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

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 an "ubiquitous vegetable". Tomatoes are produced globally, either for domestic consumption or as a commodity for international export.

*“The fruit has a large quantity of water, vitamins and minerals, low amounts of proteins and fats, and some carbohydrates. It also contains carotenes, such as lycopene and beta-Carotene”*). {, #1}

Tomato quality is divided into different aspects, commercial, organoleptic and nutritional. Market standard primarily depends on the external appeal (e.g. color, form, size), firmness and shelf life; whereas health quality relies on the nutritional value as well as on the absence of pathogenic hazards or contaminants {Nadia Bertin, 2018 #2}

Tomatoes have a sugary flavor due to their polysaccharide content, what makes them vulnerable to pathogens such as Penicillium, Aspergillus and Mucor. It is interesting to know how a fruit like tomato can reach a significant level of contamination by fungi. Let’s take *Penicillum* expansum as an example.

As can be expected, post-harvest spoilage fungi in tomatoes cause financial deficit for trade and customers{C. Luz, 2020 #3}. The timely identification of disease has the potential to avert losses since prompt measures can be implemented to mitigate more extensive damages.

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 {Nadia Bertin, 2018 #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.

Another simple means to assess the predisposition to microscopic fungi is by tracking the growing and handling conditions of tomato produce withing the supply chain. This correlation may be beneficial in the detection of probable origins of fungal contamination based on historical data.

Although this method can be very useful, it remains a general interconnection. The results or assessments are obtained for the batch of tomatoes as a whole. However, sometimes some tomatoes within the lot are vulnerable while others are not, which is why a more specific method is necessary that allows each sepal to be evaluated individually. Some of these analytical methods will be mentioned below.

First of all, 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 {Daniel Jiménez-Fernández, 2009 #4}.

Also, tomato samples can be tested for mycotoxins, as a high level of these compounds is caused by fungal infection{Mina Nan, 2022 #5}. Some detection solutions are, for instance, chromatography coupled with detector (i.e., mass, ultraviolet, fluorescence, HPLC, etc.) methods, electrochemical biosensors technology and immunological techniques such as such enzyme-linked immunosorbent assay (ELISA), dipsticks, flow-through membranes. 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{Mina Nan, 2022 #5}

Although these analytical methods are specific and accurate, they have several disadvantages. First of all, , they are methods for detecting the disease and not the susceptibility to the disease, that is to say, they evaluate what is happening to the fruit exactly at the moment of the analysis. In the case of disease, the future is already known (if a fruit is infected by fungus, this state will continue in the future); however, if the fruits are not yet infected, these methods cannot predict what will happen to them it in the time ahead. It should be mentioned that at first, susceptibility to fungus presence is invisible to the naked eye. It cannot be detected visually.

From everything detailed above, it is evident that a reliable, non-destructive, specific method is needed to prevent susceptibility to fungal infection in a quick way (since it is a perishable product).This would provide additional support for quality inspectors.

Infrared spectroscopy can provide a solution to this problem. Paul Skolik et al, in 2028, have studied diseased progression in whole tomatoes by AFT-FTIR and have highlighted that plant-pathogen interaction can be identified through alteration in the spectra fingerprint {Paul Skolik, 2019 #6}

Imaging Spectroscopy (or “hyperspectral imaging”,“ HSI”) can be even more useful, because in addition to spectral information, spatial information is obtained. With this technique, it is possible to acquire an image of the entire fruit (or the desired number of fruits) and extract an infrared spectrum from each pixel. This was used, for example, by the authors Huting Wang et al, in 2021, in order to detect tomatoes with early decay {Huting Wang, 2021 #7}

In a previous study carried out by Brdar et al., HSI was evaluated for early detection of weak sepals{Sanja Brdar, 2022 #8} but only one tomato cultivar was studied, no varieties were combined. In addition, no feature selection was carried out before models were trained.

The objectives of our work are: to calibrate and validate intravariety and global models (created with more than one cultivar), to grade the susceptibility to fungal infection by selecting important wavelengths for this purpose, as input for the classification models.

1. **Background**
   1. **Feature selection**

When building chemometrics models it is advisable to choose which variables to use in order to make the models more interpretable. Knowing which variables play an important role in the studied phenomenon is better than making a model incorporating hundreds or thousands of variables without knowing which ones are important, even though some methods such as PLS can tolerate the redundancy of variables.

Furthermore, if a model is built with the intention of being used routinely, it is generally more effective and economical to measure only some variables in the unknown, new samples. When it comes to NIR spectroscopy, there are two main challenges: 1) “*a large number of variables yields a very large solution space”* {J.M. Roger, 2011 #12} and 2) variables are strongly correlated. This carries a risk of overfitting the models. The use of a simple model, which has few variables, has a better chance of being generalized to a new sample than a model that has hundreds of variables and is more likely to fit perfectly to a learning set but has little generalization power.

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

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

When Embedded methods are used, variable selection is carried out at the same time as learning {Nicholas Pudjihartono, 2022 #9}

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. This consists of going back and forth between stages of selection of subsets of variables and stages of evaluation of the subgroups through the application of the selected learning method {Michael D. Sorochan Armstrong, 2022 #10}

In previous works, embedded methods (Random Forest) were used to solve a similar research problem {Sanja Brdar, 2022 #8}. 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.

When using the CovSel algorithm, the significance of each wavelength is assessed by calculating the covariance between each variable and the Y values (responses) {Alessandra Biancolillo, 2019 #11}. The CovSel method works similarly to the NIPALS algorithm in PLS. First of all, both X and Y matrices need to be scaled. 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.

Although the CovSel algorithm was originally designed for regression purposes, can be well adjusted to a discrimination problem using a Y vector with indicator variables {J.M. Roger, 2011 #12}

* 1. Outlier removal

There are typically two ways to proceed when analyzing spectra extracted from hyperspectral images. One is to gather all the spectra (pixels) corresponding to a region of interest, in this case a signal. 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 at spectra level, or 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 same region of interest were considered as a single spectrum.

It is important to mention, however, that not all pixels found in that region provide information, some pixels are considered as “noisy” or “bad pixels”. According to Chu Zhang et al, in 2020, *“Noises care mainly caused by the environment and the sensors”{Chu Zhang a b, 2020 #13}.*

According to James Burger in 2009, bad pixels show a substantial altered spectra compared to their neighbors when detected by specific algorithms. This author also provides a classification of those abnormal pixels: “dead pixels” , “hot pixels” and “stuck pixels”. Dead pixels do not react to light. According to Bingkai Liu in 2023, *“Hot pixels refer to the anomalous pixel with high dark current compared to the normal pixels with moderate dark current increase after irradiation” {Bingkai Liu, 2023 #15}.*  Finally, stuck pixels show an almost steady intermediate value. Moreover, it is interesting to note, that some pixels are always noisy, while others only sporadically; some may show a “*non- linear response to light intensity*”{Burger, 2009 #16}, while some others behave randomly.

