be discussed.

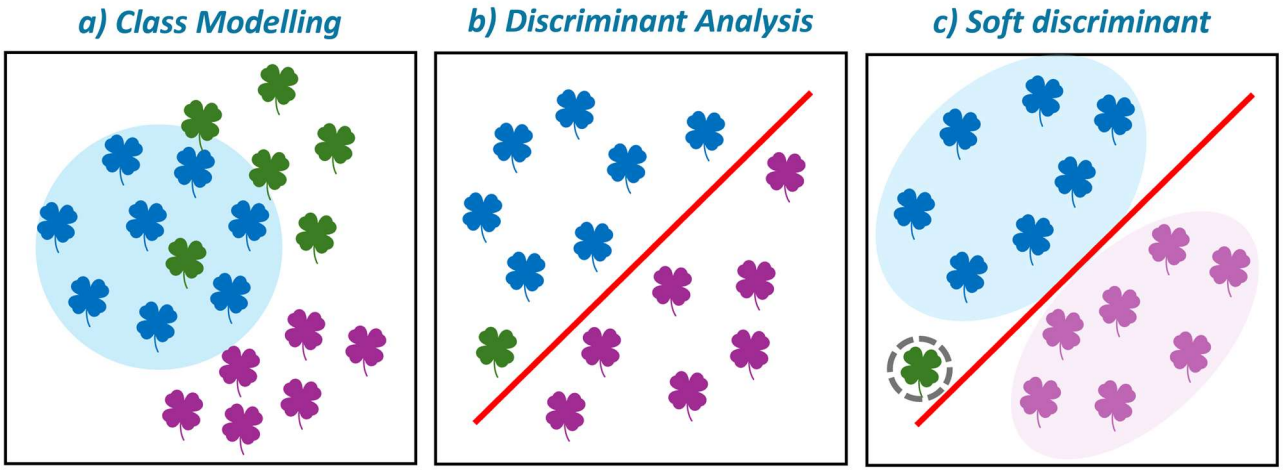


Figure 2.11 *Schematic representation of how Class Modelling a), Discriminant Analysis b) and Soft PLS-DA c) behave towards new observations belonging to unmodelled class (i.e., green clover samples). In a) new observations having similar characteristics to the modelled class are misclassified while in b) misclassifications may result from forced class assignment of new observations not belonging to any modelled class. In c) the new observation not belonging to any modelled class is handled correctly, thus being rejected by both classes.*

2.3. Data reduction of image datasets

The major drawback of spectral imaging is the processing of large amounts of data, which implies computational difficulties for real-time applications and complicates the extraction of useful information. The high dimensionality inherent in these datasets and related constraints, often referred to as the "curse of dimensionality", determine an inverse relationship between the number of input features in a model and its ability to effectively generalise [108].

Several approaches can be applied to overcome the curse of dimensionality and accelerate analysis. Among them, *feature selection* and *feature extraction* are the most common.

Feature selection identifies the most relevant features without altering their nature, while reducing redundancy and improving model efficiency. Dealing with spectral images, the features of interest can be selected both in the spectral (i.e., spectral channels) or in the spatial (i.e., pixels) domains.

As it will be discussed in **Section 2.3.1**, in this thesis interval- and sparse- based variable selection algorithms were used to identify informative spectral features, while random selection or Kennard–Stone algorithms can be used to subsample pixels to create representative spectral libraries of the classes of interest.

In contrast, feature extraction usually implies the creation of new features by combining or transforming the original ones, compressing the most important information in fewer dimensions. As presented in **Section 2.3.3**, hyperspectrograms follow the feature extraction approach.

In the previous section, the application of chemometric methods on image data was introduced under the assumption that each pixel spectrum was treated as an individual sample. However, this workflow is not always feasible: with high-dimensional image datasets, dedicated data-reduction strategies are required to ensure fast and practical processing.

Indeed, at the pixel-level each pixel spectrum is treated as an individual observation, offering the highest resolution but rapidly becoming impractical for large image datasets or real-time applications. Object-level and image-level approaches progressively reduce dimensionality by summarizing spectral information into compact feature sets, enabling faster and more feasible analyses while still retaining essential spatial context. In parallel, feature selection techniques may further refine datasets by isolating the most informative pixels, Regions of Interest (ROIs), or wavelengths.

2.3.1. Feature selection methods

Feature selection methods enable the identification of informative features from images, ensuring that the selected features represent the relevant variability for classification or calibration tasks. Feature selection can be applied in the spatial domain to select representative pixels accounting for variation

sources of the imaged samples, or in the spectral direction to identify the most influential wavelengths for classification or calibration purposes.

Spatial feature selection

When developing pixel-level calibration or classification models, a common approach is to extract a dataset of representative spectra by selecting a defined amount of them from ROIs within images. These spectra are then used for model development when they belong to the training set, or for model validation when they are part of the test set.

In this context, the preliminary evaluation of the information content of an image using exploratory tools such as PCA is crucial for characterizing image structure, highlighting variability patterns, and guiding the selection of meaningful ROIs. Indeed, the selection of ROIs directly affects the representativeness of both the training and external test set, influencing models' prediction ability.

Therefore, knowledge gained during exploratory analysis provides essential guidance on the most appropriate pixel selection strategy.

Random pixel selection is the simplest and most widespread approach, yielding the selection of a predefined number of randomly sampled pixels. However, it ignores the overall variance structure within ROIs and may lead to biased solutions, especially for small or heterogeneous datasets [109–111].

When a better representation of variability is required, the *Kennard–Stone* algorithm represents an effective alternative. This deterministic method selects samples that better approximate a uniform distribution in the feature space. At first, it identifies the pair of samples with maximal distance and then iteratively adds the furthest observations from those already selected. As a result, the final subset captures all major sources of variability, including the extremes. This configuration may also lead to overoptimistic outcomes in some cases [110,112].

Spectral feature selection

In spectral imaging systems, spectral variable selection can drastically reduce computational load, enabling faster solutions suitable for real-time applications. It is often the first step toward developing multispectral systems, which are more robust and cost-effective than hyperspectral instruments. The variable selection approaches considered in the following chapters are interval- and sparse-based methods.

Interval-based methods are iterative variable selection methods aimed at finding the most informative intervals of wavelengths (i.e., spectral bands), therefore being particularly suitable for highly multicollinear spectroscopic data. The most popular approach was introduced by Nørgaard and

colleagues [113]. These methods are based on subdividing the entire spectrum into intervals of the same width, then local models are iteratively calculated and compared to the global model, i.e., the one obtained for the full spectrum, based on the Root Mean Square Error or the Error-Rate (1-NER) in cross-validation [106].

Interval variable selection may follow two different configurations:

- *Forward selection*, which starts with single-interval models and iteratively adds intervals that minimize cross-validation error, until no further improvement is obtained;
- *Backward elimination*, which begins with all intervals included and iteratively removes the least informative interval, until no further improvement in cross-validation error is obtained.

Although its simple application, the results mainly depend on the defined interval size.

Sparse variable selection methods, by contrast, do not rely on predefined intervals or supervised information and are therefore suitable for both unsupervised and supervised applications. These methods introduce sparsity by applying a penalty term to the model's parameter vector, forcing uninformative or noisy coefficients to be equal to zero [114–117]. Among the different penalization strategies, the Least Absolute Shrinkage and Selector Operator (LASSO) [118] applies a penalization to the sum of absolute values of a parameter vector (or **L_1 norm** penalty), effectively setting several coefficients to zero. The LASSO penalty can be coupled with common chemometric tools such as PCA, PLS or PLS-DA to simultaneously perform model computation and variable selection.

Concerning the parameter vector of a linear regression model, the LASSO applies the following constraint:

$$\arg min_b = (\|\boldsymbol{y} - \boldsymbol{Xb}\|_2^2) \quad (2.13)$$

subjected to

$$\|\boldsymbol{b}\|_1^1 \leq c \quad (2.14)$$

where $\|\cdot\|_2^2$ and $\|\cdot\|_1^1$ refer to the L_2 norm and the L_1 norm penalties, respectively. In this manner, LASSO forces the sum of the absolute values of the regression vector $\boldsymbol{b}$ to be lower than the tuning scalar c, leading to a sparse solution where some coefficients are shrunk and others are set to zero. Therefore, the lower the value of the sparsity constraint c, the higher the sparsity imposed, as the LASSO converges to the same results of the unconstrained model when c is equal to the L_1 norm of $\boldsymbol{b}$. The sparsity constraint c is a tuning parameter that should be properly optimised together with model dimensionality to obtain the final model.

As it will be discussed in ***Chapter 5***, in the PCA context LASSO penalty is usually performed on PCA loadings to select spectral variables. However, LASSO penalization may also be applied to score vectors enabling pixel selection [115,119].

2.3.2. From pixel-level to object and image-level analysis

So far, the chemometric methods for MIA purposes were introduced assuming that each pixel (i.e., spectrum) is considered as an individual sample. This is not always the case: especially when dealing with high dimensional image datasets, proper data reduction strategies must be taken to enable fast and straightforward analysis.

The approach presented so far is pixel-based or *pixel-level*, as each pixel within a spectral image is treated as an individual sample or observation.

Therefore, supervised and unsupervised models are calculated starting from the information contained in the considered pixel spectra (**Figure 2.12 a**) and the model's outcomes can be visualised back to the image domain through score images or prediction images. These are pseudo-colour images where each pixel is coloured according to its PCA score value (score images) or to the class or quantitative parameter assigned by a supervised model (prediction images). Thanks to the interpretability advantages, analysis at the pixel-level is generally the first step in evaluating spectral images.

However, this approach is computationally intensive and requires substantial hardware resources, often making it impractical for large image datasets or real-time applications. Indeed, these applications often involve large numbers of samples and consequently hundreds or thousands of images, exceeding memory and time constraints.

In this context, spectral images can be seen from another *level* of perspective: for example, analysis at image-level or at object-level are able to address these limitations.

In *image-level* approaches, each image is treated as a single sample rather than analysing individual pixels. Generally, these strategies imply data dimensionality reduction by converting each image into a one-dimensional feature vector that retains the key relevant information, and then analysing the resulting matrix of feature vectors using standard chemometric methods **(Figure 2.12 b)**.

The simplest and most popular manner to summarize the information within a single image is to calculate its average spectrum. Although being suitable for homogeneous samples, this approach completely neglects spatial and spectral variability thus being ineffective when the objective is to identify localized features. Indeed, this approach determines the loss of spatial information which represents the main advantage of spectral imaging techniques.

To retrieve both spatial and spectral information, different procedures can be applied starting either from the original data matrix or from model's outputs (e.g., PCA: scores, loadings, residuals, Hotelling's T^2). Starting from these, feature profiles or derived indices can be computed for each image enabling subsequent multivariate analysis. For example, feature profiles can be calculated as the frequency distributions of spectral intensities at relevant wavelengths or score values on PCs of

interest, stacking the information in a single continuous signal. These feature profiles may also include additional descriptive statistical parameters describing location (e.g., mean, median, mode) or variance (e.g., standard deviation, kurtosis, skewness, percentile intervals) providing compact descriptors of the underlying frequency distributions [120–123]. When defining the features or indices throughout PCA quantities, it is crucial to determine the appropriate number of significant components. However, since the components that are not retained define the residuals of the PCA model, they may also be useful for the detection of anomalies or outliers.

The final matrix of feature profiles, where each row corresponds to one image while columns are related to features, is subjected to exploratory analysis and/or supervised modelling.

This strategy characterizes global properties and is mostly used when the interest lies in similarities or differences among images rather than within-image variability.

When the focus of analysis falls on discrete objects within images, not entire images or single pixels, we refer to *object-level* approaches. The first essential step is identifying ROIs, or pixels belonging to each individual object, typically through background removal and morphological operations. Once the objects have been identified, feature profiles or derived indices can be computed for each object following the same procedures discussed for the image-level approach. As for the images, the features of interest are extracted from each object identified at the pixel-level, composing a matrix of feature profiles [64,124–126] having a number of rows corresponding to the number of objects and number of columns related to the considered features (**Figure 2.12 c**).

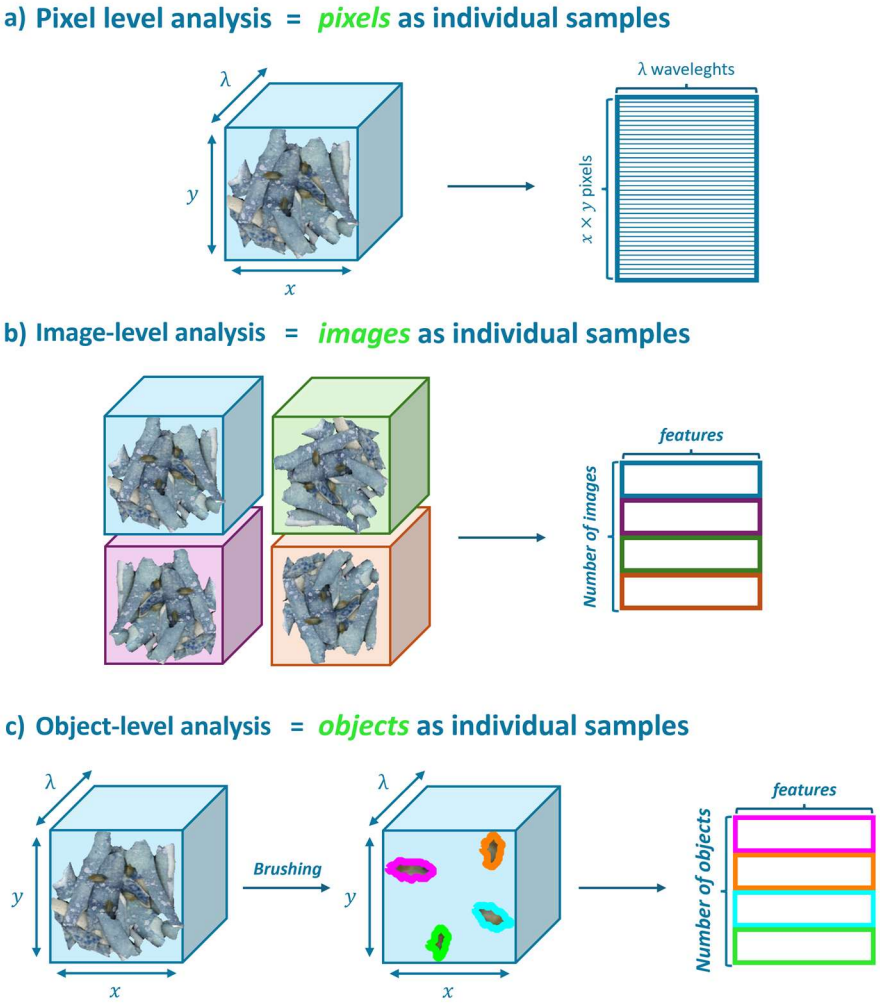


Figure 2.12 *Schematic representation of the three main approaches usually adopted for multivariate analysis of spectral images.*

2.3.3. Hyperspectrograms

As previously mentioned, spectral images are frequently converted into one-dimensional signals by calculating the average spectrum from the whole image or from specific ROIs [62,121]. However, this approach does not account for spatial information, determining an under representation of variability within heterogeneous samples. To overcome this issue, several methods enabling the extraction of relevant features, throughout the calculation of frequency distributions of spectral intensities at relevant wavelengths or score values on PCs of interest, have been emerging [125,127]. In this section it is presented an alternative method named *hyperspectrograms*, able to summarize both spatial and spectral information of spectral images [22,128,129].

Hyperspectrograms are one-dimensional signals obtained by merging in sequence the frequency distribution curves of pixel-related features determined by a PCA model (i.e., scores, Q residuals, Hotelling's T^2, loadings).

Depending on the type of useful information to be encoded within the hyperspectrograms, the PCA model used to convert the images into signals can be calculated individually for each image in the dataset, i.e., *Single Space Hyperspectrograms* (SSH), or globally, by considering all the images within the dataset, i.e., *Common Space Hyperspectrograms* (CSH).

In the first case scenario, each image is converted into the corresponding hyperspectrogram by sequentially merging the frequency distribution curves of score values, Q residuals, Hotelling's T^2 values and loading vectors related to the PCs chosen for model calculation. Therefore, each SSH signal consists of two parts: one encoding the spatial information, i.e., frequency distribution curves, and one encoding the spectral information, i.e., loading vectors (**Figure 2.13 a**).

In the second case scenario, the PCA model underlying the CSH signals calculation is a global model, in which the PC space is common to all images in the dataset. As for SSH, CSH signals are obtained by merging in sequence the frequency distribution curves of the PCs score vectors, Q residuals, and Hotelling's T^2 values (**Figure 2.13 b**). Hence, CSH signals explicitly retain only information related to the spatial variability of the corresponding image, while spectral information is implicitly retrieved by a common set of loading vectors. Since these vectors are the same for all images, they are not included in the signals.

SSH signals are more effective in identifying specific variations among the properties of pixels within an image because they account for variability within the image. Conversely, CSH signals consider both variability within individual images and among different images, thus estimating at the same time spatially resolved features within images and common properties to the entire image dataset. These advantages make CSH signals a viable tool for both the characterization of homogeneous samples and defects detection. This aspect is crucial for the detection of spatially localized features (e.g., pixels corresponding to defects), which slightly differ from other pixels based on spectral information.

As for common spectroscopic signals, the final matrix of hyperspectrograms can be subjected to unsupervised and supervised modelling, enabling the exploration of the overall dataset structure or to develop multivariate classification or calibration models working at the image-level. As for other image-level strategies, in a PCA model calculated from a matrix of CSH signals each object visualized in the score plot will represent a specific image of the dataset: this operation can highlight the presence of clusters of images having a similar characteristics and/or outliers [130,131]. Coherently, hyperspectrogram matrices can be used to develop multivariate classification models, where class assignment is performed on the entire image, considered as a single entity [128,132–134] (**Figure 2.12 b**).

As it will be further discussed in ***Chapter 3***, PCA quantities retrieved for each pixel of the investigated images can be visualized back into the original image domain, enabling a better visualization of the results. Thanks to this property, variable selection strategies may be applied to hyperspectrograms to extract relevant features and perform automatic ROIs detection.

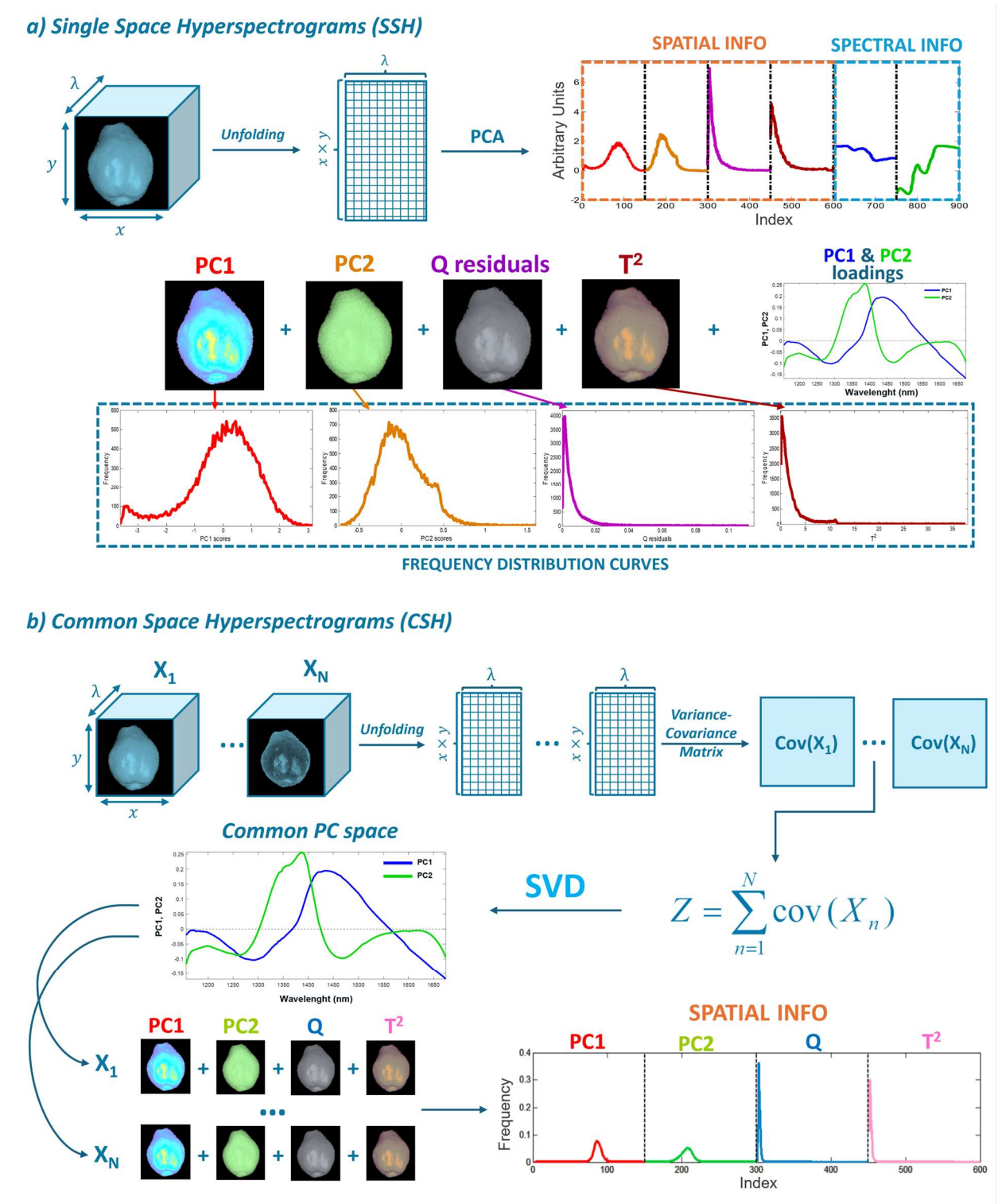


Figure 2.13 *Schematic representation of procedure followed to convert spectral images into a) SSH and b) CSH signals.*

References

[1] B.M. Nicolaï, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K.I. Theron, J. Lammertyn, Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review, *Postharvest Biology and Technology* 46 (2007) 99–118. https://doi.org/10.1016/j.postharvbio.2007.06.024.

[2] B.M. Nicolaï, T. Defraeye, B. De Ketelaere, E. Herremans, M.L.A.T.M. Hertog, W. Saeys, A. Torricelli, T. Vandendriessche, P. Verboven, Nondestructive Measurement of Fruit and Vegetable Quality, *Annual Review of Food Science and Technology*. 5 (2014) 285–312. https://doi.org/10.1146/annurev-food-030713-092410.

[3] K.H. Norris, History of NIR, *Journal of Near Infrared Spectroscopy* 4 (1996) 31–37. https://doi.org/10.1255/jnirs.941.

[4] X. Yang, P. Berzaghi, Near-Infrared Spectroscopy, in: A.M. Jiménez-Carvelo, A. Arroyo-Cerezo, L. Cuadros-Rodríguez (Eds.), Non-Invasive and Non-Destructive Methods for Food Integrity, Springer Nature Switzerland, Cham, 2024: pp. 41–59. https://doi.org/10.1007/978-3-031-76465-3_3.

[5] D.A. Burns, E.W. Ciurczak, eds., Handbook of Near-Infrared Analysis, 0 ed., CRC Press, 2007. https://doi.org/10.1201/9781420007374.

[6] Y. Ozaki, C. Huck, S. Tsuchikawa, S.B. Engelsen, Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications, Springer Nature, 2020.

[7] T.P. Czaja, S.B. Engelsen, Why nothing beats NIRS technology: The green analytical choice for the future sustainable food production, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 325 (2025) 125028. https://doi.org/10.1016/j.saa.2024.125028.

[8] J.M. Juran, Juran on Quality by Design: The New Steps for Planning Quality Into Goods and Services, Simon and Schuster, 1992.

[9] E. Skibsted, S.B. Engelsen, Spectroscopy for Process Analytical Technology (PAT), *Encyclopedia of Spectroscopy and Spectrometry* (2010) 2651–2661. https://doi.org/10.1016/B978-0-12-374413-5.00026-9.

[10] P. Iweka, S. Kawamura, T. Mitani, T. Kawaguchi, S. Koseki, Online Milk Quality Assessment during Milking Using Near-infrared Spectroscopic Sensing System, *Environmental Control in Biology* 58 (2020) 1–6. https://doi.org/10.2525/ecb.58.1.

[11] J.A. Diaz-Olivares, I. Adriaens, E. Stevens, W. Saeys, B. Aernouts, Online milk composition analysis with an on-farm near-infrared sensor*, Computers and Electronics in Agriculture* 178 (2020) 105734. https://doi.org/10.1016/j.compag.2020.105734.

[12] G. Tøgersen, T. Isaksson, B.N. Nilsen, E.A. Bakker, K.I. Hildrum, On-line NIR analysis of fat, water and protein in industrial scale ground meat batches, *Meat Science* 51 (1999) 97–102. https://doi.org/10.1016/S0309-1740(98)00106-5.

[13] C.E. Eskildsen, K.W. Sanden, S.G. Wubshet, P.V. Andersen, J. Øyaas, J.P. Wold, Estimating dry matter and fat content in blocks of Swiss cheese during production using on-line near infrared spectroscopy, *Journal of Near Infrared Spectroscopy* 27 (2019) 293–301. https://doi.org/10.1177/0967033519855436.

[14] Linee guida per l'utilizzo di tecniche spettroscopiche nella regione del visibile e vicino infrarosso nel settore agroalimentare: calibrazione quantitativa, *Ente Italiano di Normazione* (UNI), 2022. https://unistore.uni.com/uni-ts-11892-2022 (accessed December 9, 2025).

[15] Linee guida per l'utilizzo di tecniche spettroscopiche nella regione del visibile e vicino infrarosso nel settore agroalimentare: calibrazione qualitativa, *Ente Italiano di Normazione* (UNI), 2024. https://unistore.uni.com/uni-ts-11942-2024 (accessed December 9, 2025).

[16] S. Grassi, E. Casiraghi, Advances in NIR Spectroscopy Analytical Technology in Food Industries, *Foods* 11 (2022) 1250. https://doi.org/10.3390/foods11091250.

[17] S. Grassi, C. Alamprese, Advances in NIR spectroscopy applied to process analytical technology in food industries, *Current Opinion in Food Science* 22 (2018) 17–21. https://doi.org/10.1016/j.cofs.2017.12.008.

[18] C. Alamprese, S. Grassi, Food fermentations: NIR spectroscopy as a tool for process analytical technology, *Advances in Food and Nutrition Research* 115 (2025) 391–430. https://doi.org/10.1016/bs.afnr.2025.06.002.

[19] G. Gorla, A. Ferrer, B. Giussani, Process understanding and monitoring: A glimpse into data strategies for miniaturized NIR spectrometers, *Analytica Chimica Acta* 1281 (2023) 341902. https://doi.org/10.1016/j.aca.2023.341902.

[20] G. Gorla, S. Fumagalli, J.J. Jansen, B. Giussani, Acquisition strategies for fermentation processes with a low-cost miniaturized NIR-spectrometer from scratch: Issues and challenges, *Microchemical Journal* 183 (2022) 108035. https://doi.org/10.1016/j.microc.2022.108035.

[21] I. Locatelli, D. Pedrali, S. Grassi, S. Buratti, A. Giorgi, L. Giupponi, Progress in quality assessment of Italian saffron, *Scientific Reports* 15 (2025) 2175. https://doi.org/10.1038/s41598-025-86440-x.

[22] G. Foca, C. Ferrari, A. Ulrici, G. Sciutto, S. Prati, S. Morandi, M. Brasca, P. Lavermicocca, S. Lanteri, P. Oliveri, The potential of spectral and hyperspectral-imaging techniques for bacterial detection in food: A case study on lactic acid bacteria, *Talanta* 153 (2016) 111–119. https://doi.org/10.1016/j.talanta.2016.02.059.

[23] P. Mishra, F. Marini, B. Brouwer, J.M. Roger, A. Biancolillo, E. Woltering, E.H. Echtelt, Sequential fusion of information from two portable spectrometers for improved prediction of moisture and soluble solids content in pear fruit, *Talanta* 223 (2021) 121733. https://doi.org/10.1016/j.talanta.2020.121733.

[24] C. Menozzi, G. Foca, R. Calvini, L. Catellani, A. Bezzecchi, A. Ulrici, Comparison of Different Spectral Ranges to Monitor Alcoholic and Acetic Fermentation of Red Grape Must Using FT-NIR Spectroscopy and PLS Regression, *Food Analytical Methods* 17 (2024) 1171–1182. https://doi.org/10.1007/s12161-024-02636-3.

[25] Y. Lu, Y. Huang, R. Lu, Innovative Hyperspectral Imaging-Based Techniques for Quality Evaluation of Fruits and Vegetables: A Review*, Applied Sciences* 7 (2017) 189. https://doi.org/10.3390/app7020189.

[26] D. Wu, D.-W. Sun, Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part I: Fundamentals, *Innovative Food Science & Emerging Technologies* 19 (2013) 1–14. https://doi.org/10.1016/j.ifset.2013.04.014.

[27] J. Qin, Hyperspectral Imaging Instruments, in: *Hyperspectral Imaging for Food Quality Analysis and Control*, Academic Press, 2010: pp. 129–172. https://doi.org/10.1016/B978-0-12-374753-2.10005-X.

[28] J.M. Amigo, Hyperspectral and multispectral imaging: setting the scene, in: *Data Handling in Science and Technology*, Elsevier, 2019: pp. 3–16. https://doi.org/10.1016/B978-0-444-63977-6.00001-8.

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

[30] J.S. MacDonald, S.L. Ustin, M.E. Schaepman, The contributions of Dr. Alexander F.H. Goetz to imaging spectrometry, *Remote Sensing of Environment* 113 (2009) S2–S4. https://doi.org/10.1016/j.rse.2008.10.017.

[31] D. Tanzilli, M. Cocchi, J.M. Amigo, A. D'Alessandro, L. Strani, Does hyperspectral always matter? A critical assessment of near infrared versus hyperspectral near infrared in the study of heterogeneous samples, *Current Research in Food Science* 9 (2024) 100813. https://doi.org/10.1016/j.crfs.2024.100813.

[32] A. Gowen, C. Odonnell, P. Cullen, G. Downey, J. Frias, Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, *Trends in Food Science & Technology* 18 (2007) 590–598. https://doi.org/10.1016/j.tifs.2007.06.001.

[33] G. ElMasry, D.-W. Sun, P. Allen, Near-infrared hyperspectral imaging for predicting colour, pH and tenderness of fresh beef, *Journal of Food Engineering* 110 (2012) 127–140. https://doi.org/10.1016/j.jfoodeng.2011.11.028.

[34] G. ElMasry, D.-W. Sun, P. Allen, Chemical-free assessment and mapping of major constituents in beef using hyperspectral imaging, *Journal of Food Engineering* 117 (2013) 235–246. https://doi.org/10.1016/j.jfoodeng.2013.02.016.

[35] S. León-Ecay, A. López-Maestresalas, M.T. Murillo-Arbizu, M.J. Beriain, J.A. Mendizabal, S. Arazuri, C. Jarén, P.D. Bass, M.J. Colle, D. García, M. Romano-Moreno, K. Insausti, Classification of Beef longissimus thoracis Muscle Tenderness Using Hyperspectral Imaging and Chemometrics, *Foods* 11 (2022) 3105. https://doi.org/10.3390/foods11193105.

[36] 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, *Analytica Chimica Acta* 714 (2012) 57–67. https://doi.org/10.1016/j.aca.2011.11.037.

[37] C.R. Craigie, P.L. Johnson, P.R. Shorten, A. Charteris, G. Maclennan, M.L. Tate, M.P. Agnew, K.R. Taukiri, A.D. Stuart, M.M. Reis, Application of Hyperspectral imaging to predict the pH, intramuscular fatty acid content and composition of lamb M. longissimus lumborum at 24 h post mortem, *Meat Science* 132 (2017) 19–28. https://doi.org/10.1016/j.meatsci.2017.04.010.

[38] M. Kamruzzaman, G. ElMasry, D.-W. Sun, P. Allen, Non-destructive prediction and visualization of chemical composition in lamb meat using NIR hyperspectral imaging and multivariate regression, *Innovative Food Science & Emerging Technologies* 16 (2012) 218–226. https://doi.org/10.1016/j.ifset.2012.06.003.

[39] F. Ma, B. Zhang, W. Wang, P. Li, X. Niu, C. Chen, L. Zheng, Potential use of multispectral imaging technology to identify moisture content and water-holding capacity in cooked pork sausages, *Journal of the Science of Food and Agriculture* 98 (2018) 1832–1838. https://doi.org/10.1002/jsfa.8659.

[40] Q. Huang, H. Li, J. Zhao, G. Huang, Q. Chen, Non-destructively sensing pork quality using near infrared multispectral imaging technique, *RSC Adv.* 5 (2015) 95903–95910. https://doi.org/10.1039/C5RA18872E.

[41] D.F. Barbin, G. ElMasry, D.-W. Sun, P. Allen, Predicting quality and sensory attributes of pork using near-infrared hyperspectral imaging, *Analytica Chimica Acta* 719 (2012) 30–42. https://doi.org/10.1016/j.aca.2012.01.004.

[42] D.F. Barbin, G. ElMasry, D.-W. Sun, P. Allen, Non-destructive determination of chemical composition in intact and minced pork using near-infrared hyperspectral imaging, *Food Chemistry* 138 (2013) 1162–1171. https://doi.org/10.1016/j.foodchem.2012.11.120.

[43] J.-H. Cheng, D.-W. Sun, J.-H. Qu, H.-B. Pu, X.-C. Zhang, Z. Song, X. Chen, H. Zhang, Developing a multispectral imaging for simultaneous prediction of freshness indicators during chemical spoilage of grass carp fish fillet, *Journal of Food Engineering* 182 (2016) 9–17. https://doi.org/10.1016/j.jfoodeng.2016.02.004.

[44] A.R. Sigurðardóttir, H.I. Sveinsdóttir, N. Schultz, H. Einarsson, M. Gudjónsdóttir, Multispectral imaging as a predictive tool for freshness of whole Atlantic cod: Compared with sensory, chemical and microbiological analysis, *Applied Food Research* 5 (2025) 101130. https://doi.org/10.1016/j.afres.2025.101130.

[45] J.-H. Cheng, J.-H. Qu, D.-W. Sun, X.-A. Zeng, Visible/near-infrared hyperspectral imaging prediction of textural firmness of grass carp (Ctenopharyngodon idella) as affected by frozen storage, *Food Research International* 56 (2014) 190–198. https://doi.org/10.1016/j.foodres.2013.12.009.

[46] H.-J. He, D. Wu, D.-W. Sun, Potential of hyperspectral imaging combined with chemometric analysis for assessing and visualising tenderness distribution in raw farmed salmon fillets, *Journal of Food Engineering* 126 (2014) 156–164. https://doi.org/10.1016/j.jfoodeng.2013.11.015.

[47] R. Lu, Multispectral imaging for predicting firmness and soluble solids content of apple fruit, *Postharvest Biology and Technology* 31 (2004) 147–157. https://doi.org/10.1016/j.postharvbio.2003.08.006.

[48] P. Rajkumar, N. Wang, G. EImasry, G.S.V. Raghavan, Y. Gariepy, Studies on banana fruit quality and maturity stages using hyperspectral imaging, *Journal of Food Engineering* 108 (2012) 194–200. https://doi.org/10.1016/j.jfoodeng.2011.05.002.

[49] S. Qiao, Y. Tian, W. Gu, K. He, P. Yao, S. Song, J. Wang, H. Wang, F. Zhang, Research on simultaneous detection of SSC and FI of blueberry based on hyperspectral imaging combined MS-SPA, *Engineering in Agriculture, Environment and Food* 12 (2019) 540–547. https://doi.org/10.1016/j.eaef.2019.11.006.

[50] G.A. Leiva-Valenzuela, R. Lu, J.M. Aguilera, Assessment of internal quality of blueberries using hyperspectral transmittance and reflectance images with whole spectra or selected wavelengths, *Innovative Food Science & Emerging Technologies* 24 (2014) 2–13. https://doi.org/10.1016/j.ifset.2014.02.006.

[51] W. Luo, G. Fan, P. Tian, W. Dong, H. Zhang, B. Zhan, Spectrum classification of citrus tissues infected by fungi and multispectral image identification of early rotten oranges, Spectrochimica Acta Part A: *Molecular and Biomolecular Spectroscopy* 279 (2022) 121412. https://doi.org/10.1016/j.saa.2022.121412.

[52] S. Munera, J.M. Amigo, J. Blasco, S. Cubero, P. Talens, N. Aleixos, Ripeness monitoring of two cultivars of nectarine using VIS-NIR hyperspectral reflectance imaging, *Journal of Food Engineering* 214 (2017) 29–39. https://doi.org/10.1016/j.jfoodeng.2017.06.031.

[53] B. Li, B. Hou, D. Zhang, Y. Zhou, M. Zhao, R. Hong, Y. Huang, Pears characteristics (soluble solids content and firmness prediction, varieties) testing methods based on visible-near infrared hyperspectral imaging, *Optik* 127 (2016) 2624–2630. https://doi.org/10.1016/j.ijleo.2015.11.193.

[54] J. Muñoz-Postigo, E.M. Valero, M.A. Martínez-Domingo, F.J. Lara, J.L. Nieves, J. Romero, J. Hernández-Andrés, Band selection pipeline for maturity stage classification in bell peppers: From full spectrum to simulated camera data, *Journal of Food Engineering* 365 (2024) 111824. https://doi.org/10.1016/j.jfoodeng.2023.111824.

[55] Z. Schmilovitch, T. Ignat, V. Alchanatis, J. Gatker, V. Ostrovsky, J. Felföldi, Hyperspectral imaging of intact bell peppers, Biosystems *Engineering* 117 (2014) 83–93. https://doi.org/10.1016/j.biosystemseng.2013.07.003.

[56] W. Huang, J. Li, Q. Wang, L. Chen, Development of a multispectral imaging system for online detection of bruises on apples, *Journal of Food Engineering* 146 (2015) 62–71. https://doi.org/10.1016/j.jfoodeng.2014.09.002.

[57] P.M. Mehl, Y.-R. Chen, M.S. Kim, D.E. Chan, Development of hyperspectral imaging technique for the detection of apple surface defects and contaminations, *Journal of Food Engineering* 61 (2004) 67–81. https://doi.org/10.1016/S0260-8774(03)00188-2.

[58] P. Baranowski, W. Mazurek, J. Pastuszka-Woźniak, Supervised classification of bruised apples with respect to the time after bruising on the basis of hyperspectral imaging data, *Postharvest Biology and Technology 86* (2013) 249–258. https://doi.org/10.1016/j.postharvbio.2013.07.005.

[59] Y. Bu, J. Luo, J. Li, Q. Chi, W. Guo, Detection of hidden bruises on kiwifruit using hyperspectral imaging combined with deep learning, *International Journal of Food Science and Technology* 59 (2024) 5975–5984. https://doi.org/10.1111/ijfs.17256.

[60] Q. Lü, M. Tang, Detection of Hidden Bruise on Kiwi fruit Using Hyperspectral Imaging and Parallelepiped Classification, *Procedia Environmental Sciences* 12 (2012) 1172–1179. https://doi.org/10.1016/j.proenv.2012.01.404.

[61] C. Esquerre, A.A. Gowen, G. Downey, C.P. O'Donnell, Wavelength Selection for Development of a near Infrared Imaging System for Early Detection of Bruise Damage in Mushrooms (*Agaricus bisporus*), *Journal of Near Infrared Spectroscopy* 20 (2012) 537–546. https://doi.org/10.1255/jnirs.1014.

[62] A. Gowen, C.P. O'Donnell, M. Taghizadeh, P.J. Cullen, J.M. Frias, G. Downey, Hyperspectral imaging combined with principal component analysis for bruise damage detection on white mushrooms (Agaricus bisporus), *Journal of Chemometrics* 22 (2008) 259–267. https://doi.org/10.1002/cem.1127.

[63] W.-H. Lee, M.S. Kim, H. Lee, S.R. Delwiche, H. Bae, D.-Y. Kim, B.-K. Cho, Hyperspectral near-infrared imaging for the detection of physical damages of pear, *Journal of Food Engineering* 130 (2014) 1–7. https://doi.org/10.1016/j.jfoodeng.2013.12.032.

[64] Y. Li, S. You, S. Wu, M. Wang, J. Song, W. Lan, K. Tu, L. Pan, Exploring the limit of detection on early implicit bruised 'Korla' fragrant pears using hyperspectral imaging features and spectral variables, *Postharvest Biology and Technology* 208 (2024) 112668. https://doi.org/10.1016/j.postharvbio.2023.112668.

[65] I. Orrillo, J.P. Cruz-Tirado, A. Cardenas, M. Oruna, A. Carnero, D.F. Barbin, R. Siche, Hyperspectral imaging as a powerful tool for identification of papaya seeds in black pepper, *Food Control* 101 (2019) 45–52. https://doi.org/10.1016/j.foodcont.2019.02.036.

[66] A.I. Ropodi, D.E. Pavlidis, F. Mohareb, E.Z. Panagou, G.-J.E. Nychas, Multispectral image analysis approach to detect adulteration of beef and pork in raw meats, *Food Research International* 67 (2015) 12–18. https://doi.org/10.1016/j.foodres.2014.10.032.

[67] M. Kamruzzaman, Y. Makino, S. Oshita, S. Liu, Assessment of Visible Near-Infrared Hyperspectral Imaging as a Tool for Detection of Horsemeat Adulteration in Minced Beef, *Food Bioprocess Technology* 8 (2015) 1054–1062. https://doi.org/10.1007/s11947-015-1470-7.

[68] M. Kamruzzaman, Y. Makino, S. Oshita, Rapid and non-destructive detection of chicken adulteration in minced beef using visible near-infrared hyperspectral imaging and machine learning, *Journal of Food Engineering* 170 (2016) 8–15. https://doi.org/10.1016/j.jfoodeng.2015.08.023.

[69] S. León-Ecay, K. Insausti, S. Arazuri, I. Goenaga, A. López-Maestresalas, Combination of spectral and textural features of hyperspectral imaging for the authentication of the diet supplied to fattening cattle, *Food Control* 159 (2024) 110284. https://doi.org/10.1016/j.foodcont.2024.110284.

[70] R. Calvini, J.M. Amigo, A. Ulrici, Transferring results from NIR-hyperspectral to NIR-multispectral imaging systems: A filter-based simulation applied to the classification of Arabica and Robusta green coffee, *Analytica Chimica Acta* 967 (2017) 33–41. https://doi.org/10.1016/j.aca.2017.03.011.

[71] S. Minaei, S. Shafiee, G. Polder, N. Moghadam-Charkari, S. van Ruth, M. Barzegar, J. Zahiri, M. Alewijn, P.M. Kuś, VIS/NIR imaging application for honey floral origin determination, *Infrared Physics & Technology* 86 (2017) 218–225. https://doi.org/10.1016/j.infrared.2017.09.001.

[72] D. Lorente, M. Zude, C. Regen, L. Palou, J. Gómez-Sanchis, J. Blasco, Early decay detection in citrus fruit using laser-light backscattering imaging, *Postharvest Biology and Technology* 86 (2013) 424–430. https://doi.org/10.1016/j.postharvbio.2013.07.021.

[73] J. Blasco, D. Lorente, V. Cortes, P. Talens, S. Cubero, S. Munera, N. Aleixos, Application of near Infrared Spectroscopy to the Quality Control of Citrus Fruits and Mango, *NIR News* 27 (2016) 4–7. https://doi.org/10.1255/nirn.1637.

[74] J. Gómez-Sanchis, J. Blasco, E. Soria-Olivas, D. Lorente, P. Escandell-Montero, J.M. Martínez-Martínez, M. Martínez-Sober, N. Aleixos, Hyperspectral LCTF-based system for classification of decay in mandarins caused by *Penicillium digitatum* and *Penicillium italicum* using the most relevant bands and non-linear classifiers, *Postharvest Biology and Technology* 82 (2013) 76–86. https://doi.org/10.1016/j.postharvbio.2013.02.011.

[75] A. Del Fiore, M. Reverberi, A. Ricelli, F. Pinzari, S. Serranti, A.A. Fabbri, G. Bonifazi, C. Fanelli, Early detection of toxigenic fungi on maize by hyperspectral imaging analysis, *International Journal of Food Microbiology* 144 (2010) 64–71. https://doi.org/10.1016/j.ijfoodmicro.2010.08.001.

[76] X. Chu, W. Wang, S.-C. Yoon, X. Ni, G.W. Heitschmidt, Detection of aflatoxin B1 (AFB1) in individual maize kernels using short wave infrared (SWIR) hyperspectral imaging, *Biosystems Engineering* 157 (2017) 13–23. https://doi.org/10.1016/j.biosystemseng.2017.02.005.

[77] B. Park, M. Kise, K.C. Lawrence, W.R. Windham, D.P. Smith, C.N. Thai, Real-time multispectral imaging system for online poultry fecal inspection using UML, *Optics East*, Boston (MA), 2006: p. 63810W. https://doi.org/10.1117/12.686379.

[78] L. Huang, J. Zhao, Q. Chen, Y. Zhang, Rapid detection of total viable count (TVC) in pork meat by hyperspectral imaging, *Food Research International* 54 (2013) 821–828. https://doi.org/10.1016/j.foodres.2013.08.011.

[79] B. Zhang, S. Fan, J. Li, W. Huang, C. Zhao, M. Qian, L. Zheng, Detection of Early Rottenness on Apples by Using Hyperspectral Imaging Combined with Spectral Analysis and Image Processing, *Food Analytical Methods* 8 (2015) 2075–2086. https://doi.org/10.1007/s12161-015-0097-7.

[80] A. Siedliska, P. Baranowski, M. Zubik, W. Mazurek, B. Sosnowska, Detection of fungal infections in strawberry fruit by VNIR/SWIR hyperspectral imaging, *Postharvest Biology and Technology* 139 (2018) 115–126. https://doi.org/10.1016/j.postharvbio.2018.01.018.

[81] B. Zhang, Y. Ou, S. Yu, Y. Liu, Y. Liu, W. Qiu, Gray mold and anthracnose disease detection on strawberry leaves using hyperspectral imaging, *Plant Methods* 19 (2023) 148. https://doi.org/10.1186/s13007-023-01123-w.

[82] J.M. Amigo, S. Grassi, Configuration of hyperspectral and multispectral imaging systems, in: *Data Handling in Science and Technology*, Elsevier, 2019: pp. 17–34. https://doi.org/10.1016/B978-0-444-63977-6.00002-X.

[83] H.F. Grahn, P. Geladi, eds., Techniques and Applications of Hyperspectral Image Analysis, 1st ed., Wiley, 2007. https://doi.org/10.1002/9780470010884.

[84] D.P. Ariana, R. Lu, Quality evaluation of pickling cucumbers using hyperspectral reflectance and transmittance imaging—Part II. Performance of a prototype, *Sensing and Instrumentation for Food Quality and Safety* 2 (2008) 152–160. https://doi.org/10.1007/s11694-008-9058-9.

[85] D.P. Ariana, R. Lu, Detection of Internal Defect in Pickling Cucumbers Using Hyperspectral Transmittance Imaging, *Transactions of the ASABE* 51 (2008) 705–713. https://doi.org/10.13031/2013.24367.

[86] Y. Lu, W. Saeys, M. Kim, Y. Peng, R. Lu, Hyperspectral imaging technology for quality and safety evaluation of horticultural products: A review and celebration of the past 20-year progress, *Postharvest Biology and Technology* 170 (2020) 111318. https://doi.org/10.1016/j.postharvbio.2020.111318.

[87] M.-H. Hu, Q.-L. Dong, B.-L. Liu, U.L. Opara, Prediction of mechanical properties of blueberry using hyperspectral interactance imaging, *Postharvest Biology and Technology* 115 (2016) 122–131. https://doi.org/10.1016/j.postharvbio.2015.11.021.

[88] J.M. Prats-Montalbán, A. De Juan, A. Ferrer, Multivariate image analysis: A review with applications, *Chemometrics and Intelligent Laboratory Systems* 107 (2011) 1–23. https://doi.org/10.1016/j.chemolab.2011.03.002.

[89] P. Geladi, H.F. Grahn, Multivariate Image Analysis, in: R.A. Meyers (Ed.), Encyclopedia of Analytical Chemistry, 1st ed., Wiley, 2000. https://doi.org/10.1002/9780470027318.a8106.

[90] Å. Rinnan, F.V.D. Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, *TrAC Trends in Analytical Chemistry* 28 (2009) 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007.

[91] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra, *Applied Spectroscoscopy* 43 (1989) 772–777. https://doi.org/10.1366/0003702894202201.

[92] Abraham. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, *Analytical Chemistry* 36 (1964) 1627–1639. https://doi.org/10.1021/ac60214a047.

[93] J.M. Amigo, C. Santos, Preprocessing of hyperspectral and multispectral images, in: *Data Handling in Science and Technology*, Elsevier, 2019: pp. 37–53. https://doi.org/10.1016/B978-0-444-63977-6.00003-1.

[94] J. Buendia Garcia, J. Gornay, M. Lacoue-Negre, S. Mas Garcia, J. Er-Rmyly, R. Bendoula, J.-M. Roger, A novel methodology for determining effectiveness of preprocessing methods in reducing undesired spectral variability in near infrared spectra, *Journal of Near Infrared Spectroscopy* 30 (2022) 74–88. https://doi.org/10.1177/09670335211047959.

[95] N.K. Afseth, A. Kohler, Extended multiplicative signal correction in vibrational spectroscopy, a tutorial, *Chemometrics and Intelligent Laboratory Systems* 117 (2012) 92–99. https://doi.org/10.1016/j.chemolab.2012.03.004.

[96] P. Gemperline, ed., Practical Guide To Chemometrics, 2 ed., *CRC Press*, 2006. https://doi.org/10.1201/9781420018301.

[97] P. Geladi, H. Isaksson, L. Lindqvist, S. Wold, K. Esbensen, Principal component analysis of multivariate images, *Chemometrics and Intelligent Laboratory Systems* 5 (1989) 209–220. https://doi.org/10.1016/0169-7439(89)80049-8.

[98] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, *Chemometrics and Intelligent Laboratory Systems* 2 (1987) 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.

[99] R. Bro, A.K. Smilde, Principal component analysis, *Analytical Methods* 6 (2014) 2812–2831. https://doi.org/10.1039/C3AY41907J.

[100] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, *Analytical Methods* 5 (2013) 3790. https://doi.org/10.1039/c3ay40582f.

[101] E. Lopez, J. Etxebarria-Elezgarai, J.M. Amigo, A. Seifert, The importance of choosing a proper validation strategy in predictive models. A tutorial with real examples, *Analytica Chimica Acta* 1275 (2023) 341532. https://doi.org/10.1016/j.aca.2023.341532.

[102] E. Szymańska, E. Saccenti, A.K. Smilde, J.A. Westerhuis, Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, *Metabolomics* 8 (2012) 3–16. https://doi.org/10.1007/s11306-011-0330-3

[103] . Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, *Chemometrics and Intelligent Laboratory Systems* 174 (2018) 33–44. https://doi.org/10.1016/j.chemolab.2017.12.004.

[104] S. Wold, Pattern recognition by means of disjoint principal components models, *Pattern Recognition* 8 (1976) 127–139. https://doi.org/10.1016/0031-3203(76)90014-5.

[105] M. Barker, W. Rayens, Partial least squares for discrimination, *Journal of Chemometrics* 17 (2003) 166–173. https://doi.org/10.1002/cem.785.

[106] R. Gosselin, D. Rodrigue, C. Duchesne, A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications, Chemometrics *and Intelligent Laboratory Systems* 100 (2010) 12–21. https://doi.org/10.1016/j.chemolab.2009.09.005.

[107] R. Calvini, G. Orlandi, G. Foca, A. Ulrici, Development of a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging, *Journal of Spectral Imaging* (2018) a13. https://doi.org/10.1255/jsi.2018.a13.

[108] J. Blasco, G. Gorla, S. Munera, R. Vitale, J.M. Amigo, Non-Destructive Spectral Systems (NDSS) for modern inspection systems in real-time: challenges and industrial perspectives, *TrAC Trends in Analytical Chemistry* 191 (2025) 118369. https://doi.org/10.1016/j.trac.2025.118369.

[109] E.W. Steyerberg, F.E. Harrell, Prediction models need appropriate internal, internal–external, and external validation, *Journal of Clinical* Epidemiology 69 (2016) 245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005.

[110] J. Ezenarro, XYOnion: a layer-based method for splitting datasets into calibration and validation subsets, *Analytica Chimica Acta* 1364 (2025) 344229. https://doi.org/10.1016/j.aca.2025.344229.

[111] K.H. Esbensen, P. Geladi, Principles of Proper Validation: use and abuse of re-sampling for validation, *Journal of Chemometrics* 24 (2010) 168–187. https://doi.org/10.1002/cem.1310.

[112] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, *Technometrics* 11 (1969) 137–148. https://doi.org/10.1080/00401706.1969.10490666.

[113] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval Partial Least-Squares Regression (i PLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy, *Applied Spectroscopy* 54 (2000) 413–419. https://doi.org/10.1366/0003702001949500.

[114] R. Calvini, A. Ulrici, J.M. Amigo, Practical comparison of sparse methods for classification of Arabica and Robusta coffee species using near infrared hyperspectral imaging, *Chemometrics and Intelligent Laboratory Systems* 146 (2015) 503–511. https://doi.org/10.1016/j.chemolab.2015.07.010.

[115] P. Filzmoser, M. Gschwandtner, V. Todorov, Review of sparse methods in regression and classification with application to chemometrics, *Journal of Chemometrics* 26 (2012) 42–51. https://doi.org/10.1002/cem.1418.

[116] R. Calvini, J.M. Amigo, Coupling randomisation and sparse modelling for the exploratory analysis of large hyperspectral datasets, *Chemometrics and Intelligent Laboratory Systems* 248 (2024) 105118. https://doi.org/10.1016/j.chemolab.2024.105118.

[117] R. Calvini, A. Ulrici, J.M. Amigo, Sparse-Based Modeling of Hyperspectral Data, in: *Data Handling in Science and Technology*, Elsevier, 2016: pp. 613–634. https://doi.org/10.1016/B978-0-444-63638-6.00019-X.

[118] M.A. Rasmussen, R. Bro, A tutorial on the Lasso approach to sparse modeling, *Chemometrics and Intelligent Laboratory Systems* 119 (2012) 21–31. https://doi.org/10.1016/j.chemolab.2012.10.003.

[119] R. Bro, E.E. Papalexakis, E. Acar, N.D. Sidiropoulos, Coclustering—a useful tool for chemometrics, *Journal of Chemometrics* 26 (2012) 256–263. https://doi.org/10.1002/cem.1424.

[120] N.C. Basantia, L.M.L. Nollet, M. Kamruzzaman, eds., Hyperspectral Imaging Analysis and Applications for Food Quality, 1st ed., *CRC Press*, 2018. https://doi.org/10.1201/9781315209203.

[121] S. Gariglio, R.R. de Oliveira, G. Canali, C. Malegori, P. Malaspina, M. Casale, P. Oliveri, P. Giordani, NIR Hyperspectral Imaging Combined with Chemometrics for Mapping Water Patterns During Dehydration of Nonvascular Epiphytic Communities, *Journal of Analysis and Testing* (2025). https://doi.org/10.1007/s41664-025-00384-9.

[122] C. Menozzi, J.M. Prats-Montalbán, R. Calvini, A. Ulrici, Comparison of colour and texture feature extraction methods to predict anthocyanins content in Sangiovese grapes, *Chemometrics and Intelligent Laboratory Systems* 263 (2025) 105446. https://doi.org/10.1016/j.chemolab.2025.105446.

[123] A. Giraudo, R. Calvini, G. Orlandi, A. Ulrici, F. Geobaldo, F. Savorani, Development of an automated method for the identification of defective hazelnuts based on RGB image analysis and colourgrams, *Food Control* 94 (2018) 233–240. https://doi.org/10.1016/j.foodcont.2018.07.018.

[124] S. Kucheryavski, K.H. Esbensen, A. Bogomolov, Monitoring of pellet coating process with image analysis — a feasibility study, *Journal of Chemometrics* 24 (2010) 472–480. https://doi.org/10.1002/cem.1292.

[125] S. Kucheryavskiy, A new approach for discrimination of objects on hyperspectral images, *Chemometrics and Intelligent Laboratory Systems* 120 (2013) 126–135. https://doi.org/10.1016/j.chemolab.2012.11.009.

[126] P. Oliveri, C. Malegori, M. Casale, E. Tartacca, G. Salvatori, An innovative multivariate strategy for HSI-NIR images to automatically detect defects in green coffee, *Talanta* 199 (2019) 270–276. https://doi.org/10.1016/j.talanta.2019.02.049.

[127] A. Antonelli, M. Cocchi, P. Fava, G. Foca, G.C. Franchini, D. Manzini, A. Ulrici, Automated evaluation of food colour by means of multivariate image analysis coupled to a wavelet-based classification algorithm, *Analytica Chimica Acta* 515 (2004) 3–13. https://doi.org/10.1016/j.aca.2004.01.005.

[128] C. Ferrari, G. Foca, R. Calvini, A. Ulrici, Fast exploration and classification of large hyperspectral image datasets for early bruise detection on apples, *Chemometrics and Intelligent Laboratory Systems,* 146 108–119. https://doi.org/10.1016/j.chemolab.2015.05.016.

[129] C. Ferrari, G. Foca, A. Ulrici, Handling large datasets of hyperspectral images: Reducing data size without loss of useful information, *Analytica Chimica Acta* 802 (2013) 29–39. https://doi.org/10.1016/j.aca.2013.10.009.

[130] R. Calvini, S. Michelini, V. Pizzamiglio, G. Foca, A. Ulrici, Evaluation of the effect of factors related to preparation and composition of grated Parmigiano Reggiano cheese using NIR hyperspectral imaging, *Food Control* 131 (2022) 108412. https://doi.org/10.1016/j.foodcont.2021.108412.

[131] R. Calvini, G. Foca, A. Ulrici, Data dimensionality reduction and data fusion for fast characterization of green coffee samples using hyperspectral sensors, *Analytical and Bioanalytical Chemistry 408* (2016) 7351–7366. https://doi.org/10.1007/s00216-016-9713-7.

[132] 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.

[133] L. Pieszczek, M. Daszykowski, Integrating hyperspectrograms with class modeling techniques for the construction of an effective expert system: Quality control of pharmaceutical tablets based on near-infrared hyperspectral imaging, *Journal of Pharmaceutical and Biomedical Analysis,* 256 (2025) 116697. https://doi.org/10.1016/j.jpba.2025.116697.

[134] J.F.I. Nturambirwe, W.J. Perold, U.L. Opara, Classification Learning of Latent Bruise Damage to Apples Using Shortwave Infrared Hyperspectral Imaging, *Sensors* 21 (2021) 4990. https://doi.org/10.3390/s21154990.

Chapter 3

From Farm to Sorting:

NIR Spectral Imaging for the management of the Brown Marmorated Stink Bug pest

3.1. Background and Aim

Globalization and intensification of human activities have driven the proliferation of invasive alien pests, which seriously affect agroecosystems, undermine agri-food production, and result in significant economic losses. In this context, *Halyomorpha halys*, commonly known as the Brown Marmorated Stink Bug (BMSB), represents one of the most worrisome threats [1].

BMSB feeding activity causes several defects, including deformities, scars, discoloration, and pitting, making a wide range of harvested products unmarketable. However, BMSB management is particularly challenging due to its high reproductive potential, mobility, polyphagia, and the limited effectiveness of available broad-spectrum insecticides, whose use also raises environmental concerns [2–4]. Given the high damage potential of this pest, effective management strategies are required both upstream, through field monitoring aimed at early detection to promptly face infestations, and downstream, along the food supply chain, to ensure the quality of harvested products. Within the framework of the European project HALY.ID (**Figure 3.1**), the present chapter explores NIR spectral imaging as a reliable and comprehensive technology to address BMSB impact along the agri-food chain, from automated pest monitoring in the field to post-harvest sorting, throughout the detection of under-peel damages compromising organic pears' quality [5].

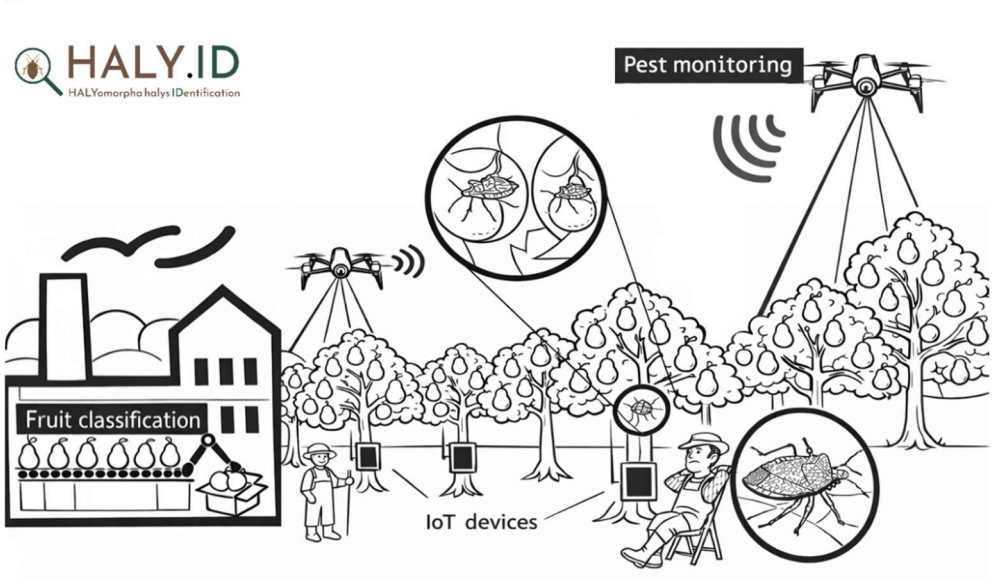


Figure 3.1 *Conceptualization of the HALY.ID project, adapted from [5]. A brief overview of the project' results can be also found in [6,7].*

With the objective of exploring innovative and sustainable strategies for pest management, **Section 3.2** focuses on the evaluation of Near-Infrared Hyperspectral Imaging (NIR-HSI) as a potential tool for field monitoring of BMSB [8]. Hyperspectral images of BMSB specimens acquired on different vegetal backgrounds were used to create a spectral library representative of both the bugs and the surrounding environment. Two complementary data analysis approaches were considered, focusing on spectral information and spatial features, respectively. In particular, chemometric models based on Soft PLS-DA combined with sparse variable selection were applied to identify the spectral regions most relevant for discrimination [9,10], paving the way for the potential implementation of multispectral imaging systems more suitable for practical on-field applications. Finally, spectral and spatial information were merged to enable automated object-level classification.

A substantial part of the research activities was dedicated to the investigation of under-peel damages on pears caused by BMSB feeding, which are not detectable using imaging systems operating in the visible range. In this context, **Section 3.3** focuses on the development of an automated and objective annotation strategy for the identification of Regions of Interest (ROIs) ascribable to BMSB punctures [11]. Since preliminary exploratory analysis using PCA was not able to clearly isolate punctures, due to their irregular shapes and blurred edges between sound and damaged tissues, an alternative annotation procedure was proposed. This approach combined the Common Space Hyperspectrograms (CSH) dimensionality reduction method with image-level classification and variable selection [12,13]. The spatial features most relevant for detecting the presence of punctures were then projected back onto the original image domain, enabling the automated identification of punctured regions.

Based on the annotated ROIs obtained in Section 3.3, **Section 3.4** addresses the development of supervised classification models for the discrimination between punctured and sound areas at the pixel level [14]. Representative spectra extracted from the annotated punctures and sound areas were used both for model calculation and for the identification of the most informative spectral regions. Finally, an additional threshold on contiguous pixels predicted as punctures was set to enable classification at the image level, providing a framework more closely aligned with the requirements of automated post-harvest sorting systems.

References

[1] M. Bariselli, R. Bugiani, L. Maistrello, Distribution and damage caused by *Halyomorpha halys* in Italy, *EPPO Bulletin* 46 (2016) 332–334. https://doi.org/10.1111/epp.12289.

