Combination of multivariate curve resolution with factorial discriminant analysis for the detection of grapevine diseases using hyperspectral imaging. A case study: flavescence dorée

Sílvia Mas Garcia, Maxime Ryckewaert, Florent Abdelghafour, Maxime Metz, Daniel Moura, Carole Feilhes, Fanny Prezman, Ryad Bendoula

Combination of Multivariate Curve Resolution with Factorial Discriminant Analysis for the detection of grapevine diseases using Hyperspectral imaging. A case study: Flavescence Dorée.

Silvia Mas Garcia[1,2*], Maxime Ryckewaert[1,2], Florent Abdelghafour[1], Maxime Metz[1,2], Daniel Moura[1], Carole Feilhes[3], Fanny Prezman[3], Ryad Bendoula[1].

[1]ITAP, INRAE, Institut Agro, University Montpellier, 34196 Montpellier, France
[2]ChemHouse Research Group, 34196 Montpellier, France
[3]IFV, 1920 Route de Lisle-sur-Tarn, 81310 Peyrole, France

[*]Author for correspondence:

E-mail: silvia.mas-garcia@inrae.fr

Abstract

Hyperspectral imaging is an emergent technique in viticulture that can potentially detect bacterial diseases in a non-destructive manner. However, the main problem is to handle the substantial amount of information obtained from this type of data, for which reliable data analysis tools are necessary. In this work, combination of multivariate curve resolution-alternating least squares (MCR-ALS) and factorial discriminant analysis (FDA) is proposed to detect the Flavescence dorée grapevine disease from hyperspectral imaging.

The main purpose of MCR-ALS in this work was providing chemically meaningful basic spectral signatures and distribution maps of the constituents needed to describe both healthy and infected images by Flavescence dorée. MCR scores (distribution maps) were used as starting information for FDA to distinguish between healthy and infected pixels/images. Such an approach is presumably more powerful than the direct use of FDA on the raw imaging data, since MCR scores are compressed and noise-filtered information on pixel properties, which makes them more suitable for discrimination analysis. High levels of correct pixels discrimination rates (CR=85,1%) for the MCR-ALS/FDA discrimination model were obtained. The model present a lesser ability to determine infected leaves than healthy leaves. Nevertheless, only two images were misclassified. Therefore, proposed strategy constitutes a good approach for the detection of the Flavescence dorée that could be potentially used to detect other phytopathologies

Keywords: Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS), Factorial Discriminant Analysis (FDA), Hyperspectral imaging, vineyard diseases, Flavescence dorée

1. Introduction

Epidemiological surveillance is a crucial issue in agriculture and especially in viticulture. As a matter of fact, the grapevine (*Vitis vinifera*) is sensitive to a wide range of biopests. To cope with these threats, preventive chemical control is required. To reduce the use of chemical inputs while ensuring the protection of the vineyard, it is necessary to implement more parsimonious spraying practices. The development of sustainable crop protection systems is closely related to the knowledge regarding the physiological state and the health status of the vineyard [1].

To date, the evaluation of sanitary risks is conducted by visual and tactile inspection, which is time consuming and labour intensive. The analysis of light-matter interaction can provide information related to physiological properties such as hydric status, nitrogen content, pigmentation or even cellular structure [2]. Therefore, optical instruments and especially multispectral (MSI) and hyperspectral imaging (HSI) are relevant tools for the automated and non-invasive detection of phytopathology [3–6]. In this context, conventional analysis of hyperspectral and multispectral images, such as determination of spectral vegetation indices (SVIs) [7], performs only a limited use of the substantial amount of information available with this type of data. Therefore, in order to successfully interpret these images, the application of advanced data processing tools is necessary. In this work, we will focus on the application of HSI to discriminate an important vine disease: the Flavescence dorée" (FD, also known as "yellowing").

