

## Classification of the physiological potential of soybean seed lots using infrared spectroscopy and chemometric methods

Júlia Martins Soares\*<sup>1</sup> , Thiago Barbosa Batista<sup>2</sup> , Martha Freire da Silva<sup>3</sup> , Natália Soares Rodrigues<sup>1</sup> , Denise Cunha Fernandes dos Santos Dias<sup>1</sup> , Laércio Junio da Silva<sup>1</sup> 

**ABSTRACT:** Near-infrared (NIR) spectroscopy is a promising tool for optimizing seed analyses quickly and assertively. The aim of this study was to investigate the viability of NIR in association with chemometric methods in classification of soybean seed lots regarding their physiological potential. We evaluated 372 soybean seed lots for vigor and obtained NIR spectra from seed samples. The original spectra were pre-processed by the following methods: Standard Normal Variate (SNV), SNV + 1st and 2nd derivatives, Gap-segment derivative, and Savitzky-Golay for the first- and second-degree derivatives, as well as combinations of the methods. The lots were divided into Class I ( $\geq 85\%$  germination after accelerated aging) and Class II ( $< 85\%$  germination after accelerated aging); and the pre-processed spectra were used to build classification models through the following methods: K-nearest neighbors (KNN), Partial Least Squares – Discriminant Analysis (PLS-DA), Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM). The PLS-DA model showed greater classification accuracy and kappa, followed by SVM. The lowest accuracy values were obtained for the NB and RF models. The regions between the wavelengths 1,000-1,200 nm and 2,200-2,500 nm were the most important for distinguishing the quality levels of soybean seeds.

**Index terms:** chemometrics, *Glycine max* L. Merrill, seed quality.

**RESUMO:** A espectroscopia no infravermelho próximo (NIR) consiste em uma ferramenta promissora para otimização das análises de sementes de forma rápida e assertiva. Este trabalho teve como objetivo investigar a viabilidade do NIR, associado a métodos quimiométricos, para classificar lotes de sementes de soja quanto ao potencial fisiológico. Foram utilizados 372 lotes de sementes de soja avaliados quanto ao vigor e obtidos espectros NIR das amostras de sementes. Os espectros originais foram submetidos aos métodos de pré-processamento *Standard Normal Variate* (SNV), SNV + 1ª e 2ª derivadas; *Gap-segment derivative*; e *Savitzky-Golay*, pelas derivadas de primeiro e segundo grau, e a combinação entre os métodos. Os lotes foram divididos em Classe I ( $\geq 85\%$  de germinação após envelhecimento acelerado), Classe II ( $< 85\%$  de germinação após envelhecimento acelerado) e os espectros pré-processados foram utilizados para a construção de modelos de classificação por meio dos métodos *K-nearest neighbors* (KNN), *Partial Least Squares – Discriminant Analysis* (PLS-DA), *Naive Bayes* (NB), *Random Forest* (RF) e *Support Vector Machine* (SVM). O modelo de classificação PLS-DA apresentou maior acurácia e kappa, seguido pelo SVM. Os menores valores de acurácia foram obtidos para os modelos NB e RF. As regiões entre os comprimentos de ondas 1.000-1.200 nm e 2.200-2.500 nm foram as mais importantes para distinguir os níveis de qualidade das sementes de soja.

**Termos para indexação:** quimiometria, *Glycine max* L. Merrill, qualidade de sementes.

Journal of Seed Science, v.46,  
e202446009, 2024



<http://dx.doi.org/10.1590/2317-1545v46278267>

\*Corresponding author  
julia.m.soares@ufv.br

Received: 09/06/2023.

Accepted: 03/14/2024.

<sup>1</sup>Universidade Federal de Viçosa,  
Departamento de Agronomia,  
36570-900, Viçosa, MG, Brasil.

<sup>2</sup>GDM Genética do Brasil SA,  
15505-970, Votuporanga, SP, Brasil.

<sup>3</sup>Universidade Estadual de Maringá,  
Departamento de Ciências Agrárias,  
87507-190, Umuarama, PR, Brasil.

## INTRODUCTION

To meet the high demand for soybean throughout the world, successful establishment of plants in the field is of paramount importance, and that directly depends on seed quality (Baek et al., 2019). The high volume of seed lots received by seed companies requires quality evaluation methods that are faster and simpler and that optimize decision making regarding approval or rejection of seed lots.

In that respect, near-infrared (NIR) spectroscopy is a technique that has gained ground in evaluation of seed quality as it has proven to be a simple, quick, and non-destructive analytic technique that does not require previous preparation of the samples and does not use reagents (Bianchini et al., 2021; Reddy et al., 2022). In addition, initial investment in acquisition of equipment allows spectral readings to be taken, without additional expenses for each reading.

The NIR technique is based on absorption of electromagnetic radiation in wavelengths in the region of 780 to 2,500 nm. This range provides information related to the composition of the material under study, such that the reading of the spectra allows identification of the presence of functional groups (C-H, N-H, O-H). This reading depends on the interaction of the compound with the electromagnetic radiation, as the radiation is absorbed by water, carbohydrates, lipids, and proteins (Larios et al., 2020).

One of the obstacles in application of the NIR technique is processing the data, involving the choice of classification methods. The original data provided by the equipment lack application of pre-treatments for the purpose of reducing variability and enhancing the trait of interest sought in the spectra. These treatments minimize the effect of noise brought about by the equipment and by the characteristics of the samples, so as to improve the efficiency of the calibration model (Rinnan et al., 2009). Although these pre-treatments are important, there is the risk of applying an inadequate type of pre-treatment or very severe pre-processing, which will remove important information from the spectra. For that reason, it is important to evaluate the effect of different methods of pre-processing of the NIR data before defining the final model.