[2] L. Maistrello, Case Study 2: *Halyomorpha halys* (Stål) in Europe, in: A.F. Bueno, A.R. Panizzi (Eds.), *Stink Bugs (Hemiptera: Pentatomidae) Research and Management: Recent Advances and Case Studies from Brazil, Europe, and USA*, Springer Nature Switzerland, Cham, 2024: pp. 271–359. https://doi.org/10.1007/978-3-031-69742-5_15.

[3] T.C. Leskey, A.L. Nielsen, Impact of the Invasive Brown Marmorated Stink Bug in North America and Europe: History, Biology, Ecology, and Management, *Annual Review of Entomology* 63 (2018) 599–618. https://doi.org/10.1146/annurev-ento-020117-043226.

[4] L. Maistrello, P. Dioli, M. Dutto, S. Volani, S. Pasquali, G. Gilioli, Tracking the Spread of Sneaking Aliens by Integrating Crowdsourcing and Spatial Modeling: The Italian Invasion of *Halyomorpha halys*, *BioScience* (2018). https://doi.org/10.1093/biosci/biy112.

[5] HALY.ID, HALY.ID – HALYomorpha halys IDentification: Innovative ICT tools for targeted monitoring and sustainable management of the brown marmorated stink bug and other pests, https://www.haly-id.eu/ (accessed January 7, 2026).

[6] L. Almstedt, F.B. Sorbelli, B. Boom, R. Calvini, E. Costi, A. Dinca, V. Ferrari, D. Giannetti, L. Ichim, A. Kargar, C. Lazar, L. Maistrello, A. Navarra, D. Niederprüm, P. Offermans, B. O'Flynn, L. Palazzetti, N. Patelli, C.M. Pinotti, D. Popescu, A.K. Rangarajan, L. Serghei, A. Ulrici, L. Wolf, D. Zorbas, L. Zurek, A Comprehensive Pest Monitoring System for Brown Marmorated Stink Bug, *IEEE Transactions on AgriFood Electronics* (2024) 1–11. https://doi.org/10.1109/TAFE.2024.3469538.

[7] L. Almstedt, F.B. Sorbelli, B. Boom, R. Calvini, E. Costi, A. Dinca, V. Ferrari, D. Giannetti, L. Ichim, A. Kargar, C. Lazar, L. Maistrello, A. Navarra, D. Niederprüm, P. Offermans, B. O'Flynn, L. Palazzetti, N. Patelli, C.M. Pinotti, D. Popescu, A.K. Rangarajan, L. Serghei, A. Ulrici, L. Wolf, D. Zorbas, L. Zurek, Beyond the Naked Eye: Computer Vision for Detecting Brown Marmorated Stink Bug and Its Punctures, *IEEE Transactions on AgriFood Electronics* (2024) 1–12. https://doi.org/10.1109/TAFE.2024.3429537.

[8] Ferrari, V., Calvini, R., Boom, B., Menozzi, C., Rangarajan, A.K., Maistrello, L., Offermans, P., Ulrici, A. Evaluation of the potential of near infrared hyperspectral imaging for monitoring the invasive brown marmorated stink bug, *Chemometrics and Intelligent Laboratory Systems* 234 (2023), 104751;

[9] R. Calvini, G. Orlandi, G. Foca, A. Ulrici, Development of a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging, *J. Spectral Imaging* (2018) a13. https://doi.org/10.1255/jsi.2018.a13.

[10] P. Filzmoser, M. Gschwandtner, V. Todorov, Review of sparse methods in regression and classification with application to chemometrics, *Journal of Chemometrics* 26 (2012) 42–51. https://doi.org/10.1002/cem.1418.

[11] Ferrari, V., Calvini, R., Menozzi, C., Costi, E., Giannetti, D., Offermans, P., Maistrello, L., Ulrici, A. NIR hyperspectral imaging to identify damage caused by *Halyomorpha halys* on pears: Automated identification of Regions of Interest related to punctured areas, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 343 (2025), 126543;

[12] R. Calvini, G. Foca, A. Ulrici, Data dimensionality reduction and data fusion for fast characterization of green coffee samples using hyperspectral sensors, *Analytical and Bioanalytical Chemistry* 408 (2016) 7351–7366. https://doi.org/10.1007/s00216-016-9713-7.

[13] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval Partial Least-Squares Regression (i PLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy, *Applied Spectroscopy* 54 (2000) 413–419. https://doi.org/10.1366/0003702001949500.

[14] Ferrari, V., Calvini, R., Menozzi, C., Costi, E., Offermans, P., Maistrello, L., Ulrici, A. NIR hyperspectral imaging to identify damage caused by *Halyomorpha halys* on pears: development of classification models, submitted for publication.

3.2. Evaluation of the potential of near infrared hyperspectral imaging for monitoring the invasive Brown Marmorated Stink Bug

What follows is the integral content of: Ferrari, V., Calvini, R., Boom, B., Menozzi, C., Rangarajan, A.K., Maistrello, L., Offermans, P., Ulrici, A. (2023). Evaluation of the potential of near infrared hyperspectral imaging for monitoring the invasive brown marmorated stink bug, *Chemometrics and Intelligent Laboratory Systems*, 234, 104751. DOI: 10.1016/j.chemolab.2023.104751

Evaluation of the potential of Near Infrared Hyperspectral Imaging for monitoring the invasive Brown Marmorated Stink Bug

Veronica Ferrari [a], Rosalba Calvini [a,*], Bas Boom [b], Camilla Menozzi [a], Aravind Krishnaswamy Rangarajan [b], Lara Maistrello [a], Peter Offermans [b], Alessandro Ulrici [a]

[a] *University of Modena and Reggio Emilia, Department of Life Sciences, Pad. Besta, Via Amendola, 2, 42122, Reggio Emilia, Italy*

[b] *IMEC OnePlanet, Bronland 10, Wageningen, The Netherlands*

**Corresponding author*

Abstract

The brown marmorated stink bug (BMSB), *Halyomorpha halys*, is an invasive insect pest of global importance that damages several crops, compromising agri-food production. Field monitoring procedures are fundamental to perform risk assessment operations, in order to promptly face crop infestations and avoid economical losses. To improve pest management, spectral cameras mounted on Unmanned Aerial Vehicles (UAVs) and other Internet of Things (IoT) devices, such as smart traps or unmanned ground vehicles, could be used as an innovative technology allowing fast, efficient and real-time monitoring of insect infestations.

The present study consists in a preliminary evaluation at the laboratory level of Near Infrared Hyperspectral Imaging (NIR-HSI) as a possible technology to detect BMSB specimens on different vegetal backgrounds, overcoming the problem of BMSB mimicry. Hyperspectral images of BMSB were acquired in the 980–1660 nm range, considering different vegetal backgrounds selected to mimic a real field application scene. Classification models were obtained following two different chemometric approaches. The first approach was focused on modelling spectral information and selecting relevant spectral regions for discrimination by means of sparse-based variable selection coupled with Soft Partial Least Squares Discriminant Analysis (s-Soft PLS-DA) classification

algorithm. The second approach was based on modelling spatial and spectral features contained in the hyperspectral images using Convolutional Neural Networks (CNN). Finally, to further improve BMSB detection ability, the two strategies were merged, considering only the spectral regions selected by s-Soft PLS-DA for CNN modelling.

Keywords: *Halyomorpha halys*, hyperspectral imaging, pest management, precision agriculture, multivariate image analysis

1. Introduction

In the last decades, the increase of anthropogenic activities determined the spread of invasive insect pests, which can seriously affect agroecosystems compromising agri-food production and resulting in severe economic losses. One of the most worrisome pests of global importance is the brown marmorated stink bug (BMSB), *Halyomorpha halys,* which causes serious damages to several agricultural crops [1]. The damage occurs mainly to fruits and seeds as a result of BMSB feeding activity: its piercing-sucking mouth apparatus determines deformities, scars, discolorations and pitting which make products unmarketable [2]. In Southern Europe, suitable climate and high density of crops provided excellent conditions for the establishment of large populations of BMSB [3]. As a result of BMSB activity, in 2019 economical losses in fruit orchards of Northern Italy were estimated to be equal to €590 million [4].

The management of BMSB is very challenging due to high reproductive potential, high mobility, polyphagy and ineffectiveness of available broad-spectrum insecticides, which determine a negative impact on the environment [5–8]. According to Integrated Pest Management (IPM) practices, field monitoring of insect pests is of fundamental importance to gain information about their presence and to timely adopt proper actions to face the infestation and avoid economical losses. Although it represents a crucial step, field monitoring is time and money consuming for farmers, since it requires direct field inspection by technicians [2,7].

With the aim of improving crop field pest management, automated monitoring systems based on spectral cameras mounted on Unmanned Aerial Vehicles (UAVs) and other Internet of Things (IoT) devices, such as smart traps or unmanned ground vehicles, can be used as an innovative technology allowing fast, efficient, and real-time monitoring [9–11]. As previous studies reported, multispectral imaging (MSI) and hyperspectral imaging (HSI) cameras in the visible and near infrared regions mounted over UAVs proved to be effective tools for a fast and reliable assessment of crop infection

and infestation [12,13]. Consequently, the time needed to detect the presence of possible crop infections is strongly reduced and it is possible to employ targeted pest management strategies.

The present study performed an initial assessment at the laboratory level of hyperspectral imaging as a possible method to identify BMSB specimens on different background types simulating a real field application scene. In this case, the effectiveness of spectral cameras working in the near infrared range (NIR) has been evaluated in order to overcome BMSB mimicry, as the brown marmorated colour makes this insect hardly detectable on dark brown vegetal backgrounds with spectral cameras operating in the visible range.

Hyperspectral images are three-dimensional matrices composed of one spectral (λ) and two spatial (x and y) dimensions, obtained by stacking together hundreds of (x, y) grey-scale images acquired at different successive wavelengths, λ. Therefore, this technique couples the advantages of spectroscopic methods with the possibility of visualizing spectral data at each pixel of an image, allowing the visualization of the chemical composition of the sample surface [14,15]. Despite the great potential of this technique, its data-richness represents at the same time the main advantage and disadvantage of HSI: a large amount of data permits a detailed representation of the analysed samples, but, at the same time, it involves issues related to data handling, storage and analysis.

Chemometric techniques are mandatory to unravel the curse of dimensionality in HSI and to develop classification or calibration models able to predict the qualitative or quantitative properties of interest from hyperspectral data [16]. Simple but effective applications of chemometric techniques to hyperspectral data include the use of linear calibration or classification methods, such as Partial Least Squares (PLS) or Partial Least Squares Discriminant Analysis (PLS-DA) [17,18]. An advantage of these methods consists in the fact that the models are easily interpretable, especially when it is necessary not only to obtain good model performances but also to highlight the relevant spectral regions for the problem at hand. In this context, it is also possible to apply spectral variable selection algorithms, which allows the subsequent elimination of spectral regions that are not pertinent, leading at the same time to better results in classification or calibration issues and to an increased chemical interpretation of the results [18,19]. Considering HSI applications, the identification of spectral bands relevant for the problem at hand allows to further develop faster and cheaper multispectral imaging systems. In MSI systems only a limited number of specific wavebands are considered, reducing the time needed for the acquisition and the efforts for data management. In addition, MSI systems are characterized by higher resistance and stability of the optical components, which makes them more suitable for application in the field [14,20].

Recent progresses in computer technology led to the development of advanced Deep Learning (DL) techniques, which require large datasets to be trained and are able to face complex applications.

Among DL methods, Artificial Neural Networks (ANNs) progressively emerged as valuable tools to solve complex and highly non-linear regression and classification tasks, besides becoming popular in countless fields [21].

In particular, Convolutional Neural Networks (CNNs) achieved state-of-the-art performances in the domain of Computer Vision applications. In this context, CCNs models have been successfully employed for image classification, object detection, and image segmentation [22]. To optimize CNNs modelling capabilities, multiple architectures have been proposed for image segmentation and classification [23–25] including the U-Net [25], which is of particular interest for the implementation on edge devices such as UAVs, due to its ability to achieve good model performances with reduced computation efforts [26]. U-Net was initially designed for computer vision applications, however recent works also demonstrated the advantages of this algorithm in the segmentation of hyperspectral images [27,28].

In contrast to PLS-DA, the U-Net employs both spatial and spectral information to perform the classification by virtue of its 2D convolutional layers which are designed to exploit spatial relationships between pixels. Because of this advantage, the U-Net is able to suppress false positive pixels by considering neighbouring pixels values, which should make the U-Net more robust for application in the field. In this study, NIR hyperspectral imaging was used to develop classification models able to discriminate BMSB specimens and the vegetal backgrounds following different Machine Learning (ML) strategies. Firstly, the spectral information was modelled using Soft Partial Least Squares - Discriminant Analysis (Soft PLS-DA), an extension of the classical PLS-DA algorithm, and by s-Soft PLS-DA [29], a version of Soft PLS-DA where sparse-based variable selection was also implemented [18,20, 30]. Then, a deep learning method based on U-Net was implemented and adapted for hyperspectral images, focusing on the spatial features of BMSB. Finally, in order to reduce the complexity of the network in the spectral dimension, the U-Net was adjusted to use the relevant spectral regions previously identified by s-Soft PLS-DA. This also allowed us to investigate whether merging together a spectral-based and a spatial-based method could result in improved classification performances.

2. Materials and Methods

2.1. Samples

BMSB specimens used in the present work were provided by the Applied Entomology Lab, University of Modena and Reggio Emilia. All the progenitor individuals had been previously captured in urban parks of the city of Reggio Emilia using the tree beating technique. The bugs were reared in

climatic chambers at 26 ◦C, 60% relative humidity, L16: D8 photoperiod, inside clear mesh cages (30 × 30 × 30 cm, approximately 40 individuals/cage) with organic tomatoes, carrots, green bean pods and raw peanuts as food. A bottle cap with a water-soaked cotton swab was used as water supply. Food and water were replaced twice per week.

The samples of vegetal backgrounds were collected in the University campus surroundings (Via Amendola, 2, San Maurizio, Reggio Emilia) on the same day of image acquisition.

To develop the classification models and to identify the spectral bands able to discriminate BMSB from the different backgrounds, 20 specimens of BMSB and seven types of vegetal backgrounds were considered, corresponding to green leaves, yellow leaves, dry leaves, grass, soil, bark and tree branches. These background types were selected to mimic real field conditions. BMSB specimens were randomly divided in five different groups (G1-G5) with 4 insects in each group; the bugs belonging to each group were kept together and always acquired in the same image of the different vegetal background types.

2.2. Image acquisition and elaboration

The hyperspectral images were acquired using a HSI line-scan system composed of a desktop NIR Spectral Scanner (DV Optic) embedding a Specim N17E reflectance imaging spectrometer, coupled to a Xenics XEVA 1.7–320 camera (320 × 256 pixels) embedding Specim Oles 31 f/2.0 optical lens and covering the spectral range from 900 to 1700 nm (5 nm resolution, 150 spectral channels). To enable a better evaluation of the stability of the acquisition system over time, a setup composed of a silicon carbide sandpaper as sample background, which is characterized by a very low and constant reflectance spectrum [31], a 99% reflectance standard, and two ceramic tiles with two different grayscale tones and intermediate reflectance values were used for the acquisition of all the images.

The raw data were then converted into reflectance values by applying the instrument calibration procedure based on the high-reflectance standard reference and on the estimate of the dark current [32]. Furthermore, to minimize the variability among images over time, an additional internal calibration was performed [33].

The wavelengths at the extremes of the spectral range are characterized by low S/N values and were excluded, considering only the spectral range between 980 and 1660 nm (137 wavelengths) for further analysis.

On the whole, 35 hyperspectral images were acquired (= 7 vegetal backgrounds × 5 BMSB groups). The BMSB specimens belonging to each group were positioned in different image area and considering different orientations of the bugs when acquiring the samples on the different background types.

Firstly, the pixels related to the black sandpaper background were removed from each image by excluding all the pixels with reflectance values lower than 0.3 reflectance units at 1000 nm, which was identified as the most discriminant wavelength between the sample area and the black sandpaper background. Then, Principal Component Analysis (PCA) was applied to each image to perform a masking procedure, in order to separate the pixels belonging to the bugs from the pixels belonging to the different vegetal backgrounds. In this case, the hyperspectral images were preprocessed using standard normal variate (SNV) and mean center. Therefore, for each hyperspectral image two masks were obtained: one to identify the pixels belonging to the vegetal backgrounds and one to identify the pixels belonging to the BMSB specimens.

To perform image elaboration, the acquired hyperspectral images were converted to .mat format and further analysed in MATLAB environment (R2020b, The MathWorks Inc., USA). Image correction based on internal calibration was performed using ad hoc routines written in MATLAB environment, while the masking procedure based on PCA was performed using the HYPER-Tools software package [34] (version 3.0, https://www.hypertools.org).

2.3. Data analysis

In this study the acquired hyperspectral images were used to develop classification models able to discriminate BMSB bugs from vegetal backgrounds following two main strategies. The former approach was focused on modelling the different spectral features of the considered classes using the Soft PLS-DA [29] algorithm. In addition, this kind of approach also allowed to select the more relevant spectral regions by coupling sparse-based variable selection with Soft PLS-DA [18,30,35].

The latter approach utilized a U-Net, a Convolutional Neural Network (CNN) based deep learning model, that allows to exploit the features based on spatial relationships between neighbouring pixels of BMSB.

Then, the advantages of the two different strategies were combined to further implement a more effective detection method. As a matter of fact, spectral variable selection enabled the identification of a reduced subset of spectral bands relevant for BMSB detection, which was used to improve the detection ability of the U-Net classification model based on BMSB spatial features (**Figure 1**).

Finally, the performance of all the classification models, that were developed at the pixel-level, was also evaluated at the object-level. In pixel-level classification, each pixel spectrum of an image is considered as a separate entity and it is classified by the model into the corresponding predicted class. Instead, in the case of object-level classification, each BMSB specimen is considered as a single object and classification performance was determined by the correct identification of the considered BMSB specimens.

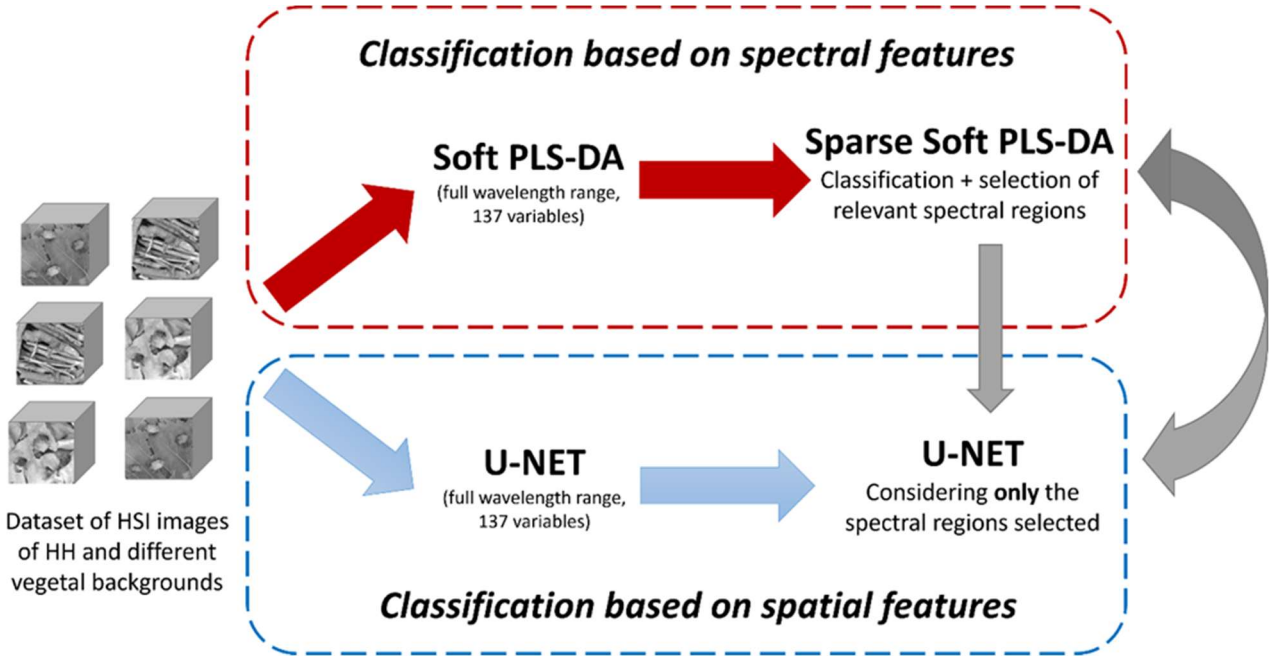


Figure 1 *Schematic representation of the classification strategies adopted to detect BMSB on different vegetal backgrounds using NIR-HSI.*

2.3.1. Classification based on spectral features

For the development of the classification models using Soft PLS-DA and s-Soft PLS-DA it is necessary to select a set of representative spectra belonging to both BMSB and vegetal backgrounds. This phase is crucial since it determines the representativity of spectral signatures considered for the two classes to implement robust and reliable classification models.

For each image, a PCA model was calculated using mean center as preprocessing method, considering only the pixels belonging to the vegetal background and retaining 3 PCs. The number of PCs to retain in the model was chosen based on a preliminary evaluation performed on some representative sample images. Then, outlier pixels were removed considering the 99.9% confidence limit for both Hotelling T2 values and Q residuals. Finally, a new PCA model was calculated considering again 3 PCs and the Kennard-Stone algorithm [36] was applied in the PCs space to select for each image 200 spectra representative of the vegetal background. As a result, 7000 spectra belonging to the vegetal backgrounds were collected (= 200 spectra × 35 hyperspectral images).

The same procedure was also adopted to select from each image 200 spectra belonging to the bugs, obtaining also in this case 7000 spectra representative of the BMSB class. Therefore, the dataset of spectra belonging to the modelled classes and used to develop the classification models was composed of 14,000 spectra.

The dataset was split into a training set (TR-spectra), used for model calculation, and a test set (TS-spectra) used for external validation. To this aim, BMSB specimen groups (previously cited in **Section 2.1**) were considered: the spectra belonging to images containing G1-G3 groups were used for the TR-spectra dataset, including 8400 spectra (= 3 BMSB groups × 7 backgrounds × 200 spectra × 2 classes), while the spectra belonging to images containing G4 and G5 groups were considered for the TS-spectra dataset, including the remaining 5600 spectra.

As an additional external validation, the classification models were also applied to the test set images (TS-images), i.e., to the whole hyperspectral images of the G4 and G5 BMSB specimen groups. The corresponding prediction images (i.e., the images with the pixels coloured according to the predicted class) were used to visualize the prediction performances directly on the images and to obtain a quantitative evaluation of the classification performances over the whole images.

The classification models were calculated both using Soft PLS-DA and combining this algorithm with a sparse-based variable selection approach.

Soft PLS-DA combines the advantages of classical discriminant analysis and class modelling techniques, in order to increase the flexibility of classification models for field application.

Like PLS-DA, the Soft PLS-DA algorithm is based on a discriminant approach, which maximizes the discrimination between samples belonging to the investigated classes, but, unlike PLS-DA, class assignment is performed by fixing additional limits both on the Y predicted values and on the Q residuals. More in detail, a new sample is assigned to a defined class according to the following criteria:

- having Q residuals values falling inside the 99.9% confidence limit of the model. This limit has been chosen to set boundaries large enough to consider different classes' variability as much as possible while being able to exclude samples with a very low fit to the model;
- having y predicted values falling inside an acceptability range for the considered class, whose lower limit is defined by PLS-DA threshold for the investigated class while the upper limit allows to reject objects found at the extremes of the Gaussian probability density function.
- for classification problems with more than two classes, the samples must be unambiguously assigned only to one class.

Samples that do not match all the three criteria defined by Soft PLS-DA decision rule are not assigned to any class and automatically labelled as "not assigned" samples (NA). In this manner, Soft PLS-DA overcomes PLS-DA limited ability to correctly handle new objects not belonging to the target classes while maximizing discrimination between the classes of interest [29,37,38]. For a detailed description of Soft PLS-DA algorithm the reader is referred to Ref. [29].

Moreover, in s-Soft PLS-DA sparse-based variable selection was coupled with Soft PLS-DA, to maintain high model performances while selecting only the most representative wavelengths involved in the classification. The main idea behind sparse methods in the context of linear regression and classification is to decrease the computational load while increasing the robustness of the prediction models [17]. The sparsity is achieved by adding a penalty term to the computation of the model coefficients: in this case, a Least absolute shrinkage and selection operator (Lasso) penalty approach was applied [20,29,30,39].

Both Soft PLS-DA and s-Soft PLS-DA classification models were calculated considering different row-preprocessing methods, i.e., SNV, detrend, first derivative and second derivative, followed by mean center.

The optimization of the classification models was performed using venetian blinds cross-validation with 3 deletion groups.

In s-Soft PLS-DA it is necessary to also optimize the sparsity of the model, i.e., the number of variables to be selected, in addition to the proper number of LVs. Different models were calculated considering all the possible combinations between a number of LVs ranging from 1 to 10 and a number of variables selected for each LV ranging from 5 to 137, with a step equal to 5. The best combination between the number of LVs and the number of selected variables was identified by maximizing classification efficiency (EFF, *see* **Section 2.3.5**) estimated in cross-validation [29].

Soft PLS-DA and s-Soft PLS-DA were calculated using *ad hoc* routines written in MATLAB environment (ver. 2020b, The MathWorks, USA). The MATLAB routine to run Soft PLS-DA algorithm [29] is freely downloadable from http://www.chimslab.unimore.it/downloads/.

2.3.2. Classification based on spatial features

Classification based on spatial features using Deep Learning approaches relies on big data: a large amount of training data is required to maximize the generalizability of the model in order to avoid overfitting [22]. Therefore, when dealing with smaller datasets it is a common practice to apply Data Augmentation techniques such as rotation, flipping, and scaling to obtain an augmented experimental dataset. In this process, the total amount of available images used for the development of the U-Net model was increased by 10 times based on some preliminary evaluations, obtaining a total of 231 images from the original 21 images in the TR-image dataset. For each original image, the same augmentation techniques were also applied to the corresponding masks, identifying the pixels labelled as vegetal background or BMSB. In this manner, it was possible to associate each augmented image with the respective masks.

The dataset of augmented images was used to calculate a classification model using U-Net algorithm. A typical U-Net (**Figure 2**) architecture is built up from several convolutional neural layers that reduce spatial dimensionality using MaxPool layers [25] in a first stage, while in the second stage increase spatial dimensionality again using either Deconvolution or Upsampling layers [40] The U-Net's ability to transfer the entire feature map to the second stage, allows the use of higher resolution details in the later decision layers.

In this work, Upsampling layers were used instead of the Deconvolution layers to retain the original spatial resolution of the image. Moreover, to deal with the hyperspectral images, the first convolutional neural layer, which normally has three input dimensions (RGB channels), was adapted to the input dimensionality of the HSI image (137 spectral channels). Although this step increased the size of the network, its impact on model complexity is negligible, since later feature layers in the network contain more weights. This is clearly shown in **Figure 2**, where the first convolution layer would normally contain 3×3×C×64 wt, where C increases from 3 dimensions for RGB to 137 dimensions for full hyperspectral. However, the tensor of the additional weights in the subsequent encoding layers are much large than the weight tensor of the first convolution layer.

In order to deal with the class unbalance between a greater amount of vegetal background pixels and the annotated BMSB pixels, a balanced loss function was introduced. This procedure allowed to give more weight to the BMSB class which would otherwise be under-represented in the loss function. Empirically, it was found that an exact reweighting based on the number of annotated BMSB and background pixels in the training set resulted in oversized BMSB detected regions due to the assignment of the border pixels between BMSB and background. Accurately sized BMSB regions were obtained by reducing the importance of annotated BMSB pixels (empirically, the annotated BMSB pixel weight was divided by 4).

The performance of the classification model was then evaluated using the same external test set images (TS-images) described in **Section 2.3.1**, corresponding to images containing different BMSB specimens from those contained in the TR-image dataset.

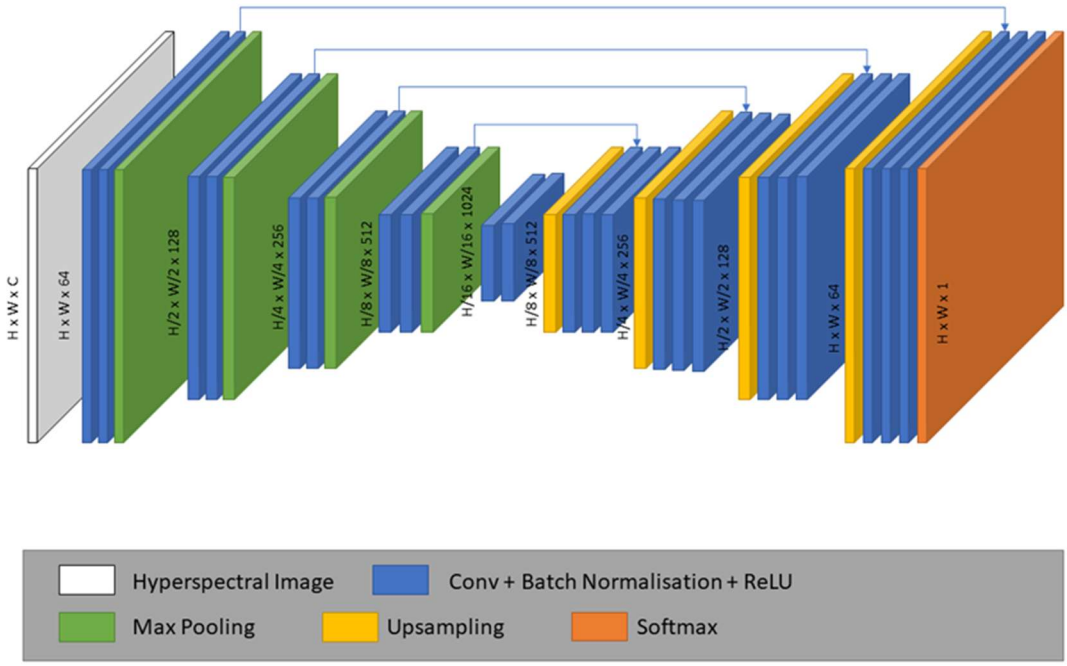


Figure 2 *Schematic representation of the U-Net architecture used in this work, where H x W x C are respectively the image height, width and number of hyperspectral channels. The colour legend describes the different kind of layer used in this architecture. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article).*

2.3.3. Classification based on merged methods

The two different classification strategies mentioned in **Section 2.3.1** and **Section 2.3.2** were merged in order to implement a more effective detection method. In particular, the reduced subset of relevant spectral bands for BMSB detection identified using s-Soft PLS-DA was used to further improve the detection ability of U-Net models. In particular, from the analysis of the s-Soft PLS-DA results, two sets of spectral bands were selected: a more restricted one, named Selection 1, and a wider one, named Selection 2. U-Net was then tested on both the spectral ranges, after adapting the convolutional neural layer to the input dimensionality of the two sets of spectral bands.

2.3.4. Object-level classification

In practical applications of BMSB detection, the prediction images obtained by classifying each pixel of the considered hyperspectral image can be subjected to further elaboration to obtain the final classification output at the object-level. In the present case, once the pixel-level classification models have been obtained, a further evaluation was made by focusing on the presence of clusters of neighbouring pixels predicted as BMSB. At first, defining the expected size in terms of number of

pixels of BMSB specimens, it was possible to remove from the prediction images clusters of pixels predicted as BMSB with size lower than the established threshold. Then, each retained cluster of BMSB pixels was identified as a single bug.

In this study, the prediction images of test samples obtained from the different classification models (i.e., Soft PLS-DA, s-Soft PLS-DA, U-Net on full spectral range, U-Net on Selection 1 and U-Net on Selection 2) were subjected to object-level classification to simulate a real application scenario. Firstly, the clusters of neighbouring pixels predicted as BMSB with size smaller than 50 pixels were ignored. Then, the masks obtained for each image using PCA to identify the pixels belonging to the bugs were used as ground truth. The overlap between the BMSB ground truth cluster of pixels was compared with the clusters of pixels predicted as BMSB in the prediction images by computing the Intersection over Union (IoU) or Jaccard index [48]. The BMSB predicted cluster of pixels was considered to match with IoU >0.25, a slightly lower number than the standard IoU of 0.5, because the border between BMSB and background was not always clear in the hyperspectral images due to the legs of the BMSB.

In this manner, for each classification model it was possible to estimate the number of correctly identified BMSB samples (true positives), the number of actual BMSB samples not identified (false negatives), and the number of background pixels identified as BMSB (false positives).

2.3.5. Evaluation of classification performances

The performances of the classification models were evaluated both at the pixel-level and at the object-level. The statistical parameters used to evaluate the classification performances [41], considering always the BMSB class, are the following ones:

- *Sensitivity* (SENS), also referred to as *Recall* or *True Positive Rate*, defined as the ratio between the true positives (TP), i.e., the objects correctly assigned to the modelled class, and all the objects actually belonging to the considered class, i.e., the true positives and the false negatives (FN): SENS = TP / (TP + FN);
- *Specificity* (SPEC), also referred to as *Selectivity* or *True Negative Rate*, defined as the ratio between the true negatives (TN), i.e., the objects correctly rejected by the modelled class, and all the objects not belonging to the considered class, i.e., the true negatives and the false positives (FP): SPEC = TN / (TN + FP). This parameter was calculated only for the pixel-based classification models, since the object-based classification only considers the BMSB samples, thus TN cannot be defined;
- *Efficiency* (EFF), defined as the geometric mean of SENS and SPEC;

- *Precision* (PREC), defined as the ratio between the true positives and all the objects assigned to the modelled class: Precision = TP / (TP + FP);
- *F1 score*, defined as the harmonic mean of SENS and PREC.

$$\boldsymbol{F1} = \frac{\mathbf{2}}{\frac{\mathbf{1}}{\boldsymbol{SENS}} + \frac{\mathbf{1}}{\boldsymbol{PREC}}} \quad (3.1)$$

For the Soft PLS-DA and the s-Soft PLS-DA models all the statistical parameters were calculated at the pixel-level both in cross-validation (CV) and in prediction of the external test set of spectra (TS-spectra), and for all the classification models the relevant statistical parameters were estimated on the whole set of test images (TS-images), both at the pixel-level and at the object-level.

3. Results

3.1. Classification based on spectral features selection

Table 1 reports the results obtained in cross-validation (CV) and prediction of the test set (TS-spectra) from the Soft PLS-DA and s-Soft PLS-DA models. For each model, the classification performances were evaluated considering SENS, SPEC, EFF, PREC and F1 score values for BMSB class. Different spectral preprocessing methods were considered to compare both the classification performances and the selected spectral regions.

Generally, promising results were obtained in the discrimination between BMSB and all the considered vegetal backgrounds in cross-validation and prediction of the external test set. In addition, sparse variable selection considerably reduced the number of retained spectral variables, while maintaining satisfactory classification performances compared to the full spectral range.

Considering s-Soft PLS-DA method, second derivative row-preprocessing and mean center (mc) provided the highest cross-validation efficiency value, however the model calculated with SNV and mean center led to comparable classification performances considering at the same time a lower number of LVs and of selected spectral variables. Therefore, this model was chosen as the optimal classification model, since it is the one leading to best results in terms of both parsimony and classification performances. In addition, SNV row-preprocessing allows a simpler interpretation of the relevant spectral regions compared to second derivative preprocessing.

		SNV + mc		1st derivative + mc		Detrend + mc		2nd derivative + mc	
		CV	TS-spectra	CV	TS-spectra	CV	TS-spectra	CV	TS-spectra
Soft PLS-DA 137 spectral variables	LVs	3		4		5		6	
	SENS (%)	93.4	95.4	90.9	92.4	90.8	91.5	92.0	92.4
	SPEC (%)	95.8	96.6	94.8	95.3	96.1	96.4	97.6	97.8
	EFF (%)	94.6	96.0	92.8	93.8	93.4	93.9	94.8	95.0
	PREC (%)	95.7	96.5	94.6	95.1	95.9	96.2	97.4	97.6
	F1 score (%)	94.6	96.0	92.7	93.8	93.4	93.9	95.3	95.4
s-Soft PLS-DA	LVs	3		5		7		7	
	Selected variables	60		73		123		101	
	SENS (%)	92.9	94.9	91.5	92.5	92.8	93.0	92.2	93.0
	SPEC (%)	96.0	96.6	96.4	96.0	96.0	96.3	97.6	97.4
	EFF (%)	94.5	95.7	93.9	94.2	94.4	94.6	94.8	95.2
	PREC (%)	95.9	96.5	96.2	95.9	95.9	96.1	97.4	97.2
	F1 score (%)	94.4	95.9	93.9	94.2	94.3	94.5	95.3	95.3

Table 1 *Pixel-level classification results obtained by applying Soft PLS-DA and s-Soft PLS-DA in cross-validation (CV) and prediction of the test set data matrix (TS-spectra) considering different preprocessing methods*

Figure 3 reports the regression vector of the best s-Soft PLS-DA model (i. e., the model calculated considering SNV + mean center as spectral preprocessing method) in order to evaluate the spectral regions selected

by the algorithm to discriminate BMSB from the vegetal backgrounds. It is possible to observe three main spectral bands with high absolute values in the regression vector (highlighted in green colour in **Figure 3**), and therefore with high relevance to the model. These regions correspond to the following intervals: 1220–1295 nm (C–H combination band), 1370–1410 nm (CH-combination band and O–H first overtone) and 1420–1480 nm (O–H first overtone, C–O stretch third overtone and N–H stretch first overtone).. In the following, these three spectral regions will be referred to as “Selection 1”.

Furthermore, two additional spectral regions were selected by the algorithm, even if these regions have low relevance to the model (highlighted in yellow color in **Figure 3**), i.e., they have low absolute values in the regression vector. These regions fall in the 980–1070 nm and 1330–1350 nm intervals, which correspond to N–H second overtone and C–H combination band, respectively. In the following,

all the spectral regions selected by s-Soft PLS-DA algorithm (i.e., both the yellow and the green bars in **Figure 3**) will be referred to as "Selection 2".

The selected spectral regions falling in the intervals at 1220–1295 nm and at 1420–1480 nm can be associated to the presence of cellulose, hemicellulose and lignin in the different vegetal background types [42–44].

Conversely, the selected spectral bands falling in the intervals at 980–1070 nm, 1330–1350 nm and 1370–1400 nm correspond to absorption bands ascribable to water, protein, chitin and lipids. Therefore, these spectral regions can be associated to the biochemical structure of the outer layer of insects' exoskeleton, which is rich in chitin protein chains and lipids [45–47].

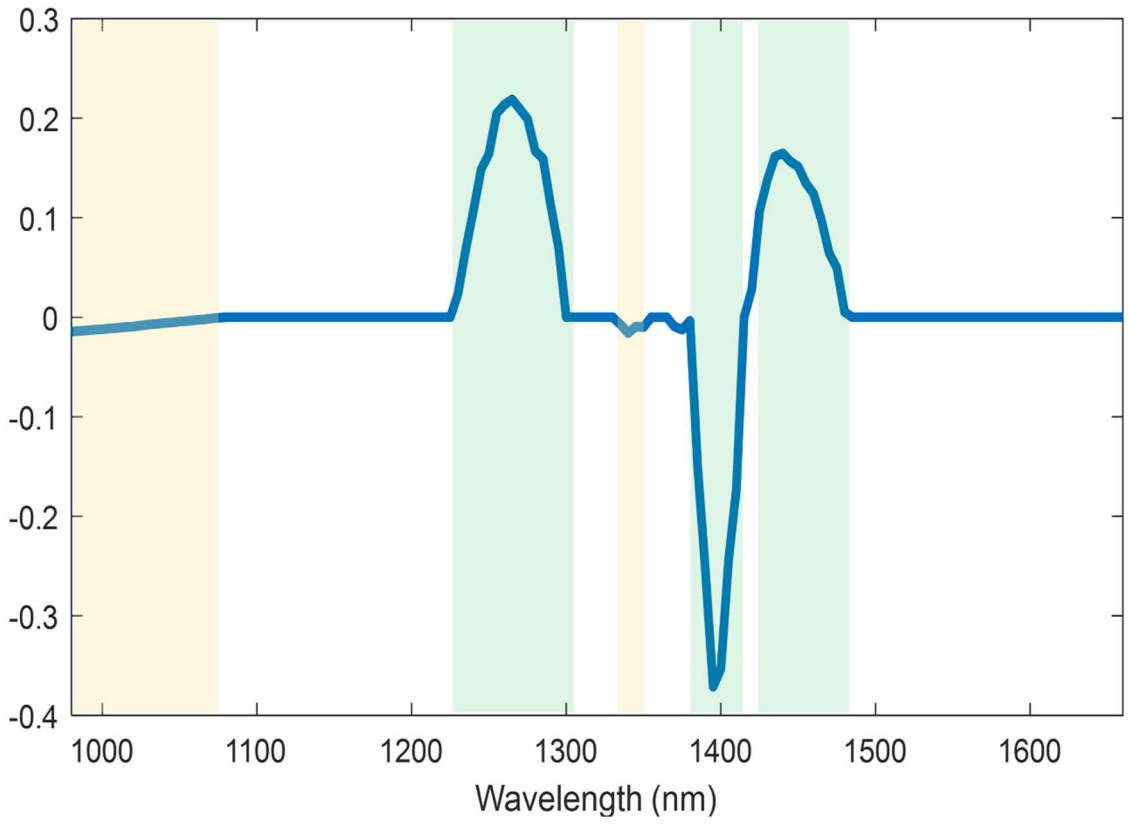


Figure 3 *Regression vector obtained by from the best s-Soft PLS-DA model and spectral regions defined as relevant for BMSB discrimination from vegetal backgrounds.*

As an additional external validation to further verify the effectiveness of the classification models, Soft PLS-DA and s-Soft PLS-DA models were applied to the whole hyperspectral images containing the bugs belonging to G4 and G5 groups (TS-images). The results obtained by applying the Soft PLS-DA and the s-Soft PLD-DA models calculated using SNV + mean center, reported in the first two columns of **Table 2**, show that s-Soft PLS-DA always leads to similar but better performances with respect to Soft PLS-DA, suggesting that variable selection provides a slight improvement of the

predictive ability. The values of SENS, SPEC and EFF are always very high, while PREC, though acceptable, is much lower due to a relatively higher number of false positives with respect to the number of true positives.

In more detail, by comparing the SPEC values of the TS-images predictions it is possible to point out some differences between background types (**Table 3**). In particular, slightly lower SPEC values were obtained with soil and tree branches as background types, i.e., the pixels belonging to these background types were more likely to be misclassified as BMSB. In addition, soil and tree branches were also the background types with a higher amount of not assigned pixels. However, it has to be considered that all the obtained SPEC values were higher than 95%, suggesting overall satisfactory classification results.

In **Figure 4** the prediction images of some sample images are reported, together with the corresponding RGB images as reference. Generally, from the comparison between prediction images and the corresponding RGB images, it is possible to verify that the pixels are correctly classified into the corresponding class. Furthermore, **Figure 4** highlights how difficult it is to detect BMSB specimens on dark brown vegetal backgrounds using only RGB images. Conversely, the bugs are clearly identified using NIR-HSI.

As reported in white circles in **Figure 4**, by comparing the prediction images obtained using Soft PLS-DA and s-Soft PLS-DA models considering tree branches and soil as background, it is possible to observe that spectral variable selection slightly improves the classification of background pixels. Indeed, the amount of misclassified background pixels (i.e., pixels belonging to the background that are wrongly classified as BMSB) is lower in the prediction images obtained from s-Soft PLS-DA model. Therefore, thanks to the possibility of selecting and considering only the spectral bands relevant for the classification, s-Soft PLS-DA facilitates the discrimination of background pixels, particularly when considering the background types more prone to misclassifications, like tree branches and soil.

Algorithm(s)	Soft PLS-DA	s-Soft PLS-DA	U-Net	s-Soft PLS-DA + U-Net	
Spectral Variables	**Full Spectrum**	**Selection 2**	**Full Spectrum**	**Selection 1**	**Selection 2**
SENS (%)	97.1	97.3	89.2	79.1	92.4
SPEC (%)	98.4	98.5	99.2	96.7	98.8
EFF (%)	97.7	97.9	94.1	87.5	95.6
PREC (%)	66.1	68.3	79.3	44.1	71.6
F1 score (%)	78.7	80.2	83.9	56.6	80.7

Table 2 *Pixel-level classification results of the spectral-based (Soft PLS-DA and s-Soft PLS-DA), spatial-based (U-Net) and spectral- & spatial-based (s-Soft PLS-DA + U-Net) classification models applied to the test images.*

	Linear Classification Models Outputs													
	Bark		**Grass**		**Dry leaves**		**Green leaves**		**Yellow leaves**		**Soil**		**Tree branches**	
	Soft PLSDA	**s-Soft PLSDA**	**Soft PLSDA**	**s-Soft PLSDA**	**Soft PLSDA**	**s-Soft PLSDA**	**Soft PLSDA**	**s-Soft PLSDA**	**Soft PLSDA**	**s-Soft PLSDA**	**Soft PLSDA**	**s-Soft PLSDA**	**Soft PLSDA**	**s-Soft PLSDA**
SENS (%)	98.0	98.0	97.9	98.6	89.1	88.2	95.8	97.3	99.9	99.9	99.4	99.4	99.7	99.4
SPEC (%)	99.4	99.4	98.3	97.9	99.4	99.4	99.5	99.4	98.6	98.4	97.1	97.2	95.4	97.0
EFF (%)	98.7	98.7	98.1	98.3	94.1	93.6	97.6	98.4	99.2	99.2	98.2	98.3	97.5	98.2
BMSB N/A (%)	0.6	0.5	0.2	0.3	0.0	0.0	0.3	0.3	0.0	0.1	0.4	0.4	0.1	0.0
BACK N/A (%)	0.2	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.3	0.23	0.5	0.4

Table 3 *Comparison between linear classification models: prediction performances for TS-images. For each background are reported the percent values of SENS, SPEC, EFF for the BMSB class, and not assigned (N/A) pixels for both BMSB and background classes, obtained considering the full wavelength range (Soft PLS-DA) and the spectral regions selected by s-Soft PLS-DA.*

	Full wavelength range													
	Bark		Grass		Dry leaves		Green leaves		Yellow leaves		Soil		Tree branches	
	Soft PLS-DA	UNET	Soft PLS-DA	UNET	Soft PLS-DA	UNET	Soft PLS-DA	UNET	Soft PLS-DA	UNET	Soft PLS-DA	UNET	Soft PLS-DA	UNET
SENS (%)	98.0	92.1	97.9	79.1.8	89.1	95.8	95.8	88.0	99.9	97.2	99.4	84.0	99.7	90.6
SPEC (%)	99.4	99.0	98.3	99.4	99.4	98.3	99.5	99.7	98.6	99.1	97.1	99.6	95.4	99.3
EFF (%)	98.7	95.5	98.1	88.7	94.1	97.0	97.6	93.7	99.2	98.2	98.2	91.5	97.5	94.9
PREC (%)	83.5	73.9	71.1	86.0	82.9	66.5	83.6	88.9	61.8	71.8	57.5	89.3	47.5	84.6
F1 score (%)	90.2	82.0	82.4	82.4	85.9	78.5	89.3	88.4	76.4	82.6	72.8	86.6	64.3	87.5

Table 4 *Comparison between linear and non-linear classification methods: prediction performances for TS-images. For each background are reported the percent values of SENS, SPEC, EFF, PREC and F1 score obtained by considering the full wavelength range.*

3.2. Classification based on spatial features

The third column of **Table 2** shows the SENS, SPEC, EFF, PREC and F1 score values calculated at the pixel level by applying the U-Net algorithm to the TS-images and considering the full wavelength range.

U-Net provided good classification performances, with an EFF value equal to 94.1%. Compared to Soft PLS-DA and s-Soft PLS-DA, U-Net led to a higher SPEC value; in more detail, U-Net led to higher values for 5 out of the 7 background types (**Table 4**), suggesting that this algorithm is less prone to provide false positive pixels. This aspect is also confirmed by the much higher PREC value of U-Net (79.3%) with respect to Soft PLS-DA (66.1%) and s-Soft PLS-DA (68.3%). On the other hand, the SENS value obtained with U-Net (89.2%) was much lower than the SENS values obtained with Soft PLS-DA (97.1%) and with s-Soft PLS-DA (97.3%). This fact can be attributed to misclassifications involving the pixels at the borders between BMSB and background; however, as it will be shown in **Section 3.4**, it did not lead to negative effects on the number of correctly detected samples. Overall, the F1 score value, accounting for both SENS and PREC, showed the highest value for U-Net (83.9%).

Figure 5 reports the prediction images obtained by applying U-Net to the TS-images. Comparing the U-Net prediction images to those obtained from Soft PLS-DA (**Figure 4**), it can be noticed that there are fewer misclassified background pixels and that the prediction masks consist of clear clusters of

pixels predicted as BMSB. However, the outlines of the BMSB samples look less detailed than those resulting from Soft PLS-DA, since U-Net also takes the neighbourhood of the pixels into account.

3.3. Results based on merged strategies

In order to merge the strengths of linear classification methods with deep learning approach, the U-Net algorithm was applied to identify BMSB specimens considering only the wavelengths selected by s-Soft PLS-DA.

As discussed in **Section 3.1**, the linear classification algorithm s-Soft PLS-DA allowed to select 60 spectral variables out of the original 137 wavelengths, which resulted to provide important information for the discrimination between BMSB and the different vegetal backgrounds. In addition, observing the regression vector of the s-Soft PLS-DA model it was possible to identify three main spectral regions with higher relevance to the classification model. The wavelengths falling in these three more relevant spectral regions are referred to as Selection 1, while all the wavelengths selected by s-Soft PLS-DA are referred to as Selection 2.

The fourth and fifth columns of **Table 2** report the U-Net classification performances of the test set images pertaining to Selection 1 and Selection 2, respectively.

Considering the results of Selection 1, the classification performances in prediction are lower than those obtained both by the spectral-based methods and by U-Net applied to the full spectral range. In particular, the decrease in the performances is more evident considering the SENS and the PREC values (79.07% and 44.1%, respectively). These results clearly indicate that considering only the spectral regions with highest absolute values of the regression coefficients of the s-Soft PLS-DA model (green bars in **Figure 3**) is not sufficient to correctly identify the BMSB samples.

As a matter of fact, the results obtained considering Selection 2 constitute an optimal compromise between those obtained using the spectral-based methods (Soft PLS-DA and s-Soft PLS-DA) and those obtained using the spatial-based U-Net method applied to the full spectrum. Actually, merging s-Soft PLS-DA with U-Net led to a higher SENS value (92.4%) with respect to U-Net alone (89.2%), and to a higher PREC value (71.6%) with respect to Soft PLS-DA and s-Soft PLS-DA (66.1% and 66.3%, respectively). Therefore, the 60 spectral variables selected by s-Soft PLS-DA are sufficient to achieve good classification performances, with satisfactory and balanced values for all the considered statistics.

The differences in the classification performances obtained including only Selection 1 and Selection 2 spectral regions can be explained considering the absorption bands falling in the different selected intervals. Indeed, only Selection 2 includes the spectral regions falling in the 980–1070 nm and 1330–1350 nm intervals, corresponding to the absorption bands of protein, chitin and lipids, which are the

main constituents of BMSB exoskeleton. Therefore, these spectral bands resulted to be fundamental in discriminating BMSB from vegetal backgrounds.

Figure 5 shows the prediction images obtained by U-net using the Full Spectrum, Selection 1 and Selection 2. The Full spectrum shows the best results followed by Selection 2, which has more false positives on wooden branches, which may be easily filtered out based on size. Selection 1 results in reduced performance due to problems with both backgrounds of leaves and wooden branches, producing false positives that are impossible to filter out

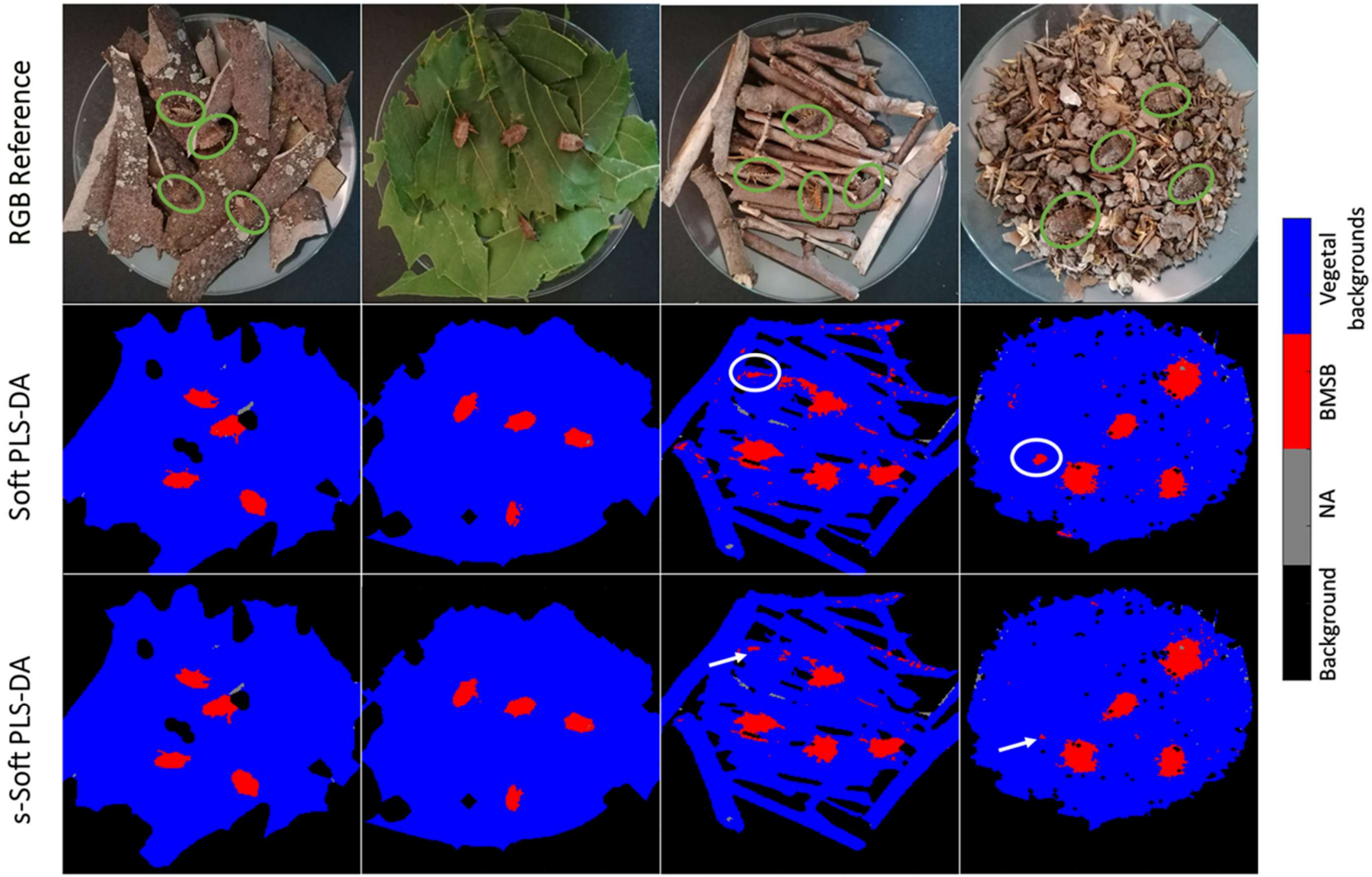


Figure 4 *Prediction images obtained by applying Soft PLS-DA and s-Soft PLS-DA models together with the corresponding RGB images of the samples.*

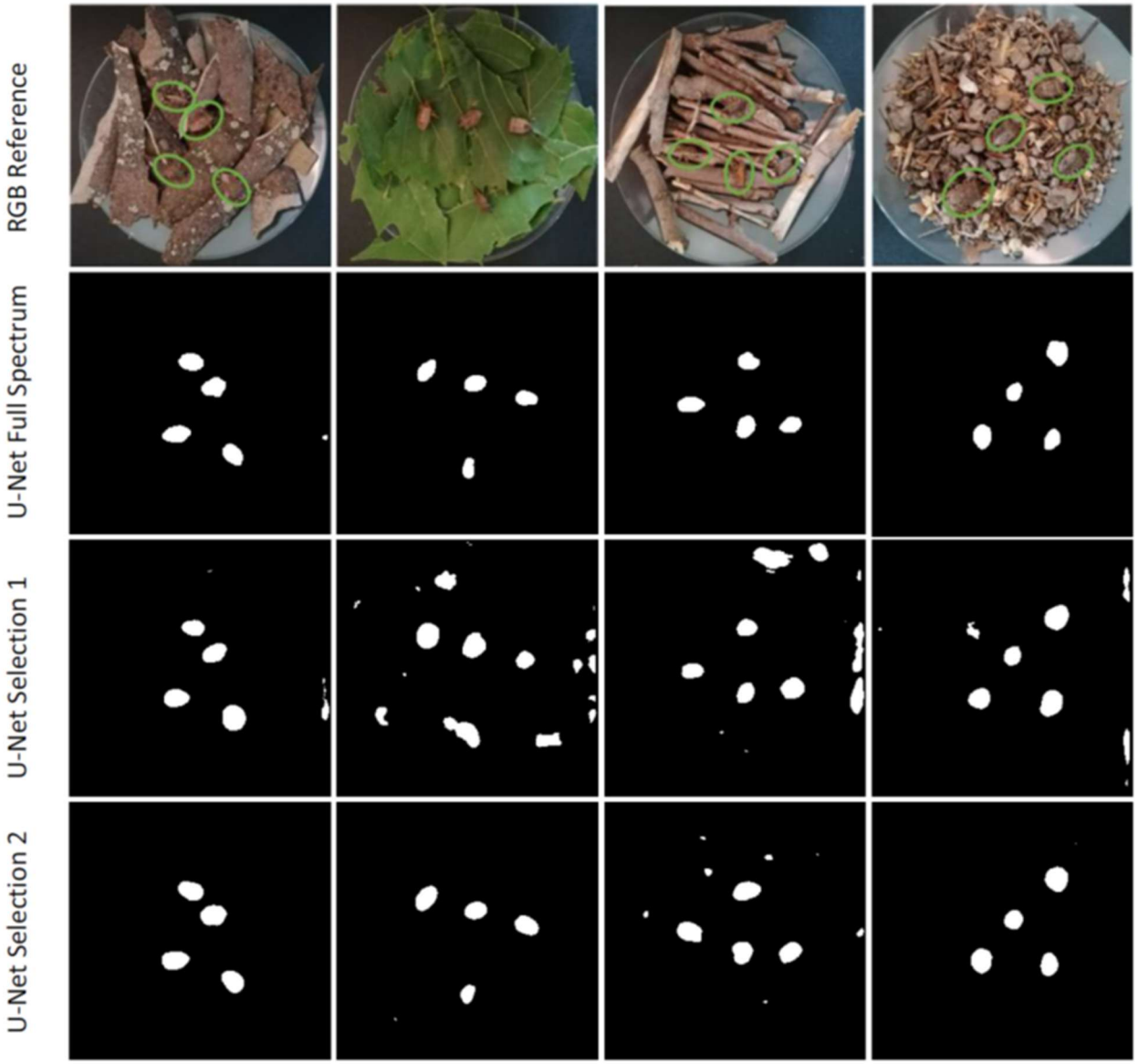


Figure 5 *Prediction masks obtained by different U-Net models together with the corresponding RGB images.*

3.4. Object-level evaluation of the classification models

Table 5 reports the performance measures of the object-level classification, where only SENS, PREC and F1 score were considered for the reason discussed in **Section 2.3.5**. Notice that **Figure 5** shows the results before potential removing the smaller clusters, however the results in **Table 5** are based on the threshold mentioned in **Section 2.3.4**.

Except for the U-Net model based on the spectral bands of Selection 1, all methods showed excellent results considering all three performance indicators. In particular, the U-Net model calculated considering Selection 1 showed a very low precision value, due to the presence of many clusters of background pixels wrongly classified as BMSB. As previously stated in **Section 3.3**, the lower performances obtained considering Selection 1 are due to the fact that the spectral regions considered in Selection 1 do not include spectral bands accounting for the main components of BMSB exoskeleton, which are instead included in Selection 2.

Comparing object-level classification performances of Soft PLS-DA and U-Net, we find that both approaches maintain high SENS values when restricting the considered wavelengths from the full wavelength range to Selection 2; in particular U-Net correctly identified all the BMSB specimens in the images in both cases. On the other hand, variable selection allowed to improve the precision value for the linear classification models (Soft PLS-DA compared to s-Soft PLS-DA or Selection 2), while the precision value slightly decreased for U-Net when considering the wavelengths of Selection 2, since in this case one false positive was identified. However, this slight decrease of performance with respect to that calculated using the whole spectral range is largely compensated by the benefits that could be obtained in practical terms. In fact, the use of a limited number of spectral bands can be the starting point for the development of classification models based on multispectral imaging systems, which require much cheaper and lighter devices than hyperspectral imaging systems.

Algorithm(s)	**Soft PLS-DA**	**s-Soft PLS-DA**	**U-Net**	**s-Soft PLS-DA + U-Net**	
Spectral variables	Full Spectrum	Selection 2	Full Spectrum	Selection 1	Selection 2
SENS (%)	98.2	98.2	100.0	82.1	100.0
Precision (%)	94.8	100.0	100.0	35.4	98.2
F1 score (%)	96.5	99.1	100.0	49.5	99.1

Table 5 *Object-level classification results of the spectral-based (Soft PLS-DA and s-Soft PLS-DA), spatial-based (U-Net) and spectral- & spatial-based (s-Soft PLS-DA + U-Net) classification models.*

4. Conclusions

NIR-HSI is a useful supporting tool for agronomists and farmers for the field monitoring of insect pests. In fact, real-time imaging techniques allow to detect consistent and sudden increase of insect populations which usually denotes an ongoing infestation. As a matter of fact, automated monitoring methods may ease the decision-making process since they permit to gather information of population dynamics and their associated ecological factors in order to develop a targeted pest control strategy. In particular, based on the outcomes of automated monitoring systems based on NIR spectral imaging, it is possible to identify specific crop areas which are more likely subjected to an ongoing infestation and require a direct inspection by technicians as well as proper pest control actions.

The present study aimed at performing a preliminary evaluation of the potential of NIR-HSI as a monitoring technique for BMSB detection. The acquired hyperspectral images were used to develop classification models able to discriminate bugs from vegetal backgrounds following different ML strategies, simulating an in-field application scene. More in detail, the classification models were calculated considering both a linear classification algorithm (Soft PLS-DA) also combined with sparse variable selection (s-Soft PLS-DA), and a deep learning approach (U-Net). While the considered linear classification methods are based on modelling the differences of the spectral response between BMSB and vegetal backgrounds, the U-Net deep learning architecture considers non-linear spatial relationships between pixels in order to provide the classification output.

Linear classification models have the great advantages of requiring a much faster training and providing easily interpretable models, which is particularly important when dealing with spectroscopic data, since it allows to identify the most relevant spectral variables for the problem at hand. On the other hand, deep learning strategies allow to face complex classification problems when the relationship between the modelled data and the final output is not linear, but in this case the interpretation of the models is quite difficult.

In this study, both linear and non-linear approaches led alone to promising results in BMSB detection, but the most relevant outcome of this work consisted in the fact that merging these two strategies allowed to combine the strengths of both methods. In particular, spectral variable selection by s-Soft PLS-DA was used in order to select a subset of relevant variables to be used for the classification with U-Net, leading to classification performances comparable to those obtained by U-Net for the full wavelength range.

The results obtained in this study can be considered as a first step toward the development of multispectral imaging systems for the detection of BMSB. Indeed, MSI systems are more suitable for applications in the field thanks to their relatively low costs and higher resistance of the optical components. Future work will focus on implementing the spectral regions selected by s-Soft PLS-DA

into a MSI-based monitoring system and on evaluating the effectiveness of the proposed approach on real field conditions. To have a preliminary assessment of the performances that can be reached with the multispectral system, hyperspectral data can be used to simulate a multispectral imaging system embedding only band-pass filters falling in the selected spectral regions [20].

