Classification and Identification of Selected Rapeseed Varieties through Handheld Near-Infrared Spectrophotometer

Khurshid Ahmad, Hafiz Munir Ahmed, Afzal Shah

Abstract:

Plant variety identification plays a vital role in maintaining genetic purity leading to improve seed business. Researchers are always in the hunt of fast, less expensive and environment friendly non-destructive methods for classification of crop varieties. This study is aimed to evaluate the potential of portable/handheld NIR to make a supervised classification model for the rapeseed varieties. The seed samples (N=225) of three (03) rapeseed varieties were scanned with handheld SCiO NIR sensor and the average of the three scans were used for classification of varieties. The classification model developed by the combination of different pre-processing and classification algorithms were tested on unknown samples (n=75). It was found that all classifiers exhibited good results except Partial Least Square–Discrimination Analysis (Rc^2=0.8). SIMCA classification was tested that correctly identified 96.4% and 93.3% samples from training and test sets respectively followed by Random Forest classifier (F1=0.97) with a success rate of 93.3% on test set. However, Support vector machine (C-SVM type) with a polynomial kernel (3^{rd} degree) gave accurate results after a combination of Standard Normal Variate (SNV) and first order Savitzky-Golay derivative (polynomial degree of 2) with number of smoothing points (window size) of 5. It classified 100% samples of training set and 97.3% samples of test set into their correct classes. Based on initial evaluation of four classification algorithm, it was found that SVM can be better utilized for varietal classification.

Keywords

NIR spectroscopy, support vector machine, SIMCA, standard normal variate, Savitzky–Golay, classification

Introduction:

Development of a plant variety is a costly task in terms of skill, infrastructure and genetic resources as well as a time-consuming, (approx. 10 years). Therefore variety identification is an important errand in breeding programs to maintain genetic purity of varieties, prevention of fraud and improvement of seed trade(Kosmowski and Worku, 2018; Xu et al., 2009).

There are several methods employed for variety identification including traditional morphological characters method and modern techniques(Morell et al., 1995). The former, although, fast and less expensive methods but some time misleading due to lack of phenotypic variation causes difficulties for variety discrimination. Modern techniques such as molecular Markers, DNA profiling techniques etc. are most precise and accurate methods yet require a well-developed laboratory to execute the discrimination study that is why they are expensive methods and cannot be performed on-site. Similarly in agriculture sector we need a large number

of samples to be the true representative of a population which also limits the use of modern sophisticated methods for the said purpose. Therefore the thrust for an accurate, fast, reliable, and cost-effective method of identification of plant varieties is always there to support agriculture research & development.

Recently many researchers used near infrared (NIR) spectroscopy for classification of several agriculture products(Lerma-García et al., 2018). NIR operates in small electromagnetic region (e.m.r) from 750-2500 nm responsible for overtones and combination bands of vibration. As the overtone are less probable, therefore, this region of e.m.r. has weak interaction with mater. The weak interactive mode of near infrared region make it difficult to be used for analytical purpose but the same nature has a benefits that NIR can penetrate deep into a sample which make it an excellent region to study interior of a sample(Lammertyn et al., 2000). The disadvantage of weak interaction (weak signals) can be overcome by use of some advanced statistical tools e.g. pretreatment of spectral data and use of some algorithm's including Partial Least Square Regression-Discrimination Analysis (PLS-DA), Support Vector Machine (SVM) and Soft Independent Modelling by Class Analogy (SIMCA) collectively called chemometrics(García-Sánchez et al., 2017; López et al., 2017).

The literature review showed that several scientists have used near infrared spectroscopy for successful classification of several fruits and crops instead of laboratory based sophisticated expensive techniques. Abu-Kalaf *et al* used NIR in reflectance mode for classification of carrot on sensory characteristics by using multivariate data analysis with an accuracy greater than 84% (Abu-Khalaf et al., 2004). Betrand and his team classified wheat varieties with 87% success using NIR in reflectance mode in combination with chemometrics (Bertrand et al., 1985). Zou *et al* used full range of VIS/NIR (300-2500 nm) for the classification of rapeseed varieties using distance discriminant analysis (DDA) and Back Propagation Neural Network (BPNN)(Zou et al., 2011). Several researchers applied different statistical tools on spectral data of NIR and developed models for quality assessments of different biochemical traits such as oils, protein, glucosinolates etc.(Hom et al., 2007).

