Title

Multiblock spectral imaging for identification of pre-harvest sprouting in Hordeum vulgare

Journal

Microchemical Journal, 191(Food Chem. 309 2020)

ISSN

0026-265X

Authors

Orth, Sebastian Helmut
Marini, Federico
Fox, Glen Patrick
et al.

Publication Date

2023-08-01

DOI

10.1016/j.microc.2023.108742

Peer reviewed

Microchemical Journal 191 (2023) 108742

Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

Multiblock spectral imaging for identification of pre-harvest sprouting in *Hordeum vulgare*

Sebastian Helmut Orth [a], Federico Marini [a,b], Glen Patrick Fox [a,c], Marena Manley [a], Stefan Hayward [a,*]

[a] Department of Food Science, Stellenbosch University, Private Bag X1, Matieland, Stellenbosch 7602, South Africa
[b] Department of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro 5, 1-00185 Rome, Italy
[c] Food Science and Technology, University of California Davis, 1Shields Ave, Davis, CA 95616, USA

ARTICLE INFO

Keywords:
Multiblock spectral imaging modelling
Multiblock variable selection
Malting barley
Preharvest germination classification
Near infrared hyperspectral imaging

ABSTRACT

A novel data fusion method based on the use of visible/near-infrared (VNIR) and shortwave infrared (SWIR) imaging sensors, to distinguish between pregerminated and ungerminated barley grain is proposed. Spectral imaging was used to fingerprint germinated and ungerminated barley grain from a total of 5640 average spectra representing single barley kernels varying with respect to germination time. Chemometric approaches utilising partial least squares-discriminant analysis (PLS-DA) and multiblock sequential and orthogonalized partial least squares-linear discriminant analysis (SO-PLS-LDA) and sequential and orthogonalized covariance selection-linear discriminant analysis (SO-CovSel-LDA) were used to build classification models. SO-PLS-LDA achieved a total classification rate of 99.88%, while SO-CovSel-LDA resulted in a classification accuracy of 97.46% when a maximum of 8 variables were selected from each data block (VNIR and SWIR) – models were validated on an independent test set. The use of multiblock approaches led to increased prediction accuracy, compared to PLS-DA, and a viable solution to address the industry problem to detect pregerminated malting barley in a rapid, non-destructive manner. This represents a significant advance with respect to the current dated methods which are hindered by time-consuming wet chemistry techniques and human subjective bias. The potential of the proposed new technique also has the further advantage of moving toward multispectral systems which can be used to detect pre-harvest germinated barley using an even more computationally rapid and affordable online sorting machine incorporating the wavebands of importance selected by SO-CovSel-LDA. The study highlights how sequential and orthogonalised data fusion approaches, in the food and agricultural sector, are powerful solutions to real world problems.

1. Introduction

Barley (*Hordeum vulgare* L.) played a significant role in the establishment of society 13,000 years ago [1] and currently ranks fourth following the most important cereal crops, wheat, maize and rice [2]. With an estimated global harvest of 140 million tonnes annually, barley contributed an estimated USD $25,6 bn to the global food sector in 2018 [3].

Although barley is used in various food sectors, it is mainly produced for use as the main ingredient in the production of beer and whiskey. However, to produce beer and whiskey, the barley must first be malted. Malting refers to the controlled, uniform germination and drying of the grain resulting in in the biosynthesis of various enzymes including hemicellulases, proteases, glucosidases, and amylases where endosperm cell walls and the endosperm protein matrix are hydrolysed [4]. Non-uniform germination during malting may result in inconsistencies in malt quality which may impact in the brewing process. Barley, intended for malt production in South Africa and the rest of the world, is therefore purchased from producers based on the germinative energy of the grain [5]. Germinative energy, a measure of grain viability, should ideally be 100 % since malting depends on rapid and predictable germination [6]. Standard germination tests can take up to 3 days. Germinative energy

Abbreviations: VNIR, Visible near infrared; SWIR, Short wave near infrared; PLS-DA, Partial least squares discriminant analysis; SO-PLS-LDA, Sequential and orthogonalised partial least squares linear discriminant analysis; SO-CovSel-LDA, Sequential and orthogonalised covariance selection linear discriminant analysis.

* Corresponding author.
E-mail address: stefanh@sun.ac.za (S. Hayward).

https://doi.org/10.1016/j.microc.2023.108742
Received 26 September 2022; Received in revised form 10 March 2023; Accepted 6 April 2023
Available online 7 April 2023

and capacity is affected by several factors, including environmental conditions during seed maturation and dormancy [7].

Dormancy is an internal characteristic of the seed which can be defined as the inability of a viable seed to germinate under optimal conditions [8,9]. In evolutionary terms, dormancy is an adaptation which promotes survival of the seed under adverse conditions [6]. For example, the presence of dormant seeds in soils could provide the opportunity for germination to occur over several seasons, thereby maximizing the chance for species survival [6]. However, although high levels of dormancy may be a positive attribute in species survival, to produce malt a low level of dormancy is desirable since uniform germination is required for malting within a few months after harvest. Barley varieties with low dormancy are therefore selected by breeding companies. However, selection pressures may result in varieties being released whose dormancy is terminated prior to harvest [8]. As a result of wet conditions, during filling and maturation, grains of such varieties can germinate before harvest, as there may be a delay between when the crop is physiological mature and when it can be harvested. This delay can lead to incipient germination or preharvest sprouting [6,8–10]. Incipient germination occurs when embryo growth is triggered, but the process is interrupted by desiccation before physical changes such as the emergence of the radicle can occur [9]. Pregerminated grain may maintain some viability, but its storage capacity is severely reduced. If wet conditions persist, the grain continues to germinate towards a 'point of no return' [9] where-after it loses its tolerance to desiccation [11]. This phenomenon is known as preharvest sprouting (PHS), and results in total loss of grain viability rendering the crop unsuitable for malting purposes [9].

Although visual inspection can be used to detect sprouted barley grains, it is not possible to detect early stage pregerminated grains when no external protrusions are present [10]. Methods for the detection of pregerminated cereal grains are mainly based on determination of hydrolysed starch due to the presence of α–amylase, synthesized during germination [12,13]. The α-amylase (and other starch degrading enzymes) breaks down starch into the base glucose which would be metabolized for the newly growing embryo. Any action of α-amylase activity on starch can be determined with the Hagberg falling number test and the Rapid Visco Analyzer (Stirring number test) or spectroscopic methods which rely on the use of a labelled substrate [14]. The presence of α-amylase can also be determined using enzyme linked immunosorbent assays (ELISA) with the use of monoclonal antibodies. However, these methods are often not sufficiently sensitive and require expensive and specialized equipment and operators [12]. These conventional methods also involve elaborate sample preparation and methodologies which require specialist training and an off-site laboratory to perform the analyses. While these methods are faster than a germination test, they cannot detect early-stage germination with certainty due to the relatively low amount of α-amylase present in the sample.

Based on the shortcomings of conventional methods, non-invasive spectroscopic techniques, that have the potential to rapidly classify between germinated and ungerminated cereal grains, have been investigated. Several studies focussed on the application of near-infrared (NIR) spectroscopy and NIR hyperspectral imaging for wheat [15–18] and barley [15,17,19]. Using these non-invasive methods, researchers were able to show the potential benefits of spectral imaging to detect pregerminated grains. However, some of these studies suffered from a small sample set used for calibration while others made use of destructive sample drying techniques which can potentially result in additional adverse biochemical changes to the grain's protein and starch structure. Such changes can affect the robustness of the classification model and the same sample preparation procedure (if any) should be followed in a controlled laboratory environment to ensure accurate classification. If the intent is to address an industry problem, in addition to the calibration model accuracy being sensitive to laboratory produced conditions, these conditions should mimic real world scenarios as close as possible. It was further highlighted, in a review of conventional and spectral imaging methods, that the development and use of a single kernel analysis approach provides information on the total variation, throughout the grain sample, leading to unbiased real-time decisions [20]. Spectral imaging also allows for the potential of laboratory sorting of individual grains in, e.g., breeding programmes.

The development and implementation of NIR technology for real-time analysis is paramount in the malting and beer brewing industry, to ensure optimum and consistent production [21]. A possible solution would be to employ multiple sensors with multiblock classification methods such as sequential and orthogonalized partial least squares-linear discriminant analysis (SO-PLS-LDA) for more robust models from supportive information, compared to partial least squares-discriminant analysis (PLS-DA). Multiblock waveband selection using sequential and orthogonalized covariance selection-linear discriminant analysis (SO-CovSel-LDA) could also strengthen the potential for use of multispectral imaging in industrial and agricultural settings. Furthermore, no studies to date considered the visible/near-infrared (VNIR) range to detect germination in barley with spectral imaging. Neither has the two multiblock methods been applied to spectral data obtained from NIR hyperspectral imaging systems. The benefit of the VNIR waveband region is that sensors are more affordable and readily available in most imaging systems. Considering the shortfalls of conventional approaches, the aim of this study was to investigate the use of spectral imaging in both VNIR and SWIR wavelength regions as a potential industry acceptable analytical approach for preharvest germinated barley classification. More specifically, spectral imaging was firstly investigated as a tool to differentiate between ungerminated and germinated barley grain using multiblock SO-PLS-LDA compared to conventional PLS-DA. Secondly, the potential to reduce the number of variables for a multispectral classifier, by obtaining the most important wavebands for both SWIR and VNIR ranges using SO-CovSel-LDA, was investigated.

2. Materials and methods

A breakdown of the data acquisition and modelling procedure undertook for this study is shown in Fig. 1 in the form of a flow diagram.

2.1. Sample acquisition and in vitro germination

Malting barley samples were kindly provided by the South African Barley Breeding Institute (SABBI, Caledon, South Africa) in collaboration with South African Breweries (SAB, Johannesburg, South Africa), a direct subsidiary of Anheuser-Busch InBev (ABInBev). At present only three varieties of malting barley, namely Kadie, Hessekwa and Elim are recommended for production in the Southern Cape of South Africa under dryland conditions. In addition, two varieties, namely Genie and Overture are recommended for use under irrigation conditions. These five commercial varieties differ in terms of dormancy (period from harvesting up to optimal malting stage), malting characteristics and phenotypical appearance. The five varieties were obtained in bulk (1 kg) and each divided into 38 sub-samples. Germination was initiated by firstly weighing 10 g of each sub-sample into a 90 × 15 mm polystyrene Petri dish. The samples were subsequently imbibed by the addition of 10 mL of deionised water (Merck, Milli-Q, Direct-Q3) with 1 ppm of Amphotericin B to inhibit fungal growth. After imbibition, the samples were allowed to swell and germinate at ambient temperature (22 °C) for 48 h. Germination and other biochemical and physiological changes of one sample of each variety were terminated cryogenically at −80 °C hourly up to 36 h and again at 48 h. The 0 h sample were cryogenically preserved immediately after imbibition. Moisture was removed by lyophilisation to preserve biochemical and physiological changes that took place within the grain during imbibition. Lyophilisation was achieved using a freeze dryer (Virtis, Benchtop 6.6, The Virtis Company, Gardiner, USA) coupled to an Edwards vacuum pump for 96 h.