Furthermore, it will also be necessary to face the issues that may arise from practical applications of multispectral systems in field. One problem consists in the fact that the images can be acquired at different distances from the camera, resulting in images with different resolutions. However, this problem can be easily tackled by creating augmented images of the insects at different scaling levels. Another problem that has to be considered in practical applications is the presence of insects different from BMSB. Future studies will consider the possibility of identifying different insect species by combining spectral information with spatial and morphological features resulting from the object-level classification.

Acknowledgements

Authors wish to thank HALY.ID, project of ERA-NET Cofund ICT-AGRI-FOOD, with funding provided by national sources (Ministero delle politiche agricole e forestali, MIPAAF) and co-funding by the European Union's Horizon 2020 research and innovation program, Grant Agreement number 862671.

Rosalba Calvini would like to thank the Italian funding programme *Fondo Sociale Europeo REACT-EU - PON "Ricerca e Innovazione" 2014 – 2020 – Azione IV.6 Contratti di ricerca su tematiche Green (D.M. 1062 del 10/08/ 2021)* for supporting her research (CUP: E95F21002330001; contract number 17-G-13884-4).

OnePlanet Research Center is supported by the Province of Gelderland.

Author contributions

V. Ferrari – Methodology; Software; Formal analysis; Investigation; Data curation; Writing – Original Draft; Writing – Review & Editing. **R. Calvini** – Conceptualization; Methodology; Software; Data curation; Writing – Original Draft; Writing – Review & Editing. **B. Boom** – Methodology; Software; Formal analysis; Investigation; Data Curation; Writing – Review & Editing. **C. Menozzi** – Investigation; Writing – Review & Editing. **A. K. Rangarajan** – Investigation; Writing – Review & Editing. **P. Offermans** – Conceptualization; Methodology; Writing – Review & Editing; Supervision; Project administration; Funding acquisition. **L. Maistrello** –Methodology; Resources; Writing – Review & Editing; Supervision; Project administration; Funding acquisition. **A. Ulrici** –

Conceptualization; Methodology; Writing – Review & Editing; Supervision; Project administration; Funding acquisition.

References

[1] T.C. Leskey, A.L. Nielsen, Impact of the Invasive Brown Marmorated Stink Bug in North America and Europe: History, Biology, Ecology, and Management, *Annu. Rev. Entomol.* 63 (2018) 599–618. https://doi.org/10.1146/annurev-ento-020117-043226.

[2] K.B. Rice, C.J. Bergh, E.J. Bergmann, D.J. Biddinger, C. Dieckhoff, G. Dively, H. Fraser, T. Gariepy, G. Hamilton, T. Haye, A. Herbert, K. Hoelmer, C.R. Hooks, A. Jones, G. Krawczyk, T. Kuhar, H. Martinson, W. Mitchell, A.L. Nielsen, D.G. Pfeiffer, M.J. Raupp, C. Rodriguez-Saona, P. Shearer, P. Shrewsbury, P.D. Venugopal, J. Whalen, N.G. Wiman, T.C. Leskey, J.F. Tooker, Biology, Ecology, and Management of Brown Marmorated Stink Bug (Hemiptera: Pentatomidae*), J. Integr. Pest Manag.* 5 (2014) 1–13. https://doi.org/10.1603/IPM14002.

[3] E. Costi, T. Haye, L. Maistrello, Biological parameters of the invasive brown marmorated stink bug, *Halyomorpha haly*s, in southern Europe, *J. Pest Sci.* 90 (2017) 1059–1067. https://doi.org/10.1007/s10340-017-0899-z.

[4] CSO, Estimation of damage from brown marmorated stink bug and plant pathologies related to climate change. http://www.csoservizi.com, 2020.

[5] T.C. Leskey, D.-H. Lee, B.D. Short, S.E. Wright, Impact of Insecticides on the Invasive *Halyomorpha halys* (Hemiptera: Pentatomidae): Analysis of Insecticide Lethality, *J. Econ. Entomol.* 105 (2012) 1726–1735. https://doi.org/10.1603/EC12096.

[6] T.C. Leskey, B.D. Short, B.R. Butler, S.E. Wright, Impact of the Invasive Brown Marmorated Stink Bug, *Halyomorpha halys* (Stål), in Mid-Atlantic Tree Fruit Orchards in the United States: Case Studies of Commercial Management, *Psyche: A J. Entomol.* 2012 (2012) 1–14. https://doi.org/10.1155/2012/535062.

[7] L. Maistrello, G. Vaccari, S. Caruso, E. Costi, S. Bortolini, L. Macavei, G. Foca, A. Ulrici, P.P. Bortolotti, R. Nannini, L. Casoli, M. Fornaciari, G.L. Mazzoli, P. Dioli, Monitoring of the invasive *Halyomorpha halys*, a new key pest of fruit orchards in northern Italy, *J. Pest Sci.* 90 (2017) 1231–1244. https://doi.org/10.1007/s10340-017-0896-2.

[8] L. Maistrello, P. Dioli, M. Dutto, S. Volani, S. Pasquali, G. Gilioli, Tracking the Spread of Sneaking Aliens by Integrating Crowdsourcing and Spatial Modeling: The Italian Invasion of *Halyomorpha halys*, *BioScience* (2018). https://doi.org/10.1093/biosci/biy112.

[9] F.B. Sorbelli, F. Coro, S.K. Das, E. Di Bella, L. Maistrello, L. Palazzetti, C.M. Pinotti, A Drone-based Application for Scouting *Halyomorpha halys* Bugs in Orchards with Multifunctional Nets, in: 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (*PerCom Workshops*), IEEE, Pisa, Italy, 2022: pp. 127–129. https://doi.org/10.1109/PerComWorkshops53856.2022.9767309.

[10] O. Friha, M.A. Ferrag, L. Shu, L. Maglaras, X. Wang, Internet of Things for the Future of Smart Agriculture: A Comprehensive Survey of Emerging Technologies, *IEEE/CAA Journal of Automatica Sinica* 8 (2021) 718–752. https://doi.org/10.1109/JAS.2021.1003925.

[11] A. Milella, G. Reina, M. Nielsen, A multi-sensor robotic platform for ground mapping and estimation beyond the visible spectrum, *Precision Agriculture* 20 (2019) 423–444. https://doi.org/10.1007/s11119-018-9605-2.

[12] D. Caballero, R. Calvini, J.M. Amigo, Hyperspectral imaging in crop fields: precision agriculture, in: *Data Handling in Science and Technology*, Elsevier, 2019: pp. 453–473. https://doi.org/10.1016/B978-0-444-63977-6.00018-3.

[13] R. Calvini, A. Ulrici, J.M. Amigo, Growing applications of hyperspectral and multispectral imaging, in: *Data Handling in Science and Technology*, Elsevier, 2019: pp. 605–629. https://doi.org/10.1016/B978-0-444-63977-6.00024-9.

[14] A. Gowen, C. Odonnell, P. Cullen, G. Downey, J. Frias, Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, *Trends in Food Sci. Technol.* 18 (2007) 590–598. https://doi.org/10.1016/j.tifs.2007.06.001.

[15] A.A. Gowen, F. Marini, C. Esquerre, C. O'Donnell, G. Downey, J. Burger, Time series hyperspectral chemical imaging data: Challenges, solutions and applications, *Anal. Chim. Acta* 705 (2011) 272–282. https://doi.org/10.1016/j.aca.2011.06.031.

[16] D. Saha, A. Manickavasagan, Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review, *Curr. Res. Food Sci.* 4 (2021) 28–44. https://doi.org/10.1016/j.crfs.2021.01.002.

[17] J. Burger, A. Gowen, Data handling in hyperspectral image analysis, *Chemometr. Intell. Lab. Syst.* 108 (2011) 13–22. https://doi.org/10.1016/j.chemolab.2011.04.001.

[18] R. Calvini, A. Ulrici, J.M. Amigo, Practical comparison of sparse methods for classification of Arabica and Robusta coffee species using near infrared hyperspectral imaging, *Chemometr. Intell. Lab. Syst.* 146 (2015) 503–511. https://doi.org/10.1016/j.chemolab.2015.07.010.

[19] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, *Anal. Chim. Acta* 667 (2010) 14–32. https://doi.org/10.1016/j.aca.2010.03.048.

[20] R. Calvini, J.M. Amigo, A. Ulrici, Transferring results from NIR-hyperspectral to NIR-multispectral imaging systems: A filter-based simulation applied to the classification of Arabica and Robusta green coffee, *Anal. Chim. Acta* 967 (2017) 33–41. https://doi.org/10.1016/j.aca.2017.03.011.

[21] S. Brown, R. Tauler, B. Walczak, Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Elsevier, 2020.

[22] C. Shorten, T.M. Khoshgoftaar, A survey on Image Data Augmentation for Deep Learning, *J. Big Data* 6 (2019) 60. https://doi.org/10.1186/s40537-019-0197-0.

[23] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (2017) 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.

[24] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015: pp. 3431–3440.

[25] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015: pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

[26] S. Liu, J. Cheng, L. Liang, H. Bai, W. Dang, Light-Weight Semantic Segmentation Network for UAV Remote Sensing Images, *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* 14 (2021) 8287–8296. https://doi.org/10.1109/JSTARS.2021.3104382.

[27] F. Cervantes-Sanchez, M. Maktabi, H. Köhler, R. Sucher, N. Rayes, J.G. Avina-Cervantes, I. Cruz-Aceves, C. Chalopin, Automatic tissue segmentation of hyperspectral images in liver and head neck surgeries using machine learning, *Artificial Intelligence Surgery* (2021). https://doi.org/10.20517/ais.2021.05.

[28] M.S. Moustafa, S.A. Mohamed, S. Ahmed, A.H. Nasr, Hyperspectral change detection based on modification of UNet neural networks, *Journal of Applied Remote Sensing* 15 (2021). https://doi.org/10.1117/1.JRS.15.028505.

[29] R. Calvini, G. Orlandi, G. Foca, A. Ulrici, Development of a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging, *J. Spectr. Imaging* (2018) a13. https://doi.org/10.1255/jsi.2018.a13.

[30] P. Filzmoser, M. Gschwandtner, V. Todorov, Review of sparse methods in regression and classification with application to chemometrics, *Journal of Chemometrics* 26 (2012) 42–51. https://doi.org/10.1002/cem.1418.

[31] J. Burger, P. Geladi, Hyperspectral NIR image regression part II: dataset preprocessing diagnostics, *Journal of Chemometrics* 20 (2006) 106–119. https://doi.org/10.1002/cem.986.

[32] J. Burger, P. Geladi, Hyperspectral NIR image regression part I: calibration and correction, *Journal of Chemometrics* 19 (2005) 355–363. https://doi.org/10.1002/cem.938.

[33] A. Ulrici, S. Serranti, C. Ferrari, D. Cesare, G. Foca, G. Bonifazi, Efficient chemometric strategies for PET–PLA discrimination in recycling plants using hyperspectral imaging, *Chemometr. Intell. Lab. Sys.* 122 (2013) 31–39. https://doi.org/10.1016/j.chemolab.2013.01.001.

[34] N. Mobaraki, J.M. Amigo, HYPER-Tools. A graphical user-friendly interface for hyperspectral image analysis, *Chemometr. Intell. Lab. Sys.* 172 (2018) 174–187. https://doi.org/10.1016/j.chemolab.2017.11.003.

[35] M.A. Rasmussen, R. Bro, A tutorial on the Lasso approach to sparse modeling, *Chemometr. Intell. Lab. Sys.* 119 (2012) 21–31. https://doi.org/10.1016/j.chemolab.2012.10.003.

[36] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, *Technometrics* 11 (1969) 137–148. https://doi.org/10.1080/00401706.1969.10490666.

[37] M. Barker, W. Rayens, Partial least squares for discrimination, *Journal of Chemometrics* 17 (2003) 166–173. https://doi.org/10.1002/cem.785.

[38] A.L. Pomerantsev, O.Ye. Rodionova, Multiclass partial least squares discriminant analysis: Taking the right way—A critical tutorial, *Journal of Chemometrics* 32 (2018) e3030. https://doi.org/10.1002/cem.3030.

[39] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society Series B: *Statistical Methodology* 58 (1996) 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[40] S.M. Kamrul Hasan, C.A. Linte, U-NetPlus: A Modified Encoder-Decoder U-Net Architecture for Semantic and Instance Segmentation of Surgical Instruments from Laparoscopic Images, in: *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC), IEEE, Berlin, Germany, 2019: pp. 7205–7211. https://doi.org/10.1109/EMBC.2019.8856791.

[41] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, *Chemometr. Intell. Lab. Sys.* 174 (2018) 33–44. https://doi.org/10.1016/j.chemolab.2017.12.004.

[42] T.J. Bruno, P.D.N. Svoronos, CRC Handbook of Fundamental Spectroscopic Correlation Charts, CRC Press, Boca Raton, 2005. https://doi.org/10.1201/9780849332500.

[43] X. Jin, X. Chen, C. Shi, M. Li, Y. Guan, C.Y. Yu, T. Yamada, E.J. Sacks, J. Peng, Determination of hemicellulose, cellulose and lignin content using visible and near infrared spectroscopy in Miscanthus sinensis, *Bioresource Technology* 241 (2017) 603–609. https://doi.org/10.1016/j.biortech.2017.05.047.

[44] X. Li, C. Sun, B. Zhou, Y. He, Determination of Hemicellulose, Cellulose and Lignin in Moso Bamboo by Near Infrared Spectroscopy, *Sci. Rep.* 5 (2015) 17210. https://doi.org/10.1038/srep17210.

[45] F.E. Dowell, J.E. Throne, D. Wang, J.E. Baker, Identifying Stored-Grain Insects Using Near-Infrared Spectroscopy, *J. Econ. Entomol.* 92 (1999) 165–169. https://doi.org/10.1093/jee/92.1.165.

[46] J.B. Johnson, An overview of near-infrared spectroscopy (NIRS) for the detection of insect pests in stored grains, *J. Stored Prod. Res.* 86 (2020) 101558. https://doi.org/10.1016/j.jspr.2019.101558.

[47] C. Ridgway, J. Chambers, Detection of external and internal insect infestation in wheat by near-infrared reflectance spectroscopy, *J. Sci. Food Agric.* 71 (1996) 251–264.

[48] P. Jaccard, The Distribution of the flora in the alpine zone, *New Phytologist* 11 (1912) 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

3.3. NIR Hyperspectral Imaging to identify damage caused by *Halyomorpha halys* on pears: Automated identification of Regions of Interest related to punctured areas

What follows is the integral content of: Ferrari, V., Calvini, R., Menozzi, C., Costi, E., Giannetti, D., Offermans, P., Maistrello, L., Ulrici, A. (**2025**). NIR hyperspectral imaging to identify damage caused by *Halyomorpha halys* on pears: Automated identification of Regions of Interest related to punctured areas, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 343, 126543. DOI: 10.1016/j.saa.2025.126543

NIR Hyperspectral Imaging to identify damage caused by *Halyomorpha halys* on pears: automated identification of Regions of Interest related to punctured areas

Veronica Ferrari[1], Rosalba Calvini[1,2*], Camilla Menozzi[1], Elena Costi[1,2], Daniele Giannetti[1], Peter Hoffermans[3], Lara Maistrello[1,2], Alessandro Ulrici[1,2]

[1] *University of Modena and Reggio Emilia, Department of Life Sciences, Pad. Besta, Via Amendola, 2, 42122, Reggio Emilia, Italy*

[2] *Interdepartmental Research Centre BIOGEST-SITEIA, University of Modena and Reggio Emilia, Piazzale Europa, 1, Reggio Emilia, 42122, Italy*

[3] *IMEC OnePlanet, Bronland 10, Wageningen, the Netherlands*

* *Corresponding author.*

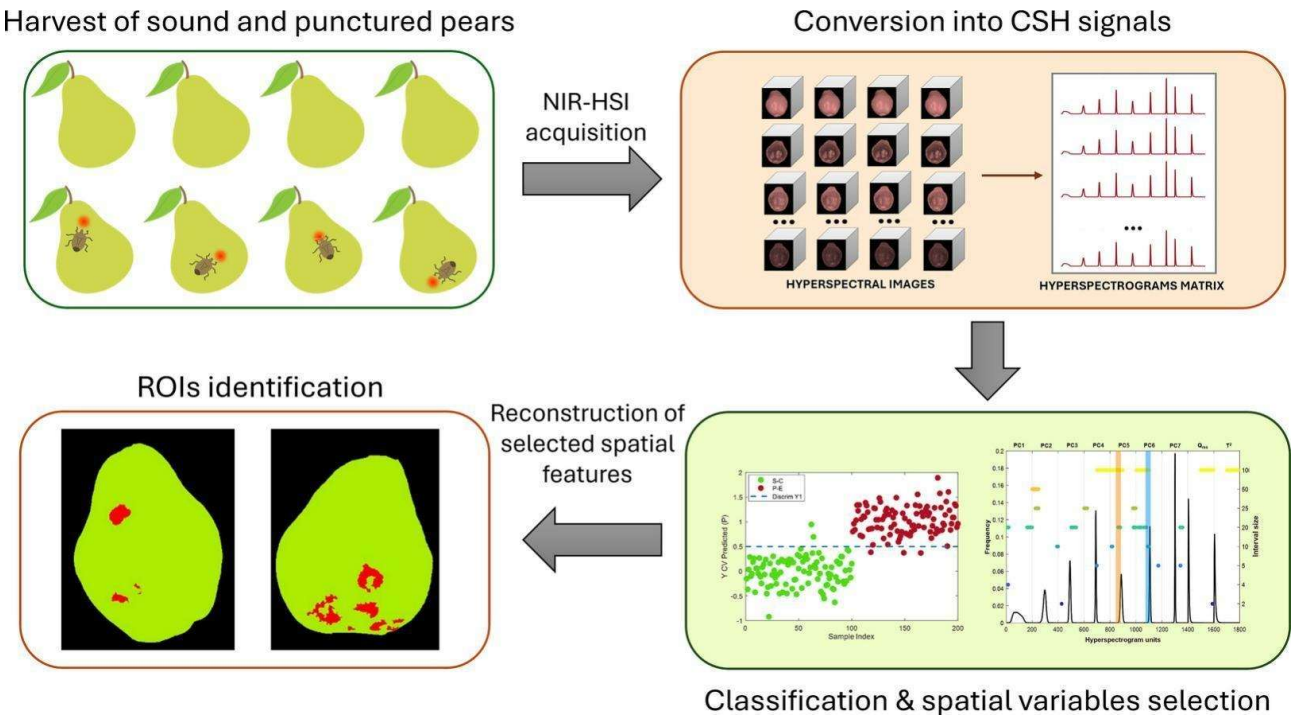
Abstract

Halyomorpha halys, commonly known as the Brown Marmorated Stink Bug (BMSB), is an emerging pest in pear orchards determining major economic losses. BMSB feeding on fruits close to harvest ripening cause internal damage invisible to the naked eye, therefore undetectable using RGB image acquisition systems. To face this issue, in the present work Near-Infrared Hyperspectral Imaging (NIR-HSI) is proposed as a non-destructive technique to automatically discard damaged fruits in post-harvest sorting lines.

In this context, the identification of Regions of Interest (ROIs) ascribable to the punctures is a crucial step affecting the outcomes of supervised classification models. Due to irregular shapes and blurred edges between sound and punctured areas, most popular thresholding techniques are not able to automatically detect the ROIs while, on the other hand, manual thresholding is arbitrary and time consuming on large hyperspectral image datasets.

This paper provides an innovative method for the automated ROIs selection based on image data dimensionality reduction (DDR) and image-level classification coupled with spatial feature selection. To this aim, the hyperspectral images were compressed into Common Space Hyperspectrograms (CSH), signals summarising both spatial and spectral information of the original images. The CSH features highly correlated with the presence of BMSB punctures and more frequently selected by interval Partial Least Squares – Discriminant Analysis (iPLS-DA) models allowed the identification of ROIs of punctured areas. Indeed, the reconstruction of the selected features back into the original image domain led to a successful identification of ROIs ascribable to BMSB punctures in an automated and objective way.

Graphical abstract



Keywords: Hyperspectral imaging, data dimensionality reduction, Regions of Interest, fruit punctures, post-harvest sorting

1. Introduction

In the last decades, the increase *Halyomorpha halys* (Hemiptera, Pentatomidae), also known as the Brown Marmorated Stink Bug (BMSB), is native to East Asia and has become an invasive pest in North America and Europe, posing a significant threat to agriculture. This polyphagous pest attacks a wide range of crops, including fruit, vegetables and ornamentals, leading to significant economic losses [1,2]. In Europe, where BMSB has spread rapidly in almost all countries since 2004, it is causing considerable damage, particularly to fruit crops and hazelnut trees [3]. In northern Italy, BMSB has caused severe damage to fruit crops, with pears being particularly susceptible to this pest [3,4].

BMSB primarily inflicts damage on fruits through its piercing-sucking mouthparts during feeding, causing types of injuries such as deformities, discoloration and internal damage [5–7].

BMSB punctures that occur at a late stage of pear development often result in internal damage that is not externally visible, manifested by the presence of suberified or necrotic areas in the pulp, characterized by a brownish, pithy and corky tissue beneath the skin [7]. Suberification and necrosis significantly reduce the marketability of pears, as they lead to unattractive, poor-quality fruit that is often discarded. Furthermore, internal damage is particularly problematic as it can be extensive yet remaining hidden beneath the fruit skin, making it difficult to detect by visual inspection until the fruit is cut or begins to rot [1].

Near Infrared Hyperspectral Imaging (NIR-HSI) has rapidly emerged as an analytical tool for fruit quality evaluation, allowing to obtain in a fast, cheap and non-destructive manner a detailed insight into the chemical composition and its variation over the sample surface [8–11].

The possibility of coupling both spatial and spectral information makes NIR-HSI particularly suitable for the early detection of fruit defects that are not visible to the naked eye, such as bruises, punctures and various types of damage [12–16]. Furthermore, NIR hyperspectral or multispectral cameras can easily be used for online quality control in post-harvest sorting lines to automatically discard damaged or not compliant fruit.

Given the great potential of this technique for fruit defect detection, in this study we applied NIR-HSI to identify damage caused by BMSB punctures on pears. For the practical application of NIR-HSI systems, it is necessary to develop supervised classification models able to process the hyperspectral images and provide a class assignment for each pixel. This kind of approach is known as pixel-level classification.

Therefore, pixel-level classification models able to discriminate sound and punctured pear areas must be trained using a library of representative spectra belonging to both classes [17,18]. The selection of such spectra is a crucial aspect for the effectiveness and robustness of the classification models, and

it is usually performed by extracting the pixel spectra from Regions Of Interest (ROIs) of the investigated classes [19,20]. Therefore, ROIs identification and labelling are extremely important image elaboration steps affecting the outcomes of supervised classification models. In image analysis, this procedure is often referred to as image annotation and it can be performed using different strategies according to the problem under investigation [21].

Image segmentation based on thresholding procedures is one of the most common approaches to perform ROIs selection, due to its simplicity and efficiency [22,23]. Usually, the grey scale image acquired at the wavelength showing the highest contrast between the area of interest and the remainder pixels is selected, and the corresponding histogram is used to determine a threshold value able to select the pixels of the ROIs.

According to the problem under investigation, a single wavelength may not be sufficient to perform ROIs segmentation, but it is necessary to consider the information brought by different wavelengths at the same time. In this case, it is possible to apply Multivariate Image Analysis (MIA) by calculating a Principal Component Analysis (PCA) model on the investigated image. In this manner, ROIs selection can be performed by thresholding procedures on the score images or by selecting clusters of pixels in the score plots [24,25].

The identification of the proper threshold value is therefore crucial for ROIs selection. The threshold value can be identified either manually by visual inspection of the histogram of grey scale images (i.e., images of single wavelengths or PCA score images) or using automated algorithms [26].

Manual identification of the threshold value can be quite easy when there is a clear separation between ROIs and the remainder parts of the image or when ROIs correspond to specific objects located in the images with a distinctive shape. However, very often ROIs cannot be clearly distinguished and manual thresholding can be a challenging task that may be affected by the operator choice, since different operators can identify different threshold values.

Furthermore, when many images have to be analyzed altogether it is very difficult to define a threshold value appropriate for all the images of the dataset, and manually defining the threshold for each image can be very laborious and time consuming.

Given the drawbacks of manual ROIs selection, it is preferable to use automated procedures [27,28]. Otsu algorithm [29] is the most popular automated thresholding method due to its effectiveness and simplicity. This algorithm identifies a threshold value that separates the pixels of an image in two classes. The threshold value is defined by minimizing the intra-class variance and, at the same time, maximizing inter-class variance.

Munera et al. [30] used an adaptive thresholding based on Otsu method to automatically select ROIs of damaged areas on hyperspectral images of persimmon fruits. This methodology was applied on

PC6 score images and then the pixel spectra extracted from the ROIs of damaged areas were used to train supervised classification models.

However, Otsu algorithm provides satisfactory results only when the grey scale image histogram is bimodal and the two groups of pixels are easily separated [31,32]. This condition is rarely met in practical applications of hyperspectral image analysis for fruit damage detection since the distinction between damaged and sound areas is often unclear, in particular when the samples are in an early damage stage.

Improved modifications of the Otsu method or different thresholding algorithms have been applied for automatic segmentation of bruises or other types of injuries in fruits [33–36]. However, they require some assumptions about pixel distributions to be fulfilled or the optimization of several image elaboration steps. Due to these challenges, manual ROIs selection is still a common practice for fruit damage segmentation [14,37].

In this study, the identification of ROIs related to pears areas damaged by BMSB resulted a very challenging task. First of all, it has to be considered that during the experimental part of the study we acquired about 2000 hyperspectral images of sound and punctured pears, also considering different acquisition times from harvest up to five weeks after.

By a preliminary investigation of some representative hyperspectral images using PCA, it was possible to visually identify the pixels related to punctured areas thanks to the observation of the score images. However, there was not a clear distinction between score values of sound and punctured areas. Furthermore, it was not possible to manually find interval(s) of PCA score values applicable to all the images of the dataset, due to the heterogeneity between the different fruits and the variability caused by the varying ripening levels over the acquisition times.

In addition, the shape of punctured areas was irregular and different from fruit to fruit, therefore it was not possible to couple thresholding with object or shape detection algorithms. Moreover, considering the high number of images acquired in this study, it was not feasible to manually identify the ROIs by elaborating each single image.

To face all these challenges, this paper is focused on the development of an innovative method for the automated identification of ROIs based on image data dimensionality reduction and image-level classification coupled with spatial feature selection.

In a previous study developed by some of the authors [16], a large dataset of hyperspectral images of sound and bruised apples was firstly subjected to data dimensionality reduction using the Single Space Hyperspectrograms (SSH) method and then the resulting data matrix was analyzed using interval Partial Least Squares Discriminant Analysis Algorithm (iPLS-DA) [38]. SSH approach consists in reducing the relevant spectral and spatial information of each image into a one-

dimensional signal, which is obtained by merging in sequence the frequency distribution curves of quantities derived from a PCA model calculated separately for each image [39]. iPLS-DA applied to the SSH data matrix allowed to perform image-level classification between sound and bruised fruits and to select the hyperspectrogram variables more correlated with the presence of the bruises. The selected variables were then visualized back into the original image domain obtaining the exact localization of bruised areas.

The algorithm for automated ROIs identification developed in this study follows a similar workflow. Firstly, the hyperspectral images of sound and punctured pears were converted into Common Space Hyperspectrograms (CSH) [40], a modification of SSH method. CSH are based on the same principle as SSH, but the signals are obtained by merging in sequence the frequency distribution curves of score values of a global PCA model common to all the images of the dataset used to train the subsequent classification models. Then, iPLS-DA models considering different interval sizes were calculated for image-level classification of sound and punctured fruits. The spatial variables selected more frequently from the iPLS-DA models, corresponding to specific image pixels, were then reconstructed back into the image domain, allowing an automated and objective selection of the pixels related to punctured areas.

2. Materials and methods

2.1. Experimental protocol

The pear samples considered in this study were harvested from an organic orchard located in Carpi (Modena, Italy). The fruits belonged to *Williams* variety and were harvested in summer 2022 and summer 2023.

In both years, 40 tree branches were selected from different plants located in different areas of the orchard. Immediately after fruit set, the selected branches were covered with cylindrical inclusion cages made of a semi-rigid plastic mesh covered with a sleeve of flexible white fabric, each one containing 2 to 4 fruits. The inclusion cages were used to protect the ripening fruit from uncontrolled biotic and abiotic adversities [7,41].

On the same day of commercial pear harvest (August 7th in 2022 and August 4th in 2023), BMSB specimens were manually placed inside half of the inclusion cages. All progenitors of these BMSB specimens had been previously captured in the field from urban parks of the city of Reggio Emilia (Italy) using the tree-beating technique. The bugs were reared in climatic chambers at 26 °C, 60% relative humidity, L16:D8 photoperiod inside clear mesh cages with organic tomatoes, carrots, green bean pods and raw peanuts as food. A bottle cap with a water-soaked cotton swab was used as water

supply. Food and water were replaced twice per week. Feeding of BMSB specimens was interrupted two days before placing them in the inclusion cages.

More in detail, 3 BMSB specimens were placed inside each one of the 20 inclusion cages used to expose the fruits to the bugs. In this manner, about half of the fruits considered were exposed to BMSB feeding punctures, while the remaining fruits were used as control fruits.

In the 2022 season, the bugs were kept in the inclusion cages for two days, from August 7th until August 9th, and then both the exposed and the control fruits were harvested. In particular, 53 control fruits and 40 exposed fruits were harvested in 2022, for a total of 93 pear samples. Unfortunately, only 11 of the 40 exposed fruits were found to have damage ascribable to BMSB punctures. Indeed, the summer of 2022 was characterised by exceptionally high temperatures which reached up to 40°C at the beginning of August. These extreme climatic conditions negatively affected BMSB specimens' vitality in the field and their propensity to feed.

Based on the outcomes of the summer 2022 harvest, some minor changes were made to the experimental protocol carried out in summer 2023. In this case, the BMSB specimens were kept in the inclusion cages for five days, from August 4th to August 8th. All the fruits were then left on the tree for one week before harvesting, which took place on August 16th. In the 2023 season, a total of 68 fruits were collected: 38 control fruits and 30 BMSB exposed fruits. Considering the exposed pears, 18 out of 30 fruits resulted to have damage ascribable to BMSB punctures.

All the collected fruits presented the typical characteristics of *Williams* variety, and they showed a great size variability, with fruit diameter ranging from 35 mm to 75 mm.

2.2. Hyperspectral image acquisition

In both 2022 and 2023 years, the fruit samples were transported to the laboratory for image acquisition immediately after harvest. During transport, the fruits were placed in boxes covered with egg crate foam to prevent damage.

To monitor the evolution of fruit damage over time, 8 subsequent acquisition times were considered from harvest (T1) until five weeks after (T2-T8).

The schedule followed for hyperspectral image acquisition in 2022 and 2023 is reported in **Table 1**.

	2022	2023
BMSB exposure	August 7th – August 9th	August 4th – August 8th
T1	August 9th	August 16th
T2	August 11th	August 23rd
T3	August 16th	August 29th
T4	August 22nd	September 5th
T5	August 29th	September 11th
T6	September 5th	September 14th
T7	September 9th	September 18th
T8	September 12th	September 21st

Table 1. *Schedule for exposure of the fruits to BMSB in the orchard and image acquisition times (T1 – T8); T1 corresponds to the harvesting day.*

Between the different acquisition sessions, the pears were stored at refrigerated temperatures of 0 – 2°C to simulate post-harvest storage conditions [42,43].

In order to acquire the whole fruit surface, each pear was divided into four longitudinal sections labelled as A, B, C and D (**Figure 1 A**). Therefore, one hyperspectral image for each section was acquired for a total of four images for each fruit.

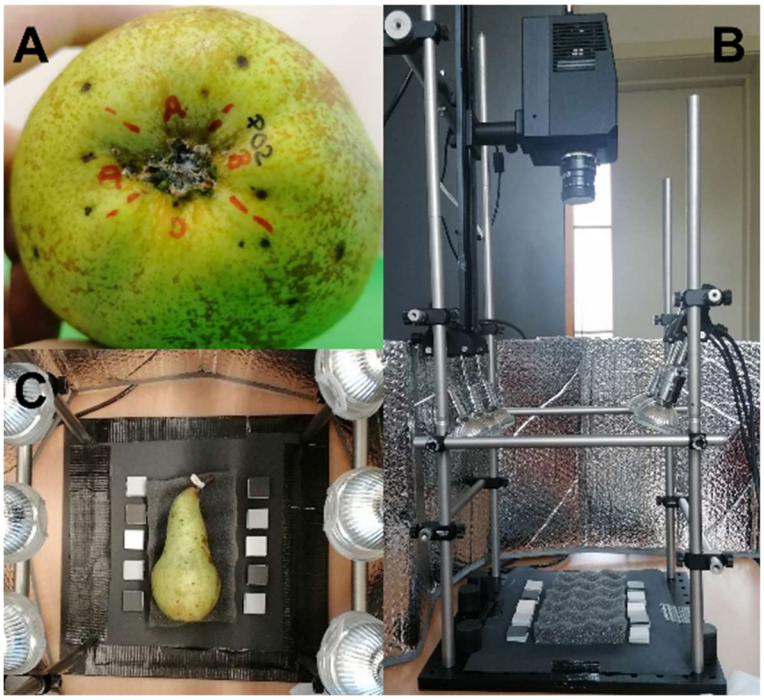


Figure 1. *In A) example of one fruit divided onto four longitudinal sections, in B) and C) image acquisition set-up.*

The hyperspectral images were acquired using a line-scanning hyperspectral camera (SnapScan SWIR, IMEC One Planet, The Netherlands) working in the 1156 – 1674 nm spectral range with a

spectral resolution of 100 spectral channels [44]. The images were acquired considering a lens aperture equal to f/2.8 and an integration time equal to 2 ms. Four small halogen lamps with diffusors were utilized for the homogeneous illumination of the samples (**Figure 1 B**). The acquired images have a spatial resolution of 492×820 pixels. Before data acquisition, the camera was calibrated using the acquisition software. This calibration was performed considering the dark current of the device and a of a white tile with high reflectance.

A black silicon carbide paper was used as image background and the fruits were placed on a black foam rubber holder to prevent them from rolling during image acquisition. The image scene also included 10 ceramic tiles with different grayscale tones and intermediate reflectance values (**Figure 1 C**). These ceramic tiles were added to correct possible time-depended drifts if necessary. However, the comparison of some representative images acquired at different acquisition times in both 2022 and 2023 years allowed to verify the absence of unwanted variations over time. Therefore, it was not necessary to apply any correction procedure beside standard calibration procedure performed by the acquisition device.

Due to the large number of fruit samples to be acquired at each acquisition time and the need to acquire four images for each pear, corresponding to the four sections, the collected fruits were randomly divided into two groups. The fruits belonging to the first group (Group 1) were acquired at each acquisition time, while those belonging to the second group (Group 2) were acquired only at a specific acquisition time. At each acquisition time, the pears belonging to Group 2 were peeled immediately after hyperspectral image acquisition in order the verify the presence of the damage due to BMSB punctures. Conversely, the pears belonging to Group 1 were peeled only after hyperspectral image acquisition at T8.

At the end of the image acquisition procedure, 952 and 1012 hyperspectral images were acquired in 2022 and 2023, respectively. Therefore, the final dataset of hyperspectral images was composed of 1964 images, corresponding to 1.5 TB.

It was necessary to peel the collected fruits after image acquisition in order to verify the presence of damage due to BMSB punctures. Indeed, this damage affect the fruit pulp, and is not visible at the naked eye on the intact fruits. During this procedure we noticed that not all the fruits exposed to BMSB were actually punctured and, in some cases, signs of damage were not clearly visible and could also be ascribable to mild damage caused by other factors. In addition, some of the control fruits as well as some of the exposed fruits had severe damage due to other biotic agents (e.g., brown spots caused by moulds such as *Stemphylium vesicarium* or *Alternaria spp*., pear scab).

Thanks to the visual inspection of the peeled fruits, it was possible to divide the acquired images into the following categories:

- images of sound sections of control fruits (S-C), corresponding to the images of fruit sections of control samples without any sign of damage;
- images of damaged sections of control fruits (D-C), corresponding to the images of fruit sections of control samples showing damage;
- images of damaged sections of exposed fruits (D-E), corresponding to the images of fruit sections of samples exposed to BMSB with damage;
- images of sound sections of exposed fruits (S-E), corresponding to the images of fruit sections of samples exposed to BMSB but not showing any kind of damage.

Superficial discolorations or little black spots, which are typical of *Williams* variety, were not considered as damage. It was considered as damage any kind of spot, suberification, necrosis, moulded or deliquescent area present in the fruit pulp, underneath the fruit peel, visibly different from a completely sound area. The number of damaged areas was irrelevant since a fruit section was labelled as damaged if at least one damaged spot was identified.

Table 2 summarises the number of images belonging to the different categories for both 2022 and 2023 harvest years.

		2022		**2023**	
		# of images	%	# of images	%
Control fruits	Sound sections (S-C)	242	53.5%	340	69.1%
	Damaged sections (D-C)	210	46.5%	152	30.9%
	Total	452	-	492	-
Exposed fruits	Sound sections (S-E)	174	34.8%	157	30.2%
	Damaged sections (D-E)	326	65.2%	363	69.8%
	Total	500	-	520	-

Table 2. *Number of images acquired in 2022 and 2023 harvest years of sound sections of control fruits (S-C), damaged sections of control fruits (D-C), sound sections of exposed fruits (S-E) and damaged sections of exposed fruits (D-E). The percentage values are calculated considering for each year the total number of images of control and exposed fruits, respectively.*

Furthermore, based on visual inspection, the damage found in exposed fruits was classified into three categories based on damage type (**Table 3**):

- Type 1: mild damage of unknown origin;
- Type 2: damage ascribable to BMSB punctures (e.g., suberifications);
- Type 3: severe damage not caused by BMSB punctures (e.g., moulds or other diseases).

Some examples of the three damage types are reported in **Figure 2**. In some cases, the same fruit section showed several damaged areas pertaining to different damage types. If Type 1 and Type 2 damage were simultaneously found on a fruit section, that section was classified as Type 2. Conversely, if Type 1 or Type 2 damage were found together with a Type 3 damage, the considered section was classified as Type 3.

As reported in **Table 3**, in 2022 the number of images showing damage related to BMSB punctures (Type 2) was much lower than expected. Indeed, as previously mentioned in **Section 2.1**, the weather conditions of summer 2022 when the pears were exposed to BMSB were not conducive to bugs vitality and feeding. Conversely, thanks to more favourable weather conditions in summer 2023, it was possible to obtain a higher number of images of fruit sections with punctures due to BMSB.

From here onwards, the images belonging to D-E class with Type 2 damage will be referred to as P-E images (i.e., images of punctured sections of fruits exposed to BMSB).

		2022		**2023**	
		# of images	%	# of images	%
Type 1	Mild damage of unknown origin	229	70.2%	34	9.4%
Type 2	Damage ascribable to BMSB punctures	54	16.6%	298	82.1%
Type 3	Severe damage not ascribable to BMSB	43	13.2%	31	8.5%
Total		326	-	363	-

Table 3. *Subdivision of D-E images according to damage type based on visual inspection of the corresponding peeled fruit samples.*

Figure 2. *Example of fruit sections assigned to different damage categories.*

2.3. Image elaboration

The first step of image elaboration consisted in background removal. To this aim, a PCA model was calculated on each image considering mean center as preprocessing method. All the pixels with negative score values were ascribable to the background composed by the black silicon carbide sheet and the black foam rubber support, and thus they were removed.

An additional step was necessary to remove the pixels ascribable to the plastic label attached on pear pedicel, used to identify the fruit samples. These pixels were characterised by reflectance values higher than 0.5 reflectance units at 1398 nm; therefore, a thresholding procedure was performed considering this wavelength.

Finally, morphological erosion using a disk-shaped structuring element with radius equal to 2 pixels [45] was carried out to remove uninformative pixels belonging to fruit pedicel and calyx [46].

All these steps were performed to each image in an automated manner using routines written *ad hoc* in MATLAB environment (R2020a, The MathWorks, USA) based on Image Processing Toolbox (v. 11.1) and PLS_Toolbox (v. 8.8.1, Eigenvector Research Inc., USA).

Subsequently, *pixel-level* PCA models were calculated on some representative images belonging to P-E class to evaluate the differences between sound and punctured areas. These PCA models were calculated considering linear detrend and mean center as preprocessing methods.

2.4. Method for automated ROIs identification

For the development of the algorithm able to automatically identify the ROIs ascribable to punctured areas, we considered only the images of sound sections of control fruits (S-C) and of punctured sections of exposed fruits (P-E) acquired in 2022 harvest years. This choice was due to technical reasons, since the annotation method was developed between 2022 and 2023 harvesting campaigns. In this manner, the images acquired in 2023 were used in a second moment for a further validation of the method developed in this study for automated identification of punctured areas.

As shown in **Table 2** and **Table 3**, 2022 dataset consists of 242 S-C images and only 54 P-E images. Given the large imbalance in the number of available images for the two classes, 159 S-C images out of 242 were randomly selected and used for model development.

Therefore, the final dataset used for the development of the annotation algorithm was composed by 213 images: 159 S-C images and 54 P-E images. This dataset of hyperspectral images was divided into training set and test set as follows:

- 150 training images (TR), including 111 S-C images (belonging to 10 fruits) and 39 P-E images (belonging to 6 fruits);
- 63 test images (TS), including 48 S-C images (belonging to 6 fruits) and 15 P-E images (belonging to 5 fruits).

The subdivision of the images into TR and TS images was done by keeping into the same set images of the same fruit sample, also when it was acquired at the different acquisition times.

The algorithm developed in this study to perform the automated identification of ROIs can be summarized in the following key steps:

1. Conversion of the images into Common Space Hyperspectrograms (CSH);
2. Calculation of interval Partial Least Squares Discriminant Analysis (iPLS-DA) models on CSH dataset;
3. Image reconstruction of CSH intervals selected by iPLS-DA.

These steps are outlined in **Figure 3**, and they will be described in more detail in the following sections.

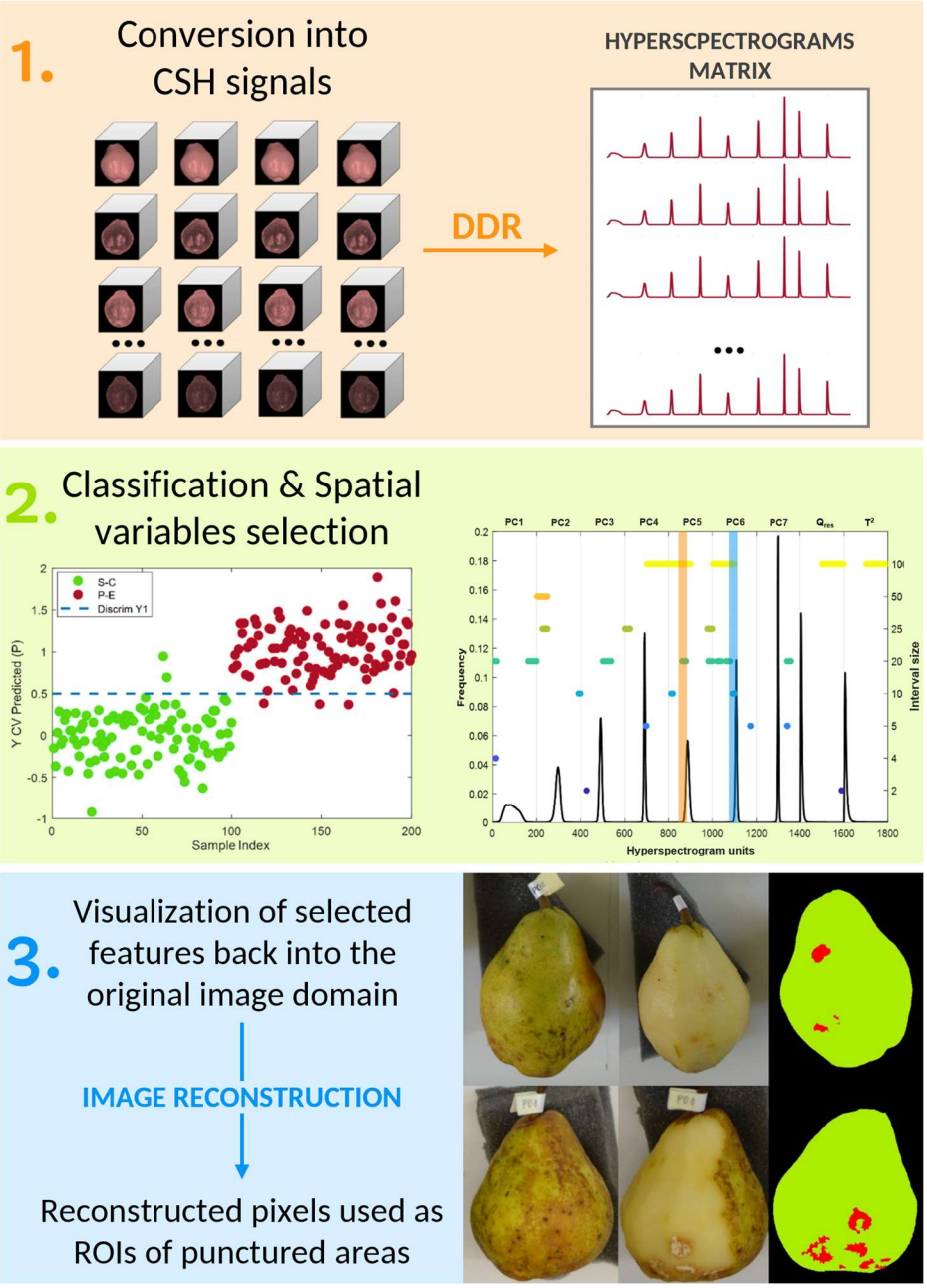


Figure 3. *Schematic representation of the key steps for the automated annotation of ROIs of punctured areas: 1) conversion of the hyperspectral images in CSH signals, 2) application of iPLS-DA for development of image-level classification models (sound vs punctured) and selection of the most relevant variables, 3) visualization of the selected variables back into the original image domain and use of the reconstructed pixels as ROIs ascribable to punctured areas.*

2.4.1. Conversion into *Common Space Hyperspectrograms*

Firstly, the hyperspectral images were converted into CSH signals. This operation allows to drastically reduce the dimensionality of the dataset since each image is converted into a features vector, i.e., the hyperspectrogram, which accounts for the relevant spatial and spectral information contained in the corresponding image [39,40,47,48]. CSH signals are generated by combining in

sequence the frequency distribution curves of quantities obtained from a PCA model common to all the TR images.

This conversion initially involved unfolding each hyperspectral image into two-dimensional data matrices, where the rows correspond to the pixels retained after background removal and erosion (*see* **Section 2.3**), and the columns are the spectral variables. Then, the unfolded hypercubes were preprocessed row-wise with linear detrend and centred according to the global average spectrum obtained by averaging all the pixel spectra retained in the TR images. For each image, the variance–covariance matrix was calculated, and the kernel variance–covariance matrix was then obtained by summing the single variance–covariance matrices of all the TR images [49]. The loading vectors of the PCA space common to all the TR images were obtained by applying singular value decomposition (SVD) to the kernel variance–covariance matrix, and 7 PCs were retained.

Subsequently, both TR and TS images were projected onto this common PC space and the corresponding score vectors, Q residuals and Hotelling's T^2 values were obtained. The CSH signal of each image was obtained by merging in sequence the frequency distribution curves of the seven score vectors, of Q residuals and of T^2 values (*see* key step 2 in **Figure 3**). To evaluate the effect of the bin size used in the computation of the frequency distribution curves, different CSH datasets were calculated considering bin size values equal to 100 and 200.

Therefore, considering bin size equal to 100 the corresponding TR and TS datasets were composed by 900 variables (= 100 bins × [7 score vectors + Q residuals + T^2]), while for bin size equal to 200 the corresponding TR and TS datasets were composed by 1800 variables (= 200 bins × [7 score vectors + Q residuals + T^2]).

For each bin size, the range of the different frequency distribution curves was separately defined considering the minimum and maximum values of the corresponding quantity across all the TR images. In addition, for each CSH signal the frequency distribution curves were normalised by the number of pixels retained after background removal and erosion.

The CSH signals were calculated using *ad-hoc* routines written in MATLAB environment and based on a graphical user-friendly interface (Hyperspectrograms GUI) freely downloadable from https://www.chimslab.unimore.it/downloads/.

For further details about the procedure used to obtain CSH signals the reader is referred to [40].

2.4.2. Calculation of iPLS-DA models

CSH datasets were analysed by means of iPLS-DA algorithm [38] to discriminate images of sound and punctured fruits, using autoscaling as signal pretreatment. iPLS-DA is a wrapper variable selection method based on the subdivision of the whole signal into a defined number of intervals of

equal length. In the *forward* mode used in this study, classification models are iteratively calculated by adding intervals until the minimum classification error in cross-validation is reached [50]. According to the number of bins used to obtain the frequency distribution curves included in CSH signals, different interval sizes for iPLS-DA were tested. For the CSH dataset obtained considering 100 bins, the iPLS-DA models were calculated with interval sizes equal to 50, 25, 20, 10, 5, 4 and 2, while the iPLS-DA models developed for the CSH dataset with 200 bins considered interval sizes equal to 100, 50, 25, 20, 10, 5, 4 and 2.

A Leave-One-Fruit-Out cross-validation scheme was adopted: in each cross-validation iteration all the signals obtained from the images of the same fruit were left out.

The performances of the iPLS-DA models were evaluated in terms of sensitivity (SENS), specificity (SPEC) and efficiency (EFF) in cross-validation and in prediction [51].

However, it has to be highlighted that the iPLS-DA models were calculated with the main aim of selecting in an automated and objective manner intervals of hyperspectrogram variables that are mostly related to the presence of punctured areas. Therefore, the main aim of this part of the study was not to obtain classification models with satisfactory performances, but to check their ability to identify the ROIs ascribable to punctures. This latter task was performed using the image reconstruction procedure explained in the following section.

2.4.3. Image reconstruction of selected variables

Since CSH are obtained by merging in sequence the frequency distribution curves of score, Q residuals and T^2 vectors, performing variable selection on this kind of signals means selecting groups of pixels sharing similar features, which in turn are related to the problem of interest (i.e., the presence of punctures in this study).

For example, suppose that, after variable selection using iPLS-DA, the model selects one interval of 10 variables falling in the CSH region of the frequency distribution curve of PC3 and that the selected variables correspond to PC3 score values ranging from 0.35 to 0.45. Image reconstruction allows to visualise back into the original image domain the selected CSH variables by visualising only the pixels falling in the intervals of interest. Considering the previous example, we are interested in visualising back into the original image domain only the pixels with PC3 score values in the range 0.35 – 0.45. In our case these pixels are ideally ascribable to the punctures and constitute the ROIs.

Further details about the image reconstruction procedure are reported in [39] and [16].

In addition to comparing the ability to identify ROIs by individual iPLS-DA models, also the overall frequency of selection of the CSH regions resulting by the different interval sizes was taken into account. In fact, since the position and width of the useful CSH features are not known in advance, it

is advisable to consider different interval size values and then to focus on the regions that are the most frequently selected ones [52]. Therefore, the reconstruction of the CSH regions most frequently selected was also performed by analysing the two CSH datasets (100 bins and 200 bins) separately from each other. In this way it was possible to highlight the most useful CSH regions regardless of the specific iPLS-DA interval size, and to use them to improve ROIs selection.

3. Results and discussion

3.1. Pixel-level exploratory analysis

A preliminary exploratory analysis with PCA was performed on some representative images of P-E fruits harvested in 2022 to assess the spectral differences between pixels ascribable to punctured and sound areas. For this investigation, the segmented hyperspectral images were preprocessed row-wise considering different methods (i.e. SNV, derivatives, detrend) combined with mean centering. The linear-detrend row preprocessing was the one allowing a better separation between pixels related to punctured and sound areas.

According to the outcomes of the PCA models calculated on single P-E images, a clear separation between sound areas and punctures was hard to achieve. First of all, the information related to the presence of actual damage related to BMSB activity represented less than 0.10 % of the explained variance and, secondly, in some images this kind of information was retrieved by different PCs or diverse intervals of PC score values.

As an example, **Figure 4** reports the PCA results obtained from two images of different punctured fruits acquired at T6 and T7 acquisition times, respectively. For both images, PC5 is the component allowing to better locate the punctures thanks to the visualisation of the corresponding PC5 score images and their comparison with the RGB images of peeled fruits. However, observing the PC1 *vs* PC5 score plots and the PC5 histograms, it is clear that pixels of punctured areas do not form a well-defined cluster distinct from the pixels of sound areas.

Using the brushing approach [24], for both images it was possible to manually identify PC5 score values mainly characterising punctured areas thanks to the possibility of interactively visualising the corresponding pixels (highlighted in magenta colour in **Figure 4 B-D** and **G-I**). For the image acquired at T6, the pixels of punctured areas fall into the -0.012 – -0.048 PC5 score range (**Figure 4 B-D**) while for the image acquired at T7 pixels of punctured areas have PC5 score values from -0.018 – -0.040 (**Figure 4 G-I**). Even if these two intervals partially overlap, it is almost impossible to identify threshold values suitable for both images and the absence of a clear cluster of pixels of the punctures makes it difficult to apply traditional automated thresholding methods like the

Otsu algorithm. Moreover, the areas ascribable to punctures are characterized by strongly irregular shapes not particularly dense in pixels, determining blurred edges between sound and punctured areas (**Figure 4 B** and **G**), thus not allowing a clear separation using morphological operators.

Another aspect that further complicated the selection of punctured areas was that in some images, PCs other than PC5 were also involved in punctures detection, making the thresholding procedure even more difficult since multiple PCs had to be considered at the same time.

All these aspects made the correct identification of ROIs ascribable to the punctures a very challenging task. Basically, it would have been necessary to perform PCA at the pixel-level on each image of P-E fruits to manually identify the ROIs, but considering the high number of images acquired, this operation was not feasible. In addition, the threshold selection on relevant score values and related pixels ascribable to ROIs would have been subjective and strictly dependent on the operator. Therefore, to perform a selection of ROIs in an objective and automatized way, it was necessary to adopt image-level analysis strategies (*see* **Section 2.4**).

Interestingly, the spectral information related to the presence of punctures reported by the loadings of PC5 in **Figure 4 E** and **J** seemed coherent on the majority of investigated P-E images, for different types of BMSB related damage (i.e. suberifications and necrosis of fruit pulp, which are shown in **Figure 4 A** and **F** respectively). The more relevant spectral regions in the loadings of PC5 fall into the 1100-1250 nm spectral range (C−H second overtone), in the 1370-1420 nm spectral range ($C-H_2$ combination band, aromatic C−H combination band and O−H stretch first overtone), in the 1320-1370 nm ($C-H_3$ combination band) and in the 1430-1500 nm region (C=O stretch third overtone and N-H stretch first overtone) [53,54]. These outcomes suggest that the spectral differences between damaged and sound areas reflect the different chemical composition of the tissues. Indeed, different studies reported a change of sugars and polyphenols content in fruits as a metabolic response to BMSB punctures [55–57]. Interestingly, Zamljen et al., 2021 [57] demonstrated that the metabolic response of apples fruits (*Malus* spp.) towards BMSB punctures involves an increase in sugars and polyphenols (i.e. hydroxycinnamic acids, flavanols) content only in the damaged areas. To our knowledge, similar studies have not been performed on pears. Nonetheless, since both apple (*Malus* spp.) and pear (*Pyrus* spp.) fruits belong to the *Maloideae* sub-family, they may exhibit a similar metabolic behaviour [58,59]. It has to be considered that different damaging agents or stresses (e.g., moulds or other diseases) may induce different metabolic responses in fruits, determining different chemical modifications of the fruit pulp. However, these evaluations are out of the aim of current manuscript as this study is focused on identifying damage caused by BMSB punctures.

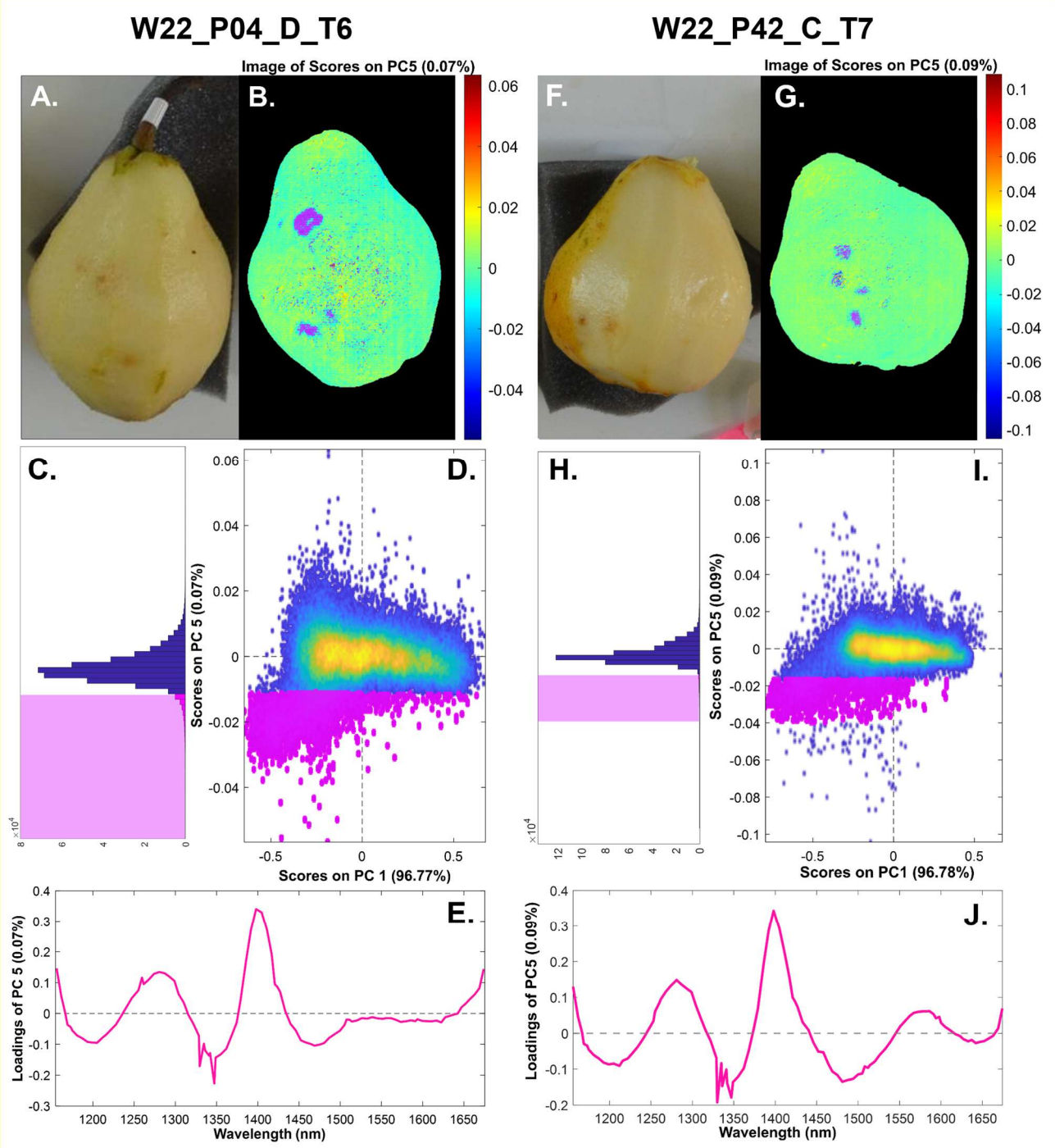


Figure 4. *PCA results of different P-E fruits images acquired at T6 and at T7. In (A, F) the RGB of the peeled fruit used as reference, in (B, G) PC5 score image, in (D, I) PC1-PC5 score plot, in (C, H) histogram of PC5 scores, in (E, J) loading vector of PC5. In (B-D, G-I) the pixels highlighted in magenta correspond to those selected as ROIs of punctured areas using the brushing tool.*

3.2. Image-level classification and variable selection by iPLS-DA

Starting from CSH datasets calculated considering 100 and 200 bins, both PLS-DA and iPLS-DA classification models were calculated to discriminate between P-E and S-C hyperspectrograms. Considering iPLS-DA, different interval sizes were tested (*see* **Section 2.4.2**). The classification

results of all the calculated models are summarized as heatmap of cross-validation efficiency (EFF CV) values in **Figure 5**. Unfortunately, PLS-DA led to overall poor classification results, with EFF CV values equal to 0.381 and 0.417 for the CSH datasets calculated with 100 and 200 bins, respectively.

Moving to iPLS-DA results, it is possible to observe an improvement of the classification results of some models with respect to PLS-DA. More in detail, the iPLS-DA model calculated on the CSH 200 bins dataset with interval size of 20 variables led to the highest EFF CV value, corresponding to 0.766. Acceptable results were obtained also for interval sizes of 50 and 100 variables for CSH 200 bins dataset, and interval sizes of 4, 5 and 10 variables for CSH 100 bins dataset. For each CSH dataset, the three best performing models were selected based on EFF CV values and further investigated by calculating their prediction performances using the corresponding CSH test set.

The cross-validation and test set prediction results of the selected models are reported in **Table 4**, in terms of SENS, SPEC and EFF values referred to the P-E class.

It is possible to observe that these models generally have high SPEC but poor SENS values for the P-E class, suggesting that they do not correctly recognise the punctured areas of the fruits. Moving to the prediction results, there is a general slight improvement in the model performances, in particular considering the SENS values. The only exception is the model calculated from the CSH 200 bins dataset and considering an interval size of 50 variables, which led to very poor classification results.

Unfortunately, not completely satisfactory results were obtained from the iPLS-DA classification models (**Table 4**). This outcome can be motivated also considering notable imbalance of samples between the two classes. As shown in **Table 3**, the sections of fruits exposed to BMSB that actually exhibited punctures (P-E) were significantly lower than the sections of sound control fruits (S-C). Moreover, the overlapping of the two classes and the rather low variance explained by the actual presence of damage may concur to these outcomes (**Figure 4**). However, the main focus of this part of the study was not to obtain satisfactory image-level classification results, but to develop an objective and automated method to identify the ROIs ascribable to BMSB damage.

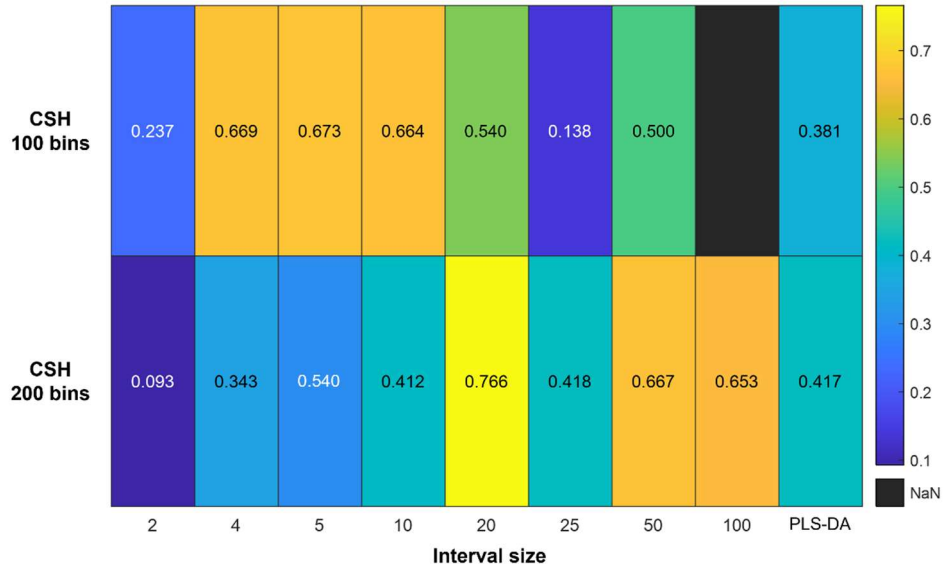


Figure 5 *Heatmap of efficiency values obtained in cross-validation (EFF CV) for the PLS-DA and iPLS-DA models calculated from 100 bins and 200 bins CSH datasets.*

			Classification results						Number of images with identified ROIs			
			CV			PRED			PC selected			Tot.
CSH	Interval size	LVs	SENS	SPEC	EFF	SENS	SPEC	EFF	PC1	PC5	PC6	
100 bins	4	6	0.513	0.874	0.669	0.600	0.833	0.707	7	12	-	**15**
	5	6	0.513	0.863	0.673	0.733	0.792	0.762	8	24	0	**26**
	10	5	0.538	0.820	0.664	0.600	0.813	0.698	8	24	0	**26**
200 bins	20	6	0.692	0.847	0.766	0.800	0.604	0.695	3	12	0	**15**
	50	3	0.462	0.964	0.667	0.133	1.00	0.365	-	-	-	**0**
	100	5	0.539	0.793	0.653	0.867	0.854	0.860	-	0	10	**10**

Table 4. *Classification performances of the three best iPLS-DA models calculated for each CSH dataset (SENS and SPEC values are referred to the P-E class) and corresponding number of P-E images for which the ROIs of punctured regions were successfully reconstructed. Only the selected PCs leading to ROIs identification are reported. In the columns related to the number of images with identified ROIs, "0" indicates that the interval was selected but no ROIs were reconstructed while "-" indicates no interval selection.*

3.3. ROIs visualisation based on iPLS-DA selected variables

The hyperspectrogram variables selected by iPLS-DA correspond to intervals of score values that are likely to be related to the P-E class, which in turn are more likely to correspond to the pixels of punctured areas. Indeed, the selected variables can be visualised back into the original image domain using the image reconstruction procedure illustrated in **Section 2.4.3**.

