

“Predicting fungal infection sensitivity of sepals in harvested tomatoes using Hyperspectral Imaging 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

¹Wageningen University and Research, Netherlands

²Biosense Institute, Serbia



Table of contents

1) Introduction

2) Materials and methods

Samples

Data description and analysis

3) Results

4) Conclusions

Introduction



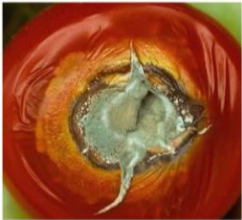
Problem definition



Problem definition

- **Motivation:** tomato - highly sensitive during harvest and post-harvest processes (transport, packaging and storage), susceptible to pathogenic fungi
- This leads to post-harvest losses, reaching **up to 30%** in some developing countries
- Early weakness of the sepals is not visible to the naked eye - no method exists of detecting this automatically prior to the infection

Aspergillus



Penicillium



Mucor



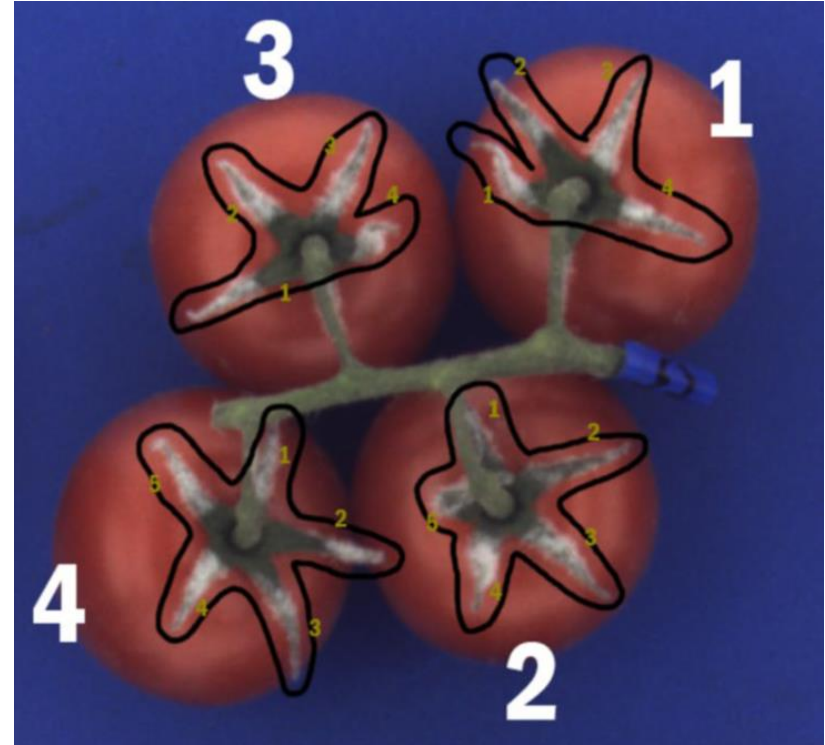
????



It begins ...



... and it gets worse (after 5 days)



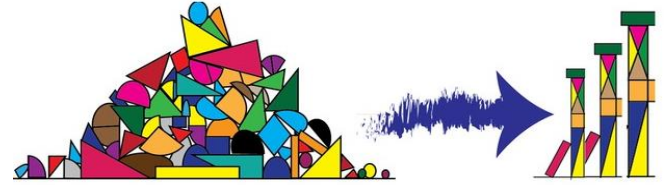
Objectives of this feasibility study

- Investigate if there is a correlation between the hyperspectral data captured at harvest and the fungal infection observed 3 and 4 days later by chemometrics
- Calibrate and validate a global model to grade the susceptibility to fungal infection, and to identify important wavelengths as input for the model

Background theory



Feature selection



Increasing the number of variables:

Introduces unnecessary **NOISE** for discrimination, especially if they are strongly correlated

Carries a risk of **OVERFITTING** the models

Using a simple model with few variables has a better chance of being generalized to a new sample than a model with hundreds of variables, which may fit the training set perfectly well but has limited generalization power

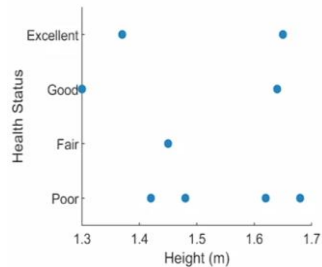
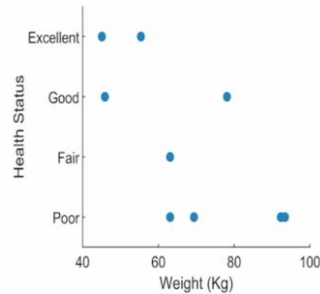
3 Approaches:

1. Variable Transformation and/or Selection
2. Discretization
3. Group Summary

Variable transformation

This involves applying an equation to existing variables to create a new feature

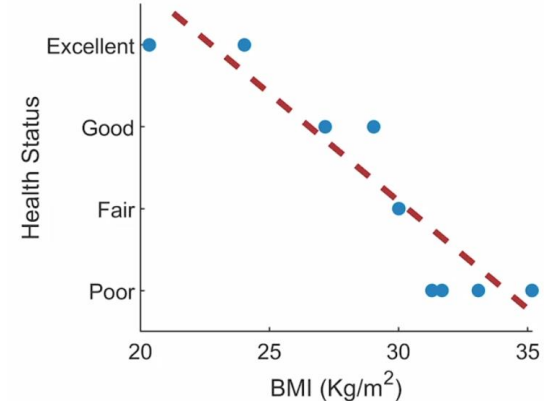
Example: Is it possible to accurately classify the health status of each individual from the original variables of age, height, weight and location?