However the main advantage of at-site application can only be obtained by application of portable/handheld NIR devices. As the handy NIR usually operate in a narrow portion of near infrared range due using simple light detector for lowering the cost of device. Which means that much less number of spectral data point is used for model development which usually compromise the high accuracy of analysis. However a reasonable high accuracy can be achieved, if proper statistical tools is applied for smoothing and removal of noise and scattering effect. Recently several manufacturer claimed for the development of pocket size NIR sensor with different spectral ranges. Cen *et al* used handheld VIS/NIR for classification of orange varieties using Back Propagation Neural Network (BPNN) combine with PCA and PLS-DA and concluded that PCA+BPNN gave better result (Cen et al., 2007). In another study, a portable NIR was used for successful classification of grapevine varieties using SVM, ANN and PLS-DA as classifier algorithms (87.25%). They found that SVM and ANN outperformed PLS-DA (Gutierrez et al., 2015). He *et al* developed a fast method for successful identification of apple

varieties using NIR and chemometrics such as Wavelength Transform (WT), and ANN and claimed a classification model development with 100% accuracy (He et al., 2007).

In the light of the literature cited and to develop a model for at-site rapeseed variety differentiation and confirmation, a study was designed to evaluate the efficiency of pocket size NIR device to make classification model for the rapeseed varieties developed at NIFA. The study will also intended to help the breeders for the selection of true to type varieties, if combine with other phenotypic and agronomic character.

Materials and methods:

Rapeseed samples:

Three hundred samples of three rapeseed varieties viz Durr-e-NIFA (DN), NIFA Gold (NG) and NIFA Sarson-T20 (NS) developed by Plant Breeding and Genetics Division of Nuclear Institute for Food and Agriculture (NIFA), Peshawar, were collected under the supervision of breeders during harvest season 2019-20 and 2020-21 and stored at oilseed store at room temperature. Three scans of intact grain samples were recorded by using SCiO sample holder and averaged. The sample holder was shaken well before recording each scan to ensure averaged scan data would be representative of the larger sample. Three hundred samples were randomly split into seventy five samples of each variety as training set (N= 225) and twenty five samples as test set (n=75).

NIR spectrometer sensor and Chemometrics:

NIR sensor (make: Consumer Physics SCIO) having dimension 18.8 × 40.2 × 67.7 mm and weighing 35 g , was used for recording spectral finger print (McGrath et al., 2021). An android app "The Lab" installed on smartphone was used to operate and record the spectra. The device operate in wavelength range of 740–1070 nm with 1nm resolution (331 data points as dependent variables). The spectra were stored in Consumer Physics cloud database and can be accessed through a window-based app "The Lab web". The Lab Web can also be used to view the data in PCA view and to remove the outliers. The Lab Web was used for pre-processing data and developing and testing classification models. The Lab Web allow some pretreatment method such as SNV, Derivatives, de-trending etc. and Random Forest (RF) algorithm to classify the samples with F1 value and confusion matrix to check the performance of the model. The data were downloaded to PC and exported to "The Unscrambler X", v10.4 (CAMO software) for further analysis.

Pre-processing and classification algorithms

In this study several pretreatment methods combined with predictive algorithms were compared by using the Unscrambler X software. The performance of four advanced algorithms including Partial Least Squares Discriminant Analysis (PLS-DA), Support Vector Machine (SVM), soft independent modelling by class analogy (SIMCA) and Random Forest (RF) were checked to develop predictive models of grain cultivar identification. Different preprocessing methods including Baseline Correction, Standard Normal Variate (SNV) and smoothing with Savitzky-Golay derivative (SG-d) were used to convert the raw spectra to more informative curve. It was

found that the combination of SNV and derivative yielded best results in all classification models.

Results and Discussion:

Rapeseed is considered to be the second important oilseed crop having a rich source of oil, proteins and many saturated and unsaturated fatty acid (Raymer, 2002; Reddy, 2017). The varietal identification by naked eye is a real daunting task and most of the time leads to incorrect identification, until and unless some clear cut morphological markers are not available. Instead of costly and destructive methods of seed identification, researchers are trying to make fast, cheap and non-destructive methods for this purpose. The intact seeds of all the three varieties were scanned with pocket size NIR sensor (Fig. 01). Due to poor interaction of near infrared region with matter, the raw spectra usually shows no apparent peaks however, after applying some pretreatments, the spectra can be converted into more informative curve. The raw spectra were pretreated with Standard Normal Variate (SNV), and first order Savitzky-Golay derivative with polynomial order of 2 and different window size followed by model building. As the spectra were recorded on whole grain the scattering effect can be minimized by SNV while derivatives are good mean to elaborate a weak peaks by resolving the overlapped peaks (Barnes et al., 1989; Huang et al., 2010). Four classification methods were used along with different pretreatments to get best combination of pretreatment and classification algorithm for classification.