Based on these considerations, for each one of the best performing iPLS-DA models reported in **Table 4**, image reconstruction of P-E images was carried out considering the selected hyperspectrograms variables.

The image reconstruction results are shown in **Table 4**, reporting the number of P-E images for which the ROIs of punctured regions were reconstructed considering the selected intervals of a specific frequency distribution curve composing the CSH signals, and the overall number of reconstructed images.

Considering the CSH 100 bins dataset, the interval size 5 and interval size 10 models allowed to identify the highest number of ROIs for 26 P-E images out of 54; the selected intervals fall in the PC1 and PC5 frequency distribution curves of the CSH signals. These are the best performing models in terms of number of reconstructed images, although they are not the models leading to the best image-level classification results.

While being the best in terms of EFF CV values obtained globally, the iPLS-DA model calculated on the CSH 200 bins dataset considering interval size of 20 variables led to the visualization of ROIs on only 15 images out of 54. Also in this case, the relevant intervals selected fall into the PC1 and PC5 frequency distribution curves of CSH signals.

The iPLS-DA model calculated considering an interval size of 50 intervals on CSH 200 bins dataset led to acceptable results in cross-validation but very poor in prediction. Interestingly, this model led to the selection of only one interval falling in the PC2 scores frequency distribution curves which didn't allow to visualize any ROI.

Lastly, the iPLS-DA model with interval size of 100 variables on CSH 200 bins dataset allowed the visualization of ROIs on 10 P-E images out of 54. Although not the best in cross-validation, this model led to the best results in prediction and, more importantly, allowed the ROIs visualization considering an interval falling in the frequency distribution curve of PC6. This hyperspectrogram region has been frequently selected also for other models, but did not result in ROIs visualization.

In order to improve the results in terms of number of P-E images with ROIs of punctures automatically identified, it is crucial to gather all the relevant spatial features (i.e., intervals selected from CSH signals) related to the presence of damage. Therefore, considering the 100 bins and 200 bins CSH datasets separately, the selected intervals were also evaluated based on their frequency of selection by the different iPLS-DA models with varying intervals sizes.

This operation was done based on the assumption that intervals of variables selected more frequently are more likely related to the presence of BMSB damage, regardless of the performance of the single models.

Figure 6 displays the intervals selected by iPLS-DA models calculated on the 100 bins and 200 bins CSH datasets: the frequencies of selection are shown in **Figures 6 A** and **6 B**, while **Figures 6 C** and **6 D** report the variables selected from each iPLS-DA model.

For each CSH dataset we considered the variables selected most frequently by the different iPLS-DA models, and for each interval we visualised the corresponding reconstructed images to verify the correct identification of ROIs related to punctures. The results are reported in **Table 5** as number of P-E images acquired in 2022 correctly reconstructed.

Considering CSH 100 bins dataset, the most frequently selected intervals correspond to variables selected three times and they fall in CSH regions of the frequency distribution curves of PC1 and PC5. The reconstruction of these intervals led to the correct visualisation of the ROIs of punctured regions in 26 images out of the 54 P-E images acquired in 2022, corresponding to 6 fruits out of 11 for which it was possible to annotate the punctures. For the majority of the reconstructed images, the interval selected on PC5 is the one providing the visualisation of the ROIs, however for some images also the interval selected on PC1 allowed to reconstruct the pixels of punctured areas. This finding is coherent with the outcomes of the single iPLS-DA models reported in **Table 4**. Variables selected at least twice were also evaluated, but did not result in additional ROIs visualisation.

Considering the results obtained for CSH 200 bins dataset, image reconstruction was performed with the variables selected twice by the different iPLS-DA models and corresponding to CSH regions falling in the PC1, PC2, PC5, PC6, PC7, and Q residuals frequency distribution curves. Among these intervals, only those falling in the PC5 and PC6 frequency distribution curves (highlighted in **Figure 6 D**) led to the correct visualisation of the ROIs of punctured areas in 32 P-E images out of 54 P-E images of 2022 (**Table 5**), corresponding to 8 fruits out of 11. These interval variables correspond to PC5 score values ranging from -0,025 to -0,016 and PC6 score values ranging from -0,020 e -0,015.

The selection based on the variables selected most frequently from the different iPLS-DA models on CSH 200 bins dataset allowed to outperform the image reconstruction results obtained by the individual iPLS-DA models. It is likely that a higher number of bins for the calculation of the frequency distribution curves composing the hyperspectrograms eased the separation between spatial features related to pixels of punctured areas and the ones related to sound pixels, thus allowing a better ROIs visualization for the CSH 200 bins dataset with respect to the 100 bins dataset.

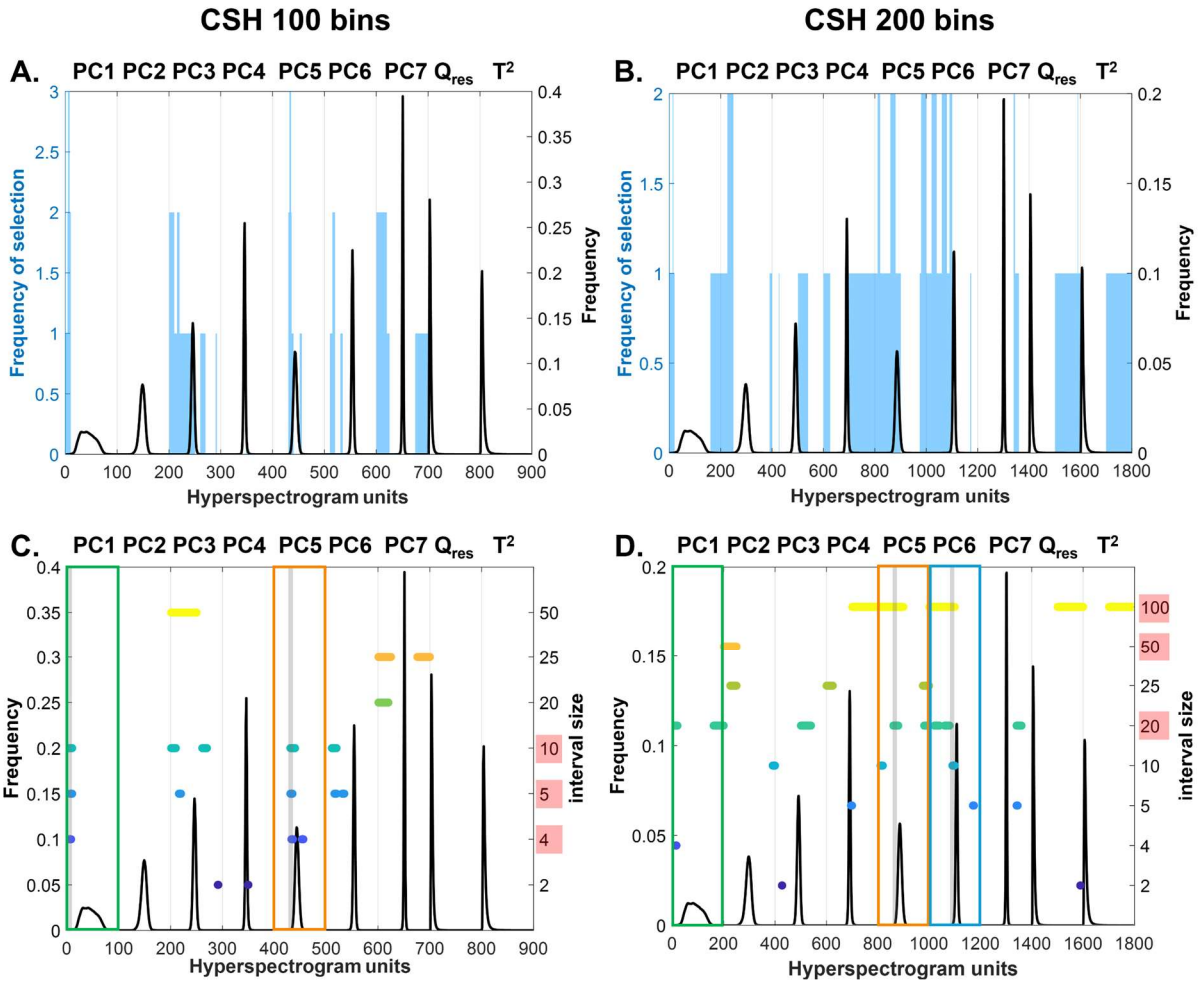


Figure 6. *Frequency of selection of the CSH variables considering the different iPLS-DA models calculated for CSH 100 bins (A) and CSH 200 bins (B) datasets and CSH variables selected by the iPLS-DA models calculated with different interval sizes (C and D). In C) and D) the highlighted interval size values correspond to those leading to the correct visualisation of ROIs of punctured regions using the image reconstruction procedure. The framed frequency distribution curves correspond to the PCs that led to at least one ROI reconstruction (see Table 4).*

CSH	Frequency	PC1	PC5	PC6	# P-E Images /Tot. 54	# P-E fruits /Tot. 11
200 *bins*	2	0	21	11	**32**	**8**
100 *bins*	3	8	24	-	**26**	**6**

Table 5. *Number of P-E images with reconstructed ROIs obtained considering the most frequently selected intervals of score values for the CSH 100 bins and CSH 200 bins datasets. The number of P-E images is subdivided based on the corresponding PC. The total number of P-E images and P-E fruits with identified ROIs is also reported.*

Figure 7 shows some examples of image reconstruction based on the PC5 and PC6 most frequently selected intervals considering the CSH 200 bins dataset. The image reconstruction and ROIs visualisation are shown not only for P-E images, but also for a S-C image, a D-E "Type 3" image

(i.e., damaged-exposed with severe damage not ascribable to BMSB), and a D-C image. Furthermore, the examples of ROIs reconstruction are provided with the corresponding RGB images of unpeeled and peeled fruits as reference.

More in detail, **Figure 7 A** reports the image reconstruction results obtained for a S-C image (i.e., sound section of a control fruit), where only few sparse pixels falling in the PC6 selected interval were reconstructed back into the original image domain. However, these pixels can be easily filtered out by applying morphological operators to the reconstructed image. In particular, the combination of *erosion* (disk-shaped structuring element with radius of 2 pixels) and *filling holes* morphological operators allowed to completely remove these pixels, as shown in the final ROIs visualization.

On the other hand, the spatial features selected on CSH signals allowed to reconstruct image regions corresponding to damaged areas of P-E images (**Figure 7 B-C**). Generally, as shown in orange pixels in **Figure 7 B**, the reconstruction of the selected PC5 interval corresponds to the pixels belonging to punctures, specifically ascribable to the edges between sound and punctured areas. In some other cases, the selected interval on PC6 scores brought relevant information about the presence of punctured areas, and the corresponding reconstructed pixels usually fell inside the edges of the punctured areas (blue pixels in **Figure 7 C**).

Also in the case of P-E images, after image reconstruction of the selected spatial features of CSH signals, the same morphological operators described above were applied to further optimize ROIs selection. The ROI visualization is reported in red colour in the lowest row of **Figure 7 B-C**. Comparing the final ROI annotation of punctured areas with the corresponding RGB image of the peeled fruit section, it is possible to verify the correct identification of punctured regions. Furthermore, considering the RGB images of the unpeeled fruit sections it is possible to observe that the damage occurs in the pulp under the peel and they are not visible at the naked eye.

The procedure developed in this study allowed the correct reconstruction of ROIs of both suberified and necrotic areas, which are the typical types of damage caused by BMSB punctures. However, suberification and necrosis of the fruit pulp may also be due to other causes as they are a common response of plants and corresponding fruits to stress. For this reason, we tested the ability of the proposed approach to visualize this damage when it is not caused by BMSB punctures. To this aim, **Figure 7 D** shows the correct reconstruction of a suberification damage found in a section of a control fruit (D-C), therefore not due to BMSB. Suberified tissues have the same chemical composition regardless of the agents that caused them to emerge, and fruits with suberified areas are considered of low quality. Therefore, we verified that the developed approach is capable of reconstructing damage with the same chemical composition, although not directly caused by BMSB.

Finally, we also tested the proposed approach in identifying damage different from suberification and necrosis, such as mould. **Figure 7 E** shows one section of a pear exposed to BMSB (D-E) which exhibits both BMSB-induced suberification and a moulded area. In this case, the identified ROI is only related to the suberified area, suggesting that the selected hyperspectrogram intervals are specific only for damage similar to that caused by BMSB.

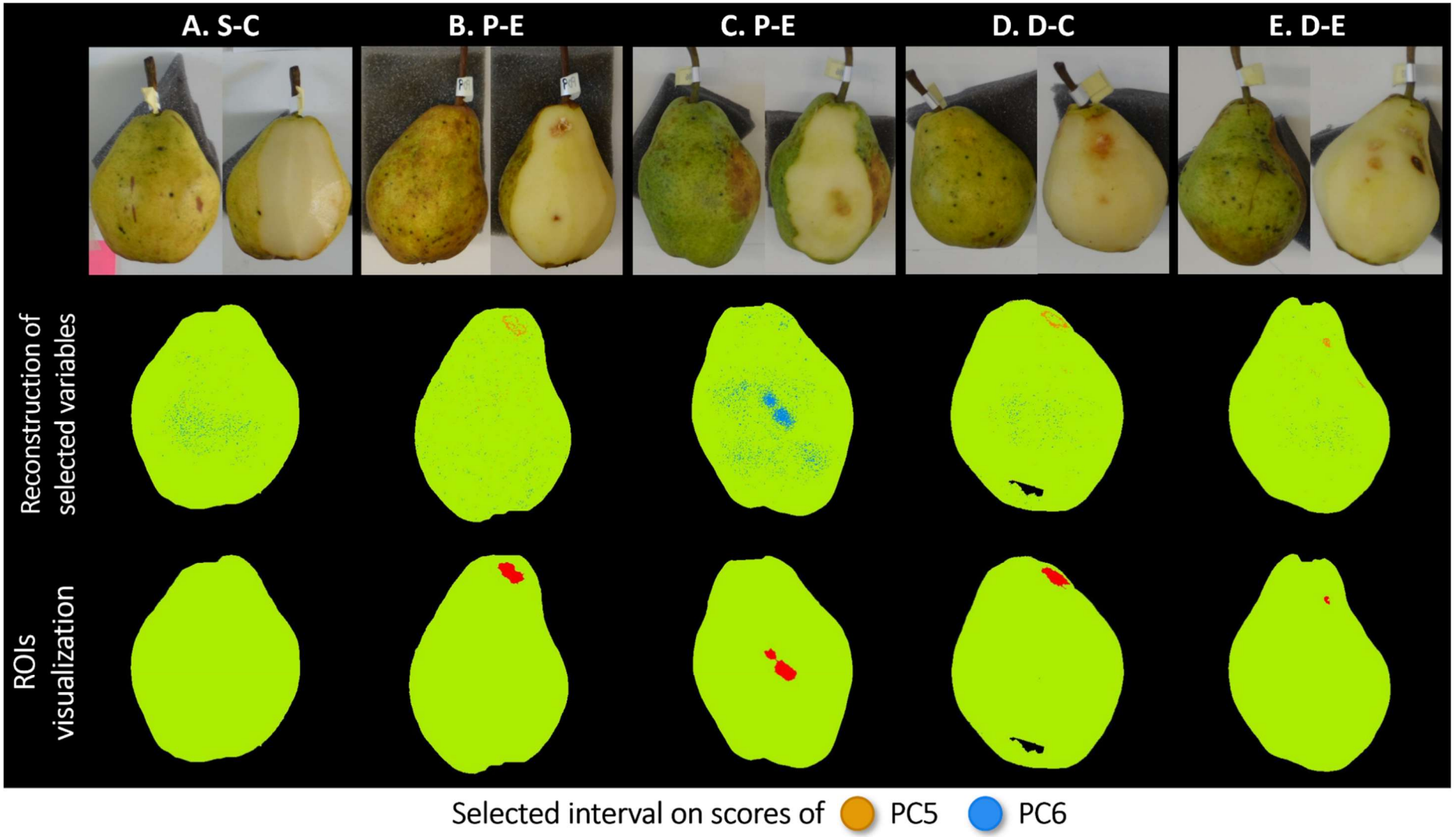


Figure 7. *Reconstruction in the original image domain of PC5 (orange) and PC6 (blue) selected intervals ascribable to punctures for S-C (Sound-Control), P-E (Punctured-Exposed), D-C (Damaged-Control) and D-E (Damaged-Exposed) pears. The final ROIs annotation obtained after erosion and filling holes morphological operators is shown in red colour. Reconstructed images are reported together with the corresponding RGB images of the fruits before and after peeling.*

3.4. Validation of the ROIs selection method

As previously described in **Section 2.4**, the automated ROIs visualization method was developed based on the images acquired in 2022. Specifically, the CSH intervals falling in the PC5 and PC6 frequency distribution curves used for image reconstruction were selected considering the best results in terms of correctly reconstructed P-E images.

Based on the developed approach, we correctly annotated the ROIs of punctured areas in 32 out of 54 P-E images acquired in 2022, corresponding to 59.2% (**Table 6**). Since part of the collected fruits were acquired at different acquisition times, we also evaluated the number of fruits for which we had the correct visualization of damaged areas. Considering 2022 samples, we annotated images corresponding to 8 damaged by BMSB exposed fruits out of 11 fruits (72.7%).

The method for automated identification of ROIs of punctured areas was further validated considering the P-E hyperspectral images acquired in 2023. The corresponding CSH 200 bins dataset was calculated by projecting the 2023 images on the common PC space calculated for 2022 images (*see* **Section 2.4.1**); then, the selected intervals on PC5 and PC6 frequency distribution curves were visualised back into the original image domain. As shown in **Table 6**, satisfactory results were obtained in this validation step, with 240 out of 298 of 2023 P-E images (80.5%) for which it was possible to correctly reconstruct the ROIs of damaged areas. The outcomes out this validation step are even more satisfactory considering that we were able to correctly visualise the ROIs of punctures in at least one image of all the punctured fruits collected in 2023.

The different performances obtained in ROIs reconstruction for the P-E images acquired in 2022 and 2023 are probably due to the different conditions occurred in the two years. Indeed, climatic conditions in 2022 were not favourable for BMSB feeding activity, resulting in the collection of a limited number of punctured fruits. Conversely, climatic conditions in summer 2023 were more favourable for BMSB vitality, leading to a greater intensity and incidence of damage compared to summer 2022, thus allowing a greater number of P-E fruits to be collected.

	Harvest year	
	2022	**2023**
Reconstructed P-E images	32	240
Total P-E images	54	298
% Reconstructed P-E images	**59.2 %**	**80.5 %**
Reconstructed P-E fruits	8	18
Total P-E fruits	11	18
% Reconstructed P-E fruits	**72.7 %**	**100 %**

Table 6. *Results obtained from the ROIs visualisation method for P-E images acquired in 2022 and 2023.*

The ability of the methodology towards the identification of punctures during post-harvest storage was evaluated on P-E pears acquired at each acquisition time (*see* **Section 2.2**).

Figure 8 reports an example of the reconstructed pixels and the final ROIs annotation for P-E pears harvested in 2022 and 2023. For instance, the damaged areas reconstructed by the selected PC5 score interval on P-E images of the pear harvested in 2023 are consistent. The identification of ROIs ascribable to the punctures was feasible at any time, from harvest to six weeks later in post-harvest storage conditions. Nonetheless, the different rotation of the fruits during the subsequent HSI acquisition times had a major impact on the visualization of the ROIs, especially for the image acquired at T2 of P-E harvested in 2022, where the punctured area was inadvertently not imaged.

For both P-E fruits shown in **Figure 8**, the reconstruction of the selected PC5 interval in orange corresponds to the pixels belonging to the punctured area, specifically related to the edges between sound and punctured regions. As already observed, PC5 seems retaining the predominant information able to separate the sound areas from the punctured ones, thus allowing the reconstruction of a greater number of P-E images with ROIs correctly identified. In addition, the wavelengths characterized by the highest absolute values of loadings of PC5 calculated on the 200 bins CSH dataset (**Figure 9**) are coherent with the higher absolute values of loadings of PC5 obtained with a PCA model calculated on an individual image (**Figure 4 E** and **J**). As described in **Section 3.1**, the more relevant spectral regions seem to confirm again a possible variation in sugars and polyphenols content in punctured tissues [55–57].

Table 7 reports the outcomes for reconstructed images belonging to P-E fruits, subdivided by acquisition time (T1-T8) and harvest year. Overall, the percentage of P-E images harvested in 2022 with ROIs reconstructed is lower than the ones harvested in 2023: the worst performances for 2022 samples are due to the fewer P-E images collected for each acquisition time and to the three P-E fruits for which ROIs were not annotated. In a few cases, the areas ascribable to the punctures are partially or not annotated due to the fragmented reconstruction of ROIs with lower pixels density, which are subsequently discarded as noise.

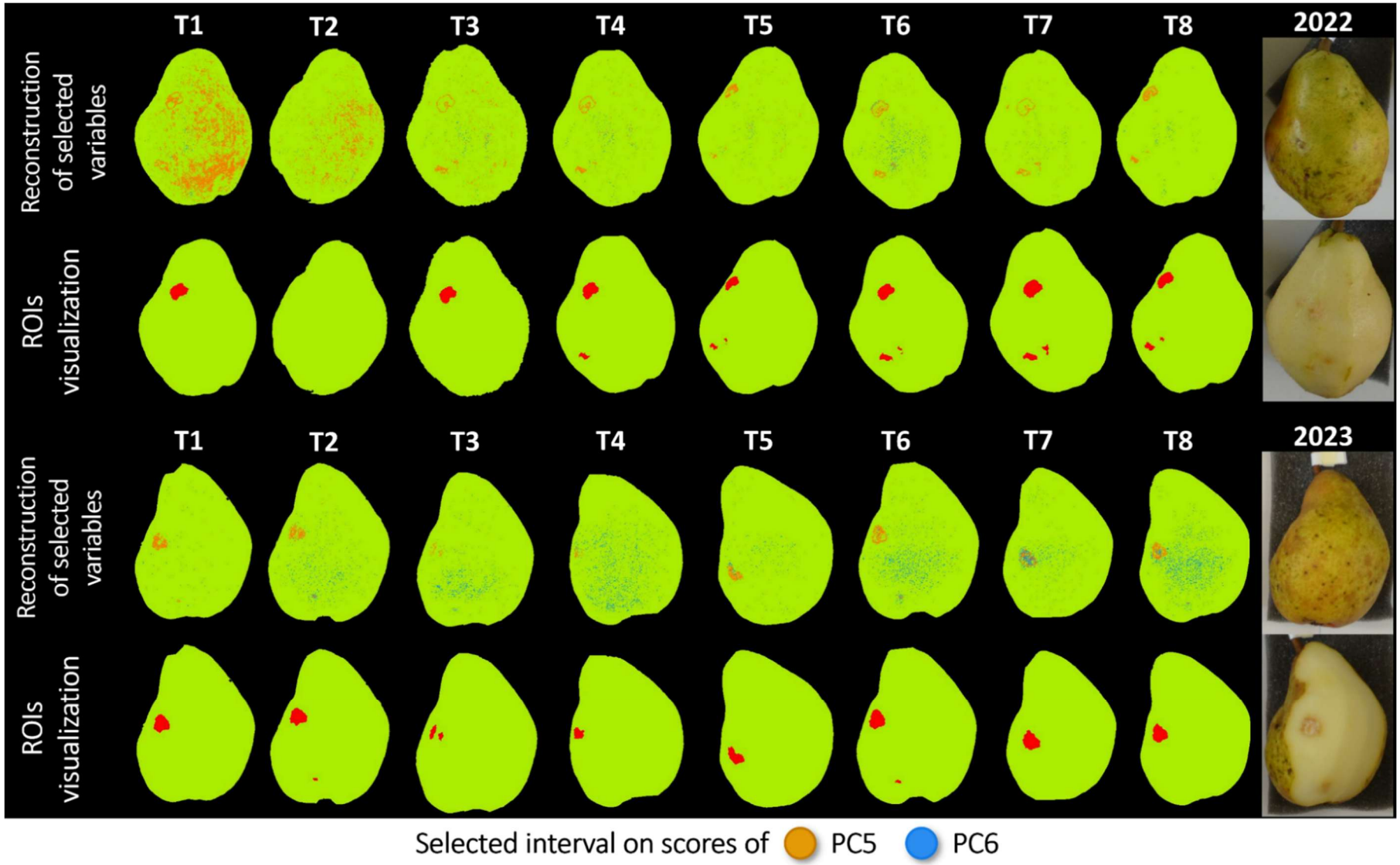


Figure 8. *Reconstruction in the original image domain of score values of PC5 (orange) and PC6 (blue) ascribable to punctures for P-E (Punctured-Exposed) pears harvested in 2022 and 2023 over time. Along with the RGB references of the fruits before and after peeling, the final reconstructions are shown in red.*

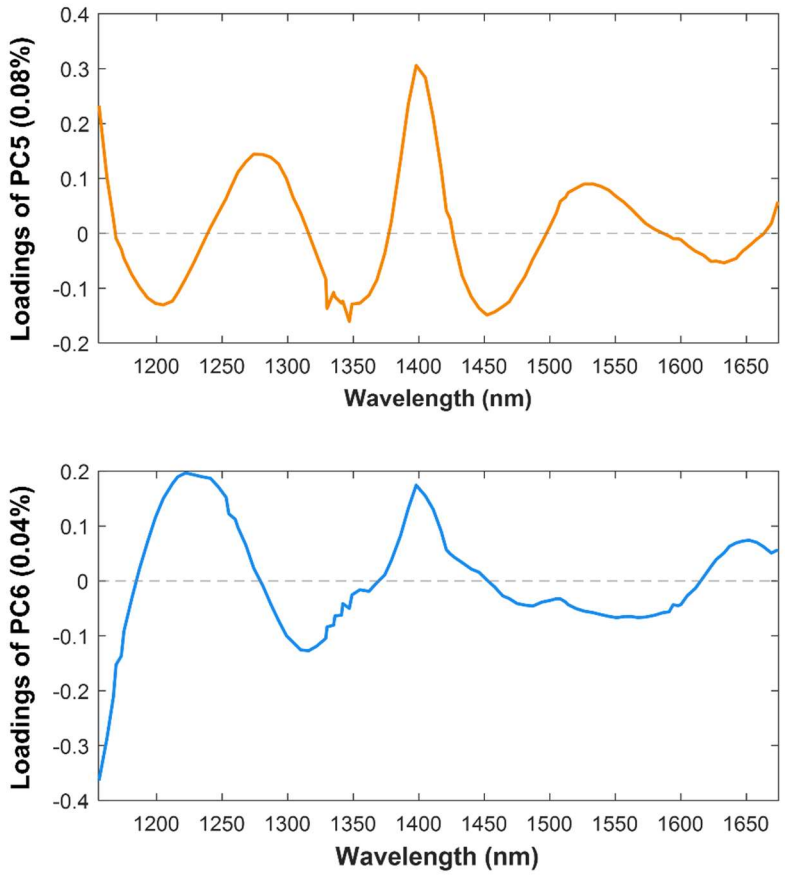


Figure 9. *Loadings of PC5 and PC6 resulted by the global PCA model used for the calculation of CSH 200 bins.*

		Harvest year					
		2022			2023		
		# Reconstructed P-E images	# P-E images	%	# Reconstructed P-E images	# P-E images	%
Acquisition Times	T1	3	4	75.0%	34	41	82.9%
	T2	2	5	40.0%	25	34	73.5%
	T3	4	9	44.4%	26	33	78.8%
	T4	3	4	75.0%	34	41	82.9%
	T5	3	4	75.0%	27	35	77.1%
	T6	6	8	75.0%	34	38	89.5%
	T7	8	14	57.1%	33	40	82.5%
	T8	3	6	50.0%	27	36	75.0%
	Tot.	**32**	**54**	**59.2%**	**240**	**298**	**80.5%**

Table 7. *Reconstructed P-E images and corresponding fruits grouped according to acquisition time and harvest year.*

4. Conclusions

In the present work, an innovative approach is proposed for the automated and objective identification of ROIs related to punctured areas on NIR hyperspectral images of pears.

In the investigated case study, the identification of one or more threshold values to select the pixels of punctured regions was a challenging task due to the complexity of the dataset.

Indeed, a preliminary pixel-level analysis performed by PCA on individual images highlighted the absence of a clear separation between pixels of punctured regions and pixels of sound areas, therefore not allowing the identification of a unique threshold on PC scores common to all the images. On the other hand, performing manual selection of ROIs by means of the evaluation of PCA models calculated on each single image was not feasible given the high number of images and its arbitrariness.

To overcome these limits, we developed an innovative method for the objective and automated identification of ROIs based on the calculation of Common Space Hyperspectrograms (CSH) and image-level classification coupled with spatial features selection using iPLS-DA. The dimensionality reduction of the image data enabled by CSH was fundamental to handle the large size of the investigated dataset, allowing to compress the relevant information into a global PCA space common to the training set images.

Despite not leading to satisfactory classification results at the image-level, the spatial variables selected more frequently by iPLS-DA models led to the selection of spatial features more correlated to the presence of punctures. Then, these spatial features were visualized back into the original domain, and this image reconstruction procedure, coupled with morphological operators, allowed to correctly annotate the ROIs of punctured areas in the majority of P-E images. Puncture-like damage not related to BMSB activity (e.g., suberifications) was successfully annotated while, vice versa, damage not attributable to BMSB (e.g., mould) was not reconstructed.

Even if the image compression performed by CSH approach can be seen as a loss of pixel-related resolution of the images, the outcomes of this study demonstrate that this approach to data reduction preserves the spatial information, allowing to effectively identify spatially resolved features like damaged spots.

Furthermore, it has to be highlighted that the proposed approach is not to be intended as a classification method, but it was developed only as a strategy to effectively identify ROIs of punctured areas in the hyperspectral images of the collected pears. Starting from the ROIs obtained following this procedure, the second part of the study will focus on the development of pixel-level classification models. Indeed, the selected ROIs will be used to create a dataset of representative spectra belonging

to both punctured and sound areas. In turn, the so obtained dataset will be used to develop supervised pixel-level classification models for predictive purposes in post-harvest sorting systems.
More in general, the proposed annotation algorithm can be used in all those situations were many images have to be analysed altogether and the identification of common threshold values for ROIs selection is very difficult due the complexity of the considered problem.

Acknowledgements

Authors wish to thank HALY.ID, project of ERA-NET Cofund ICT-AGRI-FOOD, with funding provided by national sources (Ministero delle politiche agricole e forestali, MIPAAF) and co-funding by the European Union's Horizon 2020 research and innovation program, Grant Agreement number 862671.
Dr. Rosalba Calvini and Dr. Elena Costi would like to thank the Italian funding programme Fondo Sociale Europeo REACT-EU - PON "Ricerca e Innovazione" 2014 – 2020 – Azione IV.6 Contratti di ricerca su tematiche Green (D.M. 1062 del 10/08/ 2021) for supporting their research (CUP: E95F21002330001).
The authors wish to express their gratitude to Dr. Niccolò Patelli (Applied Entomology Lab, UNIMORE) for the support on-field and Enrico Giovanella for the valuable technical support during the image acquisition and investigation.

CRediT Author Statement

Veronica Ferrari: Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Writing – Original draft; Visualisation. **Rosalba Calvini:** Conceptualisation; Methodology; Software; Investigation; Data curation; Writing – Review & editing; Visualisation. **Camilla Menozzi:** Investigation; Data curation; Writing – Review & editing. **Elena Costi:** Resources; Investigation. **Daniele Giannetti:** Resources; Investigation. **Peter Hoffermans:** Resources; Investigation; Project administration; Funding acquisition. **Lara Maistrello:** Resources; Writing – Review & editing; Project administration; Funding acquisition. **Alessandro Ulrici:** Conceptualisation; Methodology; Writing – Review & editing; Supervision; Project administration; Funding acquisition.

References

[1] K.B. Rice, C.J. Bergh, E.J. Bergmann, D.J. Biddinger, C. Dieckhoff, G. Dively, H. Fraser, T. Gariepy, G. Hamilton, T. Haye, A. Herbert, K. Hoelmer, C.R. Hooks, A. Jones, G. Krawczyk, T. Kuhar, H. Martinson, W. Mitchell, A.L. Nielsen, D.G. Pfeiffer, M.J. Raupp, C. Rodriguez-Saona, P. Shearer, P. Shrewsbury, P.D. Venugopal, J. Whalen, N.G. Wiman, T.C. Leskey, J.F. Tooker, Biology, Ecology, and

Management of Brown Marmorated Stink Bug (Hemiptera: Pentatomidae), *J. Integr. Pest Manag.* 5 (2014) A1–A13. https://doi.org/10.1603/IPM14002.

[2] T.C. Leskey, A.L. Nielsen, Impact of the Invasive Brown Marmorated Stink Bug in North America and Europe: History, Biology, Ecology, and Management, *Annu. Rev. Entomol.* 63 (2018) 599–618. https://doi.org/10.1146/ANNUREV-ENTO-020117-043226/CITE/REFWORKS.

[3] L. Maistrello, Case Study 2: *Halyomorpha halys* (Stål) in Europe, in: A.F. Bueno, A.R. Panizzi (Eds.), Stink Bugs (Hemiptera: Pentatomidae) Research and Management, Springer, 2024: pp. 271–359. https://doi.org/10.1007/978-3-031-69742-5_15.

[4] L. Maistrello, G. Vaccari, S. Caruso, E. Costi, S. Bortolini, L. Macavei, G. Foca, A. Ulrici, P.P. Bortolotti, R. Nannini, L. Casoli, M. Fornaciari, G.L. Mazzoli, P. Dioli, Monitoring of the invasive *Halyomorpha halys*, a new key pest of fruit orchards in northern Italy, *J. Pest Sci.* (2004) 90 (2017) 1231–1244. https://doi.org/10.1007/s10340-017-0896-2.

[5] A.L. Nielsen, G.C. Hamilton, Seasonal Occurrence and Impact of *Halyomorpha halys* (Hemiptera: Pentatomidae) in Tree Fruit, *J. Econ. Entomol.* 102 (2009) 1133–1140. https://doi.org/10.1603/029.102.0335.

[6] M. Bariselli, R. Bugiani, L. Maistrello, Distribution and damage caused by *Halyomorpha halys* in Italy, EPPO Bulletin 46 (2016) 332–334. https://doi.org/10.1111/epp.12289.

[7] A.L. Acebes-Doria, T.C. Leskey, J.C. Bergh, Injury to apples and peaches at harvest from feeding by *Halyomorpha halys* (Stål) (Hemiptera: Pentatomidae) nymphs early and late in the season, *Crop Protection* 89 (2016) 58–65. https://doi.org/10.1016/J.CROPRO.2016.06.022.

[8] N.-N. Wang, D.-W. Sun, Y.-C. Yang, H. Pu, Z. Zhu, Recent Advances in the Application of Hyperspectral Imaging for Evaluating Fruit Quality, *Food Anal. Methods* 9 (2016) 178–191. https://doi.org/10.1007/s12161-015-0153-3.

[9] Y. Lu, Y. Huang, R. Lu, Innovative Hyperspectral Imaging-Based Techniques for Quality Evaluation of Fruits and Vegetables: A Review, *Applied Sciences* 7 (2017), 189. https://doi.org/10.3390/APP7020189.

[10] D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O.L. García-Navarrete, J. Blasco, Recent Advances and Applications of Hyperspectral Imaging for Fruit and Vegetable Quality Assessment, *Food and Bioprocess Technology* 5 (2012) 1121–1142. https://doi.org/10.1007/s11947-011-0725-1.

[11] J. Wieme, K. Mollazade, I. Malounas, M. Zude-Sasse, M. Zhao, A. Gowen, D. Argyropoulos, S. Fountas, J. Van Beek, Application of hyperspectral imaging systems and artificial intelligence for quality assessment of fruit, vegetables and mushrooms: A review, *Biosyst. Eng.* 222 (2022) 156–176. https://doi.org/10.1016/J.BIOSYSTEMSENG.2022.07.013.

[12] B. Li, T. Ma, L. Bai, T. Inagaki, H. Seki, S. Tsuchikawa, Three-dimensional visualization and detection of early bruise in apple based on near-infrared hyperspectral imaging coupled with geometrical influence correction, *Postharvest Biol. Technol.* 210 (2024) 112753. https://doi.org/10.1016/j.postharvbio.2023.112753.

[13] Y. Bu, J. Luo, J. Li, Q. Chi, W. Guo, Detection of hidden bruises on kiwifruit using hyperspectral imaging combined with deep learning, *Int. J. Food Sci. Technol.* 59 (2024) 5975–5984. https://doi.org/10.1111/ijfs.17256.

[14] N.K. Mahanti, R. Pandiselvam, A. Kothakota, P. Ishwarya S., S.K. Chakraborty, M. Kumar, D. Cozzolino, Emerging non-destructive imaging techniques for fruit damage detection: Image processing and analysis, *Trends Food Sci. Technol.* 120 (2022) 418–438. https://doi.org/10.1016/J.TIFS.2021.12.021.

[15] M. Jiang, Y. Li, J. Song, Z. Wang, L. Zhang, L. Song, B. Bai, K. Tu, W. Lan, L. Pan, Study on Black Spot Disease Detection and Pathogenic Process Visualization on Winter Jujubes Using Hyperspectral Imaging System, *Foods* 12 (2023) 435. https://doi.org/10.3390/foods12030435.

[16] C. Ferrari, G. Foca, R. Calvini, A. Ulrici, Fast exploration and classification of large hyperspectral image datasets for early bruise detection on apples, *Chemometr. Intell. Lab.* 146 (2015) 108–119. https://doi.org/10.1016/j.chemolab.2015.05.016.

[17] J.M. Amigo, H. Babamoradi, S. Elcoroaristizabal, Hyperspectral image analysis. A tutorial, *Anal. Chim. Acta* 896 (2015) 34–51. https://doi.org/10.1016/j.aca.2015.09.030.

[18] R. Calvini, J.M. Amigo, A. Ulrici, Transferring results from NIR-hyperspectral to NIR-multispectral imaging systems: A filter-based simulation applied to the classification of Arabica and Robusta green coffee, *Anal. Chim. Acta* 967 (2017) 33–41. https://doi.org/10.1016/J.ACA.2017.03.011.

[19] A. Gowen, J.-L. Xu, A. Herrero-Langreo, Comparison of spectral selection methods in the development of classification models from visible near infrared hyperspectral imaging data, *J. spectr. imaging* (2019) a4. https://doi.org/10.1255/jsi.2019.a4.

[20] S.R. Delwiche, I. Baek, M.S. Kim, Does spatial region of interest (ROI) matter in multispectral and hyperspectral imaging of segmented wheat kernels?, *Biosyst. Eng.* 212 (2021) 106–114. https://doi.org/10.1016/j.biosystemseng.2021.10.003.

[21] D. Zhang, Md.M. Islam, G. Lu, A review on automatic image annotation techniques, *Pattern Recognition* 45 (2012) 346–362. https://doi.org/10.1016/j.patcog.2011.05.013.

[22] M. Sezgin, B. Sankur, Survey over image thresholding techniques and quantitative performance evaluation, *J. Electron. Imaging* 13 (2004) 146–168. https://doi.org/10.1117/1.1631315.

[23] M. Huang, J. Tang, B. Yang, Q. Zhu, Classification of maize seeds of different years based on hyperspectral imaging and model updating, *Comput. Electron. Agric.* 122 (2016) 139–145. https://doi.org/10.1016/j.compag.2016.01.029.

[24] K. Esbensen, P. Geladi, Strategy of multivariate image analysis (MIA), *Chemometr. Intell. Lab.* 7 (1989) 67–86. https://doi.org/10.1016/0169-7439(89)80112-1.

[25] J.M. Prats-Montalbán, A. De Juan, A. Ferrer, Multivariate image analysis: A review with applications, *Chemometr. Intell. Lab.* 107 (2011) 1–23. https://doi.org/10.1016/j.chemolab.2011.03.002.

[26] M. Vidal, J.M. Amigo, Pre-processing of hyperspectral images. Essential steps before image analysis, *Chemometr. Intell. Lab.* 117 (2012) 138–148. https://doi.org/10.1016/j.chemolab.2012.05.009.

[27] P.K. Sahoo, S. Soltani, A.K.C. Wong, A survey of thresholding techniques, *Comput. Graph. Image Process* 41 (1988) 233–260. https://doi.org/10.1016/0734-189X(88)90022-9.

[28] H.D. Cheng, X.H. Jiang, Y. Sun, J. Wang, Color image segmentation: advances and prospects, *Pattern Recognition* 34 (2001) 2259–2281. https://doi.org/10.1016/S0031-3203(00)00149-7.

[29] N. Otsu, A threshold selection method from gray-level histograms., *Automatica* 11 (1975) 23–27.

[30] S. Munera, A. Rodríguez-Ortega, N. Aleixos, S. Cubero, J. Gómez-Sanchis, J. Blasco, Detection of Invisible Damages in 'Rojo Brillante' Persimmon Fruit at Different Stages Using Hyperspectral Imaging and Chemometrics, *Foods* 10 (2021) 2170. https://doi.org/10.3390/foods10092170.

[31] X. Xu, S. Xu, L. Jin, E. Song, Characteristic analysis of Otsu threshold and its applications, *Pattern Recognition Letters* 32 (2011) 956–961. https://doi.org/10.1016/j.patrec.2011.01.021.

[32] T.Y. Goh, S.N. Basah, H. Yazid, M.J. Aziz Safar, F.S. Ahmad Saad, Performance analysis of image thresholding: Otsu technique, *Measurement* 114 (2018) 298–307. https://doi.org/10.1016/j.measurement.2017.09.052.

[33] J. Li, L. Chen, W. Huang, Detection of early bruises on peaches (Amygdalus persica L.) using hyperspectral imaging coupled with improved watershed segmentation algorithm, *Postharvest Biol. Technol.* 135 (2018) 104–113. https://doi.org/10.1016/j.postharvbio.2017.09.007.

[34] A. Siedliska, P. Baranowski, W. Mazurek, Classification models of bruise and cultivar detection on the basis of hyperspectral imaging data, *Comput. Electron. Agric.* 106 (2014) 66–74. https://doi.org/10.1016/j.compag.2014.05.012.

[35] H. Yin, B. Li, Y. Liu, F. Zhang, C. Su, A. Ou-yang, Detection of early bruises on loquat using hyperspectral imaging technology coupled with band ratio and improved Otsu method, *Spectrochim Acta A Mol Biomol. Spectrosc.* 283 (2022) 121775. https://doi.org/10.1016/j.saa.2022.121775.

[36] Y. Lu, R. Lu, Histogram-based automatic thresholding for bruise detection of apples by structured-illumination reflectance imaging, *Biosyst. Eng.* 160 (2017) 30–41. https://doi.org/10.1016/j.biosystemseng.2017.05.005.

[37] N. Vélez Rivera, J. Gómez-Sanchis, J. Chanona-Pérez, J.J. Carrasco, M. Millán-Giraldo, D. Lorente, S. Cubero, J. Blasco, Early detection of mechanical damage in mango using NIR hyperspectral images and machine learning, *Biosyst. Eng.* 122 (2014) 91–98. https://doi.org/10.1016/J.BIOSYSTEMSENG.2014.03.009.

[38] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval Partial Least-Squares Regression (i PLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy, *Applied Spectroscopy* 54 (2000) 413–419. https://doi.org/10.1366/0003702001949500.

[39] C. Ferrari, G. Foca, A. Ulrici, Handling large datasets of hyperspectral images: Reducing data size without loss of useful information, *Anal. Chim. Acta* 802 (2013) 29–39. https://doi.org/10.1016/j.aca.2013.10.009.

[40] 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.

[41] J.C. Bergh, S. V. Joseph, B.D. Short, M. Nita, T.C. Leskey, Effect of pre-harvest exposures to adult *Halyomorpha halys* (Hemiptera: Pentatomidae) on feeding injury to apple cultivars at harvest and during post-harvest cold storage, *Crop Protection* 124 (2019) 104872. https://doi.org/10.1016/J.CROPRO.2019.104872.

[42] H.R. El-Ramady, É. Domokos-Szabolcsy, N.A. Abdalla, H.S. Taha, M. Fári, Postharvest Management of Fruits and Vegetables Storage, in: E. Lichtfouse (Ed.), *Sustainable Agriculture Reviews*, Springer International Publishing, Cham, 2015: pp. 65–152. https://doi.org/10.1007/978-3-319-09132-7_2.

[43] FAO, Manual for the preparation and sale of fruits and vegetables: from fi eld to market, Food and Agriculture Organization of the United Nations, Rome, 2004.

[44] P. Gonzalez, J. Pichette, B. Vereecke, B. Masschelein, A. Lambrechts, L. Krasovitski, L. Bikov, An extremely compact and high-speed line-scan hyperspectral imager covering the SWIR range, in: N.K. Dhar, A.K. Dutta (Eds.), *Image Sensing Technologies: Materials, Devices, Systems, and Applications* V, SPIE, Orlando, United States, 2018: p. 19. https://doi.org/10.1117/12.2304918.

[45] R. Van Den Boomgaard, R. Van Balen, Methods for fast morphological image transforms using bitmapped binary images, *CVGIP: Graphical Models and Image Processing* 54 (1992) 252–258. https://doi.org/10.1016/1049-9652(92)90055-3.

[46] M. Hickey, C. King, The Cambridge Illustrated Glossary of Botanical Terms, Cambridge University Press, 2000.

[47] S. Kucheryavskiy, A new approach for discrimination of objects on hyperspectral images, *Chemometr. Intell. Lab.* 120 (2013) 126–135. https://doi.org/10.1016/j.chemolab.2012.11.009.

[48] L. Pieszczek, M. Daszykowski, Integrating hyperspectrograms with class modeling techniques for the construction of an effective expert system: Quality control of pharmaceutical tablets based on near-infrared hyperspectral imaging, *J. Pharm. Biomed. Anal.* 256 (2025) 116697. https://doi.org/10.1016/J.JPBA.2025.116697.

[49] P. Geladi, H.F. Grahn, Multivariate Image Analysis, in: R.A. Meyers (Ed.), Encyclopedia of Analytical Chemistry, 1st ed., Wiley, 2000. https://doi.org/10.1002/9780470027318.a8106.

[50] M. Cocchi, A. Biancolillo, F. Marini, Chemometric Methods for Classification and Feature Selection, in: *Comprehensive Analytical Chemistry*, Elsevier, 2018: pp. 265–299. https://doi.org/10.1016/bs.coac.2018.08.006.

[51] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, *Chemometr. Intell. Lab.* 174 (2018) 33–44. https://doi.org/10.1016/j.chemolab.2017.12.004.

[52] G. Marrubini, A. Papetti, E. Genorini, A. Ulrici, Determination of the Sugar Content in Commercial Plant Milks by Near Infrared Spectroscopy and Luff-Schoorl Total Glucose Titration, *Food Anal. Methods* 10 (2017) 1556–1567. https://doi.org/10.1007/s12161-016-0713-1.

[53] D.A. Burns, E.W. Ciurczak, Handbook of Near-Infrared Analysis, 3rd Edition, CRC Press, 2007. https://doi.org/10.1201/9781420007374.

[54] Jr. Workman Jerry, L. Weyer, Practical Guide to Interpretive Near-Infrared Spectroscopy, 2 ed., CRC Press, 2007. https://doi.org/10.1201/9781420018318.

[55] S. Gacnik, D. Rusjan, M. Mikulic-Petkovsek, Metabolic Response of Peach Fruit to Invasive Brown Marmorated Stink Bug (*Halyomorpha halys* Stål.)'s Infestation *Int. J. Mol. Sci.* 25 (2024) 606. https://doi.org/10.3390/ijms25010606.

[56] N.C. Weber, J. Razinger, J. Jakopič, V. Schmitzer, M. Hudina, A. Slatnar, R. Veberič, F. Štampar, T. Zamljen, Brown Marmorated Stink Bug (*Halyomorpha halys* Stål.) Attack Induces a Metabolic Response in Strawberry (Fragaria × ananassa Duch.) Fruit, *Horticulturae* 7 (2021) 561. https://doi.org/10.3390/horticulturae7120561.

[57] T. Zamljen, A. Medič, R. Veberič, M. Hudina, F. Štampar, A. Slatnar, Apple Fruit (Malus domestica Borkh.) Metabolic Response to Infestation by Invasive Brown Marmorated Stink Bug (*Halyomorpha halys* Stal.), *Horticulturae* 7 (2021) 212. https://doi.org/10.3390/horticulturae7080212.

[58] J.-M. Celton, D. Chagné, S.D. Tustin, S. Terakami, C. Nishitani, T. Yamamoto, S.E. Gardiner, Update on comparative genome mapping between Malus and Pyrus, BMC Res Notes 2 (2009) 182. https://doi.org/10.1186/1756-0500-2-182.

[59] K. Kubitzki, Flowering Plants. Dicotyledons: Celastrales, Oxalidales, Rosales, Cornales, Ericales, Springer Science & Business Media, 2013.

3.4. NIR Hyperspectral Imaging to identify damage caused by *Halyomorpha halys* on pears: development of classification models

What follows is the integral content of: Ferrari, V., Calvini, R., Menozzi, C., Costi, E., Offermans, P., Maistrello, L., Ulrici, A. NIR hyperspectral imaging to identify damage caused by *Halyomorpha halys* on pears: development of classification models, *submitted for publication*.

NIR Hyperspectral Imaging to identify damage caused by *Halyomorpha halys* on pears: development of classification models

Veronica Ferrari[1], Rosalba Calvini[1*], Camilla Menozzi[1], Elena Costi[1], Peter Offermans[2], Lara Maistrello[1], Alessandro Ulrici[1]

[1] *University of Modena and Reggio Emilia, Department of Life Sciences, Pad. Besta, Via Amendola, 2, 42122, Reggio Emilia, Italy*

[2] *IMEC OnePlanet, Bronland 10, Wageningen, the Netherlands*

**Corresponding author*

Abstract

Halyomorpha halys, the Brown Marmorated Stink Bug (BMSB), has become a major threat for pear orchards, causing severe economic losses throughout Southern Europe. Italy, one of the leading European pear producers, has experienced a dramatic reduction in yields of economically relevant cultivars over the past decade due to extreme climate events and outbreaks of pests and pathogens.

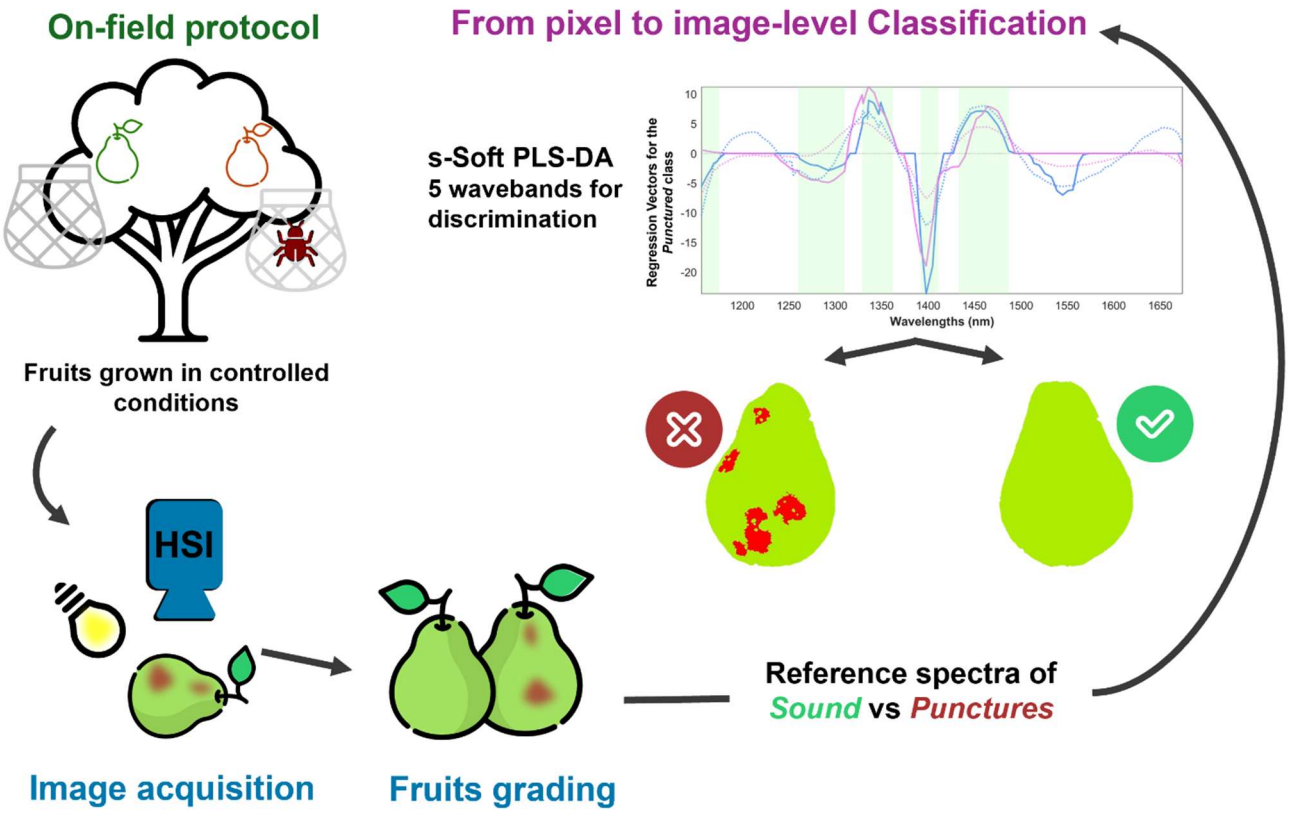
To address this issue, Near-Infrared Hyperspectral Imaging (NIR-HSI) is proposed as a non-destructive technique to automatically discard damaged fruits in post-harvest sorting lines. The study was performed on organic pears (*cv*. Abate Fétel and *cv.* Williams) harvested over two consecutive years.

Following an automated annotation procedure that identified punctured and sound areas of the fruits, representative spectra of both *Punctured* and *Sound* classes were extracted to build pixel-level classification models. Classification models were developed using Soft Partial Least Squares Discriminant Analysis (Soft PLS-DA), a hybrid approach that combines the advantages of both discriminant and class modelling techniques, and its sparse variant (s-Soft PLS-DA), which incorporates variable selection to achieve a parsimonious solution more suitable for online post-harvest sorting systems.

An additional decision criterion was optimized starting from the prediction images belonging to the training set, enabling image-level classification. Besides being advantageous for automated sorting

systems, image-level classification allowed a comprehensive assessment of models' classification performances towards modelled and unmodelled (i.e., damages not ascribable to BMSB) classes.

Graphical abstract



Keywords: Hyperspectral imaging, damage detection, post-harvest sorting, Brown marmorated stink bug, Multivariate classification, Soft PLS-DA, Sparse methods.

1. Introduction

In recent decades, increasing anthropogenic activities eased the spread of invasive insect pests capable of altering agroecosystems and compromising agri-food production, leading to substantial economic losses [1,2]. Among these, *Halyomorpha halys*, commonly known as the Brown Marmorated Stink Bug (BMSB), is particularly noteworthy. Native to East Asia, it rapidly spread to Europe and North America, becoming a significant agricultural threat. This highly polyphagous pest damages a wide range of crops, including fruits, vegetables, and ornamental plants, with considerable economic impact [3–5].

BMSB feeds using its piercing-sucking mouthparts, causing deformities, discoloration, surface lesions, and internal tissue damage in fruits [6–8]. In Southern Europe, favourable climatic conditions and high crop density have supported large pest populations [9]. Northern Italy has been particularly affected, especially its pear orchards, where economic losses were estimated at €590 million [10].

Italy has historically been the leading European country for pear production, with an average annual yield of approximately 707,000 tons, ranking third worldwide. However, frequent frosts at fruit set and drought during summer along with an increased activity of pathogens and pests such as BMSB have reduced the production to approximately 343,000 tons over the past three years, with only 22,000 hectares of cultivated land remaining, and constantly decreasing [11,12]. In Northern Italy, the production is predominantly concentrated in Lombardy, Emilia-Romagna, and Veneto, where several cultivars of economic relevance are grown, some regulated by the European Protected Geographical Indication (PGI) designation [13]. In particular, Abate Fétel and Williams are among the most widely cultivated cultivars in the Emilia-Romagna region, accounting for approximately 37% and 27% of national production in 2019, respectively [14]. The importance of these cultivars is also characterized by their strong link with the territory, guaranteed by the "*Pere dell'Emilia-Romagna"* PGI designation [15].

Damage caused by BMSB during the early stages of fruit development is easily visible, since it often determines fruit deformation. Conversely, BMSB punctures occurring in the late stages typically result in internal suberification and necrosis of the pulp, producing brown, corky tissue beneath an externally undamaged skin, and are therefore not detectable through visual inspection [6]. Moreover, the feeding activity of BMSB has been associated with increased susceptibility to pathogens including yeast, *Monilinia spp*. and other fungi, which may lead to fruit rot [5,7,16,17]. As a result, internal defects caused by BMSB late damages severely reduce fruit quality and commercial value, while remaining difficult to detect through visual inspection before fruit consumption or deterioration during post-harvest [5].

Currently, post-harvest sorting practices mainly rely on visual inspection or computer vision systems, which are unable to detect the internal injuries caused by BMSB late feeding as they are not visible at the naked eye. At the same time, Near-InfraRed (NIR) spectroscopy is widely used to assess fruit quality and safety thanks to the possibility of characterising fruits chemical properties [18,19]. However, traditional spectroscopic techniques do not provide information related to the spatial distributions of chemical compounds, therefore they are not useful in the detection of localized defects such as punctures [20].

In this context, NIR spectral imaging represents a viable alternative to conventional computer vision systems operating in the visible range, enabling rapid, non-destructive, and efficient characterization of the chemical and physical properties of food products. Hyperspectral imaging (HSI) combines the strengths of spectroscopy and imaging, allowing the visualization of spatial distribution of the chemical features across a sample's surface [21,22]. This makes NIR-HSI especially suitable for detecting defects in heterogeneous food matrices, as it allows both the assessment of chemical composition and its spatial differences on the sample surface [20,23–31]. However, HSI data-richness is a double-edged sword: on the one hand, it allows a detailed sample characterization while on the other hand, it poses challenges in data handling, storage, and analysis [32,33]. In this context, Multivariate Image Analysis plays a crucial role in addressing the curse of dimensionality and extracting relevant information for the development of robust classification or calibration models able to predict qualitative and quantitative properties of food products.

Within this frame, we evaluated NIR-HyperSpectral Imaging (NIR-HSI) as a possible post-harvest sorting system method for the early detection of BMSB under-peel damage on pears, which is not detectable using cameras working in the visible range. For the practical application of NIR-HSI systems, it is necessary to develop supervised classification models able to process the hyperspectral images and provide a class assignment for each pixel (i.e., pixel-level classification). To this aim, the identification in the hyperspectral images of Regions of Interests (ROIs) of both sound and damaged areas is a crucial aspect affecting the effectiveness and robustness of the classification models [34,35]. This procedure is often referred to as image annotation.

In a previous study reported as **Section 3.3** of this thesis [36], we developed an innovative approach for the automated and objective annotation of BMSB punctured areas. The proposed method involved image dimensionality reduction and image-level classification coupled with spatial features selection. In the present study, starting from annotated ROIs of punctured and sound areas of the fruits, representative spectra belonging to both classes were extracted to build pixel-level classification models. To achieve a parsimonious solution suitable for the implementation of online post-harvest sorting systems, variable selection methods were applied to retrieve only the most relevant spectral

regions able to distinguish sound areas of the fruits from those affected by BMSB punctures. Indeed, multispectral cameras are more suitable to be used in practical implementations of post-harvest sorting systems as they are faster, cheaper and their optical components are highly robust [37,38].
For the calculation of classification models, the soft discriminant algorithm Soft Partial Least Squares Discriminant Analysis (Soft-PLS-DA) was considered since it allows to reject spectra that do not belong to any of the modelled classes, thus offering a more robust and flexible solution in sorting systems. Moreover, sparse-based variable selection was coupled with Soft-PLS-DA (s-Soft PLS-DA) algorithm to select relevant spectral regions able to discriminate between punctured and sound areas. Finally, Soft PLS-DA and s-Soft PLS-DA [39] classification models were applied to all the acquired images: from the resulting prediction images belonging to the training set a decision criterion was optimized to provide a classification at the image-level. The criterion was applied to all the prediction images, enabling a global evaluation of models' classification performances towards modelled (i.e., punctured and sound areas) and unmodelled (i.e., other damages) classes.

2. Materials and methods

2.1. Experimental protocol on-field

Pear samples belonging to Abate Fétel and Williams cultivars were harvested in an organic orchard located in Carpi (Modena, Italy) during summer 2022 and summer 2023.
In both years, 40 tree branches belonging to different plants located in diverse areas of the orchard were considered. At fruit set, the branches were covered with cylindric inclusion cages made to protect the fruits from uncontrolled biotic and abiotic adversities [6,40].
The same day of commercial harvest, 20 inclusion cages were selected and 3 BMSB specimens were manually placed inside each one of the selected cages. In this manner, about half of the fruits were exposed to BMSB feeding for about 3 – 5 days while the remainder fruits were considered as control samples. Following this procedure, a total of 156 fruits and 160 fruits were collected for *cv.* Abate Fétel and *cv.* Williams, respectively.
The reader is refereed to [36], reported integrally in **Section 3.3** of ***Chapter 3***, for further insights about the experimental protocol followed on field.

2.2. Image acquisition and elaboration

Image acquisition was performed at 8 subsequent times to monitor the evolution of fruit damage over time, from harvest day (T1) to five weeks later (T2-T8), following the schedule reported in **Table 1**.

	Williams		Abate Fétel	
	2022	**2023**	**2022**	**2023**
BMSB exposure	August 7th – August 9th	August 4th – August 8th	September 2nd – September 6th	September 1st – September 5th
T1	August 9th	August 16th	September 13th	September 12th
T2	August 11th	August 23rd	September 20th	September 19th
T3	August 16th	August 29th	September 27th	September 16th
T4	August 22nd	September 5th	October 4th	October 3th
T5	August 29th	September 11th	October 11th	October 10th
T6	September 5th	September 14th	October 13th	October 12th
T7	September 9th	September 18th	October 17th	October 17th
T8	September 12th	September 21st	October 20th	October 19th

Table 1 *Schedule for BMSB exposure of the fruits in the orchard and image acquisition times (T1 – T8); T1 corresponds to the harvest day.*

Between the different acquisition sessions, the fruits were stored at refrigerated temperatures of 0 – 2°C to simulate post-harvest storage conditions [41,42].