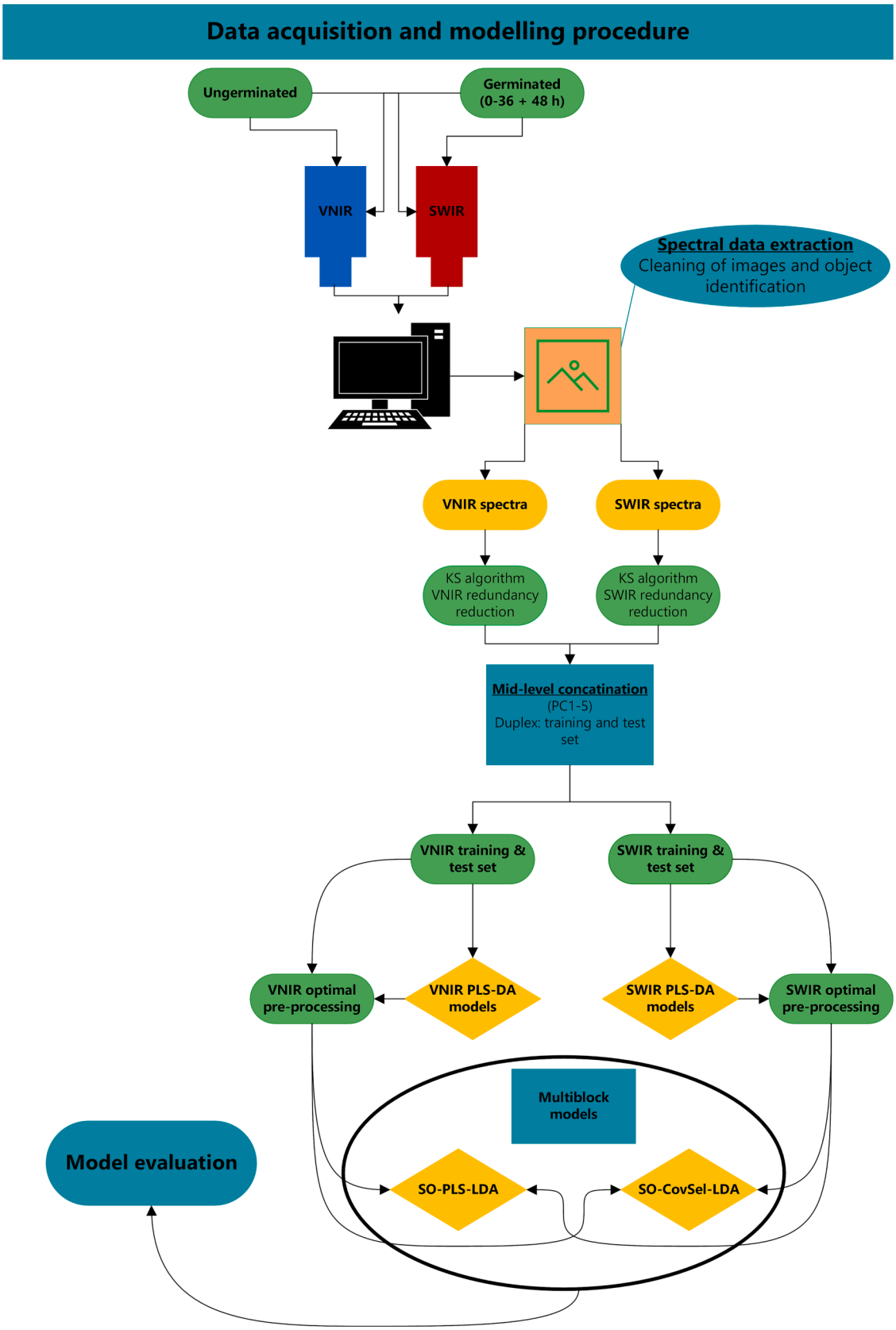

Fig. 1. A summary of the data acquisition and modelling strategy followed and used for this study, presented as schematic diagram.

2.2. Spectral imaging camera setup

The VNIR camera (HySpex VNIR-1800; Norsk Elektro Optikk, Norway), with a cooled and stabilised scientific grade Complementary Metal Oxide Semiconductor (CMOS) sensor, had a spectral range of 400–1000 nm, with 1800 spatial pixels and 186 spectral channels allowing for spectral resolution of 3.26 nm at 100 frames per second (FPS). The camera had a field-of-view (FOV) of 17° and a pixel FOV across and along the object of 0.16/0.32 mrad. The selected working distance was 0.3 m, allowing for a linear FOV of 86 mm and a pixel size of 0.05/0.1 mm. The lens was further fitted with a circular polariser to minimise spectral scattering reflected from the object. The camera system was equipped with a translation stage with a variable speed drive to allow for constant translation speed of the object past the sensor, which was set to travel 120 mm past the FOV of the sensor. Two linear direct current (DC) 150 W halogen light sources emitting light with a wavelength ranging from 400 to 2500 nm were used to illuminate the barley samples and to obtain the correctly optimised integration time for the camera sensor to detect the illuminated object of interest.

The SWIR camera (Hyspex SWIR-384; Norsk Elektro Optikk, Norway) system, with a cooled Mercury Cadmium Telluride (MCT) sensor, had a spectral range of 930 to 2500 nm. The camera had 384 spatial pixels with 288 spectral channels and allowed for a spectral resolution of 5.45 nm at 100 FPS. The FOV of the camera was 16° and the pixel FOV across and along the object was 0.73/0.73 mrad. The working distance chosen for this setup was 0.3 m allowing for a FOV of 84 mm and a pixel size of 0.22/0.22 mm. As with the VNIR camera setup, the lens was fitted with a circular polariser to minimise spectral scattering effects reflected from the object. The camera system was equipped with the same translation stage and the same light sources as used for the VNIR camera setup. A 50 % absorbance/reflectance external grey Zenith Allucore diffuse reflectance standard (SphereOptics GmbH, Germany) and internal dark reference (closing of camera shutter) were used to collect reference images intermittently every 30 min throughout imaging for both the VNIR and SWIR camera setups. Reference images were used for colorimetric and radiometric calibration.

2.3. Image acquisition and data processing

The selected 38 sub-samples per variety were imaged, in both VNIR and SWIR, prior to imbibition. These were used as the ungerminated class. Subsequently these samples were imbibed for the indicated time points (hourly 0 h to 36 h plus 48 h) as described earlier, preserved and imaged in both VNIR and SWIR. These were used as the germinated class.

Spectral images were collected in reflectance mode using the Breeze 2021.1 (Predictera AB, Umeå, Sweden) software package from all samples before and after imbibition. Spectral images were obtained with both the VNIR and SWIR camera systems. The images were converted from reflectance to pseudo-absorbance using the Breeze software package exported in Envi file format. Images were subsequently imported into the Evince 2.7.12 (Prediktera AB, Sweden) software package which was primarily used to remove background, spectral scattering, and regions of over absorbance from the images. This was achieved by evaluating principal component (PC) scores plots and images interactively. A region-of-interest (ROI) was selected from each individual kernel and the average spectrum of each ROI determined. Spectral data obtained for each kernel in the respective samples, for both the VNIR and SWIR images, were exported in MATLAB file format. To reduce data redundancy and to maintain maximum sample variance, 15 representative average spectra were selected with the Kennard-Stone algorithm from each sample using MATLAB software (version R2021a, MathWorks). This method captures the necessary and most important variance for each sample. This approach is similar to applying convex hull [22–23] or other data reduction strategies [24], subsequent to PCA. The 15 selected spectra of each sample, for both the VNIR and SWIR data sets, were concatenated into two data blocks resulting in a total of 5700 spectra each. Each VNIR and SWIR data block therefore comprised 5 barley varieties, two classes (germinated and ungerminated) which included 38 samples (time points) each from which 15 spectra were obtained. However, due irregular data that had to be removed and missing data, data set balancing was performed resulting in two balanced data sets of 5640 spectra each. Binary dummy classes were predefined for the two data sets.

2.4. Spectral pre-treatment

Different spectral pre-treatments were applied to the VNIR and SWIR spectral data sets using MATLAB 2021a software. The pre-treatments applied were mean centring (MC), standard normal variate (SNV), Savitzky-Golay first (2nd order interpolating polynomials and 19 points window; D1) and second derivative (3rd order interpolating polynomials and 19 points window; D2), SNV in combination with Savitzky-Golay first derivative (2nd order interpolating polynomials and 19 points window; SNV + D1) and SNV in combination with Savitzky-Golay second derivative (3rd order interpolating polynomials and 19 points window; SNV + D2). As a standard approach for NIR spectral data pre-treatment, MC was always used in combination with the other pre-treatment methods.