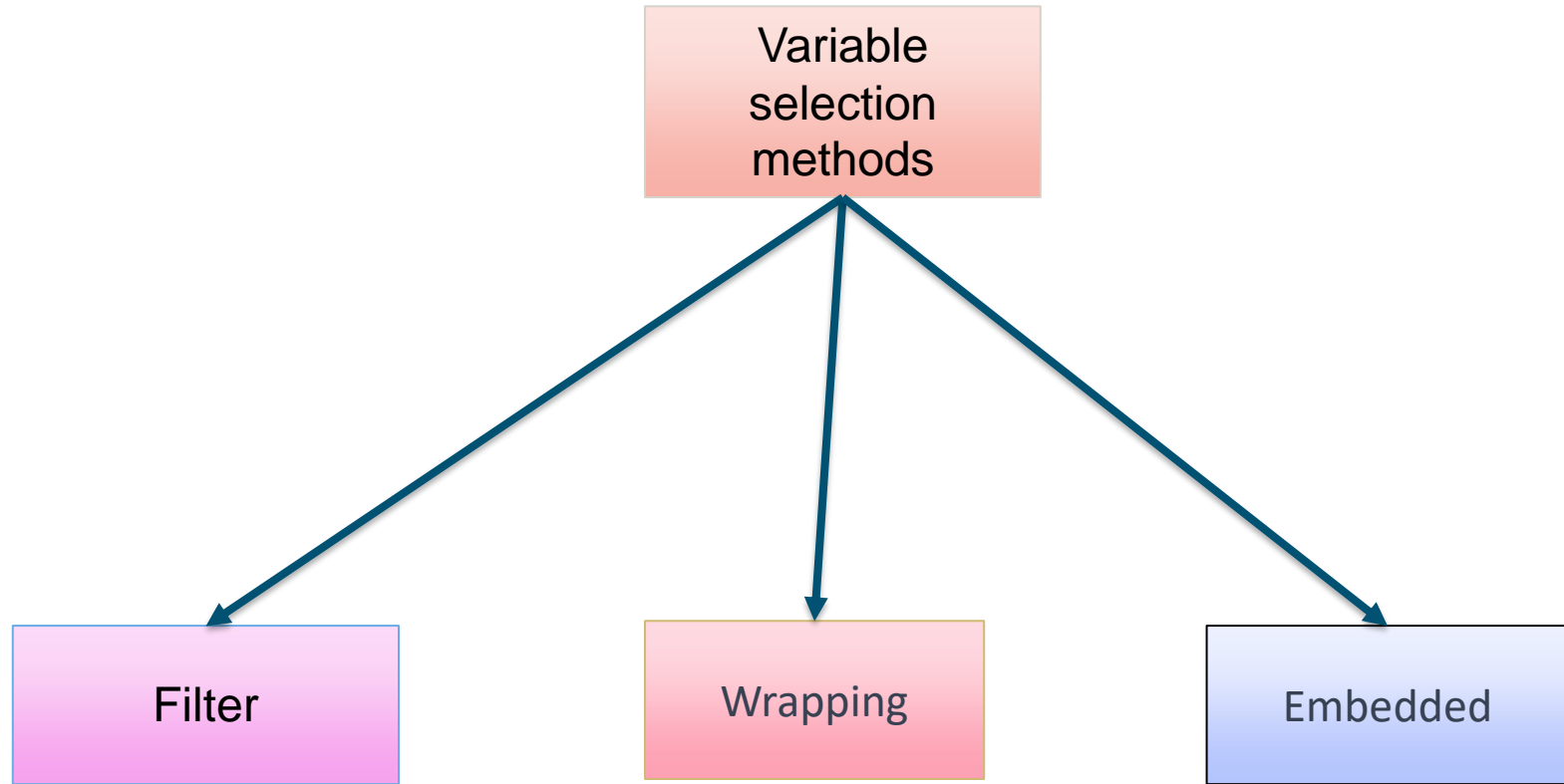


No clear correlation

Correlation

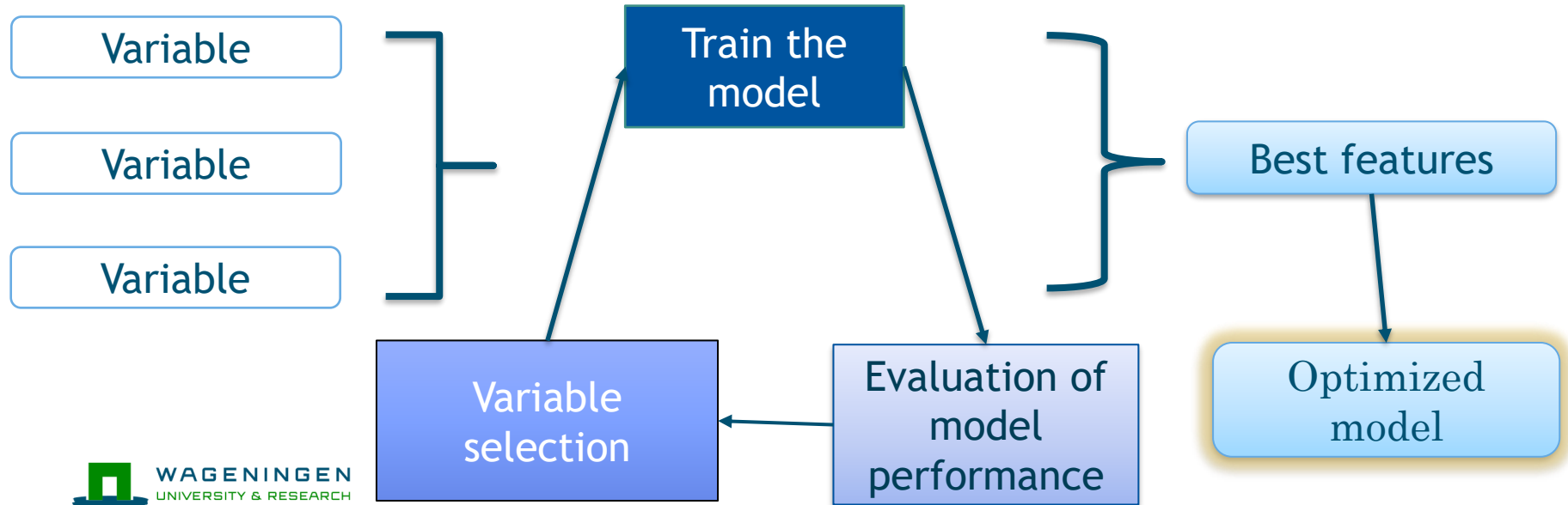
$$\text{BMI} = \frac{\text{Weight (Kg)}}{\text{Height (m)}^2}$$





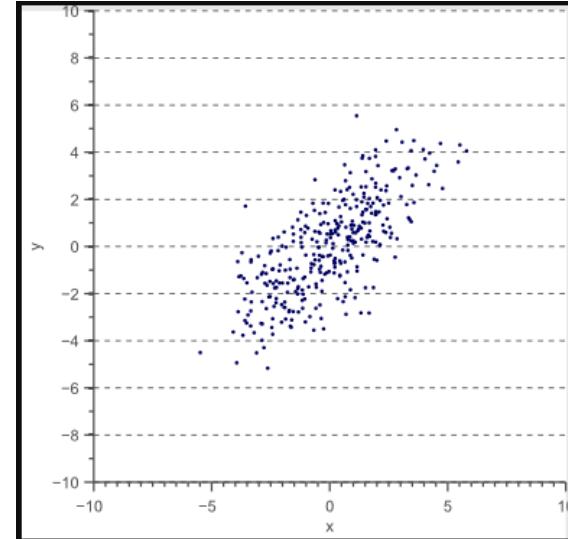
Wrapping methods

These variable selection methods depend on model performance. It is an iterative process of back and forth, where the chosen features are evaluated in relation to the final model performance in sequential stages.



Covariance

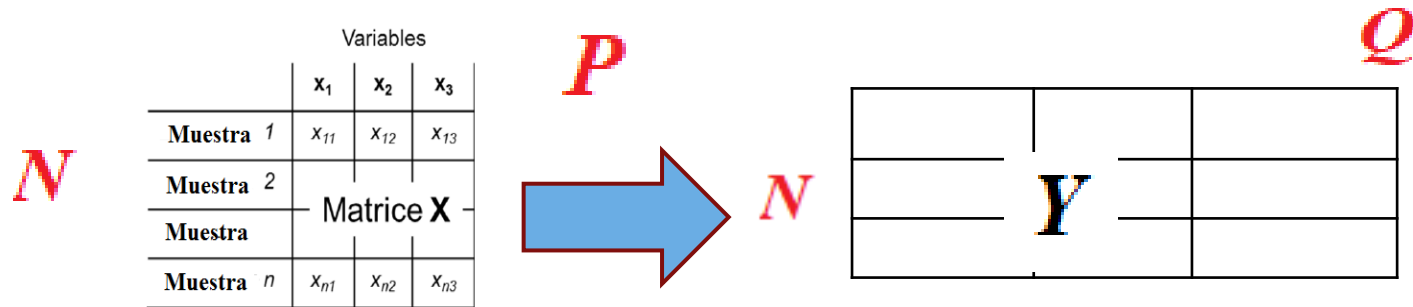
- ❑ Indicates the degree of joint variation between two random variables with respect to their means
- ❑ It can be used to understand the direction of the relationship between two variables
- ❑ The correlation coefficient is equal to the covariance divided by the product of the standard deviations of the variables



$$Cov(X, Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Wrapping method: Covariance selection (CovSel)

In each iteration, one variable **X** is selected on a criterion of the maximization of the covariances with **Y**



Once the variable with the highest covariance is isolated and selected, all other predictive factors and responses are orthogonalized with respect to it, and the process is repeated until the fixed number of variables has been selected

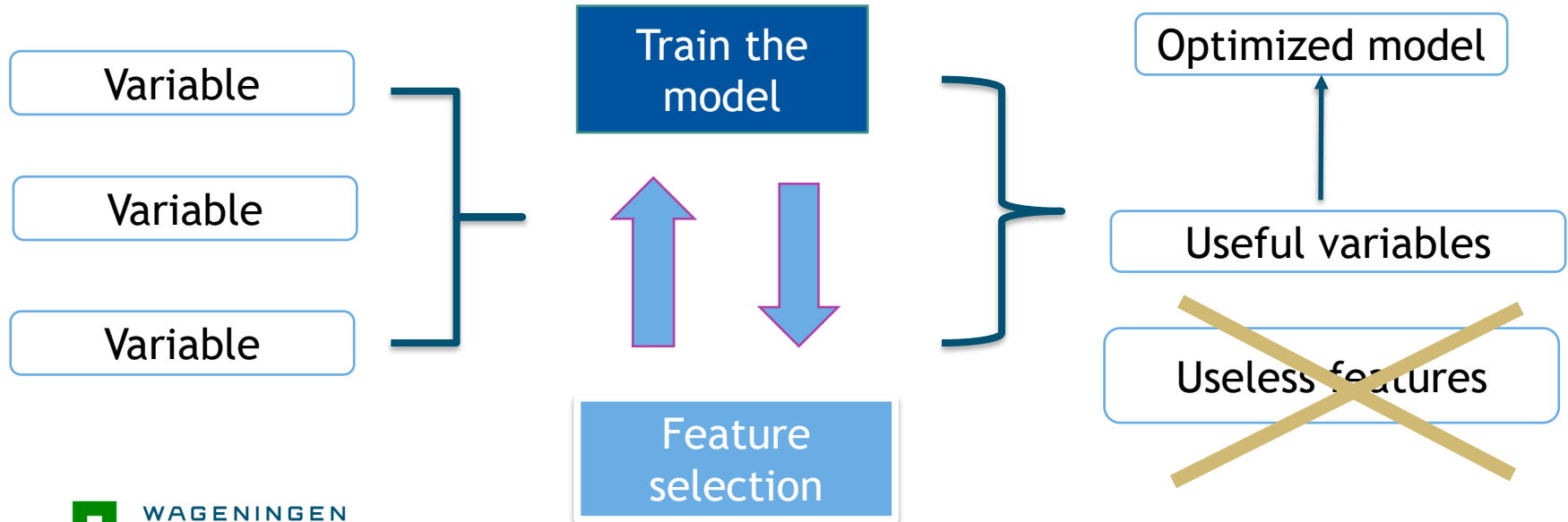
Covariance selection (CovSel)