The hyperspectral camera used for the acquisitions was a line-scanning system (SnapScan SWIR, IMEC One Planet, The Netherlands) working in the 1156 – 1674 nm spectral range with 5 nm spectral resolution (100 channels) [43], as described in detail in [36].

To cover the whole pear surface during acquisition, the investigated fruits were divided into four sections (A – D) and each section was imaged, obtaining four hyperspectral images for each fruit.

As previously mentioned, damage due to BMSB punctures is not visible to the naked eye on the intact fruits' surface as it affects the fruit pulp. Therefore, to verify the actual presence of damages it was necessary to peel the fruits after image acquisition.

For this reason and considering the high number of collected fruits, the pears were randomly divided into two groups: the first group (Group 1) composed of fruits to be imaged at each acquisition time and peeled only at T8, and a second group of fruits (Group 2) to be imaged at specific acquisition times and peeled immediately after. Over the two harvest years, 3644 hyperspectral images were acquired, corresponding to 1680 images of Abate Fétel fruits and 1964 images of Williams fruits. The storage space of the whole dataset was equal to 3.1 TB.

Due to adverse climate conditions and the activity of other biotic agents, neither all the control fruits were sound, nor the fruits exposed to BMSB actually showed damages. Therefore, an expert visual inspection was necessary to divide the acquired images into the following categories:

- images of sound sections of control fruits (S-C), corresponding to the images of fruit sections of control samples without any sign of damage;
- images of damaged sections of control fruits (D-C), corresponding to the images of fruit sections of control samples showing damage;
- images of damaged sections of exposed fruits (D-E), corresponding to the images of fruit sections of samples exposed to BMSB with damage;
- images of sound sections of exposed fruits (S-E), corresponding to the images of fruit sections of samples exposed to BMSB but not showing any kind of damage.

Table 2 summarises the number of images belonging to the different categories for both Abate Fétel and Williams cultivars.

Based on visual inspection, the fruits showing damage were additionally classified into three categories based on damage type:

- Type 1: mild damage of unknown origin (e.g., superficial signs of the fruit pulp);
- Type 2: damage ascribable to BMSB punctures (e.g., suberifications);
- Type 3: severe damage not caused by BMSB punctures (e.g., moulds or other diseases).

Table 3 reports the number of images belonging to D-E class with Type 2 damage, which will be referred to as P-E images (i.e., images of punctured sections of fruits exposed to BMSB).

Preliminary image elaboration steps were necessary to remove the background and mask the fruit area. These steps involved thresholding procedures based on Principal Component Analysis (PCA) and morphological erosion, as further detailed in [36].

		Abate Fétel			Williams		
		# of images			# of images		
		2022	2023	Total	2022	2023	Total
Control fruits	Sound sections (S-C)	399	250	649	242	340	582
	Damaged sections (D-C)	69	74	139	210	152	362
Exposed fruits	Sound sections (S-E)	80	180	260	174	157	331
	Damaged sections (D-E)	444	188	632	326	363	689

Table 2 *Number of images of Abate Fétel and Williams cultivars acquired in this study. The count of the images is performed according to the following categories: sound sections of control fruits (S-C), damaged sections of control fruits (D-C), sound sections of exposed fruits (S-E) and damaged sections of exposed fruits (D-E).*

		Abate Fétel			Williams		
		# of images			# of images		
		2022	2023	Total	2022	2023	Total
Type 1	Mild damages of unknown origin	192	55	247	229	34	263
Type 2	Damages ascribable to BSBM punctures	**221**	**101**	**322**	**54**	**298**	**352**
Type 3	Severe damages not ascribable to BMSB	31	32	63	43	31	74

Table 3. *Subdivision of D-E images according to damage type based on visual inspection of the corresponding peeled fruit samples.*

2.3. Development of pixel level classification models

2.3.1. Annotation of ROIs belonging to punctured areas

The development of pixel-level classification models able to discriminate between pixel spectra of sound and punctured areas started with the construction of a library of reference spectra representative of both classes. This task requires the availability of the ground truth, i.e., the exact location of the punctured regions on P-E fruits, quite challenging to achieve in this scenario.

Indeed, to mimic a real situation, the pears were exposed to BMSB directly in the field, therefore it was possible to verify the presence of the punctures and locate them only after peeling the fruits. In addition, standard thresholding approaches for ROIs selection, like e.g., PCA score values of the hyperspectral images, were not easily applicable due to the slight differences between sound and damaged areas, the irregular shape of the punctures, and the high number of acquired images.

To solve this issue, in a previous study we proposed an innovative approach for image annotation that allows the direct comparison of a high number of images altogether and the automated identification of ROIs related to target spatial features [36]. The proposed algorithm follows several subsequent steps. Firstly, data dimensionality reduction (DDR) is performed using the Common Space Hyperspectrograms (CSH) approach, which converts each image into a signal, named hyperspectrogram, built by merging in sequence the frequency distribution curves of quantities (i.e., score, Hotelling T^2 and Q residuals vectors) obtained from a global PCA model common to all the considered images [45]. In particular, the images belonging to the two pear cultivars were analysed separately due to the considerably different physical characteristics (i.e., shape, dimensions, peel rustiness). Therefore, two distinct CSH data matrices were obtained for Abate Fétel and Williams samples, respectively. Then, interval PLS-DA (iPLS-DA) models were calculated considering different interval sizes at the image-level to discriminate images of sound and punctured fruits, thus selecting the hyperspectrograms variables more relevant for classification. Finally, the variables selected most frequently by the different iPLS-DA models were reconstructed back into the image domain.

Since the selected hyperspectrograms variables correspond to intervals of score values more likely related to P-E class, visualising them back into the original image domain allowed to select the pixels related to punctured areas, thus accounting the selected areas as punctures' annotation. A more detailed description of the image annotation algorithm can be found in [36].

For each cultivar, the annotation models were developed between 2022 and 2023 harvest campaigns. In particular the S-C and P-E images of 2022 samples were divided into training (TR) set and test (TS) set images, as reported in **Table 4**. The subdivision of the images into TR and TS was done by

keeping into the same set the images related to the same fruit sample, also across different acquisition times. The TR set images were used to develop the models and the TS set images were used for validation. Concerning 2022 harvest campaign, 140 out of 221 P-E images of *cv.* Abate Fétel samples and 32 out of 54 P-E images of *cv.* Williams samples were correctly annotated.

Then, the annotation procedure was applied to the P-E images acquired in 2023, leading to 70 out of 101 P-E images of *cv.* Abate Fétel samples and 240 out of 298 P-E images of *cv.* Williams samples being successfully annotated. Further details about the outcomes of image annotation are reported in **Table 4**.

		Abate Fétel			**Williams**		
	# images	S-C	P-E Annotated	P-E Total	S-C	P-E Annotated	P-E Total
2022	TR	246	82	136	112	22	39
	TS	153	58	85	130	10	15
	Total	**399**	**140**	**221**	**242**	**32**	**54**
2023	TR	166	47	-	228	159	-
	TS	84	23	-	112	81	-
	Total	**250**	**70**	**101**	**340**	**240**	**298**

Table 4 *S-C and P-E images collected in 2022 and 2023, subdivided in TR and TS for cv. Abate Fétel and cv. Williams. For P-E images, both the total and annotated P-E images counts are shown. Concerning 2022 images, the subdivision into TR and TS images was used to develop the image annotation algorithm and to calculate the classification models in this study. Regarding 2023 images, the annotation algorithm developed using the images acquired in 2022 was applied to all P-E images, and only those successfully annotated were then used for further development of the classification models. For this reason, for 2023 the TR/TS partition only refers to the annotated P-E images.*

2.3.2. Spectra selection and dataset structure

After annotating the ROIs of punctured regions as described in **Section 2.3.1**, representative spectra of both punctured and sound areas were randomly selected from S-C and annotated P-E hyperspectral images of both cultivars. The spectra selection phase is crucial in determining the representativeness of spectral signatures for punctured and sound areas, which holds deeply the robustness and reliability of the classification models.

The images belonging to other categories (i.e., D-C, S-E, D-E, not annotated P-E) were not considered in the development of pixel-level classification models, but they were used in a second moment to evaluate the robustness of the models (*see* **Section 2.4**).

The images were initially divided into TR and TS images. For the 2022 samples of both cultivars, the same TR/TS partition considered for the annotation procedure was applied. To increase the number and representativeness of the samples, also the 2023 images were divided into TR and TS images, as reported in **Table 4**. Also in this case, images belonging to the same fruit sample, even when acquired at different acquisition times, were kept in the same set.

			S-C			P-E annotated		
			# images	**# fruits**	**# spectra**	**# images**	**# fruits**	**# spectra**
Abate Fètel	TR	2022	246	28	6218	82	20	6274
		2023	166	9	1954	47	5	1576
		Total	**412**	**37**	**8172**	**129**	**25**	**7850**
	TS	2022	153	17	5314	58	14	6000
		2023	84	7	2200	23	3	1158
		Total	**237**	**24**	**7514**	**81**	**17**	**7158**
Williams	TR	2022	112	10	2990	22	5	2542
		2023	228	17	5117	159	9	5571
		Total	**340**	**27**	**8107**	**181**	**14**	**8113**
	TS	2022	130	25	5348	10	4	2263
		2023	112	12	2600	81	9	5114
		Total	**240**	**37**	**7948**	**91**	**13**	**7377**

Table 5 *S-C and P-E images, along with the corresponding number fruits and the number of selected spectra considered for the creation of the TR and TS set, subdivided by harvest year and cultivar*

The spectra signatures for the *Sound* class were selected exclusively from hyperspectral images of S-C fruit sections, while the representative spectra of the *Punctured* class were extracted only from ROIs of punctured areas in the annotated P-E images, generated using the annotation procedure described in **Section 2.3.1**.

For each pear cultivar, a global dataset comprising approximately 30,000 representative spectra of *Sound* and *Punctured* classes was obtained. In each dataset, about half of the spectra originated from TR images and were used as the TR set for model development, while the remainder spectra belonged to TS images and were employed for external validation.

As mentioned in **Section 2.2**, some fruits were imaged at all the eight acquisitions times (Group 1), while the others were acquired only at specific times (Group 2). Consequently, Group 1 fruits had a much higher number of images compared to Group 2 fruits. Therefore, extracting a fixed number of randomly selected spectra from each S-C and P-E image would have implied an over representation of Group 1 fruits and an under representation of Group 2 fruits. To address this issue, we decided to fix the number of spectra to consider for each S-C or P-E fruit, and then we calculated accordingly the number of spectra to extract from each hyperspectral image.

In more detail, for each S-C fruit about 260 spectra were randomly selected from the corresponding hyperspectral images, while for each P-E fruit about 460 spectra were randomly selected from the corresponding ROIs in the images.

When only one image was available per fruit, 260 and 460 spectra were selected for S-C and P-E images, respectively. Conversely, for fruits with multiple images (e.g., those belonging to Group 1), the number of spectra to extract from each image was determined by dividing the total number of spectra per fruit by the number of images of the fruit.

It has to be considered that, in some cases, the size of the ROIs annotated on P-E fruits was smaller than 460 spectra; in such cases, all the spectra belonging to the annotated ROI were selected.

More detailed information about the subdivision of fruits, images and spectra in TR and TS for each cultivar can be found in **Table 5**.

Figure 1 reports the average spectrum of the selected pixel spectra related to the *Punctured* and *Sound* classes belonging to the TR for both cultivars.

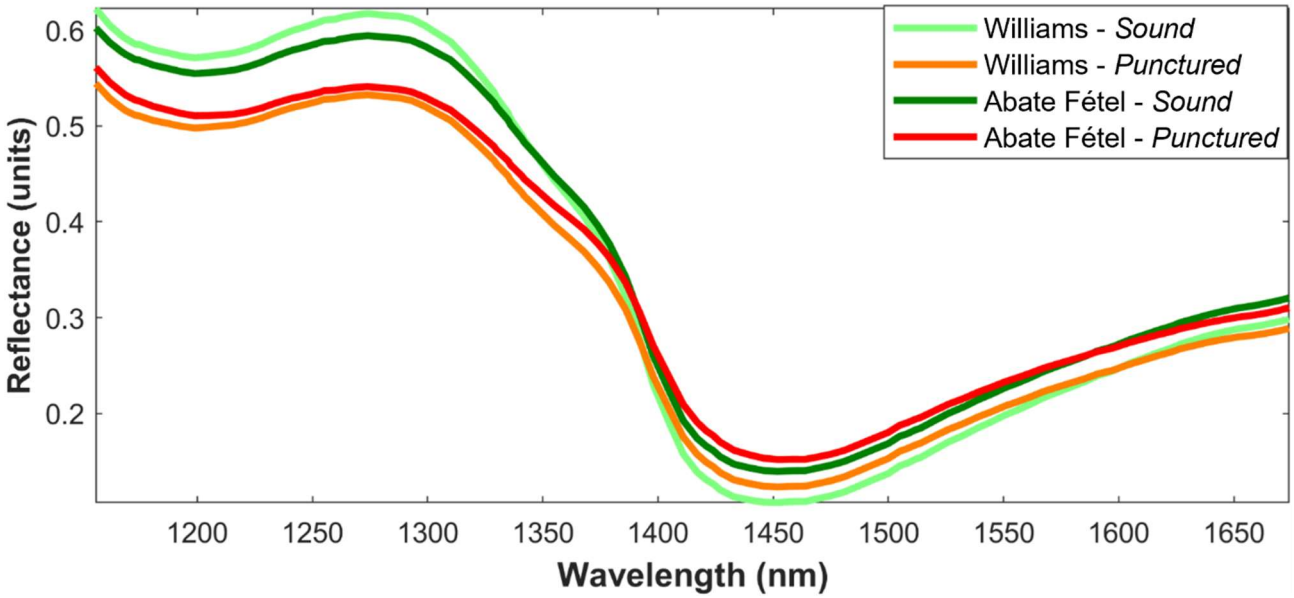


Figure 1 *Average raw spectra representative for the Punctured and Sound classes belonging to the TR sets related to cv Abate Fétel and cv. Williams.*

2.3.3. Model calculation and validation

The datasets of representative spectra of *Sound* and *Punctured* classes were used to develop pixel-level classification models using Soft PLS-DA algorithm. Coherently with the annotation procedure described in **Section 2.3.1**, separate classification models were developed for *cv.* Abate Fètel and *cv.* Williams, as they show considerably different physical characteristics.

Soft PLS-DA is a soft discriminant algorithm that combines the advantages of discriminant analysis and class modelling approaches. Its configuration allows for increased flexibility and robustness in practical applications: indeed, by applying additional constraints for class assignment, it effectively identifies possible outliers. In this manner, Soft PLS-DA overcomes the limitations of PLS-DA in handling new objects not belonging to the target classes, maximizing at the same time the discrimination between the classes of interest [39,46,47].

In Soft PLS-DA, model computation is the same as PLS-DA, but a new sample is assigned to a defined class according to the following criteria:

- Q residuals values falling within the 99.9 % confidence limit of the model. This limit was set wide enough to consider as much as possible within classes variability, but allowing at the same time to exclude samples with a very poor fit to the model;
- y predicted values falling within an acceptability range for the considered class. The lower limit of this range is defined by the PLS-DA threshold value for the considered class, while the upper limit allows the rejection of objects located at the extremes of the Gaussian probability density function;
- the samples must be unambiguously assigned to one class only: this aspect is rather useful for multiclass classification issues since it solves ambiguities in class attribution.

The samples that do not match all the criteria defined by the Soft PLS-DA decision rules are not assigned to any class and automatically labelled as *Not Assigned* (NA). The reader is referred to [39] for a detailed description of Soft PLS-DA algorithm.

The classification models were built considering the whole spectral range and, secondly, taking into account the most relevant wavelengths for classification selected by sparse-based variable selection with sparse Soft PLS-DA (s-Soft PLS-DA).

s-Soft PLS-DA couples Soft PLS-DA algorithm described above with Least Absolute Shrinkage and Selection Operator (LASSO) penalisation, which constrains the sum of the absolute values of the regression vector to be lower than a tuning parameter. In this manner, the coefficients of irrelevant variables are shrunk to zero, effectively enabling variable selection [48–52].

To obtain a good sparse solution, i.e., many regression coefficients shrunk to zero, it is necessary to choose the proper number of latent variables (LVs) and the level of sparsity, i.e., the number of spectral variables to be selected. To this aim, grid search was performed testing all the possible combinations between a number of LVs ranging from 1 to 10 and a number of spectral variables selected for each LV ranging from 5 to 100, with a step equal to 5.

Both Soft PLS-DA and s-Soft PLS-DA classification models were calculated considering the same preprocessing methods leading to the best results in the annotation procedure (**Section 2.3.1**), i.e.,

Multiplicative Scatter Correction, MSC (median spectrum as reference) and mean center for *cv.* Abate Fétel, and linear-detrend and mean center for *cv.* Williams.

The classification models were optimized using a custom cross-validation scheme, obtained by splitting the TR spectra based on pear sample into six deletion groups. For both cultivars, the optimal Soft PLS-DA and s-Soft PLS-DA models were selected as those leading to the highest Non-Error Rate (NER) values. Afterwards, the classification performances were assessed though external validation using the TS set of spectra.

The classification performances of the models were evaluated in calibration (CAL), cross-validation (CV) and prediction (PRED) of the external TS set by calculating the statistical parameters sensitivity (SENS), specificity (SPEC), efficiency (EFF), and Non-Error Rate (NER) [53].

2.4. From pixel-level classification models to image-level predictions

For each cultivar, the best performing Soft PLS-DA and s-Soft PLS-DA models were applied to the original hyperspectral images obtaining the corresponding prediction images, i.e., images where each pixel is assigned to *Sound* or *Punctured* class, or not assigned to any class (NA), according to the classification outcome of the corresponding spectrum. This allows to visualize the spatial distribution of the predicted classes over the fruit surface and have a qualitative assessment of model performances.

To simulate the application of the proposed approach to a fruit sorting line, it is necessary to move from the pixel-level prediction images to an image-level classification by classifying each image fruit as *Sound* or *Punctured*.

For the sake of clarity, from here onwards we will use the term *object* to refer to any cluster of pixels in a prediction image whose spectra are classified as belonging to the *Punctured* class by the selected Soft PLS-DA or s-Soft PLS-DA model.

For image-level classification, the prediction images were further elaborated by implementing an additional classification rule based on the size of the objects predicted as punctured areas. More in detail, based on the comparison between the size of objects wrongly predicted as punctures and the size of the objects correctly assigned to punctured regions, a threshold value of 105 contiguous pixels was set as the minimum object size for punctures. This threshold value was chosen since it represents the best compromise in terms of SENS, SPEC and EFF values for the TR images belonging to both cultivars.

Consequently, objects predicted as *Punctured* smaller than the established threshold were removed from the prediction images.

Based on this rule, images containing at least one object with more than 105 pixels predicted as *Punctured* were automatically categorized as "*Punctured images*", while images with no punctured objects or punctured objects smaller than 105 pixels were assigned to the "*Sound images*" category. The image-level classification performances were quantitatively assessed considering the percentages of S-C images predicted as *Sound* and P-E annotated images predicted as *Punctured*.
Moreover, the same classification rule was applied to the prediction images of all the other fruit images to evaluate the behaviour of the models towards unknown damage, i.e., damage not caused by BMSB. The resulting prediction images were used to visualize the prediction performances directly on the original image domain in order to perform a global evaluation of the classification performances.

2.5. Software

Background removal and spectra extraction were performed on each image in an automated manner using routines written in MATLAB environment (R2020a, The MathWorks, USA), based on the Image Processing Toolbox (v. 11.1) and the PLS_Toolbox (v. 8.8.1, Eigenvector Research Inc., USA). Soft PLS-DA models were calculated using the MATLAB function freely downloadable from https://www.chimslab.unimore.it/downloads/. s-Soft PLS-DA models and the final image-level classification step were implemented using *ad hoc* routines written in MATLAB environment.

3. Results and Discussion

3.1. Classification at pixel-level

Table 6 reports the results obtained in CAL, CV and PRED of the external TS of representative spectra for the Soft PLS-DA and s-Soft PLS-DA models. For each model, the classification performances were evaluated considering SENS, SPEC, EFF, NA and NER expressed as percentage values for both *Punctured* and *Sound* classes.
Overall promising classification results were obtained in CV and PRED of the external TS set considering both cultivars. Starting from *cv*. Abate Fétel, the best performances in terms of NER were obtained retaining 6 LVs, both considering the entire spectrum or only the wavebands most relevant for discrimination purposes selected by sparse based methods. In the case of *cv*. Williams, the best performances were obtained by choosing 3 LVs for the calculation of Soft PLS-DA and 5 LVs for the calculation of the sparse model.

Despite the objective of the work is to correctly detect *Punctured* areas, the least possible amount of *Sound* areas must be misclassified to avoid waste and economical losses. In this context, NER could be a suitable indicator to assess the overall ability of the model to correctly assign both *Punctured* and *Sound* spectra to the class they belong to. In prediction of the external TS set (PRED), the models achieving the best results for both cultivars led to NER values greater than 90.7% at the pixel-level, which is promising due to the complexity of the problem at hand.

The number of spectra rejected as NA is mostly higher for *cv*. Abate Fétel, in particular for the Soft PLS-DA model. Indeed, for *cv*. Abate Fétel, the percentage of spectra predicted as NA decrease for the s-Soft PLS-DA model for both classes investigated. Generally, s-Soft PLS-DA leads to slightly better results in PRED, reaching a NER of 91.5% at pixel-level. Conversely, considering *cv.* Williams, an increase of the percentage of spectra predicted as NA is shown for the s-Soft PLS-DA model. Although s-Soft PLS-DA exhibited a minor reduction in PRED performance, especially for the *Sound* class, it still represents a valid solution, achieving a NER of 91.2% while considering only half of the original spectral variables.

Generally, it has to be highlighted that s-Soft PLS-DA models led to classification results comparable to those of Soft PLS-DA, but proposing a more parsimonious solution in terms of considered spectral variables.

			Abate Fétel			Williams		
			MSC (median)+ mc			Linear-detrend + mc		
			CAL	CV	PRED	CAL	CV	PRED
Soft PLSDA	**LVs**		6			3		
	Spectral variables		100			100		
	SENS (%)	Punctured	91.9	90.4	**93.3**	93.8	92.1	**91.8**
		Sound	94.2	93.8	**88.0**	95.2	95.6	**97.2**
	SPEC (%)	Punctured	96.8	96.5	**90.8**	95.7	96.1	**97.4**
		Sound	94.5	93.7	**95.6**	94.1	92.9	**91.8**
	EFF (%)	Punctured	94.3	93.4	**92.0**	94.7	94.1	**94.5**
		Sound	94.4	93.7	**91.7**	94.6	94.2	**94.5**
	NA (%)	Punctured	2.7	3.2	**2.1**	0.3	0.8	**0.1**
		Sound	2.6	2.7	**2.6**	0.5	0.5	**0.2**
	NER		93.0	92.1	**90.7**	94.5	93.8	**94.5**
s-Soft PLSDA	**LVs**		6			5		
	Spectral variables		**51**			**54**		
	SENS (%)	Punctured	90.5	89.1	**91.8**	93.0	92.8	**93.0**
		Sound	95.8	95.4	**91.2**	94.6	95.7	**89.3**
	SPEC (%)	Punctured	96.3	96.1	**91.9**	95.4	97.1	**91.6**
		Sound	92.4	91.6	**93.4**	94.3	95.2	**94.2**
	EFF (%)	Punctured	93.4	92.6	**91.9**	94.2	94.9	**92.3**
		Sound	94.1	93.5	**92.3**	94.5	95.5	**91.7**
	NA (%)	Punctured	1.9	2.5	**1.3**	1.3	2.4	**1.1**
		Sound	0.5	0.7	**0.8**	0.8	1.4	**2.3**
	NER (%)		93.1	92.3	**91.5**	93.8	94.3	**91.2**

Table 6 *Pixel-level classification results obtained by applying Soft PLS-DA and s-Soft PLS-DA in calibration (CAL), cross-validation (CV) and prediction (PRED) of the TS set. The results, expressed in terms of percentages of SENS, SPEC, EFF, and NA, are reported for both Punctured and Sound classes while NER provides a global evaluation of the classification performances.*

Figure 2 reports the regression vectors for the *Punctured* class of the models calculated considering the whole spectrum (dotted line) or the sparse solution (solid line), enabling the interpretation of the spectral regions that differentiate *Punctured* from *Sound* areas of the fruits. Notably, the spectral variables selected by the two s-Soft PLS-DA models, derived from the spectra of the two pear cultivars, are largely consistent. The spectral regions selected by both sparse models are highlighted in green colour in **Figure 2**.

The main spectral regions with high absolute values for both varieties fall into the 1150−1180 nm (C=O, C−H second overtone), in the 1260−1310 nm (C−H second overtone), in the 1330−1370 nm ($C-H_3$ combination band), in the 1390-1420 nm ($C-H_2$ combination band, aromatic C−H combination band and O−H stretch first overtone), and in the 1430−1490 nm (O−H first overtone, C=O stretch third overtone and N−H stretch first overtone) spectral regions. Moreover, the region selected only for *cv*. Abate Fétel falling in the interval at 1510-1570 nm may be related to starch content, which is gradually broken down into soluble sugars during fruit ripening. Conversely, for *cv*. Williams there are not actual additional spectral regions considered, just isolated wavebands that slightly expand the intervals already selected [54–56].

The spectral regions falling in the intervals at 1330–1370 nm and at 1430–1490 nm can be associated with a higher absorption of bonded water, carbohydrates and fibres which characterize the *Sound* areas of the fruits. Conversely, the spectral bands falling in the intervals at 1150−1180 nm, 1260−1310 nm and 1390-1420 nm may refer to a higher absorption of free water, sugars and phenolic compounds in *Punctured* areas. Interestingly, the spectral regions selected seem coherent with the most informative spectral regions identified in the previous study related to the annotation of BMSB punctures (*see* **Figure 9** reported in **Section 3.3** of ***Chapter 3***, [36]).

Several studies reported changes in the sugars and polyphenols in fruits punctured by BMSB as a metabolic response [57–60]. Interestingly, Zamljen and colleagues [60] reported that the metabolic response of apples fruits (*Malus* spp.) towards BMSB punctures involves an increase in sugars and polyphenols (i.e., hydroxycinnamic acids, flavanols) content only in the damaged areas. To our knowledge, similar studies have not been performed on pears. However, both apple (*Malus* spp.) and pear (*Pyrus* spp.) fruits may exhibit a similar metabolic behaviour since they belong to the *Maloideae* sub-family [61,62].

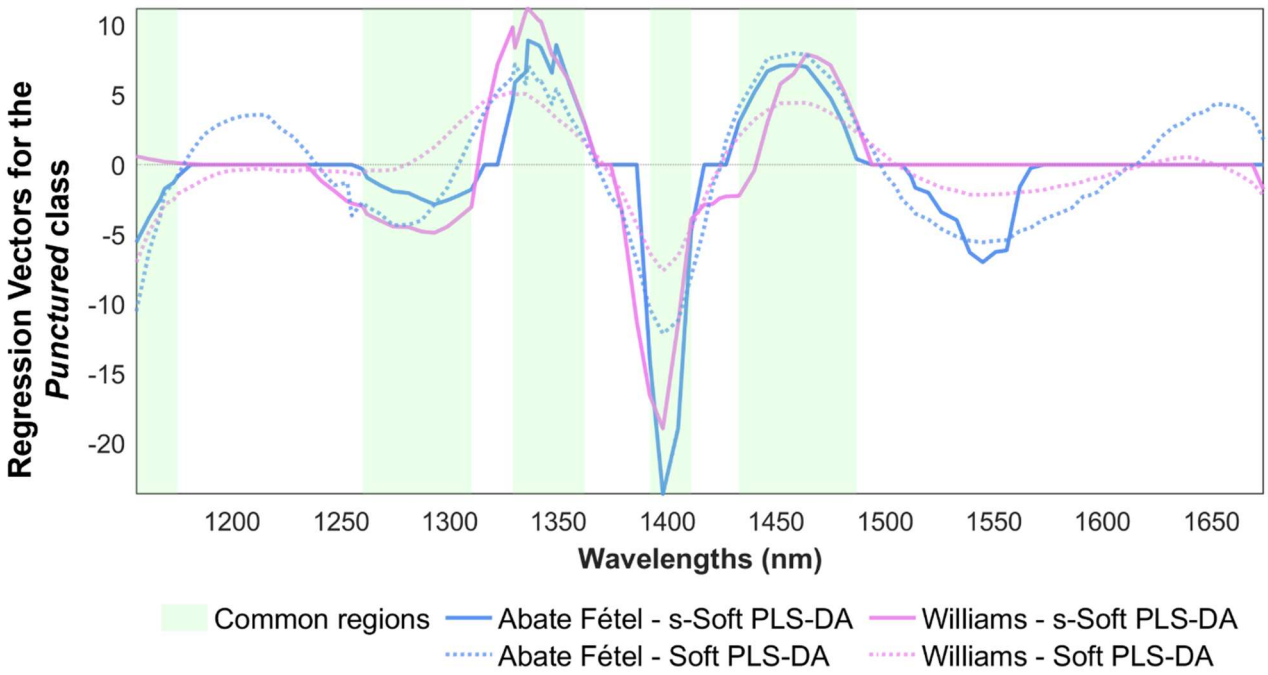


Figure 2 *Regression vectors for cv. Abate Fétel and cv. Williams: the dotted lines refer to the Soft PLS-DA regression vectors while the solid lines refer to the regression vectors obtained with sparse based variable selection. The areas highlited in green refer to common spectral regions selected for both cultivars.*

3.2. Classification at image-level

Table 7 reports the results obtained by applying Soft PLS-DA and s-Soft PLS-DA models to the TR (CAL) and TS (PRED) hyperspectral images, which belong to S-C and P-E categories. The results are expressed as percentages of images predicted as *Punctured* or predicted as *Sound*, concerning both P-E annotated and S-C images individually. As mentioned in **Section 2.4**, the prediction images obtained by applying the classification models to the hyperspectral images are further elaborated to obtain an image-level prediction. In particular, only those prediction images containing at least one *Punctured* object larger than 105 pixels were classified as *Punctured images*, while the other images were considered as *Sound images*.

			Abate Fétel			Williams		
			# images	Predicted as *Punctured* (%)	Predicted as *Sound* (%)	# images	Predicted as *Punctured* (%)	Predicted as *Sound* (%)
Soft PLSDA	P-E Annotated	TR	**129**	**87.6**	12.4	**182**	**97.3**	2.7
		TS	**81**	**81.5**	18.5	**91**	**94.5**	5.5
	S-C	TR	**412**	10.2	**89.8**	**340**	10.3	**89.7**
		TS	**237**	36.7	**63.3**	**240**	5.4	**94.6**
s-Soft PLSDA	P-E Annotated	TR	**129**	**86.8**	13.2	**182**	**97.3**	2.7
		TS	**81**	**82.7**	17.3	**91**	**96.7**	3.3
	S-C	TR	**412**	9.2	**90.8**	**340**	15.6	**84.4**
		TS	**237**	25.7	**74.3**	**240**	13.3	**86.7**

***Table 7**. Image-level classification results of Soft PLS-DA and s-Soft PLS-DA models. For each cultivar, the prediction results for the P-E annotated and the S-C images, subdivided in TR and TS images, are reported. SENS values, i.e., P-E annotated images correctly predicted as Punctured images, are coloured in red while SPEC values, i.e., S-C images correctly rejected by the Punctured class, are coloured in green.*

Despite better global performances for the *cv*. Williams, the ability of the models in correctly identifying the P-E images as *Punctured* are generally higher than the ability in recognizing the S-C images as *Sound*. This is the major downside of the *cv*. Abate Fétel classification model considering the whole spectrum, that allowed to recognize as *Sound* only 89.8% and 63.3% of the S-C images belonging to the TR set and external TS set, respectively. A slight improvement was achieved considering only the spectral regions selected by sparse based methods (s-Soft PLS-DA), enabling the correct assignment of 90.8% and 74.3% of the S-C images belonging to the TR set and external TS set, respectively. The models' ability in recognizing correctly the P-E annotated images as *Punctured images* is definitely higher, achieving 81.5% and 82.7% in prediction of the external TS set images considering the whole spectrum and sparse variable selection, respectively.

On the other hand, the classification results for *cv*. Williams were far more promising. Considering the whole spectrum, the ability of *cv*. Williams Soft PLS-DA model outperformed the *cv*. Abate Fétel one, leading to the correct assignment of 94.5% of the P-E annotated images and 94.6% of the S-C images belonging to the external TS set. Throughout sparse based variable selection, the classification results were slightly worse, leading to the correct assignment of 96.7% of the P-E annotated images and of 86.7% of the S-C images belonging to the external TS set.

In **Figures 3-4** some representative prediction images for both cultivars are reported along with the corresponding RGB images of peeled and unpeeled fruits used as reference. For *cv.* Abate Fétel, the prediction images of S-C and P-E images belonging to the TS are shown in **Figure 3 A-C**.

As a result of the image level decision rule, the S-C image in **Figure 3 B** is wrongly assigned to the *Punctured* class by Soft PLS-DA, by predicting a rusty superficial spot on the fruit peel as puncture. Conversely, s-Soft PLS-DA is able to correctly predict the S-C image as *Sound*, and also reduce the number of NA pixels. Both S-C and P-E images shown in **Figure 3 A** and **Figure 3 C** are correctly predicted by the models. Nonetheless, the prediction image obtained applying Soft PLS-DA on the P-E image shows some additional areas wrongly predicted as *Punctured*, whose extension is reduced when considering s-Soft PLS-DA results.

Figure 3 D shows the prediction images of a P-E image whose punctured areas were not previously annotated by the methodology followed in [36]. In this case, despite being quite small, the area ascribable to the puncture was correctly seen by both Soft PLS-DA and s-Soft PLS-DA. In **Figure 3 E** an image of a Damaged-Control (D-C) fruit section is shown as an example of how the models may behave towards other types of damage. In this case, the pear is affected by pear scab (i.e., other type of damage, not ascribable to the BMSB): Soft PLS-DA effectively predicts several areas of the fruit as NA, while the sparse solution assigns them as *Punctured* areas. Despite handling this situation in a different manner, both models seem to partially identify areas of the fruit that are not compliant to the *Sound* class.

The same evaluation can be done for *cv*. Williams prediction images shown in **Figure 4**. In this case, S-C and P-E images in **Figure 4 A-B** belong to the TS set: both images were correctly predicted as *Sound* and *Punctured images* by Soft PLS-DA and s-Soft PLS-DA, with no significant differences. As for *cv.* Abate Fétel, the prediction images of a P-E image whose punctured areas were not previously annotated is shown in **Figure 4 C**. Throughout the comparison with RGB reference images, both Soft PLS-DA and s-Soft PLS-DA models enable the identification of punctured area.

Nonetheless, suberification and necrosis of the fruit pulp are a common response of plants and corresponding fruits to stress, that may be due to different causes. For *cv*. Williams, an additional evaluation of models' robustness was done towards D-C images showing suberification damages (**Figure 4 D**). Regardless of the agents that caused them to emerge, suberified tissues have the same chemical composition therefore compromising fruits' quality. As for the annotation methodology in [36], the calculated classification models are capable to detect damages sharing the same chemical composition as *Punctured* areas, although not directly being caused by BMSB.

Another interesting scenario is shown in **Figure 4 E**, in which a fruit exposed to BMSB exhibits both BMSB-induced suberification and a moulded area. The corresponding prediction images calculated applying Soft PLS-DA and s-Soft PLS-DA partially detect the suberified areas, correctly predicting the image as *Punctured*. Coherently with **Figure 3 E**, part of the moulded area is predicted as NA by

both models, suggesting that different damages not previously considered for model training may be present.

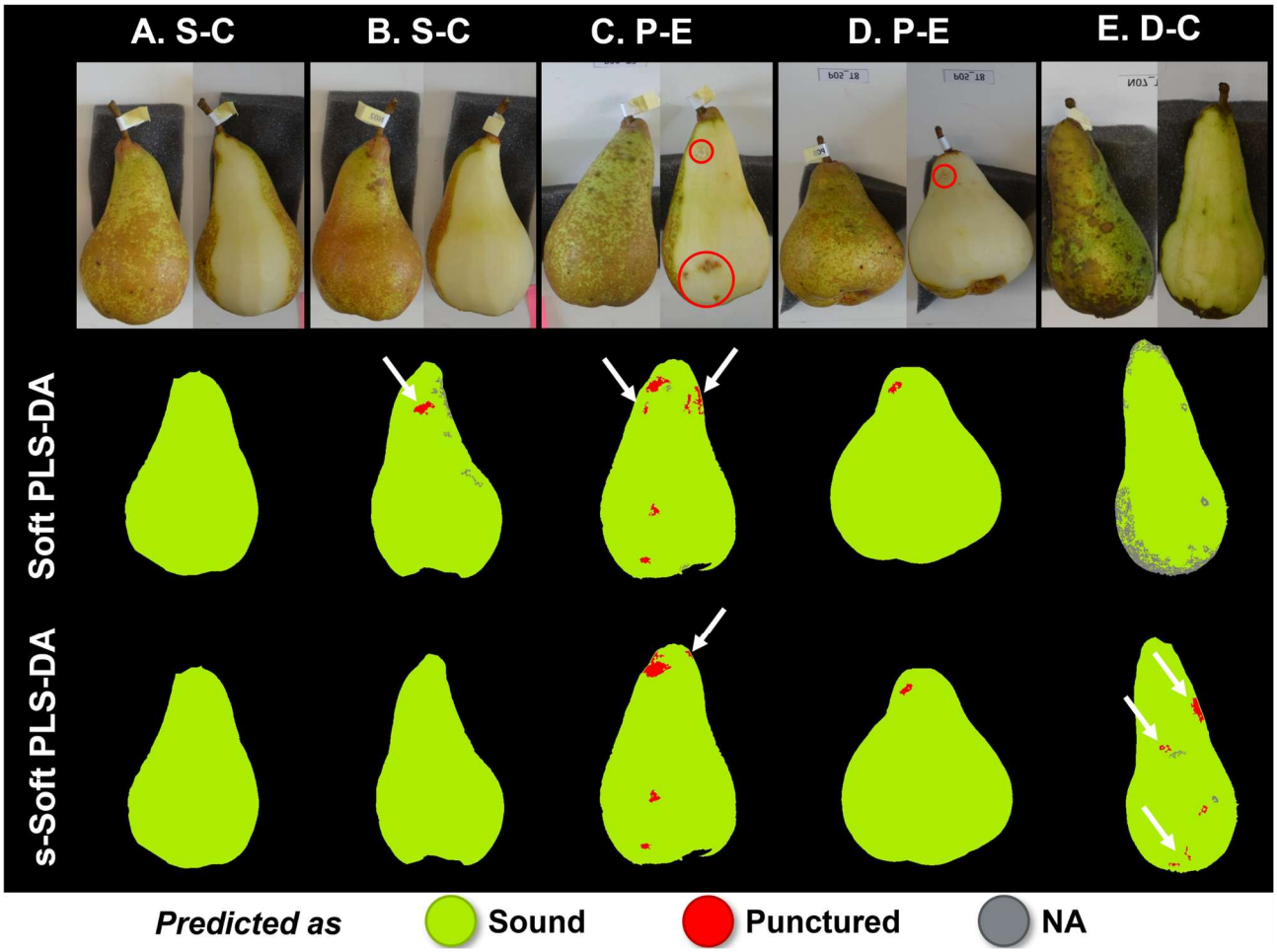


Figure 3 *cv. Abate Fétel prediction images obtained by applying Soft PLS-DA and s-Soft PLS-DA models together with the corresponding RGB images of the samples.*

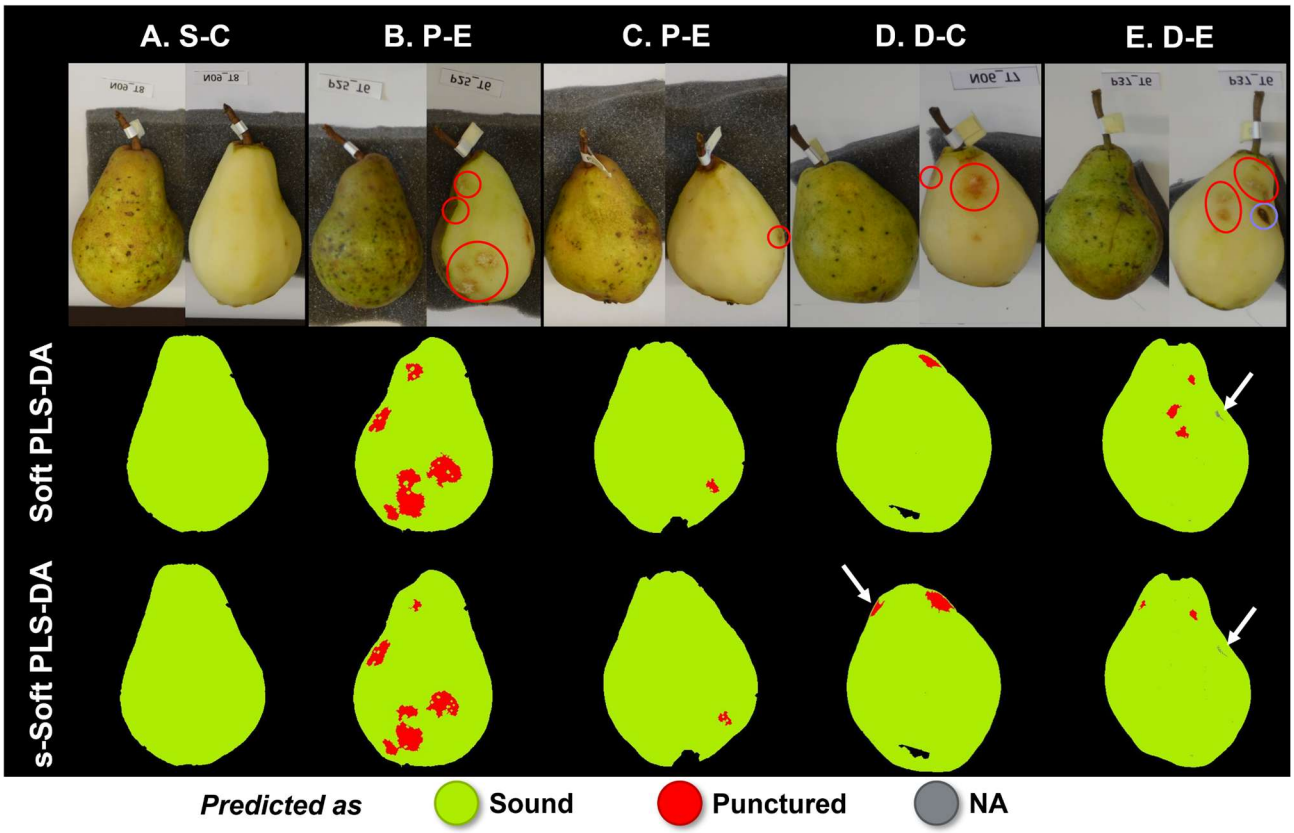


Figure 4 *cv. Williams prediction images obtained by applying Soft PLS-DA and s-Soft PLS-DA models together with the corresponding RGB images of the samples.*

The ability of the classification models towards the identification of punctures during post-harvest storage is consistent, despite the major impact of the different fruits' rotation during the subsequent image acquisition times [36].

Figures 5-6 display the predictions for each P-E and S-C section over time for *cv.* Abate Fétel and *cv.* Williams, respectively. The predictions are presented in a heatmap-style layout, where each column represents a specific fruit section and the rows correspond to the different acquisition times. Accordingly, each cell represents the image-level classification of a single hyperspectral image: red cells indicate images classified as *Punctured*, green cells correspond to images classified as *Sound*, and grey cells indicate that no image was acquired for that section at the corresponding acquisition time.

Overall, punctures seem detectable from the harvest day (T1), becoming slightly more defined around the edges over time. Concerning both cultivars, the P-E sections predicted as *Punctured* at least one time by Soft PLS-DA and s-Soft PLS-DA are basically the same. On the other hand, the number of times that P-E sections are predicted as *Punctured* is slightly different depending on the model, since s-Soft PLS-DA improves the predictions for the *Punctured* class.

The same conclusion cannot be drawn for the S-C images. Concerning *cv.* Abate Fétel, s-Soft PLS-DA model seem to misclassify more often S-C sections acquired at T1 (day of harvest): interestingly, the same behaviour is shown for the prediction of S-E sections (**Figure 7**). Conversely, this trend is not shown for Soft PLS-DA, where the wrong assignations are repeated for the same S-C sections over time, probably due to superficial spots on the fruit peel (e.g., section C of sample N02 harvested in 2022, also reported in **Figure 3 B**). This behaviour may be partially explained by the relevancy of the spectral region at 1510-1570 nm for *cv.* Abate Fétel s-Soft PLS-DA model. Interestingly, this spectral region seem related to starch content [55], compound that gradually decrease during fruit ripening since being converted in soluble sugars (*see* **Figure 2**).

For *cv.* Williams, while the predictions for S-C sections obtained with Soft PLS-DA do not show significant differences between vintages, s-Soft PLS-DA tends to misclassify more often S-C sections of fruits harvested in 2022. A similar behaviour is shown in predictions for the S-E sections (**Figure 8**), where the sections predicted as *Punctured* at least one time are more or less the same for fruits harvested in 2023 while the frequency of misclassifications is higher for fruits harvested in 2022.

Table 8 reports the global classification performances at image level towards modelled classes (i.e., P-E and S-C images) and unknown classes (i.e., D-C, D-E, S-E images), in terms of percentages of images seen as *Punctured* or seen as *Sound* by the models.

The D-C and D-E images showing mild damage (Type 1) of both cultivars are generally seen as *Sound* by the models, and this can be explained considering low severity of the damage. On the other hand, for the D-C images showing severe damage not due to BMSB (Type 3) the percentages of images seen as *Punctured* by the models tend to increase. In this case, this category comprehends all the severe damages determined by abiotic or biotic agents such as moulds, scab but also suberifications or pulp necrosis not due to BMSB. As a matter of fact, the percentages of D-C (Type 3) images seen as *Punctured* exceed the 60% for *cv.* Williams, for which the occurrence of suberified areas on fruits is a common plant response to adversities (**Figure 4 D**).

The D-E images showing severe damage not ascribable to BMSB (Type 3) are generally affected by moulds, scab or chilling injuries. These injuries are mainly seen as *Sound* by the models, which may be due to the different chemical composition from suberified areas. An particular case is depicted in **Figure 4 E**, in which a D-E image is showing a punctured and a moulded area: the two images falling into this scenario are seen as *Punctured*. Analogous considerations can be made comparing the predictions for P-E sections and other damaged sections (i.e., D-C, D-E) reported in **Figures 9-10**.

Concerning the ability of the models to correctly identify *Sound* pears, the percentages of images labelled as S-C and S-E predicted as *Sound* can be compared. The worst results were obtained for *cv.*

Williams when using s-Soft PLS-DA, where the percentage of S-E images predicted as *Sound* was up 21% less than S-C images. The lower classification performances seem emphasized for S-E images of fruits harvested in 2022 (**Table 9**). Interestingly, the worst outcomes coincide with *cv.* Williams fruits harvested in 2022, which were deeply affected by extreme high temperatures registered at the beginning of August 2022. The adverse climatic conditions may had a negative impact, not only for BMSB specimens' vitality in the field and the amount of P-E images collected [36], but on the overall quality of the harvested fruits.

Conversely, the results for *cv*. Abate Fétel are more promising: indeed, the percentage of images predicted as *Sound* of S-C and S-E are comparable for both Soft PLS-DA and s-Soft PLS-DA (**Tables 8** and **10**).

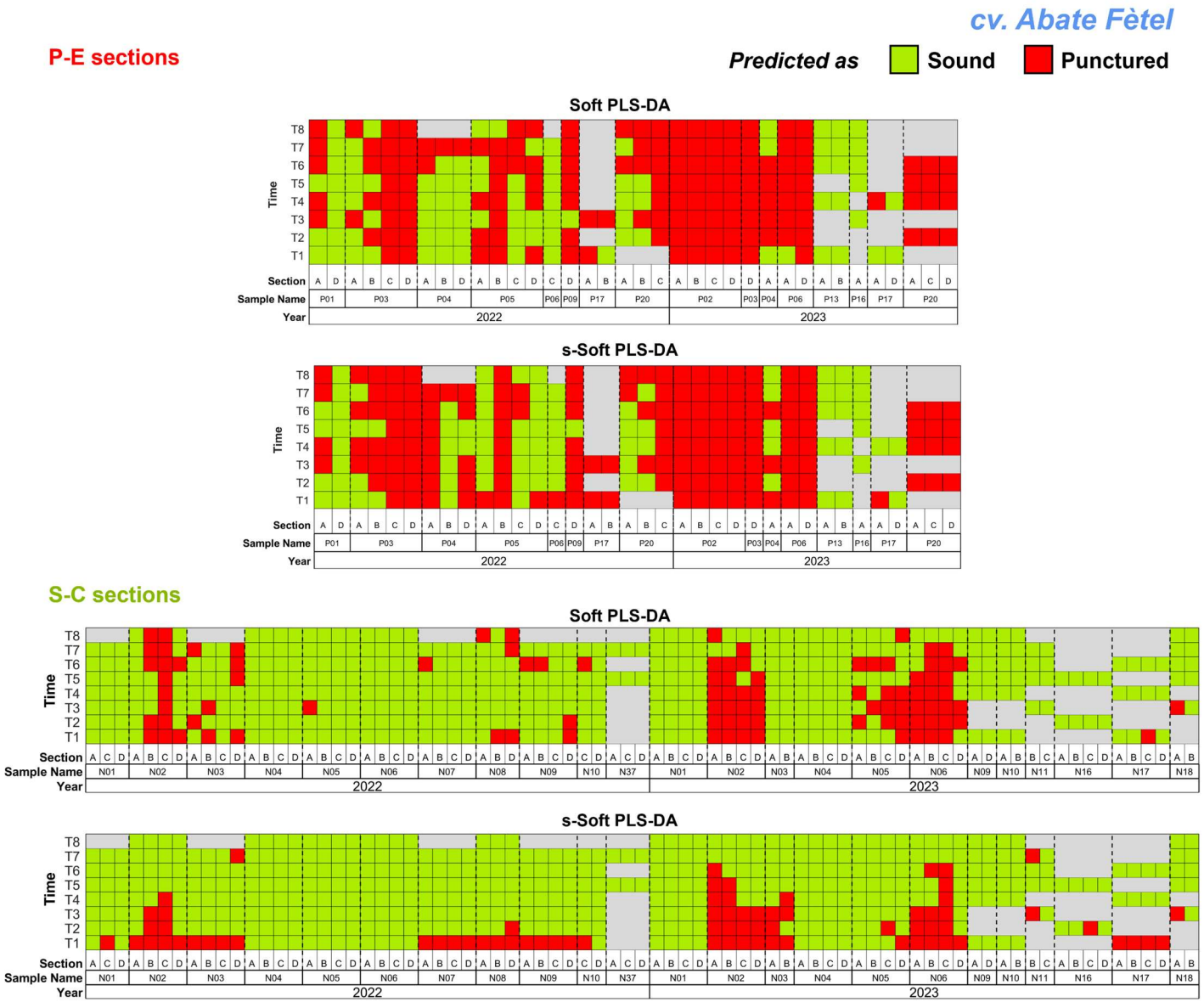


Figure 5 *cv. Abate Fétel: image-level predictions for P-E and S-C sections over time, considering Soft PLS-DA and s-Soft PLS-DA models.*

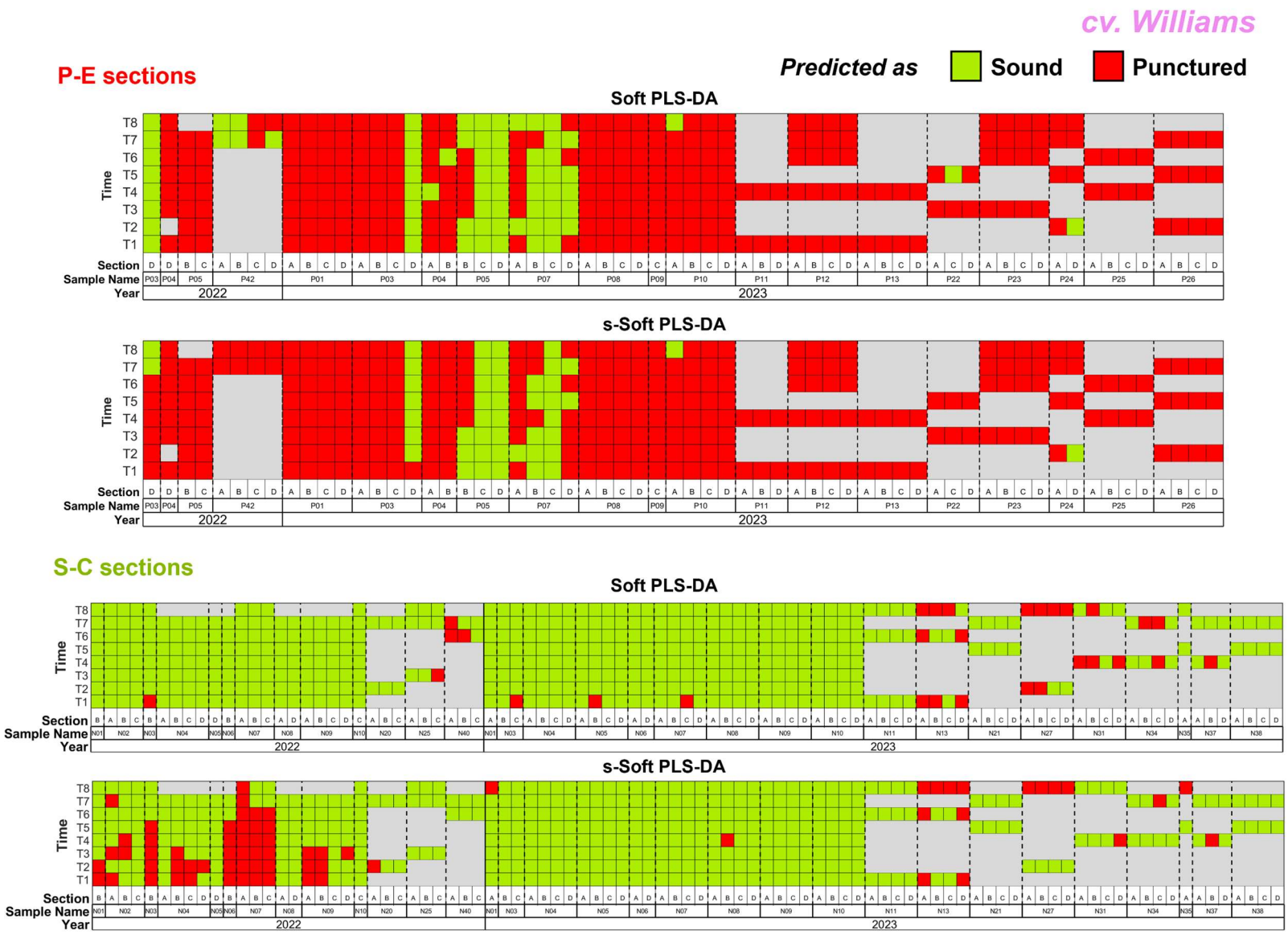


Figure 6 *cv. Williams: image-level predictions for P-E and S-C sections over time, considering Soft PLS-DA and s-Soft PLS-DA models.*

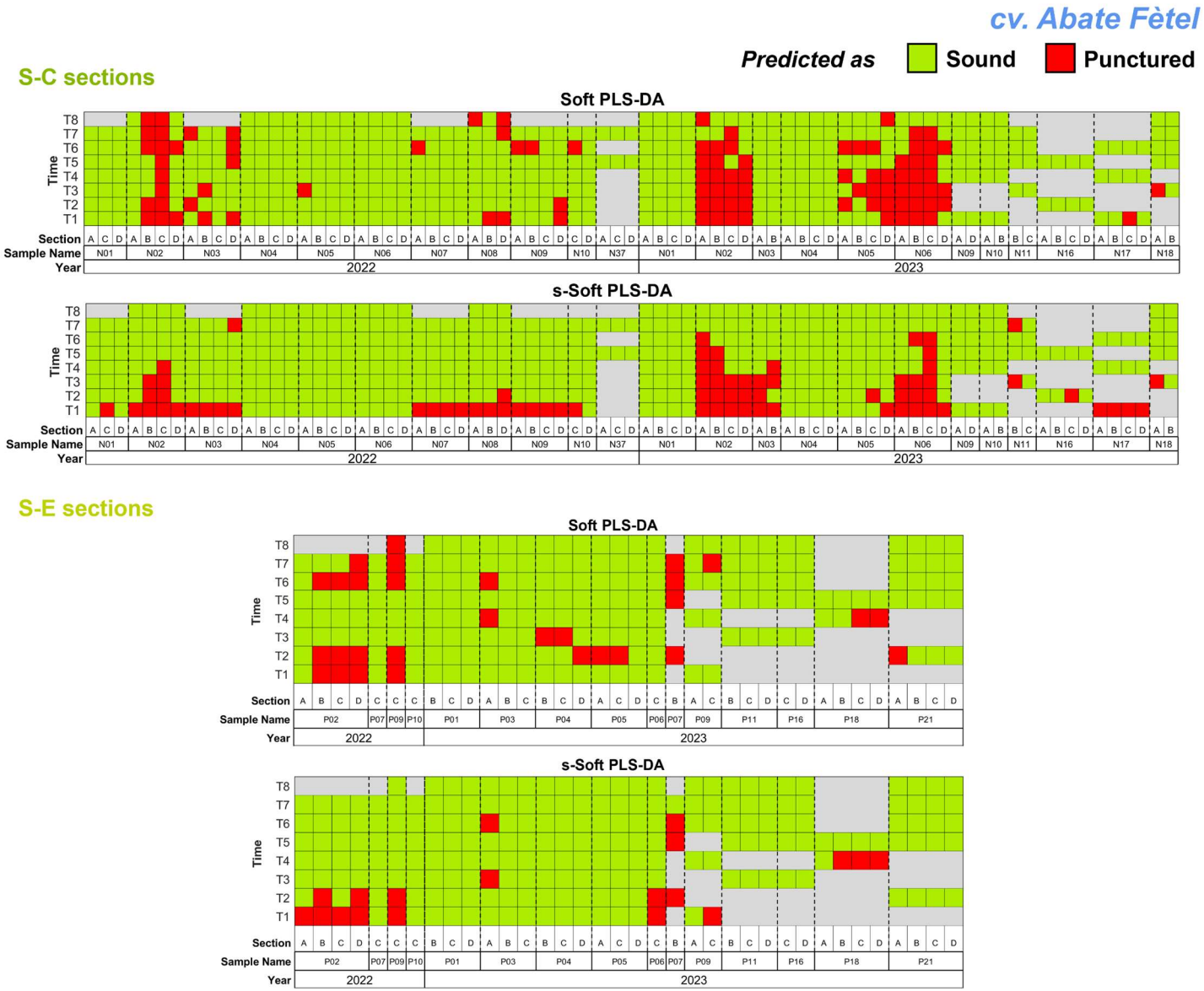


Figure 7 *cv. Abate Fétel: image-level predictions for S-C and S-E sections, considering Soft PLS-DA and s-Soft PLS-DA models.*

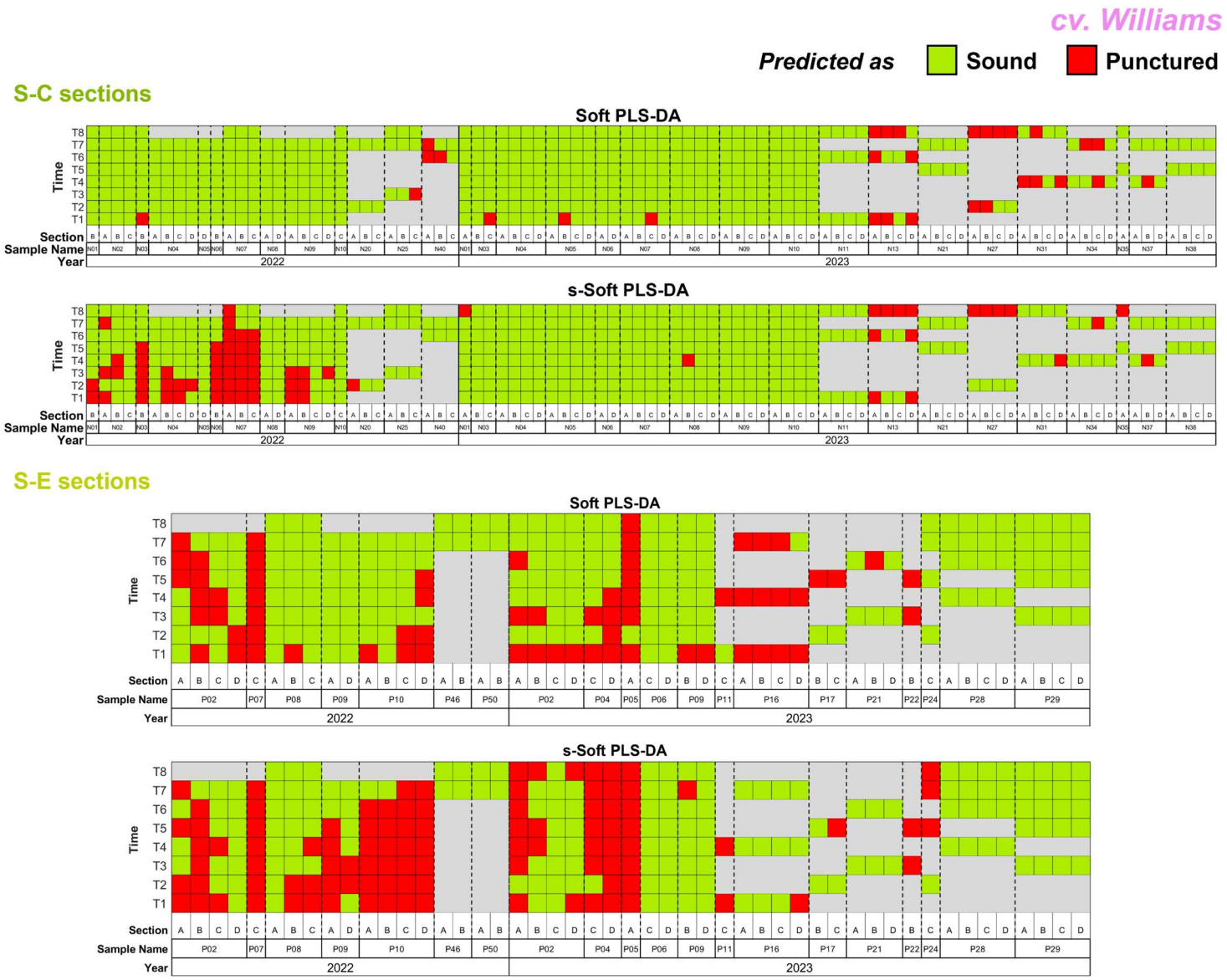


Figure 8 *cv. Williams: image-level predictions for S-C and S-E sections, considering Soft PLS-DA and s-Soft PLS-DA models.*

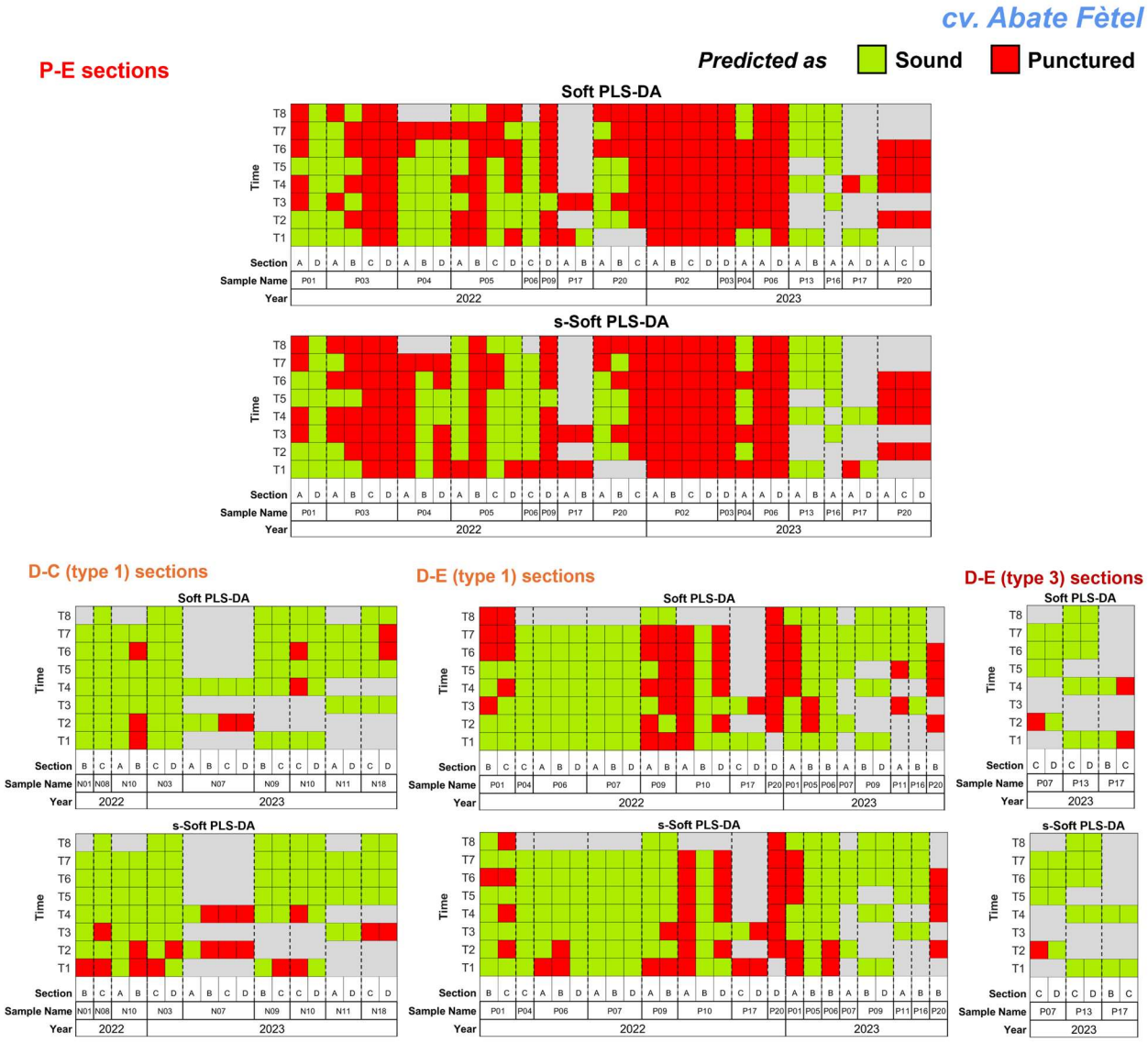


Figure 9 *cv. Abate Fétel: image-level predictions of Soft PLS-DA and s-Soft PLS-DA models for P-E sections and images of fruit sections with damage not ascribable to BMSB.*

Figure 10 *cv. Williams image-level predictions of Soft PLS-DA and s-Soft PLS-DA models for P-E sections and images of fruit sections with damage not ascribable to BMSB.*

			Soft PLSDA		s-Soft PLSDA	
		# images	Predicted as *Punctured* (%)	Predicted as *Sound* (%)	Predicted as *Punctured* (%)	Predicted as *Sound* (%)
Abate Fétel	D-C (Type 1)	122	14.8%	85.2%	20.5%	79.5%
	D-C (Type 3)	12	41.7%	58.3%	50.0%	50.0%
	D-E (Type 1)	247	37.7%	62.3%	34.4%	65.6%
	D-E (Type 3)	63	41.3%	58.7%	39.7%	60.3%
	P-E	322	67.4%	32.6%	68.6%	31.4%
	S-C	653	19.9%	80.1%	15.6%	84.4%
	S-E	260	13.5%	86.5%	10.4%	89.6%
Williams	D-C (Type 1)	207	18.8%	81.2%	28.5%	71.5%
	D-C (Type 3)	152	61.8%	38.2%	68.4%	31.6%
	D-E (Type 1)	263	27.8%	72.2%	39.5%	60.5%
	D-E (Type 3)	74	43.2%	56.8%	23.0%	77.0%
	P-E	352	78.7%	21.3%	84.1%	15.9%
	S-C	582	8.2%	91.8%	14.6%	85.4%
	S-E	331	25.1%	74.9%	35.6%	64.4%

Table 8 *Global evaluation of image-level predictions of Soft PLS-DA and s-Soft PLS-DA models for both cv. Abate Fétel and cv. Williams.*

Williams		# images	Soft PLSDA		s-Soft PLSDA	
			Predicted as *Punctured* (%)	Predicted as *Sound* (%)	Predicted as *Punctured* (%)	Predicted as *Sound* (%)
2022	D-C (Type 1)	146	11.0%	89.0%	24.7%	75.3%
	D-C (Type 3)	64	67.2%	32.8%	85.2%	14.8%
	D-E (Type 1)	229	29.3%	70.7%	44.5%	55.5%
	D-E (Type 3)	43	53.5%	46.5%	23.3%	76.7%
	P-E	54	63.0%	37.0%	79.6%	20.4%
	S-C	242	4.1%	95.9%	24.8%	75.2%
	S-E	174	24.7%	75.3%	41.4%	58.6%
2023	D-C (Type 1)	61	37.7%	62.3%	37.7%	62.3%
	D-C (Type 3)	91	58.2%	41.8%	57.1%	42.9%
	D-E (Type 1)	34	17.6%	82.4%	5.9%	94.1%
	D-E (Type 3)	31	29.0%	71.0%	22.6%	77.4%
	P-E	298	81.5%	18.5%	84.9%	15.1%
	S-C	340	11.2%	88.8%	7.4%	92.6%
	S-E	157	25.5%	74.5%	29.3%	70.7%

Table 9 *Global evaluation of image level predictions for cv. Williams images subdivided by harvest year.*

Abate Fétel		# images	Soft PLSDA		s-Soft PLSDA	
			Predicted as *Punctured* (%)	Predicted as *Sound* (%)	Predicted as *Punctured* (%)	Predicted as *Sound* (%)
2022	D-C (Type 1)	51	23.5%	76.5%	17.6%	82.4%
	D-C (Type 3)	10	55.6%	44.4%	44.4%	55.6%
	D-E (Type 1)	192	42.2%	57.8%	38.5%	61.5%
	D-E (Type 3)	31	67.7%	32.3%	71.0%	29.0%
	P-E	221	62.9%	37.1%	65.2%	34.8%
	S-C	403	16.1%	83.9%	11.9%	88.1%
	S-E	80	21.3%	78.8%	16.3%	83.8%
2023	D-C (Type 1)	71	8.5%	91.5%	22.5%	77.5%
	D-C (Type 3)	3	0%	100%	66.7%	33.3%
	D-E (Type 1)	55	21.8%	78.2%	20.0%	80.0%
	D-E (Type 3)	32	15.6%	84.4%	9.4%	90.6%
	P-E	101	77.2%	22.8%	76.2%	23.8%
	S-C	250	26.0%	74.0%	21.6%	78.4%
	S-E	180	10.0%	90.0%	7.8%	92.2%

Table 10 *Global evaluation of image level predictions for cv. Abate Fétel images subdivided by harvest year.*

4. Conclusions

The present study aimed at developing robust and flexible classification models able to identify damage due to BMSB punctures on pears. In this context, NIR HyperSpectral Imaging (NIR-HSI) was evaluated as a possible method for detecting punctures in post-harvest sorting systems.