FD is a phytopathology caused by the bacteria *Candidatus phytoplasma vitis* that can spread fast through a leafhopper (*Scaphoideus titanus*). It represents a very serious threat, since without proper management; it can lead to the complete loss of the harvest or even the death of the vine stocks. Recently, spectral imaging have been used to detect FD. Albetis et al. [6] evaluated the potentiality of Unmanned Aerial Vehicle (UAV) multispectral imagery for the airborne detection of FD symptoms under field conditions. For this purpose, they analysed

several spectral bands, vegetation indices, and biophysical parameters. However, the specific detection of FD appears to be limited. Al-Saddik et al. [5] used a portable spectroradiometer (350–2500 nm) to collect hyperspectral reflectance data of healthy and symptomatic leaves. The aim of this study was to develop specific spectral disease indices (SDIs) for the detection of FD disease in grapevines, thereby, reaching discrimination accuracies of more than 90%. However, the SDIs were dependent on the disease infestation state and the grapevine variety considered; the best wavelengths selected were different from one case to another, and hence no single best index for FD in all situations was identified. To deal with these limitations, this work aims to propose a new general methodology (i.e. not depending of the variety) to discriminate between healthy and infected leaves based on HSI measurements and data analysis methods.

The data analysis workflow proposed in this work relies on two steps:

a) Multivariate curve resolution-Alternating Least Squares (MCR-ALS) [8,9] model to provide chemically meaningful spectral signatures and related distribution maps of the image constituents. This unmixing method allows a global differentiation between infected and healthy images. However, some components related to the two different class types (infected and healthy pixels) may overlap, and hence, a supervised discrimination method is necessary to achieve a harder separation between them.
b) Factorial discriminant analysis (FDA) [10] model on MCR scores (distribution maps). This supervised classifier will help to discriminate between infected and healthy images, using previous pixel labelling (classes infected or healthy).

Previous studies have already shown the capability of the application of MCR-ALS combined with supervised classification methods to the analysis of imaging data [11,12]. These works demonstrated that the use of MCR outputs as starting information for classification methods allows a compound-wise selection and preprocessing of the input information to be submitted to the classification algorithm. This is due to the fact that MCR results are chemically

meaningful and express concentrations or spectra of the pixel constituents in the images. Such a specificity of the method allows discarding components related to background signal contributions in the classification task. In this work, FDA is chosen as the supervised classifier because it is one of the simplest and fastest approach for discrimination that has proven its efficiency for various analytical chemistry applications [13,14]. However, to the best of our knowledge this is the first time that MCR-ALS combined with such a classification method is used as phytopathology detection model. This study demonstrates that the proposed methodology has the potential to improve disease detection in agriculture applications.

2. Material and Methods

2.1 Samples

Leaves were collected during September 2020 on previously identified plots with Flavescence Dorée. All cultivars were sampled with a similar proportion of red and white varieties. In total 109 leaves were collected on the field. The number of leaves from the different varieties selected for this study are summarised in Table 1.

TABLE 1

Infected leaves were chosen in order to represent at best the variability of the available symptoms in terms of severity and stages of infections. Leaves were selected when foliar symptoms were undoubtedly caused by FD from vines exhibiting clear symptoms of FD on other organs. Each leaf and each vine from which they were extracted were diagnosed by a phytopathology expert. Leaves were extracted from the front face, in the middle of the canopy so that to avoid the younger and older organs which can present different physiological

behaviour. Regarding the healthy leaves, they were selected in the same regions and they were asserted absent of symptoms of FD or any other visible pathology. However, some of the healthy sample can exhibit light forms of mechanical or chemical wounds (due to protection, management operations) and some slight damage caused by insects.

2.2 Image acquisition

Acquisitions of leaf images were performed with a hyperspectral camera (IQ, Specim, Finland). Imaging of grapevine leaves was carried out in the spectral range of 400-900 nm, with a spectral resolution of 7 nm. Images in RGB were also registered. Illumination was provided by a halogen lamp (Arrilite 750 Plus ARRI, Munich, Germany). Constant angles of -50° and 50° were maintained between the halogen lamp and the hyperspectral camera. These angles were chosen to optimise the intensity of the reflected beam and to reduce specular reflection.

For each sample image, the intensity of the reflected light ($I(\lambda)$) was measured. The Dark current ($I_n(\lambda)$) *i.e.* signal without light, was recorded from all measured spectra and then subtracted. A white reference (SRS99, Spectralon®) ($I_0(\lambda)$) was measured to standardise spectra and prevent nonlinearities of all the instrumentation components (light source, lens, fibbers and spectrometer). From these measurements, a reflectance image ($R(\lambda)$) was calculated for each sample, as follows:

$$R(\lambda) = \frac{I(\lambda) - I_n(\lambda)}{I_0(\lambda) - I_n(\lambda)} \qquad \text{Equation 1}$$

3. Data analysis

The proposed workflow for data analysis follows the three following steps:

a) **Image preprocessing**

b) **MCR** to recover basic spectral signatures and distribution maps of pure compounds contributions, allowing differentiation between infected and healthy images.

c) **FDA model using the MCR scores (concentration profiles) resulting from the MCR results** to predict the class (infected or healthy) of the images.