2.5. Selection of training and test sets

Training and test sets were selected using Duplex [25] on an augmented data matrix, resulting from the concatenation of both the SWIR and VNIR data sets, by means of an in house written function running under MATLAB 2021a environment. In particular, to achieve an unbiased selection, the Duplex algorithm was applied separately for each category, and a multiblock approach was subsequently used to account for the spectral and chemical variance detected by the SWIR and VNIR instruments, as follows. Each of the data blocks (SWIR and VNIR) was pre-processed using all the six pretreatments to be tested (MC, SNV, D1, D2, SNV + D1 and SNV + D2), resulting in twelve data matrices. Then, in each matrix, the spectra of the 15 kernels selected from the individual images were averaged so to ensure that, by applying the Duplex algorithm to these mean spectra, all the kernels from an image could be then included in the same subset (either training or test). PCA was then separately applied to each of the six differently preprocessed data matrices for each block, and the scores along the first five PCs for each of the six SWIR and VNIR matrices were concatenated row-wise. The Duplex algorithm was then applied to the resulting augmented matrix using a 30 % holdout threshold for the independent test set. The schematic in Fig. 2 illustrates the mid-level data augmentation approach followed. Using the indices obtained by the Duplex algorithm, the

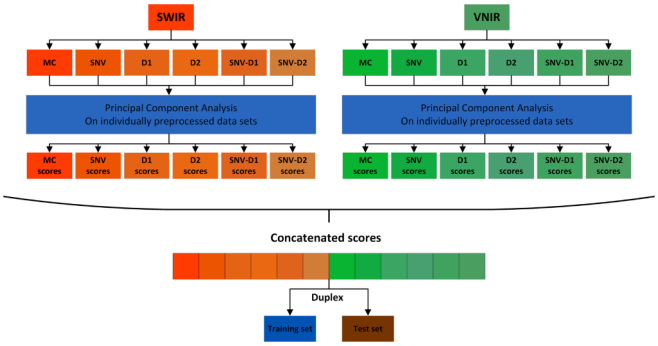

Fig. 2. Selection of a training and test sets by the Duplex algorithm using a mid-level data set augmentation. The schematic illustrates the process of applying PCA to the individual SWIR and NIR data blocks, differently pre-processed, augmenting the scores and lastly applying the Duplex algorithm to obtain the training and test sets.

selected training and test sets were extracted from the spectral data sets into two new data sets. The same training/test set splitting was used for both the PLS-DA modelling of the individual data blocks and the successive multi-block analysis to make results comparable.

2.6. Partial least squares-discriminant analysis (PLS-DA)

Partial least squares-discriminant analysis (PLS-DA) is probably the most popular discriminant classification technique used in chemometrics when dealing with ill-conditioned data matrices [26,27]. The reason for its widespread use is due to the algorithm's ability to deal with highly correlated variables, i.e., spectroscopic data. The technique exploits the possibility of coding the class assignment by means of a binary dummy response (which, for the training samples, takes the form of a vector **y** or a matrix **Y**, depending on whether two or more categories are involved) to turn a classification problem into a regression one [25], so that the PLS algorithm [28] can be used to fit the resulting model to overcome the issues posed by the predictor matrix being ill-conditioned. Considering a two-class problem such as the one in the present study, this can be shown mathematically by the linear equation (Eq. (1).

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \tag{1}$$

In this equation **X** is the (spectroscopic) data matrix collected on the samples, **b** is the vector of regression coefficients and **e** is the residuals, while **y** is the binary dummy vector encoding for the true category membership, whose elements are either 1, in correspondence to germinated samples, or 0 for ungerminated ones. Once the calibration model is set up, i.e., once the optimal value of the regression coefficients in Eq. (1) are estimated from the training data, any new measure ($\mathbf{x_{new}}$) can be classified by first calculating the values of the predicted response $\hat{y}_{\mathbf{new}}$, according to $\hat{y}_{\mathbf{new}} = \mathbf{x_{new}b}$. However, $\hat{y}_{new}$ is not categorical but real-valued so that a criterion for class-attribution is therefore needed to classify the new measures correctly. All the criteria proposed in the literature for the problems involving two classes, such as the one in the present study, are based on setting (implicitly or explicitly) a threshold to the predicted response, so that if $\hat{y}_{new}$ is higher than the threshold the sample is assigned to the class encoded as 1 (in our case, germinated samples) while if it is lower, it is predicted as belonging to the other category (the one encoded as 0; here, the ungerminated kernels). Given the binary coding, the most naïve approach is to set such threshold to 0.5, while, for instance, Perez et al. [29] proposed to couple Gaussian mixture modelling with Bayes' theorem, to translate the value of the predicted response into a posteriori probabilities of class belonging. In the present study, the classification threshold was estimated by applying linear discriminant analysis (LDA) on the predicted responses.

2.7. Sequential orthogonalized partial least-squares linear discriminant analysis (SO-PLS-LDA)

SO-PLS-LDA is a multiblock method following the basis of PLS for highly multicollinear data sets obtained from multiple sensors with recent application toward classification [30]. The technique combines linear discriminant analysis (LDA) [31] and a multiblock regression method sequential and orthogonalized- partial least squares (SO-PLS) [32–33]. By using SO-PLS to calculate a multiblock regression model between the predictor matrices and the dummy response, it is possible to fit and calculate LDA on a reduced set of features (either the scores or the predicted response). This method is especially useful when the variance/covariance structure of the data set is ill conditioned. A summary of the algorithm can be given for the case when two predictor blocks **X** and **Z** are used to predict the class membership in a classification problem involving two categories encoded by the dummy response vector **y**. This has been summarised by Biancolillo and Næs [33] and will be reiterated here in 5 steps:

a) A first PLS model between **X** and the dummy response **y** is calculated by PLS, resulting in a set of PLS regression coefficients (***b***), *X*-scores ($\boldsymbol{T}_X$) and *Y*-residuals ($\boldsymbol{e}_y$).
b) The second data block **Z** is orthogonalized with respect to the *X*-scores ($\boldsymbol{T}_X$) obtained in (a), resulting in $\mathbf{Z}_{\mathrm{Orth}}$. This step is used to remove the redundancies between the predictor blocks **X** and **Z**.
c) A second PLS model is fitted between $\mathbf{Z}_{\mathrm{Orth}}$ and the *Y*-residuals ($\boldsymbol{e}_y$). As for step (a), model parameters, such as the PLS regression coefficients for $\mathbf{Z}_{\mathrm{Orth}}$ (**c**) or the $\mathbf{Z}_{\mathrm{Orth}}$ scores ($\mathbf{T}_{\mathbf{Z}\mathrm{Orth}}$) are obtained.
d) The final predictive model can be calculated by summing the predictions of step (a) and (c) and by considering the regression equation $\hat{\mathbf{y}} = \mathbf{Xb} + \mathbf{Z}_{\mathrm{Orth}}\,\mathbf{c}$.
e) LDA is applied either to the predicted response $\hat{\mathbf{y}}$ (step (d)) or to the row-augmented scores ($\mathbf{T}_{\mathrm{SO}} = [\mathbf{T}_X \mathbf{T}_{\mathbf{Z}\mathrm{Orth}}]$)

As the method is based on sequential fitting of PLS models, the algorithm requires the selection of the optimal number of latent variables (LVs) to be extracted from each block. In the present study, the optimal model complexity was selected through a global strategy (i.e., testing all possible combinations up to a maximum specified number of components), as the one leading to the highest classification accuracy in a fivefold cross-validation. The results are usually graphically summarized in a so called Måge plot [35]. SO-PLS models were calculated by means of the MATLAB functions freely downloadable at: https://www.chem.uniroma1.it/romechemometrics/research/algorithms/so-pls/.

2.8. Sequential and orthogonalized covariance selection-linear discriminant analysis (SO-CovSel-LDA)

SO-CovSel-LDA [36] is an advanced multiblock variable selection method based on the premise of SO-PLS whilst incorporating CovSel [37]. Using the case of two-predictor block **X** and **Z** and a dummy response vector coding for two classes **y**, the algorithm can be summarised in 6 steps as iterated by Biancolillo et al. (2020) [36] and reiterated here:

a) Variables are selected by the CovSel algorithm from **X** – these variables are restructured and organised in a new matrix $\mathbf{X}_{\mathrm{sel}}$
b) y is then fitted to $\mathbf{X}_{\mathrm{sel}}$ by ordinary least squares (OLS)
c) $\mathbf{Z}_{\mathrm{Orth}}$ is obtained by orthogonalising **Z** with regard to $\mathbf{X}_{\mathrm{sel}}$
d) CovSel is then used to select variables in $\mathbf{Z}_{\mathrm{Orth}}$
e) The *Y*-residuals obtained in step (b) are then fitted to $\mathbf{Z}_{\mathrm{Orth}}$ by OLS
f) The full model is then calculated from the results obtained by steps (b) and (e).
g) Finally, classification is accomplished by applying LDA on the selected variables or on the predicted response analogously to what as described in the SO-PLS-LDA section.

Analogously, as already described for SO-PLS, selection of the optimal number of variables to be retained in each block can be carried out based on the results of a cross-validation procedure. In the present study, all possible combinations of the number of selected variables in each block (up to a specified maximum) were tested, and the one leading to the maximum accuracy in fivefold cross-validation was chosen as the optimal one. SO-CovSel-LDA models were fitted and validated by means of MATLAB functions freely downloadable at the following link: https://www.chem.uniroma1.it/romechemometrics/research/algorithms/so-covsel/.

3. Results and discussion

The spectral measurements (total as well as minimum, maximum and median) are shown in Fig. 3 for both the SWIR and VNIR data sets. The spectral data obtained was included in all the different classification models developed with PLS-DA, SO-PLS-LDA and SO-CovSel-LDA.

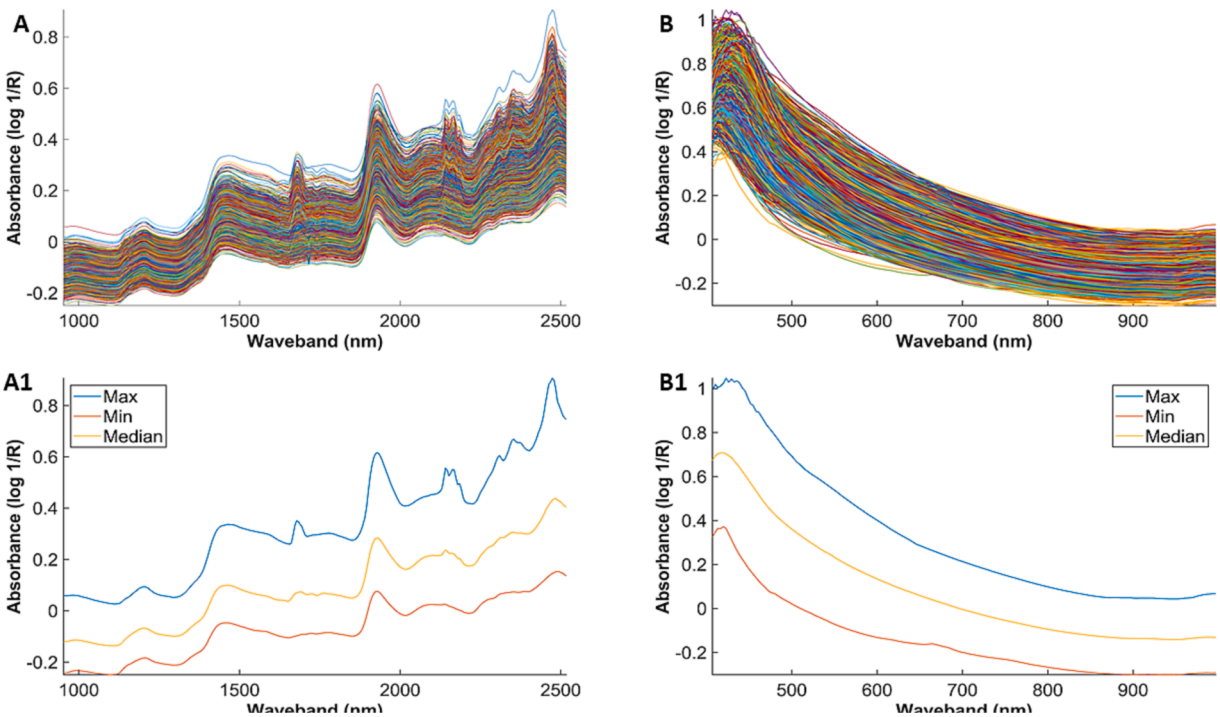

Fig. 3. Absorbance spectra obtained after the data set reduction procedure implemented using the Kennard-Stone algorithm, for (a) SWIR) and (b) (VNIR). The minimum, maximum and median spectra for (c) SWIR) and (d) (VNIR) determined from the reduced data sets.

