Received: 7 September 2015 ■ *Revised: 31 March 2016* ■ *Accepted: 1 April 2016* ■ *Publication: 10 April 2016*

JSI JOURNAL OF SPECTRAL IMAGING

openaccess

Classification of different tomato seed cultivars by multispectral visible-near infrared spectroscopy and chemometrics

Santosh Shrestha, Lise Christina Deleuran and René Gislum*

Department of Agroecology, Faculty of Science and Technology, Aarhus University, Slagelse, 4200, Denmark. E-mail: rg@agro.au.dk

The feasibility of rapid and non-destructive classification of five different tomato seed cultivars was investigated by using visible and short-wave near infrared (Vis-NIR) spectra combined with chemometric approaches. Vis-NIR spectra containing 19 different wavelengths ranging from 375 nm to 970 nm were extracted from multispectral images of tomato seeds. Principal component analysis (PCA) was used for data exploration, while partial least squares discriminant analysis (PLS-DA) and support vector machine discriminant analysis (SVM-DA) were used to classify the five different tomato cultivars. The results showed very good classification accuracy for two independent test sets ranging from 94% to 100% for all tomato cultivars irrespective of chemometric methods. The overall classification error rates were 3.2% and 0.4% for the PLS-DA and SVM-DA calibration models, respectively. The results indicate that Vis-NIR spectra have the potential to be used for non-destructive discrimination of tomato seed cultivars with an opportunity to integrate them into plant genetic resource management, plant variety protection or registration programmes.

Keywords: tomato, varietal classification, seed, chemometric methods, PLS-DA, SVM

Introduction

Tomato (*Solanum lycopersicum* L.) is one of most economically important horticultural crops worldwide. It is well known for its healthy nutrients, vitamin C and phytochemicals like lycopene and β-carotenes. Tomato is further associated with its preventive role in prostate cancer risk development.[1] Its importance in human consumption can also be estimated from the increased demand documented by FAOSTAT[2] with a more than two-fold production increase (from 77.9 to 163.9 million metric tonnes) in the last two decades (1993–2013 AD). This is accompanied by intensive breeding efforts to develop new cultivars to meet global demand.[3] As a result, several diverse tomato cultivars are released every year for commercial cultivation throughout the world. However, to release a new cultivar, compliance with DUS (distinctness, uniformity and stability) has to be ensured, which serves as a measure to protect the plant breeders' rights.[4] A descriptive characterisation of the plant cultivar is required for easy identification.[4] All the member countries of the WTO have an obligation to TRIPS (trade-related aspects of intellectual property rights) to provide the minimum intellectual property rights (IPR) through a version of plant variety protection (PVP) or patents.[4] Nepal has also drafted a bill on "Plant variety protection and Farmers' Rights", a *sui generis* of TRIPS which is under consideration in the parliament and has emphasised plant breeders' rights.[5]

The most reliable method of cultivar identification is by their morphological features, as these are distinct and stable for years.[6] However, morphological traits are very limited, and the narrow genetic diversity of modern tomato cultivars makes it more challenging to detect any new innovations like disease resistance, taste or other traits which have very few phenotypic variations.[7] Molecular markers and other biochemical methods are usually used for the identification and characteri-

ISSN: 2040-4565
doi: 10.1255/jsi.2016.a1

sation of tomato germplasm.[8,9] However, they are expensive, require experienced users and, more importantly, are destructive. Further, identification of cultivars by these methods require seeds to be planted, which is time consuming since germination and development of the plant for at least a month are necessary before it can be identified by morphological traits or any other method.[10] So, there is a need for technologies which are robust, quick, non-destructive and reliable for the identification of different tomato cultivars.

Our previous study has shown that multispectral imaging could be one of the alternatives to identify the tomato cultivars non-destructively using seeds.[5] We showed the application of normalised canonical discriminant analysis (nCDA)[11] to classify tomato cultivars by using different seed features consisting of shape and seed colour attributes.[5] However, use of nCDA analysis is limited to the instrument's (VideometerLab, Videometer A/S, Hørsholm, Denmark) built-in software which was used to capture the seed image.[5] Moreover, spectral information consisting of visible and short-wave near infrared (Vis-NIR) spectra can be extracted from the same seed image and analysed in different chemometric platforms. The Vis-NIR region contains information regarding colour attributes (visible) and chemical properties (NIR) which can be interpreted with the help of different chemometric methods by their ability to discriminate samples belonging to one or several distinct groups based on spectral properties.[12] Earlier work on Vis-NIR spectra has successfully demonstrated that they could be used for classification of tomato cultivars using the leaves[6] and tomato fruits.[13,14] They have also shown potential for identifying different cultivars of Chinese bayberry,[15] Chinese cabbage seeds[16] and rice seeds.[12] However, there is no report on the use of Vis-NIR spectra for cultivar identification of tomato using seeds. Therefore, we aimed to investigate its potential for use in classification of tomato cultivars from seeds using two different chemometric methods, PLS-DA (partial least squares discriminant analysis)[17,18] and SVM-DA (support vector machine discriminant analysis).[19] PLS-DA is a linear classification method,[17] whereas SVM-DA is well known for its non-linear classification properties.[20,21] Hence, the study also aims to compare the outcome of these two chemometric methods for classification of tomato seed cultivars.

Materials and methods

Tomato seed samples

Five tomato cultivars or accessions, viz. BL410, CL, Care Nepal, HRD17 and T9, collected from Nepal were grown in semi-field conditions in 2014 at Flakkebjerg, Denmark. Tomatoes were harvested at the red ripe stage and seeds were extracted by a natural fermentation process (pulp with seeds were collected and left overnight and later washed to extract seeds) for each cultivar. Extracted seeds were dried at room temperature for two days and subsequently fan dried for three days. The seeds were stored at 6°C until further use. A total of 1236 seeds were used for the study. The seed lots of all cultivars were subjected to a quartering sampling procedure to obtain a subsample which contained at least 200 seeds for each tomato cultivar (Table 1).