In any case, these abnormal pixels behave differently from the rest and can be detected by algorithms in an unsupervised manner.

1. **Materials and Methods**

3.1 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 Tuesday, the 10th of May, 32 samples from each cultivar arrived at the Phenomea Laboratory in Wageningen, Netherlands, without any visible indications of fungal infection.

3.2. Methods

3.2.1. 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. Subsequently, tomatoes were stored in controlled conditions encouraging fungal growth (20°C, in a closed 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 by three experts on day three and four (12th and 13th May), comprised of severity scores from zero (no fungus) to three (severe infection). Ratings of the two days and three experts were averaged.

3.2.2. Spectra extraction from hyperspectral images

Hyperspectral images were converted to false color images. These images were manually annotated with a separate polygon indicating the boundary of each individual sepal. These polygons were converted to pixel masks, which indicated whether or not a pixel was included in the set of pixels belonging to the particular sepal. The pixel spectra of each sepal were collected and then passed to analysis. The dimensions of the initial datasets can be found in Table 1.

**Table 1:** Description of initial datasets

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cultivar name/ number of | Pixels per sepal | Sepals per tomato | Tomatoes per image | Images | 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 |

3.2.3. Data analysis

A chemometric analysis was conducted with the aim of calibrating and validating models to predict the degree of disease in tomatoes as observed by specialists after 4 days of germination. This analysis involved all the standard steps in chemometric treatment, which will be explained below.

1. **Data Exploration**

In this work, different exploratory analysis procedures were carried out at two different levels: pixel and variety by Principal Component Analysis (PCA).

A.1. Pixel Level. In order to explain more clearly how the exploratory analysis was performed using PCA at the pixel level, let's consider only the Cappricia cultivar. The original table (which contained 12816 rows) was first divided into the number of images (2), then into the number of tomatoes per image (16), and finally into sepals (5 or 6). A PCA analysis was carried out on each of the final tables that belong to each sepal. The outliers were removed, the remaining pixels averaged, and the tables finally reassembled according to their labels. The final table, contained 165 spectra, each corresponding to a different sepal. This automatic analysis was repeated for the other cultivars, Brioso and Provine, generating matrices with 164 and 159 spectra respectively (Table 1).

Furthermore, after obtaining all the results, other methods for detecting outliers at pixel level were employed and compared with the previous outcomes: (i) Principal Component Analysis (PCA), (ii) Isolation Forest (IF), (iii) PCA + IF (iv) one-class Soft Independent Modeling by Class Analogy (SIMCA), (v) PCDIST algorithm, (vi) one-class Support Vector Machine and (vii) an univariate approach.

Additionally, three ways of combining relevant spectra belonging to the same area of interest were studied: (i) by taking their average, (ii) by calculating their standard deviation, adding this value as an extra variable, and finally computing the average at the row level and (iii) by pretreating these samples with the Standard Normal Variate pretreatment (SNV), and then calculating their average.

A.2. Cultivar level. Exploration was conducted in each data set by PCA. The outliers removed can be seen in Table 3.

**Table 2:** Outliers removed by PCA on each cultivar.

|  |  |
| --- | --- |
| Cultivar | Outliers removed |
| Brioso | "2599", "4038", "983", "3222", "3955", "3651", "440", "5160" |
| Cappricia | "i2T11S5", "i2T5S3", "i1T1S3", "i1T1S2" |
| Provine | "7790", "9639", "0", "7232","10340", "1002", "4117" |

1. **Pretreatments to Raw Spectra**

Various pre-treated forms of the original spectra were used to calibrate and validate different models, whose performances were compared with each other. These various methods include, for example: Standard Normal Variate (SNV), Derivatives (first and second degrees, second order polynomial and window sizes of 7, 11, 15, 17 and 25), second degree detrend, and combination of these. Only the most satisfactory results will be presented in this document.

1. **Data split**

Samples were distributed in two classes according to visual scoring.

Three different scenarios were created based on how the reference values formed classes 1 (healthy) and 2 (diseased). For example, in Scenario 1, only rating 0 was considered healthy. In Scenario 2, ratings 1 or less were considered healthy and in Scenario 3, healthy class had values equal or smaller than 0.5 (after the averages were done, by 3 experts and by 2 days).

The data sets were divided into calibration (70%) and validation (30%) sets, in a representative way for each class. In other words, Class 1 was split in a proportion of 70/30 and Class 2 was also split in the same proportion 70/30. The number of spectra resulting from this separation, according to each labeling scenario, can be seen in Table 4.

**Table 3:** 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 |

1. **Feature Selection**

In the present work an iterative process was used to select a sparse subset of important variables by CovSel algorithm before their use by the classification models. The optimal number of Important Variables (ivs), between 5 and 39, were chosen for each pretreatment, labelling and cultivar. The selected variables as given by CovSel were then used as input of the classification model.

1. **Selecting the Optimal number of Latent Variables for PLSDA**

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

**F) Calibration and Validation of PLSDA models**

The discriminant model calibrated on the training subset was applied to independent samples in the test set. The results were expressed with classification metrics such as sensitivity, specificity, precision, accuracy and balanced accuracy. Besides raw data, several preprocessing steps were performed and compared. Models were built in the training set using 5 to 39 selected variables by CovSel. PLSDA latent variables were optimized as well, by cross-validation on each tomato.

The iterative process was carried out in the Training Set, and can be summarized as follows:

3.1: First of all, a pretreatment was chosen in order to remove noise from the original spectra. The following algorithms were compared: 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.

3.2: Then, the CovSel algorithm was applied to the pretreated data, 5 important variables were chosen.

3.3: The next step consisted of dividing the Training Set again (which now contains only the important variables); in Validation and Tuning, randomly.

3.4: The selected variables as given by CovSel were then used as input of PLSDA. The Validation set was used to train the PLSDA classification model, and the Tuning set was used to validate it. This last process was carried out with the objective of choosing the optimal number of latent variables for the PLSDA algorithm. The number of latent variables that showed the least error was chosen.

3.5: Steps 3.2 to 3.4 were repeated with an increasing number of variables from 5 to 39.

3.6: Steps 3.1 to 3.5 were repeated by choosing a different pretreatment.

3.7: Steps 3.1 to 3.6 were repeated using Labeling scenarios 2 and 3.

3.8: Steps 3.1 to 3.7 were repeated using different varieties "Cappricia", "Provine" and "Brioso".