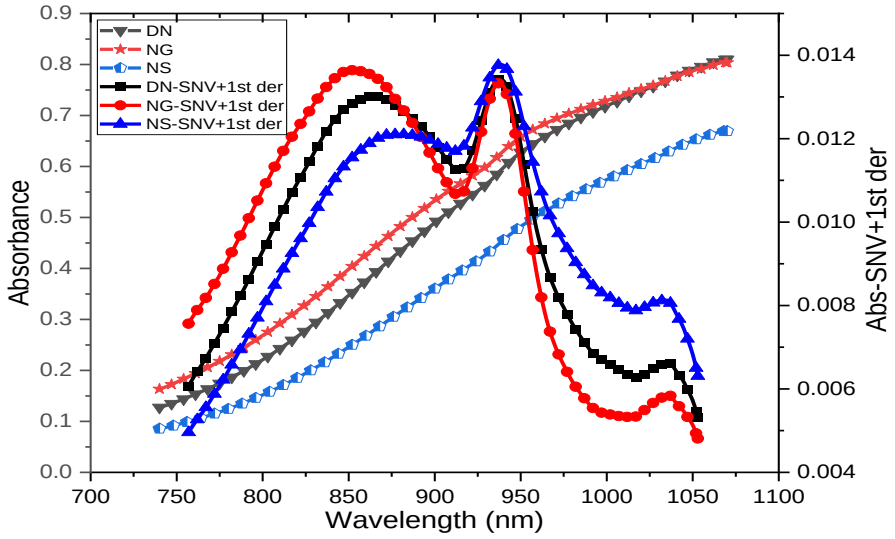

Fig. 01 Averaged raw spectra and preprocessed with SNV + 1st order S-Golay derivative (W35)

1. Partial Least Square-Discriminate Analysis (PLS-DA):

Partial Least square (PLS) regression is a well-known algorithms commonly used for estimation purposes however it can be applied for classification purposes by selecting some dummy variables for each class(Cen et al., 2007). PLS-DA with a random cross validation of twenty segments (15 samples per segment) was used to make a classification model on preprocessed (SNV + SG-d, 1st, 2, W35) spectra of N=225 samples, the score plot is shown in Fig. 02. The

values of coefficient of regression for calibration (R^2_c=0.78) and validation (r^2 =0.74) showed that the model would poorly classify the unknown samples. When same model was used to classify the unknown samples (n=75) of test set which were not included in calibration set, 65% of samples were correctly classified to their respected groups. However, by changing the pretreatment conditions i.e. decreasing the window size of SG derivative from 35 to 5 and using full cross validation, a small improvement was observed in model. The R^2c was 0.80 and 84.9% samples of training set and 68% samples of the test set were correctly classified.

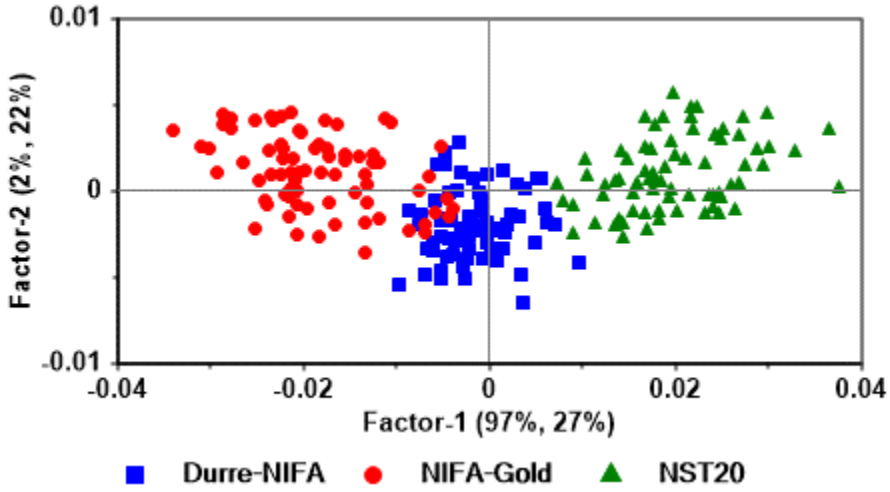

Fig. 02. Factor-1 and 2 showing that the first two factors explain 99% variation in data