The selected variables from X should have:

- ☐ **Good predictive power for Y**
- ☐ **The highest possible variability**

Embedded methods

They automatically perform feature selection as part of the model training
The result is a trained model that highlights the useful features and disregards the rest



Principles and Objectives of Discrimination

- We have a data matrix $(n \times p)$ where n samples were measured for p quantitative variables, and a vector Y of size n measured on the same samples
- This vector represents the membership of each sample to each class K
- Each class contains at least one sample, and each sample belongs to a single class

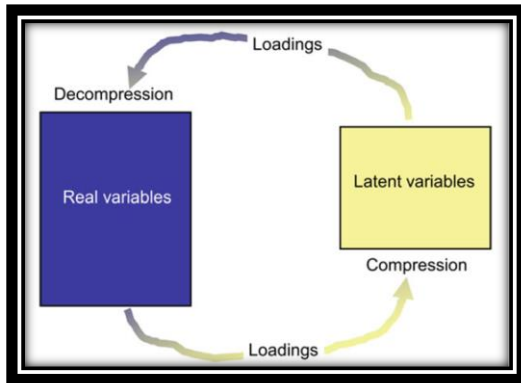
Partial Least Squares Discriminant Analysis

- Objective: To achieve a linear transformation that maps the data into a lower-dimensional space with the least possible error
- Supervised version of PCA
- In **PCA**, the transformation preserves (in its first principal component) the maximum possible variation in the original data
- In **PLS-DA**, the transformation preserves (in its first principal component) the maximum possible **covariance** between the original data and their labeling

Both can be described as iterative processes in which the error term is used to define the next principal component

Partial Least Squares Discriminant Analysis

- It consists of a classical PLS regression where the response variable is a category expressing the membership of samples in classes
- The relevant sources of data variability are modeled by the Latent Variables (LVs) which are linear combinations of the original variables



A fictitious matrix (Y) that records membership with 1s and 0s is combined with a spectral set (X), and PLS is implemented in the normal manner

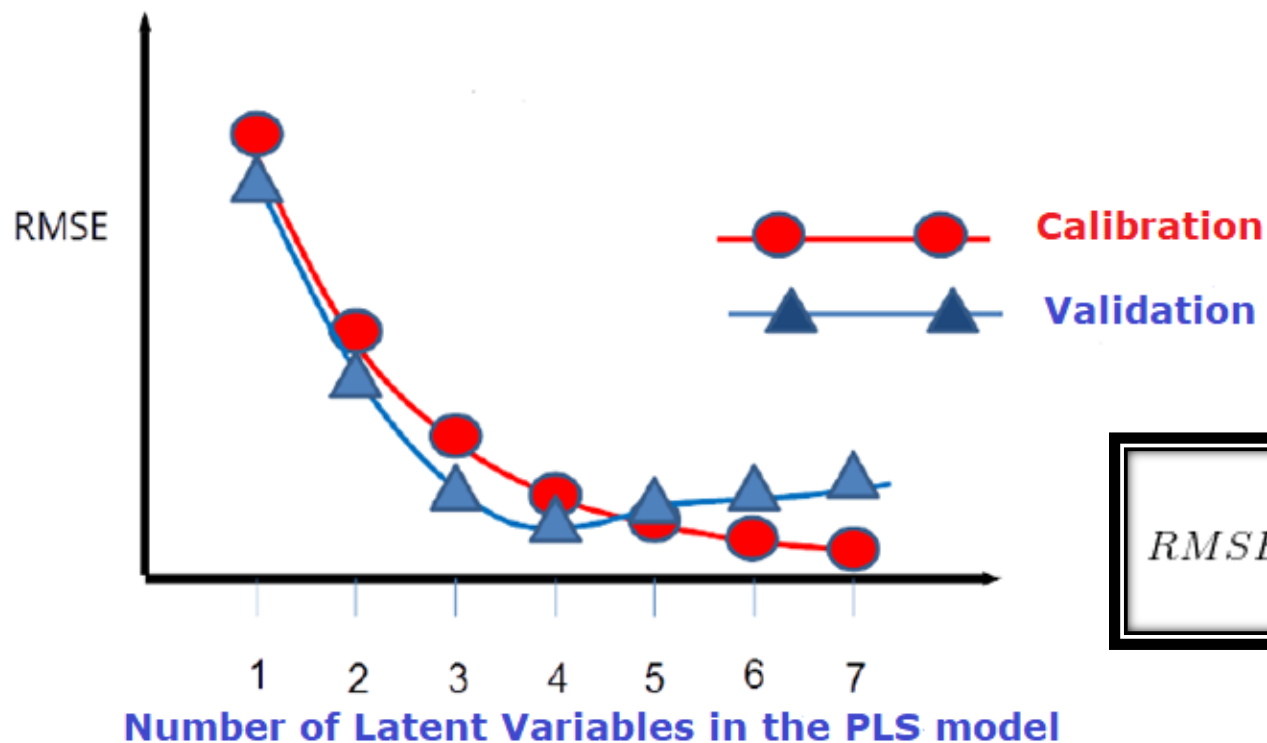


$$y = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 1 \\ 3 \\ 1 \\ 2 \\ 2 \\ 3 \end{pmatrix} \rightarrow Y = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Partial Least Squares Discriminant Analysis

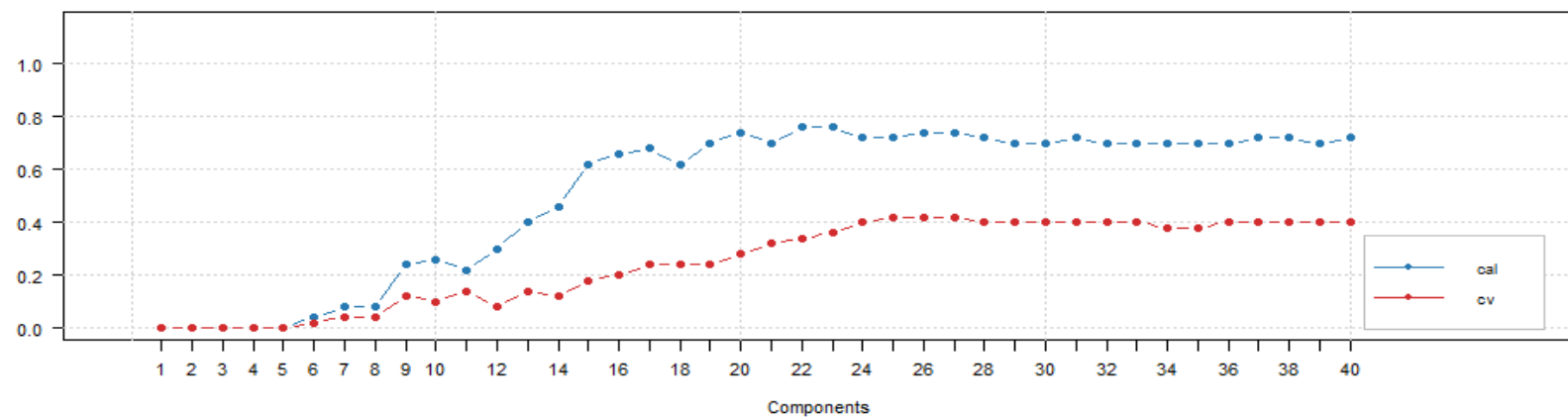
- PLS-DA provides estimated values for each sample and for each class.
- These values will not be exactly 1 or 0; however, if the calculated **y** is closer to 0, then the sample likely does not belong to that class, while a value closer to 1 would indicate the opposite
- To make a class assignment, a **threshold** can be defined for each class
- Thresholds can be calculated on the basis of the Bayes theorem