3.9: Varieties were added together (Cappricia + Provine, Cappricia + Provine + Brioso, Cappricia + Brioso, Brioso + Provine), then steps 3.1 to 3.7 were repeated.

3.10. Steps 3.1 to 3.7 were replicated, choosing the Cappricia variety as the Training set and the Brioso variety as the Test Set.

3.11. Item 3.10 was reiterated, but choosing Provine as a Test Set.

Out of all the results obtained, the model that presented the lowest Balanced Accuracy (BA) value was chosen.

1. **Results and Discussion**

**4.1 Intravariety models**

Figure 1 (A, B and C) compares the different results obtained using different classification criteria between the healthy and the diseased classes. As can be observed, 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 non-logical 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 Scenario 1 was chosen.

Another indicator of class balancing was the correlation between the accuracy (blue line) and the balanced accuracy (BA) (orange line). When the classes are balanced, these metrics should be almost identical, and their lines will overlap as we can 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.

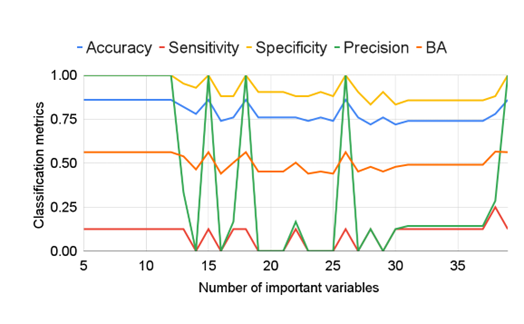
High BA values showed that the model performances were good for both classes. On the other hand, high Accuracy metric showed that the model performed well in general. When using Scenarios 1 and 3, this was the case for the majority classes only.

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.

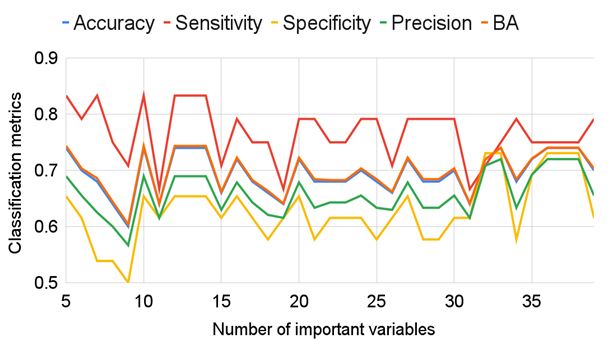
A one-class classification could also have been used, where all samples similar to the samples of one class are considered, and the others discarded by the model. In this case, it does not matter whether the classes are balanced or not, nor does it matter the number of final classes. However, these models 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.

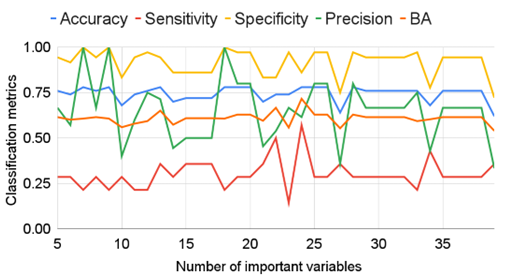
A

****

B

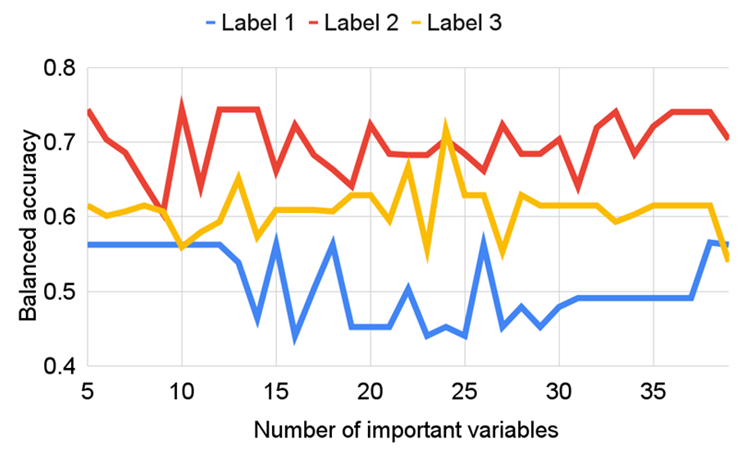
****

C

****

**Figure 1:** Cappricia classification metrics in different labelling scenarios (A: Label 1; B: Label 2; C: Label 3) 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 (iv) chosen by CovSel can be seen in Figure 2, 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.

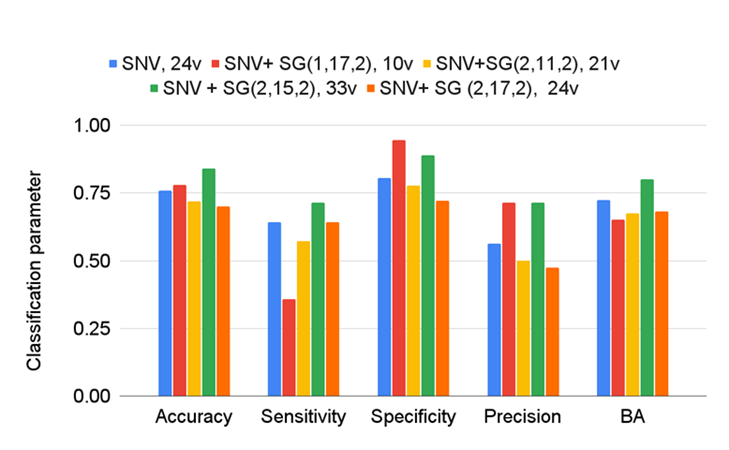


**Figure 2:** Comparison of Balanced Accuracy in different labelling scenarios according to different number of important variables as input for PLSDA.

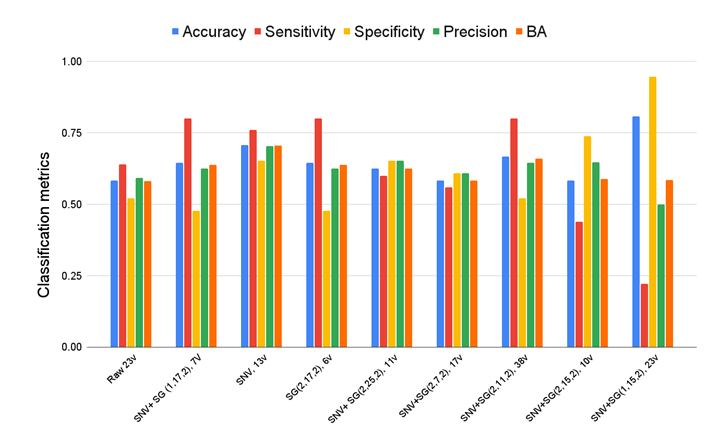
Figure 3 shows the results of the optimized models for Cappricia, Provine and Brioso, choosing Scenario 2. The following parameters generated the optimal models:

* Cappricia: SNV + Second Derivative (2, 15, 2). Then 33 iv
* Provine: SNV, then 13 iv
* Brioso: Raw spectra, then 18 iv

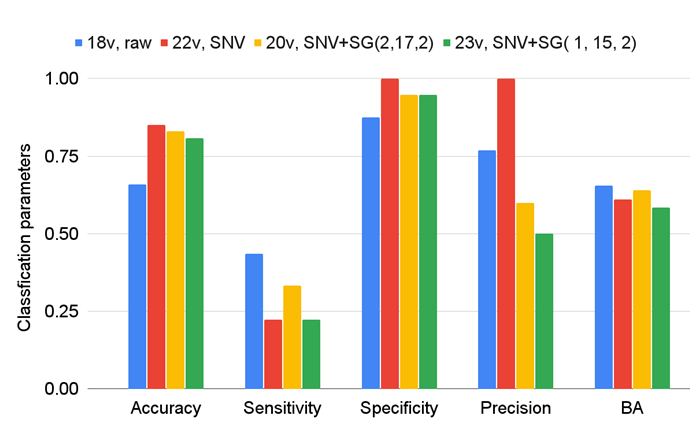
Cappricia



Provine



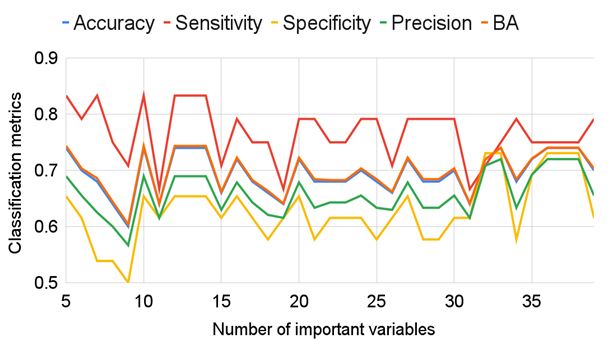
Brioso



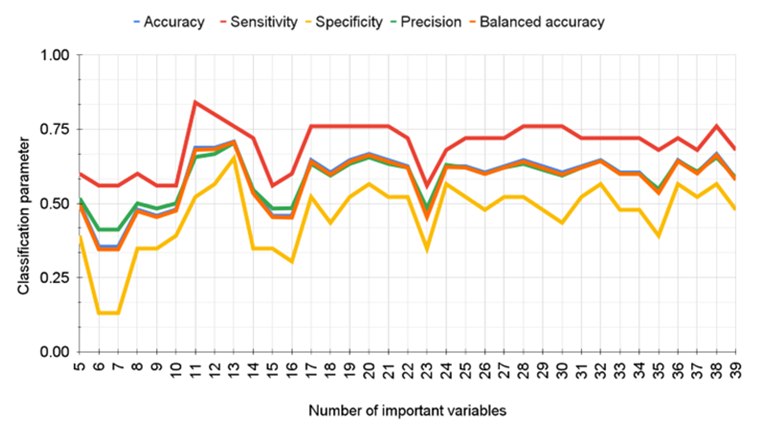
**Figure 3:** Comparison of Balanced Accuracy in different labelling scenarios according to different number of important variables as input for PLSDA.

Figure 4 shows how the values of Balanced Accuracy, Accuracy, Sensitivity, Specificity and Precision changed when different PLSDA models were calibrated with different numbers of important variables (from 5 to 39) were tested in the validation sets.

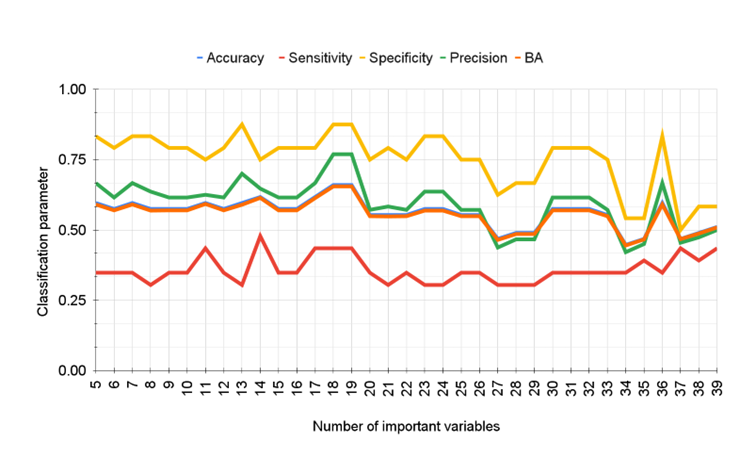
Cappricia



Provine



Brioso



**Figure 4:** Classification measures of the validation sets according to different number of important variables as input for PLSDA. 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 iv. Brioso: Optimal model was found with raw data and 18 v.

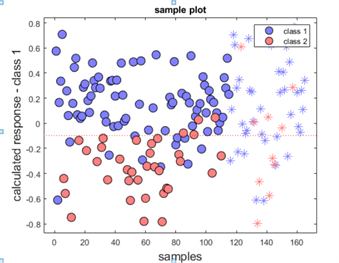
Cappricia: Optimal model was found with pretreated data (SNV+ SG (2,15,2)) and 33 most important variables. Provine: Optimal model was found with SNV pretreated data and 13 most iv Brioso: Optimal model was found with raw data and 18 ivs.

**4.2. Intervariety models**

The intervariety models presented high sensitivity for one class (0.91; 1) but not for the others (0.22, 0.03) (Table 5) in the validation sets. The score plots of the model calibrated in Cappricia showed a good separation between the classes; however, errors were observed when the validation samples were projected in this coordinate system (Figure 4). This is why, other models were investigated, calibrated and validated with combined classes.