As the SCiO NIR sensor operate in a short range of NIR region (331 data points), which is mostly the region of 2nd and 3rd overtones. The absorptions intensities of these overtones are very weak due to less probable nature as compared to 1st overtone vibration. Therefor ability of PLS-DA for classification may be enhanced by selecting spectra of broader range (i.e including the region of 1st overtone).

2. Support Vector Machine (SVM):

Support Vector Machines (SVM) is a supervised machine learning method which use a kernel functions to map the data from the original space to a new feature space. In the new feature space, the kernel function then find a support vectors for the best classification (Pierna et al., 2004). After applying pretreatment, the spectra were subjected to different SVM classifier by using different kernel and the results are summarized in Table 01 with percent accuracies of different pretreatment and SVM type.

Table 01. Different combinations of preprocessing before classification (SVM) along with accuracies

Pretreatment	Kernel Type	Calibration Accuracy (%)	Validation Accuracy (%)	Test set Accuracy (%)
Raw data	Polynomial (nu-SVM)	80.9	79.1	77.3
SG-d, W35	Polynomial (nu-SVM)	43.6	64.0	44.0
SNV	Polynomial (nu-SVM)	94.2	93.8	92.0

SNV+ SG-d, W5	Polynomial (nu-SVM)	93.8	92.9	93.3
SNV+ SG-d, W35	Polynomial (nu-SVM)	93.8	92.9	93.3
SNV+ SG-d, W35	Radial (nu-SVM)	93.8	93.8	93.9
SNV+ SG-d, W35	Polynomial (C-SVM)	98.7	98.7	96.0
SNV+ SG-d, W5	Polynomial (C-SVM)	100	99.1	97.3

Where: SG-d= 1st order S-Golay derivative of polynomial order of 2 and W= window, nu=0.5, C=1 and polynomial degree of SVM = 3

The best training and validation accuracy of 100% and 99.1% was obtained for C-SVM with capacity factor, C=1 after pretreatment combination of SNV and derivative (SG-d, 1st, 2, W5) and the kernel used was polynomial kernel type with degree of 3. The confusion matrix shown in Fig. 03 depicts that all samples of training set were predicted correctly into their rightful classes.

		Actual		
		Durre-NIFA	NIFA-Gold	NST20
Predicted	Durre-NIFA	100%	0%	0
	NIFA-Gold	0	100%	0
	NST20	0	0	100%

Fig. 03. Confusion matrix of samples of training set, classified with SVM (polynomial kernel of 3rd degree) after pretreatment of SNV and derivative (SG-d, 1st, 2, W5)

The samples from test set which were not included in calibration/training model, were then subjected to the classifier developed. As shown in Table 01, the combination of preprocessing (SNV+SG-d, 1st, 2, W5) and C-SVM algorithm of polynomial degree 3 classified maximum number of samples into their rightful classes.

3. Soft Independent Modelling by Class Analogy (SIMCA):

Soft Independent Modeling of Class Analogy (SIMCA) is another statistical method for supervised classification of data based on making a principal component analysis (PCA) model for each class separately in the training set. Unknown samples are then compared with all models and assigned to classes according to their closeness to respective PCA model in the training samples (Hibbert, 2016). The score plot of PCA (Fig. 04A) shows that 99% variance in the data could be explained by first two PCs and that NG show maximum variation as compared to DN and NS, the same is evident from the influence plot of F-residual and Hotelling's T^2 (Fig. 04B). The influence plot is used to show the two kinds of outlier. The residual value describes the sample distance to model, whereas the Leverage and Hotelling's T^2 define the ability of model to describe the sample. The samples with high leverage are well described by model but retaining that sample may span an entire component.

