**Predicting fungal infection sensitivity of sepals in harvested tomatoes using Imaging Spectroscopy and Partial Least Squares Discriminant Analysis**

Mercedes Bertotto1, Hendrik de Villiers1, Aneesh Chauhan1, Esther Hogeveen-van Echtelt1, Manon Mensink1, Zeljana Grbovic2, Dimitrije Stefanovic2, Marko Panic2, Sanja Brdar2

Wageningen University & Research1 P.O. Box 123, 6700 AB Wageningen

Biosense Institute2, Dr Zorana Diindica 1, Novi Sad 21000, Servie

Contact: Mercedes.bertotto@wur.nl

T + 31 (0)681182807

1. **Introduction**

Tomato (Solanum lycopersicum L.) is a "ubiquitous vegetable". Tomatoes are produced globally, either for domestic consumption or as a commodity for international export. The nutritional composition of this fruit includes carbohydrates, lipids and proteins. In addition, it contains vitamins, minerals, and carotenes in smaller proportions. (1)

Tomato quality is divided into different aspects, commercial, organoleptic and nutritional.2 Market standard primarily depends on the external appeal (e.g. color, form, size), firmness and shelf life, whereas health benefits rely on the nutritional value as well as on the absence of pathogenic hazards or contaminants.(2-5)

The portion of tomatoes that go to waste after the harvesting stage can reach 42% worldwide.(6) Around 30% of the harvested tomato produce may be lost during postharvest handling, primarily because of microbial decay caused by fungi such as *Rhizopus stolonifer, Alternaria alternata, and Botrytis cinerea*.(7) Pathogenic fungi can infect and spread to many different parts of a tomato plant, including the stem, calyx and skin of the fruit.6(8) In some countries tomatoes are sold including calyx. Fresh looking green parts of a tomato (calyx and vine) are a sign for dealing with fresh tomatoes. Older tomatoes show dehydration symptoms of the green parts. The calyx is also susceptible to infection by fungal spores. These spores may already be present on the tomato during cultivation. After harvest, under humid and poorly ventilated storage and transport conditions, these spores may germinate and grow further into visible mould on the calyx.(9) This negatively affects the value of the fruit and may lead to extra food loss and waste.(9, 10)

The timely identification of disease has the potential to avert losses since prompt actions can be implemented to mitigate further damages (e.g. adapt packing strategies).(9) Generally, the strategy employed in the industry to reduce pathogen attacks is the use of pesticides. However, these products can damage the food and diminish its nutritional value.2 Whenever possible, it is preferable to protect the harvested fruits by using methods that do not introduce any additional chemicals or contaminants and do not harm the food in any way.

A possible means to assess the predisposition to microscopic fungal contamination is by tracking the growing and handling conditions of tomato produce within the supply chain. This correlation may be beneficial in the detection of probable origins of fungal contamination based on historical data. However, tracking individual tomatoes or even batches from growth to harvest and later post-harvest handling and logistics is highly difficult.

Some tomatoes are more susceptible to infection and growth of spores while others are not.(9, 11) Moreover, susceptibility of individual sepals also differs. It is not known, yet, what is causing this difference. This knowledge would be useful to predict the susceptibility to this infection and growth. A more specific method is necessary which allows each calyx and sepal to be evaluated individually.

Some of the analytical methods traditionally used to evaluate the presence of fungus in plants are summarized here. Firstly, new DNA-based technology has been developed to support and replace morphology-based detections of phytopathogenic fungi. Daniel Jiménez, in 2009, developed a real-time qPCR assay for the calculation of F. oxysporum DNA in plant tissues and soil.(12) Moreover, tomato samples can be tested for mycotoxins, as a high level of these compounds is caused by fungal infection.1(13)3 Some detection solutions are, for instance, chromatography coupled with detector methods, electrochemical biosensors technology and immunological techniques such as such enzyme-linked immunosorbent assay, dipsticks and flow-through membranes.14-17(14-17) Furthermore, chromatography-mass spectrometry (GC-MS) or electronic nose (e-nose) can be used to measure the shift of the composition and concentration of volatile organic compounds (VOCs) emitted by diseased tomatoes.13(13)

Although these analytical methods are specific and accurate, they have several disadvantages. First, most of them destroy the sample during measurements. Furthermore, they are methods for detecting disease symptoms and not the susceptibility to fungal infection and growth. That is, they evaluate what is happening to the fruit exactly at the moment of the analysis. In the case of visible symptoms of the fungus, the future is already known (this state will continue and worsen in the future); however, if the fruits are not yet infected or the fungi has not germinated, these methods cannot predict what will happen to them in the future.

There is a need for a reliable, non-destructive and specific method to predict susceptibility to fungal infection in a rapid manner. This would provide additional support for quality inspectors and post-harvest management.

Infrared spectroscopy can provide a possible solution to this problem. Paul Skolik et al, in 2019, have studied diseased progression in whole tomatoes using AFT-FTIR and have highlighted that plant-pathogen interaction can be identified through alteration in the spectra fingerprint.(18)

Imaging Spectroscopy (or “hyperspectral imaging”, “HSI”) can be even more useful because spectral information can be captured across the complete product at pixel level. Huting Wang et al. in 2021 accurately classified 97.5% of healthy fruit and 100% of decayed fruit using spectral imaging.(19)

With similar motivation as the current article, in a previous study, Brdar et al. investigated ensemble machine learning methods, for early detection of sepals’ sensitivity to fungal infections on one tomato cultivar (Brioso).(11) On the one hand, their research results need to be extended to multiple cultivars. On the other hand, there is a need to investigate traditional chemometric approaches to, for instance feature selection, before training the models. To the best of our knowledge, no prior research applies HSI together with chemometric analysis to predict susceptibility to fungal infection of recently harvested tomatoes. Therefore, this is the objective of our research.

The methodology used for our aim included spectra extraction from HSI images, and model calibration and validation using Partial Least Squares Discriminant Analysis (PLSDA), focusing on the optimization of the model parameters.(20, 21)

2. Background

2.1. Feature selection

In constructing chemometric models, it is recommended to carefully select variables which refers to wavelengths or wavenumbers in spectroscopy).(22) It is preferable to identify and incorporate variables that significantly contribute to the studied phenomenon rather than include hundreds or thousands of variables blindly. While certain methods, like Partial Least Squares (PLS), can tolerate variable redundancy, understanding the importance of individual variables improves overall model effectiveness as well as model interpretability.(23)

Furthermore, if a model is built for routine use, it is generally optimal to measure only some variables in the unknown, new samples.

When it comes to NIR spectroscopy, there are two main challenges: 1) It is important to limit the outcome space, since this is directly proportional to the number of variables in a dataset (24) and 2) If variables are strongly correlated, the model carries a risk of overfitting. The use of a simple model, which has fewer but complementary variables, has a better chance of being generalized to a new sample than a model that has hundreds of highly correlated variables.

It is also important to note that some variables are important in combination with others. A variable could have average individual interest but, together with other variables, can gather important information related to a given problem.

The feature selection methods can be classified as Filters , Embedded and Wrapper, according to the interconnection with the problem to be solved.