**Table 5:** PLSDA classification model calibrated in Cappricia and validated in Brioso and Provine. Pretreatment used: SNV+ SG(2, 17, 2); all spectral range and 3 Latent Variables (LVs).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data set** | **Real/**  **predicted** | **Healthy** | **Diseased** | **NA** | **SEN** | **SPE** | **PRE** | **AC** | **BA** |
| **Calibration**  **Cappriccia** | **Healthy** | **66** | **11** | **0** | **0.86** | **0.55** | **0.63** | **0.70** | **0.71** |
| **Diseased** | **38** | **46** | **0** | **0.55** | **0.86** | **0.81** | **0.70** | **0.71** |
| **Validation**  **Brioso** | **Healthy** | **67** | **7** | **0** | **0.91** | **0.22** | **0.54** | **0.56** | **0.57** |
| **Diseased** | **57** | **16** | **0** | **0.22** | **0.91** | **0.70** | **0.56** | **0.57** |
| **Validation**  **Provine** | **Healthy** | **80** | **0** | **0** | **1** | **0.03** | **0.54** | **0.54** | **0.52** |
| **Diseased** | **69** | **2** | **0** | **0.03** | **1** | **1** | **0.54** | **0.52** |



**Figure 4:** PLSDA scores plots for the model calibrated in Cappricia.

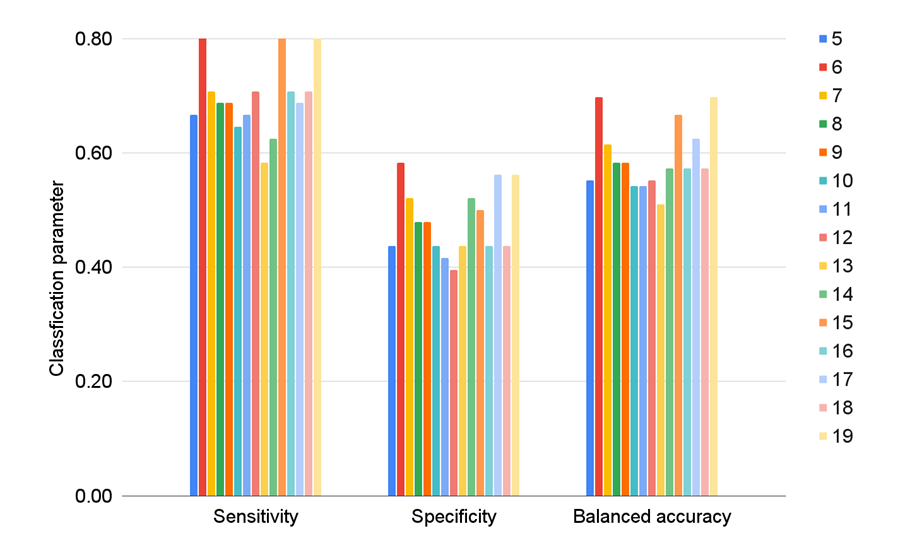
**4.3. Global models**

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 high number may seem strange, but it can be understood due to the complexity of adding two different varieties in one model.

In Figure 6, the values of Sensitivity, Specificity and Balanced Accuracy are shown for an increasing number of variables between 5 and 19. As can be seen, choosing only six variables produced similar results in terms of balanced accuracy as choosing 19.

**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** |



**Figure 6:** Classification parameters of the validation sets for the global model trained with Cappricia and Provine according to the number of important variables as input for PLSDA. Data were pretreated with SNV + SG (2,17,2).

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

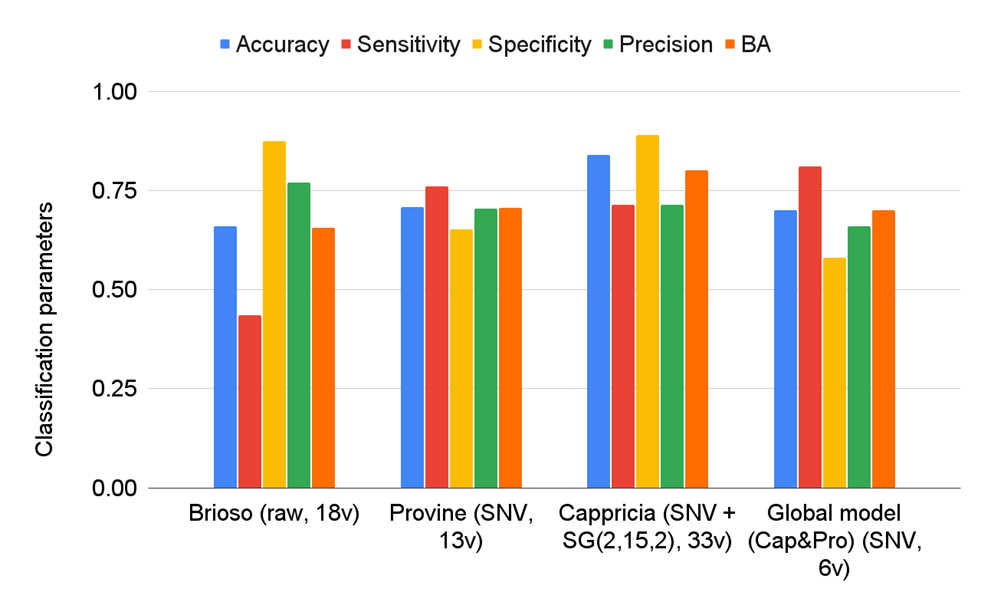
Healthy sepals that were correctly predicted as healthy: Cappricia: 0.71; Provine: 0.76; Global model: 0.81. Diseased sepals correctly classified as diseased: Cappricia: 0.89; Provine: 0.65; Global model: 0.58.

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

The intervariety models calibrated in Cappricia showed high sensitivity in class one (0.91) and low specificity in class two; this was consistent in prediction for Brioso (0.03) and Provine (0.11).

**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** | **Intervariety**  **Cal: Cap, Val: Bri**  **Raw, 10v**  **Label 3** | **Global model**  **SNV, (Cap+Pro)**  **6v**  **Label 2** |
| Accuracy | 0.83 | 0.84 | 0.71 | 0.71 | 0.66 | 0.61 | 0.70 |
| Misclassification rate | 0.17 | 0.11 | 0.29 | 0.29 | 0.34 | 0.39 | 0.30 |
| Sensitivity or recall | 0.89 | 0.71 | 0.08 | 0.76 | 0.43 | 0.32 | 0.81 |
| Specificity | 0.64 | 0.89 | 0.97 | 0.65 | 0.88 | 0.89 | 0.58 |
| Precision | 0.89 | 0.71 | 0.50 | 0.70 | 0.77 | 0.93 | 0.66 |
| Balanced accuracy | 0.77 | 0.80 | 0.52 | 0.71 | 0.65 | 0.61 | 0.70 |
| Geometric mean | 0.75 | 0.79 | 0.27 | 0.70 | 0.62 | 0.53 | 0.69 |
| F-measure | 0.89 | 0.71 | 0.14 | 0.73 | 0.55 | 0.48 | 0.73 |
| Youden’s Index | 0.53 | 0.60 | 0.05 | 0.41 | 0.31 | 0.21 | 0.39 |
| Positive likelihood ratio | 2.47 | 6.45 | 2.67 | 2.17 | 3.58 | 2.91 | 1.93 |
| Negative  likelihood ratio | 0.17 | 0.32 | 0.95 | 0.37 | 0.65 | 0.76 | 0.33 |



**Figure 7:** Classification parameters of the validation sets for the optimal models validated in this study.

Table 8 shows the important wavelengths chosen in the optimized models to discriminate susceptibility to fungal infection. When it comes to the pretreated global model, both sets of important variables are shown (19 and 6).