The pre-processing techniques commonly used for treatment of the NIR spectra are categorized in dispersion correction methods and derivative methods. Like dispersion correction methods, the Standard Normal Variate (SNV) pre-processing is designed to reduce the physical variability among the samples due to dispersion and to establish a baseline among the samples. Through this transformation, the spectra are centralized and the value of each wavelength is divided by the standard deviation of the absorbance of the spectra. Thus, all the spectra have mean zero and variance one, and as such, are independent of the characteristics of the dataset (Dhanoa et al., 1994; Rinnan et al., 2009).

The derivative methods are used when there is the need to remove additive and multiplicative effects in the spectra, in which the first derivative removes the baseline, while the second derivative removes the baseline and the linear trend. The first or second derivative of Savitzky-Golay (SG) is usually used, in which both techniques use smoothing so as not to greatly reduce the signal-to-noise ratio in the corrected spectra (Rinnan et al., 2009). In addition to each one of the methods mentioned, there is the possibility of changes in the window sizes used to estimate correction in the derivative methods, as well as the combination of more than one type of pre-processing.

Recent studies include applications of different classification models using infrared spectra combined with chemometric approaches, that is, application of mathematical and/or statistical methods to the spectral data. Such models have been described for seeds, obtaining high accuracy. Some methods used for modeling spectral data obtained from seeds are PCA (Principal Component Analysis), CDA (Canonical Discriminant Analysis), AdaBoost, Naive Bayes, PLS-DA (Partial Least Squares-Discriminant Analysis), Random Forest, and SVM (Support Vector Machine). Shrestha et al. (2016) classified tomato seeds regarding identity and varietal purity using PCA, CDA, and SVM, with accuracies ranging from 94% to 100% for five cultivars, regardless of the chemometric methods used. Kosmowski and Worku (2018) tested classification models through AdaBoost, Naive Bayes, PLS-DA, Random Forest, and SVM to classify barley, chickpea, and sorghum seeds, obtaining models with accuracy of 89%, 96%, and 87%, respectively.

For soybean, NIR has been studied as a method to distinguish genetically modified soybean from conventional varieties (Lee and Choung, 2011), to select seeds in F2 based on protein content (Lee et al., 2009), and to classify seeds based on seed health (Jaillais et al., 2015). The near-infrared spectroscopy technique allows exploitation of differences in biochemical composition, especially oil, protein, and carbohydrates; and there are changes in these compounds as seeds deteriorate, related to physiological quality. That suggests that the NIR technique may be an efficient tool for qualitative classification of the physiological potential of seed lots.

The aim of this study was to develop classification models from the spectral data of seeds from 372 soybean seed lots of different genotypes and with vigor differences for each genotype.

## MATERIAL AND METHODS

The experiment was conducted in the Seed Research Laboratory of the Department of Agronomy at the *Universidade Federal de Viçosa*, Viçosa, MG, Brazil. Commercial soybean seeds from 372 seed lots of 40 different genotypes, produced in the 2021/2022 crop season, were evaluated regarding moisture content; the NIR spectra of the samples from each lot were obtained; and germination was determined, as described below:

*Moisture content:* The Celmi® brand, model CM-500, moisture meter was used. Four replications from each lot were used at a volume sufficient to fill the entire device, and the results were expressed in percentage. Due to the effect of moisture content on spectral readings, maximum variation in moisture content among the lots was established at  $\pm 2\%$ .

*Vigor:* Vigor analysis was based on the accelerated aging test, according to the methodology proposed by Marcos-Filho (2020). The seeds were distributed in a single layer on the surface of a metal screen placed in the upper part of a *gerbox* (germination box), and 40 mL of distilled water was added to the lower part of the box. The boxes containing the seeds were maintained in BOD at 41 °C for 48 hours. This procedure promotes the deterioration process, since the seeds are exposed to a high temperature and relative humidity environment. After the aging period, the seeds were placed to germinate, with four replications of 50 seeds distributed in rolls of paper toweling moistened in the amount of 2.5 times the weight of the dry substrate. The rolls were kept in a seed germinator at 25 °C, and evaluations of the percentage of normal seedlings were made on the fifth day after sowing.

*Acquisition of NIR spectra:* Random samples of seeds from each lot were taken and divided into four replications; readings were taken in triplicate, for a total of 12 readings per lot. Each sample was composed of approximately 50 seeds so as to cover the entire surface of light emission. The spectra of the seed samples were obtained using the Fourier transform infrared spectrometer (Antaris II FT-NIR Analyzer; Thermo Scientific Co., Waltham, MA, USA), which spans the wavelength range of 1,000 to 2,500 nm. The spectrometer support software, TQ Analysis, was used to read and record the spectra.

*Pre-processing algorithms:* The original spectral data were pre-processed using dispersion correction methods, namely the Standard Normal Variate (SNV), SNV + 1<sup>st</sup> derivative and SNV + 2<sup>nd</sup> derivative, Gap-segment derivative, and Savitzky-Golay by the first- and second-degree derivatives, using window size 9 and 15.

### *Development of the classification models*

*Classes:* The soybean seed lots were divided into two classes according to the results of the vigor test by the accelerated aging methodology. Class I was formed by the lots that exhibited germination after accelerated aging greater than or equal to 85%. Class II was formed by the lots with germination after accelerated aging below 85%. In all, 58.3% (n = 217 spectral readings) of the lots corresponded to Class I (high vigor) and 41.7% (n = 155 spectral readings) to Class II (low vigor). The limit of 85% germination after seed aging was chosen for there to be equilibrium in the number of lots between the classes, which is important to increase the efficiency of the models.