(22)

Filter methods are used optionally, before wrapper and embedded methods. The user provides a threshold of feature importance, above which variables are retained in the model. These methods are univariate, and do not directly depend on the performance of the final models. (22)

When Embedded methods are used, variable selection is carried out at the same time as learning by cross-validation performance. (25)

On the other hand, wrapper methods iteratively choose a subset of features, train the models, and choose then the best combination.They use the model's error rate on a holding set to score feature subsets (22)

In previous works, embedded methods (Random Forest) were used to solve a similar research problem. (11)

This is why, in the present study, a wrapper method (the CovSel algorithm) was chosen to compare results and to evaluate the performance of a different approach.(24)

When using the CovSel algorithm, the significance of each wavelength is assessed by calculating the covariance between each variable and the Y values (responses). (26)

Firstly, X and Y matrices are centered (the mean value is deleted from each individual value in each dataset, as a result, the mean value in each dataset becomes zero). Then, the covariance of all the variables in X is calculated, with respect to the response Y, and the variable with the highest covariance is chosen. Then, all remaining variables are orthogonalized with respect to the chosen variable, and the procedure is repeated until the number of variables chosen by the user is completed. (24)

2.2. Outlier removal

There are two typical ways to proceed when analyzing spectra extracted from spectral images. One is to merge all the spectra (pixels) corresponding to a region of interest (ROI), in our case a sepal . The other is to directly make the predictions at the level of each pixel. The first option allows to drastically reduce the dimensionality of the data by reducing the number of spectra (rows), which consequently reduces the analysis time. In this work, this first option was chosen, that is to say, all the pixels coming from the ROI were merged leading to a single spectrum per ROI.

It is important to mention that not all pixels found in a ROI provide relevant information. Some pixels can be “noisy” or “bad/abnormal pixels”. According to Chu Zhang et al, in 2020, random errors are primarily generated by the surrounding conditions and the devices themselves.(27)

In 2009, James Burger found that bad pixels exhibit significantly different spectra compared to their neighbors.(28) He classified these abnormal pixels into "dead pixels," which do not respond to light, "hot pixels," characterized by high dark current, and "stuck pixels," which maintain an almost constant intermediate value.

In 2023, Bingkai Liu further explained that hot pixels have a higher dark current than normal pixels, which experience a moderate dark current increase after irradiation. (29)

Moreover, some pixels are always noisy, while others only sporadically; some may show a “non-linear response to light intensity” while some others behave randomly.(30)

In any case, these abnormal pixels behave differently from the rest and should be eliminated. In this work, they were detected in an unsupervised manner ( see Section XXXX for details). 2.3. Classification methods

There are two types of classification methods, those called "class modeling" and those called "simple discrimination".

The first type of methods create a model for each studied class. The focus of the methods remains on one individual class at a time. The samples in each class fall into a closed class space. Moreover, the shape of the space is characteristic of the class modeling method used whether the width of the class space is determined by the confidence level applied.

An object can be assigned to one class, to more than one class, or not assigned (outlier). Examples of these are Soft Independent Modeling of Class Analogies (SIMCA); Potential function methods (POTFUN) and Unequal variance methods (UNEQ).(31-34)

On the other hand, simple discrimination methods create a boundary between classes, so that they always assign each object to the class to which it most likely belongs. The limits can be linear or not. These cannot detect samples that do not correspond to any of the classes (outliers).

The aim of these discrimination methods is to find optimal boundaries between all given classes by maximizing the difference between them. A new object will always be assigned to one class, namely, to the class to which it is the most similar. This is why, these types of methods cannot be used for outlier detection, as they cannot detect samples which do not belong to any of the classes. The simplest method is the one based on the FC, the Fisher linear discriminant method. Other examples include: Discriminant partial least squares (D-PLS or PLS-DA), K-Nearest Neighbors Analysis (k-NN), Support Vector Machines (SVM), Linear Discriminant Analysis (LDA) Quadratic Discriminant Analysis (QDA).(35-38)

In this work, the SIMCA and PLSDA methods were used and compared.

1. **Materials and Methods**

2.1 Materials

Three tomato cultivars, 'Brioso,' 'Cappricia,' and 'Provine,' were used in this study. Fresh samples were harvested from different greenhouses on the 9th and 10th of May. On the 10th May, the tomatoes on the vine arrived at the Phenomea Laboratory in Wageningen, Netherlands. Tomatoes without visible fungal infection were cut from the vine (2 tomatoes from the middle of a vine, 32 samples from each cultivar). The wounds at the cut end were greased with stopcock grease to prevent dehydration at the junction.

2.2. Methods

2.2.1. Data collection

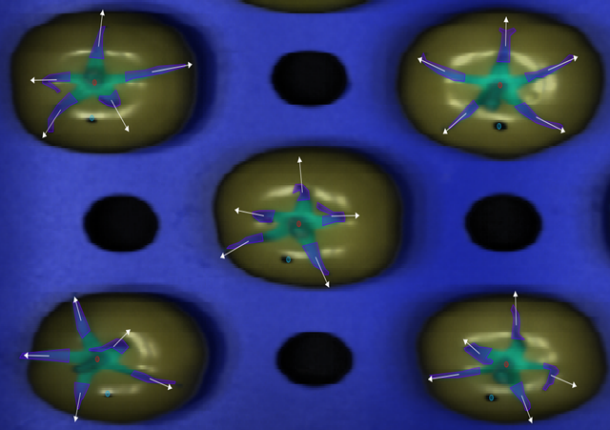
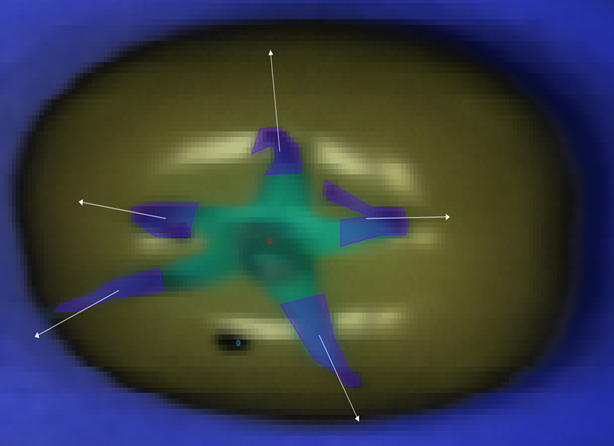
Samples were imaged in two separate groups of equal size. Hyperspectral images were recorded on day one (10th May) using a Specim FX17 NIR linescan camera with a spectral range (937.33 nm-1718 nm).(11) Subsequently, tomatoes were stored on trays (7mm blue Forex plate (35x55cm2) with holes of 2.5 cm diameter) in controlled conditions encouraging fungal growth (20°C, in a closed sanitized box reaching 100% Relative Humidity, in a room at 60% RH, lights on during 7:00-19:00h, 15 μmol·s-1·m-2).

Ground truth observations were made per sepal by three experts on day three and four (12th and 13th May), which comprised of severity scores from zero (no fungus) to four (severe infection). Ratings of the two days and three experts were averaged.

2.2.2. Spectra extraction from hyperspectral images

Hyperspectral images were converted to pseudo-color images, which were generated after manually choosing three bands which produced visibly good contrast between sepals and the background. These images were manually annotated with a separate polygon indicating the boundary of individual sepals. These polygons were converted to pixel masks, which indicated whether a pixel was included in the set of pixels belonging to a particular sepal. At sepal edges, because of blurring effects, there is some level of uncertainty in on which pixels to include. For this annotation, we favoured keeping pixels only if they were substantially sepal containing. The spectrum of each pixel was collected and then used for further analysis.

The Darwin annotation tool from V7 labs was used to perform annotations.(39) Annotations were used to extract sepal pixel spectra using a custom Python image processing pipeline.(40)

**Figure 1**: Spectra extraction from hyperspectral images. Visualization of the procedure carried out in each sepal.

2.2.3. Data analysis

A chemometric analysis was conducted with the aim of calibrating and validating models to predict the susceptibility to fungal infection in tomatoes according to their degree of disease as observed by specialists after 4 days of germination. This analysis was done using R Statistical Software (v4.3.0; R Core Team 2021) with caret, rchemo and prospectr packages, and involved the following steps:(41-44)

1. Data visualization

Firstly, spectra were plotted to have a first appreciation of the shape of the data, observe their clarity, signal-to-noise ratio, presence of obvious outliers, baseline, etc.