As discussed in the previous paper (*see* **Section 3.3** of ***Chapter 3***, [36]), the identification of regions ascribable to BMSB punctures is a challenging task due to the complexity of the dataset and the slight chemical differences between sound and damaged areas. However, a preliminary step allowed the annotation of a consistent number of regions ascribable to punctures, thus supporting the feasibility of a more detailed pixel-level classification.

To enable classification, representative spectra of sound and punctured areas were randomly sampled from sound sections of control fruits (S-C) and images of punctured sections of fruits exposed to BMSB (P-E), respectively. The classification models were calculated considering Soft PLS-DA algorithm, which provides a flexible and robust solution in sorting systems since it is able to handle

unmodelled classes. In addition, Soft PLS-DA was combined with sparse based variable selection (s-Soft PLS-DA) to identify the most important spectral variables for classification. Both Soft PLS-DA and s-Soft PLS-DA showed promising results at pixel-level, achieving prediction efficiencies of at least 91.9% for *cv.* Abate Fétel and 92.3% for *cv.* Williams for the *Punctures* class.

Furthermore, an additional threshold on the size of objects predicted as *Punctures* was set enabling image-level classification. Image-level classification is more suitable for online sorting systems, since it allows the automated assignment of the samples as a whole to a certain class. For *cv*. Williams, comparable classification performances were achieved at both pixel and image levels, with classification efficiencies in prediction of at least 92.3% and 91.6%, respectively. Conversely, image-level performance for *cv*. Abate Fétel was lower – but still satisfactory – leading to 71.8% and 78.4% classification efficiencies in prediction for Soft PLS-DA and s-Soft PLS-DA, respectively.

These results likely reflect the heterogeneity of defects, subtle chemical differences between punctures and sound areas, and inherent variability within the images. Beyond the objectives of the HALY.ID EU project for which the investigation was followed, training the models to detect other defects found on the studied fruits could improve the classification efficiency in post-harvest sorting systems.

Nevertheless, encouraging outcomes were obtained when considering only the relevant wavebands for classification selected by s-Soft PLS-DA. Importantly, most spectral regions relevant for discrimination were consistent across both cultivars and appear to be associated with slight differences in the sugars and polyphenols contents, in agreement with literature [57–60]. The parsimonious solution provided by s-Soft PLS-DA models makes this approach compatible with multispectral imaging (MSI) systems, which are more suitable for industrial sorting applications due to lower costs, higher robustness of the optical components, and reduced data complexity.

Future work may involve the implementation of the selected spectral regions into an MSI-based post-harvest sorting system to evaluate the effectiveness of the proposed approach under real conditions. In this context, the collected hyperspectral data can be used to simulate a multispectral imaging system embedding only band-pass filters falling in the selected spectral regions, enabling a preliminary assessment of the system feasibility [37].

Acknowledgements

Authors wish to thank HALY.ID, project of ERA-NET Cofund ICT-AGRI-FOOD, with funding provided by national sources (Ministero delle politiche agricole e forestali, MIPAAF) and co-funding

by the European Union's Horizon 2020 research and innovation program, Grant Agreement number 862671.

Dr. Rosalba Calvini would like to thank the Italian funding programme Fondo Sociale Europeo REACT-EU - PON "Ricerca e Innovazione" 2014 – 2020 – Azione IV.6 Contratti di ricerca su tematiche Green (D.M. 1062 del 10/08/ 2021) for supporting her research (CUP: E95F21002330001; contract number 17-G-13884–4).

The authors wish also to acknowledge Dr. Daniele Giannetti and Dr. Niccolò Patelli from the Applied Entomology Lab (UNIMORE) for the support on-field.

CRediT Author Statement

Veronica Ferrari: Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Writing – Original draft; Visualisation. **Rosalba Calvini:** Conceptualisation; Methodology; Software; Investigation; Data curation; Writing – Review & editing; Visualisation. **Camilla Menozzi:** Investigation; Data curation; Writing – Review & editing. **Elena Costi:** Resources; Investigation. **Peter Hoffermans:** Resources; Project administration; Funding acquisition. **Lara Maistrello:** Resources; Project administration; Funding acquisition. **Alessandro Ulrici:** Conceptualisation; Methodology; Writing – Review & editing; Supervision; Project administration; Funding acquisition.

References

[1] C.J.A. Bradshaw, B. Leroy, C. Bellard, D. Roiz, C. Albert, A. Fournier, M. Barbet-Massin, J. M. Salles, F. Simard, F. Courchamp, Massive yet grossly underestimated global costs of invasive insects, *Nat. Commun*. 7 (2016) 12986. https://doi.org/10.1038/ncomms12986.

[2] IPBES, Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, 2019. https://doi.org/10.5281/ZENODO.3831673.

[3] T.C. Leskey, A.L. Nielsen, Impact of the Invasive Brown Marmorated Stink Bug in North America and Europe: History, Biology, Ecology, and Management, *Annu. Rev. Entomol.* 63 (2018) 599–618. https://doi.org/10.1146/annurev-ento-020117-043226.

[4] L. Maistrello, Case Study 2: *Halyomorpha halys* (Stål) in Europe, in: A.F. Bueno, A.R. Panizzi (Eds.), Stink Bugs Hemiptera Pentatomidae *Res. Manag. Recent Adv. Case Stud. Braz. Eur. USA*, Springer Nature Switzerland, Cham, 2024: pp. 271–359. https://doi.org/10.1007/978-3-031-69742-5_15.

[5] K.B. Rice, C.J. Bergh, E.J. Bergmann, D.J. Biddinger, C. Dieckhoff, G. Dively, H. Fraser, T. Gariepy, G. Hamilton, T. Haye, A. Herbert, K. Hoelmer, C.R. Hooks, A. Jones, G. Krawczyk, T. Kuhar, H. Martinson, W. Mitchell, A.L. Nielsen, D.G. Pfeiffer, M.J. Raupp, C. Rodriguez-Saona, P. Shearer, P. Shrewsbury, P.D. Venugopal, J. Whalen, N.G. Wiman, T.C. Leskey, J.F. Tooker, Biology, Ecology, and Management of Brown Marmorated Stink Bug (Hemiptera: Pentatomidae), *J. Integr. Pest Manag*. 5 (2014) 1–13. https://doi.org/10.1603/IPM14002.

[6] A.L. Acebes-Doria, T.C. Leskey, J.C. Bergh, Injury to apples and peaches at harvest from feeding by *Halyomorpha halys* (Stål) (Hemiptera: Pentatomidae) nymphs early and late in the season, *Crop Prot.* 89 (2016) 58–65. https://doi.org/10.1016/j.cropro.2016.06.022.

[7] M. Bariselli, R. Bugiani, L. Maistrello, Distribution and damage caused by *Halyomorpha halys* in Italy, *EPPO Bull.* 46 (2016) 332–334. https://doi.org/10.1111/epp.12289.

[8] A.L. Nielsen, G.C. Hamilton, Seasonal Occurrence and Impact of *Halyomorpha halys* (Hemiptera: Pentatomidae) in Tree Fruit, *J. Econ. Entomol.* 102 (2009) 1133–1140. https://doi.org/10.1603/029.102.0335.

[9] E. Costi, T. Haye, L. Maistrello, Biological parameters of the invasive brown marmorated stink bug, *Halyomorpha halys*, in southern Europe, *J. Pest Sci*. 90 (2017) 1059–1067. https://doi.org/10.1007/s10340-017-0899-z.

[10] CSO Italy, Estimation of damage from brown marmorated stink bug and plant pathologies related to climate change, https://www.csoservizi.com/, 2020 (accessed October 28, 2025).

[11] FAOSTAT, Food and Agriculture Organization of the United Nations Statistics division, Crops and Livestock Products. Available online, https://www.fao.org/faostat/en/#data/QCL/visualize, 2025 (accessed October 28, 2025).

[12] ISTAT, Italian National Institute of Statistics. https://esploradati.istat.it/databrowser/, 2025 (accessed October 28, 2025).

[13] European Union, Regulation (EU) No. 1151/2012 of the European Parliament and of the Council of 21 November 2012 on Quality Schemes for Agricultural Products and Foodstuffs. https://eur-lex.europa.eu/eli/reg/2012/1151/oj/eng, 2012 (accessed October 28, 2025).

[14] S. Musacchi, I. Iglesias, D. Neri, Training Systems and Sustainable Orchard Management for European Pear (*Pyrus communis L.*) in the Mediterranean Area: A Review, *Agronomy* 11 (2021) 1765. https://doi.org/10.3390/agronomy11091765.

[15] Pera dell'Emilia-Romagna PGI Consortium, Production Regulations for the "*Pera dell'Emilia-Romagna*" PGI, https://peradellemiliaromagnaigp.it/wp-content/uploads/2023/11/Disciplinare_Pera_dell_Emilia_Romagna_IGP.pdf, 2023 (accessed October 28, 2025).

[16] D. Fornasiero, D. Scaccini, V. Lombardo, G. Galli, A. Pozzebon, Effect of exclusion net timing of deployment and color on *Halyomorpha halys* (Hemiptera: Pentatomidae) infestation in pear and apple orchards, *Crop Prot.* 172 (2023) 106331. https://doi.org/10.1016/j.cropro.2023.106331.

[17] L. Moore, P. Tirello, D. Scaccini, M.D. Toews, C. Duso, A. Pozzebon, Characterizing damage potential of the brown marmorated stink bug in cherry orchards in Italy, *Entomol. Gen.* 39 (2019) 271–283. https://doi.org/10.1127/entomologia/2019/0799.

[18] V. Cortés, J. Blasco, N. Aleixos, S. Cubero, P. Talens, Monitoring strategies for quality control of agricultural products using visible and near-infrared spectroscopy: A review, *Trends Food Sci. Technol*. 85 (2019) 138–148. https://doi.org/10.1016/j.tifs.2019.01.015.

[19] B. Zhang, D. Dai, J. Huang, J. Zhou, Q. Gui, F. Dai, Influence of physical and biological variability and solution methods in fruit and vegetable quality nondestructive inspection by using imaging and near-infrared spectroscopy techniques: A review, *Crit. Rev. Food Sci. Nutr.* 58 (2018) 2099–2118. https://doi.org/10.1080/10408398.2017.1300789.

[20] D. Wu, D. W. Sun, Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part I: Fundamentals, *Innov. Food Sci. Emerg. Technol*. 19 (2013) 1–14. https://doi.org/10.1016/j.ifset.2013.04.014.

[21] J.M. Amigo, I. Martí, A. Gowen, Chapter 9 - Hyperspectral Imaging and Chemometrics: A Perfect Combination for the Analysis of Food Structure, Composition and Quality, in: F. Marini (Ed.), *Data*

Handling in Science and Technology, Elsevier, 2013, pp. 343–370. https://doi.org/10.1016/B978-0-444-59528-7.00009-0.

[22] A. Gowen, C. Odonnell, P. Cullen, G. Downey, J. Frias, Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, *Trends Food Sci. Technol*. 18 (2007) 590–598. https://doi.org/10.1016/j.tifs.2007.06.001.

[23] C. Ferrari, G. Foca, R. Calvini, A. Ulrici, Fast exploration and classification of large hyperspectral image datasets for early bruise detection on apples, *Chemom. Intell. Lab. Syst*. 146 (2015) 108–119. https://doi.org/10.1016/j.chemolab.2015.05.016.

[24] M. Jiang, Y. Li, J. Song, Z. Wang, L. Zhang, L. Song, B. Bai, K. Tu, W. Lan, L. Pan, Study on Black Spot Disease Detection and Pathogenic Process Visualization on Winter Jujubes Using Hyperspectral Imaging System, *Foods* 12 (2023) 435. https://doi.org/10.3390/foods12030435.

[25] Y. Li, S. You, S. Wu, M. Wang, J. Song, W. Lan, K. Tu, L. Pan, Exploring the limit of detection on early implicit bruised 'Korla' fragrant pears using hyperspectral imaging features and spectral variables, *Postharvest Biol. Technol*. 208 (2024) 112668. https://doi.org/10.1016/j.postharvbio.2023.112668.

[26] D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O.L. García-Navarrete, J. Blasco, Recent Advances and Applications of Hyperspectral Imaging for Fruit and Vegetable Quality Assessment, *Food Bioprocess Technol*. 5 (2012) 1121–1142. https://doi.org/10.1007/s11947-011-0725-1.

[27] Y. Lu, Y. Huang, R. Lu, Innovative Hyperspectral Imaging-Based Techniques for Quality Evaluation of Fruits and Vegetables: A Review, *Appl. Sci.* 7 (2017) 189. https://doi.org/10.3390/app7020189.

[28] N.K. Mahanti, R. Pandiselvam, A. Kothakota, P. Ishwarya S., S.K. Chakraborty, M. Kumar, D. Cozzolino, Emerging non-destructive imaging techniques for fruit damage detection: Image processing and analysis, *Trends Food Sci. Technol*. 120 (2022) 418–438. https://doi.org/10.1016/j.tifs.2021.12.021.

[29] G. Wan, J. He, X. Meng, G. Liu, J. Zhang, F. Ma, Q. Zhang, D. Wu, Hyperspectral imaging technology for non-destructive identification of quality deterioration in fruits and vegetables: a review, *Crit. Rev. Food Sci. Nutr*. (2025) 1–30. https://doi.org/10.1080/10408398.2025.2487134.

[30] N. N. Wang, D. W. Sun, Y. C. Yang, H. Pu, Z. Zhu, Recent Advances in the Application of Hyperspectral Imaging for Evaluating Fruit Quality, *Food Anal. Methods* 9 (2016) 178–191. https://doi.org/10.1007/s12161-015-0153-3.

[31] J. Wieme, K. Mollazade, I. Malounas, M. Zude-Sasse, M. Zhao, A. Gowen, D. Argyropoulos, S. Fountas, J. Van Beek, Application of hyperspectral imaging systems and artificial intelligence for quality assessment of fruit, vegetables and mushrooms: A review, *Biosyst. Eng.* 222 (2022) 156–176. https://doi.org/10.1016/j.biosystemseng.2022.07.013.

[32] J. Burger, A. Gowen, Data handling in hyperspectral image analysis, *Chemom. Intell. Lab. Syst*. 108 (2011) 13–22. https://doi.org/10.1016/j.chemolab.2011.04.001.

[33] F. Marini, J.M. Amigo, Unsupervised exploration of hyperspectral and multispectral images, in: J. M. Amigo (Ed.), *Data Handling in Science and Technology*, Elsevier, 2019, pp. 93–114. https://doi.org/10.1016/B978-0-444-63977-6.00006-7.

[34] S.R. Delwiche, I. Baek, M.S. Kim, Does spatial region of interest (ROI) matter in multispectral and hyperspectral imaging of segmented wheat kernels?, *Biosyst. Eng*. 212 (2021) 106–114. https://doi.org/10.1016/j.biosystemseng.2021.10.003.

[35] A. Gowen, J. L. Xu, A. Herrero-Langreo, Comparison of spectral selection methods in the development of classification models from visible near infrared hyperspectral imaging data, *J. Spectr. Imaging* (2019) a4. https://doi.org/10.1255/jsi.2019.a4.

[36] V. Ferrari, R. Calvini, C. Menozzi, E. Costi, D. Giannetti, P. Hoffermans, L. Maistrello, A. Ulrici, NIR hyperspectral imaging to identify damage caused by *Halyomorpha halys* on pears: Automated identification of Regions of Interest related to punctured areas, *Spectrochim. Acta. A. Mol. Biomol. Spectrosc*. 343 (2025) 126543. https://doi.org/10.1016/j.saa.2025.126543.

[37] R. Calvini, J.M. Amigo, A. Ulrici, Transferring results from NIR-hyperspectral to NIR-multispectral imaging systems: A filter-based simulation applied to the classification of Arabica and Robusta green coffee, *Anal. Chim. Acta* 967 (2017) 33–41. https://doi.org/10.1016/j.aca.2017.03.011.

[38] M.Á. Martínez-Domingo, E.M. Valero-Benito, J. Hernández-Andrés, Multispectral and Hyperspectral Imaging, in: A.M. Jiménez-Carvelo, A. Arroyo-Cerezo, L. Cuadros-Rodríguez (Eds.), *Non-Invasive Non-Destr. Methods Food Integr.*, Springer Nature Switzerland, Cham, 2024: pp. 175–201. https://doi.org/10.1007/978-3-031-76465-3_9.

[39] R. Calvini, G. Orlandi, G. Foca, A. Ulrici, Development of a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging, *J. Spectr. Imaging* (2018) a13. https://doi.org/10.1255/jsi.2018.a13.

[40] J. C. Bergh, S.V. Joseph, B.D. Short, M. Nita, T.C. Leskey, Effect of pre-harvest exposures to adult *Halyomorpha halys* (Hemiptera: Pentatomidae) on feeding injury to apple cultivars at harvest and during post-harvest cold storage, *Crop Prot*. 124 (2019) 104872. https://doi.org/10.1016/j.cropro.2019.104872.

[41] H.R. El-Ramady, É. Domokos-Szabolcsy, N.A. Abdalla, H.S. Taha, M. Fári, Postharvest Management of Fruits and Vegetables Storage, in: E. Lichtfouse (Ed.), *Sustain. Agric. Rev.*, Springer International Publishing, Cham, 2015: pp. 65–152. https://doi.org/10.1007/978-3-319-09132-7_2.

[42] FAO, Manual for the preparation and sale of fruits and vegetables: from field to market, Food and Agriculture Organization of the United Nations, Rome, 2004.

[43] P. Gonzalez, J. Pichette, B. Vereecke, B. Masschelein, A. Lambrechts, L. Krasovitski, L. Bikov, An extremely compact and high-speed line-scan hyperspectral imager covering the SWIR range, in: N.K. Dhar, A.K. Dutta (Eds.), *Image Sens. Technol. Mater. Devices Syst. Appl. V*, SPIE, Orlando, United States, 2018: p. 19. https://doi.org/10.1117/12.2304918.

[44] J. M. Amigo, Hyperspectral and multispectral imaging: setting the scene, in: J. M. Amigo (Ed*.), Data Handling in Science and Technology*, Elsevier, 2019, pp. 3–16. https://doi.org/10.1016/B978-0-444-63977-6.00001-8.

[45] 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.

[46] M. Barker, W. Rayens, Partial least squares for discrimination, *J. Chemom*. 17 (2003) 166–173. https://doi.org/10.1002/cem.785.

[47] A.L. Pomerantsev, O.Ye. Rodionova, Multiclass partial least squares discriminant analysis: Taking the right way—A critical tutorial, *J. Chemom.* 32 (2018) e3030. https://doi.org/10.1002/cem.3030.

[48] R. Calvini, A. Ulrici, J.M. Amigo, Sparse-Based Modeling of Hyperspectral Data, in in: C. Ruckebusch (Ed.), *Data Handling in Science and Technology*, Elsevier, 2016, pp. 613–634. https://doi.org/10.1016/B978-0-444-63638-6.00019-X.

[49] R. Calvini, J.M. Amigo, Coupling randomisation and sparse modelling for the exploratory analysis of large hyperspectral datasets, *Chemom. Intell. Lab. Syst.* 248 (2024) 105118. https://doi.org/10.1016/j.chemolab.2024.105118.

[50] P. Filzmoser, M. Gschwandtner, V. Todorov, Review of sparse methods in regression and classification with application to chemometrics, *J. Chemom*. 26 (2012) 42–51. https://doi.org/10.1002/cem.1418.

[51] M.A. Rasmussen, R. Bro, A tutorial on the Lasso approach to sparse modeling, *Chemom. Intell. Lab*. Syst. 119 (2012) 21–31. https://doi.org/10.1016/j.chemolab.2012.10.003.

[52] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, *J. R. Stat. Soc. Ser. B Stat. Methodol*. 58 (1996) 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[53] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, *Chemom. Intell. Lab. Syst.* 174 (2018) 33–44. https://doi.org/10.1016/j.chemolab.2017.12.004.

[54] D.A. Burns, E.W. Ciurczak, eds., Handbook of Near-Infrared Analysis, 3 ed., CRC Press, 2007. https://doi.org/10.1201/9781420007374.

[55] P. Williams, M. Manley, J. Antoniszyn, Near Infrared Technology: Getting the best out of light, AFRICAN SUN MeDIA, 2019.

[56] Jr. Workman Jerry, L. Weyer, Practical Guide to Interpretive Near-Infrared Spectroscopy, 1 ed., CRC Press, 2007. https://doi.org/10.1201/9781420018318.

[57] S. Gacnik, D. Rusjan, M. Mikulic-Petkovsek, Metabolic Response of Peach Fruit to Invasive Brown Marmorated Stink Bug (*Halyomorpha halys* Stål.)'s Infestation, *Int. J. Mol. Sci.* 25 (2024) 606. https://doi.org/10.3390/ijms25010606.

[58] I.O. Ozdemir, O. Karakaya, U. Ates, B. Ozturk, M. Uluca, C. Tuncer, Characterization of hazelnut kernel responses to brown marmorated stink bug [*Halyomorpha halys* Stal (Hemiptera: Pentatomidae)] infestations: Changes in bioactive compounds and fatty acid composition, *J. Food Compos. Anal.* 124 (2023) 105696. https://doi.org/10.1016/j.jfca.2023.105696.

[59] N.C. Weber, J. Razinger, J. Jakopič, V. Schmitzer, M. Hudina, A. Slatnar, R. Veberič, F. Štampar, T. Zamljen, Brown Marmorated Stink Bug (*Halyomorpha halys* Stål.) Attack Induces a Metabolic Response in Strawberry (Fragaria × ananassa Duch.) Fruit, *Horticulturae* 7 (2021) 561. https://doi.org/10.3390/horticulturae7120561.

[60] T. Zamljen, A. Medič, R. Veberič, M. Hudina, F. Štampar, A. Slatnar, Apple Fruit (Malus domestica Borkh.) Metabolic Response to Infestation by Invasive Brown Marmorated Stink Bug (*Halyomorpha halys* Stal.), *Horticulturae* 7 (2021) 212. https://doi.org/10.3390/horticulturae7080212.

[61] J. M. Celton, D. Chagné, S.D. Tustin, S. Terakami, C. Nishitani, T. Yamamoto, S.E. Gardiner, Update on comparative genome mapping between Malus and Pyrus, *BMC Res. Notes* 2 (2009) 182. https://doi.org/10.1186/1756-0500-2-182.

[62] K. Kubitzki, Flowering Plants. Dicotyledons: Celastrales, Oxalidales, Rosales, Cornales, Ericales, Springer Science & Business Media, 2013.

Chapter 4

To SIMCA or not to SIMCA:

facing food authentication issues through NIR Hyperspectral Imaging and alternative classification strategies

4.1. Background and Aim

Near-Infrared Hyperspectral Imaging is a particularly promising technology for real-time assessment of food authenticity [1,2]. In this chapter, NIR-HSI is evaluated as a screening technique to discriminate authentic oregano samples from those suspected of adulteration with leaves of plants of lower commercial value, including olive, myrtle, strawberry tree, and sumac.

A critical aspect in addressing authentication issues lies in the choice of an appropriate supervised classification strategy. As discussed in ***Chapter 2***, supervised classification approaches can be broadly divided into Class Modelling (CM) methods, which focus on modelling similarities within a target class, and Discriminant Analysis (DA) methods, which emphasize differences between predefined classes. CM approaches, such as *Soft Independent Modelling of Class Analogy* (SIMCA, **Section 2.2.3**), are generally recommended for authentication, as they model the variability of the authentic class and allow the rejection of unknown or outlier samples [3–5]. However, they may result in poor performance when the inherent variability of the target class (i.e. authentic oregano) is greater than the variability between target and non-target classes (i.e. pure adulterants), a common situation for highly heterogeneous food matrices such as oregano [6].

Conversely, DA methods, like *Partial Least Squares Discriminant Analysis* (PLS-DA, **Section 2.2.3**), are based on maximizing differences between classes, thus being more effective in handling overlapping classes [7,8]. Nevertheless, these methods strictly depend on the representativeness of the training dataset chosen for model calibration: indeed, the assignment of unknown or extreme samples is inherently forced to one of the predefined classes. This is a fatal limitation in authentication scenarios where all possible adulterations cannot be exhaustively modelled [9].

In this context, hybrid or soft discriminant classification algorithms are a viable alternative, because they combine the strengths of both CM and DA strategies [10,11].

Accordingly, **Section 4.2** presents the development and comparison of supervised classification models based on SIMCA and a hybrid method developed by the supervisors of this Thesis, namely Soft PLS-DA ([12], **Section 2.2.3**).

For oregano authentication, supervised classification models were calculated on a training set of authentic oregano and pure adulterants using Alt-SIMCA and the hybrid method Soft PLS-DA [13].

Soft PLS-DA outperformed SIMCA, demonstrating its suitability for handling strongly overlapping classes. Moreover, Soft PLS-DA prediction images were used to quantify the percentage of pixels predicted as oregano (PPO %). Based on these results, a detection limit of NIR-HSI was established, enabling reliable discrimination of authentic and adulterated oregano samples.

References

[1] C. McVey, T.F. McGrath, S.A. Haughey, C.T. Elliott, A rapid food chain approach for authenticity screening: The development, validation and transferability of a chemometric model using two handheld near infrared spectroscopy (NIRS) devices, Talanta 222 (2021) 121533. https://doi.org/10.1016/j.talanta.2020.121533.

[2] A. Gowen, C. Odonnell, P. Cullen, G. Downey, J. Frias, Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, Trends in Food Science & Technology 18 (2007) 590–598. https://doi.org/10.1016/j.tifs.2007.06.001.

[3] R. Vitale, M. Cocchi, A. Biancolillo, C. Ruckebusch, F. Marini, Class modelling by Soft Independent Modelling of Class Analogy: why, when, how? A tutorial, Analytica Chimica Acta 1270 (2023) 341304. https://doi.org/10.1016/j.aca.2023.341304.

[4] Y.V. Zontov, O.Ye. Rodionova, S.V. Kucheryavskiy, A.L. Pomerantsev, DD-SIMCA – A MATLAB GUI tool for data driven SIMCA approach, Chemometrics and Intelligent Laboratory Systems 167 (2017) 23–28. https://doi.org/10.1016/j.chemolab.2017.05.010.

[5] P. Oliveri, V. Di Egidio, T. Woodcock, G. Downey, Application of class-modelling techniques to near infrared data for food authentication purposes, Food Chemistry 125 (2011) 1450–1456. https://doi.org/10.1016/j.foodchem.2010.10.047.

[6] Z. Małyjurek, D. de Beer, H. van Schoor, J. Colling, E. Joubert, B. Walczak, Class-modelling of overlapping classes. A two-step authentication approach, Analytica Chimica Acta 1191 (2022) 339284. https://doi.org/10.1016/j.aca.2021.339284.

[7] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods 5 (2013) 3790. https://doi.org/10.1039/c3ay40582f.

[8] M. Barker, W. Rayens, Partial least squares for discrimination, Journal of Chemometrics 17 (2003) 166–173. https://doi.org/10.1002/cem.785.

[9] O.Ye. Rodionova, A.V. Titova, A.L. Pomerantsev, Discriminant analysis is an inappropriate method of authentication, TrAC Trends in Analytical Chemistry 78 (2016) 17–22. https://doi.org/10.1016/j.trac.2016.01.010.

[10] R. Vitale, F. Marini, C. Ruckebusch, SIMCA Modeling for Overlapping Classes: Fixed or Optimized Decision Threshold?, Anal. Chem. 90 (2018) 10738–10747. https://doi.org/10.1021/acs.analchem.8b01270.

[11] Z. Małyjurek, D. De Beer, E. Joubert, B. Walczak, Combining class-modelling and discriminant methods for improvement of products authentication, Chemometrics and Intelligent Laboratory Systems 228 (2022) 104620. https://doi.org/10.1016/j.chemolab.2022.104620.

[12] R. Calvini, G. Orlandi, G. Foca, A. Ulrici, Development of a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging, J. Spectral Imaging (2018) a13. https://doi.org/10.1255/jsi.2018.a13.

[13] Ferrari, V., Calvini, R., Menozzi, C., Ulrici, A., Bragolusi, M., Piro, R., Tata, A., Suman, M., Foca, G. Addressing adulteration challenges of dried oregano leaves by NIR HyperSpectral Imaging, Chemometrics and Intelligent Laboratory Systems 249 (2024), 105133.

4.2. Addressing adulteration challenges of dried oregano leaves by NIR HyperSpectral Imaging

What follows is the integral content of: Ferrari, V., Calvini, R., Menozzi, C., Ulrici, A., Bragolusi, M., Piro, R., Tata, A., Suman, M., Foca, G. (2024). Addressing adulteration challenges of dried oregano leaves by NIR HyperSpectral Imaging, *Chemometrics and Intelligent Laboratory Systems*, 249, 105133. DOI: 10.1016/j.chemolab.2024.105133

Addressing adulteration challenges of dried oregano leaves by NIR HyperSpectral Imaging

Veronica Ferrari[1], Rosalba Calvini*[1,2], Camilla Menozzi[1], Alessandro Ulrici[1,2], Marco Bragolusi[3], Roberto Piro[3], Alessandra Tata[3], Michele Suman[4,5], Giorgia Foca[1,2]

[1] *Dipartimento di Scienze della Vita, Università di Modena e Reggio Emilia, Padiglione Besta, Via Amendola, 2 – 42122 Reggio Emilia, Italy*

[2] *Centro Interdipartimentale BIOGEST-SITEIA, Università degli Studi di Modena e Reggio Emilia, Piazzale Europa, 1 – 42122 Reggio Emilia, Italy*

[3] *Istituto Zooprofilattico Sperimentale Delle Venezie, Laboratorio di Chimica Sperimentale, Viale Fiume 78, 36100, Vicenza, Italy*

[4] *Analytical Food Science, Barilla G. e R. Fratelli S.p.A., Via Mantova, 166, 43122, Parma, Italy*

[5] *Department for Sustainable Food Process, Catholic University Sacred Heart, Piacenza, Italy*

* *Corresponding author*

Abstract

Dried oregano leaves are particularly prone to adulteration because of their widespread distribution and their easy mixing with leaves of other plants of lower commercial value, such as olive, myrtle, strawberry tree, or sumac. To reveal the presence of adulteration, in this study we considered an untargeted analytical approach, which instead of involving the *a priori* selection of specific compounds of interest is focused on defining the characteristic spectral signature of authentic oregano with respect to its most frequent adulterants. NIR HyperSpectral Imaging (NIR-HSI) represents a state-of-the-art, rapid and non-destructive technique, allowing for the collection of both spectral and spatial information from the sample, making it particularly suitable for characterizing visually heterogeneous samples.

Authentication issues are typically assessed through class modelling techniques and Soft Independent Modelling of class Analogy (SIMCA) is one of the most used algorithms in this scenario. However,

the high variability and heterogeneity within the authentic oregano class resulted in poor outcomes when SIMCA was applied. As an alternative, Soft Partial Least Squares Discriminant Analysis (Soft PLS-DA) algorithm was applied to differentiate authentic oregano samples from pure adulterants. Soft PLS-DA represents a hybrid approach that combines the advantages of both discriminant and class modelling techniques. The resultant classification model has indeed led to promising results, achieving a prediction efficiency of 92.9%. Finally, based on the percentage of pixels predicted as oregano in the Soft-PLSDA prediction images, a threshold value of 10% was established, serving as a detection limit of NIR-HSI to distinguish authentic oregano samples from adulterated ones.

Keywords: Oregano adulteration; hyperspectral imaging; NIR; classification; SIMCA; Soft PLS-DA.

1. Introduction

The global market of herbs and spices has experienced unprecedented growth in recent years, driven by an increasing demand for culinary diversity, natural flavour enhancers, and the perceived health benefits of herbal products [1]. While this growth presents excellent opportunities for the industry, it also raises concerns about authenticity and integrity of these food products. An alarming consequence of this flourishing market is the increasing risk of adulteration, which is defined by European Spice Association as “the deliberate and intentional inclusion in herbs and spices of substances whose presence is not legally declared, is not permitted or is present in form which might mislead or confuse the customer, leading to an imitated food and/or product of reduced value” [2].

In this context, we are referring to Economically Motivated Adulteration (EMA), which involves the deliberate act of altering products, particularly in the food industry, with the aim of gaining a financial advantage. This practice entails substituting expensive ingredients with lower-quality alternatives to reduce costs [3, 4]. In the case of herbs like oregano, EMA often includes adding lower-cost plant materials, which may encompass different botanical species. This unethical activity is facilitated by complexity of the supply chains, spanned in multiple stages occurring in countries that are different from that of the final sale [5]. Beyond the economic harm to consumers and the damaged reputation of honest producers and distributors, adulteration is a practice that can pose also significant health risks. Indeed, adulterated products may contain potential allergens that are not declared in the food labels, or the adulteration process may involve the introduction of different pesticides, which can accumulate dangerously in the adulterated product [6, 7].

According to a technical report by the European Commission's science and knowledge service (JCR) [5], oregano – the herb used to flavour many foods such as pizza and responsible for its characteristic "Italian" aroma – has emerged as the most adulterated herb, with 48% of samples suspected of adulteration. The botanical species introduced as substitutes typically include olive leaves, sumac, cistus, strawberry tree and myrtle, whose dried and ground leaves are visually indistinguishable from oregano. Since unintentional contamination can occur during processing, the presence of extraneous matter is still tolerated within 2% [8, 9].

While the need for a proper authentication of dried oregano is evident, the selection of the most appropriate analytical technique is not straightforward. Various methodologies have been employed to face this authentication issue, each with its own advantages and limitations. In this context, we can delineate two primary approaches that have been employed by various research groups: targeted and untargeted approaches.

Targeted approaches are based on the identification and quantification of specific chemical markers of adulteration or authenticity for a specific product. For instance, Dabrova and colleagues [6] employed advanced mass spectrometry methods to analyse 400 pesticides in both genuine and adulterated oregano samples, identifying a number of compounds that were exclusively present in those adulterated. Other studies employed liquid chromatography coupled with mass spectrometry to detect additional biomarkers of adulteration [10, 11]. Cottened and coauthors [12] used a DNA metabarcoding approach to analyse commercial samples of spices and herbs; in 22% of the examined samples, they identified undeclared species in complex mixtures containing down to 1% of adulterants. On the other hand, Pages-Rebull and coauthors [13] utilized HPLC-UV technique to determine phenolic compounds in various spices and aromatic herbs. Their study revealed that the presence and quantity of six of these compounds define their typical profile, making them suitable markers of authenticity. Targeted methodologies are highly selective and capable of detecting even very low levels of adulteration. However, they come with significant drawbacks, including high costs for the analysis, the requirement for highly specialized personnel, and the need for prior knowledge of the specific markers, which may not always be available. Additionally, the extraction and identification of markers often involve a lengthy and labour-intensive sample preparation process, particularly when dealing with complex matrices such as food samples.

Conversely, untargeted approaches entail a comprehensive examination of a sample distinctive chemical profile, without prior selection of specific compounds of interest. This approach facilitates a deep comprehension of the sample composition, essentially "recognizing" the profile of an authentic sample, much like its fingerprint, against which adulterated samples exhibit differences. Untargeted

methods, therefore, necessitate the processing of analytical results using multivariate chemometric techniques, essential for extracting pertinent information from the fingerprint [14].

Among the recent untargeted techniques employed for oregano authentication, notable examples include nuclear magnetic resonance (NMR) spectroscopy, which has been effectively utilized for initial fingerprinting to discern oregano types, geographical origins, and the presence of other plant additives [15]. In other instances, mass spectrometry has been employed, using instrumental setups that offer rapid analyses with minimal or no sample preparation, such as Proton-Transfer Reaction Time-of-Flight Mass Spectrometry (PTR-TOF-MS), enabling real-time detection of volatile organic compounds [16]. Additionally, various applications of Ambient Mass Spectrometry, such as Direct Analysis in Real Time (DART-MS) [17-19] or Atmospheric Solid Analysis Probe (ASAP-MS) [19], have been utilized. While in these cases sample preparation may not be as time-consuming, challenges persist regarding instrument costs and the high level of technical expertise required of analysts. Mid- and Near-InfraRed spectroscopic fingerprinting techniques can thus overcome these limitations, simplifying and cost-effectively enhancing the analysis process. For these reasons, in the very recent years, several studies have emerged regarding the use of FTIR (Fourier-Transform InfraRed) and NIR (Near-InfraRed) spectroscopies for the authentication of both oregano [16, 20-22] and other herbs and spices [4, 23-25]. The last relatively unexplored frontier of infrared spectroscopic analysis for herbs and spices authentication is NIR HyperSpectral Imaging (NIR-HSI) [16, 26, 27].

NIR-HSI is a cutting-edge, rapid and non-destructive technique that allows the collection of both spectral and spatial information of the sample [28-30]. Indeed, each pixel of a NIR hyperspectral image contains a complete NIR spectrum, which represents a sort of chemical fingerprint at the corresponding sample position. In this manner it is possible to obtain the so-called chemical maps of acquired samples, i.e., to characterize sample chemical composition and evaluate how it varies on sample surface. This method is particularly suitable for characterizing heterogeneous food matrices, such as ground herbs. Note that the ground herbs and spices contaminated with extraneous plant species may exhibit a variety of fragments, that potentially differ in terms of chemical composition.

In this study we used NIR-HSI to analyse authentic oregano samples, pure adulterants, and oregano samples adulterated with various types and amounts of spiked adulterants. The collected images were first explored by Principal Component Analysis (PCA) to assess spectral differences between pure oregano samples and adulterants. Afterwards, multivariate classification methods were used to obtain predictive models able to distinguish between authentic oregano samples and adulterated ones.

Ideally, authentication issues, such as the one considered in this study, should be assessed using Class Modelling (CM) classification approaches [31-33], which construct individual class models based on similarities among samples belonging to the same target class (i.e., authentic oregano). Consequently,

a new sample can be assigned to one or more of the modelled classes, or to none of them. However, CM methods usually provide poor results when the variability within the target class (i.e., authentic oregano) is greater than the variability between target and non-target classes (i.e., pure adulterants), resulting in strong overlapping of the classes.

In contrast to CM, discriminant analysis (DA) methods maximise the differences among the studied classes, even if these differences are subtle, thus providing better results when dealing with overlapping classes. One of the most used algorithms for this purpose is Partial Least Squares Discriminant Analysis (PLS-DA), a modified version of the PLS statistical regression method [34, 35]. Classical DA methods are recommended when the classes of interest are well defined as they force the class assignment of a new sample to one of the modelled classes. Therefore, they are not suitable for authentication issues where new unknown samples may belong to none of the considered classes.

To benefit from the advantages of both CM and DA methods, soft discriminant approaches may represent a valid alternative. These algorithms can be considered hybrid classification methods that combine CM and DA as they enable the classification of samples based on differences among the considered classes, while simultaneously identifying samples belonging to none of them [36-39].

Therefore, classification models were calculated on a training set of authentic oregano and of pure adulterants using both a CM approach, namely Soft Independent Modelling of Class Analogies (SIMCA) [31], and a soft discriminant method, namely Soft Partial Least Squares Discriminant Analysis (Soft PLS-DA) [39]. Essentially, Soft PLS-DA is the same as PLS-DA, however class assignment is subjected to some additional rules involving the calculation of further thresholds based on Q residuals and on *y* predictions. These additional thresholds do not constrain an unknown sample to belong to one of the modelled classes, thus facilitating effective handling of samples adulterated with extraneous herbs or materials not previously accounted for model calculation [36].

Finally, the SIMCA and Soft PLS-DA models were applied at the pixel level to all the acquired images, including those of oregano samples adulterated with different percentages of adulterants. The resulting prediction images were used to provide a final classification of the samples into genuine or adulterated oregano. Moreover, the percentage of pixels predicted as oregano from the Soft PLS-DA model allowed to define a sort of detection limit of NIR-HSI in this context.

2. Materials and Methods

2.1. Oregano samples

The sample set included forty-nine samples: in detail, we analysed twenty-six authentic oregano samples (*Origanum vulgare*, *Origanum onites*, *Coleus amboinicus* and *Origanum vulgare subsp. viridulum*, also known as Sicilian oregano), including two samples containing inflorescences and other two samples certified from the FAPAS proficiency tests 2985A-C. Four samples of pure adulterants including myrtle leaves (*Lagerstroemia indica*), sumac leaves (*Rhus coriaria*), strawberry tree leaves (*Arbutus unedo*) and olive leaves (*Olea europaea*) were also investigated. Moreover, nineteen adulterated oregano samples intentionally mixed with different percentages of myrtle leaves (*Lagerstroemia indica*), sumac leaves (*Rhus coriaria*), strawberry tree leaves (*Arbutus unedo*) and olive leaves (*Olea europaea*) and unintentionally polluted during various steps of the supply chain (with rosemary, cistus, hazelnut and sumac leaves) were tested. Among them, one certified oregano sample adulterated with olive leaves from FAPAS proficiency tests 2985A-C was included. The authentic samples originated from Italy, France, Turkey and Albany, and were harvested between 2019 and 2022. The percentages of adulterations ranged between 1.5 and 60% (*see* **Table 1** for the details of each sample).

ID	LABEL	IDENTIFICATION	YEAR
47_1	Authentic	*Origanum vulgare*	2019
47_2	Authentic	*Coleus amboinicus* (Cuban oregano) from France	2021
47_3	Authentic	*Origanum onites* pre cleaned raw material	2021
47_4	Authentic	*Origanum vulgare* from Turkey	2021
47_5	Authentic	*Origanum vulgare* from Albany	2021
47_6	Authentic	*Origanum onites* processed final product	2021
47_7	Authentic	*Origanum vulgare*	2021
47_8	Authentic	*Origanum* blend	2021
47_9	Authentic	*Origanum vulgare*	2022
47_10	Authentic	*Origanum vulgare*	2022
47_11	Authentic	*Origanum vulgare*	2018
47_12	Authentic	*Origanum vulgare subsp. viridulum,* also known as Sicilian oregano	2020
47_13	Authentic	Certified *Origanum vulgare* from FAPAS proficiency test	2020
47_14	Authentic	Certified *Origanum vulgare* from FAPAS proficiency test	2020
47_15	Authentic	*Origanum vulgare* with inflorescences	2021
47_16	Authentic	*Origanum vulgare subsp. viridulum,* also known as Sicilian oregano	2019
47_17	Authentic	*Origanum vulgare subsp. viridulum,* also known as Sicilian oregano	2019
47_18	Authentic	*Origanum vulgare subsp. viridulum,* also known as Sicilian oregano	2019
47_19	Authentic	*Origanum vulgare subsp. viridulum,* also known as Sicilian oregano	2019
47_20	Authentic	*Origanum vulgare*	2019
47_21	Authentic	*Origanum vulgare subsp. viridulum,* also known as Sicilian oregano	2019
47_22	Authentic	*Origanum vulgare subsp. viridulum,* also known as Sicilian oregano	2019
47_23	Authentic	*Origanum* blend	2019
47_24	Authentic	*Origanum vulgare*	**Unknown**
9	Authentic	*Origanum vulgare subsp. viridulum,* also known as Sicilian oregano with inflorescences	2022
10	Authentic	*Origanum vulgare* raw material	2022
48_1	Adulterated	*Origanum* blend /*Origanum vulgare*/ strawberry leaves (40%/40%/20%)	2021
48_2	Adulterated	*Origanum* blend + strawberry tree leaves (70%/30%)	2021
48_3	Adulterated	*Origanum vulgare* + sumac leaves (80%/20%)	2021
48_4	Adulterated	*Origanum* blend + myrtle leaves (40%/60%)	2021
48_5	Adulterated	*Origanum vulgare* + olive leaves (70%/30%)	2021
48_6	Adulterated	*Origanum* blend + strawberry tree leaves (80%/20%)	2021
48_7	Adulterated	*Origanum* blend + sumac leaves + strawberry tree leaves (60%/20%/20%)	2021
48_8	Adulterated	*Origanum vulgare* + olive leaves (60%/40%)	2021
48_9	Adulterated	*Origanum* blend + myrtle leaves	2021
48_10	Adulterated	*Origanum vulgare* + olive leaves (proficiency test FAPAS)	2020
48_14	Adulterated	*Oreganum vulgare* + olive leaves (80%/20%)	2019
48_16	Adulterated	Accidental pollution (rosemary, cistus, hazelnut and sumac leaves) 6%	2019
48_17	Adulterated	Hazelnut/ sumac leaves / Cistus 1.50%	2019
48_29	Adulterated	*Origanum* blend + myrtle leaves (90%/10%)	2022
48_36	Adulterated	*Origanum* blend + sumac leaves (90%/10%)	2022
1	Adulterated	*Origanum vulgare* + sumac leaves (95%/5%)	2021
2	Adulterated	*Origanum vulgare* + sumac leaves (90%/10%)	2021
3	Adulterated	*Origanum vulgare* + sumac leaves (95%/5%)	2021

4	Adulterated	*Origanum vulgare* + sumac leaves (90%/10%)	2021
5	Pure adulterant	Myrtle leaves (*Lagerstroemia indica)*	2021
6	Pure adulterant	Olive leaves (*Olea europaea)*	2021
7	Pure adulterant	Strawberry tree leaves (*Arbutus unedo)*	2021
8	Pure adulterant	Sumac leaves (*Rhus coriaria)*	2021

Table 1. *Authentic and spiked samples of dried oregano analysed by hyperspectral imaging.*

2.2. Image acquisition and elaboration

Three random aliquots of each sample, ranging between 0.2 g and 1.0 g, were placed inside a glass Petri dish of 6.0 cm diameter and acquired as an individual image using a HSI line-scan system. Such system was composed of a desktop NIR Spectral Scanner (DV Optic, Padova, Italy) embedding a Specim N17E reflectance imaging spectrometer, coupled to a Xenics XEVA 1.7-320 camera (320 × 256 pixels) embedding Specim Oles 31 f/2.0 optical lens and covering the spectral range from 900 to 1700 nm (5 nm resolution, 150 spectral channels). Due to low S/N values, the wavelengths at the extremes of the spectral range were excluded: the final hyperspectral images, covering the spectral range between 980 and 1660 nm (137 wavelengths), were considered for further analysis.

To evaluate the system's stability over time, a setup composed of a silicon carbide sandpaper as sample background – characterized by a very low and constant reflectance spectrum [40] – a 99% reflectance standard and two ceramic tiles with different grayscale tones and intermediate reflectance values, were used for the acquisition of all the images. The raw data were then converted into reflectance values by applying the instrument calibration procedure, which involved measuring the high-reflectance standard reference and the dark current. As a first step of image elaboration, an additional internal calibration was performed to minimize any residual variability among the images over time [41]. In total, 147 hyperspectral images were acquired (= 49 samples × 3 replicates).

The corrected images were then cropped to a size of 248 × 199 pixels, in order to consider only the sample area. Subsequently, the pixels associated with the black sandpaper background and the glass Petri dish were removed from each image using a fast-thresholding procedure: all the pixels with reflectance values lower than 0.50 reflectance units measured at 980 nm were ascribable to the background and removed. Finally, a morphological erosion procedure, using a disk-shaped structuring element with a radius of 2 pixels, was performed to remove the pixels placed at the edges of the samples, which were affected by scattering phenomena and specular reflections of the glass Petri dish [29].

These image elaboration steps were performed using routines written *ad hoc* in MATLAB language (R2020b, The MathWorks Inc., USA).

2.3. Data Analysis

2.3.1. Exploratory Analysis

A preliminary exploratory analysis of the images was performed by means of PCA both at the *pixel-level* and at the *image-level*. In both cases, PCA was performed using linear detrend and mean center as spectral preprocessing methods.

For *pixel-level* exploratory analysis, some representative images of authentic oregano, pure adulterants (myrtle leaves, sumac leaves, strawberry tree leaves, and olive leaves) and adulterated oregano samples (mixtures of oregano and adulterants) were selected and merged together. PCA was applied to the merged images in order to have a preliminary evaluation of the spectral differences between oregano and the considered adulterants.

Subsequently, in order to have a global evaluation of the whole dataset structure at the *image-level*, as well as to gain an overall understanding of sample characteristics and behaviour, the average spectrum was calculated from each image and PCA was applied to the average spectra dataset.

2.3.2. Classification

Classification was carried out using two classification methods: SIMCA as a class modelling technique [31] and the soft discriminant method Soft PLS-DA [39].

For both classification methods the ability to distinguish authentic and adulterated oregano samples was evaluated in two steps. Firstly, *pixel level* models able to classify genuine oregano and pure adulterants (single class including leaves of myrtle, strawberry tree, olive and sumac) were calculated. Then, both models were applied to all the acquired hyperspectral images, including those of adulterated oregano samples (i.e., mixtures of oregano and adulterant), and for each image the corresponding percentage of pixels predicted as oregano (PPO%) was calculated. Based on this value, a threshold was set in order to identify each sample as authentic or adulterated oregano (*see* **Section 2.3.2.4**). As done for the previous exploratory analysis step, the spectra were preprocessed by applying linear-detrend followed by mean centering.

The classification performances of the models were evaluated by cross-validation (CV) and prediction of the external test set (TS) by calculating the statistical parameters sensitivity, specificity and efficiency [42], where:

- sensitivity (SENS), also known as *True Positive Rate*, measures the classifier's ability to correctly identify samples belonging to a considered class. SENS is calculated as the ratio between objects correctly assigned to the modelled class (*true positives*, TP) and all objects belonging to the class.
- specificity (SPEC), also known as *True Negative Rate*, evaluates the classifier's ability to reject samples belonging to other classes. SPEC is calculated as the ratio between objects correctly rejected by the modelled class (*true negatives*, TN) and all objects that do not belong to the considered class.
- efficiency (EFF), defined as the geometric mean of SENS and SPEC, provides an overall assessment of classification performance.

These classification performances were assessed by initially applying the models to a dataset comprising spectra references of authentic oregano and pure adulterants, and subsequently, to all the acquired hyperspectral images (*see* **Sections 2.3.2.1** and **2.3.2.4**, respectively).

2.3.2.1. Dataset structure

Firstly, we developed *pixel-level* models able to classify genuine oregano and pure adulterants. To this aim, we built-up a dataset of representative spectra belonging to both classes. This phase is crucial as it determines the representativeness of spectra references for the two classes, thus affecting the robustness and reliability of the classification models.

To this aim, a PCA model was calculated on the mean centered spectra of each image considering only the pixels belonging to the sample and selecting 3 principal components (PCs). Then, the pixels outside the 99.9% confidence limit on both Hotelling T^2 and Q residuals values were excluded. Indeed, a deeper investigation these few pixels allowed to observe that they were ascribable to specular reflections or to small portions of the glass Petri dish that were not removed during background segmentation and erosion. Finally, a new PCA model was calculated, and Kennard-Stone algorithm [43] was applied in the PC space to select a representative number of pixel spectra. In particular, 600 spectra were selected from each image of pure adulterants, resulting in a total of 7200 representative spectra collected for the pure adulterants class (= 600 spectra × 12 hyperspectral images), while 100 pixel spectra were selected from each image of authentic oregano samples, for a total of 7200 representative spectra (= 100 spectra × 72 hyperspectral images). Note that two authentic oregano samples, listed as # 9 and # 10 in **Table 1,** were excluded in this phase due to the presence of branch fragments, but they were used for the final validation of the classification models (*see* **Section 2.3.2.4**). This dataset of representative spectra included therefore an overall number of 14400 spectra, and it was then used to develop the classification models.

Before calculating the classification models, the dataset was split into a training (TR) set used for model calculation and a test (TS) set used for model validation. The samples of authentic oregano were randomly subdivided in training and test samples with a ratio of 2/3 and 1/3, respectively, and the corresponding spectra were then assigned to the TR or TS dataset accordingly. Considering the pure adulterants, the subdivision into TR and TS sets was based on acquisition replicates: for each pure adulterant sample, the spectra of two of the three aliquots were included in the TR and one in the TS dataset. Therefore, the composition of TR and TS datasets can be summarised as follows:

- TR: 9600 spectra in total, 4800 spectra selected from 48 images of authentic oregano samples and 4800 spectra selected from 8 images of pure adulterants;
- TS: 4800 spectra in total, 2400 spectra selected form 24 images of authentic oregano samples and 2400 spectra selected from 4 images of pure adulterants.

Figure 1 reports the mean spectrum of the selected pixel spectra of authentic oregano belonging to the TR set together with the mean spectra of the different pure adulterants considered in this study.

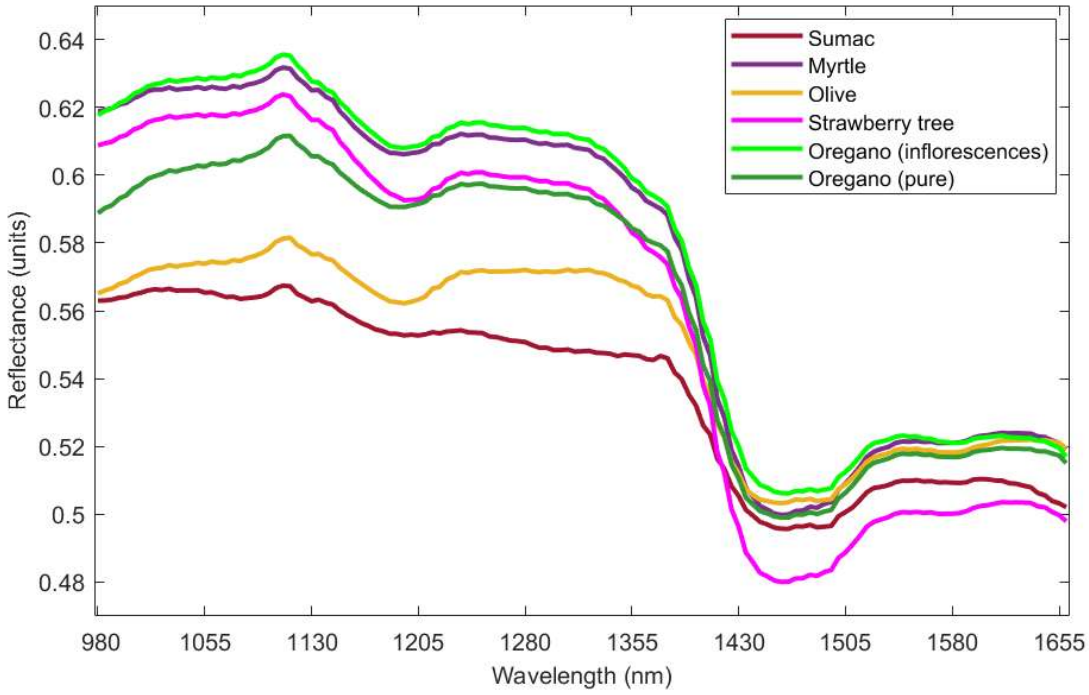


Figure 1 *Average spectrum of selected pixel spectra belonging to the TR set of authentic oregano class together with the average spectrum of each pure adulterant.*

2.3.2.2. Spectra classification by SIMCA

Under a CM perspective, the authentication issue of this study can be considered a one-class classification problem. In fact we are interested in defining the boundaries of a single target class,

i.e., authentic oregano, and predicting if a new sample belongs or not to the target class. Therefore, we developed a one-class SIMCA model considering only the spectra of authentic oregano of the TR set, while the spectra of pure adulterants of the TR set were used during cross-validation following a compliant approach [44].

SIMCA algorithm models the similarities among samples of the target class (i.e., authentic oregano), assuming that the main features of the target class can be represented by a Principal Component (PC) space of adequate dimensionality, commonly known as class subspace. Class assignment of new observations is carried out by calculating two statistical metrics accounting for the distance between the new observation and the target class subspace: the Orthogonal Distance (OD) and the Score Distance (SD). OD is the squared Euclidean distance of each new observation from its projection into the PCA model, while SD is defined as the squared Mahalanobis distance between the projection of the sample into the PCA subspace and the origin of the PCs [45].

Afterwards, OD and SD values are usually compared with the corresponding critical limits (OD_{crit} and SD_{crit}, respectively) at a defined confidence level to perform the final assignment of the new observation. Different versions of SIMCA algorithm have been developed based on how OD and SD metrics as well as the corresponding confidence limits are used to perform class assignment [31].

In this study, we used Alternative SIMCA algorithm (Alt-SIMCA), which combines OD and SD values in a single statistical parameter, *d*, that defines the limits of the acceptance subregion (**Equation 4.1**) [31, 46]:

$$d = \sqrt{\left(\frac{OD}{OD_{crit}}\right)^2 + \left(\frac{SD}{SD_{crit}}\right)^2} \quad (4.1)$$

where OD_{crit} and SD_{crit} values correspond to the critical limits at 95% confidence level. Only the observations with $d \leq \sqrt{2}$ are assigned to the target class, while those not meeting this decision rule are rejected and defined as non-target samples.