These steps are described in detail in the following subsections

3.1 Image preprocessing

In HSI (Hyperspectral Imaging), the generated data can be arranged into a data cube in which the x-and y-axis correspond to the pixel coordinates and the z-axis corresponds to the wavelengths values registered in each pixel. Data preprocessing is required to improve the signal quality and to compress the acquired raw data for further analysis.

Firstly, the pixels in the images were binned by a factor of 4 in *x* and *y*. This spatial binning produced an image of 128x128 pixels from an original image of 512x512. Afterwards, a mask was create for each image to extract only the vegetation pixels. The Spectral Angle Mapper (SAM) [15] was used for this purpose. To identify vegetation pixels, SAM compare image spectra to a reference spectrum by calculating the spectral angle between them. Smaller angles represent closer matches to the reference spectrum, and hence the corresponding pixels are classified as vegetation pixels, whereas pixels further away than the specified maximum angle threshold are not classified.

Finally, A matrix **Di** (*n,m*) of dimension *n* equal to ($x \times y$) pixels by 175 wavelengths was generated per each image.

3.2 Multivariate Curve Resolution/ Factorial Discrimination Analysis (MCR/FDA) model

Before using MCR-ALS and FDA methods, the dataset was divided into two sets of samples: training and independent test sets. The training sets were used to build the models. The test sets were left for external validation and are not used to build the models. Healthy and infected images were both divided with the same split ratio of 2/3 and 1/3 respectively for training and

test, as detailed in Table 1. This division was made randomly and assuring a similar distribution of all classes in both training and test sets.

3.2.1 Multivariate Curve Resolution- Alternating Least Squares (MCR-ALS)

The goal of the MCR-ALS algorithm is the decomposition of the image data **D** into distribution maps (relative amounts or concentration) and pure spectra of the constituents present in the imaged sample [8,9,16]. In matrix form, the hyperspectral images can be described by a bilinear model based on the Beer-Lambert law (Equation 1). Where the matrix **D** contains the pixel spectra obtained after the preprocessing described in section XX. Each spectra is then decomposed into a set of concentration profiles (**C** matrix) corresponding to pure spectra ($\mathbf{S^T}$ matrix) of the constituents present in the image. **E** is the matrix associated with noise or experimental error (residuals).

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \qquad \text{Equation 2}$$

Figure 1 shows the application of MCR-ALS to an individual image data **D**. It can be observed that every row of the resolved $\mathbf{S}^T$ matrix corresponds to the pure spectrum of an image constituent, while every column of the resolved **C** matrix of concentration profiles corresponds to the related pixel-to-pixel variation of its chemical concentration. It is worth mentioning that each column of the resolved **C** matrix can be refolded appropriately in order to recover the original two-dimensional spatial image structure and then pure distribution maps are obtained.

FIGURE 1

In order to recover the bilinear model expressed in Equation 1, MCR-ALS begins with determining the number of signal contributions in the original data set **D** by Singular Value

Decomposition (SVD) [17]. Afterwards, an initial **C** or $\mathbf{S^T}$ matrix with as many profiles as the number of components estimated for **D** is constructed to initiate the iterative resolution process. In this work, the initial $\mathbf{S^T}$ was generated by a pure variable selection method based on Simple-to-use Interactive Self-modelling Mixture Analysis (SIMPLISMA) [18]. Such estimate $\mathbf{S^T}$ and the matrix **D** are used to initialise the least squares alternating optimisation of the profiles in matrices **C** and $\mathbf{S}^T$ of the bilinear model under the constraints until convergence is achieved. The convergence criterion can be a maximum number of iterations or a value related to the difference in fit improvement between consecutive iterations.