Study of Error as a Function of Dimensionality

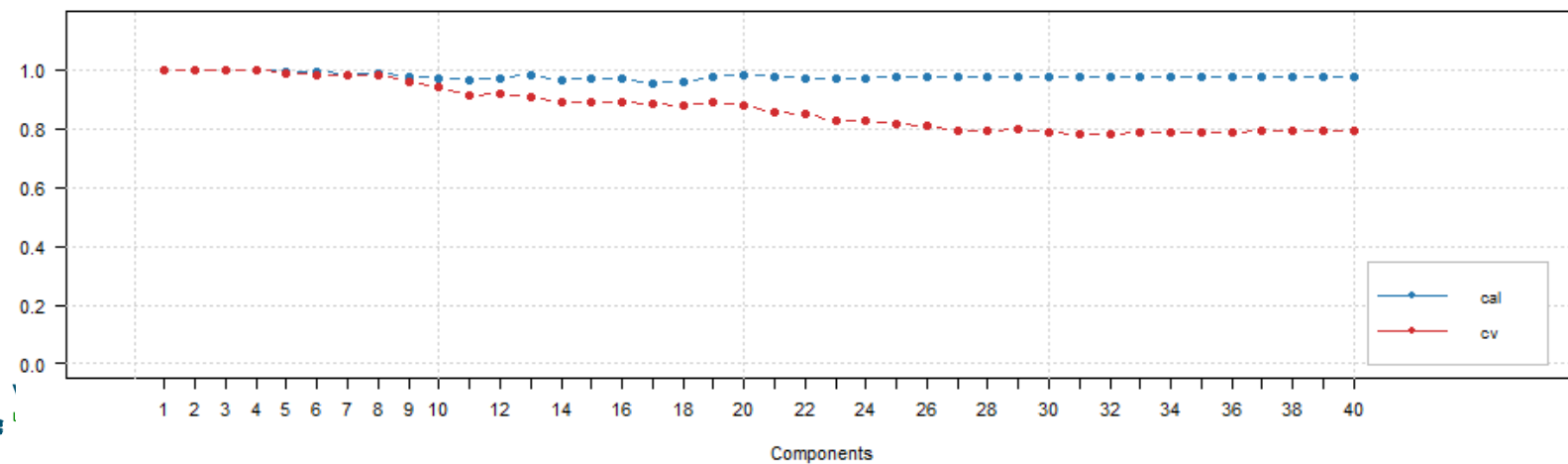


$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

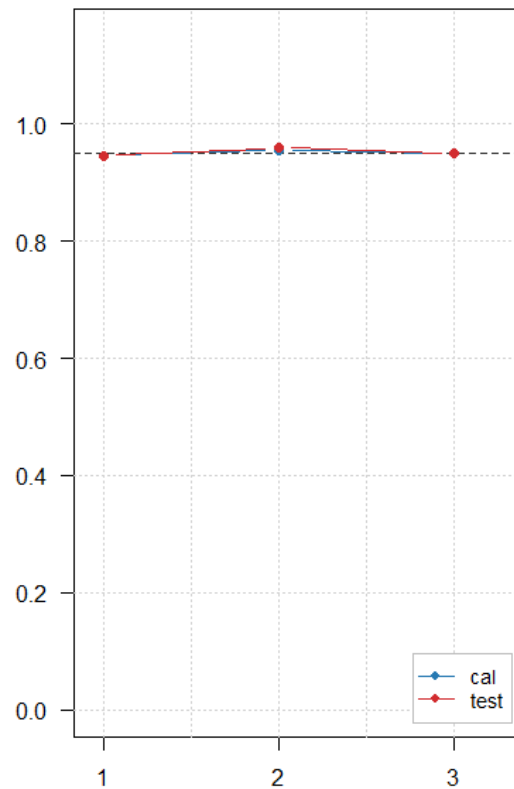
Specificity



Sensitivity

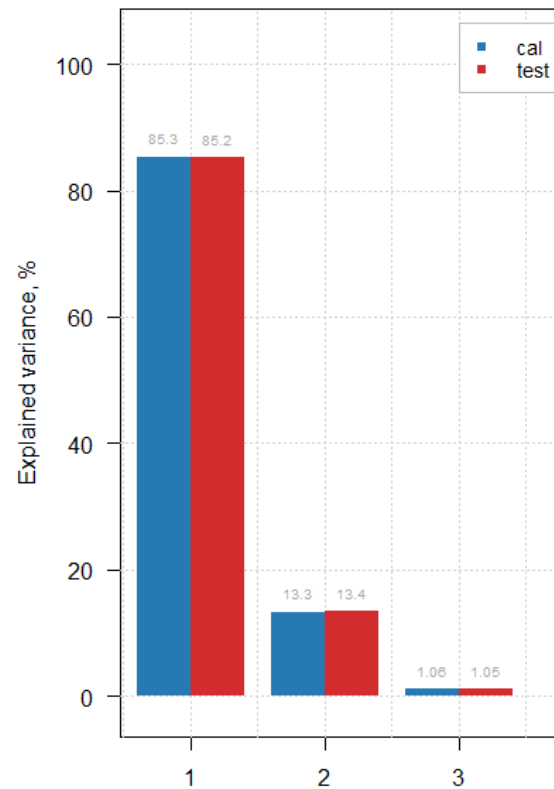


Sensitivity



Components

Variance



Components

Materials and methods



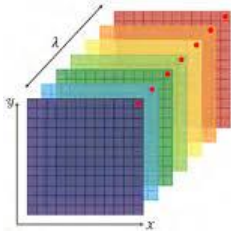
Data description

Tomatoes origin

- The samples belong to 3 cultivars – Briosso, Cappricia and Provine
- They all come from different greenhouses
- Tomatoes were harvested fresh on the 9th and the 10th of May.
- They arrived at Phenomea Laboratory, in Wageningen on Tuesday 10th May



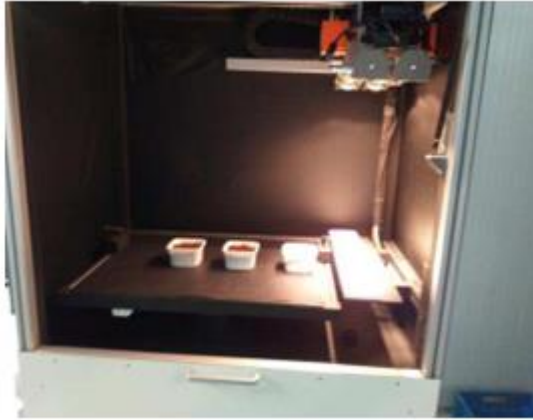
Data set



- NIR dataset: hyperspectral images (SPECIM FX17 camera) of the tomatoes from the beginning of the experiment
- Reference values from the last day of the experiment, when the fungi have infected most of the sepals

Experiment flow

1



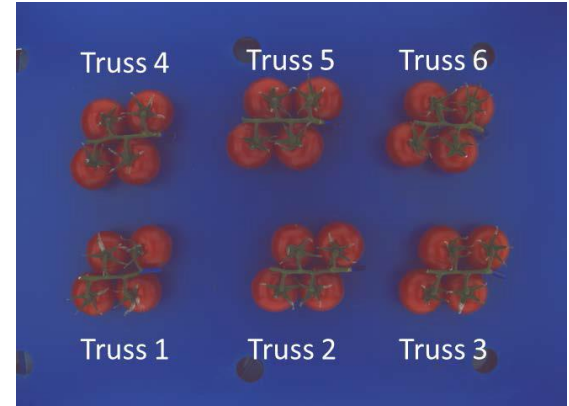
Hyperspectral imaging
of tomato trusses
using spectral cabinet*

2



Tomatoes in Hotbox -
environment for controlled
humidity and temperature
Relative humidity 100%