*Classification methods:* Five classification models were trained using the spectra of the different pre-processing techniques: K-nearest neighbors (KNN), Partial Least Squares - Discriminant Analysis (PLS-DA), Naive Bayes (NB),

Random Forest (RF), and Support Vector Machine (SVM). Classification was made using the R software (R Core Team, 2023), with adaptation of the script made available by Kosmowski and Worku (2018).

**Model validation:** The models were constructed using 70% of the data for training (calibration) and the 30% remaining for validation. In addition, cross validation was performed five times for each model in the calibration set. The models were evaluated for accuracy, given by the ratio between the number of predictions correctly made by the model and the total number of predictions, and by the Kappa coefficient of Cohen (1960), which provides a measurement of agreement adjusted by the chance for random agreement:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e}$$

where  $P_o$  is the proportion of agreements observed and  $P_e$  is the proportion of agreements expected under the supposition of random agreement. The kappa coefficient is calculated as the probability of random agreement, taking into consideration all the elements of the error matrix instead of only those that are situated on the main diagonal. An illustrational diagram for the methodology is shown in Figure 1.

## RESULTS AND DISCUSSION

Of the 372 soybean seed lots evaluated regarding vigor by the accelerated aging test, 217 lots had values greater than or equal to 85%, and 155 lots had germination lower than 85% (Figure 2). The lots were thus divided into two classes regarding vigor, and then the NIR spectra of the samples were obtained from each of the lots.

From the mean raw spectra, it was possible to identify differences among the absorption peaks of the seeds of the two classes of physiological potential. Therefore, this may be an indication that differences in seed vigor are related to biochemical composition or alterations, and that can be identified through evaluation of the NIR spectrum (Figure 3). Higher reflectance

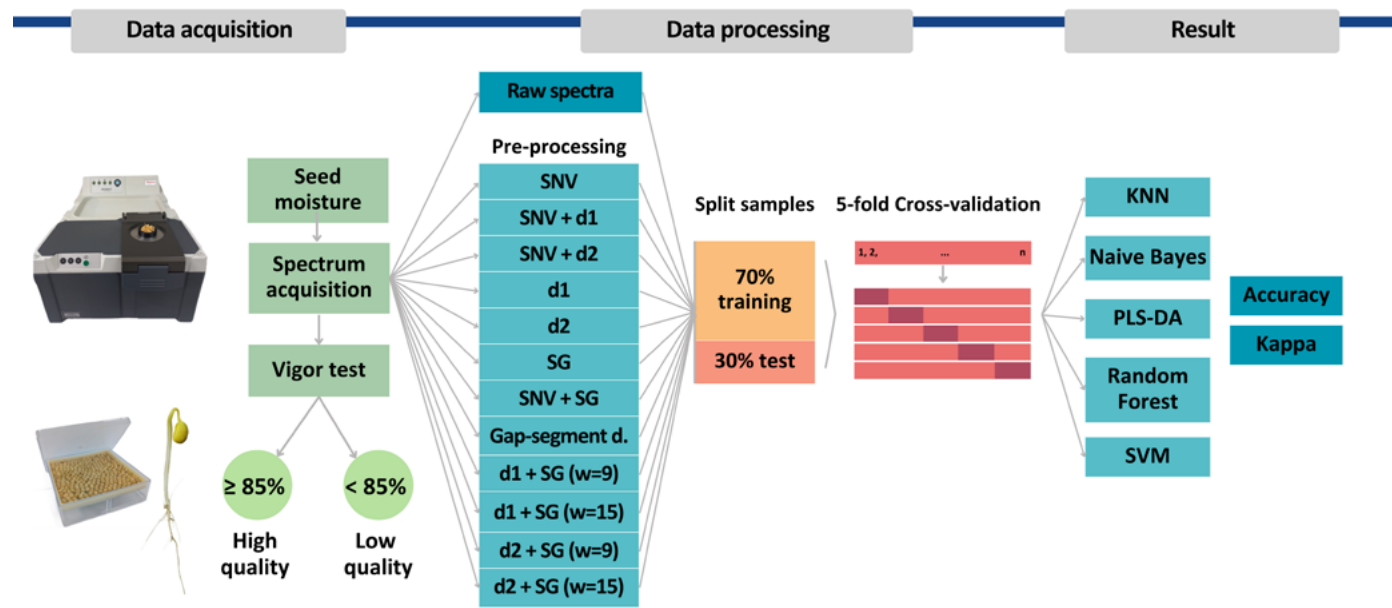


Figure 1. Flowchart of the steps for vigor classification of soybean seed lots through near-infrared spectroscopy and chemometric methods.

values of the mean raw spectra were obtained for the high quality lots over nearly the entire spectral range evaluated.

The models tested showed differences in the efficiency of classification, evaluated by the accuracy metrics (Figure 4) and kappa (Figure 5). The PLS-DA classifier achieved accuracies greater than 85% after processing of the spectra by the methods of SNV, the 1<sup>st</sup> derivative of SG window size 9, and the 2<sup>nd</sup> derivative of SG window size 15. The SVM model also exhibited accuracy values above 80% for processing by SNV and the 2<sup>nd</sup> derivative of SG window size 15. The KNN, Naive Bayes (NB), and Random Forest (RF) models did not achieve satisfactory accuracy, with values below 80% for classification.