**Table 8:** Important wavelengths found in this study to discriminate susceptibility to fungal infection. Yellow painted: Six most important variables selected by: CovSel (yellow painted); PLSDA (underlined).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| GM-Raw | GM  Snv+SG(2,17,2) | Cappricia-Raw | Cappricia  SNV + SG(2,15,2), 33v  Label 2 | Brioso  Raw, 18 v | Provine Raw, 14 v |
| "937.33" "944.25" "951.16"  "965" "1034.3" "1055.2" "1089.9" "1138.7" "1152.7" "1229.6" "1278.7"  "1384.3" "1419.6" "1554.3" "1625.5" "1689.8" "1697" "1704.1" "1718.4" | "958.08" "999.63" "1006.6" "1034.3" "1089.9" "1117.8" "1173.6" "1201.6" "1222.6" "1278.7" "1313.9"  "1363.2" "1384.3" "1476.2" "1497.5" "1589.9" "1618.4" "1632.7"  "1697" | "937.33" "944.25" "951.16" "958.08" "965" "971.92" "1041.3" "1062.1" "1083" "1089.9" "1131.8" "1145.7" "1180.6" "1208.6" "1320.9" "1384.3" "1412.6" "1426.7" "1440.8" "1462.1" "1476.2" "1497.5"  "1547.2" "1611.3" "1639.8" "1668.4" "1675.5" "1682.7" "1689.8" "1697" "1704.1" "1711.3" "1718.4" | "971.92" "992.7" "999.63" "1013.5" "1041.3" "1062.1" "1083" "1103.9" "1124.8" "1180.6" "1201.6"  "1236.6" "1292.8" "1306.8" "1327.9" "1356.1" "1370.2" "1398.4" "1419.6" "1469.2" "1483.3" "1497.5"  "1518.8" "1540.1" "1582.8" "1589.9" "1597" "1611.3" "1625.5" "1639.8" "1646.9" "1661.2" "1682.7" | "937.33" "944.25" "951.16" "971.92" "1062.1" "1089.9" "1208.6" "1292.8" "1398.4" "1433.8" "1604.1"  "1618.4" "1646.9" "1689.8" "1697" "1704.1" "1711.3" "1718.4" | "937.33" "944.25" "951.16" "1089.9" "1138.7" "1201.6" "1299.8" "1377.3" "1419.6" "1504.6" "1654.1"  "1697" "1711.3" "1718.4" |

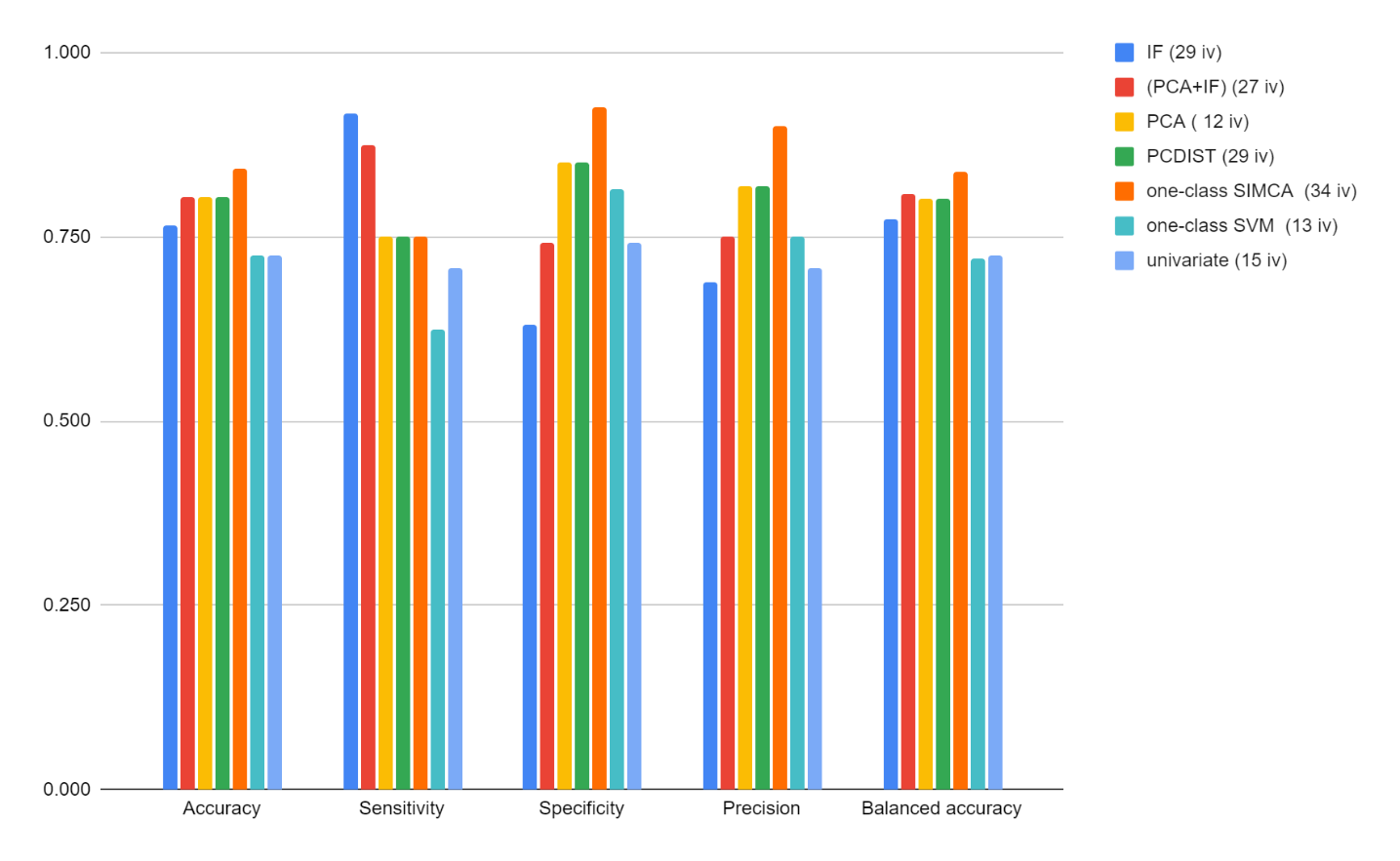
**4.4. Comparison of methods for Outlier Removal at pixel level based on the final performance of the classification models.**

In this section, only the Cappricia variety was considered for study purposes. The results of the classification errors in the validation sets for Class 1 of different outlier removal methods at pixel level can be seen in Table 9. As can be seen, the best results in terms of balanced accuracy were found by detecting outliers with one-class SIMCA, and then taking the average of the remaining pixels.