3



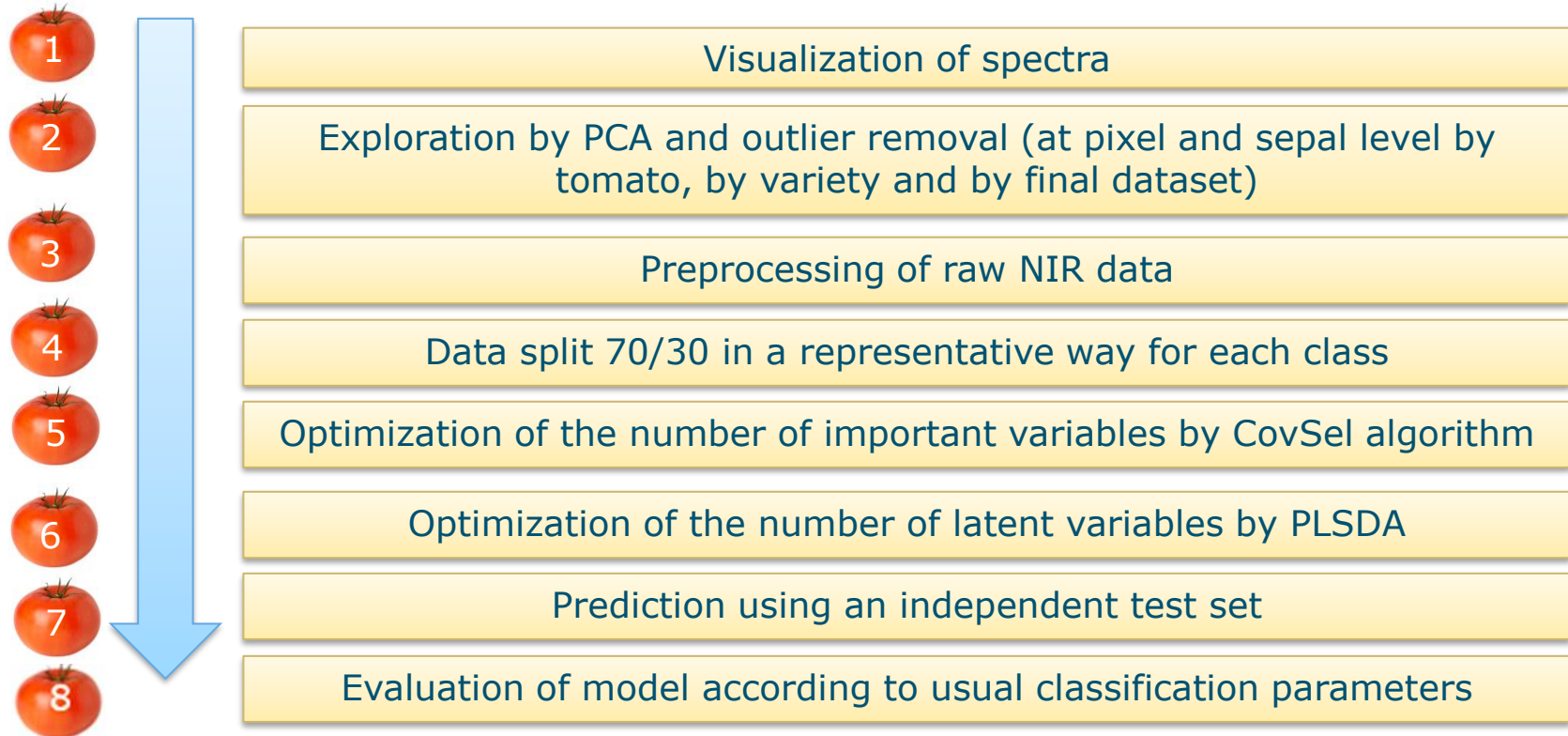
Taking phone images in
controlled environment

General analysis steps

- ❖ Ground truth observations were made by experts on day three and four, comprised of severity scores from zero (no fungus) to four (severe infection).



- ❖ Ratings of the two days, and 3 experts were averaged
- ❖ NIR spectra of sepals extracted from Hyperspectral images
- ❖ Exploration of sepals at a pixel level by Principal Component Analysis
- ❖ Removal of outlier pixels
- ❖ Creation of qualitative models by chemometrics analysis using R studio version 4.2.1 (2022-06-23 ucrt)



Classification design scenarios

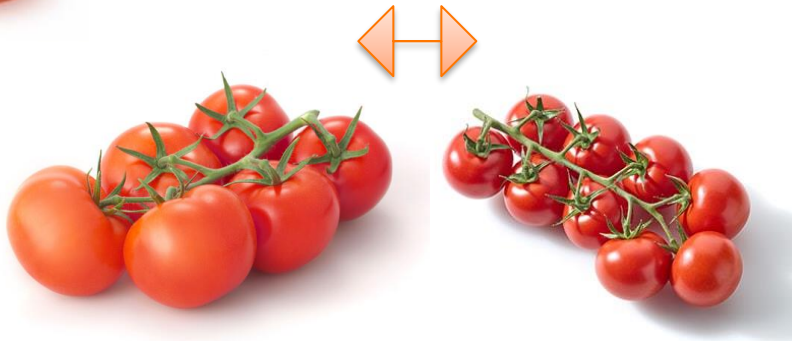
Intravariety



Global model



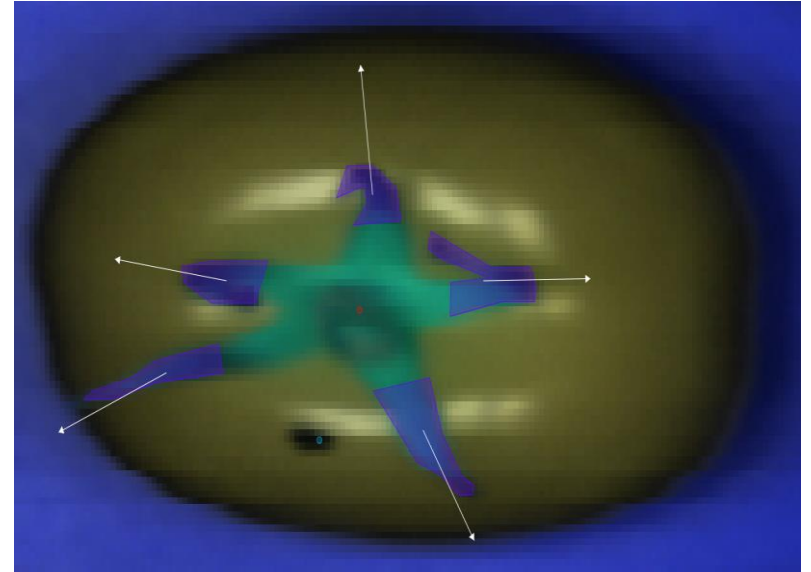
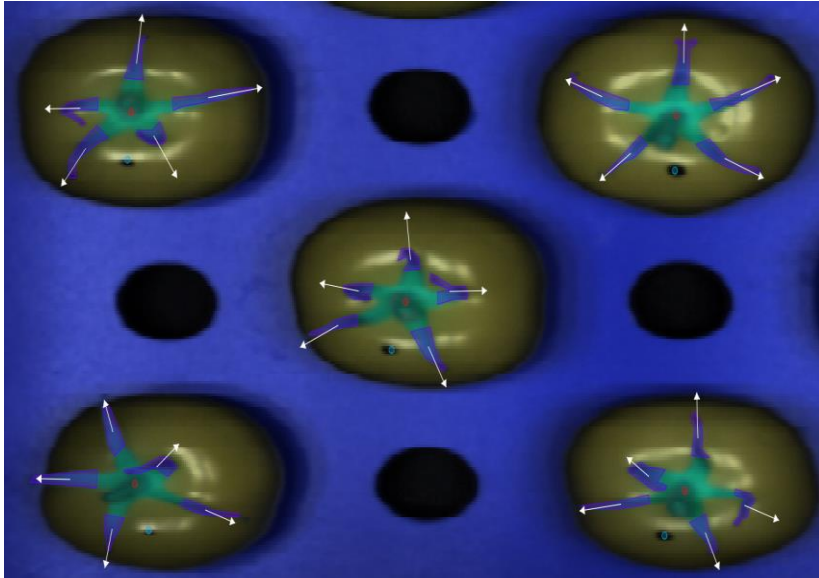
Intervariety



Results



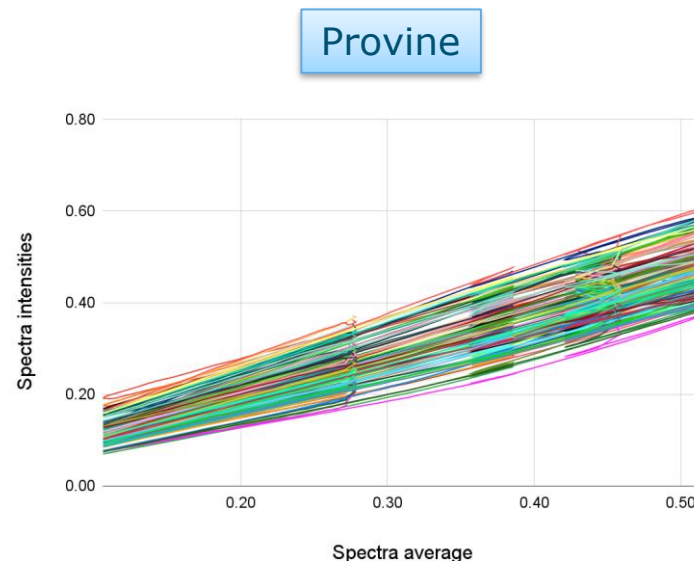
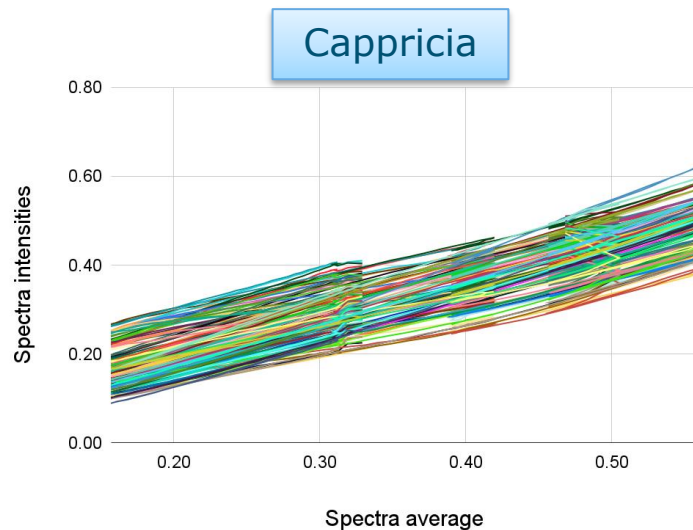
Segmentation and extraction of NIR spectra