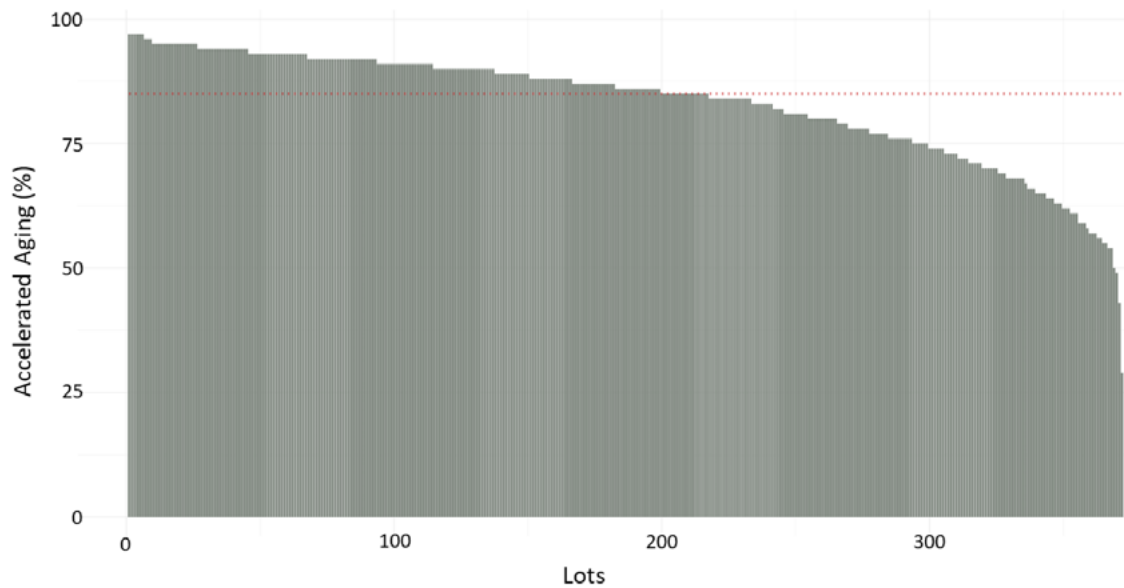


Figure 2. Germination (%) after the accelerated aging test of the 372 soybean seed lots.

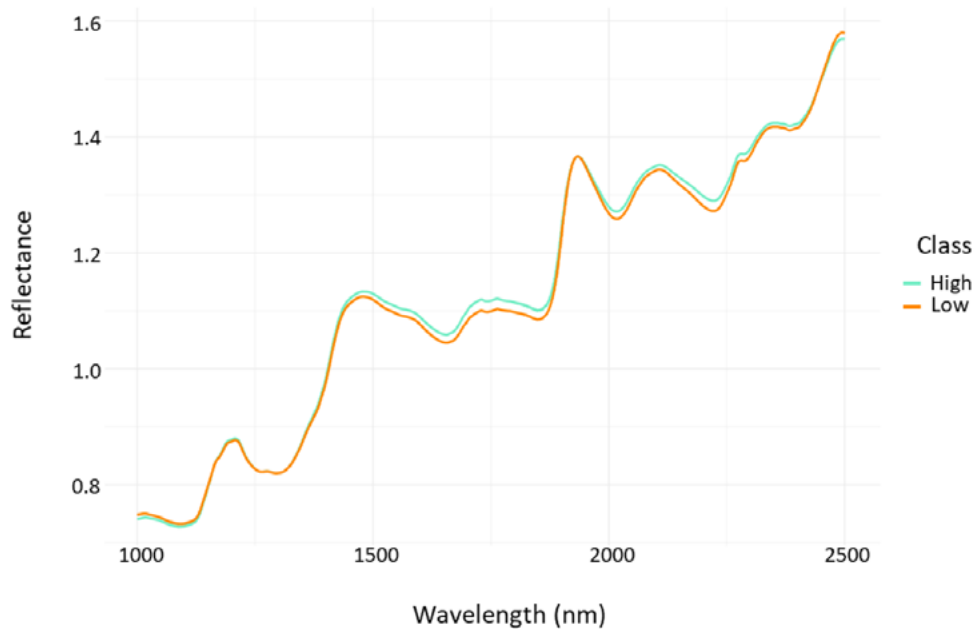


Figure 3. Mean raw spectra of soybean seed samples with different levels of physiological quality.