1. Data exploration and outlier removal

Exploratory analysis was carried out at sepal and variety level using Principal Component Analysis (PCA).(45)

PCA was applied over all the pixels for a given sepal. To detect outlier pixels, Mahalanobis distances were computed between the individual projection of each score value onto the model and the center of the model. The identification of outliers was determined based on a specified confidence level (0.95), indicating the probability that a data point lies within a certain range. The cutoff for Mahalanobis distances was employed as a threshold, beyond which data samples were classified as outliers. The confidence level played a crucial role in controlling the sensitivity of the outlier detection, with higher confidence levels leading to more stringent criteria for identifying outliers.

The outliers were removed, the remaining pixels averaged, and the datasets were finally reassembled according to their labels (see Table 1).

**Table 1:** Description of the dimensionality of the initial and final datasets before and after averaging spectra that belonged to the same sepal.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset name/ Number of | Pixels per sepal | Sepals per tomato | Tomatoes per image | HSI | Spectra in the  initial dataset | Spectra in the averaged dataset | Variables |
| Provine | Between 119 and 90 | 5 or 6 | 16 | 2 | 16156 | 159 | 112 |
| Brioso | Between 45 and 53 | 5 or 6 | 32 | 1 | 6497 | 164 | 112 |
| Cappricia | Between 81 and 124 | 5 or 6 | 16 | 2 | 12816 | 165 | 112 |

Data exploration was carried out again by PCA in order to remove outliers at variety level (in Provine, Brioso and Cappricia datasets). Score plots were created, outliers were detected visually and removed from dataset.

c) Pretreatments on raw spectra

Various pretreated forms of the original spectra were used to calibrate and validate different models, whose performances were compared with each other. These methods include: Detrend grades 1 and 2; Savitzky–Golay first and second derivatives, second polynomial degree and 9, 11, 15, 17 smoothing windows; Standard Normal Variate (SNV); and combinations of these.(46-48) Only the best results will be presented in this document.

d) Data splitThree binary-class scenarios were derived from the visual expert scoring described in Data collection (2.2.1):

* Scenario 1: Score of 0 was considered healthy, and any other value was considered infected;
* Scenario 2: A score of 1 or less was considered healthy and the rest infected;
* Scenario 3: scores from two consecutive days were averaged, and a healthy class was considered when score was 0.5 or lower, otherwise the sepal was considered infected

Stratified sampling was carried out in the following way. The datasets were divided into calibration (70%) and validation (30%) sets, in a representative way for each class, randomly. The number of spectra resulting from this separation, according to each labeling scenario, can be seen in Table 2.

**Table 2:** Number of spectra in each class (Healthy: Class 1; Diseased: Class 2) when dataset was split according to different labelling scenarios (Label 1: 0/123; Label 2: 01/23 and Label 3: 0.5/123).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cultivar | n | Label 1 | | Label 2 | | Label 3 | |
| Healthy | Diseased | Healthy | Diseased | Healthy | Diseased |
| Cappricia | 163 | 139 | 24 | 85 | 78 | 117 | 46 |
| Brioso | 153 | 145 | 8 | 78 | 75 | 126 | 27 |
| Provine | 152 | 137 | 15 | 72 | 80 | 129 | 23 |

e) Feature selection

An iterative process was used to select a sparse subset of important variables in the Training set, using CovSel algorithm.(26, 49) Iteratively top 5 to top 39 Important Variables (ivs), (numbers chosen arbitrarily), were chosen for each pretreatment, labelling and cultivar. The selected variables were then used as input for the classification model.

Dataset with measured reflectance of only the chosen variables was saved in “CovSelTrain". The same ivs were selected from the Test Set, and saved in “CovSelTest".

f) Calibration and validation of PLSDA modelsThe selected variables as given by CovSel were then used as input of PLSDA.

The Training set was split again, into Validation (70%) and Tuning (30%), randomly. Different models with different number of Latent Variables (LVs) were calibrated in Validation set and tested in the Tuning set. The number of LVs was selected according to the model that showed the lowest prediction error in the Tuning set.

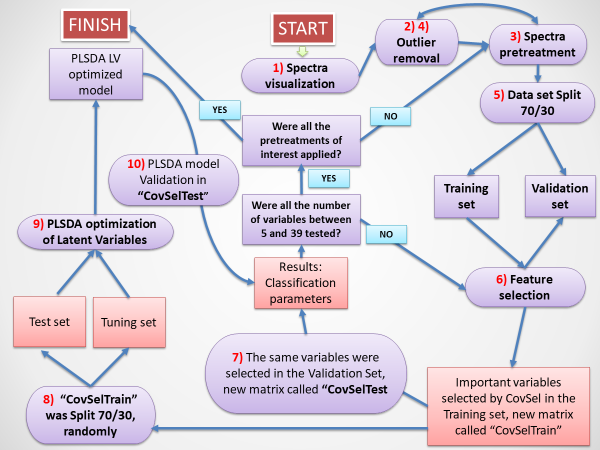
The PLSDA model, already optimized for the number of latent variables, was tested in “CovSelTest”, and classification parameters were obtained.

g) Evaluation of results

In this study, 10 parameters were used to evaluate the results: Sensitivity, Specificity, Precision, Accuracy and Balanced Accuracy, Geometric mean, F-measure, Youden index, Positive likelihood ratio, and Negative likelihood ratio. These are explained in detail in previous publications and shown in Table 3.(50, 51) The evaluation should consider all of them simultaneously, because each one of them takes into account different characteristics of the general discrimination effectiveness.

**Table 3:** Parameters commonly used to evaluate classification models. TN: True Positives; TN: True Negatives; FN: False Negatives; FP: False Positives. Source: Akosa, 2017.

|  |  |
| --- | --- |
| Measure | Formula |
| Accuracy |  |
| Misclassification rate (1-Accuracy) |  |
| Sensitivity (or Recall) |  |
| Specificity |  |
| Precision |  |
| Balanced Accuracy (BA) |  |
| Geometric mean |  |
| Positive likelihood ratio |  |
| Negative likelihood ratio |  |
| F-measure |  |
| Youden index |  |



**Figure 2**: Diagram of the basic chemometric steps carried out in this work in order to optimize the parameters of the models.

The iterative process carried out in this work is shown in Figure 2, and can be summarized as follows:

A. Spectra visualization and outlier removal.

B. Model selection.

0. Start with a cultivar from a set of cultivars. Start with no "best model” for the cultivar.

1. Select labelling scenario (from 3 scenarios).

2. Select one pretreatment or combination of pretreatments;

3. Split Dataset.

4. Select important features.

5. Apply PLSDA and select the optimal number of latent variables (LVs).

6. Repeat Steps 4 and 5 selecting from 5 to 39 variables by CovSel.

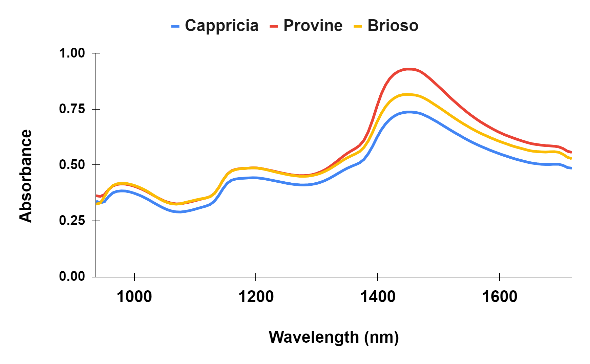
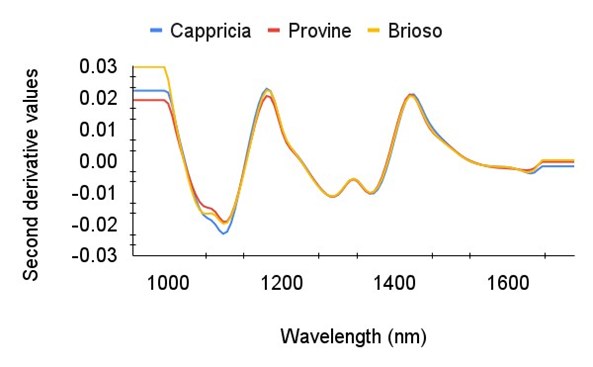
7. If a model BA is higher than the previous model, keep the current model as the “best model".