The optimal number of PCs was selected by maximising the cross-validation efficiency (as defined in **Section 2.3.2**) following the compliant strategy for one-class modelling, which consists in using also samples not belonging to the target class for model optimisation. This strategy is generally recommended when dealing with overlapping classes [44]. In particular, the authentic oregano samples of the TR set were randomly split into two deletion groups and, based on this subdivision, the corresponding spectra were assigned to the two groups; in this manner, spectra selected from replicate images of the same oregano sample were kept in the same deletion group. Conversely, the TR set spectra belonging to pure adulterants were divided into the two deletion groups based on replicates., i.e., the spectra of one of the two replicate images of the TR set were assigned to one

deletion group, while the spectra of the other replicate image were assigned to the other deletion group.

Afterwards, the classification performance of the Alt-SIMCA model was assessed though external validation using the TS set, which contains spectra belonging to both authentic oregano and pure adulterants.

Alt-SIMCA model was calculated using routines written *ad hoc* in MATLAB environment (ver. 2020b, The MathWorks, USA) based on PLS Toolbox functions (ver. 8.5, Eigenvector Research Inc., USA). The reader is referred to Vitale et al. [31] for an in-depth description of the Alt-SIMCA algorithm.

2.3.2.3. Spectra classification by Soft PLS-DA

The TR set was used to calculate a classification model to discriminate between genuine oregano and pure adulterants by means of Soft PLS-DA. Soft PLS-DA is a soft discriminant algorithm that combines the advantages of discriminant analysis and class modelling approaches. Its configuration allows for increased flexibility and robustness in classification models: by applying additional constraints for class assignment, it effectively identifies possible outliers. In this manner, Soft PLS-DA overcomes the limitations of PLS-DA in handling new objects not belonging to the target classes, maximizing at the same time the discrimination between the classes of interest [36, 39].

In details, a new sample is assigned to a defined class according to the following criteria:

- it must have Q residuals values falling within the 99.9% confidence limit of the model. This limit has been chosen to set boundaries wide enough to consider as much as possible within classes variability, but allowing at the same time to exclude samples with a very poor fit to the model;
- it must have *y* predicted values falling within an acceptability range for the considered class. The lower limit is defined by the PLS-DA threshold value for the class under investigation, while the upper limit allows for the rejection of objects located at the extremes of the Gaussian probability density function;
- for multiclass classification, the samples must be unambiguously assigned to only one class.

The samples that do not match all the three criteria defined by the Soft PLS-DA decision rules are not assigned to any class and automatically labelled as “not assigned” samples (NA).

Soft PLS-DA model was optimized by using the same custom cross-validation scheme and samples splitting criterion previously mentioned in **Section 2.3.2.2**. Furthermore, external validation was performed by predicting class assignment of the samples belonging to the TS set.

Soft PLS-DA model was calculated using routines written *ad hoc* in MATLAB environment (ver. 2020b, The MathWorks, USA). The MATLAB routine to run Soft PLS-DA algorithm [39] is freely downloadable from http://www.chimslab.unimore.it/downloads/. The reader is referred to Calvini et al. [39] for a detailed description of the Soft PLS-DA algorithm.

2.3.2.4. Validation on external images

Alt-SIMCA and Soft PLS-DA models, obtained as previously described in **Section 2.3.2.2** and **Section 2.3.2.3**, were applied to all the acquired hyperspectral images, including the images of adulterated oregano samples. The resulting prediction images, in which each pixel is coloured according to the class assignment of the corresponding spectrum, were used to directly visualize the prediction performance on the images and obtain a quantitative evaluation of the classification performances on the entire set of images. To this aim, the percentage of pixels predicted as authentic oregano (PPO%) was calculated for each prediction image.

Finally, to assess the overall ability of the classification model to differentiate between authentic and adulterated oregano, we defined a threshold value based on the percentage of pixels predicted as pure oregano in each image. Samples whose images had PPO% values higher than the threshold were considered as authentic oregano samples, while samples whose images had PPO% values lower than the threshold were considered as adulterated. This threshold value was calculated as the minimum PPO% value obtained for the images of pure oregano belonging to the training set (*see* **Section 3.4**).

3. Results and discussion

3.1. Exploratory Analysis at the pixel-level

An exploratory analysis at the *pixel-level* was performed to evaluate the spectral differences between authentic oregano, pure adulterants (myrtle leaves, sumac leaves, strawberry tree leaves, and olive leaves) and adulterated oregano. To this aim, for each adulterant type a unique hyperspectral image was obtained by merging together one image of the pure adulterant, one image of an authentic oregano sample and two images of oregano adulterated with the corresponding adulterant at different percentages. The merged hyperspectral images were then analysed by PCA.

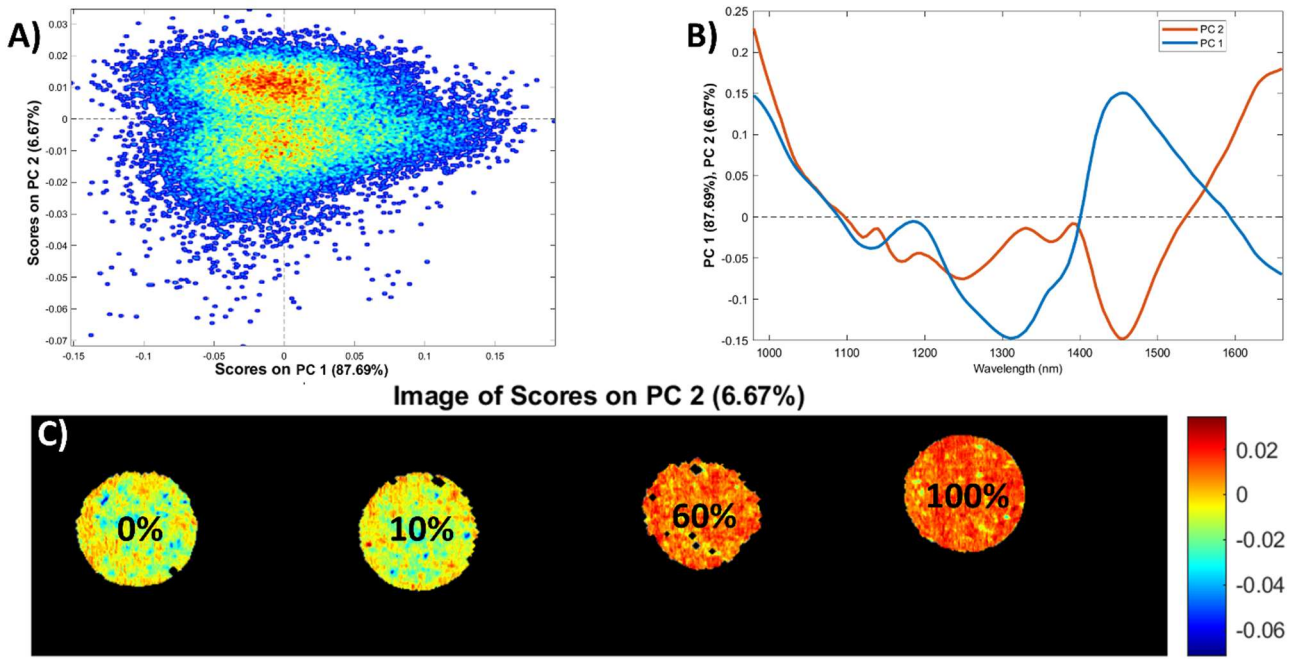


Figure 2. *Principal component analysis (PCA) results of the merged hyperspectral image containing one authentic oregano sample (0% of adulteration), two samples adulterated with different percentages of myrtle leaves (10% and 60% of adulteration) and one sample of pure myrtle leaves (100% of adulteration). In (A) PC1-PC2 score plot; in (B) PC1 and PC2 loading vectors and in (C) PC2 score image.*

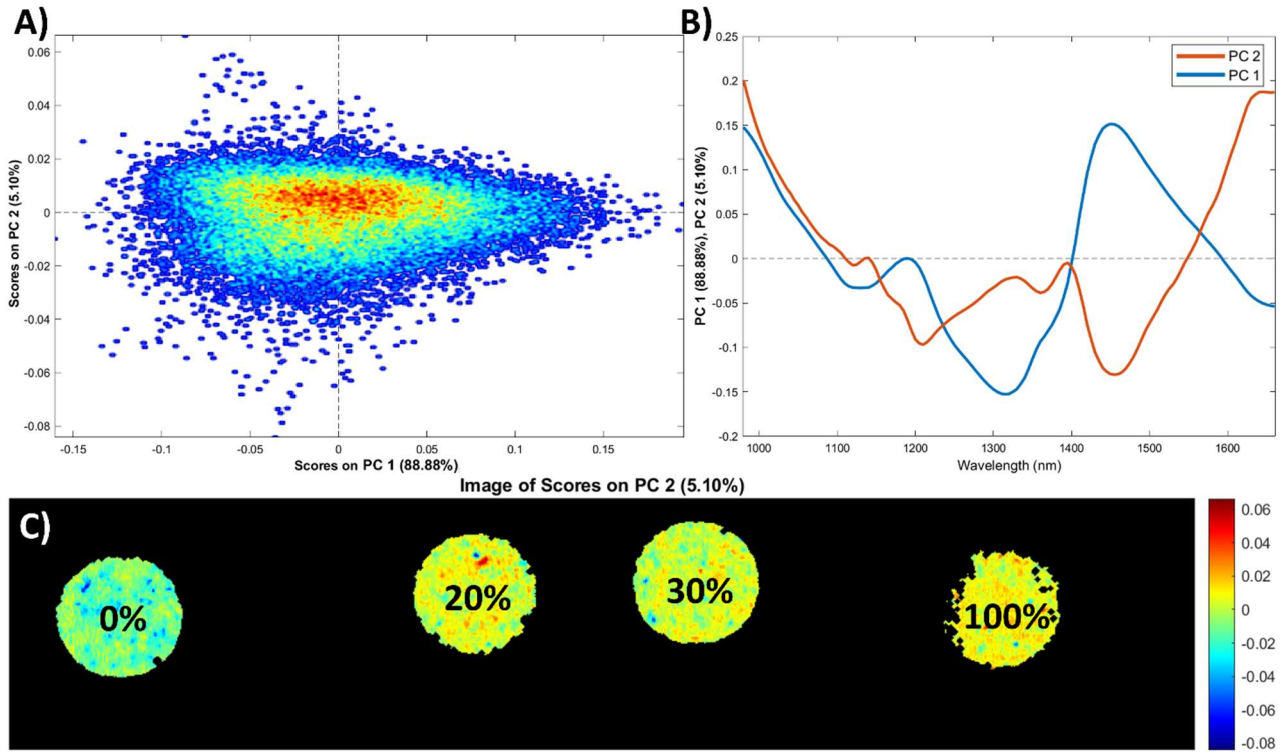


Figure 3. *PCA results of the merged hyperspectral image containing one authentic oregano sample (0% of adulteration), two samples adulterated with different percentages of strawberry tree leaves (20% and 30% of adulteration) and one sample of pure strawberry tree leaves (100% of adulteration). In (A) PC1-PC2 score plot; in (B) PC1 and PC2 loading vectors and in (C) PC2 score image.*

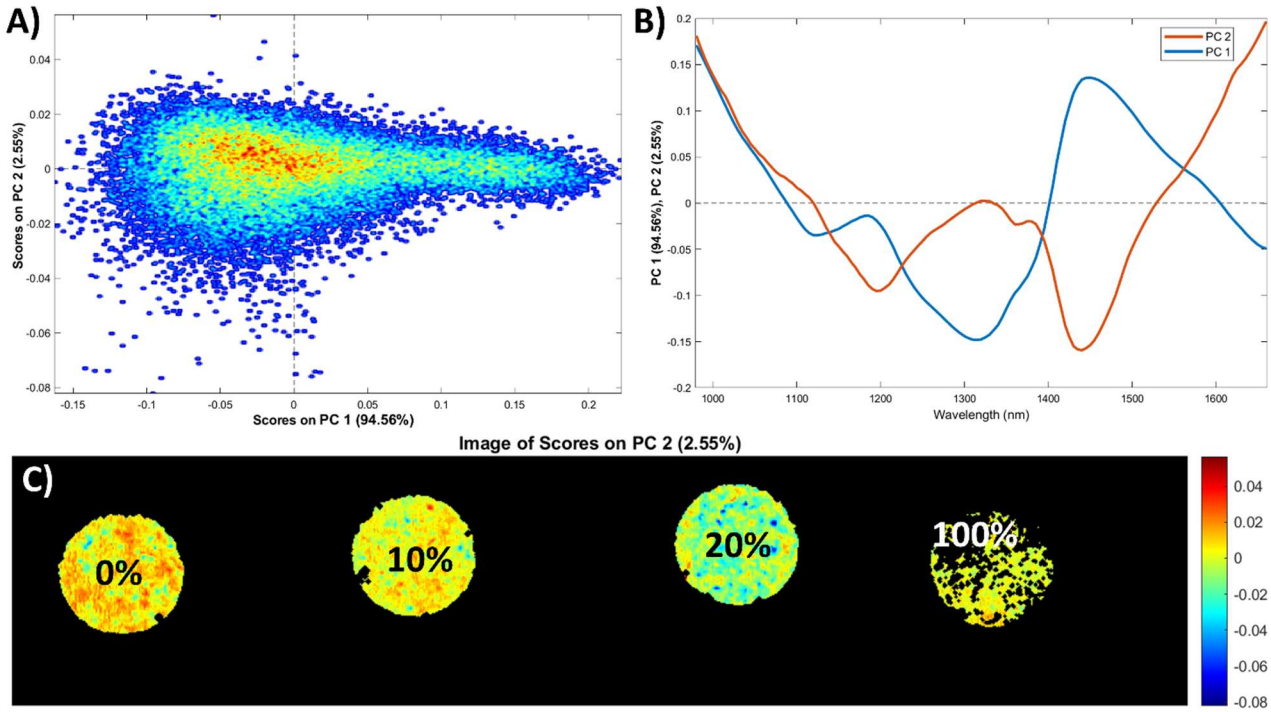


Figure 4. *PCA results of the merged hyperspectral image containing one authentic oregano sample (0% of adulteration), two samples adulterated with different percentages of sumac (10% and 20% of adulteration) and one sample of pure sumac (100% of adulteration). In (A) PC1-PC2 score plot, in (B) PC1 and PC2 loading vectors and in (C) PC2 score image.*

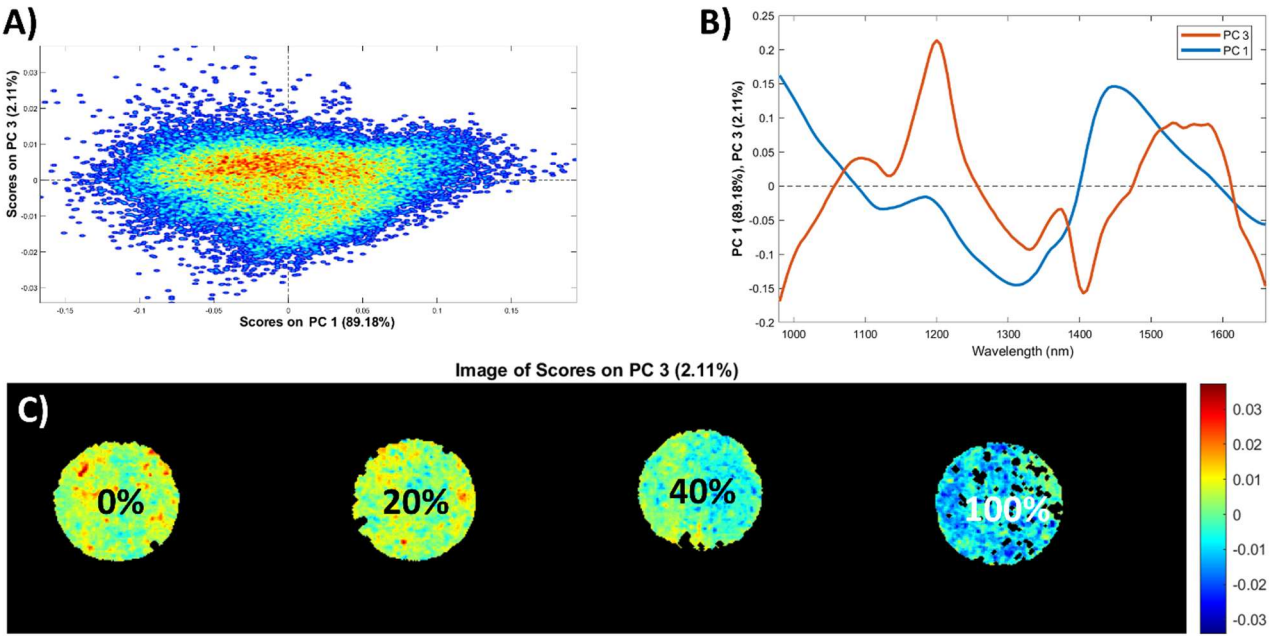


Figure 5. *PCA results of the merged hyperspectral image containing one authentic oregano sample (0% of adulteration), two samples adulterated with different percentages of olive leaves (20% and 40% of adulteration) and one sample of pure olive leaves (100% of adulteration). In (A) PC1-PC3 score plot, in (B) PC1 and PC3 loading vectors and in (C) PC3 score image.*

Figure 2 reports the results of the PCA model calculated on the merged image of an authentic oregano sample (47_04, 0% myrtle leaves), of two oregano samples adulterated with myrtle leaves (48_29, 10% myrtle leaves, and 48_04, 60% myrtle leaves) and of the pure adulterant (A_05, 100% myrtle leaves).

Figure 2 A shows the PC1-PC2 score plot, where the first two principal components account for 87.69% and 6.67% of total variance, respectively, whereas **Figure 2 B** reports the corresponding loading vectors. In the score plot each object represents a single pixel and it is coloured according to pixel density, i.e., red colour represents a region of the PC1-PC2 score space with a high density of pixels, while blue colour corresponds to low pixel density. From this score plot it is possible to observe the presence of two clusters of pixels, separated along PC2. The PC2 score image reported in **Figure 2 C** shows that the differences observed along PC2 are ascribable to the spectral differences between authentic oregano and myrtle leaves. Indeed, the pixels of the authentic oregano sample (0% of adulteration) are mainly characterised by low PC2 score values, while the pixels of the myrtle sample (100% of adulteration) show generally high PC2 score values. Note that the adulterated samples show an intermediate behaviour somehow proportional to the percentage of adulteration. In fact, the oregano sample adulterated with 10% of myrtle leaves has PC2 score values comparable to that of authentic oregano, whereas the oregano sample with 60% of adulteration generally presents positive PC2 score values, slightly lower than those of the image of the myrtle leaves.

Similar results were obtained from the PCA model calculated on the merged hyperspectral image for strawberry tree leaves as adulterant (**Figure 3**). Also in this case, PC2 allows separating the pixel spectra of the authentic oregano sample from those of sample with only strawberry tree leaves, and the oregano samples adulterated with 20% and 30% of strawberry tree leaves show an intermediate behaviour. The PCA model calculated on the merged hyperspectral image considering sumac as adulterant (**Figure 4**) provides comparable results to those previously discussed for myrtle and strawberry tree leaves. Finally, the results of the investigation regarding olive leaves as adulterant are reported in **Figure 5**. In this case, the direction which better reflects the differences between the images according to the percentage of adulterant is represented by PC3.

Therefore, the PCA models calculated on the merged images allowed capturing the presence of detectable spectral differences between authentic oregano and the adulterants investigated in this study. Considering the loading vectors of the relevant PCs for this separation (mainly PC2 for myrtle leaves, strawberry tree leaves and sumac, and PC3 for olive leaves as adulterants), it is possible to identify some common spectral regions that contribute to these findings in **Figures 2 B**, **3 B** and **5 B**. These spectral regions fall into the 980-1080 nm spectral range, corresponding to O−H third overtone and C−H second overtone, associated with polyphenol content, in the 1150-1200 nm spectral range

corresponding to asymmetric stretching of C–H second overtone, ascribable to alcohols, in the 1420-1450 nm, ascribable to O−H stretch first overtone, C=O stretch third overtone and N-H stretch first overtone which could be ascribable to terpenoids and cellulose content, and in the 1620-1660 nm region, ascribable to the stretching of aromatic C−H first overtone [21, 47-51]. Additional spectral regions that codify for myrtle, strawberry tree and sumac can be found around 1250 nm and 1380 nm, related to alcohol and methyl groups [49].

3.2. Exploratory Analysis at the *image-level*

As mentioned in **Section 2.3.1**, the whole image dataset was also evaluated at the *image-level* to gain an insight on sample characteristics and behaviour. To this aim, the average spectrum was obtained from each image and a global PCA model was calculated on the average spectra dataset using linear detrend and mean center as preprocessing methods.

Since in the previous evaluation performed at the *pixel-level* (*see* **Section 3.1**) we observed detectable differences between the spectral signatures of authentic oregano and pure adulterants, the *image-level* analysis was focused on identifying possible trends due to adulterant type and amount. Therefore, a PCA model was calculated on the average spectra of the sole authentic and adulterated oregano. The resulting PC1-PC2 score plot is reported in **Figure 6**, accounting for 92.06% of explained variance. In **Figure 6 A** the samples in the score plot are coloured according to authentic or adulterated class. It is worth noting that the oregano samples have a wide chemical variability and morphological heterogeneity, probably due to the different geographical origins and multiple harvest years. Furthermore, the two classes of authentic and adulterated oregano samples are partly overlapped. Only a limited separation of some adulterated samples, characterised by extreme (both positive and negative) PC2 score values was observed. A more in-depth investigation, based on adulterant type and percentage (**Figure 6 B**), revealed that the adulterated samples showing more marked differences from pure oregano were those characterized by adulteration percentages equal or higher than 60%, 30% and 20% with myrtle leaves, olive leaves and strawberry tree leaves, respectively.

These findings confirm the results of the PCA models calculated at the *pixel-level*, where adulterated samples with percentages lower than 20% showed similar behaviour to authentic oregano samples.

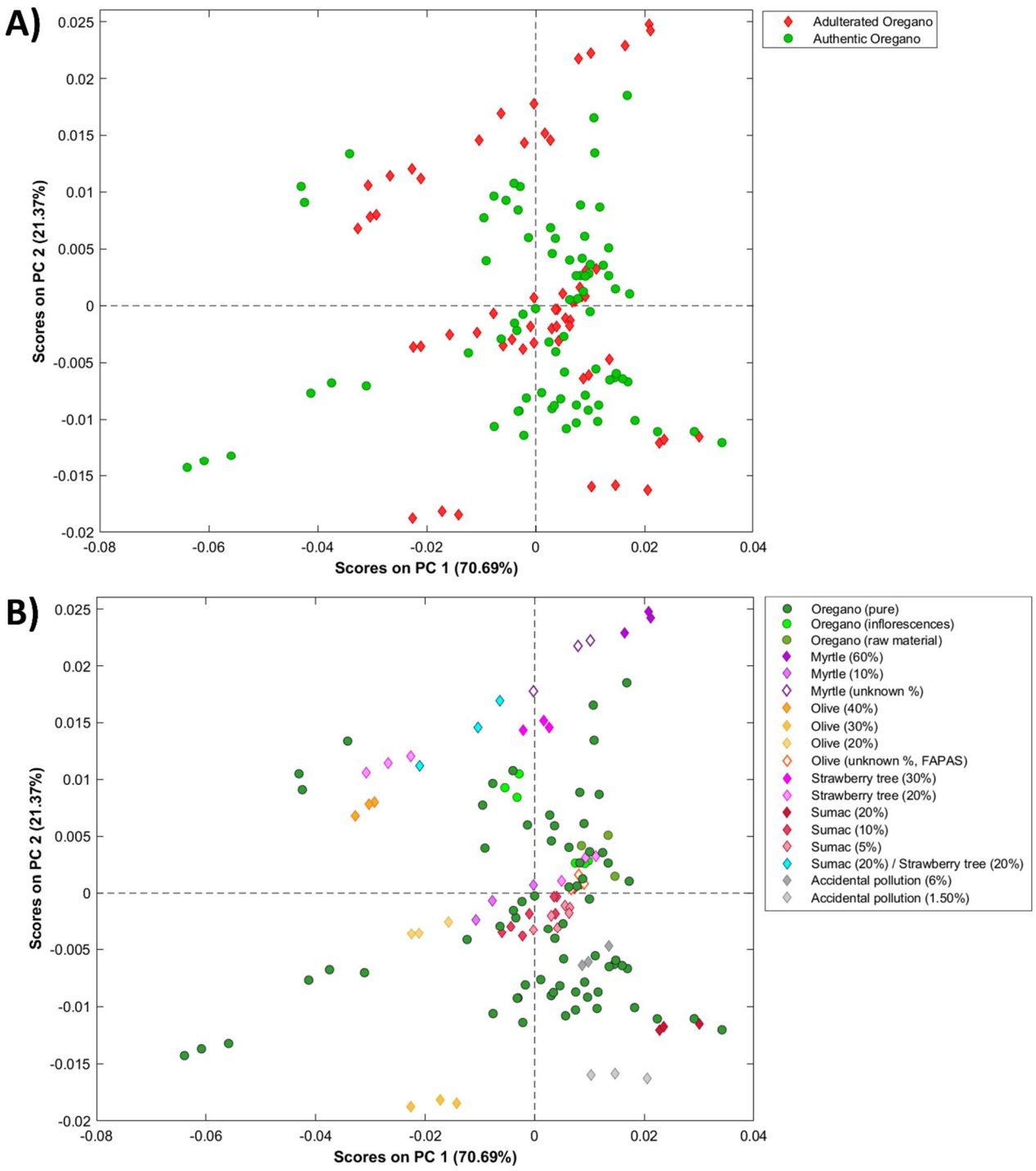


Figure 6 *PC1-PC2 score plot obtained by calculating a PCA model considering the authentic oregano (circle) and adulterated oregano (rhombus) samples average spectra. In (**A**) samples are coloured according to authentic and adulterated class; in (**B**) samples are coloured based on adulterant type and percentage.*

3.3. Classification between authentic oregano and pure adulterants

The first step in the identification of adulterated oregano samples consists in the development of *pixel-level* classification models able to distinguish genuine oregano from pure adulterants. **Table 2** reports the results obtained in calibration (CAL), cross-validation (CV) and validation of the external test set (TS) of the Alt-SIMCA and Soft PLS-DA models. The classification performances were evaluated by calculating SENS, SPEC and EFF values: for Alt-SIMCA, the results refer to authentic oregano, which is the target class of the model, while for Soft PLS-DA the results of both pure adulterants and authentic oregano classes are reported. The results in **Table 2** clearly show that the Soft PLS-DA model achieved good classification performances, with SENS and SPEC values for both classes higher than 90%.

		Alt-SIMCA	**Soft PLS-DA**	
		Authentic Oregano	**Pure Adulterants**	**Authentic Oregano**
	PCs/LVs	6	7	
Calibration (CAL)	**SENS (%)**	94.9	92.4	93.5
	SPEC (%)	-	94.3	93.8
	EFF (%)	**-**	**93.4**	**93.7**
	NA (%)	-	1.4	0.8
Cross-validation (CV)	**SENS (%)**	90.9	91.7	90.4
	SPEC (%)	50.8	91.5	93.7
	EFF (%)	**67.9**	**91.4**	**92.0**
	NA (%)	-	2.4	1.1
Prediction (TS)	**SENS (%)**	96.9	90.8	93.6
	SPEC (%)	47.1	94.1	92.3
	EFF (%)	**67.5**	**92.4**	**92.9**
	NA (%)	-	1.5	0.5

***Table 2** Classification performances of Alt-SIMCA and Soft PLS-DA models in calibration (CAL), cross-validation (CV) and prediction of the external test set (TS).*

Interestingly, about 39% of misclassified pixel spectra of authentic oregano class (i.e., spectra of authentic oregano class but predicted as pure adulterants by Soft PLS-DA model) in cross-validation belong to genuine oregano with inflorescences. This finding suggests that the presence of inflorescences may negatively affect the classification performances.

Conversely, despite the Alt-SIMCA model reached excellent SENS values both in cross-validation and TS set prediction, its ability to correctly reject not-authentic oregano samples is unsatisfactory, as indicated by SPEC values around 50%. The poor classification performances obtained with Alt-SIMCA can be explained considering the significant within-class variability both for authentic oregano and for pure adulterants, resulting in partial overlap between the two classes. In this context, class modelling approaches are generally not effective.

A more in-depth evaluation of the classification results was also performed, based on adulterant type. For this reason, considering Alt-SIMCA algorithm, **Table 3** reports for each adulterant type the percentage of spectra accepted or rejected by the authentic oregano class model. Similarly, **Table 3** also reports the results obtained for Soft PLS-DA, expressed as percentage of spectra assigned by the model to authentic oregano class, pure adulterants class and not assigned spectra. For both models, the results reported in **Table 3** are referred to cross-validation and prediction of the TS set, while for Soft PLS-DA model the results obtained also in calibration are reported in **Table 4**.

Alt-SIMCA provided satisfactory classification performances only for olive leaves, achieving a percentage of correctly rejected spectra of 79.7% for TS set prediction. Conversely, overall poor classification performances were obtained for the other adulterant types. Specifically, approximately half of the spectra belonging to strawberry tree leaves and sumac were wrongly accepted by the authentic oregano class model, and the same applies to the vast majority of myrtle spectra.

Concerning Soft PLS-DA, the best performances were obtained for strawberry tree leaves and sumac, with a percentage equal to or less than 1.7% of spectra misclassified as genuine oregano both in cross-validation and prediction of the test set. Conversely, higher misclassifications were obtained for myrtle and olive leaves, where the percentage of correctly assigned spectra ranged between 82.7% and 89.1%. Therefore, the Soft PLS-DA model better recognized sumac and strawberry tree leaves as adulterants; however, the results obtained for myrtle and olive leaves can still be considered satisfactory.

			Myrtle	Olive	Strawberry tree	Sumac
Alt-SIMCA	**CV**	**Authentic oregano (%)**	83.9	20.2	45.7	47.1
		Not Authentic oregano (%)	**16.1**	**79.8**	**54.3**	**52.9**
	TS	**Authentic oregano (%)**	89.0	20.3	52.0	50.3
		Not Authentic oregano (%)	**11.0**	**79.7**	**48.0**	**49.7**
Soft PLS-DA	**CV**	**Authentic oregano (%)**	14.7	7.9	1.1	1.7
		Pure adulterants (%)	**84.8**	**89.1**	**97.1**	**94.2**
		NA (%)	0.6	3.0	1.8	4.2
	TS	**Authentic oregano (%)**	17.2	11.7	0.7	1.5
		Pure adulterants (%)	**82.7**	**83.8**	**98.7**	**97.8**
		NA (%)	0.2	4.5	0.7	0.7

Table 3. *Classification performances of pure adulterants in cross-validation (CV) and prediction of the test set (TS). Alt-SIMCA: for each adulterant type the percentage of spectra accepted or rejected by the authentic oregano class model is reported. Soft PLS-DA: for each adulterant type the percentage of spectra predicted as authentic oregano, pure adulterants or not assigned (NA) is reported.*

			Myrtle	Olive	Strawberry tree	Sumac
Soft-PLSDA	CAL	Authentic oregano (%)	13.0	9.2	1.1	1.4
		Not Authentic oregano (%)	**86.4**	**88.8**	**97.4**	**97.2**
		NA (%)	0.6	2.1	1.5	1.4
	CV	Authentic oregano (%)	14.7	7.9	1.1	1.7
		Pure adulterants (%)	**84.8**	**89.1**	**97.1**	**94.2**
		NA (%)	0.6	3.0	1.8	4.2
	TS	Authentic oregano (%)	17.2	11.7	0.7	1.5
		Pure adulterants (%)	**82.7**	**83.8**	**98.7**	**97.8**
		NA (%)	0.2	4.5	0.7	0.7

Table 4. *Classification performances for the pure adulterants class in calibration (CAL), cross-validation (CV) and prediction of the test set (TS) using Soft PLS-DA. For each adulterant type the percentage of spectra predicted as authentic oregano, pure adulterants or not assigned (NA) is reported.*

In order to evaluate the spectral regions that contribute the most to the identification of authentic oregano, **Figure 7** reports the Variable Importance in Projection (VIP) scores of the Soft PLS-DA model. In particular, the spectral variables with VIP scores higher than 1 (red dashed line in **Figure 7**) are those with higher relevance for the classification model. These variables fall into the intervals at 980-1010 nm (O−H third overtone and C−H second overtone), 1130–1150 nm (C–H second overtone and stretching of C=O fourth overtone), 1180-1225 nm (asymmetric stretching of C−H second overtone), 1390-1420 nm (stretching of O−H first overtone for ROH and ArOH), 1445-1470 nm (stretching of O−H first overtone for water, stretching of C=O third overtone, stretching of N−H first overtone) and 1640-1650 nm (stretching of aromatic C−H first overtone). The wavebands related to C–H and C=O absorption could be associated to polyphenols and terpenoids, whereas the wavebands related to O−H and N−H absorption can be associated to water, cellulose, hemicellulose and lignin. In addition, the relevance of O−H and aromatic C–H can be associated with hydroxyl and aromatic groups, particularly present also in polyphenols. Therefore, the discriminant wavebands can be associated to differences in the aromatic content, in terms of polyphenols, terpenoids, esters and alcoholic groups, and to cellulose, hemicellulose and lignin, which are ubiquitous in plant tissues [21, 47-51].

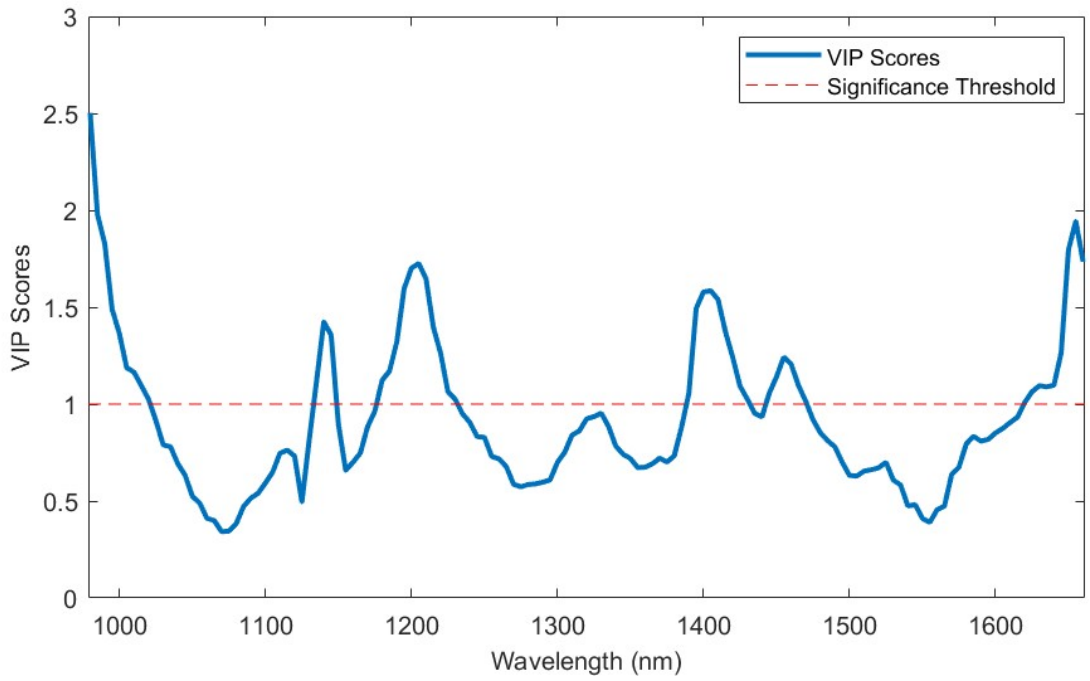


Figure 7. *VIP scores of the Soft PLS-DA model for the discrimination of authentic oregano and pure adulterants.*

3.4. Validation on external images and identification of adulterated samples

Alt-SIMCA and Soft PLS-DA classification models were applied to all the acquired hyperspectral images. For Soft PLS-DA, some representative prediction images are reported in **Figure 8**, where the pixel spectra predicted as authentic oregano are represented in green colour, the pixel spectra predicted as pure adulterant are reported in red colour while the not assigned pixels are reported in grey colour.

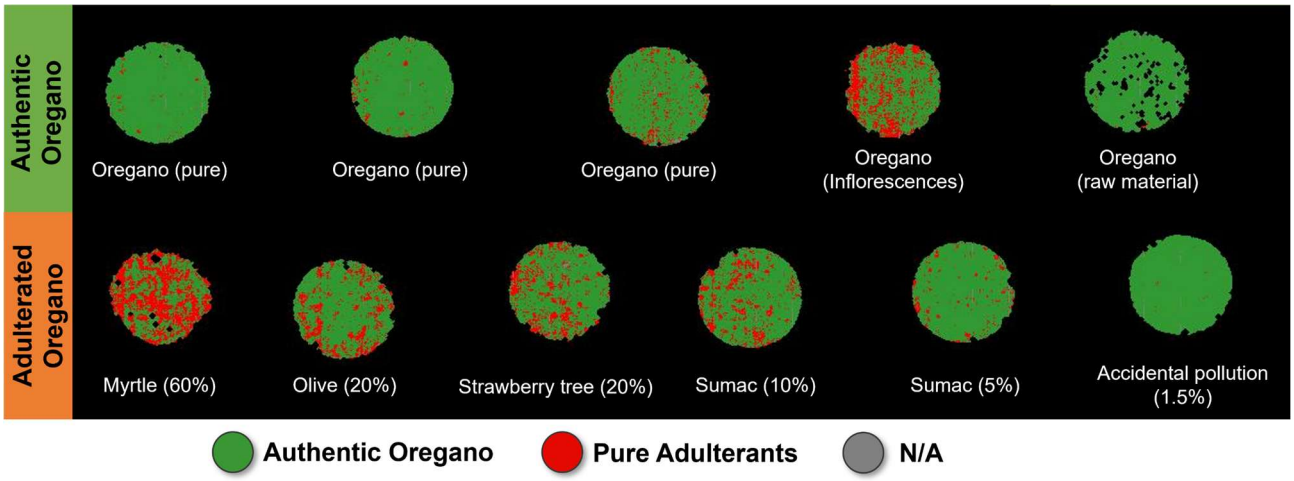


Figure 8. *Prediction images obtained by applying the Soft PLS-DA model to hyperspectral images of authentic and adulterated oregano samples. Prediction images of pure oreganos, one oregano sample with inflorescences and one sample of raw oregano are reported in the first row, whereas the prediction images of adulterated oregano samples at decreasing percentages of adulteration (from left to right) are reported in the second row.*

Specifically, the first row of images in **Figure 8** reports the prediction images obtained from five authentic oregano samples, including one sample with inflorescences. Note that almost all the pixels belonging to the pure oregano samples were correctly predicted as authentic oregano and a few misclassifications are ascribable to an intrinsic error of the classification model. The oregano sample containing inflorescences represents an exception, since it has a high number of misclassified pixels; this fact confirms what already observed in **Section 3.3**, where a relevant number of misclassified spectra of genuine oregano belonged to samples with inflorescences.

In the second row of **Figure 8** six prediction images, calculated on adulterated oregano samples, are shown: from left to right, the images are reported at decreasing concentrations of adulterants. In this case, the samples with adulterant concentrations equal to or higher than 10% have a relevant number of pixels predicted as adulterants and their amount is roughly proportional to the adulterant concentration. On the other hand, the prediction images of oregano samples adulterated at percentages lower than 10% have a number of pixels predicted as adulterant comparable to or even smaller than the number of misclassified pixels of those of the pure oregano.

In order to perform a global evaluation of the prediction ability of Alt-SIMCA and Soft PLS-DA classification models, the percentage of pixels predicted as oregano class (PPO%) was calculated for each prediction image obtained with both methods. The scatter plot in **Figure 9** shows the relationship between the actual oregano percentage in the analysed sample aliquots and the PPO% values obtained from the corresponding prediction images, calculated by applying Alt-SIMCA (**Figure 9 A**) and Soft PLS-DA (**Figure 9 B**) models.

The plot with Alt-SIMCA results (**Figure 9 A**) reveals that images of authentic oregano correctly present PPO% values higher than 90%. However, also all the images of adulterated samples have PPO% values around 90%, regardless of the actual adulterant concentration. High PPO% values are also evident for the pure adulterants, particularly for myrtle which has PPO% values around 90%. These results confirm the low specificity of Alt-SIMCA model, i.e., its poor ability of correctly rejecting samples not belonging to the authentic oregano class.

Conversely, the plot related to Soft PLS-DA results (**Figure 9 B**) shows a discrete correlation between actual oregano content and the percentage of pixels predicted as authentic oregano extracted from the prediction images. Concerning the pure adulterants, the PPO% values were found to be in the 0-10% range. Except for sample 47_08 and sample 47_15 containing inflorescences (*see* **Table 1**), the images of authentic oregano showed PPO% values around 90% or higher. In this case, the adulterated samples have PPO% values that are generally proportional to the actual concentration of pure oregano. Except for one image of a sample adulterated with 30% olive leaves, only the samples adulterated with less than 10% of adulterant show PPO% values higher than 90%, comparable to those of authentic oregano.

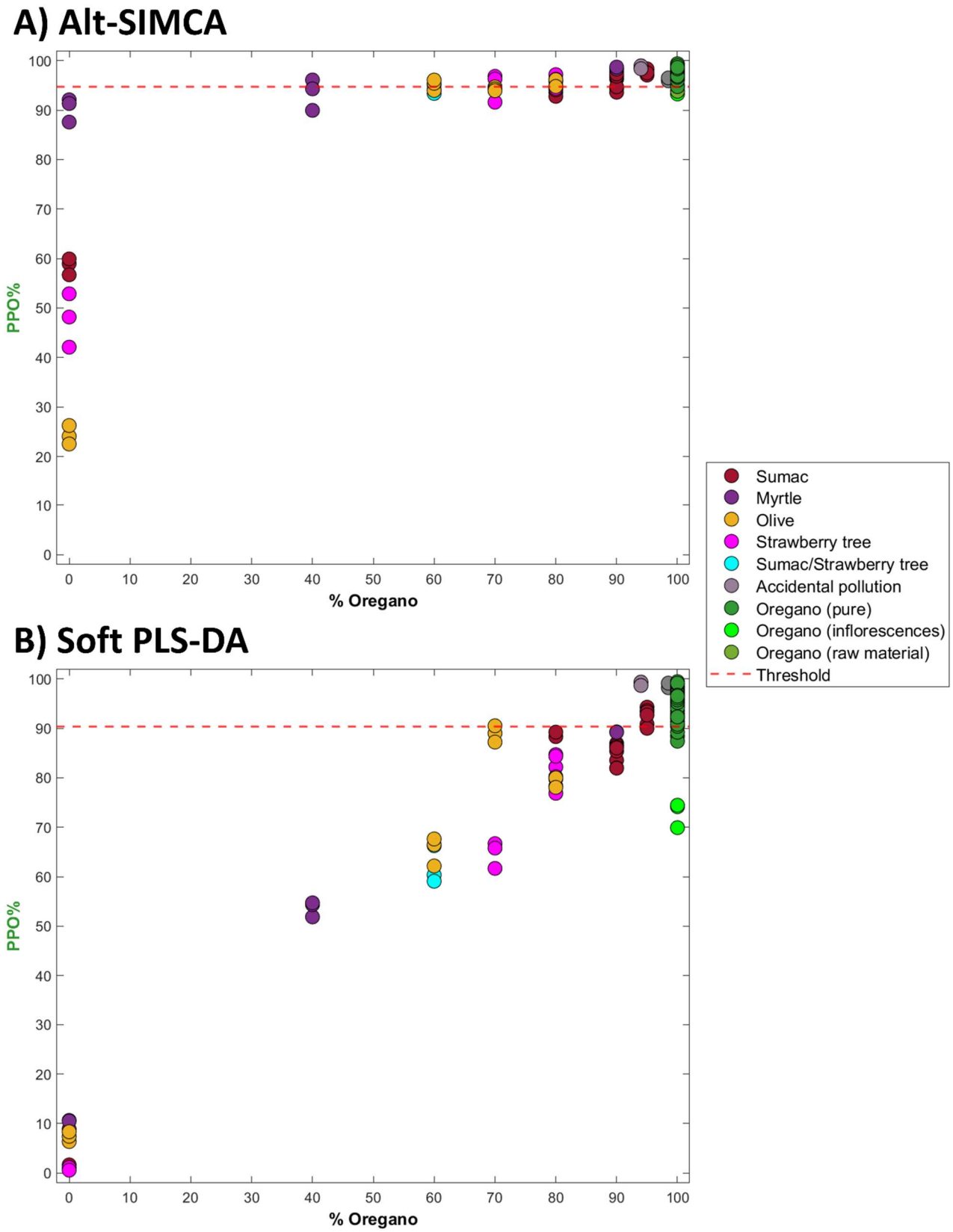


Figure 9 *Actual oregano concentration vs. percentage of pixels predicted as oregano (PPO%) by the Alt-SIMCA (A) and Soft PLS-DA (B) models. The red dashed line represents the threshold value used to discriminate images of authentic oregano from images of adulterated oregano.*

Since the final goal of this study was the identification of authentic and adulterated oregano samples, we decided to define a threshold value based on PPO% value to assign the samples to one of the two classes. This threshold value was defined as the minimum PPO% value obtained for the hyperspectral images of authentic oregano samples belonging to the training set (*see* **Section 2.3.2.1**). The threshold was set at PPO% values equal to 94.74% and 90.41% for Alt-SIMCA and Soft PLS-DA, respectively. Therefore, the hyperspectral images, whose prediction images have PPO% values equal or higher than the threshold values, were assigned to the authentic oregano class whereas those with a lower value were classified as adulterated.

The resulting outcomes are reported as confusion matrix (**Table 5**), where the columns represent the actual classes, and the rows are the assigned classes. **Table 5** shows the results obtained only for the images used as external validation, i.e., the authentic oregano images used as test images and all the images of adulterated samples. For ease of interpretation, the results of the images of adulterated oregano were reported after splitting them in three categories based on adulterant concentration: images of samples with adulterant concentration lower than 10% (adult. < 10%), with adulterant concentration equal to or higher than 10% (adult. ≥ 10%) and with unknown concentration.

Concerning the ability of the models in correctly recognising authentic oregano samples, the performances are comparable: Alt-SIMCA correctly classified 29 images of authentic oregano out of 33, which corresponds to a SENS value of 87.9%, while Soft PLS-DA correctly classified 27 images out of 33, which coincide with a SENS value of 81.8%. In both cases, the misclassified images of authentic oregano include 3 replicates of two oregano samples with inflorescences, confirming that the presence of inflorescences can negatively affect the classification performances.

Conversely, the two models have a different ability to identify adulterated oregano samples. Concerning Alt-SIMCA, only 16 adulterated oregano images out of 57 (28.1%) were correctly recognised.

On the other hand, Soft PLS-DA correctly attributed 42 adulterated oregano images out of 57 to the corresponding class (73.7%). Furthermore, among adulterated oregano there is a clear difference in the classification performances based on adulterant concentration. As expected, 11 out of 12 images with adulterant concentration lower than 10% were erroneously assigned to the authentic class, while 38 out of 39 images with adulterant concentration equal to or higher than 10% were correctly classified as adulterated oregano. The misclassified image is a replicate of one adulterated sample whose remainder replicates were correctly classified. Supposing a sample-based classification by majority voting of the assignments done on the three replicated images of each sample, we can state that all the samples with adulterant concentration equal to or higher than 10% were correctly identified by Soft PLS-DA.

The image-level classification based on the Soft PLS-DA results on the one hand confirmed the preliminary findings obtained by exploratory data analysis, and on the other hand allowed to better identify the minimal spectral differences between pure adulterants and authentic oregano. In accordance with previous studies [16], NIR-HSI is affected by a detection limit of 10% adulteration. According to the European Spice association [9], 2% of extraneous matter is tolerated; however, a detection limit of 10% may still be considered acceptable, since oregano adulteration levels are generally higher that this value to lead to a concrete economic advantage.

		Actual class			
		Authentic oregano	**Adulterated oregano**		
			adult. < 10%	adult. ≥ 10%	unknown
Assigned class (Alt-SIMCA)	**Authentic**	**29**	9	26	6
	Adulterated	4	**3**	**13**	**0**
Assigned class (Soft PLS-DA)	**Authentic**	**27**	11	1	3
	Adulterated	6	**1**	**38**	**3**

Table 5 *Classification results of the images used for external validation into authentic and adulterated oregano classes based on the threshold considering PPO% values, obtained by applying Alt-SIMCA and Soft PLS-DA. For the images of adulterated oregano, the results are reported by splitting the samples according to adulterant concentration: adulterant concentration lower than 10% (adult. < 10%), adulterant concentration equal to or higher than 10% (adult. ≥ 10%) and unknown concentration.*

4. Conclusions

The aim of the present study was to evaluate NIR-HSI as a rapid, non-destructive and untargeted method to authenticate oregano samples which, due to their heterogeneity, can benefit from the coupling of spectral and spatial information.

The initial exploratory analysis performed both at the *pixel level* on some representative images and at the *image level* on the average spectra dataset allowed to identify the presence of spectral differences between authentic oregano and pure adulterants, to point out a remarkable heterogeneity among different genuine oregano samples and to highlight the need of accounting for spatial variation of sample composition to authenticate the samples.

Based on these considerations, Alt-SIMCA and Soft PLS-DA algorithms were used to build classification models able to differentiate authentic oregano and its most frequent adulterants, i.e., myrtle, olive leaves, strawberry tree leaves and sumac.

Due to classes overlapping and heterogeneity of the authentic oregano class, Alt-SIMCA led to overall poor classification performances. Conversely, Soft PLS-DA achieved satisfactory outcomes, with efficiency values in classification higher than 91% in calibration, cross-validation and validation of the external test set. In this case, the spectra of pure strawberry tree and sumac leaves were easier to distinguish from authentic oregano, while pure myrtle and olive leaves presented higher misclassifications.

To obtain a final assignment of the acquired oregano samples into authentic and adulterated classes, both classification models were applied to all the acquired hyperspectral images and from each image the corresponding percentage of pixels predicted as oregano (PPO%) was calculated. Once defined a PPO% threshold value to differentiate authentic oregano samples from adulterated ones, it was possible to reach SENS values for authentic class equal to 87.9% and 81.8% for Alt-SIMCA and Soft PLS-DA, respectively. In both cases, the misclassifications of authentic oregano were mainly due to samples containing inflorescences.

The main differences in classification performances were encountered in the ability of correctly differentiating adulterated oregano samples. Indeed, while Alt-SIMCA was unable to correctly identify most of the adulterated samples, Soft PLS-DA successfully distinguished all adulterated oregano samples with adulterant concentrations equal to or greater than 10%. These results confirm that soft discriminant approaches like Soft PLS-DA are an effective and powerful alternative to CM and DA methods when dealing with authentication problems.

Furthermore, according to the results obtained with Soft PLS-DA, we can consider the 10% of adulteration as a sort of limit of detection of NIR-HSI to identify adulterated oregano samples. Considering that the percentage of adulteration detected on market oregano samples has been found very often at much higher levels, these results seem rather satisfactory to corroborate NIR-HSI potentialities as a screening technique able to face adulteration issues, also considering the possibility of performing the analysis in a fast and non-destructive manner.

Further developments may involve expanding the dataset to include additional authentic samples, aiming to better represent the intrinsic variability of oregano matrices and diverse types of adulterants at different percentages of adulteration as well. From the results obtained from the Soft PLS-DA model, a discernible correlation between actual oregano percentages and PPO% values emerged, suggesting the potential for developing a quantitative model with proper sampling.

Moreover, the application of spectral variable selection methods could enhance the robustness and flexibility of the model, while also offering a monitoring system that is easier to apply. Indeed, the selected wavebands can be used to implement multispectral imaging systems, which are more suitable

for industrial applications in terms of computational time, durability and lower costs of the optical components.

CRediT authorship contribution statement

V. Ferrari: Investigation, Methodology, Formal analysis, Data Curation, Software, Writing – original draft. **R. Calvini**: Conceptualization, Methodology, Formal analysis, Software, Writing – original draft. **C. Menozzi**: Investigation, Writing – review & editing. **A. Ulrici**: Supervision, Writing – review & editing. **M. Bragolusi**: Resources, Writing – review & editing. **R. Piro**: Resources, Supervision, Writing – review & editing. **A. Tata**: Resources, Writing – original draft. **M. Suman**: Resources, Supervision, Writing – review & editing. **G. Foca**: Conceptualization, Methodology, Project administration, Funding acquisition, Writing – original draft.

Acknowledgements

The authors would like to thank the funding programme "FAR_DIP2022", Department of Life Sciences – University of Modena and Reggio Emilia.

Rosalba Calvini would like to thank the Italian funding programme Fondo Sociale Europeo REACT-EU - PON "Ricerca e Innovazione" 2014 –2020 – Azione IV.6 Contratti di ricerca su tematiche Green (D.M. 1062 del 10/08/ 2021) for supporting her research (CUP: E95F21002330001; contract number 17-G-13884–4).

References

[1] P. Galvin-King, S.A. Haughey, C.T. Elliott, Herb and spice fraud; the drivers, challenges and detection, Food Control 88 (2018) 85–97. https://doi.org/10.1016/j.foodcont.2017.12.031.

[2] European Spice Association, Adulteration Awareness Document. https://www.esa-spices.org/download/esa-adulteration-awareness-document2, 2018 (accessed 9 October 2023).

[3] P.F. Ndlovu, L.S. Magwaza, S.Z. Tesfay, R.R. Mphahlele, Destructive and rapid non-invasive methods used to detect adulteration of dried powdered horticultural products: A review, Food Res. Int. 157 (2022) 111198. https://doi.org/10.1016/j.foodres.2022.111198.

[4] M. Shannon, J.L. Lafeuille, A. Frégière-Salomon, S. Lefevre, P. Galvin-King, S.A. Haughey, D.T. Burns, X. Shen, A. Kapil, T.F. McGrath, C.T. Elliott, The detection and determination of adulterants in turmeric using fourier-transform infrared (FTIR) spectroscopy coupled to chemometric analysis and micro-FTIR imaging, Food Control 139 (2022) 109093. https://doi.org/10.1016/j.foodcont.2022.109093.

[5] A. Maquet, A. Lievens, V. Paracchini, G. Kaklamanos, B. de la Calle, L. Garlant, S. Papoci, D. Pietretti, T. Zdiniakova, A. Breidbach, J. Omar Onaindia, A. Boix Sanfeliu, T. Dimitrova, F. Ulberth, Results of an EU wide Coordinated Control Plan to establish the prevalence of fraudulent practices in the marketing of herbs and spices, EUR30877EN, Publications Office of the European Union (2021) JRC126785. https://doi:10.2760/309557.

[6] L. Drabova, G. Alvarez-Rivera, M. Suchanova, D. Schusterova, J. Pulkrabova, M. Tomaniova, V. Kocoureka, O. Chevallier, C. Elliott, J. Hajslova, Food fraud in oregano: Pesticide residues as adulteration markers, Food Chem. 276 (2019) 726–734. https://doi.org/10.1016/j.foodchem.2018.09.143.

[7] S.Schaarschmidt, Public and private standards for dried culinary herbs and spices – Part I: Standards defining the physical and chemical product quality and safety, Food Control 70 (2016) 339–349. https://doi.org/10.1016/j.foodcont.2016.06.004.

[8] FAO and WHO, Standard for dried oregano, Codex Alimentarius Standard, No. CXS 342-2021. Codex Alimentarius Commission (2021).

[9] European Spice Association, European Spice Association Quality Minima Document. https://www.esa-spices.org/download/esa-qmd-rev-5-update-as-per-esa-tc-26-03-18.pdf, 2018 (accessed 15 January 2024).

[10] C. Black, S.A. Haughey, O.P. Chevallier, P. Galvin-King, C.T. Elliott, A comprehensive strategy to detect the fraudulent adulteration of herbs: The oregano approach, Food Chem. 210 (2016) 551–557. https://doi.org/10.1016/j.foodchem.2016.05.004.

[11] E. Wielogorska, O. Chevallier, C. Black, P. Galvin-King, M. Delêtre, C.T. Kelleher, S.A. Haughey, C. T. Elliott, Development of a comprehensive analytical platform for the detection and quantitation of food fraud using a biomarker approach. The oregano adulteration case study, Food Chem. 239 (2018) 32–39. https://doi.org/10.1016/j.foodchem.2017.06.083.

[12] G. Cottenet, C. Cavin, C. Blancpain, P. F. Chuah, R. Pellesi, M. Suman, S. Nogueira, M. Gadanho, A DNA metabarcoding workflow to identify species in spices and herbs, J. AOAC Int. 106(1) (2022) 65–72. https://doi.org/doi:10.1093/jaoacint/qsac099

[13] J. Pages-Rebull, C. Pérez-Ràfols, N. Serrano, M. del Valle, J.M. Díaz-Cruz, Classification and authentication of spices and aromatic herbs by means of HPLC-UV and chemometrics, Food Biosci. 52 (2023) 102401. https://doi.org/10.1016/j.fbio.2023.102401.

[14] K. Kucharska-Ambrożej, J. Karpinska, The application of spectroscopic techniques in combination with chemometrics for detection adulteration of some herbs and spices, Microchem. J. 153 (2020) 104278. https://doi.org/10.1016/j.microc.2019.104278.

[15] F. Flügge, T. Kerkow, P. Kowalski, J. Bornhöft, E. Seemann, M. Creydt, B. Schütze, U.L. Günther, Qualitative and quantitative food authentication of oregano using NGS and NMR with chemometrics, Food Chem. 145 (2023) 109497. https://doi.org/10.1016/j.foodcont.2022.109497.

[16] 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, Authenticity analysis of oregano: development, validation and fitness for use of several food fingerprinting techniques, Food Res. Int. 162 (2022) 111962. https://doi.org/10.1016/j.foodres.2022.111962.

[17] A. Massaro, A. Negro, M. Bragolusi, B. Miano, A. Tata, M. Suman, R. Piro, Oregano authentication by mid-level data fusion of chemical fingerprint signatures acquired by ambient mass spectrometry, Food Control 126 (2021) 108058. https://doi.org/10.1016/j.foodcont.2021.108058.

[18] C. Zacometti, A Massaro, T. di Gioia, S. Lefevre, A. Frégière-Salomon, J.L. Lafeuille, I. Fiordaliso Candalino, M. Suman, R. Piro, A. Tata, Thermal desorption direct analysis in real-time high-resolution mass spectrometry and machine learning allow the rapid authentication of ground black pepper and dried oregano: A proof-of-concept study. J. Mass Spectrom. 58(10) (2023) e4953. https://doi.org/10.1002/jms.4953

[19] T. Damiani, N. Dreolin, S. Stead, C. Dall'Asta, Critical evaluation of ambient mass spectrometry coupled with chemometrics for the early detection of adulteration scenarios in *Origanum vulgare* L., Talanta 227 (2021) 122116. https://doi.org/10.1016/j.talanta.2021.122116.

[20] G. Sammarco, M. Alinovi, L. Fiorani, M. Rinaldi, M. Suman, A. Lai, A. Puiu, L. Giardina, F. Pollastrone, Oregano herb adulteration detection through rapid spectroscopic approaches: Fourier transform-near

infrared and laser photoacoustic spectroscopy facilities, J. Food Compost. Anal. 124 (2023) 105672. https://doi.org/10.1016/j.jfca.2023.105672.

[21] C. McVey, T.F. McGrath, S.A. Haughey, C.T. Elliott, A rapid food chain approach for authenticity screening: The development, validation and transferability of a chemometric model using two handheld near infrared spectroscopy (NIRS) devices, Talanta 222 (2021) 121533. https://doi.org/10.1016/j.talanta.2020.121533.

[22] O.Ye. Rodionova, A.L. Pomerantsev, Chemometric tools for food fraud detection: The role of target class in nontargeted analysis, Food Chem. 317 (2020) 126448. https://doi.org/10.1016/j.foodchem.2020.126448.

[23] R. Khodabakhshian, M.R. Bayati, B. Emadi, Adulteration detection of Sudan Red and metanil yellow in turmeric powder by NIR spectroscopy and chemometrics: The role of preprocessing methods in analysis, Vib. Spectrosc. 120 (2022) 103372. https://doi.org/10.1016/j.vibspec.2022.103372.

[24] A.M. Elfiky, E. Shawky, A.R. Khattab, R.S. Ibrahim, Integration of NIR spectroscopy and chemometrics for authentication and quantitation of adulteration in sweet marjoram (Origanum majorana L.), Microchem. J. 183 (2022) 108125. https://doi.org/10.1016/j.microc.2022.108125.

[25] A. Massaro, M. Bragolusi, A. Tata, C. Zacometti, S. Lefevre, A. Frégière-Salomon, J.L. Lafeuille, G. Sammarco, I. Fiordaliso Candalino, M. Suman, R. Piro, Non-targeted authentication of black pepper using a local web platform: Development, validation and post-analytical challenges of a combined NIR spectroscopy and LASSO method, Food Control 145 (2023) 109477. https://doi.org/10.1016/j.foodcont.2022.109477

[26] J.P. Cruz-Tirado, Y. Lima Brasil, A. Freitas Lima, H. Alva Pretel, H. Teixeira Godoy, D. Barbin, R. Siche, Rapid and non-destructive cinnamon authentication by NIR-hyperspectral imaging and classification chemometrics tools, Spectrochim. Acta A Mol. Biomol. Spectrosc. 289 (2023) 122226. https://doi.org/10.1016/j.saa.2022.122226.

[27] Z. Jiang, A. Lv, L. Zhong, J. Yang, X. Xu, Y. Li, Y. Liu, Q. Fan, Q. Shao, A. Zhang, Rapid prediction of adulteration content in *Atractylodis rhizoma* based on data and image features fusions from near-infrared spectroscopy and hyperspectral imaging techniques, Foods 12 (2023) 2904. https://doi.org/10.3390/foods12152904.

[28] R. Calvini, A. Ulrici, J.M. Amigo, Growing applications of hyperspectral and multispectral imaging, in: J.M. Amigo (Ed.), Data Handling in Science and Technology – Vol. 32 Hyperspectral Imaging, Elsevier, Amsterdam, 2019, pp. 605–629.

[29] 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.

[30] V. Ferrari, R. Calvini, B. Boom, C. Menozzi, A.K. Rangarajan, L. Maistrello, P. Offermans, A. Ulrici, Evaluation of the potential of near infrared hyperspectral imaging for monitoring the invasive brown marmorated stink bug, Chemom. Intell. Lab. Syst. 234 (2023) 104751. https://doi.org/10.1016/j.chemolab.2023.104751.

[31] R. Vitale, M. Cocchi, A. Biancolillo, C. Ruckebusch, F. Marini, Class modelling by soft independent modelling of class analogy: why, when, how? A tutorial. Anal. Chim. Acta 1270 (2023) 341304. https://doi.org/10.1016/j.aca.2023.341304

[32] P. Oliveri, Class-modelling in food analytical chemistry: development, sampling, optimisation and validation issues–a tutorial. Anal. Chim. Acta 982 (2017), 9-19. http://dx.doi.org/10.1016/j.aca.2017.05.013

[33] O. Y. Rodionova, A. V. Titova, A. L. Pomerantsev, Discriminant analysis is an inappropriate method of authentication, Trends Anal. Chem. 78 (2016), 17-22. https://doi.org/10.1016/j.trac.2016.01.010

[34] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003), 166-173. https://doi.org/10.1002/cem.785

[35] 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

[36] Z. Małyjurek, D. de Beer, E. Joubert, B. Walczak, Combining class-modelling and discriminant methods for improvement of products authentication, Chemom. Intell. Lab. Syst. 228 (2022) 104620. https://doi.org/10.1016/j.chemolab.2022.104620.

[37] Z. Małyjurek, D. de Beer, H. van Schoor, J. Colling, E. Joubert, B. Walczak, Class-modelling of overlapping classes. A two-step authentication approach, Anal. Chim. Acta 1191 (2022) 339284. https://doi.org/10.1016/j.aca.2021.339284

[38] A. L. Pomerantsev, O. Y. Rodionova, Multiclass partial least squares discriminant analysis: Taking the right way—A critical tutorial, J. Chemom. 32 (2018) e3030. https://doi.org/10.1002/cem.3030

[39] R. Calvini, G. Orlandi, G. Foca, A. Ulrici, Development of a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging, J. Spectr. Imaging 7 (2018) a13. https://doi.org/10.1255/jsi.2018.a13.