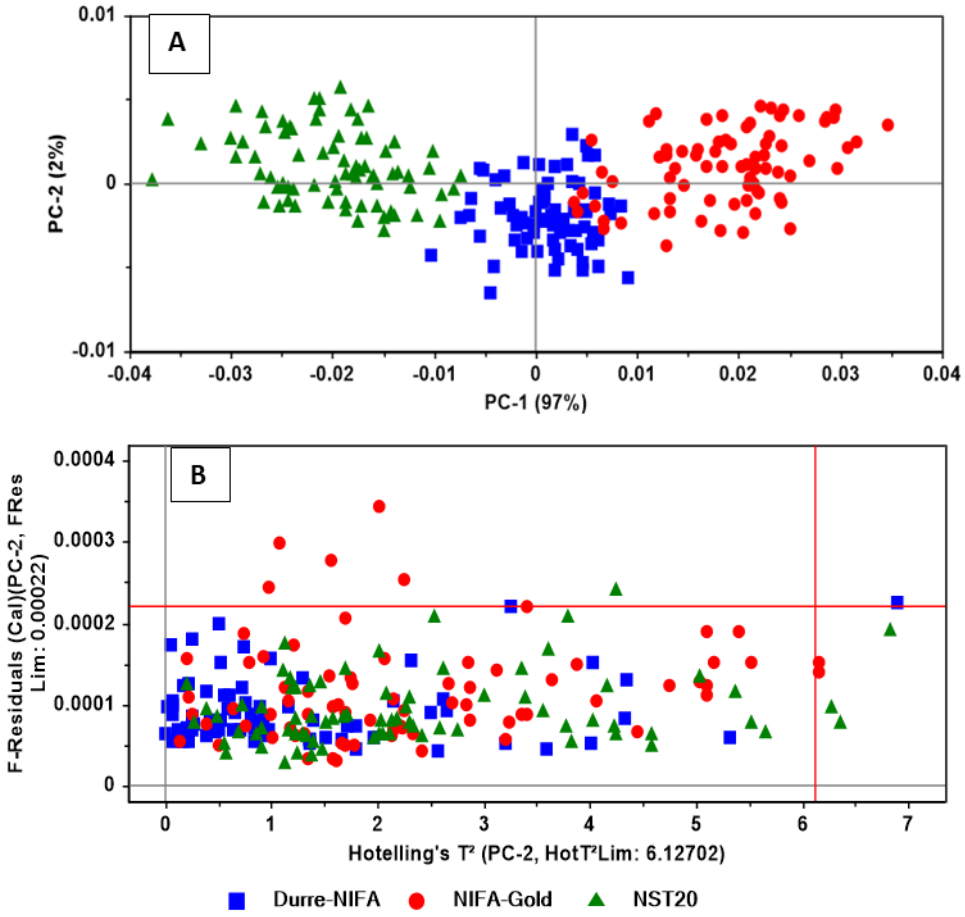

Fig. 04. PCA model of all three varieties (A) and influence plot of F-Residual and Hotelling's T^2 (B).

Separate PCAs model were built for all three varieties by using Singular Value Decomposition (SVD) algorithm on the preprocessed data. It is evident from Table 02 that almost all samples of training set were described by their respective model except one sample of DN and two samples of NS (false negative) -Table 02. Similarly 22 samples of DN were described by PCA (NG) and 9 by PCA (NS) (false positive). Similarly 5 samples of NG were false positively described by PCA (DN). By changing the pretreatment condition i.e. reducing window size from 35 to 5, all the false negative samples reduced to zero. In new condition all samples were classified by their respective PCAs. However, the doubly classified samples still exist in model (Table 02), therefore some other statistical tools must be applied to decide about their classes.

Table 02. Confusion matrix for test set classification (SIMCA-model)

Pretreatment	Variety	PCA(DN)	PCA(NG)	PCA(NS)
SNV + SG-d,1st, 2, W35)	DN	24	22	9
	NG	5	25	0
	NS	0	0	23
SNV + SG-d,1st , 2, W5	DN	25	22	10
	NG	5	25	0
	NS	2	0	25

To decide about the rightful class of a doubly classified samples, is used to calculate the distance of sample from both PCA models (Sim et al., 2004). The PCA of DN, NG and NS plotted on the Coomans' plot, to determine the classification of the unknown sample by measuring its distance from the either of the classes along vertical and horizontal axis (Fig. 05). The shortest distance from any of the classifier from the point of the unknown sample will decide its rightful class. By comparing the distances of all samples from the three models, it was found that 217 samples (96.4%) of training set and 70 (93.3%) samples of test set were correctly classified into their rightful classes.