3.1. Performance of PLS-DA classification models

3.1.1. SWIR PLS-DA calibration models

Classification models using PLS-DA as the classifier and the SWIR region as the variables showed good discriminant power, with cross-validated classification accuracy for all tested pre-processing techniques being above 99 % (Table 1). Considering the number of LVs, and the overall classification accuracy, sensitivity and specificity not only in cross-validation but also in calibration, as shown in Table 1, the model using only MC as spectral pre-treatment was taken as being the simplest and sufficiently accurate at distinguishing between the assigned classes. Moreover, the consistency between the results obtained in the calibration stage (e.g., with the optimal model applied to the training data) and in cross-validation suggests the absence of overfitting. Once the optimal pre-processing was selected based on the cross-validation results, the corresponding model (i.e., the one using MC only) was then validated using an independent test set comprising of 1695 spectra, each representing a barley kernel germinated to a different time point or not germinated at all. Test set classification accuracy (99.53 %), sensitivity (99.53 %) and specificity (99.52 %), as shown in Table 1, made for a highly selective and sensitive classification model using the SWIR waveband region.

The benefit of using the PLS algorithm is that multiple different plots can be obtained to visualise model performance and characteristics, and for interpretation. Fig. 4A and 4B are an example of such a plot. The variable importance in projection (VIP) plot [38] (Fig. 4A and 4B) shows the variables which contribute the most to the definition of the classification model, i.e., in this case, in differentiating between ungerminated and germinated barley seeds. The stability of VIP scores at high model complexity and how this influences interoperability is still

Table 1
Classification accuracy, sensitivity, specificity and confusion matrix for calibration, cross- validation and the independent test set results, obtained from the SWIR data set when PLS-DA was used.

Model	Pre-processing	LV	Calibration			
			Accuracy (%)	Sensitivity	Specificity	Confusion matrix
1	MC	18	99.85	100.00	99.69	[1995,0;6,1944]
2	SNV + MC	18	99.82	100.00	99.64	[1995,0;7,1943]
3	D1 + MC	17	99.70	99.95	99.44	[1994,1;11,1939]
4	D2 + MC	18	99.54	99.95	99.13	[1994,1;17,1933]
5	SNV + D1 + MC	16	99.62	99.90	99.33	[1993,2;13,1937]
6	SNV + D2 + MC	18	99.72	100.00	99.44	[1995,0;11,1939]
Model	**Pre-processing**	**LV**	**Cross validation: venetian blinds (5 folds)**			
			Accuracy (%)	**Sensitivity**	**Specificity**	**Confusion matrix**
1	MC	18	99.80	100.00	99.59	[1995,0;8,1942]
2	SNV + MC	18	99.70	99.95	99.44	[1994,1;11,1939]
3	D1 + MC	17	99.65	99.95	99.33	[1994,1;13,1937]
4	D2 + MC	18	99.57	99.95	99.18	[1994,1;16,1934]
5	SNV + D1 + MC	16	99.54	99.85	99.23	[1992,3;15,1935]
6	SNV + D2 + MC	18	99.67	100.00	99.33	[1995,0;13,1937]
Model	**Pre-processing**	**LV**	**Independent test set**			
			Accuracy (%)	**Sensitivity**	**Specificity**	**Confusion matrix**
1	MC	18	99.53	99.53	99.52	[851,4;4,836]

MC - mean centring; SNV - standard normal variate; D1 – first derivative (Savitzky-Golay); D2 – second derivative (Savitzky- Golay).

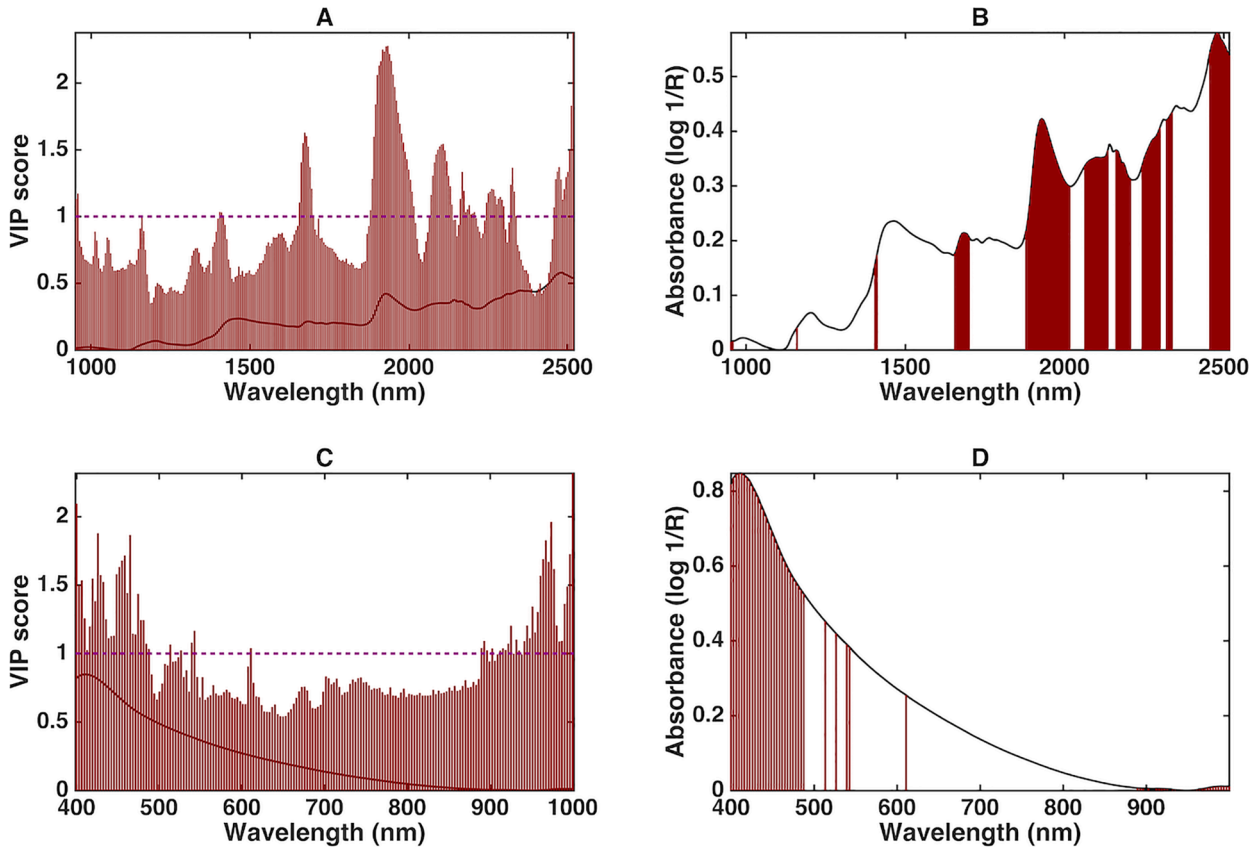

Fig. 4. Variables of importance as extracted by the PLS data transformation approach in a so-called variable importance in projection (VIP) plot for the (A) SWIR and (C) VNIR data sets used in PLS-DA. Variables with a score of above 1 are commonly accepted as being the most important toward the classification problem and can be further used in a variable selection approach. Panels (B) and (D) highlight the variables identified as relevant for the SWIR and VNIR data sets, respectively, on the average spectra of the training samples.

debated. A high number of LV's used is often associated with fitting more noise and artefacts. It is commonly agreed that more observations in comparison with variables result in more stable VIP scores. To confirm stability of VIP scores, evaluation of model over- or under fitting is important and needs to be taken into account during model development as elaborated on by Geladi and Kowalski [39]. Fig. 4 is easily interpretable due to minimum noise being fitted and easy identification of the most important variables contributing to the model.

As can be seen in Fig. 4A and 4B, the results suggest that classification is mostly based on water, cellulose, starch and protein damage. Indeed, since a greater than one criterion is assumed for significance, it is evident how the wavelength ranges contributing the most to the PLS model are 1600–1900 nm (starch and cellulose), 1940–2050 nm (water and protein) and 2100–2500 nm indicative of a combination of protein, water, cellulose and starch overtone and combination bands [17,40–43]. Specifically, the PLS-DA technique shows good potential to

Table 2
Classification accuracy, sensitivity, specificity and the confusion matrix for calibration, cross-validation and the independent test set results, obtained from the VNIR data set when using PLS-DA.