Villiers, Hendrik de
<hendrik.devilliers@wur.nl>



Part 3: Spectra pretreatments



$$A(\lambda) = k\varepsilon(\lambda)LC + Af(\lambda) + Ab(\lambda)$$

Where, A : Absorbance; λ : Wavelength; $A(\lambda)$: Absorbance at a particular wavelength k : multiplicative effect; ε : Molar absorption coefficient ($M^{-1} cm^{-1}$); L : Optical path length (cm) C : Molar concentration; $Af(\lambda)$. Additive effect at a particular wavelength caused by a certain number of photons escaping from the sensor; $Ab(\lambda)$. Additive effect at a particular wavelength caused by random phenomena



Part 4: Sampling

Labelling
scenarios:



1: 0/123, 2: 01/23, 3: 0.5/123

Variety	n	Healthy (Label 1)	Diseased (Label 1)	Healthy (Label 2)	Diseased (Label 2)	Healthy (Label 3)	Diseased (Label 3)
Cappriccia	163	139	24	85	78	117	46
Brioso	153	145	8	78	75	126	27
Provine	152	137	15	72	80	129	23

Data split was carried out in a representative way for each class 70/30, at a sepal level
All the sepals that belong to the same tomatoes were kept together



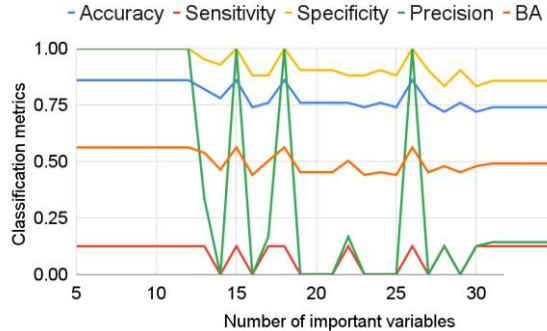
Part 5: Feature selection Most important variables for each variety

Label 2

Cappricia	937.3 3	944.25	951.1 6	965	1062.1	1083	1208.6	1320.9	1391.4	1426.7	1462.1	1561.4	1654.1	1704.1	1711.3
Provine	937.3 3	944.25	951.1 6	978.85	1089.9	1138.7	1201.6	1299.8	1370.2	1391.4	1440.8	1554.3	1654.1	1697	1718.4
Brioso	937.3 3	944.25	951.1 6	1089.9	1299.8	1391.4	1462.1	1589.9	1625.5	1661.2	1689.8	1697	1704.1	1711.3	1718.4
Global Model	937.3 3	944.25	958.0 8	1055.2	1089.9	1208.6	1278.7	1363.2	1405.5	1462.1	1589.9	1654.1	1704.1	1711.3	1718.4

Intravariety: Cappricia classification metrics according to different number of important variables as input for PLSDA

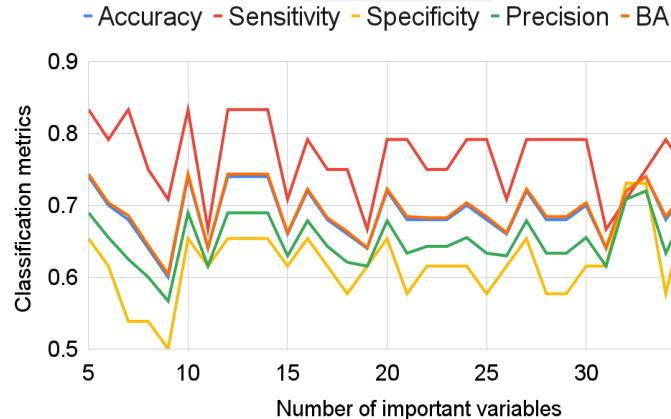
Label 1



$$ACC = (TP + TN) / (TP + TN + FP + FN)$$

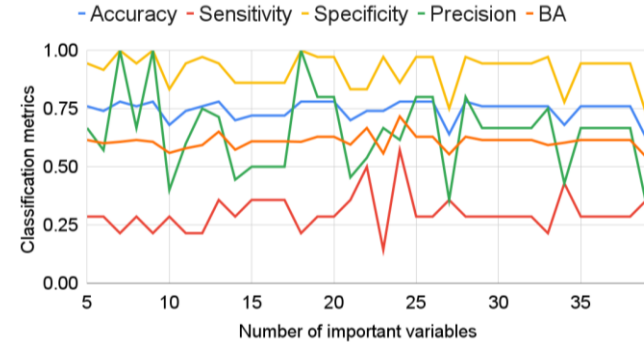
$$SEN = TP / TP + FN$$

Label 2



$$PRE = TP / TP + FP$$

Label 3

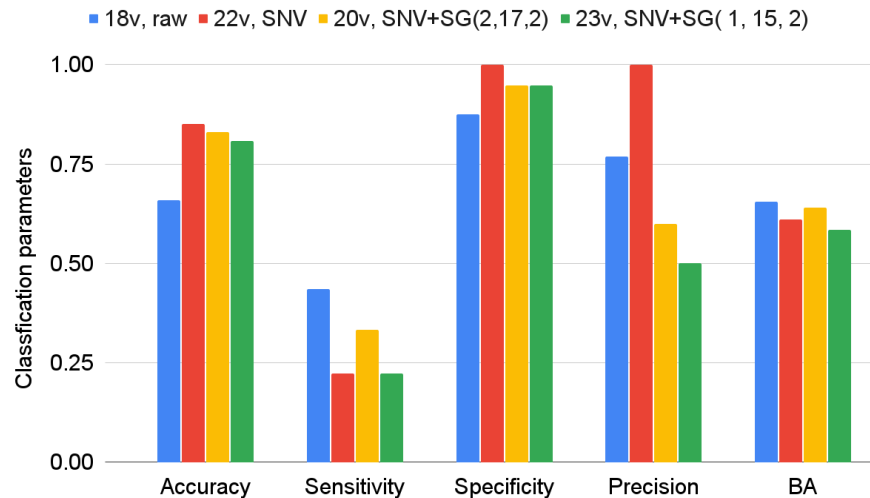
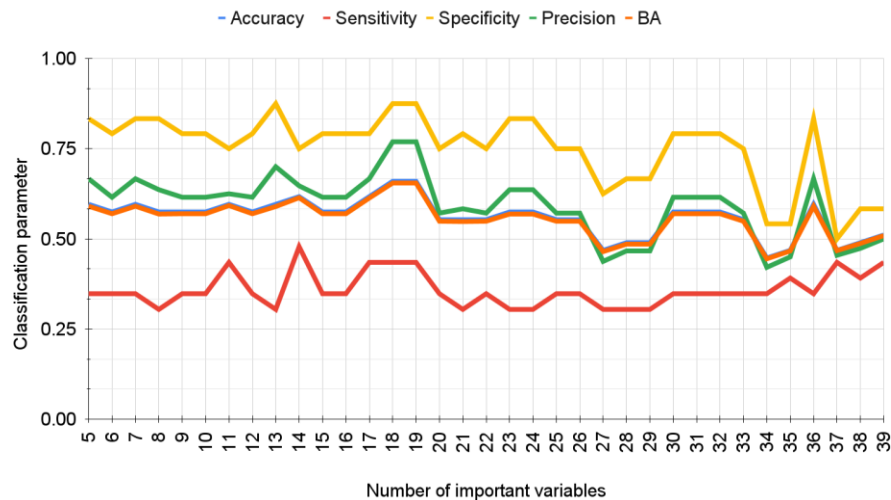


$$SPE = TN / TN + FP$$

Classification metrics according to different number of important variables as input for PLSDA