The quality of the MCR results are described by the explained variance (% r^2), which are calculated according to the following expressions:

$$\% \, r^2 = 100 \times \left(1 - \frac{\Sigma e_{ij}^2}{\Sigma d_{ij}^2}\right) \qquad \text{Equation 3}$$

where e_{ij} is equal to d_{ij}-d_{ij}^*, d_{ij}^* are the values of the data set reproduced by the bilinear model and d_{ij} the original values in the original data set **D**. In order to consider that MCR results of an analysis are adequate, the variance explained must be sufficiently high and the concentration profiles and spectra obtained must be chemically meaningful and show shapes consistent with the variation in the original data sets.

MCR-ALS can also be used to analyse simultaneously several images in a single multiset structure to provide more reliable results [9,19]. Resolved features would define much better general traits analysed together than if they were analysed individually. In this study, the multiset structures were obtained by setting different images **Di** one on top of each other to form a column-wise augmented matrix **Daug**. The bilinear model in Equation 1 is now extended to the augmented data set as shown in Equation 4:

$$\mathbf{Daug=[D_1;D_2;...;D_n]=[C_1;C_2;...;C_n]S^T+[E_1;E_2;...;En]=CaugS^T+Eaug} \qquad \text{Equation 4}$$

where **Caug** is a column-wise augmented matrix formed by as many submatrices **Ci** as images in the multiset, and $\mathbf{S^T}$ is a single data matrix of pure spectra, assumed to be common and valid for all the images in the multiset. The concentration profiles in each of these submatrices can be also refolded conveniently to recover the related distribution maps of each image (see Figure 1b).

The MCR-ALS analysis of a single image or an image multiset takes the benefit of the use of constraints on **C** or/and $\mathbf{S}^T$ to obtain chemically meaningful and more accurate spectral signatures and distribution maps. In this study, the most common constraints in image resolution, such as non-negativity and normalisation, were used. Moreover, the constraint of correspondence among species to encode the information related to the presence/absence of some components in the different **Ci** submatrices in the multiset structures was also applied [9,16].

MCR-ALS distribution maps (**C** matrix) and pure spectra ($\mathbf{S^T}$ matrix) are excellent low dimension, noise-filtered meaningful basis of the pixel and the spectral space of the image, which may be further used to obtain additional information. In this work, the MCR scores (distribution maps) were fed into the FDA to predict the type-class (healthy or infected) of the images.

It is worth mentioning that a multiset structure containing all the training dataset from both healthy and infected images (**DtrFH**) was used for the MCR approach. Then, the distribution maps related to the multiset structure containing all the test dataset (**DtestFH**) were calculated by a single non-negative least-squares step taking MCR pure spectra obtained in the training stage ($\mathbf{S^Ttr}$).

3.2.2 Factorial Discrimination Analysis (FDA)

The aim of FDA [10] is to predict the membership of an individual to a group of samples according to pre-defined groups. This method searches for relationships between a qualitative variable

(healthy or infected) and a group of quantitative explanatory variables (wavelengths, intensities…). The use of the qualitative variable within a population allows the division of this population into different groups, with each individual assigned to one group. Discrimination of the groups consists of maximising the variance between their gravity center. For each group, the distance from the different gravity center of the groups is calculated and then, the sample is assigned to the group where its distance between the centre of gravity is the nearest. Comparison of the assigned group to the real group is an indicator of the quality of the model, and hence, discrimination rate (CR) is taken as a criterion of goodness for the developed model

In this work, FDA was performed to determine the affiliation of each pixel/image whether to the healthy or to the infected class. High correlations can occurred among the wavelengths or intensities of the pixels/images, therefore, MCR scores (distribution maps) coming from the augmented **CtrFH** matrix obtained by MCR-ALS have been used as the pixel input information for FDA. Therefore, no variable reduction algorithm such as PCA or ICA need to be done due to the fact that MCR scores (concentration profiles) are compressed and noise-filtered information on pixel properties. The gravity centre of each sample type in the model was calculated from these training sample scores. The Mahalanobis distance [20] from each to each level of the gravity centres was measured. Finally, test samples were assigned to the group with the nearest gravity centre.

3.3 Software

All data processing has been performed in MATLAB platform (Version 2015b, MathWorks Inc., Natick, MA, USA). The application of MCR-ALS has been performed using the MCR GUI (multivariate curve resolution graphical user interface) developed by the chemometrics group of Universitat de Barcelona and IDAEA-CSIC [21], which is can be downloaded from the MCR webpage http://www.mcrals.info/. FDA analysis method has been applied using in-house routines, partly based on the PLS Toolbox (Eigenvector Research Inc., Manson, WA, USA).