Spectral imaging and acquisition of Vis-NIR spectra

Spectral images (Figure 1) from each seed sample were captured using a VideometerLab instrument (Videometer A/S,

Table 1. Details of samples used for classification of tomato cultivars. Seed samples for test set one were randomly selected, whereas seed samples for test set two were selected by automatic data split using the Onion algorithm of PLS Toolbox ver. 7.9.

Cultivars	Calibration set	Test set 1	Test set 2	Total seeds
BL410	160	50	16	226
CL	120	96	14	230
Care Nepal	198	66	27	291
HRD17	170	91	22	283
T9	152	37	17	206
Total	800	340	96	1236

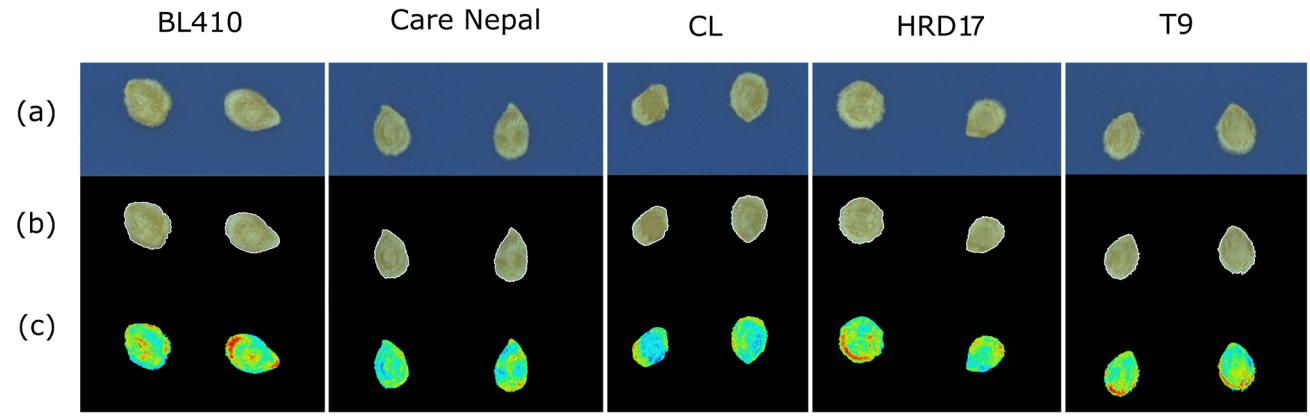

Figure 1. The captured multispectral images of five tomato seeds cultivars. (a) The images after blue background segmentation; the white margin on the seeds shows the selection of the ROI (b) seed images at 525 nm (c).

Hørsholm, Denmark). This instrument acquires multispectral images in 19 Vis-NIR wavelengths: 375, 405, 435, 450, 470, 505, 525, 570, 590, 630, 645, 660, 700, 780, 850, 870, 890, 940 and 970 nm. The wavelengths from 375 nm to 700 nm are from the visible range, and the wavelengths from 780 nm to 970 nm are in the NIR region. The instrument consists of a five-megapixel CCD camera, mounted inside the top of the integrating sphere, coated with matte titanium paint, with illumination from 19 light emitting diodes (LEDs) placed along the rim of the sphere. The instrument was calibrated to absolute reflectance using a bright and dark reference object (NIST traceable targets) and geometrically aligned using the dotted plate before capturing the seed images.[22] The seeds were placed at the bottom of the integrating sphere on a "blue circular disc" and a high resolution multispectral image of 2056 × 2056 pixels was captured.

The captured image contains the information from the seeds, which are the region of interest (ROI) (Figure 1b), and the background (on which the seeds were placed to capture the image, in this case a "blue circular disc") is noise and irrelevant to the analysis. So, a default "blue background mask" of the VideometerLab software was applied to only obtain images of the seeds. Mean reflectance spectra were calculated from each image by averaging the intensity of pixels within the ROIs at each wavelength. The resulting data consisted of 19 mean values of the reflectance from the seed which were later used for the classification of tomato cultivars.

Spectral pre-processing and sample set partition

The Vis-NIR spectra were preprocessed using SNV[23] and detrended[23] before mean centring. Principal component analysis (PCA)[24] was used on preprocessed Vis-NIR spectra to examine the grouping of the tomato cultivars and possible outlier detection. The whole Vis-NIR data set was divided into three sets consisting of one calibration set and two test sets. The first test set consisted of 340 seeds, which were randomly selected from the data. The second test set consisted of 96 seeds and was obtained using the automatic data split Onion algorithm of PLS Toolbox version 7.9, and the remaining 800 seeds were used to develop a calibration model. The details of the sample set partition are shown in Table 1.

Partial least squares discriminant analysis (PLS-DA)

Partial least squares discriminant analysis, a linear classification method,[18] is a derivative of the standard PLS regression algorithm[17] which uses class variables instead of numeric variables. PLS1 and PLS2 algorithms are commonly used based on the number of classes; for two-class problems, the former is used, while the latter is used when there are more than two classes of samples. In PLS, the dummy variable *Y* is used as a response variable, and it is set to 1 if the sample is one of either class and 0 if not. For instance, in our work comprising five classes, each sample is coded as one of the following five vectors: [1 0 0 0 0], [0 1 0 0 0], [0 0 1 0 0], [0 0 0 1 0], [0 0 0 0 1] designating the classes 1, 2, 3, 4 and 5, respectively. The model seldom predicts either 1 or 0 perfectly, so a cut-off value was set at 0.5, above which the sample is predicted as 1 and below which it is predicted as 0. In this study, the optimal number of latent variables (LVs) was chosen on the basis of minimal classification error for calibration and cross-validation of the model. The model was cross-validated by Venetian blinds of 10 data splits with 10 samples in each split. Further information on PLS-DA can be found in the work of Barker and Rayens[25] and Ballabio and Consonni.[18]