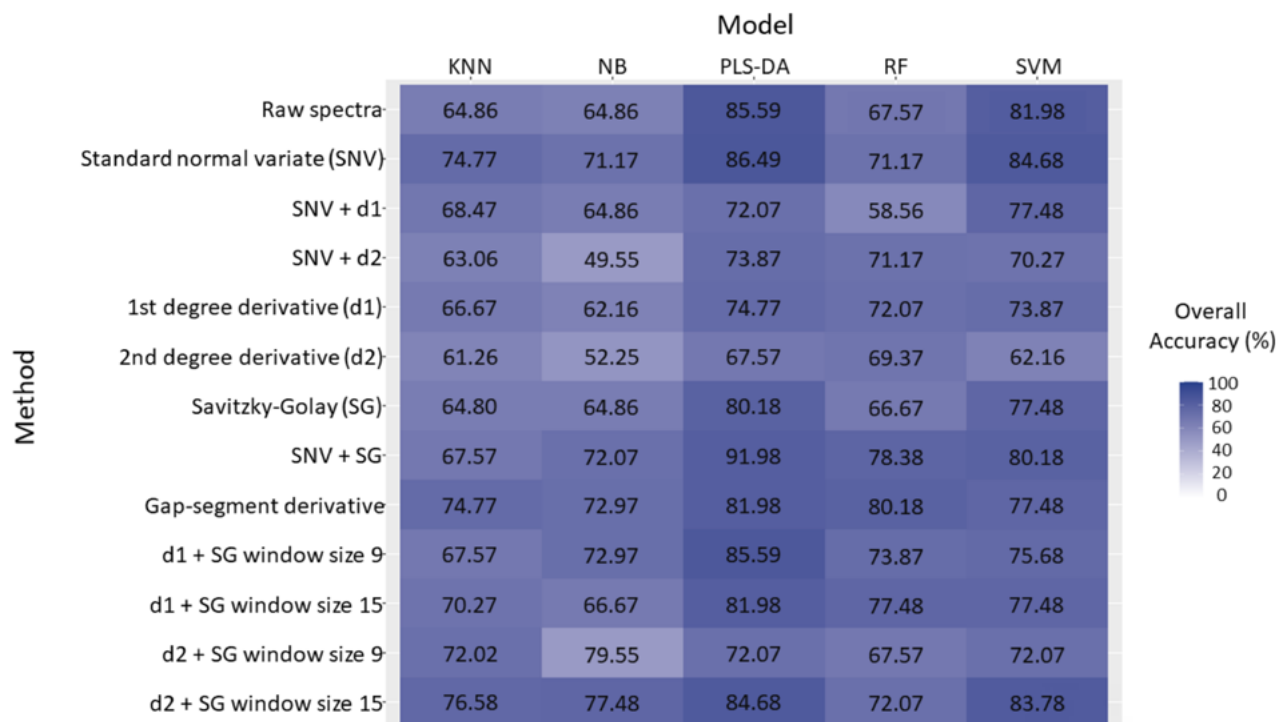


Figure 4. Heatmap of the overall accuracy of classification of soybean seeds in two vigor levels using different pre-processing methods (rows) and models (columns) on FT-NIR spectra. KNN – K-Nearest Neighbors; NB – Naive Bayes; PLS-DA – Partial Least Squares - Discriminant Analysis; RF – Random Forest; SVM – Support Vector Machine.

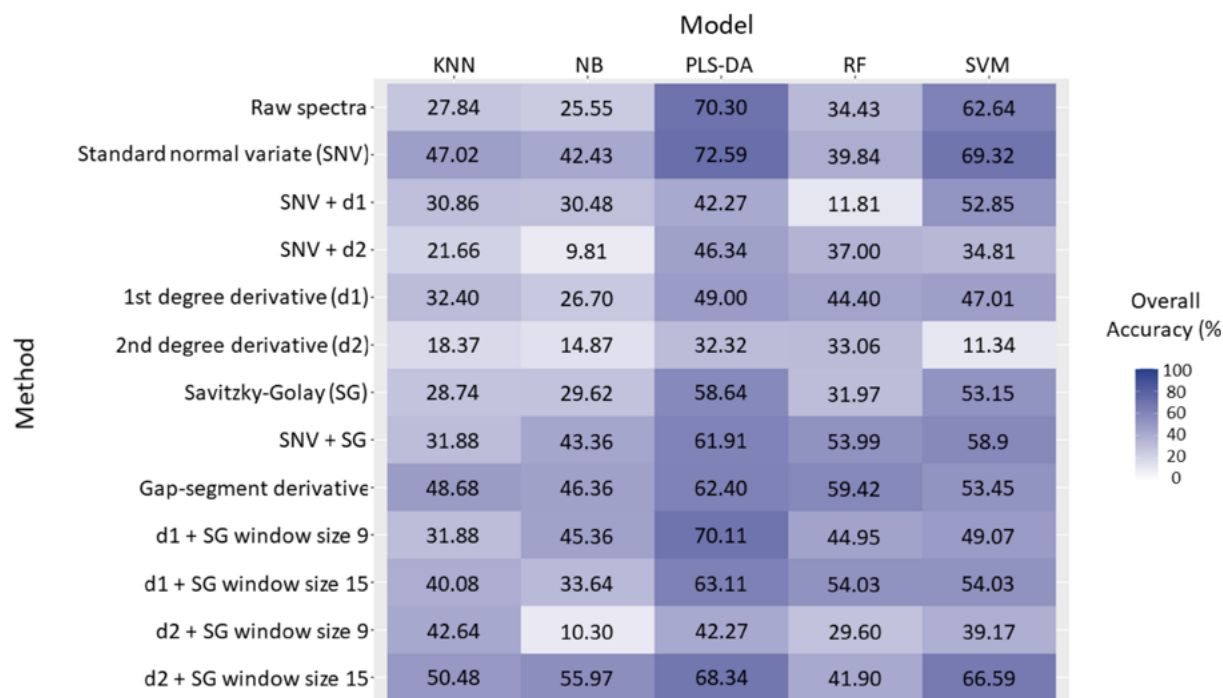


Figure 5. Heatmap of the kappa of classification of soybean seeds in two vigor levels using different pre-processing methods (rows) and models (columns) on FT-NIR spectra. KNN – K-Nearest Neighbors; NB – Naive Bayes; PLS-DA – Partial Least Squares - Discriminant Analysis; RF – Random Forest; SVM – Support Vector Machine.



The kappa analysis developed by Cohen (1960) is used in evaluation of the efficiency of the model; the objective of this analysis is checking the proportion of agreement upon removing the effect of attribution to a specific class having been at random. The result provided by kappa analysis reveals how well the classification data are aligned with the reference data. It is important to emphasize that the kappa of Cohen generates a value lower than accuracy, since it considers all the discrepancies beyond the diagonal of the confusion matrix, whereas accuracy considers only the diagonal. Landis and Koch (1977) established categories for the levels of agreement, ranging from 0 to 0.20 for minimum agreement, from 0.21 to 0.40 for reasonable agreement, from 0.41 to 0.60 for moderate agreement, from 0.61 to 0.80 for substantial agreement, and from 0.81 to 1.0 for nearly perfect agreement. In this study, substantial kappa agreements were obtained especially for the PLS-DA model; for the SVM classifier, agreements were obtained only in the processing methods SNV and d2+SG (w=15) (Figure 5).