[40] J. Burger, P. Geladi, Hyperspectral NIR image regression part II: Dataset preprocessing diagnostics, J. Chemom. 20(3-4) (2006) 106–119. https://doi.org/10.1002/cem.986.

[41] A. Ulrici, S. Serranti, C. Ferrari, D. Cesare, G. Foca, G. Bonifazi, Efficient chemometric strategies for PET-PLA discrimination in recycling plants using hyperspectral imaging, Chemom. Intell. Lab. Syst. 122 (2013) 31–39. https://doi.org/10.1016/j.chemolab.2013.01.001.

[42] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, Chemom. Intell. Lab. Syst. 174 (2018) 33-44. https://doi.org/10.1016/j.chemolab.2017.12.004.

[43] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics 11 (1969) 137–148. https://doi.org/10.1080/00401706.1969.10490666.

[44] O. Y. Rodionova, P. Oliveri, A. L. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemom. Intell. Lab. Syst. 159 (2016) 89-96. http://dx.doi.org/10.1016/j.chemolab.2016.10.002

[45] S. Wold, Pattern recognition by means of disjoint principal components models, Pattern Recogn. 8 (1976) 127–139. https://doi.org/10.1016/0031-3203(76)90014-5

[46] A. Biancolillo, R. Bucci, A. L. Magrì, A. D. Magrì, F. Marini, Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication. Anal. Chim. Acta 820 (2014), 23-31. http://dx.doi.org/10.1016/j.aca.2014.02.024

[47] T.J. Bruno, P.D.N. Svoronos, CRC Handbook of fundamental spectroscopic correlation charts. CRC Press, Boca Raton, 2005, chapter 2.

[48] J.S. Shenk, J.J. Workman, Jr., M.O. Westerhaus, Application of NIR Spectroscopy to Agricultural Products, in: D.A. Burns, E.W. Ciurczak, E. W. (Eds.), Handbook of near-infrared analysis, CRC Press, Boca Raton, 2007, pp. 356–365.

[49] J.J. Workman, Jr., L. Weyer, Practical guide to interpretive near-infrared spectroscopy. CRC Press, Boca Raton, 2007, Chapters 5-6.

[50] J.J. Workman, Jr., The handbook of organic compounds, three-volume set: NIR, IR, and UV-Vis spectra featuring polymers and surfactants – Vol. 2, Academic Press, Cambridge, 2000, pp-160-165.

[51] I. Oniga, C. Pușcaș, R. Silaghi-Dumitrescu, N.K. Olah, B. Sevastre, R. Marica, I. Marcus, A.C. Sevastre-Berghian, D. Benedec, C.E. Pop, D. Hanganu, *Origanum vulgare ssp. vulgare*: Chemical composition and biological studies, Molecules 23(8) (2018) 2077. https://doi.org/10.3390/molecules23082077.

Chapter 5

Changing perspectives:
applying sparsity in the spatial direction for the analysis of hyperspectral data

As previously discussed in ***Chapter 2***, hyperspectral images represent an example of big data arrays, since each image is a three-dimensional matrix comprising millions of numbers.
The high dimensionality inherent to hyperspectral data leads to substantial computational, storage, and interpretability issues, making it difficult to extract meaningful and interpretable information. Multivariate Image Analysis (MIA) is therefore essential to unravel this *curse of dimensionality*; however, classical pixel-level approaches rapidly become impractical for large datasets in real-time applications, motivating the need for dedicated dimensionality reduction strategies.

Within this framework, this chapter evaluates the benefits of sparse methods applied in the spatial direction as an alternative feature selection strategy for identifying relevant pixels of interest. In particular, sparse Principal Component Analysis (sPCA) with sparsity imposed on the score vectors (score-level sPCA) has been investigated as a promising exploratory tool to deal with high-dimensional data, facilitating the identification of pixels corresponding to Regions of Interest.

5.1. Introduction

Data analytical problems are becoming increasingly complex as a result of advances in process analytical technologies (PAT). While these developments enable the routine acquisition of high-dimensional datasets, they also make the extraction of meaningful and interpretable information increasingly challenging [1,2].
A representative example of this issue is hyperspectral imaging (HSI). HSI combines the strengths of spectroscopy and imaging techniques, allowing the simultaneous acquisition of spatial and spectral information across a sample surface. A hyperspectral image can be described as a three-dimensional data array with two spatial dimensions and one spectral dimension. Owing to this structure, a single hyperspectral image may consist of millions of data points [3,4].
Multivariate Image Analysis (MIA) is therefore essential for the analysis of hyperspectral data. However, the data-richness associated with HSI poses significant challenges in terms of computational cost, data handling, and real-time applicability, ultimately complicating the extraction

of meaningful information. Moreover, this *curse of dimensionality* determines an inverse relationship between the number of input features in a model and its ability to generalise effectively [5].

To address these issues, proper data dimensionality reduction strategies are required to retain only relevant spectral or spatial features ascribable to Regions Of Interest (ROIs), which describe localized sources of variability within the sample. Among the most widely used data dimensionality reduction techniques, Principal Component Analysis (PCA) summarizes the original data into a lower-dimensional subspace that captures the main sources of variance [6,7].

However, standard MIA approaches such as PCA typically produce dense models, in which a large number of spectral and spatial features contribute to each component. As a consequence, identifying which features are truly relevant for further investigation becomes difficult, especially when the information of interest is partially covered by other sources containing non pertinent information or noise.

A possible strategy to reduce model complexity and enhance interpretability is to impose sparsity to the models, by forcing the contribution of less influential features to zero. Methods adopting this principle are referred to as sparse methods. These approaches extend traditional multivariate techniques by inducing sparsity to the estimated model parameters, which are forced to contain many zero entries and only a limited number of coefficients have non-zero values. In this way, irrelevant or noise-related features are discarded, while the retained features highlight the most relevant sources of information within the data [8–10].

Sparse methods offer clear advantages in terms of interpretability and model simplification. Nevertheless, their application requires proper optimization of additional parameters, such as the sparsity level and the number of components [3,11–13]. Beyond interpretability, the growing interest in sparse methods is also motivated by the observation that many underlying data-generating processes are inherently sparse [11]. This assumption is particularly relevant in hyperspectral imaging, where chemically meaningful information is often confined to specific spectral bands or spatial regions, while the remaining data primarily reflect noise or irrelevant variability.

When dealing with hyperspectral data, sparsity is often imposed in the spectral domain in order to select relevant wavelengths for a specific problem. Numerous sparse-based extensions of traditional unsupervised or supervised chemometric algorithms have been applied to hyperspectral datasets to perform variable selection, improving model performance while providing more interpretable results from the spectral point of view [14–18].

However, the benefit of imposing sparsity in the spatial domain to identify and extract relevant spatial features is still underexplored. Within this framework, this study investigates the advantages of adding a sparsity penalisation term to the score vectors of a sPCA model calculated on hyperspectral data to

select spatial features, i.e., pixels within a hyperspectral image. As a result, the corresponding score images will be sparse, with only a limited number of pixels containing non-zero values.

Sparsity in the spatial domain may be advantageous for hyperspectral data, as it enables the direct extraction of ROIs while suppressing non-informative and redundant information. In addition, this strategy favours data compression by reducing storage and computational efforts.

Therefore, in this study, we explore the benefits of a score-level sPCA approach for spatial feature selection in Near Infrared (NIR) hyperspectral images considering two distinct case studies of different complexity. For both case studies the main goal of the analysis consists in isolating and selecting the pixels corresponding to specific spatial features, and identifying their spectral characteristics.

5.2. Theory

5.2.1. Sparse methods

Sparse methods aim to identify informative features in high-dimensional datasets by enforcing sparsity to the model parameters. Sparsity is introduced through the inclusion of a penalty term, which forces uninformative or noisy coefficients to zero [8–10].

Different penalization strategies can be adopted to obtain sparse models. Ridge regression [19] introduces an L_2 norm regularisation term on the sum of squared values of the model parameters, resulting in coefficient shrinkage. Conversely, the Least Absolute Shrinkage and Selection Operator (LASSO) [20,21] applies an L_1 norm penalty on the sum of absolute values of the model parameters, promoting sparsity by driving many coefficients exactly to zero. An alternative solution was introduced in [22], namely elastic net, which offers a good compromise, allowing the selection of features of interest like the LASSO while enabling at the same time coefficients shrinkage like Ridge regression.

Regardless of the penalisation strategy adopted, the penalisation term is added to the objective function of the model; therefore, feature selection and model computation are performed in a single step. Both model complexity (i.e., the number of PCs in sPCA) and the sparsity level have to be defined prior to model computation and this represents a crucial step for data analysis.

5.2.2. Score-level sparse PCA (sl-sPCA)

Sparse Principal Component Analysis (sPCA) extends classical PCA by inducing sparsity on the model parameters, namely scores and/or loadings, thereby improving interpretability while retaining the main variance structure of the data [20,23–25].

In the majority of sPCA applications sparsity is induced on the loading, while sparsity at the score-level is still underexplored. To avoid confusion, in this chapter the term score-level sPCA (sl-sPCA) is used to specifically denote sPCA formulations in which sparsity is imposed on the score vectors.

Several algorithms have been proposed to calculate sPCA models, the majority of them focusing on sparsity induced on the loading vectors, as already highlighted. A widely used approach is to compute one sparse component at a time by deflating each component from the data, in order to compute residuals to fit the next component. However, a recent review of sPCA algorithms showed that deflation-based methods may determine the appearance of artefacts or inaccurate residual estimation, potentially leading to redundancy and double-counting of variance across multiple components [12]. An alternative approach consists in simultaneously extracting all the sparse components altogether [13]. The Alternating Shrunken Least Squares (ASLS) algorithm proposed by Rasmussen and Bro is an example of this approach, as it estimates all the sparse PCs simultaneously by iterating between scores and loadings until convergence [20]. This strategy is particularly effective in avoiding redundancy of the retrieved information, as sparsity is imposed by fulfilling the L_1 norm constraint component-wise on the loading vectors during the iterations, eventually enforcing orthogonality on the scores.

ASLS algorithm was used in this study to calculate sl-sPCA. Following a workflow similar to that proposed in [20,23], the sparse solution was obtained by estimating altogether a fixed number of sparse components by iterating between scores loadings until convergence, while imposing the LASSO penalisation to each column of the score matrix.

Let $\boldsymbol{X}$ matrix denote an unfolded and preprocessed spectral image with size $\{n,\ m\}$, where n corresponds to the number of pixels included in the image and m corresponds to the number of spectral variables. Given a fixed number of components (A), ASLS algorithm finds the score-level sparse solution to:

$$\arg min_{\boldsymbol{T},\boldsymbol{P}} = (\|\boldsymbol{X} - \boldsymbol{T}\boldsymbol{P}^T\|_F^2) \tag{5.1}$$

subjected to

$$\|\boldsymbol{t}_a\|_1 \leq c \qquad \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}_A \qquad for\ a = 1, \dots, A \tag{5.2}$$

where $\boldsymbol{T}\{n, A\}$ is the scores matrix and $\boldsymbol{t}_a$ denotes the a-th column of the score matrix $\boldsymbol{T}$, $\boldsymbol{P}\{m, A\}$ is the loadings matrix and $\|\|\ \ \|_F^2$ is the sum of squared elements (also known as Frobenius norm).

Considering the loadings, the constraint $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}_A$ ensures both orthogonality and unit-norm normalization for each loading vector, as for classical PCA. Most importantly, the constraint $\|\boldsymbol{t}_a\|_1 \leq c$ represent the L_1 norm penalisation on each score vector $\boldsymbol{t}_a$, where the scalar parameter c corresponds to the sparsity constraint, a tuning parameter that defines the level of sparsity. In more detail, the sparsity constrain can range from 1 to the square root of the number of elements in $\boldsymbol{t}_a$, which in this case corresponds to the number of pixels included in the image (n). The lower the sparsity constraint value, the higher sparsity is imposed on the score vectors. In other words, sl-sPCA models calculated with c value close to one correspond to highly sparse models with only few pixels having non-zero score values.

The algorithm estimates the loading matrix ($\boldsymbol{P}$) based on the current scores ($\boldsymbol{T}$), which are initialised using the score matrix of standard PCA. Then, the updated score matrix is estimated based on current loadings as least squares estimate subjected to column-wise soft thresholding due to the application of the LASSO penalisation.

The estimation of score and loading matrices is alternated until convergence, defined as negligible changes in reconstruction error between successive iterations. Therefore, the formulation of sl-sPCA can be interpreted as a particular case of Sparse Matrix Regression (SMR), where an L_1 norm penalty is applied column-wise to the score matrix and additional orthonormality constraints are imposed on the loading vectors [23].

In the context of image analysis, sparsity induced on the score vectors allows to obtain the corresponding sparse score images, that highlight spatially resolved features or ROIs.

5.3. Materials and Methods

5.3.1. Datasets

The effectiveness of sl-sPCA as a pixel selection method was evaluated considering two different well-established NIR-HSI datasets with increasing levels of complexity. The first dataset comprised NIR hyperspectral images of plastics pellets made of different polymers, representing a relatively controlled scenario as the different plastic polymers have distinct signatures in the NIR range. The second dataset included images of Brown Marmorated Stink Bugs (BMSB) acquired on diverse vegetal backgrounds, thus representing a more challenging situation. For both datasets, the main goal

of sl-sPCA analysis consisted in the selection of pixels belonging to spatially resolved ROIs, such as plastic pellets of an extraneous polymer in the first dataset and BMSB in the second dataset.

Dataset 1: Plastic pellets

The plastics dataset consists of two NIR hyperspectral images acquired in the 950 – 1700 nm spectral range. Both images contain polystyrene (PS) plastic pellets as the main material; one image also includes two polyethylene pellets (PS + PE 2) while the other also contains one polyethylene terephthalate (PET) pellet (PS + PET 1). The plastic pellets were obtained by chopping household plastic waste items made of the considered polymers. The aim of the analysis is to identify and characterise the chemical profile of the outlier objects, corresponding to the PE or PET pieces, respectively. A masking procedure was carried out prior to data analysis in order to remove the pixels related to the black sandpaper background used for image acquisition.

For further information about this benchmark dataset created *ad hoc* for outlier detection, the reader is referred to [8].

The average spectra corresponding to pure PS, PE and PET pellets can be compared in **Figure 5.1 A**. Some spectral differences are glaring, especially between PE and PS and PET, where the most peculiar spectral regions are related to the C−H aromatic second overtone (1143 nm), absorption band common to PS and PET, and to C—H aliphatic second overtone (1215 nm) corresponding to absorption bands of PE. Additional slight differences between PS and PET are related to the spectral regions between 1670 nm -1700 nm, ascribable to the C−H aromatic stretching first overtone [26].

Dataset 2: BMSB on vegetal backgrounds

This dataset consists of three NIR hyperspectral images acquired in the 980 – 1660 nm spectral range of BMSB specimens placed on different vegetal backgrounds. Specifically, each image contains four insects; in two images the specimens are positioned on homogeneous backgrounds, i.e., green leaves (BMSB + LEV) and tree branches (BMSB + TBR), while in the third image the BMSB specimens are placed on a heterogenous background (BMSB + MIX) including bark, grass, leaves, soil and tree branches.

These three images are a subset of a dataset acquired in a previous study with the aim of assessing the ability of NIR-HSI in detecting BMSB on field, and the different vegetal backgrounds were chosen in order to mimic real field conditions [17]. After image acquisition, a masking procedure was implemented to remove the pixels of the black sandpaper background of the acquisition set up and include only the sample pixels (i.e., BMSB and vegetal backgrounds). The detailed description of this

dataset, including image acquisition setup and masking procedure, is reported in **Section 3.2** of ***Chapter 3*** [17].

Figure 5.1 B reports a subset of representative spectra of BMSB, LEV and TBR. Concerning vegetal backgrounds, the spectral differences are mostly related to water, cellulose, hemicellulose and lignin content (1220–1295 nm, 1420–1480 nm) while, between backgrounds and bug specimens, the differences are mainly due to absorption bands ascribable to protein, chitin and lipids (980–1070 nm,1330–1350 nm), which can be associated to the biochemical structure of the outer layer of insects' exoskeleton [26].

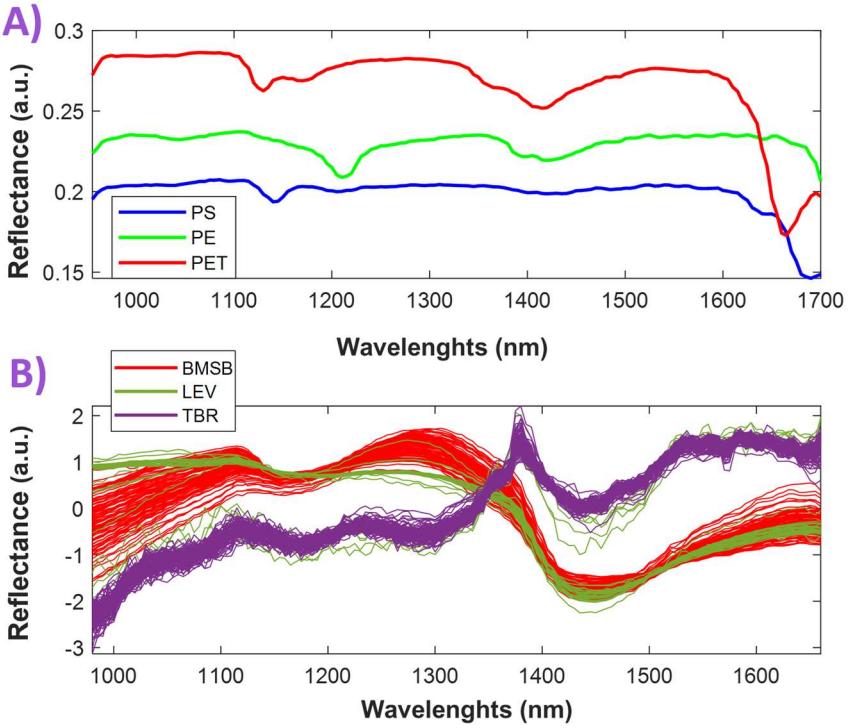


Figure 5.1 *Representative spectra of the investigated datasets. In A): average spectra of pure PS, PE and PET pellets while in B): 100 representative spectra of BMSB, green leaves (LEV) and tree branches (TBR).*

5.3.2. Data analysis

All the hyperspectral images of the two datasets considered in this study were pre-processed with SNV and mean centering prior to data analysis.

For each image, a series of sl-sPCA models was calculated considering a fixed number of PCs and varying values of the sparsity constraint. In particular, the number of PCs was chosen for each image based on a preliminary assessment of the main sources of data variability.

To evaluate the effect of different sparsity levels on pixel selection performed by sl-sPCA, 20 different values for the sparsity constraint were tested for each image. As stated in **Section 5.2.2**, in sl-sPCA

the sparsity constrain can range between 1 and the square root of the number of pixels. A value of *c* equal to one corresponds to a highly sparse model, in which only one pixel is selected, while *c* equal to the square root of the number of pixels corresponds to the standard PCA model.

Since the five images used in this study have a different number of pixels included after masking, using the same *c* value for all of them would result in models with different effective sparsity levels. To ensure a meaningful comparison across images, the values of the sparsity constraint to be tested were therefore defined in a relative manner by setting 20 sparsity constraint levels common to all the images moving from high sparsity (sparsity constraint level equal to 1) to very low sparsity (sparsity constraint level equal to 20). Each sparsity constraint level was linked to a specific *c* value for each image, as reported in **Table 5.1**. Specifically, for each image, 22 equally-spaced values of *c* were generated ranging between 1 and the square root of the number of pixels. The two extreme values were subsequently discarded, as they correspond to models of limited interest (i.e., a single selected pixel and standard PCA, respectively). The remaining 20 sparsity values were retained and used for sl-sPCA model computation.

A detailed description of the dimensions of the considered images and the values of the sparsity constraints tested for each image are reported in **Table 5.1**.

5.3.3. Software

The code for sl-sPCA was implemented in MATLAB environment (2024a, The MathWorks, USA) by adapting the ASLS algorithm for sPCA freely available at https://ucphchemometrics.com/186-2/algorithms/ [20,23].

All the additional calculations were carried out in MATLAB 2024a (The MathWorks, USA) environment, using HYPER-tools v.4 [27], Image Processing Toolbox (v. 11.1) and PLS_Toolbox (v. 8.8.1, Eigenvector Research Inc., USA) functions.

		PS+PE2	PS+PET1	BMSB+LEV	BMSB+TBR	BMSB+MIX
Image size		145×196	145×196	241×245	241×245	241×245
# pixels after background removal		**17672**	**20381**	**37550**	**25559**	**31136**
# spectral variables		150	150	137	137	137
# sPCs		**3**	**4**	**3**	**4**	**4**
		Sparsity constraint (c)	Sparsity constraint (c)	Sparsity constraint (c)	Sparsity constraint (c)	Sparsity constraint (c)
Sparsity constraint level	1	**7**	**8**	**10**	9	**9**
	2	14	**15**	19	**16**	18
	3	**20**	21	**29**	24	**26**
	4	26	28	38	**31**	34
	5	**32**	35	**47**	39	43
	6	39	**42**	56	46	**51**
	7	45	48	65	**54**	59
	8	**51**	55	**74**	62	68
	9	58	**62**	84	69	76
	10	64	69	93	**77**	**85**
	11	70	75	102	84	93
	12	76	82	111	92	101
	13	83	89	120	99	110
	14	89	96	130	107	118
	15	95	102	139	114	126
	16	102	109	148	122	135
	17	108	116	157	130	143
	18	114	123	166	137	151
	19	120	129	175	145	160
	20	127	136	185	152	168

Table 5.1 *Specifics of the investigated datasets, in terms of data size, number of included pixels, number of sPCs and the 20 sparsity constraint values tested, listed in ascending order from higher sparsity to lower sparsity. Sparsity constraints values corresponding to the results that will be presented in the Results and Discussion section are reported in bold and highlighted in green colour.*

5.4. Results and Discussion

5.4.1. Dataset 1: Plastic pellets

Considering PS + PE 2 image, **Figure 5.2** compares the score images and loading vectors of the first three principal components obtained from sl-sPCA models with those of the standard PCA model. To illustrate the impact of sparsity on the spatial distribution of scores, for each component four sparse score images corresponding to four sparsity constraint values are shown alongside the one of the reference PCA model (*see* **Table 5.1**, values highlighted in green). The sl-sPCA models whose score images are shown in **Figure 5.2** correspond to increasing values of c, from higher (i.e., $c = 7$) to lower ($c = 51$) sparsity.

Overall, sl-sPCA produces results that are consistent with classical PCA while substantially reducing the number of considered pixels. In the PS+PE 2 score images of standard PCA, PC1 and PC3 primarily capture variance related to light scattering due to the strongly irregular shape of the plastic pellets, and therefore contribute limited information for the detection of the two PE outlier pellets. Moving to the sPC1 and sPC3 score images of ssPCA, only the extreme pixels located at the edges of the plastic pellets are consistently preserved as non-zero entries by the score-level sparse models. The main differences across the sl-sPCA models are driven by the imposed sparsity level, which directly determines the number of non-zero pixels retained in the score images.

The selection of the sparsity constraint is crucial for the identification of ROIs, here associated with the two PE pellets. As shown in **Figure 5.2 A**, PC2 captures the relevant information associated with the presence of PE pellets in both the PCA reference model and all the sl-sPCA models. Notably, an accurate selection of PE pixels is achieved even at high sparsity levels: indeed, the pixels corresponding to PE are retained as non-zero entries across the entire range of sparsity constraints investigated, from the most parsimonious ($c = 7$) to the least sparse solutions ($c = 51$). In this scenario, the best sl-sPCA model, i.e., the model allowing to select all the pixels related to PE pellets while discarding as much as possible the other pixels, resulted to be the sl-sPCA model calculated with c equal to 20, resulting in a good masking of the spatially resolved areas ascribable to PE pellets. On the whole, this optimal sl-sPCA model selected only the 9.9% of included pixels in the PS + PE 2 image, which were automatically selected by the models and presented at least one non-zero entry for each score vector.

The last row of **Figure 5.2 A** reports the false-colour images obtained by combining the sPC1, sPC2 and sPC3 score maps of each sl-sPCA model, where the red channel corresponds to sPC1 score image, the green channel to sPC2 score image, and the blue channel to sPC 3 score image. Considering the sl-sPCA false-colour score image of the model calculated with $c = 20$, it is possible to observe that the majority of the overall selected pixel correspond to PE pellets while only fewer extreme pixels of PS were selected.

Figure 5.2 B reports the loading vectors obtained from the different sl-sPCA models together with those of the standard PCA as reference. Considering sPC1, highly-sparse solutions exhibit increased absolute values in the spectral regions 960–1000 nm, 1140–1150 nm, and 1660–1700 nm. These regions correspond to characteristic PS absorption bands, due to C–H aromatic second overtone (~1143 nm) and the C–H aromatic stretching first overtone (1660–1700 nm) [26], suggesting that high-sparsity models prioritize the selection of PS pixels.

In contrast, sPC2 loadings highlight spectral regions around 1215 nm and 1390–1420 nm [26], corresponding to the C–H aliphatic second overtone and the C–H combination band, respectively.

These absorption bands characteristic of PE are consistently emphasized in high sparsity sl-sPCA models, in contrast with the reference PCA model.

Conversely to sPC1 and sPC2, sPC3 loadings drastically change depending on the sparsity constraint chosen, thus not representing the same contribution consistently. For example, highly-sparse solutions (c~10) are heavily affected by light scattering and noise while the sparse constraint (c=20) leading to the best sl-sPCA model onward the spectral regions characterized by higher absolute values correspond to the absorption bands of both PS and PE.

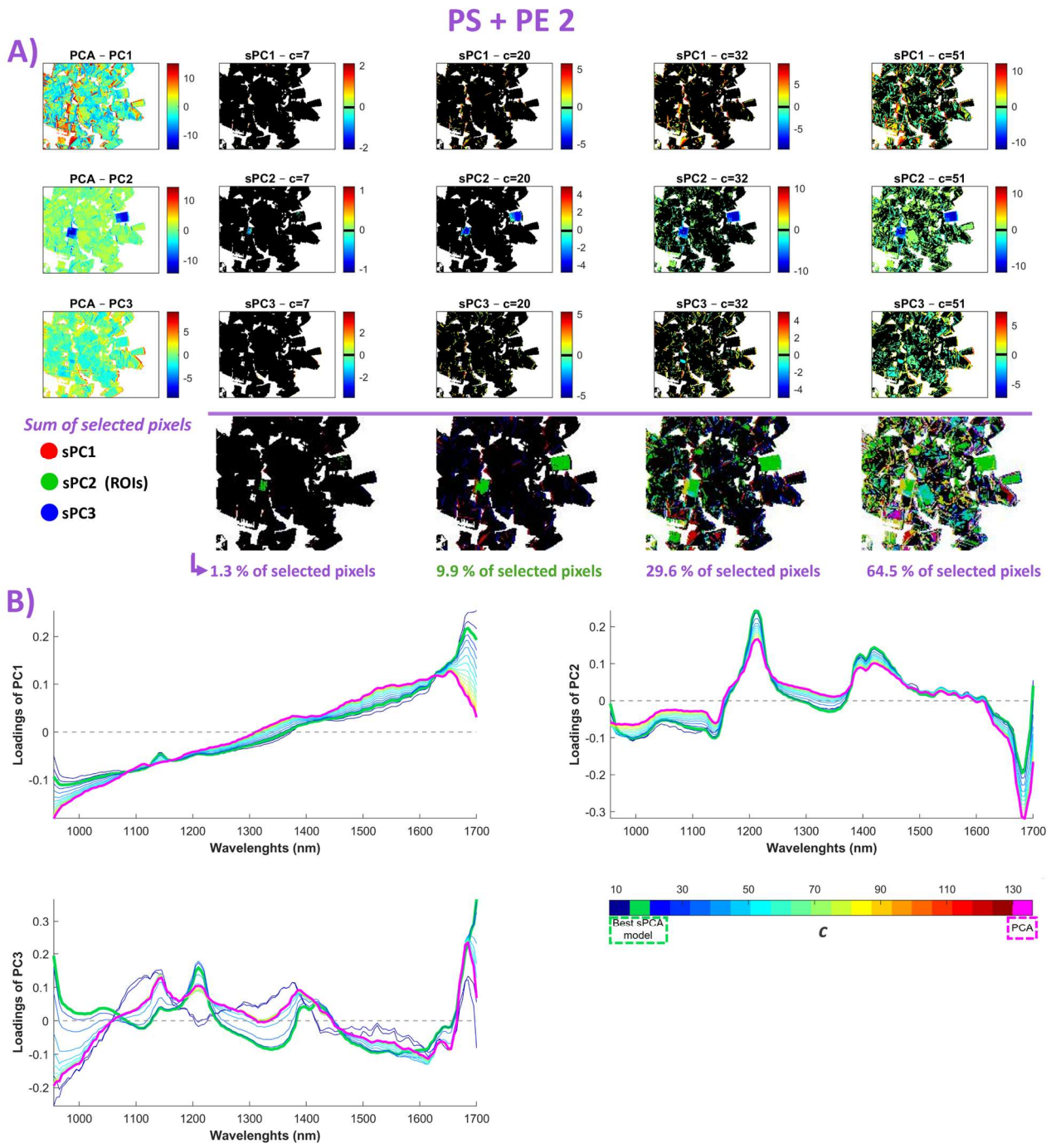


Figure 5.2 *Comparison between standard PCA and sl-sPCA for PS + PE 2 image. In A): PCA score images of PC1, PC2 and PC3 together with the corresponding score images of sl-sPCA models calculated with different values of the sparsity constraint c; the last row reports, for each sl-sPCA model, the false-color image obtained by assigning the sPC1, sPC2 and sPC3 score images to the red, green and blue channels, respectively. The percentage of selected pixels corresponds to the percentage of image pixels that contain at least one non-zero entry in the corresponding score matrix of the sl-sPCA model. In B): loading plots of the standard PCA model and of the corresponding sl-sPCA models calculated with varying values of the sparsity constraint c.*

Similar trends are observed for the PS + PET 1 image (**Figure 5.4**). Consistently with the previous case, the variance associated with light scattering strongly influences the information of the components from PC1 to PC3, resulting in a high number of noisy and non-informative pixels to stand out. Besides light scattering, these components also retain information related to PS spectral fingerprint. Indeed, for both sPC1 and sPC3 loading vectors, increasing the sparsity level progressively suppresses light scattering contributions while enhancing the relevance of PS absorption bands at 1140–1150 nm and 1660–1700 nm.

PC4 describes the variability related to the presence of the PET pellet, and this information is kept constant across all the sl-sPCA models (**Figure 5.3 A**). As in the PS + PE 2 case, successful masking of PET-related pixels is achieved at high sparsity (c = 15), with only 5.2% of the original pixels retained by the corresponding sl-sPCA model. In this scenario, when non-zero pixels are limited to those ascribable to PET, the spectral regions at 1380 – 1400 nm and around 1660 nm gain relevance compared with the unconstrained PCA model (**Figure 5.3 B**). These absorption bands, associated with the C–H aromatic combination band and the C–H aromatic stretching first overtone, are specific for PET, further reflecting the almost exclusive selection of pixels ascribable to PET.

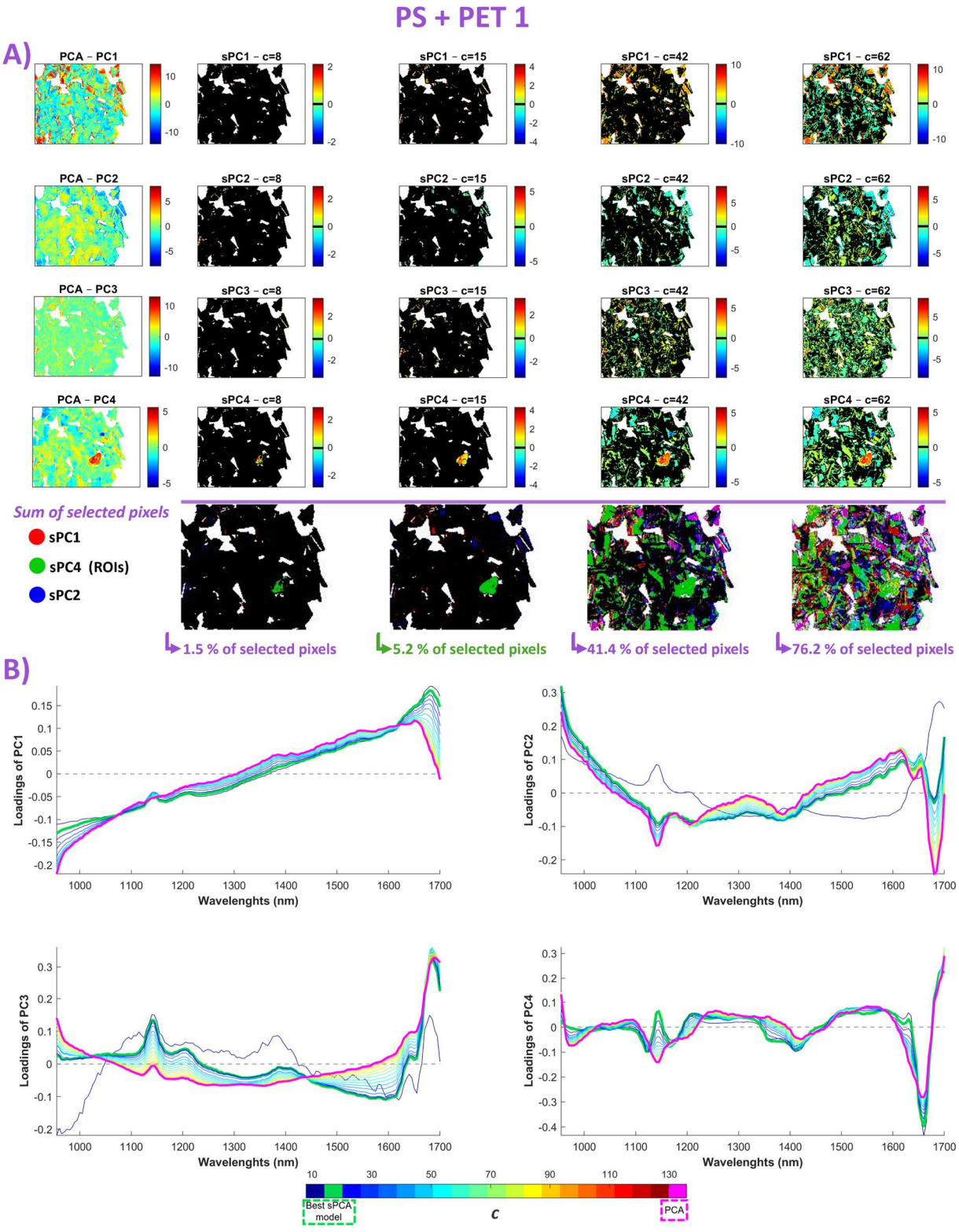


Figure 5.3 *Comparison between standard PCA and sl-sPCA for PS + PET 1 image. In A): PCA score images of PC1, PC2 and PC4 together with the corresponding score images of sl-sPCA models calculated with different values of the sparsity constraint c; the last row reports, for each sl-sPCA model, the false-color image obtained by assigning the sPC1, sPC4 and sPC2 score images to the red, green and blue channels, respectively. The percentage of selected pixels corresponds to the percentage of image pixels that contain at least one non-zero entry in the corresponding score matrix of the sl-sPCA model. In B): loading plots of the standard PCA model and of the corresponding sl-sPCA models calculated with varying values of the sparsity constraint c.*

5.4.2. Dataset 2: BMSB on vegetal backgrounds

Figures 5.4–5.6 compare the score images and loading vectors of the considered principal components obtained from sl-sPCA models with those of the standard PCA as reference. In particular, **Figures 5.4–5.6** report the results obtained with BMSB + LEV, BMSB + TBR and BMSB + MIX images, respectively.

Consistently with the analysis of the plastics dataset, **Figures 5.4 A-5.6 A** show the score images of four sl-sPCA models calculated with different sparsity constraints, alongside with the corresponding PCA score images in order to illustrate the impact of sparsity on the spatial distribution of scores. More details about the models reported in **Figures 5.4 A-5.6 A** can be found in **Table 5.1** (values highlighted in green).

Generally, sl-sPCA produces results that are consistent with classical PCA while substantially reducing the number pixels for calculation and summarizing relevant information. An exception is observed for the BMSB + TBR image. For this image, PC4 is the component describing the variation due to the presence of BMSB specimens in standard PCA. When moving to sl-sPCA, high sparsity levels ($c \leq 16$) lead to sparse score images whose non-zero pixels are not associated with the presence of BMSB specimens but with spurious pixels bringing not relevant information, as also confirmed by the shape of the sl-sPCA loading vectors of the models calculated with high sparsity (**Figure 5.5 B**). Given the pronounced heterogeneity of the BMSB + TBR image, this behaviour is likely ascribable to clusters of extreme or highly variable pixels, further emphasizing the importance of selecting an appropriate sparsity constraint for reliable ROIs annotation. Moving to less sparse sl-sPCA models, the model calculated with c equal to 31 enabled a good selection of BMSB specimens (**Figure 5.5 A**), with only the 12% of selected pixels.

Similar results can be visualized for BMSB + LEV and BMSB + MIX, where the best-performing sl- sPCA models ($c = 29$ and $c = 26$, respectively) led to a good annotation of the spatially resolved areas ascribable to BMSB (**Figures 5.4 A** and **5.6 A**).

For BMSB + LEV, the main source of variance is related to BMSB presence, thus allowing a good separation from the background. Conversely to other images investigated, the information describe by PC1 is mainly due to physical differences which, as reported in the corresponding loading plot, it doesn't lead to any particular change in contribution of meaningful spectral regions related to ROIs presence.

Considering BMSB + TBR and BMSB + MIX images, the comparison of loading vectors provides additional insights. In both cases, PC4 captures the variance associated with BMSB detection, exhibiting similar and consistent spectral features across sparse and unconstrained models (**Figures**

5.5 B–5.6 B). Notably, high-sparsity solutions emphasize spectral regions at 1080–1120 nm and 1380–1410 nm, which correspond to absorption bands of protein, chitin, lipids and water [17].

Moreover, information related to variations in water content is mainly described by the score images of PC3. For instance, in the BMSB + MIX dataset, comparison between the sPC3 score image and the corresponding RGB image reveals that pixels associated with drier backgrounds (e.g. bark, branches, and dry leaves) exhibit positive PC3 values, whereas pixels corresponding to grass and green leaves are characterized by negative PC3 values.

These findings further support the ability of sl-sPCA to enhance chemically meaningful spectral information while effectively suppressing non-informative pixel variability.

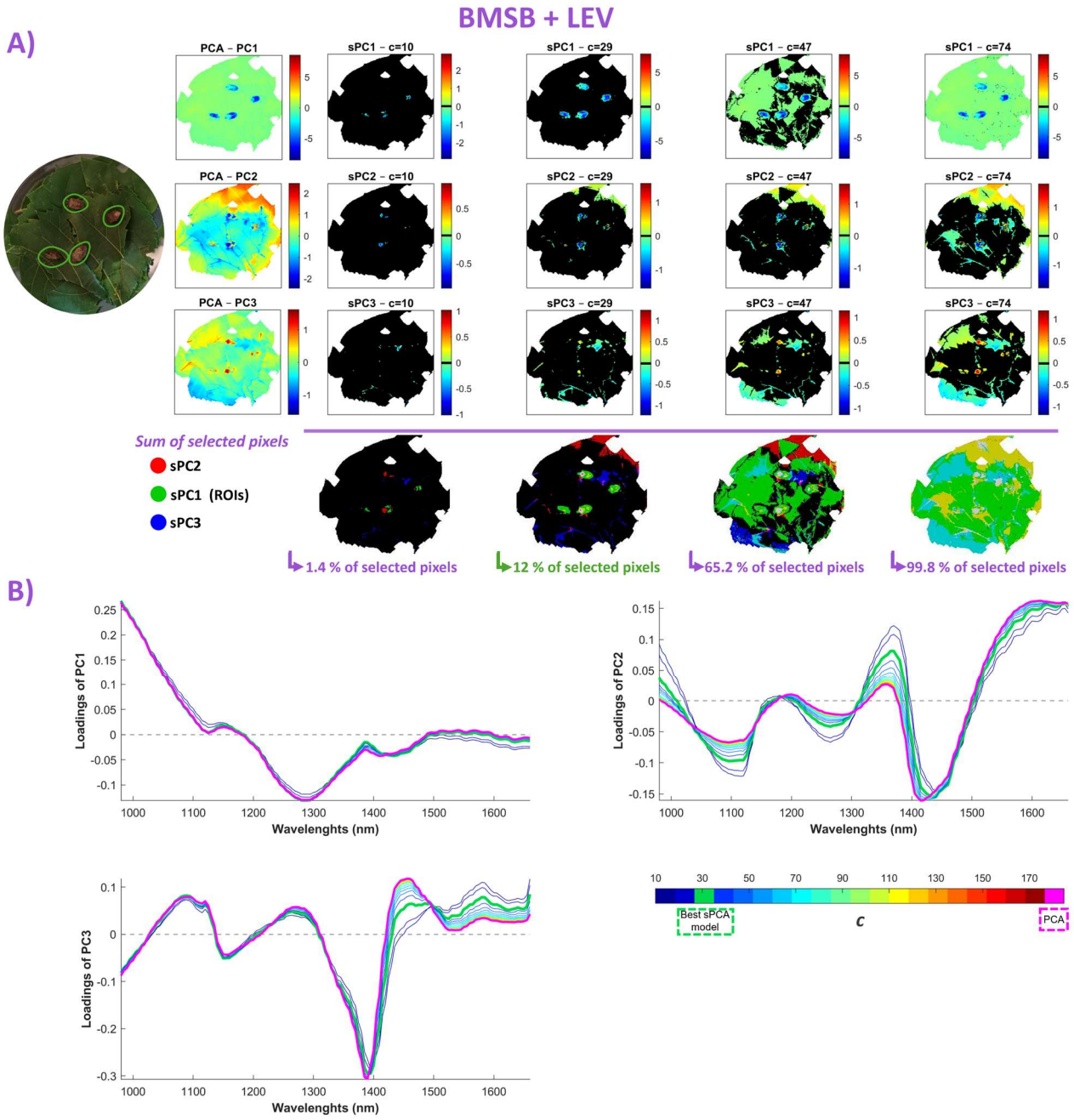


Figure 5.4 *Comparison between standard PCA and sl-sPCA for BMSB + LEV image. In A): PCA score images of PC1, PC2 and PC3 together with the corresponding score images of sl-sPCA models calculated with different values of the sparsity constraint c; the last row reports, for each sl-sPCA model, the false-colour image obtained by assigning the sPC2, sPC1 and sPC3 score images to the red, green and blue channels, respectively. The percentage of selected pixels corresponds to the percentage of image pixels that contain at least one non-zero entry in the corresponding score matrix of the sl-sPCA model. In B): loading plots of the standard PCA model and of the corresponding sl-sPCA models calculated with varying values of the sparsity constraint c.*

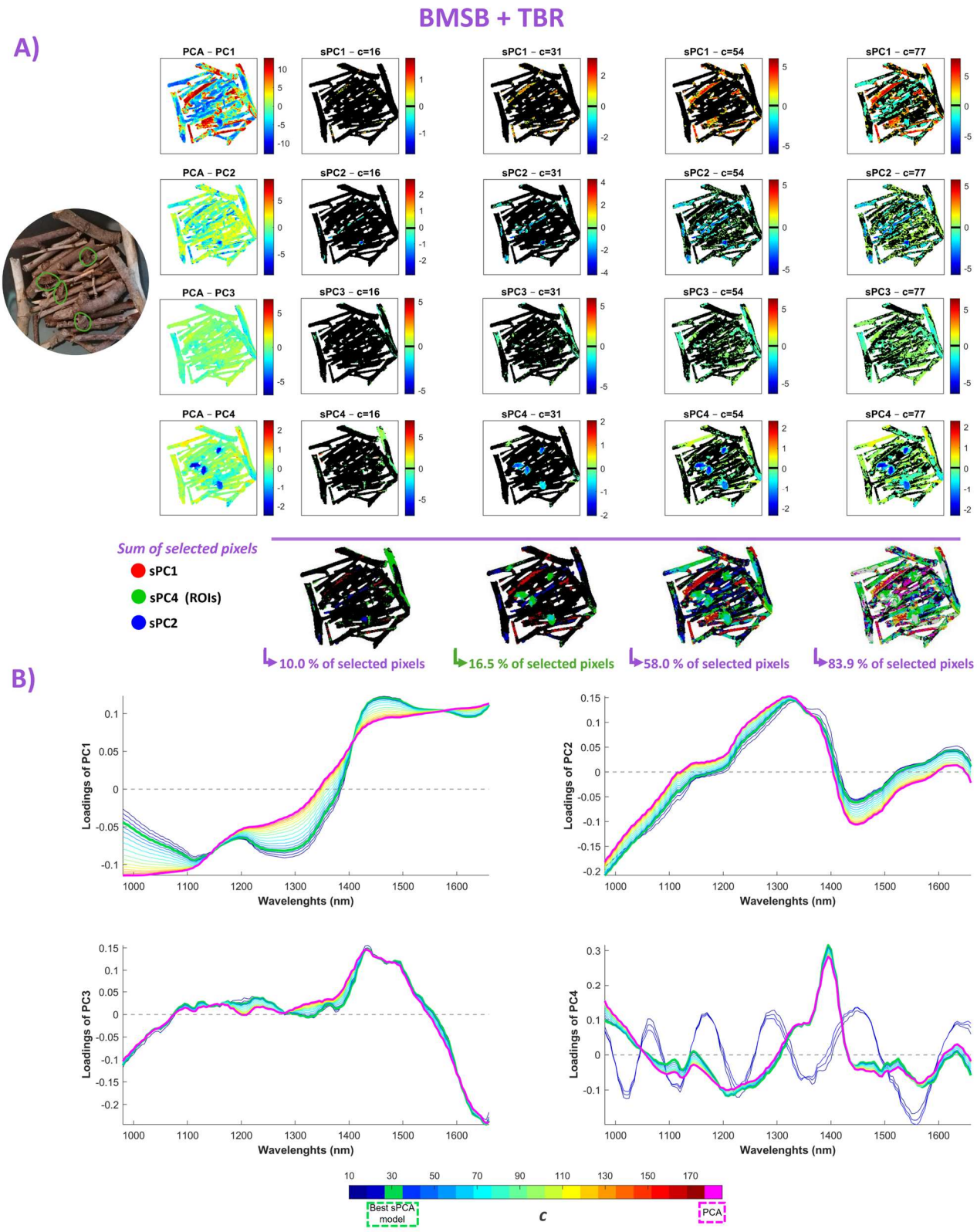


Figure 5.5 *Comparison between standard PCA and sl-sPCA for BMSB + TBR image. In A): PCA score images of PC1, PC2 and PC4 together with the corresponding score images of sl-sPCA models calculated with different values of the sparsity constraint c; the last row reports, for each sl-sPCA model, the false-colour image obtained by assigning the sPC1, sPC4 and sPC2 score images to the red, green and blue channels, respectively. The percentage of selected pixels corresponds to the percentage of image pixels that contain at least one non-zero entry in the corresponding score matrix of the sl-sPCA model. In B): loading plots of the standard PCA model and of the corresponding sl-sPCA models calculated with varying values of the sparsity constraint c.*

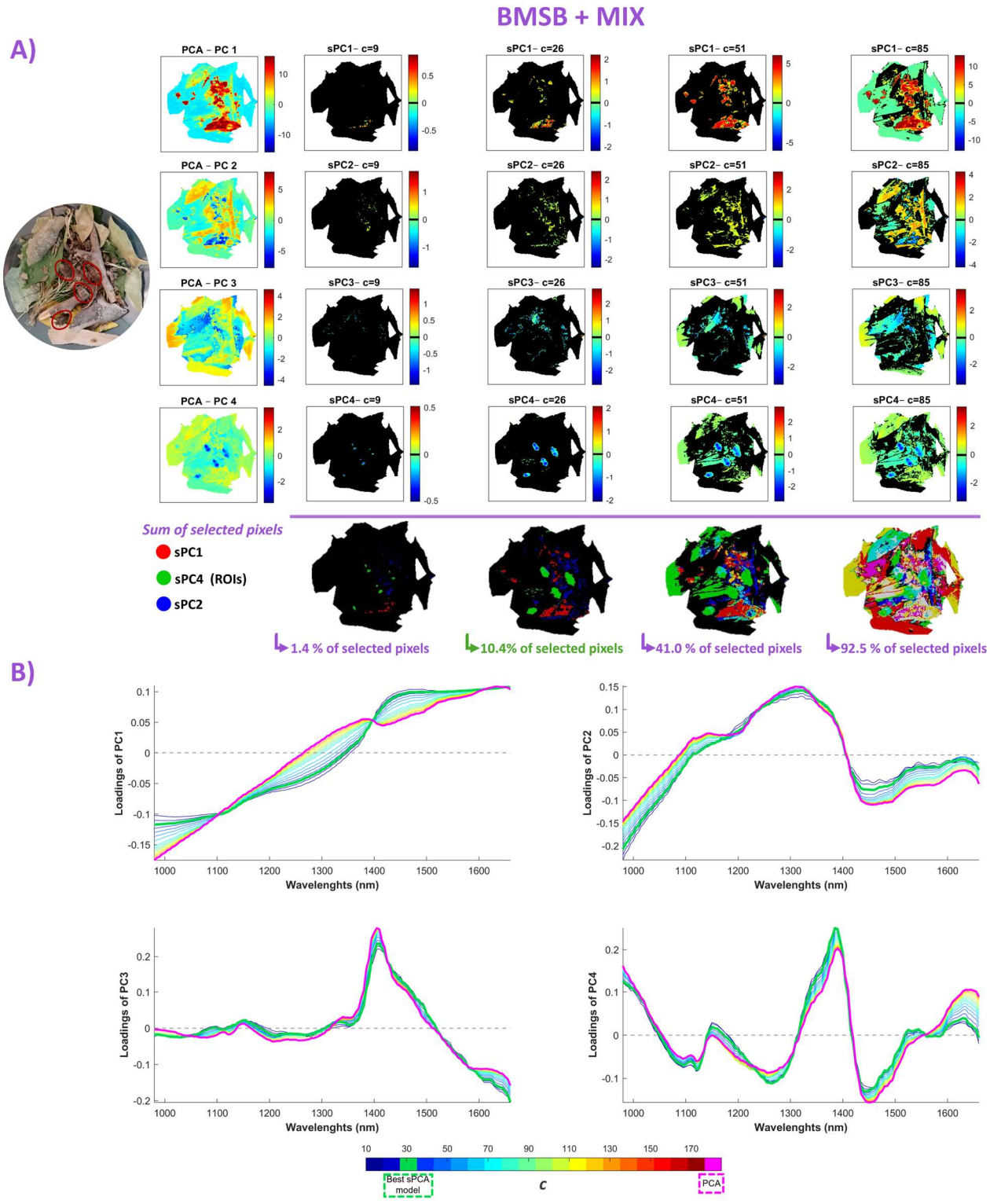


Figure 5.6 *Comparison between standard PCA and sl-sPCA for BMSB + MIX image. In A): PCA score images of PC1, PC2, PC3 and PC4 together with the corresponding score images of sl-sPCA models calculated with different values of the sparsity constraint c; the last row reports, for each sl-sPCA model, the false-color image obtained by assigning the sP1, sPC4 and sPC2 score images to the red, green and blue channels, respectively. The percentage of selected pixels corresponds to the percentage of image pixels that contain at least one non-zero entry in the corresponding score matrix of the sl-sPCA model. In B): loading plots of the standard PCA model and of the corresponding sl-sPCA models calculated with varying values of the sparsity constraint c.*

5.5. Conclusions

This study performed a preliminary evaluation on the benefits of elaborating hyperspectral images with sPCA algorithm where sparsity is applied to the spatial domain, i.e, to the scores. The effectiveness of sl-sPCA was evaluated on benchmark NIR hyperspectral images of increasing complexity, ranging from datasets of plastic pellets composed of different polymers to datasets of BMSB specimens on homogeneous or mixed vegetal backgrounds [8,23]. Particular attention was paid to the ability of the proposed approach in identifying spatially resolved ROIs in the images. The influence of varying sparsity levels was investigated, as the identification of appropriate sparsity represents a key aspect for refining the selection of pixels associated with ROIs.

Overall, sl-sPCA effectively reduced redundant information while enhancing model interpretability. This improvement was observed directly in the score images through the clear spatial localization of ROIs and, indirectly, in the loading vectors, where sparsity increased the contribution of absorption bands associated with chemical interpretation of the selected spatial features.

These results highlight sl-sPCA as a valuable exploratory tool for hyperspectral image analysis, enabling efficient data compression while preserving meaningful spatial and spectral information. As a further development, sPCA can be used as an alternative approach to select representative pixels from classes of interest to develop supervised classification models.

Moreover, further studies will be addressed to obtain a deeper knowledge of sl-sPCA properties, focusing on the quantitative assessment of model performances considering for example the explained variance and percentage of selected pixels, which strictly depend on the sparsity constraint and the abundance of pixels related to ROIs within the image.

As a final consideration, the aim of sl-sPCA is not to exactly reproduce the original data, but rather to enhance interpretability and extract meaningful, spatially localized information. As highlighted in the series *"All sparse models are wrong, but some are useful"* [11–13], sparse models represent approximations of reality that can provide valuable insight when properly applied.

References

[1] S. Grassi, C. Alamprese, Advances in NIR spectroscopy applied to process analytical technology in food industries, Current Opinion in Food Science 22 (2018) 17–21. https://doi.org/10.1016/j.cofs.2017.12.008.

[2] D. Tanzilli, M. Cocchi, J.M. Amigo, A. D'Alessandro, L. Strani, Does hyperspectral always matter? A critical assessment of near infrared versus hyperspectral near infrared in the study of heterogeneous samples, Current Research in Food Science 9 (2024) 100813. https://doi.org/10.1016/j.crfs.2024.100813.

[3] R. Calvini, J.M. Amigo, Coupling randomisation and sparse modelling for the exploratory analysis of large hyperspectral datasets, *Chemometrics and Intelligent Laboratory Systems* 248 (2024) 105118. https://doi.org/10.1016/j.chemolab.2024.105118.

[4] A. Gowen, C. Odonnell, P. Cullen, G. Downey, J. Frias, Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, *Trends in Food Science & Technology* 18 (2007) 590–598. https://doi.org/10.1016/j.tifs.2007.06.001.

[5] J. Blasco, G. Gorla, S. Munera, R. Vitale, J.M. Amigo, Non-Destructive Spectral Systems (NDSS) for modern inspection systems in real-time: challenges and industrial perspectives, *TrAC Trends in Analytical Chemistry* 191 (2025) 118369. https://doi.org/10.1016/j.trac.2025.118369.

[6] R. Bro, A.K. Smilde, Principal component analysis, *Analytical Methods* 6 (2014) 2812–2831. https://doi.org/10.1039/C3AY41907J.

[7] C. Ferrari, G. Foca, A. Ulrici, Handling large datasets of hyperspectral images: Reducing data size without loss of useful information, *Analytica Chimica Acta* 802 (2013) 29–39. https://doi.org/10.1016/j.aca.2013.10.009.

[8] R. Calvini, A. Ulrici, J.M. Amigo, Sparse-Based Modeling of Hyperspectral Data, in: *Data Handling in Science and Technology*, Elsevier, 2016: pp. 613–634. https://doi.org/10.1016/B978-0-444-63638-6.00019-X.

[9] P. Filzmoser, M. Gschwandtner, V. Todorov, Review of sparse methods in regression and classification with application to chemometrics, *Journal of Chemometrics* 26 (2012) 42–51. https://doi.org/10.1002/cem.1418.

[10] E. Andries, S. Martin, Sparse Methods in Spectroscopy: An Introduction, Overview, and Perspective, *Applied Spectroscopy* 67 (2013) 579–593. https://doi.org/10.1366/13-07021.

[11] J. Camacho, A.K. Smilde, E. Saccenti, J.A. Westerhuis, All sparse PCA models are wrong, but some are useful. Part I: Computation of scores, residuals and explained variance, *Chemometrics and Intelligent Laboratory Systems* 196 (2020) 103907. https://doi.org/10.1016/j.chemolab.2019.103907.

[12] J. Camacho, A.K. Smilde, E. Saccenti, J.A. Westerhuis, R. Bro, All sparse PCA models are wrong, but some are useful. Part II: Limitations and problems of deflation, *Chemometrics and Intelligent Laboratory Systems* 208 (2021) 104212. https://doi.org/10.1016/j.chemolab.2020.104212.

[13] J. Camacho, A.K. Smilde, E. Saccenti, J.A. Westerhuis, R. Bro, All sparse PCA models are wrong, but some are useful. Part III: Model interpretation, *Chemometrics and Intelligent Laboratory Systems* 266 (2025) 105498. https://doi.org/10.1016/j.chemolab.2025.105498.

[14] C.-Y. Cao, M.-T. Li, Y.-J. Deng, L. Ren, Y. Liu, X.-H. Zhu, Joint sparse local linear discriminant analysis for feature dimensionality reduction of hyperspectral images, *Remote Sensing* 16 (2024) 4287.

[15] W. Zhang, A. Yuan, J. Tang, X. Li, Sparse Principal Component Analysis and Adaptive Multigraph Learning for Hyperspectral Band Selection, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17 (2024) 1419–1433. https://doi.org/10.1109/JSTARS.2023.3335286.

[16] K.Y. Peerbhay, O. Mutanga, R. Ismail, Does simultaneous variable selection and dimension reduction improve the classification of Pinus forest species?, *Journal of Applied Remote Sensing* 8 (2014) 085194–085194.

[17] V. Ferrari, R. Calvini, B. Boom, C. Menozzi, A.K. Rangarajan, L. Maistrello, P. Offermans, A. Ulrici, Evaluation of the potential of near infrared hyperspectral imaging for monitoring the invasive brown marmorated stink bug, *Chemometrics and Intelligent Laboratory Systems* 234 (2023) 104751. https://doi.org/10.1016/j.chemolab.2023.104751.

[18] R. Calvini, G. Orlandi, G. Foca, A. Ulrici, Development of a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging, *Journal of Spectral Imaging* (2018) a13. https://doi.org/10.1255/jsi.2018.a13.

[19] A.E. Hoerl, R.W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics* 12 (1970) 55–67. https://doi.org/10.1080/00401706.1970.10488634.

[20] M.A. Rasmussen, R. Bro, A tutorial on the Lasso approach to sparse modeling, *Chemometrics and Intelligent Laboratory Systems* 119 (2012) 21–31. https://doi.org/10.1016/j.chemolab.2012.10.003.

[21] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58 (1996) 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[22] H. Zou, T. Hastie, Regularization and Variable Selection Via the Elastic Net, *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67 (2005) 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[23] R. Bro, E.E. Papalexakis, E. Acar, N.D. Sidiropoulos, Coclustering—a useful tool for chemometrics, *Journal of Chemometrics* 26 (2012) 256–263. https://doi.org/10.1002/cem.1424.

[24] I.T. Jolliffe, N.T. Trendafilov, M. Uddin, A Modified Principal Component Technique Based on the LASSO, *Journal of Computational and Graphical Statistics* 12 (2003) 531–547. https://doi.org/10.1198/1061860032148.

[25] H. Zou, T. Hastie, R. Tibshirani, Sparse Principal Component Analysis, *Journal of Computational and Graphical Statistics* 15 (2006) 265–286. https://doi.org/10.1198/106186006X113430.

[26] D.A. Burns, E.W. Ciurczak, eds., Handbook of Near-Infrared Analysis, 3rd ed., CRC Press, 2007. https://doi.org/10.1201/9781420007374.

[27] N. Mobaraki, J.M. Amigo, HYPER-Tools. A graphical user-friendly interface for hyperspectral image analysis, *Chemometrics and Intelligent Laboratory Systems* 172 (2018) 174–187. https://doi.org/10.1016/j.chemolab.2017.11.003.

Chapter 6
General conclusions

The investigations carried out in this PhD Thesis focused on the development and application of sustainable and non-destructive analytical strategies for food quality and safety assessment based on spectral imaging systems and chemometric analysis. In modern agri-food systems, Near-Infrared Hyperspectral Imaging (NIR-HSI) has emerged as a promising tool for non-destructive inspection, as it combines the advantages of spectroscopic measurements with those of imaging techniques. However, the high cost and sensitivity of optical components, together with computational demands required to manage the large volume of data generated, still limit the widespread adoption of NIR-HSI as a process analytical technology.

Within this framework, the present PhD Thesis aimed at applying chemometric strategies to overcome the curse of dimensionality and to extract meaningful information from high-dimensional image datasets. Accordingly, three multivariate image analysis approaches were investigated: feature selection, enabling the identification of relevant spectral and spatial features; feature extraction, aimed at data compression; and supervised classification using Soft PLS-DA, a soft discriminant algorithm which proved to be particularly flexible and robust . These approaches were applied to the development of practical solutions for diverse objectives, including pest monitoring, post-harvest fruit sorting, and product authentication.

The potential of NIR-HSI as a tool for both pest monitoring and post-harvest fruit sorting was assessed in the context of the management of the Brown Marmorated Stink Bug (BMSB), an invasive pest severely affecting pear production. For field monitoring applications, sparse-based variable selection combined with Soft PLS-DA enabled the identification of the most informative spectral regions for discriminating BMSB specimens from the surrounding environment, paving the way for the future implementation of multispectral imaging systems more suitable for on-field applications.

Concerning post-harvest quality assessment, a substantial part of the research activities was dedicated to the investigation of under-peel damages on pears caused by BMSB feeding, which are not detectable using imaging systems operating in the visible range. Similarly to pest detection, sparse-based variable selection coupled with Soft PLS-DA was successfully applied to discriminate punctured and sound areas at both pixel-level and at image-level, providing a framework compatible with the requirements of automated sorting systems.

To this aim, the annotation of Regions of Interest (ROIs) associated with BMSB punctures was a crucial step for building datasets of representative spectra related to sound and damaged areas. To

address this issue, a novel automated annotation strategy was proposed, combining *Common Space Hyperspectrograms* with image-level classification and variable selection. This approach enabled the objective identification of ROIs associated with punctures and their visualization in the original image domain.

In the context of food authentication, the soft discriminant algorithm Soft PLS-DA was proposed as a viable classification strategy capable of handling outlier samples while maximizing differences between classes. NIR-HSI was evaluated as a screening tool for oregano authentication by comparing supervised classification models based on a traditional class modelling approach, i.e. Alt-SIMCA, and Soft PLS-DA. The best overall results were achieved with Soft PLS-DA, demonstrating its suitability for handling strongly overlapping classes. Moreover, a 10% threshold was established as the detection limit of NIR-HSI, allowing reliable discrimination between authentic and adulterated oregano samples.

To further address the computational limitation associated with NIR-HSI, the potential of sparse methods as an alternative feature selection strategy for the identification of pixels associated with ROIs was explored. In this context, sparse Principal Component Analysis (sPCA) with sparsity imposed on the score vectors (score-level sPCA) was evaluated on benchmark NIR hyperspectral datasets of increasing complexity, ranging from plastic pellets composed of different polymers to images of BMSB specimens on homogeneous and mixed vegetal backgrounds. This approach effectively reduced redundant information while enhancing model interpretability and preserving meaningful spatial and spectral features. Overall, the results highlight sl-sPCA as a valuable exploratory tool for hyperspectral image analysis, enabling clear spatial localization of ROIs on score images and, indirectly, enhancing the contribution of chemically relevant absorption bands in the loading vectors.

Overall, the chemometric strategies adopted throughout this thesis consistently favoured parsimonious and interpretable models, thereby simplifying their transferability to multispectral imaging systems that are far more suitable for on-line applications. Moreover, the results presented demonstrate that the integration of NIR-HSI with advanced, interpretable chemometric strategies provides an effective and sustainable framework for addressing complex agri-food challenges, enabling the development of practical tools for monitoring, sorting, and authentication.