**Table 9:** Results of the classification errors in the validation sets for Class 1 of different outlier removal methods at pixel level. The optimal number of important variables as input for PLSDA is included.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Average** | | | | | | |
|  | **IF**  **(29 iv)** | **(PCA+IF) (27 iv)** | **PCA ( 12 iv)** | **PCDIST (29 iv)** | **one-class SIMCA**  **(34 iv)** | **one-class SVM (13 iv)** | **univariate (15 iv)** |
| **Accuracy** | 0.765 | 0.804 | 0.804 | 0.804 | 0.843 | 0.725 | 0.725 |
| **Sensitivity** | 0.917 | 0.875 | 0.750 | 0.750 | 0.750 | 0.625 | 0.708 |
| **Specificity** | 0.630 | 0.741 | 0.852 | 0.852 | 0.926 | 0.815 | 0.741 |
| **Precision** | 0.688 | 0.750 | 0.818 | 0.818 | 0.900 | 0.750 | 0.708 |
| **Balanced accuracy** | 0.773 | 0.808 | 0.801 | 0.801 | 0.838 | 0.720 | 0.725 |
|  | **Standard deviate as a feature + average** | | | | | | |
|  | **IF**  **(15 iv)** | **PCA + IF (17 iv)** | **PCA (28 iv)** | **PCDIST (30 iv)** | **one-class SIMCA (33 iv)** | **one class- SVM (28 iv)** | **univariate**  **(5 iv)** |
| **Accuracy** | 0.745 | 0.804 | 0.725 | 0.725 | 0.765 | 0.765 | 0.725 |
| **Sensitivity** | 0.583 | 0.875 | 0.750 | 0.750 | 0.750 | 0.833 | 0.833 |
| **Specificity** | 0.889 | 0.741 | 0.704 | 0.704 | 0.778 | 0.704 | 0.630 |
| **Precision** | 0.824 | 0.750 | 0.692 | 0.692 | 0.750 | 0.714 | 0.667 |
| **Balanced accuracy** | 0.736 | 0.808 | 0.727 | 0.727 | 0.764 | 0.769 | 0.731 |
|  | **SNV + average** | | | | | | |
|  | **IF**  **(38 iv)** | **(PCA+IF) 30 iv** | **PCA (15 iv)** | **PCDIST (14 iv)** | **one-class SIMCA**  **(38 iv)** | **one-class SVM (28 iv)** | **univariate (24 iv)** |
| **Accuracy** | 0.804 | 0.725 | 0.667 | 0.765 | 0.745 | 0.725 | 0.784 |
| **Sensitivity** | 0.708 | 0.792 | 0.833 | 0.667 | 0.875 | 0.792 | 0.875 |
| **Specificity** | 0.889 | 0.667 | 0.519 | 0.852 | 0.630 | 0.667 | 0.704 |
| **Precision** | 0.850 | 0.679 | 0.606 | 0.800 | 0.677 | 0.679 | 0.724 |
| **Balanced accuracy** | 0.799 | 0.729 | 0.676 | 0.759 | 0.752 | 0.729 | 0.789 |

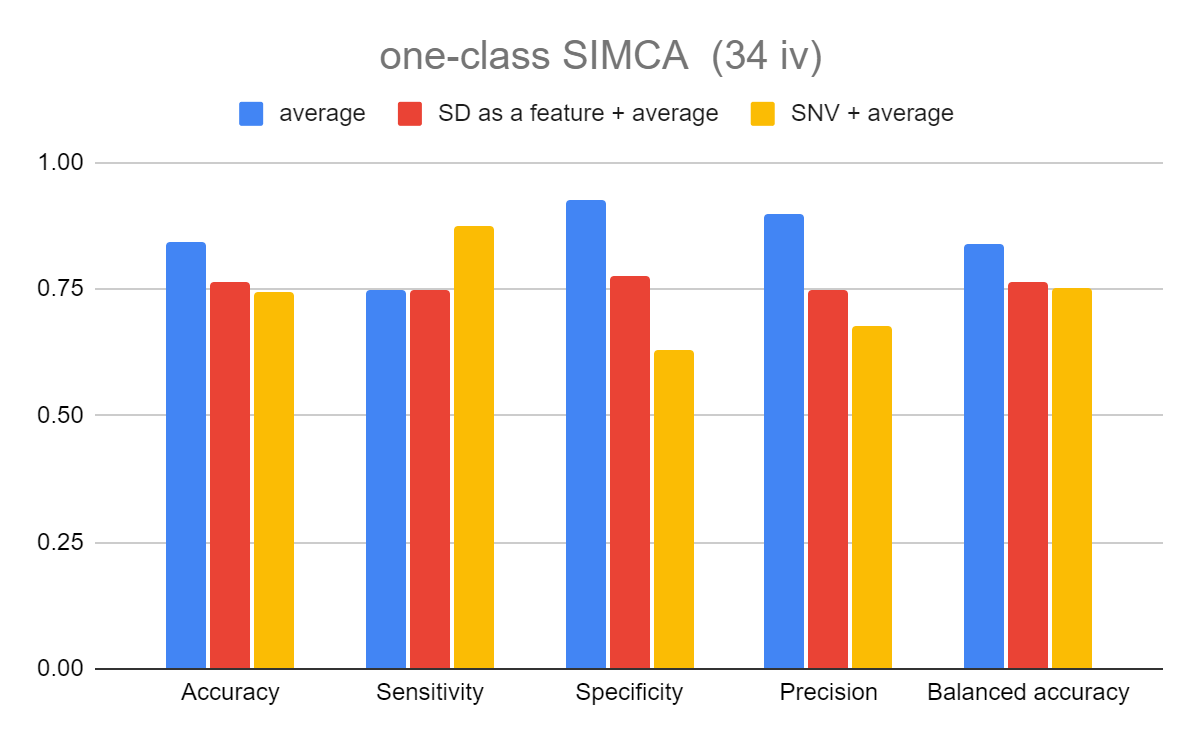
Figure 8 shows prediction errors of the comparison of different outlier removal methods, when the remaining pixels were averaged.