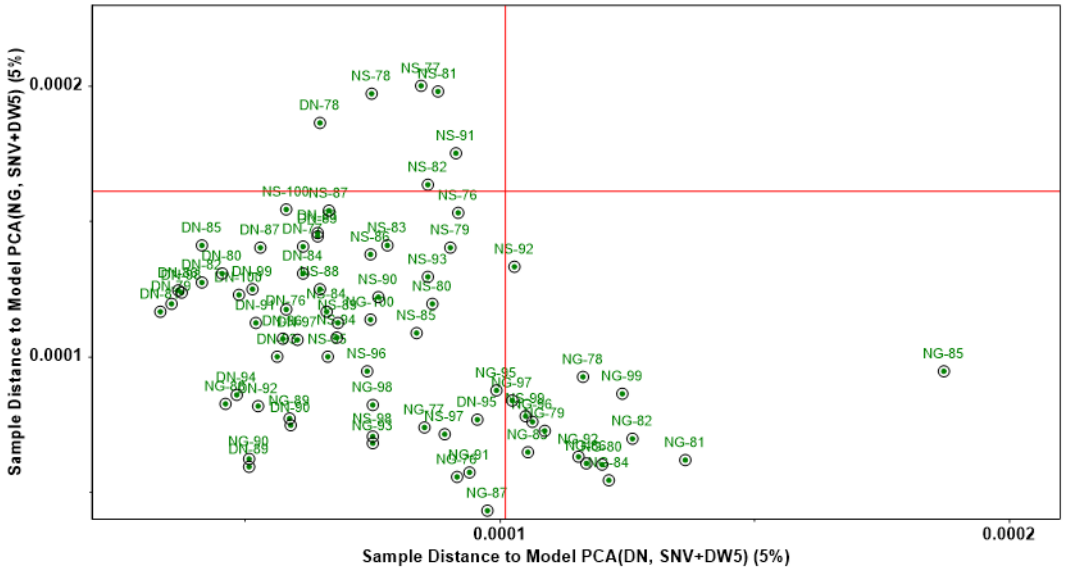

Fig. 05. Coomans' Plot of PCA DN and PCA NG showing the distance of samples from each model.

4. Classification model using the "The Lab web"

The spectra collected were also analyzed by SCiO Lab, a web based app of consumer physics. An android app "the Lab" was used to collect the spectra and data stored at Consumer Physics cloud was accessed by "the Lab Web". The lab web allows the users to select a set of pre-processing functions prior to development of classification models. After preprocessing the averaged spectra are shown in Fig. 06. However, there is a single algorithm available known as Random Forest (RF) algorithm which make a more accurate and stable prediction based on numerous decision trees (Shannon et al., 2021). The efficiency is checked by F1 score (also called F score) and confusion matrix. F1 score can be applied to multiclass classification and is calculated from the precision and recall of the test (Chicco and Jurman, 2020).