Model	Pre-processing	LV	Calibration			
			Accuracy %	Sensitivity	Specificity	Confusion matrix
1	MC	18	98.91	99.85	97.95	[1992,3;40,1910]
2	SNV + MC	18	98.50	99.60	97.38	[1987,8;51,1899]
3	D1 + MC	17	97.47	99.05	95.85	[1976,19;81,1869]
4	D2 + MC	18	98.05	99.50	96.56	[1985,10;67,1883]
5	SNV + D1 + MC	18	98.00	98.95	97.03	[1974,21;58,1892]
6	SNV + D2 + MC	18	98.33	99.50	97.13	[1985,10;56,1894]
Model	**Pre-processing**	**LV**	**Cross validation: venetian blinds (5 folds)**			
			Accuracy %	**Sensitivity**	**Specificity**	**Confusion matrix**
1	MC	18	98.73	99.90	97.54	[1993,2;48,1902]
2	SNV + MC	18	98.33	99.45	97.18	[1984,11;55,1895]
3	D1 + MC	17	97.24	98.90	95.54	[1973,22;87,1863]
4	D2 + MC	18	97.77	99.45	96.05	[1984,11;77,1873]
5	SNV + D1 + MC	18	97.74	98.90	96.56	[1973,22;67,1883]
6	SNV + D2 + MC	18	98.07	99.35	96.77	[1982,13;63,1887]
Model	**Pre-processing**	**LV**	**Independent test set**			
			Accuracy %	**Sensitivity**	**Specificity**	**Confusion matrix**
1	MC	18	98.70	99.77	97.62	[853,2;20,820]

MC - mean centring; SNV - standard normal variate; D1 – first derivative (Savitzky-Golay); D2 – second derivative (Savitzky- Golay).

be used as a classification method to determine if barley seed has undergone pre-harvest sprouting or not.

3.1.2. VNIR PLS-DA calibration models

When PLS-DA analysis was conducted on the VNIR data, comparably good results were obtained, though slightly less accurate. In particular, when looking at the cross-validation outcomes of the candidate models built on the differently pre-processed data (Table 2), it can be observed how classification accuracy was always higher than 97 % but lower than 99 %. Here too MC was a sufficient spectral pre-treatment method for a model with minimal mathematical strain to obtain accurate classification, resulting in an overall correct classification rate of 98.91 % in calibration and 98.73 % in cross-validation (Table 2). When the optimal model was applied to the independent test set for the external validation step, a classification accuracy of 98.70 % was obtained. These results give a clear indication that the VNIR region can be used within a 97 % confidence interval framework to classify between ungerminated and germinated barley seed. It was of interest to note that the sensitivity of all the models was slightly better than the specificity. This is acceptable as the class 'ungerminated' is simply being incorrectly assigned as 'germinated', which in turn does not influence the outcome of the classification model as the objective was to correctly classify for germinated, relating to model sensitivity.

The VIP plot (Fig. 4C and 4D) obtained for the PLS-DA model using MC applied to the VNIR wavebands shows that variables of interest for making the correct classification were colour in the visible region of the electromagnetic spectrum and also protein (910 nm) and starch (990 nm) [41,43]. The colour region is of interest since classification could be due to the barley seed visually having a change of colour as the radicle emerges, or even as the plumule starts to develop. The biochemical changes regarding the protein and starch structure of the endosperm and germ end are as for the SWIR region and were to be the expected changes that would be detected spectroscopically. It could be hypothesised that the slightly lower classification accuracy obtained when using the VNIR spectral region vs the SWIR region could be due to the lack in detecting a change in the copious amount of cellulose which forms part of the seed carapace interlaced and woven with linear proteins [44].

3.2. Performance of multiblock classification models

3.2.1. SO-PLS-LDA calibration models

It is to be expected that a multiblock method such as SO-PLS-LDA could lead to better classification accuracy when compared to using the SWIR or VNIR data sets individually. This is because complementary information is extracted from both of the data blocks, enabling the class assignment to be made more easily [45]. This is akin to the human senses which all play a role in making cognitive decisions, so too does SO-PLS-LDA. Up to 99.85 % classification accuracy was obtained when using this method (Table 3), however, due to the high classification accuracy obtained when using the blocks individually and applying PLS-DA this was to be expected (Tables 1 & 2).

Multiple combinations of spectral pre-treatment were investigated in the model selection stage and the best one was chosen as that leading to the maximum classification accuracy in a fivefold cross-validation approach. The results obtained for the SWIR (Table 1) and VNIR (Table 2) PLS-DA models suggest that MC on the SWIR and VNIR data sets would be the most mathematically simple spectral pre-treatment method to obtain adequate classification accuracy, also applying the same pre-processing prior to using SO-PLS-LDA should result in good calibration accuracy. This hypothesis was confirmed by the outcomes of the model selection stage (summarized in Table 3). Indeed, although all the tested combinations of pre-treatments resulted in a cross-validation accuracy higher than 99.5 %, the highest discriminant ability (with 99.85 % and 99.65 % specificity in calibration and cross-validation, respectively, and 100 % sensitivity in both stages) was registered for the model calculated on both data sets pre-treated with mean centring only. These results were achieved by using 17 SWIR and 8 VNIR LVs from a calibration data set of 3948 SWIR and VNIR spectra each representing a single barley kernel. The maximum number of LVs used over all the models was 17 from SWIR and 13 from VNIR. In contrast, when SNV or SNV and second derivative were applied to the two data sets, 9 SWIR and 9 VNIR LVs were used, respectively, which represent the minimum complexity of the calculated models (Table 3). It should be noted that, in general, a relatively high number of latent variables are selected both for the individual PLS-DA and for the multi-block SO-PLS-LDA models, but this can be explained by the vast number of spectral observations which were analysed and by the wide range of sources of variabilities which were considered in the design. This can also be visually illustrated by observing the PC scores images of the VNIR and SWIR that were germinated at 0 h and 36 h (Fig. 5). No visual sign of chemical changes can be observed in the first two PCs (approximately 98 % variance) even after 36 h of imbibition. This is a preliminary indication of the complexity of the data and of the subsequent developed models.

When applied to the independent test set, the best model, i.e., the one built on simply mean-centred blocks was highly sensitive and had no false negative predictions, it was also very selective towards the classification problem and only 2 out of the 1695 test set samples were

Table 3
Classification accuracy, sensitivity, specificity and confusion matrix for calibration, cross- validation and independent test set results, obtained from the SWIR and VNIR data sets when SO-PLS-LDA was used.

Model	Pre-processing	LV	Calibration			
			Accuracy	Sensitivity	Specificity	Confusion matrix
1	{MC};{MC}	[17,8]	99.92	100	99.85	[1995,0;3,1947]
2	{SNV + MC};{SNV + MC}	[9,9]	99.82	100	99.64	[1995,0;7,1943]
3	{D1 + MC};{D1 + MC}	[15,13]	99.82	99.95	99.69	[1994,1;6,1944]
4	{D2 + MC};{D2 + MC}	[11,10]	99.80	100	99.59	[1995,0;8,1942]
5	{SNV + D1 + MC};{SNV + D1 + MC}	[12,10]	99.65	99.95	99.33	[1994,1;13,1937]
6	{SNV + D2 + MC};{SNV + D2 + MC};	[9,9]	99.77	100	99.54	[1995,0;9,1941]
Model	**Pre-processing**	**LV**	**5-Fold cross-validation**			
			Accuracy	**Sensitivity**	**Specificity**	**Confusion matrix**
1	{MC};{MC}	[17,8]	99.85	100	99.69	[1995,0;6,1944]
2	{SNV + MC};{SNV + MC}	[9,9]	99.70	99.95	99.44	[1994,1;11,1939]
3	{D1 + MC};{D1 + MC}	[15,13]	99.77	99.90	99.64	[1993,2;7,1943]
4	{D2 + MC};{D2 + MC}	[11,10]	99.67	99.95	99.38	[1994,1;12,1938]
5	{SNV + D1 + MC};{SNV + D1 + MC}	[12,10]	99.62	99.90	99.33	[1993,2;13,1937]
6	{SNV + D2 + MC};{SNV + D2 + MC};	[9,9]	99.65	99.38	99.90	[1993,2;12,1938]
Model	**Pre-processing**	**LV**	**Independent test set**			
			Accuracy	**Sensitivity**	**Specificity**	**Confusion matrix**
1	{MC};{MC}	[17,8]	99.88	100	99.76	[855,0;2,838]

MC - mean centring; SNV - standard normal variate; D1 – first derivative (Savitzky-Golay); D2 – second derivative (Savitzky- Golay).