Support vector machine discriminant analysis (SVM-DA)

Support vector machine (SVM) is a robust machine learning algorithm developed by Cortes and Vapnik,[19] and it is based on a structural risk minimisation (SRM) strategy which reduces the risk of overfitting the data.[26] SVM constructs a hyperplane as a decision line, which separates the classes with the largest distance from the nearest training data points. The samples used for defining the boundary of classes are termed as support vectors (SVs) and are the only ones used for the model development. SVM maps the dataset of *"n"* observation and *"k"* variables (in this case Vis-NIR) into a higher dimensional feature space by use of a kernel function. Radial basis function (RBF), a kernel function which is used for non-linear problems, was used in the study to reduce computational complexity of the training procedure and to obtain good prediction results. Two parameters, γ (RBF kernel width) and c (SVM cost factor) of RBF are needed to be tuned *a priori*. The former is used as a regulation constant affecting the generalisation performance of SVM models and the latter is the cost factor which controls the trade-off between training errors and model complexity of SVM models. The search limits of γ and c were set to the default LIBSVM algorithm of PLS Toolbox version 7.9, which were 10^{-6} to 10 with 15 values spaced uniformly and from 10^{-3} to 100 with 11 values spaced uniformly, respectively. These two parameters were decided based on the minimal classification error through a two-dimensional grid search coupled with cross-validation by Venetian blinds of 10 data splits with 10 samples in each split.

The above mentioned chemometric methods, viz. PCA, PLS-DA and SVM-DA, were performed using MATLAB version 8.1.0.604 (R2013a) (The Math Works, Inc., Natick, MA, USA) along with the PLS Toolbox 7.9 (Eigenvector Research, Inc., WA, USA).

Model evaluation measures

To evaluate the performance of the classification models, the classification error rate (*ER*) for each cultivar, overall classification *ER* (*OER*), sensitivity (*Sn*), specificity (*Sp*) and accuracy were calculated as per Ballabio and Consonni.[18] The equations used for the calculations are as given below:

Sensitivity (*Sn*),

$$Sn = \frac{TP}{TP + FN} \quad (1)$$

where *TP* is true positive samples, *FN* is false negative samples.

Specificity (*Sp*),

$$Sp = \frac{TN}{FP + TN} \quad (2)$$

where *TN* is the true negative and *FP* is false positive.

Classification error rate (*ER*),

$$ER = 1 - \frac{Sn + Sp}{2} \quad (3)$$

Overall classification error rate,

$$OER = \frac{\sum_{i=1}^{n} ER}{n} \quad (4)$$

where *n* is the number of classes (cultivars in our study) and *ER* is classification error rate.

Classification accuracy,

$$Accuracy = \frac{Correctly\,classified\,samples}{Total\,samples} \times 100\% \quad (5)$$

Results and discussion

Exploratory analysis

Vis-NIR spectral data from seeds of the five tomato cultivars showed variations, but exhibited similar trends of reflectance in each wavelength (Figure 2). The variations in the spectra indicate the differences among tomato cultivars with regard to physical and chemical properties of tomato seeds. The variations in the visible range can be attributed to colour of the seed samples, whereas the variations in the NIR region are due to chemical differences in seeds of the cultivars.[12,22] These spectral variations indicated that Vis-NIR can be exploited for qualitative classification using chemometric methods.

PCA was initially performed on the Vis-NIR spectra without any data pre-treatment to explore the possible clustering of the tomato cultivars and to identify possible outliers. However, distinct discriminations among the tomato cultivars were not observed (data not shown). This is not a surprising observation as the spectral properties of the seeds might have influence on the physical phenomena like light scattering, particle size distribution and alignment with the incident beam of light which add noise to the data.[27] Therefore, mathematical data pre-processing algorithms SNV and detrend were used to eliminate or minimise the physical effects for further data analysis.[23] The PCA performed on the preprocessed Vis-NIR spectra revealed few outliers (data not shown) in the calibration set. However, removing the outliers did not improve the model and they were subsequently retained for further development of the classification models. Figure 3 shows the three-dimensional principal component (PC) score plot using the first three score vectors, PC 1, PC 2 and PC 3, which contributed most of the spectral variations of 96.5%, i.e. 47.8%, 42.6% and 6.1%, respectively. It showed the clustering of seed samples of the same cultivars, though some overlaps between the clusters of cultivars were observed. The results indicate that discrimination between the five tomato cultivars is possible based on reflectance from seeds. Further, it signifies that different spectral attributes from samples can be associated with characteristics of the seed from each cultivar.

PLS-DA model

The PLS-DA model was developed using six LVs to classify the tomato cultivars. The developed PLS-DA model explained 99.7% of the variation of the Vis-NIR spectra, out of which