The models obtained for the spectra pre-processed by SNV exhibited better accuracies and kappa using PLS-DA and SVM. Correction of the spectra by SNV removes the variability among the samples due to light dispersion, mainly caused by physical effects that induce variations in the spectra. Light dispersion in the studies with NIR is highly affected by the difference in seed size, by the roughness of the seed coat, and by the spherical surface of the soybean seed samples.

The derivative methods by Savitzky-Golay also stood out in the accuracy values in relation to the other techniques and combinations for the PLS-DA model. The window refers to the interval of points along the data in which a polynomial adjustment is made to smooth the data and calculate derivatives, with the aim of minimizing the deviations between the fitted values and the real values (Savitzky and Golay, 1964; Chen et al., 2004). Derivative methods reduce the noise, based on smoothing the data through the convolution process between two points in the spectrum. That way, the number of variables is reduced and regions of interest are highlighted. As the window increases, the noise tends to be removed, but if the window is very large, peaks of interest can also be distorted and lose importance in the classification. In contrast, smaller windows may retain more noise (Chen et al., 2004). In this study, the adjustment with window size 15 showed greater accuracy than with window size 9 for classification of the lots.

The derivatives and the dispersion correction methods, such as SNV, remove different effects, and the combination of the methods can be beneficial for extracting information from the spectra by removing a larger number of variables not related to the content of interest. Nevertheless, in this study, it was observed that the combination of SNV with any other pre-processing methods reduced the accuracy compared to the data processed by the SNV alone, and the combinations removed important information for separation of the seed quality classes.

In relation to the classification models, the PLS-DA was superior to the others for all types of pre-processing, both in the accuracy metric and for kappa. Secondly, classification by the SVM model exhibited higher accuracy values for the SNV and second-degree derivative of SG (window size 15) types of processing, followed by classification by Random Forest and KNN with the derivative pre-processing methods, Gap-segment, SNV + SG, and first-degree derivative of SG (window size 15). In last place, the Naive Bayes model exhibited the lowest performance among the models in prediction of the classes.

In studies on seed quality of other species, the PLS-DA algorithm also showed better performance for the spectroscopy data. Medeiros et al. (2020) more accurately classified the germination capacity of brachiaria grass seeds, with 82% accuracy using PLS-DA; whereas SVM, Random Forest, and Naive Bayes exhibited accuracies of 79%, 77%, and 69%, respectively.

In the model obtained using the Random Forest algorithm, greater capacity was observed for separation of the germination classes of *Urochloa brizantha* seeds for NIR spectroscopy data when combined with X-ray data (Medeiros et al., 2020). The RF and NB models were more efficient than SVM in identification of damage by *Fusarium* in wheat seeds (Zhang et al., 2020a). The use of SVM and NB was also efficient in classification of rice seeds with different degrees of heat damage (Zhang et al., 2020b).

Although efficiency was obtained through SVM, RF, and NB in classification of NIR data of seeds in other studies, these models were not satisfactory for reliably classifying soybean seed samples (Figures 4 and 5). Several factors can

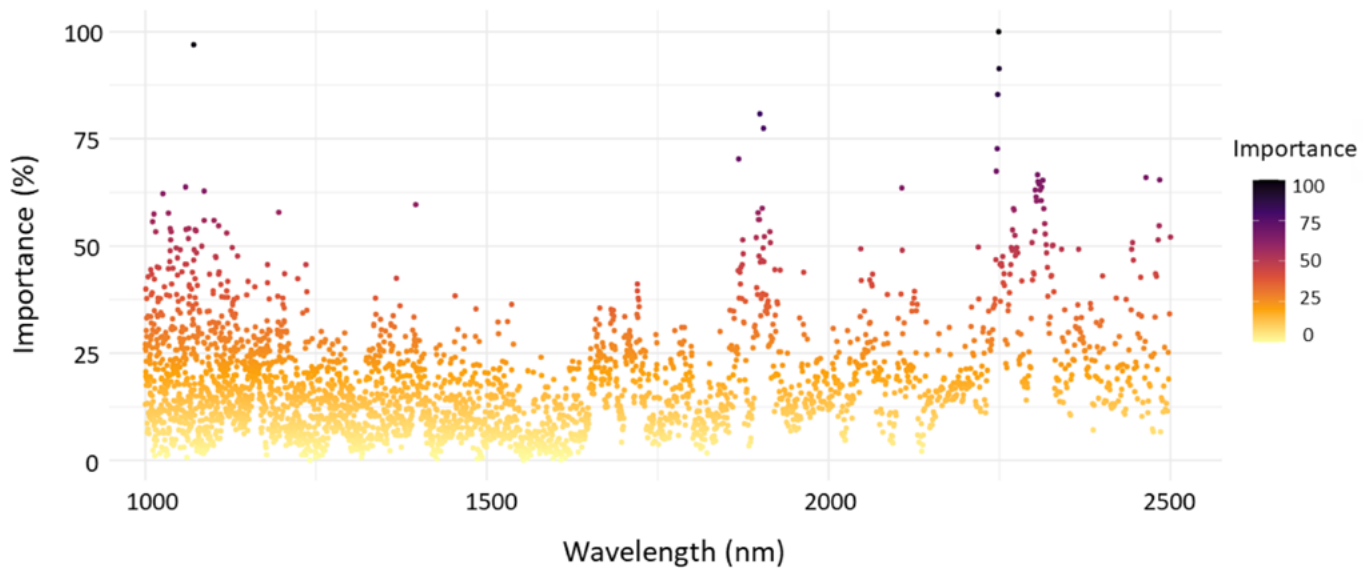


Figure 6. Overall importance of the wavelength variables for classification via PLS-DA of the quality levels of soybean seeds. The spectra were pre-processed using SNV method.