Note: Results of steps 1 to 7 will give the best model per cultivar.

C. The same process was repeated for global modeling where different scenarios of variety combinations were investigated: Cappricia+Provine, Cappricia + Brioso, Brioso + Provine and Cappricia + Provine + Brioso.

1. **Results and Discussion**

3.1 Intravariety models

**Figure 3**. left: Average of the raw spectra for each variety, right: Average of the spectra pretreated with SNV and second derivative (2, 17, 2).

First, the raw spectra were plot immediately after being extracted from the images. This first visualization allowed us to have a first appreciation of how the spectra looked in relation to noise and scattering effects, distortions in the baselines, signal-to-noise ratio, in addition to the presence of clear outliers. In order to understand the presence of multiplicative and/or additive effects in the spectra, their intensities were plot as a function of the average spectra (graphs not shown here). The shape of these graphs (millefeuille or cone) helped distinguish effects in the spectra. In all cases, combined effects (multiplicative and additive) were found in the analyzed spectra. Figure 3 (on the left) shows the average of the raw spectra for each variety, and Figure 3 (right) shows the average of the spectra pretreated with SNV and second derivative (2, 17, 2). As mentioned above, other pretreatments were applied and compared as well.

It can be observed that raw spectra of Provine and Brioso overlapped up to approximately 1400 nm, while at longer wavelengths the average spectra were clearly differentiated. Cappricia spectra showed greater intensity throughout the entire spectral range.

Three bands are observed in the pure spectra of all varieties. In the following paragraphs, tentative assignments will be mentioned along with their bibliographic sources.

The maximum intensities observed were 0.7373 (Cappicia), 0.9296 (Provine) and 0.8163 (Brioso); at 1455nm, 1447.9nm and 1447.9nm respectively, at 1455nm (6872.85cm-1) in Cappricia and at 1447.9nm (6906.08cm-1) in the other two varieties. These bands can be attributed to the symmetric and asymmetric stretching vibrations of water molecules at the first harmonic of the OH stretching vibrations of water.(52) More specifically, those wavelengths are included into two well-defined wavelength ranges where water shows the greatest variation of energy absorbance in response to disturbances, (Water Matrix Coordinates, “WAMACS”), called C8 and C9. *“WAMACS describe different conformations of water such as water dimers, trimers, superoxides, water solvation shells, etc”* .(52, 53)

The mentioned peaks in spectra can be found in WAMACS C8 and C9.

*C9 1458-1468nm: Water molecules with 2 hydrogen bonds (S2)*

*C8 1448-1454nm: ν2 + ν3, Water solvation shell, OH-(H2O)4,5.(53, 54)*

The other peak in raw spectra was located at 1194.6 nm (8368.20cm-1) in all three varieties. According to M. Jakubíková et al. in 2016*: “the region from 8300 to 8600 cm-1 corresponds to the third overtone band of the bond CH”.(55)*

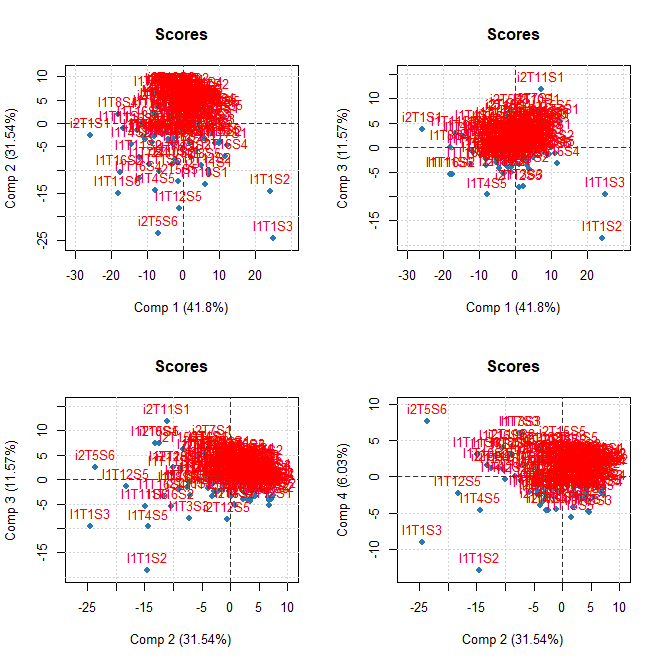
In 2003, Dalimov et al. explained that tomato has approximately 11% lignin, which has carboxylic groups that distinguish it from other plants.40 Moreover, these authors analyzed IR spectra of suspended tomato particles, and found typical absorption bands for lignin and carbohydrates. They assigned the 1194.6 nm wavelength to the second C–H str overtone of methyl groups, CH3-groups, as well as the lignin component of tomatoes.

However, other publications assign this band to glucose. Tanaka et al, in 2021, measured several glucose anomers in light and heavy water by NIR, and found a peak at 1195 nm in both solvents.(56) Furthermore, Lopez et al in 2016, performed carbohydrate analysis by NIR, and assigned the same peak to the OH stretch 1st overtone of glucose.(57)

Finally, the three raw spectra have a peak at 978.85nm (10241.5 cm-1). It has been assigned in literature to the O–H stretching second overtones, to the hydroxide ion (980 nm) and to the hydrogen-bonded –OH, 2nd overtone (980.4 nm).(58, 59)

Figure 3 (on the right) shows average spectra of the measured reflectance for each cultivar, after they were corrected using both the SNV and second derivative (grade=2, window width=17 and polynomial degree=2) algorithms. It should be mentioned that in this study, the most appropriate pretreatments were chosen according to the way in which they modify the performance of the models. Pretreatments depend on data, sometimes they can be beneficial but other times not.

In this example, SNV was used to remove both the scattering effects caused by the diffusion of photons and the measurement noise (random phenomena present throughout the entire measurement chain). The resulting spectra had media equal to zero and standard deviation equal to one. Furthermore, second derivative was used to find the exact location (center) of the shoulders in the original spectra, by deconvoluting and highlighting the peaks. As a result, we observed significantly narrower bands. Peaks appeared in the same locations as peaks in the original spectra.