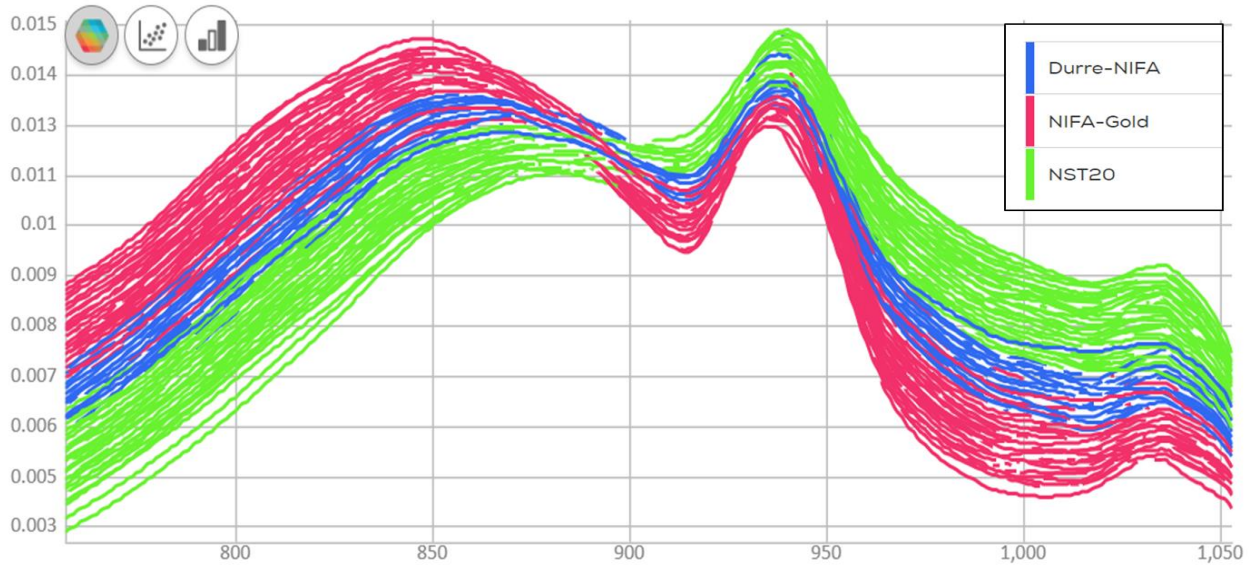

Fig. 06. Spectra of three varieties using SNV and 1st order derivative on "The Lab Web" view.

After checking the several preprocessing methods and their combination, it was revealed that the combination of SNV and derivative (SG-d, 1st, 2, W35) gave best result with F1 score of 0.97 for the classification of three varieties (confusion matrix is given in Fig. 07) on training set with 10 fold cross validation by samples.

		Actual		
		Durre-NIFA	NIFA-Gold	NST20
Predicted	Durre-NIFA	98%	04%	0
	NIFA-Gold	0	95%	0
	NST20	01%	0	99%

Fig. 07. Confusion matrix by Random Forest classifier using SNV and 1st order derivative

Using the "Test model" option in "The Lab" android app, the model was checked on 75 samples (25 of each) of all the three varieties. 70 of total samples (93.3%) were correctly classified into their rightful classes with a confidence level of 70-99%. However the 4 samples of NG were classified with a confidence level 54-70% which was re-checked thrice to confirm the class.

Table 03. Summary of all combinations of preprocessing and classification algorithms.

Classification Algorithm	Pretreatment	Correctly classified samples: Training set N=225	Correctly classified samples: Test set n=75
PLS-DA (NIPLAS)	SNV+ SG-d, W35	82.6%	65%
	SNV+ SG-d, W5	84.9%	70%
C-SVM (polynomial kernel)	SNV+ SG-d, W35	98.7%	96.0%
	SNV+ SG-d, W5	100%	97.3%
SIMCA (SVD)	SNV+ SG-d, W35	96.0%	92.6%
	SNV+ SG-d, W5	96.4%	93.3%
Random Forest	SNV+ SG-d, W35	--	93.3%

Where: SG-d= 1st order S-Golay derivative of polynomial order of 2 and W= window size (smoothing points)

Conclusions:

Based on the results, it was concluded that C-SVM (C=1) classification in combination with pretreatment of SNV and 1st order derivative exhibited accurate results on both training and test sets compared to SIMCA and Random Forest classifications. In this study classification models were developed for selected three varieties and sample analyzed qualified over 90% for their rightful class. Therefore, it is suggested that handheld SCiO NIR sensor can be best utilized for classification of all available rapeseed varieties and spot analysis to determine the purity and rightfulness of unknown seed samples class through C-SVM (C=1).

References:

Abu-Khalaf, Nawaf, Bent S Bennedsen and Gitte K Bjørn 2004. Distinguishing carrot's characteristics by near infrared (NIR) reflectance and multivariate data analysis. Agricultural Engineering International: CIGR Journal.

Barnes, RJ, Mewa Singh Dhanoa and Susan J Lister 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied Spectroscopy 43: 772-777.

Bertrand, Dominique, Paul Robert and William Loisel 1985. Identification of some wheat varieties by near infrared reflectance spectroscopy. Journal of the Science of Food and Agriculture 36: 1120-1124. doi: https://doi.org/10.1002/jsfa.2740361114

Cen, Haiyan, Yong He and Min Huang 2007. Combination and comparison of multivariate analysis for the identification of orange varieties using visible and near infrared reflectance spectroscopy. European Food Research and Technology 225: 699-705. doi: 10.1007/s00217-006-0470-2

Chicco, Davide and Giuseppe Jurman 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics 21: 1-13.

García-Sánchez, Francisco, Luis Galvez-Sola, Juan J Martínez-Nicolás, Raquel Muelas-Domingo and Manuel Nieves 2017. Using near-infrared spectroscopy in agricultural systems. Kyprianidis KG, Skvaril J. Developments in near-infrared spectroscopy. London: IntechOpen: 97-127.