Brioso

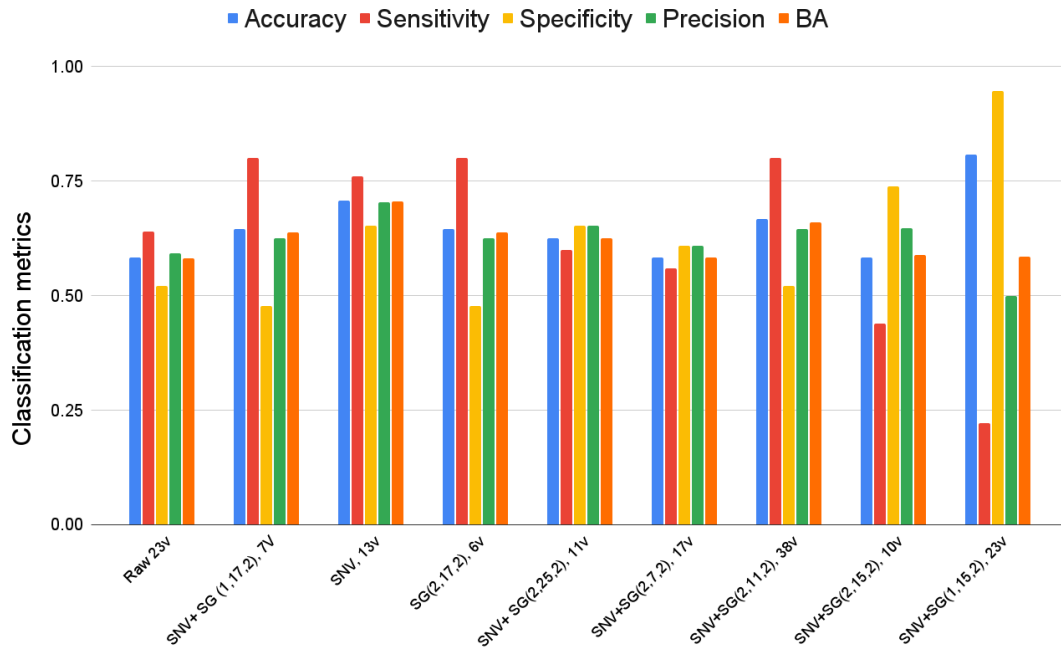
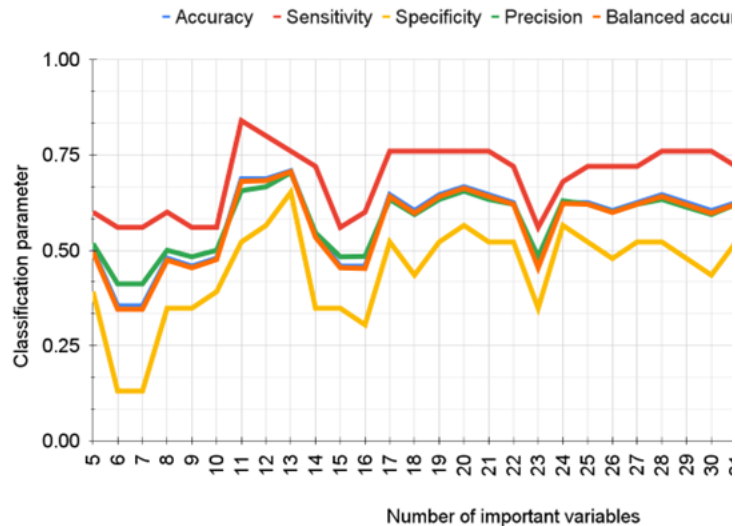
Raw



Classification metrics according to different number of important variables as input for PLSDA

Provine

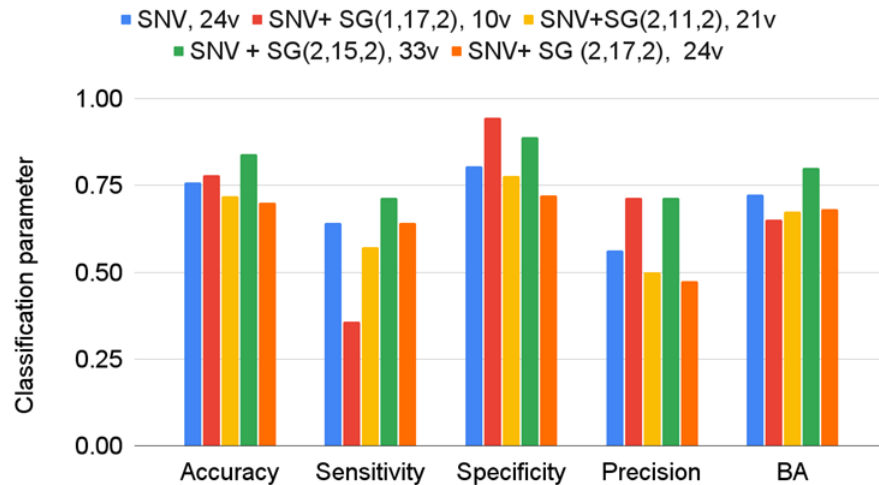
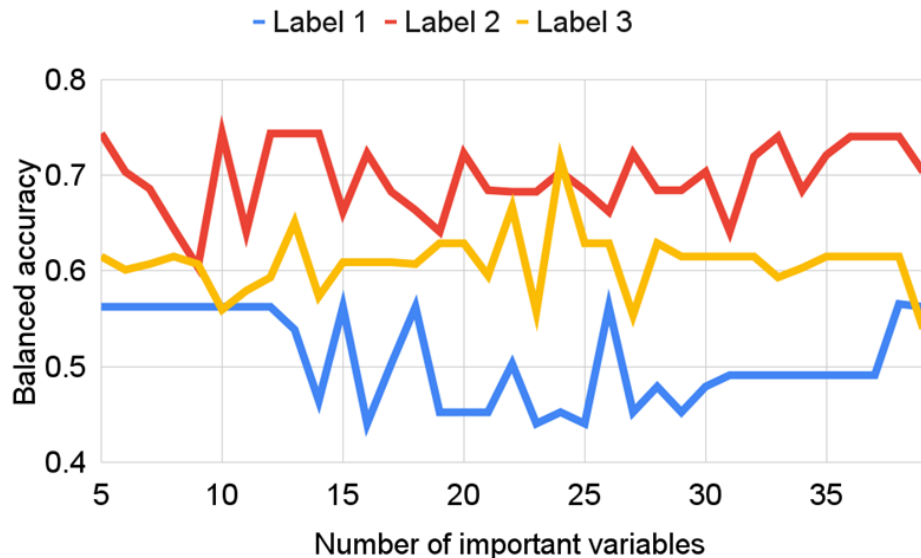
SNV



Classification metrics according to different number of important variables as input for PLSDA

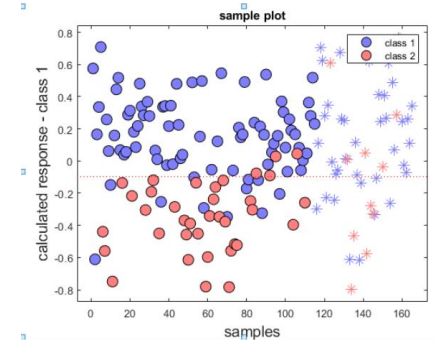
Cappricia

Raw



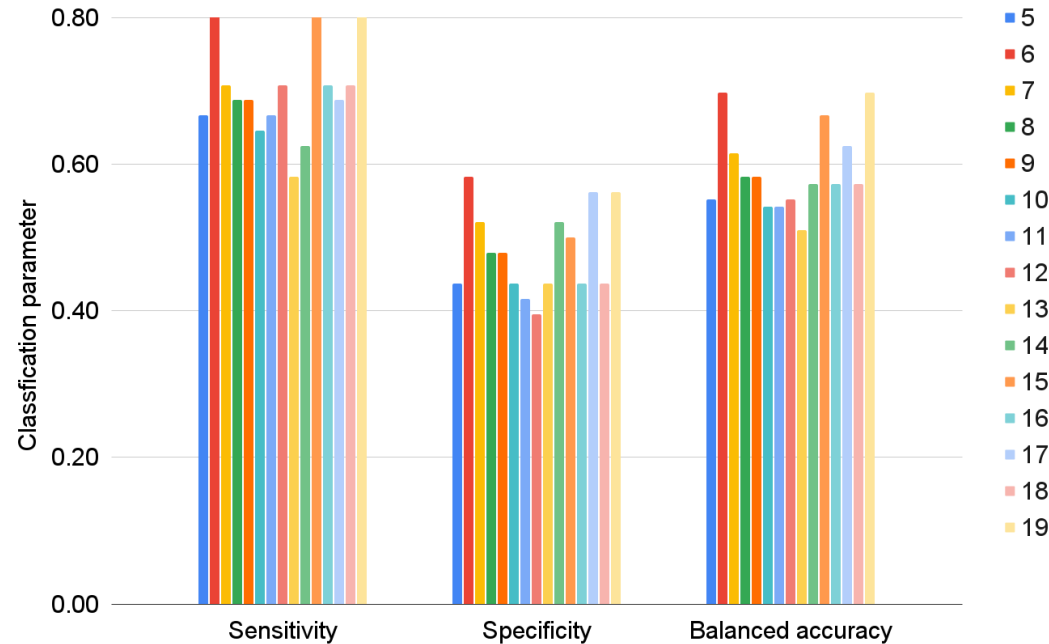
Intervariety Label 2 SNV + SG (2,17,2) LV=3