interfere in spectral reading, such as the equipment used, the crop, and the type and volume of the sample. In this study, different soybean genotypes were used, with different lots in each genotype; that is, the seeds naturally have physical and biochemical differences in their composition. Genetic difference is a variable that can be qualified by NIR when related to protein, oil, and carbohydrate content, and it affects spectral readings (Lee and Choung, 2011).

For the PLS-DA model in which SNV pre-processing was used, the accuracy value (86.5%) was among the highest obtained. Thus, from the model obtained for the spectra, the graph of overall importance of the variables for classification of the seeds was constructed in two levels of physiological quality (Figure 6).

The wavelengths exhibited peaks of importance in the regions of 1,000 – 1,200 nm, 1,800 – 1,900 nm, 2,200 – 2,300 nm, and near 2,500 nm (Figure 5). These wavelength regions correspond to functional groups that contributed to the classification of the lots (Zhang et al., 2020b). These functional groups may be indicative of water, proteins, oil, and carbohydrates, and generally overlap in broad spectral ranges. For example, the wavelength regions from 1,000 to 1,200 nm, identified as important in this study, include absorption peaks that have already been identified with oil (1,185, 1,220 nm) and protein structure (1,100 – 1,185 nm) (Al-Almery et al., 2018). Peaks in wavelengths from 2,200 to 2,300 nm were associated with lipid absorbance (2,140, 2,138, 2,347 nm); 2,300 nm was related to lipids and proteins; and 2,282 and 2,330 nm were related to carbohydrates (Xu et al., 2019).

In soybean, protein and oil are the main seed reserve components, and stresses in the field, such as high temperatures and water deficit, are able to alter the protein and oil content, synchronizing with reductions in germination and vigor (Jumrani and Bhatia, 2018). In addition, these compounds undergo changes as a result of the growing environment (Capelin et al., 2022) and the seed deterioration process (Brzezinski et al., 2022), which affects physiological potential. Thus, proteins, lipids, and carbohydrates are essential constituents in the structures of the seeds, and any change in the composition of the embryo and its reserves can affect the germination process and the vigor of the material.

Finally, the NIR spectroscopy technique in association with chemometric analysis was efficient in identifying qualitative change in seed vigor in spectral bands related to compounds that determine physiological quality. The expectation is that this may be a tool that facilitates decision making regarding rejection or approval of seed lots in analysis laboratories. It is important to highlight that this technique does not dispense with or replace traditional tests for seed quality evaluation. However, it may be useful for assisting in rapid decision making, for example, for rejecting low quality lots before processing (which means savings in time and resources) or for monitoring seed quality during storage.



## CONCLUSIONS

NIR spectroscopy in combination with chemometrics are valuable tools for classification of soybean seed lots regarding physiological potential, and they provide information that can be related to the biochemical composition of the seeds.

## ACKNOWLEDGMENTS

To the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG) and *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) for their financial support. This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) - Finance Code 001.

## REFERENCES

- AL-ALMERY, M.; GENEVE, R.L.; SANCHES, M.F.; ARMSTRONG, P.R.; MAGHI-RANG, E.B.; LEE, C.; VIEIRA, R.D.; HILDEBRAND, D.F. Near-infrared spectroscopy used to predict soybean seed germination and vigour. *Seed Science Research*, v.28, p.245-252, 2018. <https://doi.org/10.1017/S0960258518000119>
- BAEK, I.; KUSUMANINGRUM, D.; KANDPAL, L.M.; LOHUMI, S.; MO, C.; KIM, M.; CHO, B.K. Rapid measurement of soybean seed viability using kernel-based multispectral image analysis. *Sensors*. v.19, n.271, s19020271, 2019. <https://doi.org/10.3390/s19020271>
- BIANCHINI, V.M.; MASCARIN, G.M.; SILVA, L.C.A.S.; ARTHUR, V.; CARSTENSEN, J.M.; BOELT, B.; SILVA, C.B. Multispectral and X-ray images for characterization of *Jatropha curcas* L. seed quality. *Plant Methods*, v.17, n.9, s13007, 2021. <https://doi.org/10.1186/s13007-021-00709-6>
- BRZEZINSKI, C.R.; ABATI, J.; ZUCARELI, C.; KRZYZANOWSKI, F.C.; HENNINH, A.A.; HENNING, F.A. Quality and chemical composition of soybean seeds with different lignin contents in the pod and seed coat subjected to weathering deterioration in pre-harvest. *Journal of Seed Science*, v.44, e202244024, 2022. <https://doi.org/10.1590/2317-1545v44257665>
- CAPELIN, M. A.; MADELLA, L.A.; PANHO, M.C.; MEIRA, D.; BARRIONUEVO, F.; RODRIGUES, A.P.D.C.; BENIN, G. Physiological quality and seed chemical composition of soybean seeds under different altitude. *Bragantia*, v.81, e1022, 2022. <https://doi.org/10.1590/1678-4499.20210244>
- CHEN, J.; JÖNSSON, P.; TAMURA, M.; GU, Z.; MATSUSHITA, B.; EKLUNDH, L. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. *Remote Sensing of Environment*, v.91, n.3-4, p.332–344, 2004. <https://doi.org/10.1016/j.rse.2004.03.014>
- COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, v.20, n.1, p.37–46, 1960. <https://doi.org/10.1177/001316446002000104>
- DHANOVA, M.S.; LISTER, S.J.; SANDERSON, R.; BARNES, R.J. The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. *Journal of Near Infrared Spectroscopy*, v.2, n.1, p.43–47, 1994. <https://doi.org/10.1255/jnirs.30>
- JAILLAIS, B.; ROUMET, P.; PINSON-GADAIS, L.; BERTRAND, D. Detection of *Fusarium* head blight contamination in wheat kernels by multivariate imaging. *Food Control*, v.54, p.250-258, 2015. <https://doi.org/10.1016/j.foodcont.2015.01.048>
- JUMRANI, K.; BHATIA, V. S. Combined effect of high temperature and water-deficit stress imposed at vegetative and reproductive stages on seed quality in soybean. *Indian Journal of Plant Physiology*, v.23, n.2, p.227–244, 2018. <https://doi.org/10.1007/s40502-018-0365-9>
- KOSMOWSKI, F.; WORKU, T. Evaluation of a miniaturized NIR spectrometer for cultivar identification: The case of barley, chickpea and sorghum in Ethiopia. *PLOS ONE*, v.13, n.3, e0193620, 2018. <https://doi.org/10.1371/journal.pone.0193620>
- LANDIS, R.J.; KOCH, G. The measurement of observer agreement for categorical data. *Biometrics*, v.33, n.1, p.159–174, 1977. <https://doi.org/10.2307/2529310>