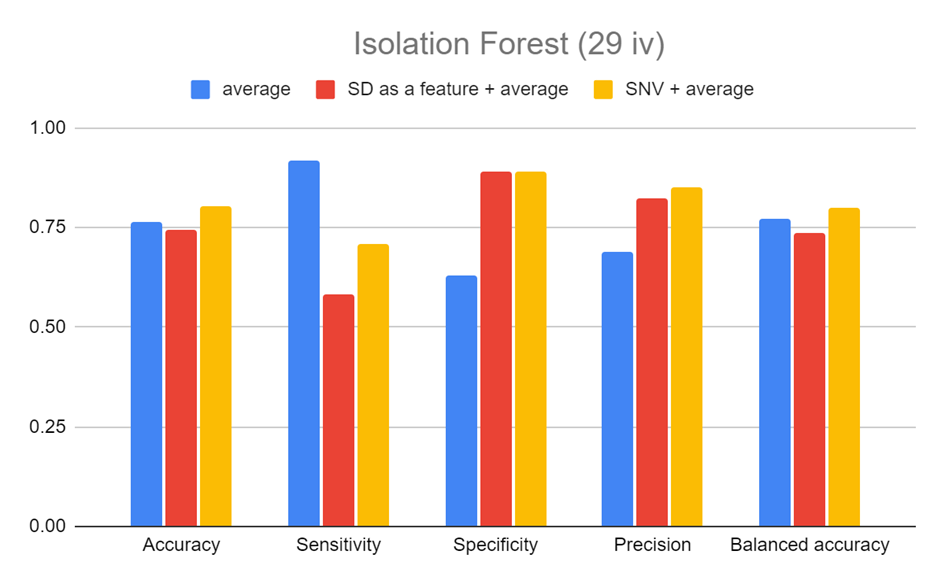
**Figure 8:** Prediction errors of the comparison of different outlier removal methods, when the remaining pixels were averaged.

In the case of one-class SIMCA, it is evident that simply averaging the inliers provides better results, except for sensitivity, which benefits from performing a Standard Normal Variate (SNV) beforehand.

A



B

****

**Figure 9:** Classification metrics obtained when different methods for outlier removal were used, and their comparison between different ways of averaging the Inlier pixels: A) One Class SIMCA B) Isolation Forest.

1. **Conclusion**

A new global model was calibrated on two different tomato cultivars: Cappricia and Provine; and evaluated in independent samples of the same varieties, using Standard Normal Variate and 6 important variables chosen by CovSel. The optimized model achieved a sensitivity of 0.81, specificity of 0.58 and balanced accuracy of 0.70. This model presented potential as a fast alternative method to grade recently harvested tomatoes before the fungal infection is visually observed.

Novelty of this work - investigate HSI to capture the sepal susceptibility to fungal infection by chemometric analysis of different varieties of tomatoes.

The results from this research reaches to a 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. J.M. Roger, B. P., D. Bertrand, E. Fernandez-Ahumada (2011). "CovSel: Variable selection for highly multivariate and multi-response calibration. Application to IR spectroscopy." Chemometrics and Intelligent Laboratory Systems 106(2).

2) Chapter 2. Tomato. DOI: <https://www.oecd-ilibrary.org/docserver/9789264279728-6-en.pdf?expires=1694684540&id=id&accname=oid006406&checksum=62D03306F89450B59CFEE6DA4F9811F5>

3) C. Luz, V. D. O., J.M. Quiles, R. Romano, J. Mañes, G. Meca (2020). Biopreservation of tomatoes using fermented media by lactic acid bacteria. 130. doi:https://doi.org/10.1016/j.lwt.2020.109618

4) Alessandra Biancolillo, F. M., Jean-Michel Roger. (2019). SO-CovSel: A novel method for variable selection in a multiblock framework. Journal of Chemometrics, 34(2). doi: <https://doi.org/10.1002/cem.3120>

5) Daniel Jiménez-Fernández, M. M.-B., Juan A. Navas-Cortés, & Rafael M. Jiménez-Díaz, B. B. L. (2009). Identification and quantification of Fusarium oxysporum in planta and soil by means of an improved specific and quantitative PCR assay. Applied Soil Ecology, 46(3), 372-382. doi:https://doi.org/10.1016/j.apsoil.2010.10.001

6) Huting Wang, R. H., Mengyun Zhang, Zhiqiang Zhai, Ruoyu Zhang. (2021). Identification of tomatoes with early decay using visible and near infrared hyperspectral imaging and image-spectrum merging technique. Journal of Food Process Engineering 44(4). doi:https://doi.org/10.1111/jfpe.13654

7) J.M. Roger, B. P., D. Bertrand, E. Fernandez-Ahumada. (2011). CovSel: Variable selection for highly multivariate and multi-response calibration. Application to IR spectroscopy. Chemometrics and Intelligent Laboratory Systems, 106(2). doi:ff10.1016/j.chemolab.2010.10.003f

8) Michael D. Sorochan Armstrong, P. d. l. M., J. Harynuk\*. (2022). Review of Variable Selection Methods for Discriminant-Type Problems in Chemometrics. Frontiers. doi: <https://doi.org/10.3389/frans.2022.867938>

9) Mina Nan, 2 Huali Xue,1,\* and Yang Bi3,\*. (2022). Contamination, Detection and Control of Mycotoxins in Fruits and Vegetables. Toxins, 14(5), 309. doi:10.3390/toxins14050309

10) Nadia Bertin, M. G. (2018). Tomato quality as influenced by preharvest factors. Scientia Horticulturae, 233(15). doi:https://doi.org/10.1016/j.scienta.2018.01.056

11) Nicholas Pudjihartono, F., W. Kempa-Liehr, M. O'Sullivan. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers, 2. doi: <https://doi.org/10.3389/fbinf.2022.927312>

12) Paul Skolik, M. R. M., Francis L. Martin. (2019). ATR-FTIR spectroscopy non-destructively detects damage-induced sour rot infection in whole tomato fruit. Planta, 249(3), 925-939. doi:https://www.jstor.org/stable/48701885

13) Sanja Brdar, M. P., Esther Hogeveen‑van Echtelt, Manon Mensink, ŽeljanaGrbović, Ernst Woltering, Aneesh Chauhan. (2022). Predicting sensitivity of recently harvested tomatoes and tomato sepals to future fungal infections. Nature, 11:23109 doi: <https://doi.org/10.1038/s41598-021-02302-2>

14) Chu Zhang a b, L. Z. a. b., Yiying Zhao a b, Susu Zhu a b, Fei Liu a b, Yong He a b (2020). "Noise reduction in the spectral domain of hyperspectral images using denoising autoencoder methods." **203**(104063).

15) Burger, J. (2009). "Bad Pixel Detection in Hyperspectral Staring Camera Systems." NIR news **20**(1): 9-12.

16) Bingkai Liu, Y. L., Lin Wen, Xiang Zhang and Qi Guo (2023). "Effects of Hot Pixels on Pixel Performance on Backside Illuminated Complementary Metal Oxide Semiconductor (CMOS) Image Sensors." Sensing and Imaging **23**(6159).