Cal: Cappricia Val1: Brioso Val2: Provine

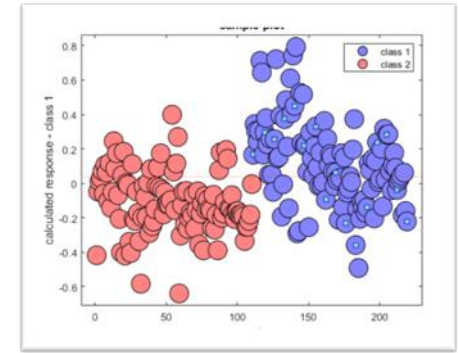
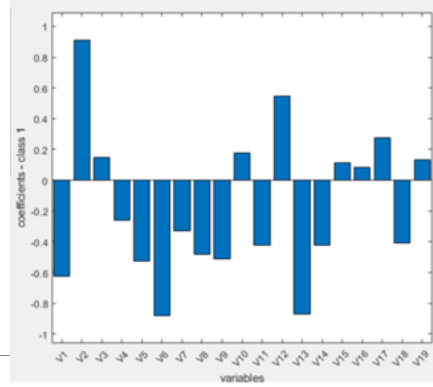
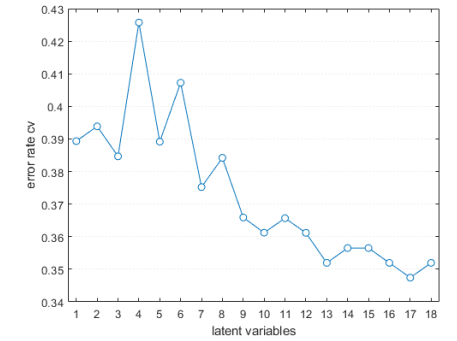


Data set	Real/ predicted	Healthy	Diseased	N A	SEN	SPE	PRE	AC	BA
Calibration Cappricia	Healthy	66	11	0	0.86	0.55	0.63	0.70	0.71
	Diseased	38	46	0	0.55	0.86	0.81	0.70	0.71
Validation Brioso	Healthy	67	7	0	0.91	0.22	0.54	0.56	0.57
	Diseased	57	16	0	0.22	0.91	0.70	0.56	0.57
Validation Provine	Healthy	80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

Classification parameters for the global model "Cap&Pro" $SNV + SG(2,17,2)$ according to the number of important variables as input for PLSDA



Global Model

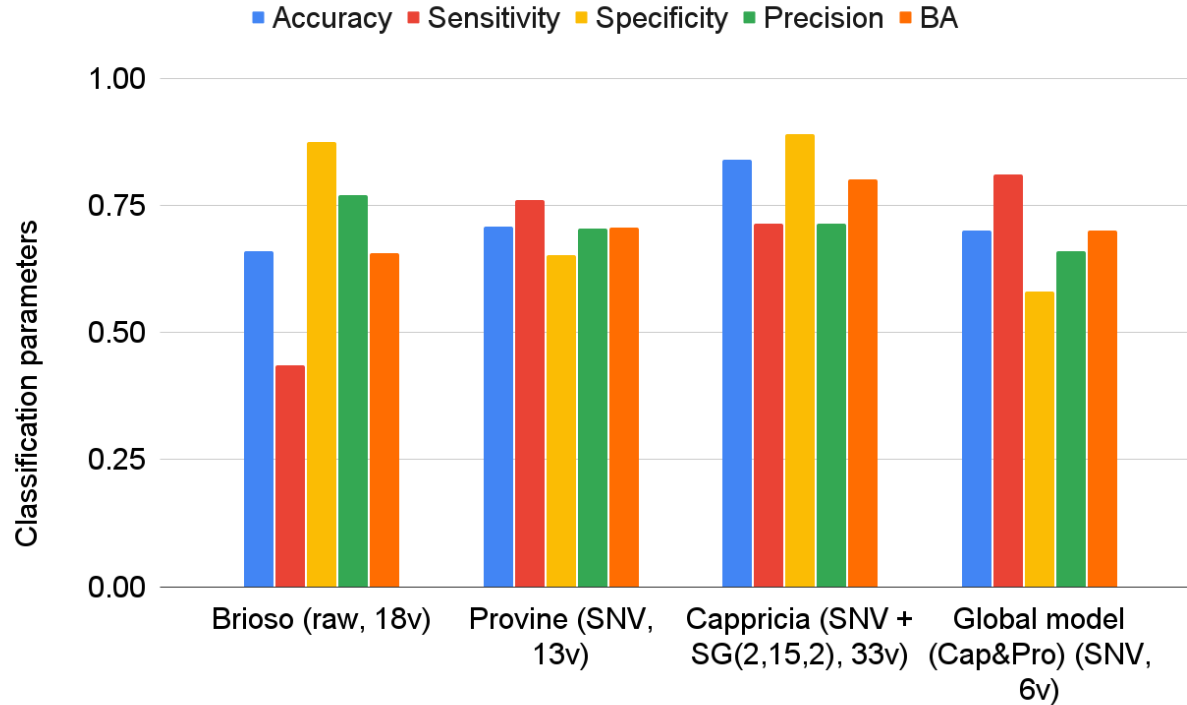


Data set	Real/ predicted	Heal thy	Disea sed	NA	Sensitiv ity	Specifi city	Precis ion	BA
Calibrati on	Health y	62	47	0	0.57	0.74	0.68	0.66
	Diseas ed	29	81	0	0.74	0.57	0.63	0.66
Validati on	Health y	23	25	0	0.48	0.71	0.62	0.60
	Diseas ed	14	34	0	0.71	0.48	0.58	0.60

Results- Summary

Parameter/ Model	Cappricia	Cappricia	Provine	Provine	Brioso	Intervariety	Global model
	Raw, 15v	SNV + SG(2,15,2), 33v	Raw, 14v	SNV, 13v	Raw, 18v	Cal: Cap, Val: Bri	SNV, (Cap+Pro)
	Label 3	Label 2	Label 3	Label 2	Label 2	Raw, 10v	6v
						Label 3	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

Classification metrics of the optimized models



Results

- Out of all the sepals that did not have the disease, how many got negative test results?

High proportion of actually healthy sepals that were correctly predicted out of all positive predictions made by the models

- ❖ Cappricia: 0.71
- ❖ Provine: 0.76
- ❖ Global model: 0.81

Results

- Out of all the sepals that do have the disease, how many got positive results?

The ability of a test to correctly classify an individual sepal as diseased

- ❖ Cappricia: 0.89
- ❖ Provine: 0.65
- ❖ Global model: 0.58

Results

- A good performance on positive and negative classes respectively was found in Capparicia 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
- BA reduced to the traditional accuracy (0.70) in the global model showing that the classifier performed equally well on either classes
- The intervariety models calibrated in Capparicia 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)

Conclusions



Conclusions

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

Conclusions

- Novelty of this work - investigate HSI to capture the sepal susceptibility of fungal infection by chemometric analysis of different varieties of tomatoes
- The results from this research reaches to a strong conclusion that discrimination between more susceptible and less susceptible samples is feasible under controlled conditions
- Unanswered questions:
 - What is the information that the spectra captures in these samples to help discriminate the susceptibility to fungus?
 - How will the global model perform when predicting samples from another harvesting time?
- These questions are the likely to set the directions for new investigations on this subject

Thank you for your kind attention



Antares



Horizon 2020
Programme

References

- Brdar, S., Panić, M., Hogeveen-van Echtelt, E. *et al.* Predicting sensitivity of recently harvested tomatoes and tomato sepals to future fungal infections. *Sci Rep* **11**, 23109 (2021). <https://doi.org/10.1038/s41598-021-02302-2>
- Paper 942-2017 Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data Josephine S Akosa, Oklahoma State University <https://support.sas.com/resources/papers/proceedings17/0942-2017.pdf>
- Roger, J.-M.; Mallet, A.; Marini, F. Preprocessing NIR Spectra for Aquaphotomics. *Molecules* **2022**, 27, 6795. <https://doi.org/10.3390/molecules27206795>