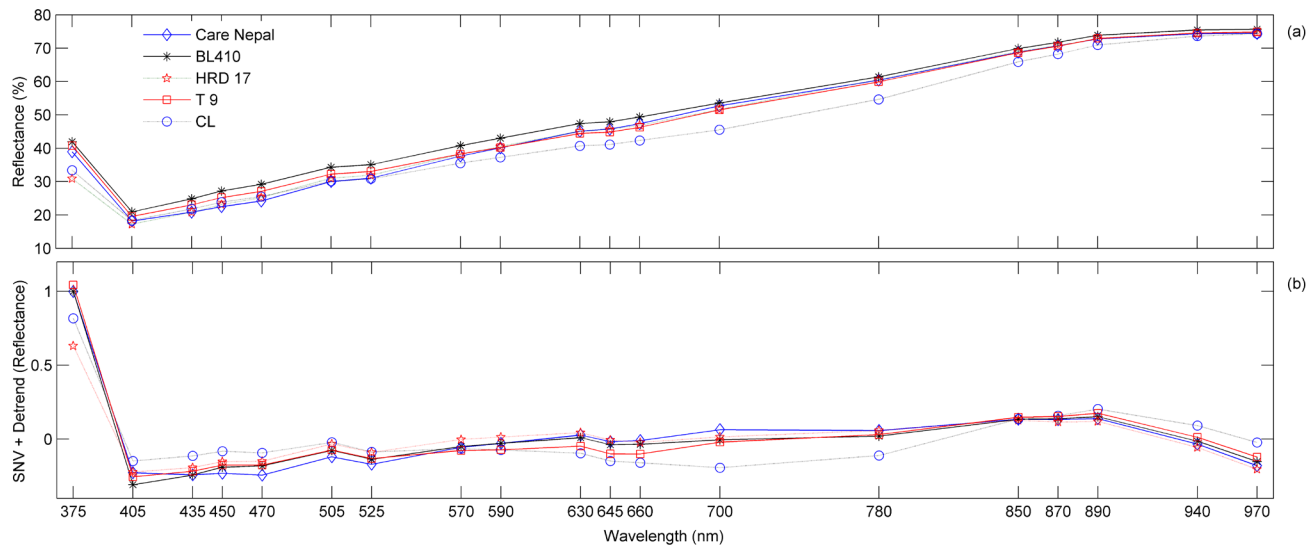

Figure 2. The mean Vis-NIR spectrum of the five tomato cultivars extracted from the ROI of the seed images in 19 wavelengths. The wavelengths from 375 nm to 700 nm are from the visible range and wavelengths from 780 nm to 970 nm are from the NIR region (a) averaged SNV and detrend pre-processed Vis-NIR spectra of tomato seeds (b).

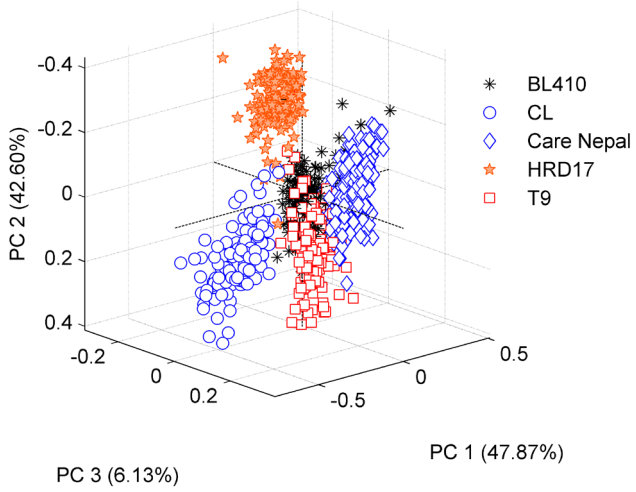

Figure 3. First three principal component (PC) scores plot shows the clustering of five tomato cultivars towards their group membership. The values in the parentheses indicate the information for variation contained in the respective PCs.

96.4% variation information came from the first three LVs. The model was able to classify all the cultivars of the calibration set with an overall classification *ER* of 3.2%, with the least *ER* for HRD17 and CL of 0.3% and 0.8%, respectively (Table 2). The calibration model was relatively poor in classifying cultivars BL410, Care Nepal and T9 as the misclassification rate was higher for each and contributed significantly to the overall *ER* (Table 2). This could be plausible as the clusters of these three cultivars were found to overlap in the exploratory analysis (Figure 3). However, the model was able to predict test sets of samples with a classification accuracy of 94% to 100% for both test sets (Table 3). The proportion of misclassified seeds was almost similar in the two test sets (Table 3). Figure 4 shows the classification accuracy of the PLS-DA model for test set one. Overall ERs were also consistent for test set one and test set two with 1.8% and 2.1%, respectively (Table 2). Further, the sensitivity of the model, i.e. the ability to correctly identify the positive samples belonging to the class, was reasonably higher for all cultivars with absolute classification for CL and HRD17 (Table 4). The specificity of the model, i.e. ability to reject samples of all other cultivars, was also adequately higher and very comparable to its ability to correctly classify samples, which signifies the robustness of the model. In general, PLS-DA showed the potential of Vis-NIR spectral data for classifying seeds of tomato cultivars.

Table 2. Classification error rate for each tomato cultivar and overall classification error rate for different data sets of two chemometric methods.

Chemometric method	Data set	BL410	CL	Care Nepal	HRD17	T9	Overall *ER*
PLS-DA	Calibration set	7.3%	0.8%	4.6%	0.3%	2.8%	3.2%
	Cross-validation	7.7%	0.8%	4.5%	0.6%	3.0%	3.3%
	Test set one	1.8%	0.6%	2.1%	0.0%	4.8%	1.8%
	Test set two	3.7%	1.5%	2.4%	1.1%	2.2%	2.1%
SVM-DA	Calibration set	0.3%	0.4%	0.6%	0.1%	0.5%	0.4%
	Cross-validation	1.5%	1.8%	1.5%	1.1%	1.5%	1.5%
	Test set one	1.1%	1.7%	0.2%	0.6%	2.7%	1.2%
	Test set two	0.6%	0.0%	2.5%	0.0%	2.9%	1.2%

Table 3. The number of misclassified seeds and classification accuracy results for the two independent test sets predicted by PLS-DA and SVM-DA.