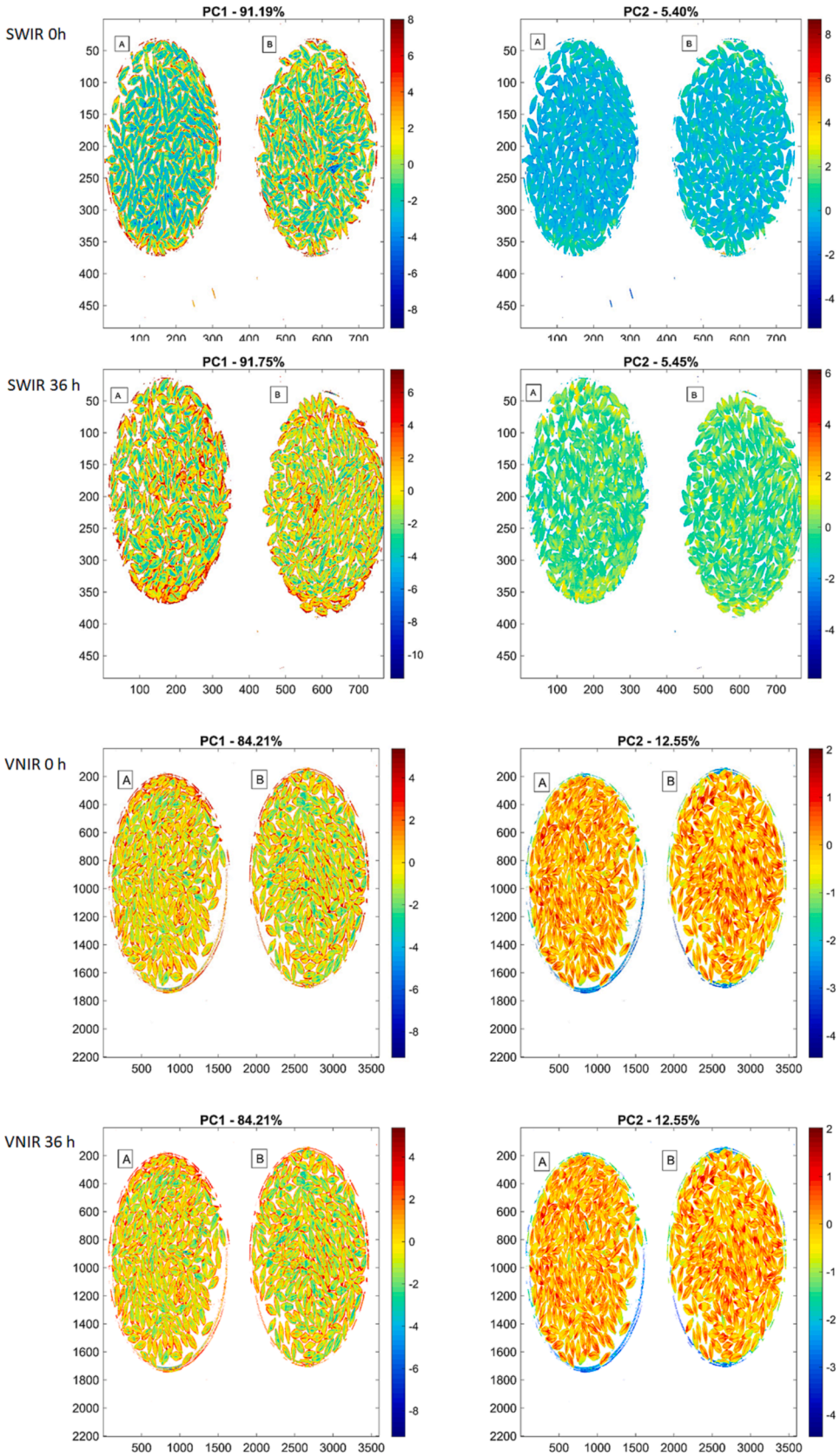

Fig. 5. PC scores images of barley grain that has been (A) germinated and (B) ungerminated captured by SWIR and VNIR HSI cameras. The example samples are from the Genie variety and show a scores image of PC 1 and 2 that has not been germinated at 0 h and 36 h for SWIR and VNIR images respectively.

incorrectly classified as being germinated when the true class was ungerminated. These outcomes correspond to 100 % sensitivity, 99.76 % specificity and, consequently, 99.88 % accuracy. Using the waveband regions in both SWIR and VNIR together with a multiblock approach shows that very early-stage pre-harvest germination can be classified with confidence – this represents a critical focus point as conventional methods are either not sensitive enough, suffer from operator bias or require a skilled laboratory technician to perform the analysis.

3.2.2. SO-CovSel-LDA calibration models

Multispectral approaches are beneficial for the food manufacturing and agricultural sectors as they bring about the possibility of applying the technique directly in an in-line or on-line system. In this context, a variable selection and data reduction technique can only be considered a suitable solution when the number of variables selected allow classification accuracy within the confidence limits of conventional methodology. Based on these considerations, in the last stage of our study, the possibility of achieving accurate classifications at the same time including a limited number of wavelengths in the model was evaluated using a recently proposed multi-block variable selection approach, namely SO-CovSel-LDA [36]. Also in this case, different data pre-treatments were tested and the optimal ones were selected as those leading to the highest classification accuracy in 5-fold cross-validation. The results are shown in Table 4, where it is immediately obvious how, despite the great reduction in the number of total variables, the classification accuracy remains comparable to that obtained on the full-spectrum data sets. Indeed, calibration and cross-validation accuracy was above 99 % for all the models with different spectral pre-treatment techniques being used on both the SWIR and VNIR data blocks. The results reported in Table 4 suggest that the best outcomes can be obtained using only mean-centring (with 21 selected variables, 18 from SWIR and 3 from VNIR). This is also promising in the light of a possible real-world application (e.g., on filter instruments, where applying pre-processing such as SNV or differentiation could not be as efficient as with the full spectrum, and possibly create artifacts).Therefore, we decided to select the model built on mean-centred blocks as the final one, which was then applied to the test set for external validation. The corresponding accuracy was 98.76 %, with a sensitivity of 99.53 % and a specificity of 97.98 %, resulting from incorrectly classifying 4 samples as false negatives and 17 as false positives with the number of wavebands reduced to 18 for SWIR and 3 for VNIR. From an original 288 variables for the SWIR region and 186 for the VNIR region, only 6.2 % of the original SWIR and 1.6 % of the VNIR wavebands were used to obtain a similar classification accuracy when using PLS-DA and SO-PLS-LDA The selected wavebands used in the selected model are shown in Fig. 6A (SWIR) and 6B (VNIR). For the SWIR region, the selected wavebands cover almost the entire spectral interval, together with the extreme points which are often chosen so to compensate for additive/multiplicative effects [36]. Selected wavebands include the one at 1056 nm which could be assigned to second overtone of CH_2 bonds and the N—H stretch, and those at 1143 and 1323 nm, which can be ascribed to the second overtone of aromatic C—H stretches. For the band at 1677 nm, assignment is normally made to the first overtone of the stretching vibrations of more aromatic C—H bonds, while that at 1890 nm can be ascribed to O—H stretches and C—O stretches of sstarch molecules The selected variable of 1928 nm corresponds to second overtone stretches assigned to CONH bonds and water, while 1977 nm is assigned to $CONH_2$ and asymmetric N—H stretches of proteins. The assignment of the signals at 2108, 2141 and 2168 nm can be made to starch O—H and C—O stretching, alkene stretching and amide interactions of the form CONHR. The band selected at 2293 nm is assigned to amino acid with N—H and carbonyl group stretching, and the bands at 2326 nm is assigned to C—H terminal stretching in the form CH_2. The last assignment for 2473 nm is C—H and C—C stretching of starch molecules [31,46]. With reference to the VIP plots obtained for PLS-DA classification models (Fig. 4A and 4B SWIR and 4C and 4D VNIR), using the SWIR waveband region there is some information overlap and that areas of importance extracted by the two methods are similar. The first two wavelengths selected in the VNIR region, 445 nm and 547 nm, are both assigned to colour, with 445 nm being more towards the blue spectrum and 547 nm within the yellow region of the spectrum – both of these are primary colours which in combination makes green. Chlorophyll, the green pigment of plants, also shows strong absorbance in the 445 nm waveband region. On the other hand, the third one, 1000 nm, is probably selected to implicitly account for baseline correction/normalization. These results suggest that the developing plumule of germinating barley grain, which is green in colour due to chlorophyll, may be the driver for selection of these specific wavebands in the VNIR region. This is in agreement with a study conducted by Nakaji et el. [47], where classification of growing (live) and non-growing (dead) rhizosphere components where classified using VNIR images.

Further waveband reduction was implemented on the spectral data

Table 4
Number of reduced variables, classification accuracy, sensitivity, specificity and confusion matrix for calibration, cross-validation and the independent test set results, obtained from the SWIR and VNIR data sets when SO-CovSel-LDA was used.

Model	Pre-processing	Reduced variables	Calibration Accuracy	Sensitivity	Specificity	Confusion matrix
1	{MC};{MC}	[18;3]	99.29	99.85	98.72	[1992,3;25,1925]
2	{SNV + MC};{SNV + MC}	[13;4]	98.66	99.70	97.59	[1989,6;47,1903]
3	{D1 + MC};{D1 + MC}	[13;7]	99.21	99.85	98.56	[1992,3;28,1922]
4	{D2 + MC};{D2 + MC}	[13;7]	99.06	99.95	98.15	[1994,1;36,1914]
5	{SNV + D1 + MC};{SNV + D1 + MC}	[13;6]	99.11	99.90	98.31	[1993,2;33,1917]
6	{SNV + D2 + MC};{SNV + D2 + MC};	[13;7]	99.26	99.95	98.56	[1994,1;28,1922]
Model	**Pre-processing**	**Reduced variables**	**5-Fold cross-validation**			
			Accuracy	**Sensitivity**	**Specificity**	**Confusion matrix**
1	{MC};{MC}	[18;3]	99.21	99.80	98.62	[1991,4;27,1923]
2	{SNV + MC};{SNV + MC}	[13;4]	98.88	99.75	98.00	[1990,5;39,1911]
3	{D1 + MC};{D1 + MC}	[13;7]	98.63	99.60	97.64	[1987,8;46,1904]
4	{D2 + MC};{D2 + MC}	[13;7]	98.91	99.90	97.90	[1993,2;41,1909]
5	{SNV + D1 + MC};{SNV + D1 + MC}	[13;6]	98.38	99.75	96.97	[1990,5;59,1891]
6	{SNV + D2 + MC};{SNV + D2 + MC};	[13;7]	99.11	99.90	98.31	[1993,2;33,1917]
Model	**Pre-processing**	**Reduced variables**	**Test set**			
			Accuracy	**Sensitivity**	**Specificity**	**Confusion matrix**
1	{MC};{MC}	[18;3]	98.76	99.53	97.98	[851,4;17,823]
1.1	{MC};{MC}	[10;2]	97.29	99.42	95.12	[850,5;41,799]
1.2	{MC};{MC}	[8;5]	97.46	99.53	95.36	[851,4;39,801]
1.3	{MC};{MC}	[5;5]	92.92	98.71	87.02	[844,11;109,731]

MC - mean centring; SNV - standard normal variate; D1 – first derivative (Savitzky-Golay); D2 – second derivative (Savitzky- Golay).