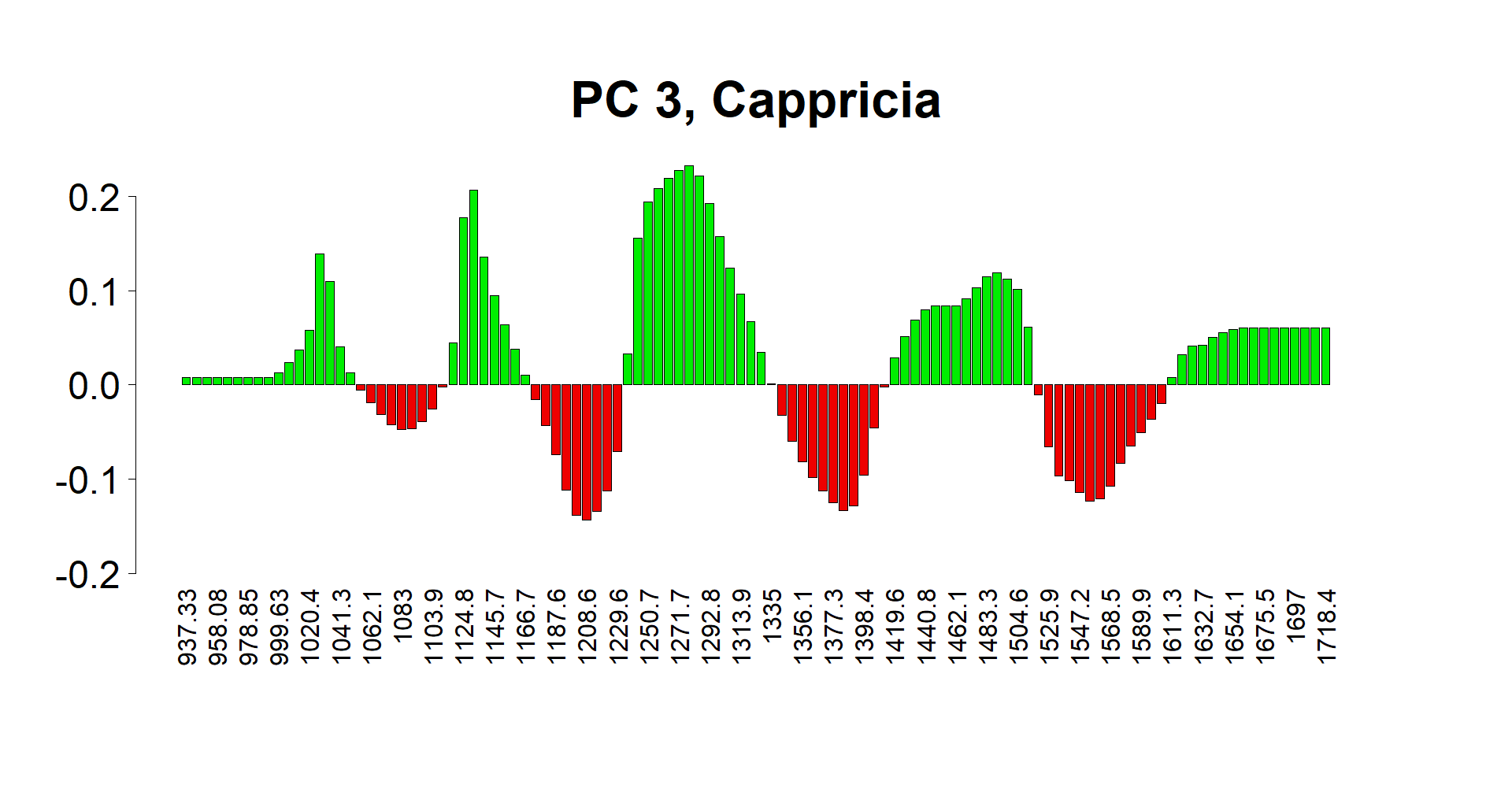


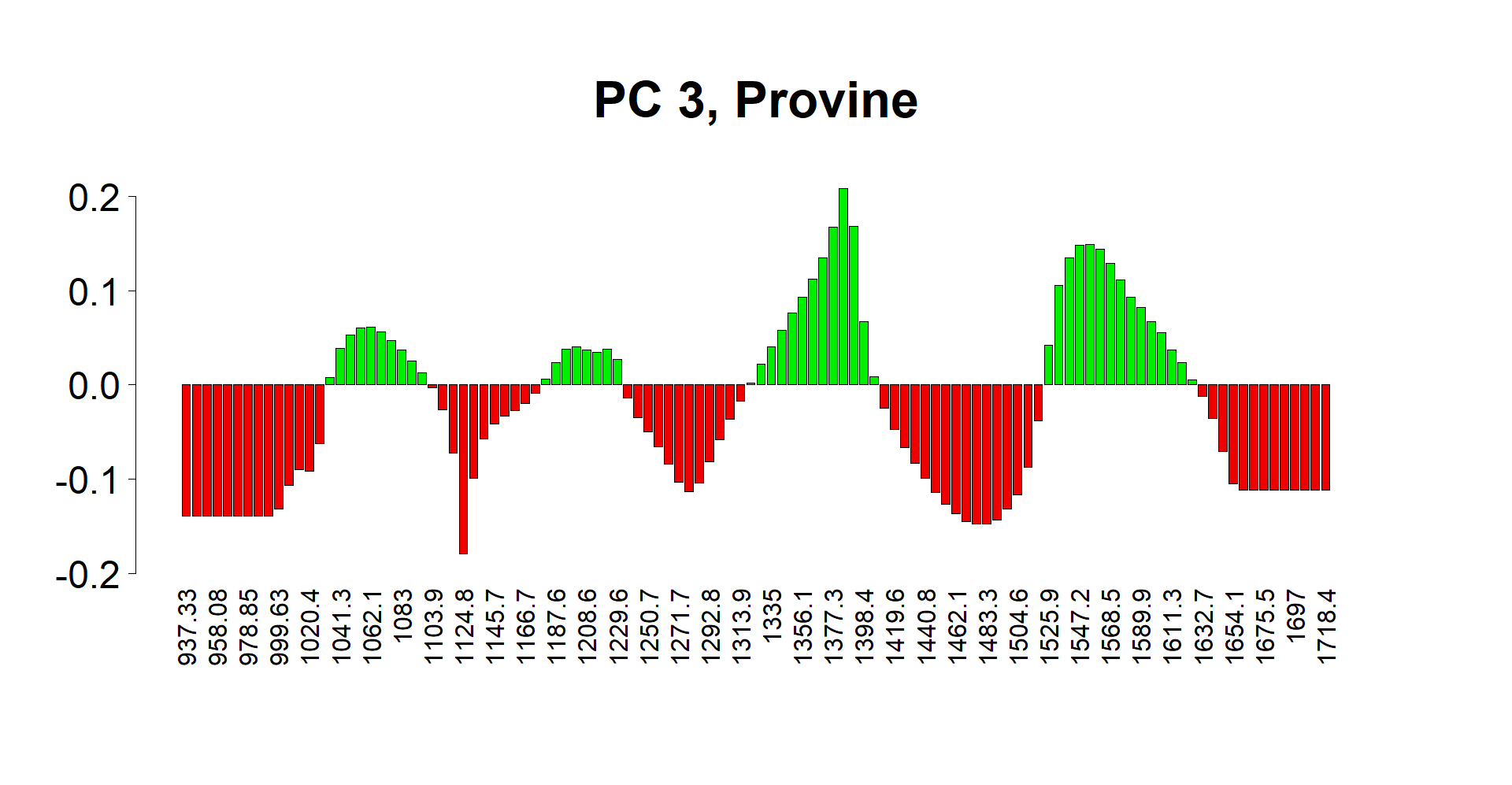
**Figure 4**: PCA score plots of Cappricia cultivar.