4. Results and discussion

4.1. MCR. Global differentiation between infected and healthy images

The first MCR-ALS analysis was focused on identifying significant contributions with a specific reflectance signature for each leaf type (healthy and infected). For that purpose, two multisets were built, one formed by the 47 training images corresponding to the infected leaves from all varieties (**DaugtrF**), and the other multiset formed by the 25 training images corresponding to the healthy leaves from all varieties (**DaugtrH**). MCR-ALS was applied separately to each of these multiset structures using non-negativity constraints in concentration and spectra profiles and spectra normalisation.

Table 2 summarises the number of resolved components and the explained variance obtained from the MCR-ALS analyses of both multisets. Resolution of three contributions was necessary in both cases. The inclusion of a different number of contributions gave solutions worse mathematically or unreliable spectra or distribution maps.

TABLE 2

Figure 2a and b show the MCR-ALS resolved distribution maps (with their corresponding RGB images) and pure spectra of each analysed multiset, respectively. To simplify, resolved distribution maps of only one image per variety is shown. It can be seen that the blue and red contributions present resolved pure spectra rather similar in both multisets, with a Pearson correlation coefficient higher than 0.90. The blue contribution shows a low intensity plateau in visible region (from 400 to 700 nm) and then an increment of the intensity in the near infrared region that ends rather stable from 750 to 900 nm. This contribution seems to present the typical

profile related to the cell structure of the leaf. The red contribution presents a peak at 550 nm and a low intensity between 600 and 640 nm, which could correspond to the pigment content, especially anthocyanins and chlorophyll. The green contribution presents a greater spectral dissimilarity between the two multisets. Remarkably, the component from the infected leaf multiset (Figure 2a) has a characteristic peak located at 700 nm, a second peak located at 650 nm and lower intensity values at 400 nm and 500 nm. This green contribution could be attributed to a difference in slope level in the red-edge region, an imbalance between chlorophyll a and chlorophyll b, and an appearance of carotenoids. For the healthy multiset (Figure 2b), the green present values in the visible region that oscillate between 0.6 and 0.4 and the slope in the near infrared region increases from 750 to 800 nm. Therefore, this component seems to reflect an intensity level in the pigment region.

The distribution maps use a graduated colour scale per column, where the blue colour corresponds to small concentration values and the red colour to large values. Differences between the scores of white and red wine varieties can be observed for the infected multiset (Figure 2a). Unlikely, there is no visible differences between the white and red grape varieties on the healthy multiset (Figure 2b). This seems to show that the spectra obtained vary according to the grape variety at the onset of the disease. For example, the white varieties (Chardonnay, Colombars and Loin de l'oeil) have abnormally high values for the red component (figure 2a) might due to the fact that these varieties have low anthocyanin levels but still retain the Chlorophyll pigments. Therefore, it could explain why these leaves retain their green colouring in contrast to the red grape varieties (see RGB images in Figure 2). Indeed, very low scores for the third component will translate into a redder and greener colouration of the leaves.

FIGURE 2

Once, the basic spectral signatures that differentiate between infected and healthy images are resolved, MCR-ALS analysis of the multiset formed by both infected and healthy training images (**DaugtrFH**) was performed. In this case, the correspondence among species constraint was also used since the presence/absence of constituents in each sample was known. From this information, a matrix containing 72 blocks, (representing the 47 infected and the 25 healthy training images analysed simultaneously) and 4 columns (representing the number of constituents: both common blue and red contributions and the specific contributions for each multiset) coding the presence (1) or absence (0) of each constituent in each image was introduced as information in the resolution process. The absent constituents in the image were then forced to have null concentration profiles.

	1	**2**	**3**	**4**
47 blocks Flavescence Dorée	1	1	1	0
25 blocks Healthy	1	0	1	1

Figure 3 shows the MCR-ALS resolved distribution maps (corresponding to the same images in Figure 2) and pure spectra of the **DaugtrFH** multiset. The resolved spectra in Figure 3 are rather similar to the pure spectra obtained from the MCR-ALS analyses of both infected and healthy multisets (see Figure 2). Blue distribution maps refer to absent constituents in images and the rest of the maps are consistent with those obtained in Figure 2, matching the relative concentration of the different constituents in the images. A rather similar fit to previous MCR-ALS analysis (see Table 2) was obtained (r2% = 99.96), strongly supporting the MCR results. The introduction of the correspondence among species constraint does not perturb the natural

behaviour of the dataset. On the contrary, it improves the accuracy of the resolved profiles and reduces ambiguity.