Data set	Cultivars	Misclassified seeds		Accuracy (%)		Total seeds
		PLS-DA	SVM-DA	PLS-DA	SVM-DA	
Test set one	BL410	2	1	96	98	50
	CL	2	3	98	97	96
	Care Nepal	1	0	98	100	66
	HRD17	2	0	98	100	91
	T9	2	2	95	95	37
Test set two	BL410	1	0	94	100	16
	CL	0	0	100	100	14
	Care Nepal	1	1	96	96	27
	HRD17	0	0	100	100	22
	T9	1	1	94	94	17

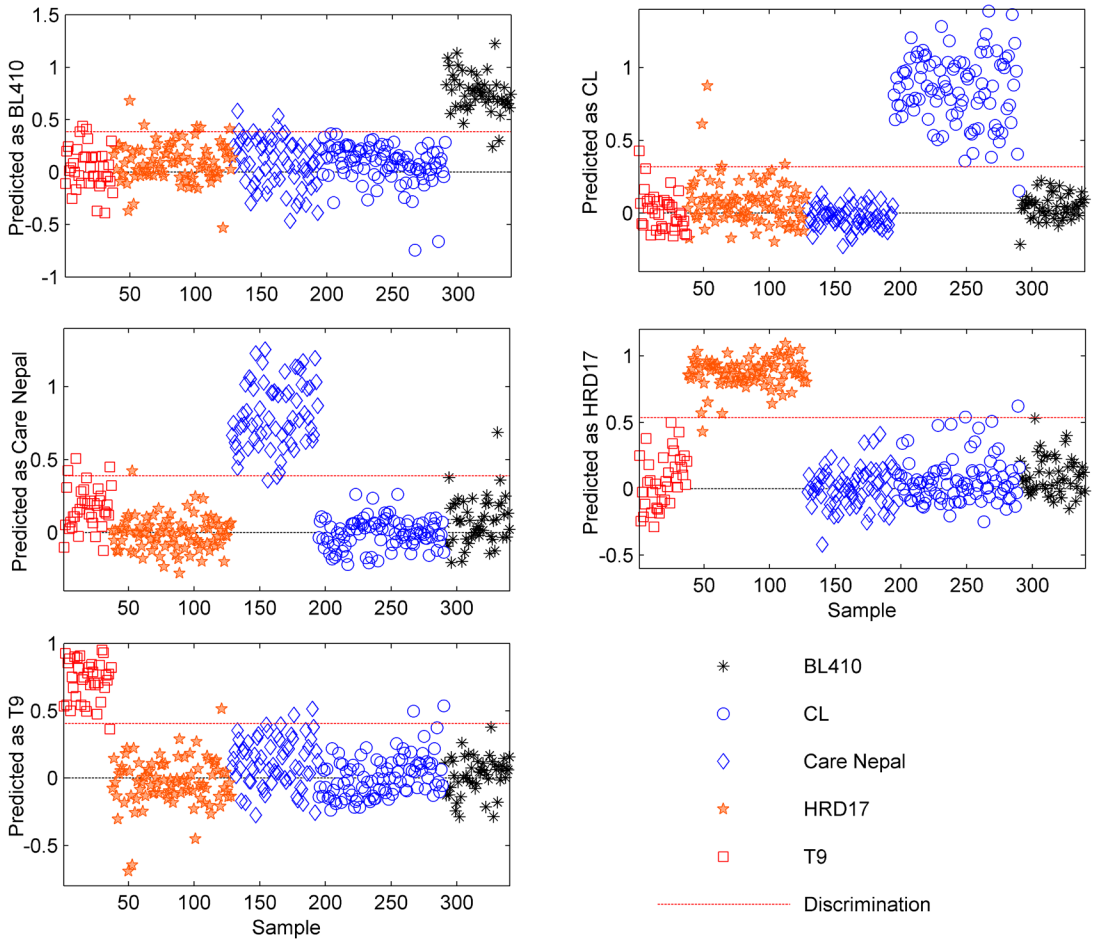

Figure 4. Prediction results of test set one for five tomato seed cultivars by PLS-DA.

SVM-DA

The SVM-DA model was developed without any data compression. SVM parameters, i.e. γ and c optimisation based on the minimum misclassification error through a grid search method, were specified by the position of "X" in Figure 5, where the values for the parameters were 31.62 and 10, respectively. The value of c determines the trade-off between the complexity of the boundary indicating the importance given to misclassified samples or samples near the boundary.[21] Usually, a lower value of c is preferred as a degree of misclassification is tolerated, i.e. it permits some samples to be ignored or placed on other side of the classifier's margin and balances the classification error against the complexity of the model.[19,21] The model contained 126 support vectors (SVs), which were used to define the decision plane for classification of the five tomato seed cultivars. This is a considerable number of SVs to be included in the development of a classification model with an average of around 16% of the total samples. These support vectors contributed in defining the class boundaries for prediction of seeds in the test sets.

The overall classification *ER* of the SVM-DA model was very low, 0.4% for calibration, and comparatively stable in prediction of the two test sets with an overall classification *ER* of 1.2% and 1.2%, respectively (Table 2). The SVM-DA model predicted the cultivars with a very low number of misclassified seeds for all cultivars (Table 3), which can be also be observed in Figure 6 which shows the prediction results of test set one. The classification accuracy of the two test sets was also very high for all the tomato cultivars with the lowest being 94% for T9 (Table 3). The capacity of the model to identify the positive samples was very similar to its ability to reject the other cultivar samples (Table 4). The values for sensitivity and specificity were more or less equal to one for all cultivars, indicating the strength of the model[18] (Table 4). The promising result from the SVM-DA analysis shows the potential of Vis-NIR spectral data for discriminating seeds of tomato cultivars.