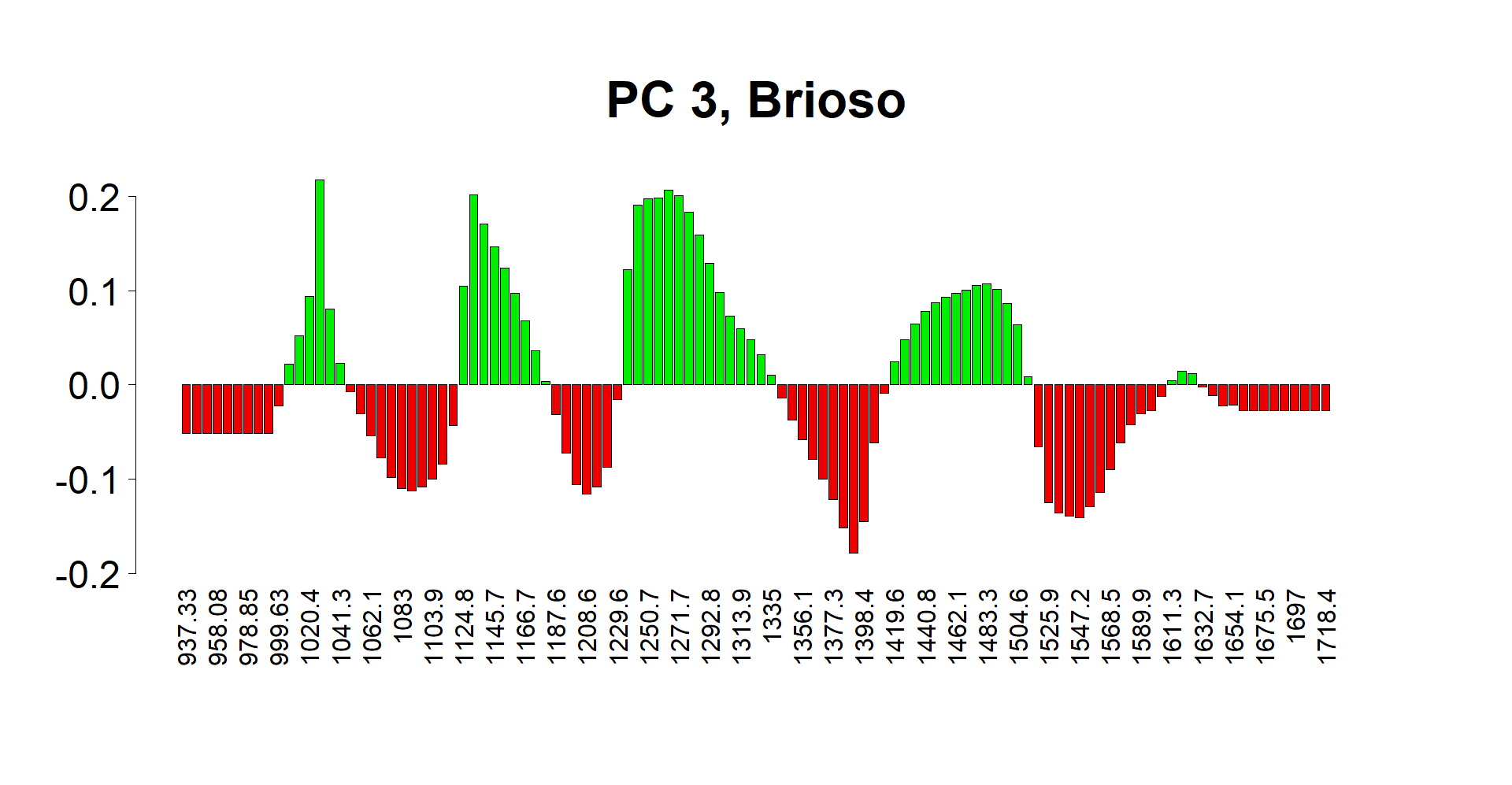
The PCA analysis was performed in this case on the pretreated spectra, first with SNV and then with the second derivative (2, 17, 2), in all three varieties. Scores plots for Cappricia are shown in Figure 4. The number of principal components to accumulate the variance explained by the model, together with the outliers detected can be seen in Table 4.

**Table 4.** Results of exploration by PCA, to detect outliers at cultivar level.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cultivar** | **Number of Principal Components** | **% Variance explained by the model** | **Number of extreme outliers detected visually** |
| Brioso | 8 | 99.15 | 6 |
| Cappricia | 7 | 99.02 | 6 |
| Provine | 8 | 99.13 | 4 |







**Figure 5**: PC 3 Loading plots of cultivars Cappricia, Provine and Brioso.

The Loadings Plots showed that there is no notable difference between varieties in PC1 and PC2 in all three varieties (graphs not shown). However, PC3 was markedly different in each variety (Figure 5), indicating that this main component is a significant carrier of the inter-varietal distinctiveness of this product.

**One-class SIMCA**

**Table 5.** Results of One-class SIMCA on "Healthy” Class. Exp.Var: % Variance explained by the model.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cultivar** | **Exp.Var** | **TP**  **Cal** | **FP**  **Cal** | **TN**  **Cal** | **FN**  **Cal** | **Spec. Cal** | **Sens Cal.** | **Acc. Cal.** | **PC** | **Spec. Val** | **Acc. Val** |
| **Brioso** | **99.14** | **74** | **0** | **0** | **2** | **NA** | **0.974** | **0.974** | **6** | **0.157** | **0.157** |
| **Provine** | **99.83** | **72** | **0** | **0** | **4** | **NA** | **0.947** | **0.947** | **6** | **0.301** | **0.301** |
| **Cappricia** | **99.1** | **71** | **0** | **0** | **6** | **NA** | **0.922** | **0.922** | **6** | **0.270** | **0.270** |

A SIMCA one-class analysis was carried out, with the aim of developing a classification model in the "Healthy" class. All samples that do not correspond to this category should be rejected by the model.

The One-Class SIMCA model gave good results for the healthy class in the three varieties (Table). However, when each trained model was validated in its corresponding diseased class, it was not able to reject samples with high specificity (Table 5).

This is why, in this study, the “class modeling approach” was not adequate to solve the problem and the PLSDA algorithm was used instead.

**PLS-DA**

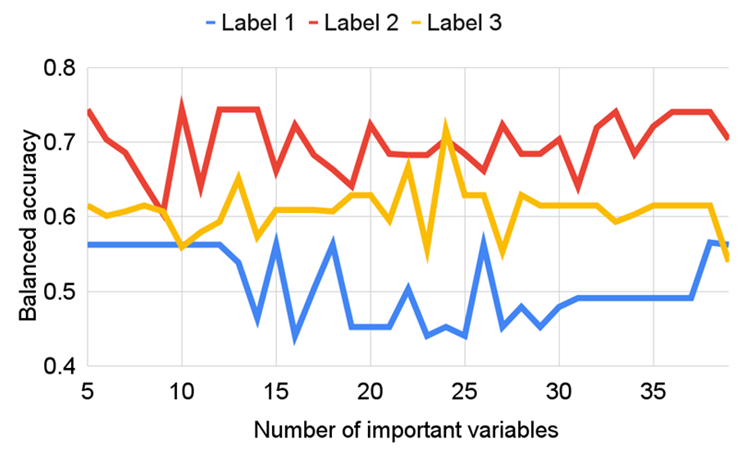
Different results obtained using different labeling scenarios between the healthy and the diseased classes for Cappricia cultivar were compared (plots not shown). As a result, the precision metric behaved erratically when Scenario 1 was selected. This metric showed high values when less than 13 variables were chosen, but then decreased abruptly with 14 variables; and increased again when 15 variables were chosen. This counterintuitive behavior was due to the fact that the precision metric took into account false positives in the denominator, which changed abruptly with different splits. In other words, the behavior of the precision metric showed that the data were not uniformly distributed in both classes, when Scenarios 1 and 3 were chosen.Another indicator of class balancing was the correlation between the accuracy (blue line) and the balanced accuracy (BA) (orange line). When the classes were balanced, these metrics were almost identical, and their lines overlapped as we could see in Scenario 2 (Figure B). On the other hand, in Figures 1. A and 1. C, a clear separation was observed between them, the accuracy was higher than BA.

As a rule, high BA values showed that the model performances were good for both classes. On the other hand, high accuracy metric showed that the model performed well in general. When using Scenarios 1 and 3, this was the case for the majority classes only.

It should be mentioned that there are several ways to solve data imbalance. One of them is with oversampling (adding samples from the least represented class); or with undersampling (deleting samples from the majority class). In the first case, the risk of overfitting increases, since during cross-validation, the same samples that are in the model can be used to validate it; in the second case, important information is removed from the model.

A one-class classification could also have been used, where all samples similar to the samples of one class are considered, and the others discarded by the model. However, these models are always less specific, and in the case of the present study, they showed poorer classification parameters.

Due to these reasons, Scenario 2 was chosen to calibrate and validate the models. No addition or removal of samples was made, except for the aforementioned outliers.



**Figure 6:** Comparison of Balanced Accuracy in different labelling scenarios according to different number of important variables as input for PLSDA.

The relationship between the BA of the models versus the number of important variables (ivs) chosen by CovSel can be seen in Figure 6, for the different Scenarios 1, 2 and 3. Once again, we can see that Scenario 2 was the best option, because it showed higher BA values.

The following parameters generated the optimal models:

Cappricia: SNV + Second Derivative (2, 15, 2) and 33 ivs

Provine: SNV and 13 ivs

Brioso: Raw spectra and 18 ivs

Cappricia: Optimal model was found with pretreated data (SNV+ SG (2,15,2)) and 33 most important variables. Provine: Optimal model was found with SNV pretreated data and 13 most ivs Brioso: Optimal model was found with raw data and 18 ivs.