FIGURE 3

In order to validate this model, distribution maps related to the multiset structure containing all the test dataset (**DtestFH**) were calculated by a single non-negative least-squares step using the pure spectra obtained in Figure 3 (**S^TtrFH**). Satisfactory results (calculated r2%= 99.95) with consistent distributions maps (data not shown) are obtained, validating the MCR results.

Now, the basic spectral signatures and distribution maps of pure compounds contributions of both infected and healthy images can be estimated. However, some components related to the two different class types (infected and healthy pixels) may overlap. Thus, this unsupervised method is not sufficient to distinguish between these two class-type, and hence an appropriate method for discrimination is required. Therefore, the MCR scores of both training (**CaugtrFH**) and test (**CaugtestFH**) sets were used for discriminant analysis. In practice, the FDA enables to determine the relation between these scores and the most probable class of the samples.

4.2. FDA model. Class assignement at pixel and leaf scale.

The FDA model is calibrated based on the MCR scores of training dataset (**CaugtrFH**). Once the MCR/FDA model is estimated, it was used to predict the class (infected or healthy) of each pixel in the test dataset (CaugtestFH). At pixel level, the discrimination rate (CR) of test set is equal to 85.1%. Both infected and healthy test pixels were correctly classified into their corresponding class with more than of 75% and 95% in accuracy, respectively. For infected pixels, a lower CR value is obtained, consequently, the model present a lesser ability to determine infected than healthy pixels. This can be attributed to the labelling process. Indeed,

leaves affected by Flavescence dorée are entirely labelled as infected, *i.e.* every pixel of the leaf is labelled the same. Indeed, at early stages, infected leaf images most likely include healthy pixels or pixels presenting slight symptoms that were labelled in a single infected class and then used in the calibration. Therefore, the model for FD could be depreciated by the presence of healthy samples, hence the lesser accuracy in discrimination.

In order to better evaluate the capacity of the MCR/FDA strategy to discriminate between infected and healthy leaves, pixel-wise decisions are summarised at the scale of the leaf. Indeed the chemical information is relevant/ consistent at the scale of the spectrum/pixel. However, on a pythopathological view, it is more sensible to consider at the scale of an organ (*i.e.* the leaf in this case). Considering the characteristics of the symptoms and the development of the disease, it is proposed to consider that a leaf is infected if more than 50% of its pixels are classified as such. Therefore the CRs for images are calculated as the percentage of correct predictions to the total number of pixels for each image (see Table 3). On the other hand, a healthy leaf should exhibit in total very few abnormal pixels. Therefore, it is considered that a leaf is healthy if 75% of its pixels are healthy (to take into account the fact that some part of the leaf could be less vigorous but still unaltered by any disease).

TABLE 3

From Table 3, a satisfactory CR higher than 74 % for all healthy leaves can be observed except for the image ***DtestgH2*** (63.3 %). Similar results are obtained for infected leaves. Only the image ***DtestcoF2*** presents a CR lower than 50 %. However, ***DtestgF7, DtestfF1, DtestcF1, Dtest1F2*** and ***DtestF3*** show also CR lower than 60 %. This lesser accuracy for infected images could be explained by the greater variability induced by the diversity of severity stages of the pathology. In addition, the model was calibrated using a single multiset, that included diverse

red and white grape varieties that exhibit different visible symptoms. Nevertheless, since discrimination results obtained for almost all images are satisfactory, the MCR/FDA strategy could be considered adequate and future leaves are expected to be properly classified into their corresponding class.