The study demonstrates the potential of Vis-NIR spectra for discrimination of tomato seed cultivars without any sample pre-treatment. Both of the chemometric methods showed good classification accuracy with a low classification error rate. The efficiency of the models to accept the positive samples (sensitivity) and reject the negative samples (specificity) for each tomato cultivars were very high and comparable between the two methods (Table 2). Choice of selecting either of the two methods would therefore merely depend on individual preference. SVMs are often considered robust classification methods and usually produce better results even if the data are noisy.[20] However, they are non-linear and might be complex if the user does not have prior knowledge of the choice of kernels and parameter optimisations.[21,26,28] On the other hand,

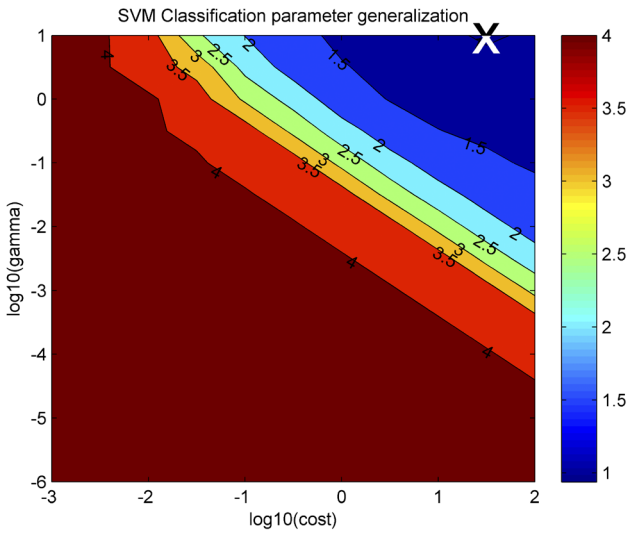

Figure 5. The contour plot of the optimisation parameters γ and *c* for discrimination of five tomato cultivars, and the position of "X" indicate the optimal result. The position "X" is a logarithmic value of 31.62 and 10 for γ and *c*, respectively.

PLS-DA is a simple linear classification method, and it is easy to understand and interpret the results.[18,28,29] The results from these two chemometric approaches are consistent with the outcomes on the prediction of cultivars from Vis-NIR datasets of rice seeds,[12] cabbage seeds,[16] tomato fruits[14] and leaves[6] etc. with accuracies ranging from 94% to 100%. The choice of the chemometric methods could also be extended to other linear classification methods like soft independent modelling of class analogy (SIMCA) and discriminant analysis (DA), as they also demonstrated higher classification accuracies of 94% and 97% in identifying cultivars of cabbage seeds[16] and tomato fruits,[14] respectively. Artificial neural network (ANN), a non-linear chemometric method,[12] has also been used in classifying rice seeds[12] and Chinese bayberry[15] with >96% and 95% prediction accuracies, respectively. All these methods give us good alternatives to choose from, however, priority should be given to the methods which are simple, easy to understand and interpret results.[18,29]

The results obtained from Vis-NIR are promising and comparable to the previous study which included seed colour attributes like intensity, hue, saturation, CIELab L*, CIELab a*, CIELab b*, nCDA-based trimmed mean pixel value and seed shape features.[5] Nevertheless, the purpose of the current study was not to compare results with the previously used method nCDA,[5] rather more to highlight the features of the instrument which can be used in other chemometric platforms for better classification. Inclusion of Vis-NIR and morphological attributes of seeds has been reported to increase the classification accuracy to a certain extent.[12] However, morphological traits from tomato seeds have been found to be inconclusive in discrimination of tomato cultivars.[5] The results from the Vis-NIR spectra are robust enough for classification of tomato seed cultivars, and the inclusion of the morphological traits may only add more noise to the model. Furthermore, the results are from only one

Table 4. The sensitivity and specificity results obtained from two chemometric methods for five different tomato cultivars. Sensitivity is the ability of the model to correctly identify samples of the cultivar, whereas the specificity is the capacity to reject the samples of other cultivars. The values range from 0 to 1; high values indicate better classification results.

Data set	Cultivars	PLS-DA		SVM-DA	
		Sensitivity	Specificity	Sensitivity	Specificity
Calibration	BL410	0.93	0.93	0.99	1.00
	CL	1.00	0.98	0.99	1.00
	Care Nepal	0.95	0.96	0.99	1.00
	HRD17	1.00	0.99	1.00	1.00
	T9	0.97	0.97	0.99	1.00
Cross-validation	BL410	0.93	0.92	0.97	1.00
	CL	1.00	0.98	0.97	1.00
	Care Nepal	0.95	0.96	0.98	1.00
	HRD17	0.99	0.99	0.99	0.99
	T9	0.97	0.97	0.98	0.99
Test set one	BL410	1.00	0.96	0.98	1.00
	CL	1.00	0.99	0.97	1.00
	Care Nepal	1.00	0.96	1.00	1.00
	HRD17	1.00	1.00	1.00	0.99
	T9	0.94	0.96	0.95	1.00
Test set two	BL410	0.96	0.97	1.00	0.99
	CL	0.99	0.98	1.00	1.00
	Care Nepal	0.97	0.98	0.96	0.99
	HRD17	0.99	0.99	1.00	1.00
	T9	0.97	0.98	0.94	1.00