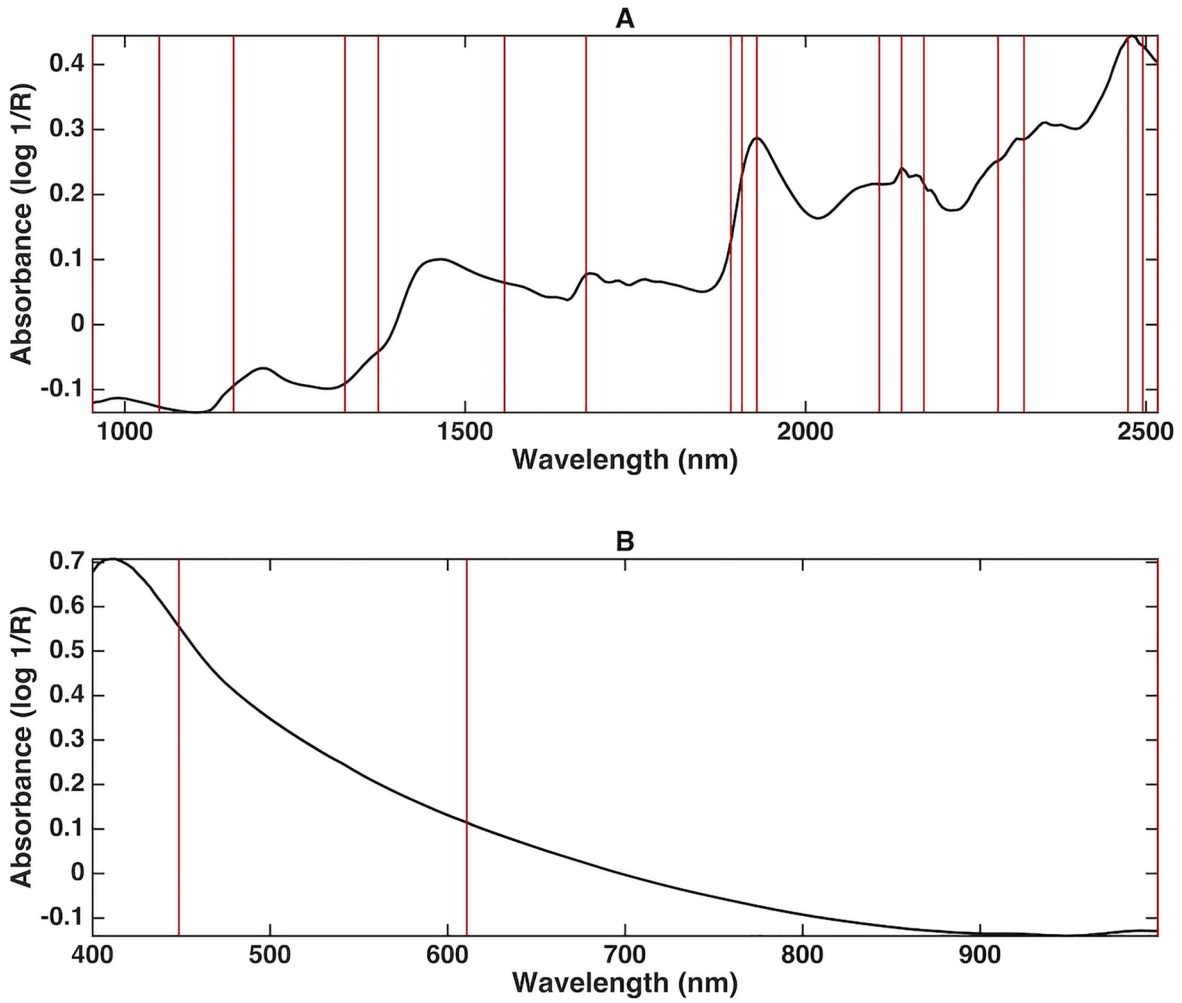

Fig. 6. Variables selected using the multiblock SO-CovSel-LDA procedure for (A) SWIR and (B) VNIR data blocks.

sets that were only pre-treated with MC to obtain the minimum number of wavebands which can be used from the SWIR and VNIR regions while still achieving a classification accuracy above 95 %. The threshold was set to use a maximum of 10, 8 and 5 wavebands from each data block, the results of which are shown in Table 4. Using a maximum of 10 wavebands, from both the SWIR and VNIR data sets, achieved a test set classification accuracy of 97.29 % with a sensitivity of 99.42 % and specificity of 95.12 %, classifying 5 false negatives and 41 false positives. The total number of variables used for this classification was 10 from the SWIR and 2 from the VNIR region. When the model was limited to include a maximum of 8 variables per block, 8 SWIR and 5 VNIR variables were used to obtain a test set classification accuracy of 97.46 % % with a sensitivity of 99.53 % and specificity of 95.36 %. The number of kernels classified as false negatives was 4 and false positives were 39 out of a total of 1695 representative kernels. Limiting to 5 useable variables from each data block (SWIR and VNIR) resulted in a test set classification accuracy of 92.92 %, with sensitivity of 98.71 % and specificity of 87.02 % using 5 SWIR and 5 VNIR variables. Having decided to consider a 95 % threshold as the minimum accuracy for a model to be acceptable for practical use, with the scope of classifying between ungerminated and germinated barley grain, this last model, though still presenting a good sensitivity, resulted unsuitable. The true benefit of a multispectral imaging approach is the significant reduction in sensor cost and the gained benefit of implementation of such a system for real time monitoring of industry problems such as barley pre-harvest germination classification [48]. The benefit of such technology is further exploited by a variable selection procedure selecting variables (wavebands) of importance from two unique imaging systems in the VNIR and SWIR regions. This can ultimately lead to the design of industrial grain grading equipment using only light sources of specific waveband which have been identified by this dual imaging sensor approach.

Average spectra (object-wise) were used for more efficient model development from a data set comprising 760 images. The method used in this study thus offers a rapid way to obtain and retain reliable data from multiple samples and to capture intra- and inter-sample variability more easily. Furthermore, using a robust multiblock variable selection strategy, enables selection of important variables (wavelengths) from both imaging systems. This will enable one to determine the need for variables in either or both the VNIR and SWIR regions to develop for example a single multispectral system.

Moving toward agri-industrial application for precision farming and rapid grading, the foundation work with respect to variable reduction through waveband selection and, ultimately, the proof of concept has been shown. Multiple platforms could be used to accommodate an instrument, these being in field grading by combine harvesters, during harvest, or post-harvest grading at barley storage facilities and malting facilities. However, for implementation on these platforms, a multispectral instrument using the specific wavebands indicated in this study will have to be designed and built to specifications tailored to the housing platform. A possible solution will be to use light emitting diodes, of specific wavelength, arranged in a linear series, similar to current high throughput food and agricultural product grading and sorting instruments tailored to cereal grain. Indeed, an array of light filters

where only specific light is allowed to reach the sensor, much like a Red Green Blue (RGB) mosaic filter used in conventional cameras, can also be designed and used. Unsurprisingly, an instrument with the capacity to function in agricultural environments will have to be built in a robust manner. Thus such an instrument will have to include preventative measures to counteract environmental artefacts such as dust and other foreign objects, whilst maintaining selective targeting toward barley grain. Furthermore, instrument recalibration will have to commence annually prior to harvest, with data captured from the southern hemisphere supporting that of northern hemisphere harvests and vice versa. Consequently due to natural sample variation, model calibration and recalibration, by capturing data from multiple seasons and geographical growing locations, will be fundamental to ensure that the most robust model for precise concurrent harvest decisions and post-harvest grading is used.

4. Conclusion

The proposed methods, investigated to classify pre-harvest germination in malting barley, showed good results with classification accuracies of above 99 % being obtained in all objectives. Using NIR-HSI systems in the SWIR and VNIR waveband regions shows that it is an analytical tool sufficiently sensitive to be used in solving classification problems when suitable chemometric methods are applied to spectral data. Using PLS-DA to build classification models, in the SWIR and VNIR waveband regions individually, good accuracy was achieved when using only the standard MC spectral pre-treatment technique. Allowing for a novel multiblock approach showed that an increase in classification accuracy can be obtained when the information from multiple sensors are used to address a global agriculture problem. SO-PLS-LDA and SO-CovSel-LDA, using the fused SWIR and VNIR data blocks, achieved good classification accuracy. These results also demonstrated that industrial application is possible. Using the multiblock variable selection procedure, SO-CovSel-LDA, the number of wavebands to be included in the model could be reduced to 8 and 5 for SWIR and VNIR, which corresponds to using only 2.8 % and 2.7 % of the original SWIR and VNIR. This still allows a classification accuracy higher than 97 % on an independent test set to be obtained. This further proves the concept that an online multispectral instrument can be built using the fundamental information obtained in this study. Such an instrument or grading system using a multispectral approach will allow for rapid throughput of barley grain at malthouses and silos and allow for the assessment of degree of pre-harvest sprouting and ultimately barley seed viability. With commercial examples of optical sorters being able to sort grain at up to 40 tonnes/h per processing line, imaging systems such as the one proposed have the added advantage of being able to detect, classify and sort grain based on other characteristics (not considered in this study but well researched) such as protein and moisture content, fungal contamination (e.g., due to fusarium and ergot) and grain total friability. The financial benefit of such a system is thus justified for the use of spectral imaging in the vis/NIR regions in the beer brewing and malting sector. This will enable barley farmers to gain unbiased grading information with regard to their crops and harvests, allowing them to make data driven decisions as to the quality of the grain they are selling. To the best of our knowledge no similar work has been performed with regards to using NIR-HSI coupled to SO-PLS-DA and SO-CovSel-LDA to classify between germinated and ungerminated barley seed using two instruments and separate sensors in the SWIR and VNIR waveband regions, respectively. The SO-PLS-LDA and SO-CovSel-LDA methods have, up until now, also not been implemented toward any classification problem using data obtained from two NIR-HSI instruments. The ground-breaking nature of this study makes for a truly novel approach to address an industry problem.

CRediT authorship contribution statement

Sebastian Helmut Orth: Methodology, Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing. **Federico Marini:** Supervision, Conceptualization, Software, Resources, Writing – review & editing. **Glen Patrick Fox:** Supervision, Conceptualization, Resources, Writing – review & editing. **Marena Manley:** Supervision, Conceptualization, Resources, Writing – review & editing. **Stefan Hayward:** Supervision, Conceptualization, Resources, Writing – review & editing, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

All samples used in this study was donated by the South African Barley Breeding Institute (SABBI) in Caledon, Western Cape, South Africa. The authors would like to thank Dr. Nikki Else, Mr. Cobus Berner and Mrs. Marna Esterhuizen from ABinBev for their assistance during this study.

Funding

This work was supported financially by ABinBev Africa Pty Ltd.