For a better evaluation of the lowest CR results of both **DtestcoF2** and **DtestgH2** images**,** Figure 4 a and b shows their predicted distribution maps (**CtestcoF2** and **CtestgH2**, respectively) alongside their corresponding RGB images. It can be seen that **DtestcoF2** which presents only 25.2 % of infected pixels exhibits early and slight symptoms (as shown by a general low-level green hue), and hence its uncertain state could explain its low accuracy. Likewise, **DtestgH2** image does not have the appearance of a healthy leaf due to the presence of some stains (possibly confounding factors such as stresses or fungal diseases). This suggests that possibly a binary discrimination assignment without an external class for confounding factors is insufficient for this application, which could also explain the lower CR for this particular leaf. Moreover, as way of example Figure 4 c and d shows two examples of good predictions for both **DtestgF1** and **DtestgH1** images*.*

FIGURE 4

In summary, it can be said that the combination of HSI and the method MCR-ALS with FDA model proved to be efficient to distinguish between infected and healthy images. However, to evaluate the discriminant potential of the proposed approach, larger data sets showing a greater variability of symptoms and infection stages is required. Moreover, confounding factors such as abiotic stresses or other phytopathology exhibiting similar symptoms should be also tested.

401 Ultimately, the processing of images representing canopies rather than isolated single leaves
402 should be taken into consideration to guarantee its feasibility in field conditions.

403

404 **5. Conclusions**

405 The strategy of combining MCR-ALS and FDA proved its interest for the discrimination
406 between healthy and infected leaves by Flavescence dorée based on the use of hyperspectral
407 images. For the first time, this strategy was applied as a phytopathology detection approach.

408 MCR-ALS enables to extract some relevant signatures that can discriminate healthy leaves from
409 leaves infected by Flavescence dorée. The pure component resulting from this model can be
410 interpreted concerning the visible symptoms of FD and to some associated physiochemical
411 disruptions. The relative abundances of these components within the leaves (MCR scores) can
412 be processed with FDA and provide an efficient discrimination of the leaves.

413 To improve the proposed strategy and reach a practical application in viticulture, some aspects
414 such as confounding factors, progressive infection stages and feasibility in the field should be
415 taken into account. Another development to improve these results, would be to upgrade the
416 labelling process, e.g. by selecting areas of the leaves clearly identified as infected rather than
417 assigning a class to the whole leaf. Nonetheless, Hyperspectral imaging combined with the
418 proposed data processing approach has the potential to be a valuable strategy to detect grapevine
419 diseases.

420 **6. Acknowledgements**

421 This work was supported by the INTERREG SUDOE SOE3/P2/E0911 Viniot projet

422 **7. Conflict of interest**

423 The authors report there are not conflicts of interest

8. Bibliography

1 E. Tona, A. Calcante and R. Oberti, *Precis. Agric.*, 2018, **19**, 606–629.

2 G. A. Carter and A. L. A. N. K. K. Napp, 2001, **88**, 677–684.

3 A. Mahlein, U. Steiner, C. Hillnhütter, H. Dehne and E. Oerke, 2012, 1–13.

4 S. Sankaran, A. Mishra, R. Ehsani and C. Davis, *Comput. Electron. Agric.*, 2010, **72**, 1–13.

5 F. Dor, G. Disease and H. Al-saddik, , DOI:10.3390/s17122772.

6 J. Albetis, A. Jacquin, M. Goulard, H. Poilvé, J. Rousseau, H. Clenet, G. Dedieu and S. Duthoit, , DOI:10.3390/rs11010023.

7 C. F. Jordan, *Ecology*, 1969, **50**, 663–666.

8 R. Tauler, *Chemom. Intell. Lab. Syst.*, 1995, **30**, 133–146.

9 R. de Juan, Anna; Rutan, S. and Tauler, in *Comprehensive Chemometrics*, ed. B. Brown, S. D. ; Tauler, R. and Walczak, Elsevier, 2010, vol. 2, pp. 325–344.

10 D. . Hand, *Biometrical J.*, 1985, **27**, 148.

11 V. Olmos, M. Marro, P. Loza-Alvarez, D. Raldúa, E. Prats, F. Padrós, B. Piña, R. Tauler and A. de Juan, *J. Biophotonics*, 2018, **11**, e201700089.

12 S. Mas, A. Torro, L. Fernández, N. Bec, C. Gongora, C. Larroque, P. Martineau, A. De Juan and S. Marco, *Talanta*, 2020, 208, 120455

13 S. Hennessy, G. Downey and C. P. O′Donnell, *J. Agric. Food Chem.*, 2009, **57**, 1735–1741.

14 R. Karoui, M. Hammami, H. Rouissi and C. Blecker, *Food Chem.*, 2011, **127**, 743–748.

15 F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. T. Shapiro, P. J. Barloon and A. F. H. Goetz, *Remote Sens. Environ.*, , DOI:10.1016/0034-4257(93)90013-N.