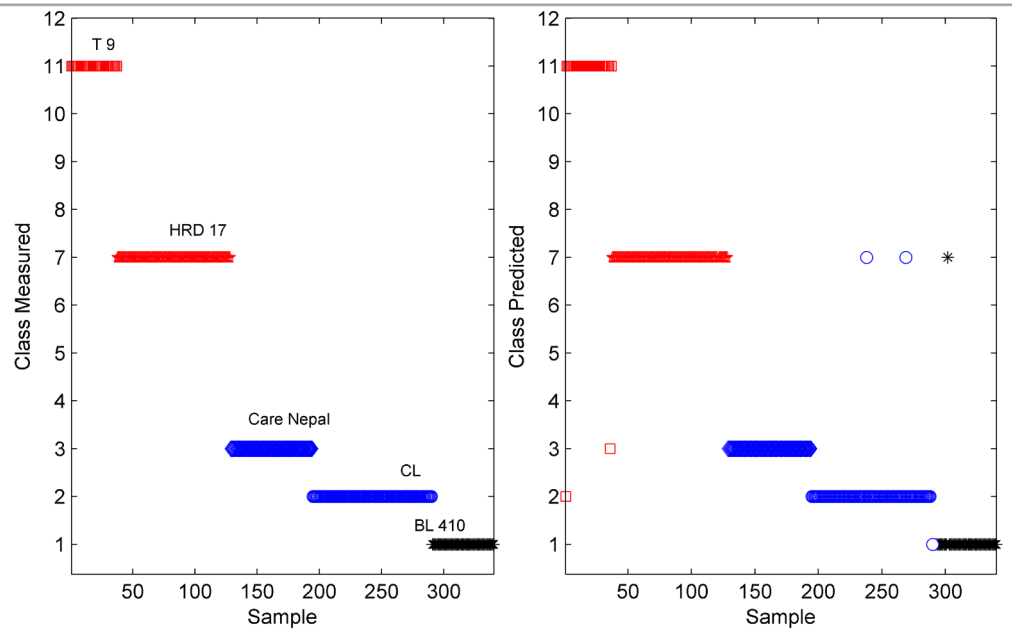

Figure 6. The measured (left) and predicted classes (right) of test set one for each tomato cultivar obtained from SVM-DA.

growth condition and further validation would be required to test the robustness of the technology and the methods using the samples from seed lots containing variation in growing conditions and seed age. Growing conditions and seed age have been associated with differences in the physical and chemical properties of the seeds,[30–33] which in turn has an effect on the spectral properties of the seed.[34] Moreover, Vis-NIR spectra have also shown their ability to discriminate or identify tomato cultivars based on their reflectance from leaves[6] and tomato fruits.[12,14–16] Further, Vis-NIR spectra have been successfully used to segregate transgenic and non-transgenic tomato genotypes using reflectance from tomato fruits[13] and also for rice seed genotypes.[12] These studies suggest an alternative option to the use of highly sophisticated molecular markers for cultivar identification or for discrimination, which are often expensive and time-consuming. Therefore, the significance of these studies advocates for at least integration of the technology for initial assessment of the materials, which will provide wider flexibility of the use of the plant materials without any sample preparation from leaves,[6] fruits[13,14] or seeds for discrimination of tomato cultivars.

Conclusion

This study presents the novelty of using Vis-NIR spectra for classification of different tomato cultivars using seeds together with chemometric approaches. PLS-DA and SVM-DA were both equally good for prediction of the unknown samples with the highest classification accuracy. The seed samples used in the study were from the same harvest year with similar growing conditions (climate and cultivation practices). So, further research is needed to test the performance of the Vis-NIR spectra for seeds of different tomato cultivars having variations in harvest year and growing conditions.

References

1. V. Er, J.A. Lane, R.M. Martin, P. Emmett, R. Gilbert, K.N. Avery, E. Walsh, J.L. Donovan, D.E. Neal and F.C. Hamdy, "Adherence to dietary and lifestyle recommendations and prostate cancer risk in the prostate testing for cancer and treatment (Protect) trial", *Cancer Epidem. Biomar.* **23,** 2066 (2014). doi: http://dx.doi.org/10.1158/1055-9965.EPI-14-0322
2. FAOSTAT, http://faostat3.fao.org (2013).
3. H.W.M. Hilhorst, "The tomato seed as a model system to study seed development and germination", *Acta Bot. Neerl.* **47,** 169 (1998).
4. R. Tripp, N. Louwaars and D. Eaton, "Plant variety protection in developing countries. A report from the field", *Food Policy* **32,** 354 (2007). doi: http://dx.doi.org/10.1016/j.foodp01.2006.09.003
5. S. Shrestha, L.C. Deleuran, M.H. Olesen and R. Gislum, "Use of multispectral imaging in varietal identification of tomato", *Sensors* **15,** 4496 (2015). doi: http://dx.doi.org/10.3390/s150204496
6. H.-r. Xu, P. Yu, X.-p. Fu and Y.-b. Ying, "On-site variety discrimination of tomato plant using visible-near infrared reflectance spectroscopy", *J. Zhejiang Univ. Sci. B* **10,** 126 (2009). doi: http://dx.doi.org/10.1631/jzus.B0820200
7. J. Smith and J. Register III, "Genetic purity and testing technologies for seed quality: a company perspective", *Seed Sci. Res.* **8,** 285 (1998). doi: http://dx.doi.org/10.1017/S0960258500004189
8. M.R. Foolad and D.R. Panthee, "Marker-assisted selection in tomato breeding", *Crit. Rev. Plant Sci.* **31,** 93 (2012). doi: http://dx.doi.org/10.1080/07352689.2011.616057
9. M. Caramante, G. Corrado, L.M. Monti and R. Rao, "Simple sequence repeats are able to trace tomato culti-

vars in tomato food chains", *Food Control* **22,** 549 (2011). doi: http://dx.doi.org/10.1016/j.foodcont.2010.10.002

10. M.B. Santos, A.A. Gomes, W.T. Vilar, P. Almeida, M. Milani, M. Nóbrega, E.P. Medeiros, R.K. Galvão and M.C. Araújo, "Non-destructive NIR spectrometric cultivar discrimination of castor seeds resulting from breeding programs", *J. Brazil. Chem. Soc.* **25,** 969 (2014).

11. J.G. Cruz-Castillo, S. Ganeshanandam, B.R. MacKay, G.S. Lawes, C.R.O. Lawoko and D.J. Woolley, "Applications of canonical discriminant analysis in horticultural research", *HortScience* **29,** 1115 (1994).