Gutierrez, Salvador, Javier Tardaguila, Juan Fernandez-Novales and Maria P Diago 2015. Support vector machine and artificial neural network models for the classification of grapevine varieties using a portable NIR spectrophotometer. PloS one 10: e0143197.

He, Yong, Xiaoli Li and Yongni Shao 2007. Fast Discrimination of Apple Varieties Using Vis/NIR Spectroscopy. International Journal of Food Properties 10: 9-18. doi: 10.1080/10942910600575666

Hibbert, David B 2016. Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations 2016). Pure and Applied Chemistry 88: 407-443.
Hom, Nang Hseng, Heiko C Becker and Christian Möllers 2007. Non-destructive analysis of rapeseed quality by NIRS of small seed samples and single seeds. Euphytica 153: 27-34.
Huang, J, S Romero-Torres and M Moshgbar 2010. Practical Considerations in Data Pre-treatment for NIR and Raman Spectroscopy. American Pharmaceutical Review http://www.americanpharmaceuticalreview.
Kosmowski, Frédéric and Tigist Worku 2018. Evaluation of a miniaturized NIR spectrometer for cultivar identification: The case of barley, chickpea and sorghum in Ethiopia. PloS one 13: e0193620.
Lammertyn, Jeroen, Ann Peirs, Josse De Baerdemaeker and Bart Nicolaï 2000. Light penetration properties of NIR radiation in fruit with respect to non-destructive quality assessment. Postharvest Biology and Technology 18: 121-132. doi: https://doi.org/10.1016/S0925-5214(99)00071-X
Lerma-García, María J, Victoria Cortés, Pau Talens and José M Barat 2018. Variety discrimination of fruits, edible plants, and other foodstuffs and beverages by infrared spectroscopy. In Comprehensive Analytical Chemistry, 127-163: Elsevier.
López, Mercedes G, Ana Sarahí García-González and Elena Franco-Robles 2017. Carbohydrate analysis by NIRS-Chemometrics. Developments in Near-Infrared Spectroscopy 10: 67208.
McGrath, Terence F, Maeve Shannon, Olivier P Chevallier, Ratnasekhar Ch, Fei Xu, Fanzhou Kong, Hong Peng, Ernest Teye, Selorm Akaba and Di Wu 2021. Food Fingerprinting: Using a two-tiered approach to monitor and mitigate food fraud in rice. Journal of AOAC International 104: 16-28.
Morell, MK, R Peakall, R Appels, LR Preston and HL Lloyd 1995. DNA profiling techniques for plant variety identification. Australian Journal of Experimental Agriculture 35: 807-819.
Pierna, JA Fernández, V Baeten, A Michotte Renier, RP Cogdill and P Dardenne 2004. Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds. Journal of Chemometrics: A Journal of the Chemometrics Society 18: 341-349.
Raymer, Paul L 2002. Canola: an emerging oilseed crop. Trends in new crops and new uses 1: 122-126.
Reddy, Gadi VP 2017. Integrated management of insect pests on canola and other Brassica oilseed crops: CABI.
Shannon, Maeve, CH Ratnasekhar, Terence F McGrath, Arun P Kapil and Christopher T Elliott 2021. A two-tiered system of analysis to tackle rice fraud: The Indian Basmati study. Talanta 225: 122038.
Sim, Chew Oon, Mohammad Razak Hamdan, Zhari Ismail and Mohd Noor Ahmad 2004. Assessment of herbal medicines by chemometrics–assisted interpretation of FTIR spectra. J Analytica Chimica Acta 1: 14.
Xu, Hui-rong, Peng Yu, Xia-ping Fu and Yi-bin Ying 2009. On-site variety discrimination of tomato plant using visible-near infrared reflectance spectroscopy. Journal of Zhejiang University Science B 10: 126-132.
Zou, Qiang, Hui Fang, Fei Liu, Wenwen Kong and Yong He 2011. Comparative Study of Distance Discriminant Analysis and Bp Neural Network for Identification of Rapeseed Cultivars Using Visible/Near Infrared Spectra. In Comparative Study of Distance Discriminant Analysis and Bp Neural Network for Identification of Rapeseed Cultivars Using Visible/Near Infrared Spectra, eds. Daoliang Li, Yande Liu and Yingyi Chen, Computer and Computing Technologies in Agriculture IV, 124-133. Berlin, Heidelberg: Springer Berlin Heidelberg.