16 A. de Juan, in *Data Handling in Science and Technology*, 2020.

17 N. Lord, G. H. Golub and C. F. Van Loan, *Math. Gaz.*, , DOI:10.2307/3621013.

18 W. Windig, C. E. Heckler, F. A. Agblevor and R. J. Evans, *Chemom. Intell. Lab. Syst.*, , DOI:10.1016/0169-7439(92)80104-C.

19 R. Tauler, A. Smilde and B. Kowalski, *J. Chemom.*, 1995, **9**, 31–58.

20 P. . Mahalanobis, in *On the generalized distance in statistics*, Proceedings of the National Institute of Sciences (Calcutta) (1936), 1936.

21 J. Jaumot, A. de Juan and R. Tauler, *Chemom. Intell. Lab. Syst.*, 2015, **140**, 1–12.

9. Figure captions

Figure 1. MCR application to a) an individual hyperspectral image, b) an image multiset structure.

Figure 2. MCR-ALS results for a) the multiset of training infected dataset (**DaugtrF**) and b) the multiset of training healthy datset (**DaugtrH**). Left plots: related MCR-ALS distribution maps with their corresponding RGB images. Right plots: resolved pure MS spectra. Varieties in italics correspond to *white varieties.*

Figure 3. MCR-ALS results for the multiset of both infected and healthy training (**DaugtrFH**). Left plots: related MCR-ALS distribution maps with their corresponding RGB images. Right plots: resolved pure MS spectra. Varieties in italics correspond to *white varieties.*

Figure 4 Predicted distribution maps of: a) **DtestcoF2**,b) **DtestgH2, c) DtestgF1 and d) DtestgH1** images with their corresponding RGB images.

Table 1. Total number of leaves images selected from the different varieties and the number of images both in the training and in the independent test set for the MCR-ALS/FDA models (see section 3.2 for more information).

Varieties	**Flavesc. Dorée**			**Healthy**		
	Total	**Training**	**Test**	**Total**	**Training**	**Test**
Gamay (g)	19	12	7	10	7	3
Fer (f)	10	7	3	5	3	2
Duras (d)	9	6	3	3	2	1

Chardonnay (c)	*12*	*8*	*4*	*11*	*7*	*4*
Colombard (co)	*10*	*7*	*3*	*5*	*3*	*2*
Loin de l'œil (l)	*10*	*7*	*3*	*5*	*3*	*2*

Varieties in italics correspond to *white varieties*

Table 2. Number of resolved components and variance explained by MCR-ALS analysis of **DaugtrF** and **DaugtrH** multiset structures.

Multiset	Explained variance	Resolved components
DaugtrF	99.91 %	3
DaugtrH	99.97 %	3

Table 3. Discrimination Rate (CR) of infected and healthy images of the test dataset.

Infected Dabcd*	**CR**	**Healthy Dbacd***	**CR**
DtestgF1	95.1	**DtestgH1**	97.7
DtestgF2	94.2	**DtestgH2**	*63.3*
DtestgF3	93.2	**DtestgH3**	*74.7*
DtestgF4	90.8	**DtestfH1**	94.6
DtestgF5	73.5	**DtestfH2**	99.9
DtestgF6	78.6	**DtestdH1**	99.3
DtestgF7	53.8	***DtestcH1***	*99.9*

DtestfF1	*53.1*	***DtestcH2***	*99.9*
DtestfF2	88.4	***DtestcH3***	*99.4*
DtestfF3	83.9	***DtestcH4***	*99.9*
DtestdF1	93.9	***DtestcoH1***	*97.2*
DtestdF2	90.4	***DtestcoH2***	*98.9*
DtestdF3	94.0	***DtestlH1***	*100*
DtestcF1	*53.5*	***DtestlH2***	*99.8*
DtestcF2	*73.8*		
DtestcF3	*86.4*		
DtestcF4	*91.3*		
DtestcoF1	*51.7*		
DtestcoF2	*25.2*		
DtestcoF3	*94.2*		
DtestlF1	*81.7*		
DtestlF2	*55.6*		
DtestlF3	*50.1*		

*Image code Dabcd; a= training (tr) or test ; b=variety; c=flavescence doré (F) or healthy (H) and d=sample number

Varieties in italics correspond to *white variety*. Images in red present low CR results.