12. C. Liu, W. Liu, X. Lu, W. Chen, J. Yang and L. Zheng, "Nondestructive determination of transgenic *Bacillus thuringiensis* rice seeds (*Oryza sativa* L.) using multispectral imaging and chemometric methods", *Food Chem.* **153,** 87 (2014). doi: http://dx.doi.org/10.1016/j.foodchem.2013.11.166

13. L.J. Xie, Y.B. Ying, T.J. Ying, H.Y. Yu and X.P. Fu, "Discrimination of transgenic tomatoes based on visible/near-infrared spectra", *Anal. Chim. Acta* **584,** 379 (2007). doi: http://dx.doi.org/10.1016/j.aca.2006.11.071

14. L. Xie, Y. Ying and T. Ying, "Classification of tomatoes with different genotypes by visible and short-wave near-infrared spectroscopy with least-squares support vector machines and other chemometrics", *J. Food Eng.* **94,** 34 (2009). doi: http://dx.doi.org/10.1016/j.jfoodeng.2009.02.023

15. X. Li, Y. He and H. Fang, "Non-destructive discrimination of Chinese bayberry varieties using vis/NIR spectroscopy", *J. Food Eng.* **81,** 357 (2007). doi: http://dx.doi.org/10.1016/j.jfoodeng.2006.10.033

16. D. Wu, L. Feng, Y. He and Y. Bao, "Variety identification of Chinese cabbage seeds using visible and near-infrared spectroscopy", *Trans. ASABE* **51,** 2193 (2008). doi: http://dx.doi.org/10.13031/2013.25382

17. M. Sjöström, S. Wold and B. Söderström, "PLS discriminant plots", in *Pattern Recognition in Practice*, Ed by E.S.G.N. Kanal. Elsevier, Amsterdam, p. 461 (1986). doi: http://dx.doi.org/10.1016/B978-0-444-87877-9.50042-X

18. D. Ballabio and V. Consonni, "Classification tools in chemistry. Part 1: linear models. PLS-DA", *Anal. Meth.* **5,** 3790 (2013). doi: http://dx.doi.org/10.1039/c3ay40582f

19. C. Cortes and V. Vapnik, "Support-vector networks", *Mach. Learn.* **20,** 273 (1995). doi: http://dx.doi.org/10.1007/BF00994018

20. S. Mahadevan, S.L. Shah, T.J. Marrie and C.M. Slupsky, "Analysis of metabolomic data using support vector machines", *Anal. Chem.* **80,** 7562 (2008). doi: http://dx.doi.org/10.1021/ac800954c

21. R.G. Brereton and G.R. Lloyd, "Support vector machines for classification and regression", *Analyst* **135,** 230 (2010). doi: http://dx.doi.org/10.1039/B918972F

22. M.H. Olesen, P. Nikneshan, S. Shrestha, A. Tadayyon, L.C. Deleuran, B. Boelt and R. Gislum, "Viability prediction of *Ricinus cummunis* L. seeds using multispectral imaging", *Sensors* **15,** 4592 (2015). doi: http://dx.doi.org/10.3390/s150204592

23. R.J. Barnes, M.S. Dhanoa and S.J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra", *Appl. Spectrosc.* **43,** 772 (1989). doi: http://dx.doi.org/10.1366/0003702894202201

24. S. Wold, K. Esbensen and P. Geladi, "Principal component analysis", *Chemometr. Intell. Lab. Syst.* **2,** 37 (1987). doi: http://dx.doi.org/10.1016/0169-7439(87)80084-9

25. M. Barker and W. Rayens, "Partial least squares for discrimination", *J. Chemometr.* **17,** 166 (2003). doi: http://dx.doi.org/10.1002/cem.785

26. Y. Xu, S. Zomer and R.G. Brereton, "Support vector machines: a recent method for classification in chemometrics", *Crit. Rev. Anal. Chem.* **36,** 177 (2006). doi: http://dx.doi.org/10.1080/10408340600969486

27. J. Duckworth, "Mathematical data preprocessing", in *Near-Infrared Spectroscopy in Agriculture*, Ed by C.A. Roberts, J. Workman Jr and J.B. Reeves III. ASA, CSSA and SSSA, Madison, WI, p. 115 (2004). doi: http://dx.doi.org/10.2134/agronmonogr44.c6

28. H. Li, Y. Liang and Q. Xu, "Support vector machines and its applications in chemistry", *Chemometr. Intell. Lab. Syst.* **95,** 188 (2009). doi: http://dx.doi.org/10.1016/j.chemolab.2008.10.007

29. S. Wold, M. Sjöström and L. Eriksson, "PLS-regression: a basic tool of chemometrics", *Chemometr. Intell. Lab. Syst.* **58,** 109 (2001). doi: http://dx.doi.org/10.1016/S0169-7439(01)00155-1

30. G.W. Rathke, O. Christen and W. Diepenbrock, "Effects of nitrogen source and rate on productivity and quality of winter oilseed rape (*Brassica napus* L.) grown in different crop rotations", *Field Crops Res.* **94,** 103 (2005). doi: http://dx.doi.org/10.1016/j.fcr.2004.11.010

31. J. Vollmann, C.N. Fritz, H. Wagentristl and P. Ruckenbauer, "Environmental and genetic variation of soybean seed protein content under Central European growing conditions", *J. Sci. Food Agric.* **80,** 1300 (2000). doi: http://dx.doi.org/10.1002/1097-0010(200007)80:9<1300::aid-jsfa640>3.3.co;2-9

32. A. Francis and P. Coolbear, "Changes in the membrane phospholipid-composition of tomato seeds accompanying loss of germination capacity caused by controlled deterioration", *J. Exp. Bot.* **35,** 1764 (1984). doi: http://dx.doi.org/10.1093/jxb/35.12.1764

33. A. Francis and P. Coolbear, "A comparison of changes in the germination responses and phospholipid composition of naturally and artificially aged tomato seeds", *Ann. Bot.* **59,** 167 (1987).

34. B.G. Osborne, T. Fearn and P.H. Hindle, *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis*. Longman Scientific and Technical, Harlow, UK (1993).