3.2. Global models

**Table 6:** PLSDA Classification results of a global model, calibrated and validated with Cappricia and Provine.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data set** | **Real/**  **predicted** | **Healthy** | **Diseased** | **NA** | **Sensitivity** | **Specificity** | **Precision** | **BA** |
| **Calibration** | **Healthy** | **62** | **47** | **0** | **0.57** | **0.74** | **0.68** | **0.66** |
| **Diseased** | **29** | **81** | **0** | **0.74** | **0.57** | **0.63** | **0.66** |
| **Validation** | **Healthy** | **23** | **25** | **0** | **0.48** | **0.71** | **0.62** | **0.60** |
| **Diseased** | **14** | **34** | **0** | **0.71** | **0.48** | **0.58** | **0.60** |

Table 6 shows PLSDA classification results of a global model, calibrated and validated with cultivars Cappricia and Provine. In this model, spectra were pretreated with SNV, and then with second derivative (2, 17, 2). Then, 19 important variables were chosen by the CovSel algorithm. Finally, the PLSDA model was trained with the calibration data, and 17 latent variables were chosen. This relatively high number can be understood as being due to the complexity of adding two different varieties in one model.

**Table 7:** PLSDA Classification results of all the optimal models in this study.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter/**  **Model** | **Cappricia**  **Raw, 15v**  **Label 3** | **Cappricia**  **SNV + SG(2,15,2), 33v**  **Label 2** | **Provine**  **Raw, 14v**  **Label 3** | **Provine**  **SNV, 13v**  **Label 2** | **Brioso**  **Raw, 18v**  **Label 2** | **Global model**  **SNV, (Cap+Pro)**  **6v**  **Label 2** |
| Accuracy | 0.83 | 0.84 | 0.71 | 0.71 | 0.66 | 0.70 |
| Sensitivity or recall | 0.89 | 0.71 | 0.08 | 0.76 | 0.43 | 0.81 |
| Specificity | 0.64 | 0.89 | 0.97 | 0.65 | 0.88 | 0.58 |
| Precision | 0.89 | 0.71 | 0.50 | 0.70 | 0.77 | 0.66 |
| Balanced accuracy | 0.77 | 0.80 | 0.52 | 0.71 | 0.65 | 0.70 |
| Geometric mean | 0.75 | 0.79 | 0.27 | 0.70 | 0.62 | 0.69 |
| F-measure | 0.89 | 0.71 | 0.14 | 0.73 | 0.55 | 0.73 |
| Youden’s Index | 0.53 | 0.60 | 0.05 | 0.41 | 0.31 | 0.39 |
| Positive likelihood ratio | 2.47 | 6.45 | 2.67 | 2.17 | 3.58 | 1.93 |
| Negative  likelihood ratio | 0.17 | 0.32 | 0.95 | 0.37 | 0.65 | 0.33 |

Table 7 and Figure 8, show PLSDA modeling results of all the optimal models in this study. Balanced accuracy was comparable to traditional accuracy in all models created with Scenario 2, showing that the classifier performed equally well on either class.

Accurate predictions for healthy sepals were as follows: Cappricia (0.71), Provine (0.76), Global model (0.81). Similarly, for diseased sepals correctly classified as such: Cappricia (0.89), Provine (0.65), Global model (0.58).

Moreover, good performances on both positive and negative classes were found in the Cappricia Intravariety model. High positive likelihood ratio of 6.45 (above 1: increased evidence for disease-free) for the Healthy class; Low negative likelihood ratio of 0.32 (increased evidence for disease) for the Infected class.

For two-class classification, the geometric mean was calculated as the square root of the product of specificity and sensitivity. As a rule, If one of the classes cannot be recognized by the model, this parameter tends to zero. (60) This parameter showed this behavior, when its values were less than 0.5. This is was observed in the case of sample classification of the Provine variety using Scenario 3: although the specificity of this model was high, the sensitivity was very low (0.08), and the GM was 0.28. In all other cases, this parameter was greater than 0.5 showing that the models were able to recognize both classes.

**4) Conclusion**

This work was carried out with the objective of developing a method to detect the susceptibility of freshly harvested tomatoes to the presence of fungi, in non-destructive ways, and before the disease can be observed visually. For this, hyperspectral images of the samples were measured, and models were developed based on their relationships with ground truth data.

The models can be divided into three general categories: those calibrated and validated using a single variety (intravariety), and finally those calibrated and validated with several varieties together (global models). In all cases, the best results were found using scenario 2 as a reference.

Within the first category, the optimal model was created with the Cappricia variety: Balanced accuracy= 0.80; Sensitivity= 0.71 and Specificity= 0.89.

As for the global models, the optimal models were calibrated using Cappricia and Provine together: Balanced accuracy=0.70, Sensitivity=0.81; Specificity=0.58.

The results from this research suggest the conclusion that discrimination between more susceptible and less susceptible samples is feasible under controlled conditions.

**Acknowledgements**

The authors would like to thank the EU commission for funding obtained through the ANTARES Horizon 2020 project (grant 664387).

**Conflicts of interest**

The authors declare no conflicts of interest.

**References**

1. OECD. Tomato (Solanum lycopersicum). In: Publishing O, editor. Safety Assessment of Transgenic Organisms in the Environment. 7. Paris: OECD Publishing; 2017.

2. Nadia Bertin MG. Tomato quality as influenced by preharvest factors. Scientia Horticulturae. 2018;233(15).

3. STANDARD FOR TOMATOES (CODEX STAN 293-2008). Codex Alimentarius2008.

4. Unit AS, Division ECaT, Europe UNECf. UNECE STANDARD FFV-36 concerning the marketing and commercial quality control of TOMATOES 2017.

5. Regulation (EU) No 543/2011.

6. Isaac Kojo Arah HA, Ernest Kodzo Kumah, and Hayford Ofori. Preharvest and Postharvest Factors Affecting the Quality and Shelf Life of Harvested Tomatoes: A Mini Review. International Journal of Agronomy. 2015;2015.

7. Yeimmy Peralta-Ruiz CDGT, Angie Sinning-Mangonez, Edgar A. Coronell, Marcos F. Marino and Clemencia Chaves-Lopez. Reduction of Postharvest Quality Loss and Microbiological Decay of Tomato “Chonto” (Solanum lycopersicum L.) Using Chitosan-E Essential Oil-Based Edible Coatings under Low-Temperature Storage. Polymers (Basel). 2020;1822.

8. E.J. Smid LH, H.A.M. Boerrigter, L.G.M. Gorris. Surface disinfection of tomatoes using the natural plant compound trans-cinnamaldehyde. Postharvest Biology and Technology. 1996;9(3):343-50.

9. Manon Mensink AC, Najim El Harchioui, Esther Hogeveen. Kwaliteit van tomatenkronen na oogst. In: Research WFB, editor. projectnummer 62341268002019.

10. Jan Janse HB. Kroonschimmel bij tomaat : consultancyonderzoek. In: Wageningen UR Glastuinbouw WUAaFSG, editor. 2007.

11. Sanja Brdar MP, Esther Hogeveen‑van Echtelt, Manon Mensink, ŽeljanaGrbović, Ernst Woltering, Aneesh Chauhan. Predicting sensitivity of recently harvested tomatoes and tomato sepals to future fungal infections. Nature. 2022; 11:23109

12. Daniel Jiménez-Fernández MM-B, Juan A. Navas-Cortés, Rafael M. Jiménez-Díaz BBL. Identification and quantification of Fusarium oxysporum in planta and soil by

means of an improved specific and quantitative PCR assay. Applied Soil Ecology. 2009;46(3):372-82.