References

[1] F. Ghahremaninejad, E. Hoseini, S. Jalali, The cultivation and domestication of wheat and barley in Iran, brief review of a long history, Bot. Rev. 87 (2021) 1–22, https://doi.org/10.1007/S12229-020-09244-W/FIGURES/6.
[2] A.C. Newton, A.J. Flavell, T.S. George, P. Leat, B. Mullholland, L. Ramsay, C. Revoredo-Giha, J. Russell, B.J. Steffenson, J.S. Swanston, W.T.B. Thomas, R. Waugh, P.J. White, I.J. Bingham, Crops that feed the world 4. Barley: a resilient crop? Strengths and weaknesses in the context of food security, Food Secur. 32 (3) (2011) 141–178, https://doi.org/10.1007/S12571-011-0126-3.
[3] FAOSTAT, FAOSTAT Statistical Database, (2019). http://www.fao.org/faostat/en/#data/QC/visualize (accessed March 14, 2021).
[4] B. Contreras-Jiménez, A. Del Real, B.M. Millan-Malo, M. Gaytán-Martínez, E. Morales-Sánchez, M.E. Rodríguez-García, Physicochemical changes in barley starch during malting, J. Inst. Brew. 125 (2019) 10–17, https://doi.org/10.1002/JIB.547.
[5] R. No, 443, Regulations relating to the grading, packing and marking of malting barley intended for sale in the Republic of South Africa, Government Notice Department of Agriculture, Forestry and Fisheries, 2013 https://www.gov.za/documents/agricultural-product-standards-act-regulations-grading-packing-and-marking-malting-barley.
[6] F. Gubler, A.A. Millar, J.V. Jacobsen, Dormancy release, ABA and pre-harvest sprouting, Curr. Opin. Plant Biol. 8 (2005) 183–187, https://doi.org/10.1016/j.pbi.2005.01.011.
[7] S.S. Chen, J.L. Chang, Does gibberellic acid stimulate seed germination via amylase synthesis? Plant Physiol. 49 (1972) 441–442, https://doi.org/10.1104/pp.49.3.441.
[8] R. Lin, R.D. Horsley, P.B. Schwarz, Associations between caryopsis dormancy, α-amylase activity, and pre-harvest sprouting in barley, J. Cereal Sci. 48 (2008) 446–456, https://doi.org/10.1016/j.jcs.2007.10.009.
[9] N.A. Gualano, R.L. Benech-Arnold, Predicting pre-harvest sprouting susceptibility in barley: Looking for "sensitivity windows" to temperature throughout grain filling in various commercial cultivars, F, Crop. Res. 114 (2009) 35–44, https://doi.org/10.1016/j.fcr.2009.06.016.
[10] W.T. Buckley, M.S. Izydorczyk, W.G. Legge, Detection of incipient germination in malting barley with a starch viscosity method and a proposed ethanol emission method, Cereal Chem. 93 (2016) 450–455, https://doi.org/10.1094/CCHEM-07-15-0147-R.
[11] P. Schopfer, D. Bajracharya, C. Plachy, Control of seed germination by abscisic acid, Plant Physiol. 64 (5) (1979) 822–827.
[12] R.L. Benech-arnold, R.A. Sánchez, Handbook of seed physiology: applications to agriculture, Choice Rev, Online. 42 (2005).

[13] D.J. Mares, K. Mrva, Wheat grain preharvest sprouting and late maturity alpha-amylase, Planta. 240 (2014) 1167–1178, https://doi.org/10.1007/S00425-014-2172-5.
[14] V.A. McKie, B.V. McCleary, A rapid, automated method for measuring α-amylase in pre-harvest sprouted (sprout damaged) wheat, J. Cereal Sci. 64 (2015) 70–75, https://doi.org/10.1016/J.JCS.2015.04.009.
[15] O.A.H. Jones, Assessing pre-harvest sprouting in cereals using near-infrared spectroscopy-based metabolomics, NIR news. 28 (2017) 15–19, https://doi.org/10.1177/0960336016687945.
[16] S. Grassi, G. Cardone, D. Bigagnoli, A. Marti, Monitoring the sprouting process of wheat by non-conventional approaches, J. Cereal Sci. 83 (2018) 180–187, https://doi.org/10.1016/J.JCS.2018.08.007.
[17] C.M. McGoverin, P. Engelbrecht, P. Geladi, M. Manley, Characterisation of non-viable whole barley, wheat and sorghum grains using near-infrared hyperspectral data and chemometrics, Anal. Bioanal. Chem. 401 (2011) 2283–2289, https://doi.org/10.1007/s00216-011-5291-x.
[18] J.G.A. Barbedo, E.M. Guarienti, C.S. Tibola, Detection of sprout damage in wheat kernels using NIR hyperspectral imaging, Biosyst. Eng. 175 (2018) 124–132, https://doi.org/10.1016/j.biosystemseng.2018.09.012.
[19] M. Arngren, P. Waaben Hansen, B. Eriksen, J. Larsen, R. Larsen, Analysis of Pregerminated Barley Using Hyperspectral Image Analysis, J. Agric. Food Chem. 59 (2011) 11385–11394, https://doi.org/10.1021/jf202122y.
[20] G. Fox, M. Manley, Applications of single kernel conventional and hyperspectral imaging near infrared spectroscopy in cereals, J. Sci. Food Agric. 94 (2014) 174–179, https://doi.org/10.1002/jsfa.6367.
[21] G. Fox, The brewing industry and the opportunities for real-time quality analysis using infrared spectroscopy, Appl. Sci. 10 (2020) 1–12, https://doi.org/10.3390/app10020616.
[22] C. Ruckebusch, R. Vitale, M. Ghaffaria, S. Hugelier, N. Omidikia, Perspective on essential information in multivariate curve resolution, TrAC Trends in Analytical Chemistry 132 (2020), 116044, https://doi.org/10.1016/j.trac.2020.116044.
[23] M. Ghaffari, N. Omidikia, C. Ruckebusch, Joint selection of essential pixels and essential variables across hyperspectral images, Analytica Chimica Acta 1141 (2021) 36–46, https://doi.org/10.1016/j.aca.2020.10.040.
[24] C. Ferraria, G. Focaa, A. Ulrici, Handling large datasets of hyperspectral images: Reducing data size without loss of useful information, Analytica Chimica Acta 802 (2013) 29–39, https://doi.org/10.1016/j.aca.2013.10.009.
[25] R.D. Snee, Validation of Regression Models: Methods and Examples, Technometrics. 19 (1977) 415–428, https://doi.org/10.1080/00401706.1977.10489581.
[26] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166–173, https://doi.org/10.1002/cem.785.
[27] L. Stahle, S. Wold, Partial least squares analysis with cross-validation for the two class problem: A Monte Carlo study, J. Chemom. 1 (1987) 185–196, https://doi.org/10.1002/cem.1180010306.
[28] S. Wold, H. Martens, H. Wold, The multivariate calibration problem in chemistry solved by the PLS method (1983) 286–293, https://doi.org/10.1007/BFB0062108.
[29] N.F. Pérez, J. Ferré, R. Boqué, Calculation of the reliability of classification in discriminant partial least-squares binary classification, Chemometr. Intell. Lab. Syst. 95 (2009) 122–128, https://doi.org/10.1016/j.chemolab.2008.09.005.
[30] A. Biancolillo, I. Måge, T. Næs, Combining SO-PLS and linear discriminant analysis for multi-block classification, Chemom. Intell. Lab. Syst. 141 (2015) 58–67, https://doi.org/10.1016/j.chemolab.2014.12.001.
[31] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (1936) 179–188, https://doi.org/10.1088/0004-6256/139/6/2200.
[32] T. Næs, O. Tomic, B.H. Mevik, H. Martens, Path modelling by sequential PLS regression, J. Chemom. 25 (2011) 28–40, https://doi.org/10.1002/cem.1357.
[33] A. Biancolillo, T. Næs, The sequential and orthogonalized PLS regression for multiblock regression: theory, examples, and extensions, in: M. Cocchi (Ed.), Data Fusion Methodology and Applications, Data Handling in Science and Technology, vol. 31, Elsevier, Amsterdam, 2019, pp. 157–177, https://doi.org/10.1016/B978-0-444-63984-4.00006-5.
[35] I. Måge, T. Næs, Split-plot design for mixture experiments with process variables: A comparison of design strategies, Chemom. Intell. Lab. Syst. 78 (2005) 81–95, https://doi.org/10.1016/j.chemolab.2004.12.010.
[36] A. Biancolillo, F. Marini, J.M. Roger, SO-CovSel: A novel method for variable selection in a multiblock framework, J. Chemom. 34 (2020) 1–21, https://doi.org/10.1002/cem.3120.
[37] J.M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada, CovSel: Variable selection for highly multivariate and multi-response calibration. Application to IR spectroscopy, Chemom. Intell. Lab. Syst. 106 (2011) 216–223, https://doi.org/10.1016/j.chemolab.2010.10.003.
[38] S. Wold, E. Johansson, M. Cocchi, PLS: partial least squares projections to latent structures, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design, ESCOM Science Publishers, Leiden, The Netherlands, 1993, pp. 523–550.
[39] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Analytica Chimica Acta 185 (1986) 1–17, https://doi.org/10.1016/0003-2670(86)80028-9.
[40] G.P. Fox, K. Onley-Watson, A. Osman, Multiple linear regression calibrations for barley and malt protein based on the spectra of hordein, J. Inst. Brew. 108 (2002) 155–159, https://doi.org/10.1002/J.2050-0416.2002.TB00534.X.
[41] B.G. Osborne, T. Fearn, P.H. Hindle, Practical NIR spectroscopy with Applications in Food and Beverage Analysis, 2nd ed,, Longman Scientific & Technical, 1993, p. 227 p..
[42] E. Pigorsch, Spectroscopic cationic quaternary ammonium starches, Starch - Stärke. 61 (2009) 129–138, https://doi.org/10.1002/STAR.200800090.
[43] B.G. Osborne, Near-Infrared Spectroscopy in Food Analysis, Encycl. Anal. Chem. (2000), https://doi.org/10.1002/9780470027318.a1018.
[44] C.D. Elvidge, Visible and near infrared reflectance characteristics of dry plant materials, Int. J. Remote Sens. 11 (1990) 1775–1795, https://doi.org/10.1080/01431169008955129.
[45] A. Biancolillo, R. Boqué, M. Cocchi, F. Marini, Data fusion strategies in food analysis, in: M. Cocchi (Ed.), Data Fusion Methodology and Applications, Data Handling in Science and Technology, vol. 31, Elsevier, Amsterdam, 2019, pp. 271–310, https://doi.org/10.1016/B978-0-444-63984-4.00010-7.
[46] P. Williams, J. Antoniszyn, M. Manley, Near-infrared Technology: Getting the Best Out of Light, AFRICAN SUN MeDIA, 2019.
[47] T. Nakaji, K. Noguchi, H. Oguma, Classification of rhizosphere components using visible-near infrared spectral images, Plant Soil. 310 (2008) 245–261, https://doi.org/10.1007/s11104-007-9478-z.
[48] W.-H. Su, D.-W. Sun, Multispectral imaging for plant food quality analysis and visualization, Compr. Rev. Food Sci. 17 (1) (2018) 220–239, https://doi.org/10.1111/1541-4337.12317.

Further reading

[34] P. Firmani, A. Nardecchia, F. Nocente, L. Gazza, F. Marini, A. Biancolillo, Multi-block classification of Italian semolina based on near infrared spectroscopy (NIR) analysis and alveographic indices, Food Chem. 309 (2020), 125677, https://doi.org/10.1016/j.foodchem.2019.125677.