- LARIOS, G.; NICOLODELLI, G.; RIBEIRO, M.; CANASSA, T.; REIS, A.R.; OLIVEIRA, S.L.; ALVES, C.Z.; MARANGONI, B.S.; CENA, C. Soybean seed vigor discrimination by using infrared spectroscopy and machine learning algorithms. *Analytical Methods*, v.12, n.35, p.4303–4309, 2020. <https://doi.org/10.1039/D0AY01238F>
- LEE, J.H.; CHOUNG, M.G. Nondestructive determination of herbicide-resistant genetically modified soybean seeds using near-infrared reflectance spectroscopy. *Food Chemistry*, v.126, n. 1, p. 368–373, 2011. <https://doi.org/10.1016/j.foodchem.2010.10.106>
- LEE, J.-D.; SHANNON, J.G.; CHOUNG, M.-G. Selection for protein content in soybean from single F2 seed by near infrared reflectance spectroscopy. *Euphytica*, v.172, n.1, p.117–123, 2009. <https://doi.org/10.1007/s10681-009-0067-5>
- MARCOS-FILHO, J. Teste de envelhecimento acelerado. In: KRZYZANOWSKI, F.C.; VIEIRA, R.D.; FRANÇA-NETO, J.B.; MARCOS-FILHO, J. (Eds.) *Vigor de sementes: conceitos e testes*. Londrina: ABRATES, p.182-244, 2020.
- MEDEIROS, A.D.; SILVA, L.J.; RIBEIRO, J.P.O.; FERREIRA, K.C.; ROSAS, T.J.F.; SANTOS, A.A.; SILVA, C.B. Machine Learning for seed quality classification: An advanced approach using merger data from FT-NIR spectroscopy and X-ray imaging. *Sensors*, v.20, n.15, p.4319, 2020. <https://doi.org/10.3390/s20154319>
- R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. 2023. <https://www.R-project.org>
- REDDY, P.; GUTHRIDGE, K.M.; PANOZZO, J.; LUDLOW, E.J.; SPANGENBERG, G.C.; ROCHFORD, S.J. Near-infrared hyperspectral imaging pipelines for pasture seed quality evaluation: An overview. *Sensors*, v.22, n.5, s22051981, 2022. <https://doi.org/10.3390/s22051981>
- RINNAN, Å.; BERG, F.V.D.; ENGELSEN, S.B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, v.28, n.10, p.1201–1222, 2009. <https://doi.org/10.1016/j.trac.2009.07.007>
- SAVITZKY, A.; GOLAY, M.J.E. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, v.36, n.8, p.1627–1639, 1964.
- SHRESTHA, S.; DELEURAN, L.C.; GISLUM, R. Classification of different tomato seed cultivars by multispectral visible-near infrared spectroscopy and chemometrics. *Journal of Spectral Imaging*, v.5, a.1, 2016. <https://doi.org/10.1255/jsi.2016.a1>
- XU, J.; NWAFOR, C.C.; SHAH, N.; ZHOU, Y.; ZHANG, C. Identification of genetic variation in Brassica napus seeds for tocopherol content and composition using near-infrared spectroscopy technique. *Plant Breeding*, v.138, n.5, p.624–634, 2019. <https://doi.org/10.1111/pbr.12708>
- ZHANG, D. CHEN, G.; ZHANG, H.; JIN, N.; GU, C.; WENG, S.; WANG, Q.; CHEN, Y. Integration of spectroscopy and image for identifying fusarium damage in wheat kernels. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, v.236, p.118344, 2020a. <https://doi.org/10.1016/j.saa.2020.118344>
- ZHANG, L.; RAO, Z.; JI, H. Hyperspectral imaging technology combined with multivariate data analysis to identify heat-damaged rice seeds. *Spectroscopy Letters*, v.53, n.3, p.207–221, 12, 2020b. <https://doi.org/10.1080/00387010.2020.1726402>