13. Mina Nan, 2 Huali Xue,1,\* and Yang Bi3,\*. Contamination, Detection and Control of Mycotoxins in Fruits and Vegetables. Toxins. 2022;14(5):309.

14. Margot Delavy ARDS, Clara M. Heiman and Alix T. Coste. Investigating Antifungal Susceptibility in Candida Species With MALDI-TOF MS-Based Assays. Frontiers in Cellular and Infection Microbiology. 2019;9(19).

15. Kazbek Dyussembayev PS, Ido Bar, Jeremy C. Brownlie, Muhammad J. A. Shiddiky, Rebecca Ford. Biosensor Technologies for Early Detection and Quantification of Plant Pathogens. Frontiers in Chemistry. 2021;9.

16. N. Rajeshwari MDS, M. Krishnappa, H.S. Shetty, C.N. Mortensen & S.B. Mathur Development of ELISA for the detection of Ralstonia solanacearum in tomato: its application in seed health testing. World Journal of Microbiology and Biotechnology 1998;14:697-704.

17. Shuodan Hu CY, Hong Yu, Yu Zhang, and Chuan-Qing Zhang. Establishment of the Recombinase Polymerase Amplification–Lateral Flow Dipstick Detection Technique for Fusarium oxysporum. Plant Disease. 2023;107(9):2665-72.

18. Paul Skolik MRM, Francis L. Martin. ATR-FTIR spectroscopy non-destructively detects damage-induced sour rot infection in whole tomato fruit. Planta. 2019;249(3):925-39.

19. Huting Wang RH, Mengyun Zhang, Zhiqiang Zhai, Ruoyu Zhang. Identification of tomatoes with early decay using visible and near infrared hyperspectral imaging and image-spectrum merging technique. Journal of Food Process Engineering 2021;44(4).

20. Ståhle L WS. Partial least squares analysis with cross-validation for the two-class problem: A monte carlo study. Journal of chemometrics. 1987;1(3):185-96.

21. Barker M RW. Partial least squares for discrimination. Journal of Chemometrics. 2003;17(3):166-73.

22. Michael D. Sorochan Armstrong PdlM, J. Harynuk\*. Review of Variable Selection Methods for Discriminant-Type Problems in Chemometrics. Frontiers. 2022.

23. Chapter 2. Tomato.

24. J.M. Roger BP, D. Bertrand, E. Fernandez-Ahumada. CovSel: Variable selection for highly multivariate and multi-response calibration. Application to IR spectroscopy. Chemometrics and Intelligent Laboratory Systems. 2011;106(2).

25. Nicholas Pudjihartono F, W. Kempa-Liehr, M. O'Sullivan. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers. 2022;2.

26. Alessandra Biancolillo FM, Jean-Michel Roger. SO-CovSel: A novel method for variable selection in a multiblock framework. Journal of Chemometrics. 2019;34(2).

27. Chu Zhang LZ, Yiying Zhao, Susu Zhu, Fei Liu, Yong He Noise reduction in the spectral domain of hyperspectral images using denoising autoencoder methods. Chemometrics and Intelligent Laboratory Systems. 2020;203(104063).

28. Burger J. Bad Pixel Detection in Hyperspectral Staring Camera Systems. NIR news. 2009;20(1):9-12.

29. Bingkai Liu YL, Lin Wen, Xiang Zhang and Qi Guo. Effects of Hot Pixels on Pixel Performance on Backside Illuminated Complementary Metal Oxide Semiconductor (CMOS) Image Sensors. Sensors. 2023;23(13):6159.

30. Burger J. Bad Pixel Detection in Hyperspectral Staring Camera Systems. NIR news. 2009;20(1).

31. Coomans D, & Broeckaert, I. . Potential Pattern Recognition in Chemical and Medical Decision Making. Research Studies Press, England Letchworth. 1986.

32. Derde MP, & Massart, D.L. . UNEQ: a disjoint modelling technique for pattern recognition based on normal distribution. Analytica Chimica Acta. 1986;184:33-51.

33. Hotelling H. Multivariate quality control. C.Eisenhart MWHWAWE, editor: McGraw-Hill, N.Y.; 1947.

34. Wold S. SM. Chemometrics: Theory and Applications1977.

35. Ruiyang Wu NH. Quadratic discriminant analysis by projection Journal of Multivariate Analysis. 2022;190(104987).

36. Kramer O. K-Nearest Neighbors. In: Heidelberg S-VB, editor. Dimensionality Reduction with Unsupervised Nearest Neighbors 512013. p. 13-23.

37. ∗ AT. Linear Discriminant Analysis: A Detailed

Tutorial. 2017.

38. Ricci NCE. Support Vector Machines. Encyclopedia of Algorithms 2008:928-32.

39. The Darwin annotation tool from V7 labs. Available from: (<https://darwin.v7labs.com>). [Last accessed 24th January 2024].

40. Van Rossum G, Drake Jr FL. Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam; 1995.

41. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2016. Available from: <https://www.R-project.org/>.

42. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. <https://doi.org/10.18637/jss.v028.i05>.

43. Lesnoff M (2023). \_rchemo: Dimension reduction, Regression and Discrimination for Chemometrics\_. R package version 0.0-18, <<https://github.com/mlesnoff/rchemo>>.

44. Antoine Stevens and Leornardo Ramirez-Lopez (2022). An introduction to the prospectr package. R package Vignette R package version 0.2.6.

45. Fearn T. Principal Components Analysis is Not a Classification Method. NIR news. 2016;27(3):3-27.

46. Smith, J., et al. “Polynomial Detrending in Near-Infrared Spectroscopy.” Journal of Spectroscopic Techniques, 20(3), 2010, pp. 123-135.

47. Antonov L. An alternative for the calculation of derivative spectra in the near-infrared spectroscopy. Journal of Near Infrared Spectroscopy. 2017;25(2):145-8.

48. R. J. Barnes MSD, and Susan J. Lister. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. Applied Spectroscopy. 1989;43(5).

49. Jean-Michel Roger AM, and Federico Marini. Preprocessing NIR Spectra for Aquaphotomics. Molecules. 2022;27(20):6795.

50. Akosa JS. Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data. Oklahoma State University; 2017.

51. C. Luz VDO, J.M. Quiles, R. Romano, J. Mañes, G. Meca Biopreservation of tomatoes using fermented media by lactic acid bacteria. 2020;130.

52. RT JM. Aquaphotomics—From Innovative Knowledge to Integrative Platform in Science and Technology. Molecules. 2019;24(15).

53. 2024. AMAfhwacLaottJ.

54. M. Jakubíková JS AK, P. Májek. Near-infrared spectroscopy for rapid classification of fruit spirits. Journal of Food Science and Technology 2016;56(6).

55. D. N. Dalimov GND, and M. Bhatt3. CHEMICAL COMPOSITION AND LIGNINS OF TOMATO AND POMEGRANATE SEEDS. Chemistry of Natural Compounds. 2003;39(1):37-40.

56. Sae Tanaka RT, Masato Yasui Details of glucose solution near-infrared band assignment revealed the anomer difference in the structure and the interaction with water molecules. Journal of Molecular Liquids. 2021;324(114764).

57. Mercedes G. López ASG-G, Elena Franco- Robles. Carbohydrate Analysis by NIRS-Chemometrics. Skvaril KGKaJ, editor. Developments in Near-Infrared Spectroscopy2016.

58. Zoltan Kovacs JM, Petya Veleva , Mitsue Oshima, Shogo Shigeoka, Roumiana Tsenkova Aquaphotomics for monitoring of groundwater using short-wavelength near-infrared spectroscopy. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. 2022;279(121378).

59. Tsenkova RN, I.K. Iordanova, K. Toyoda, and D.R. Brown. Prion protein fate governed by metal binding. Biochemical and biophysical research communications. 2004;325(3):1005-12.

60. . <https://imbalancedlearn.org/stable/references/generated/imblearn.metrics.geometric_mean_score>. Last accesed on the 30th January 2024.