Shahid Bahonar University of Kerman

Biomechanism and Bioenergy Research

Online ISSN: 2821-1855
Homepage: https://bbr.uk.ac.ir

Iranian Society of Agricultural Machinery Engineering and Mechanization

Subspace/Discriminate Ensemble-based Machine Learning on Visible/Near-infrared Spectra as an Effective Procedure for Non-destructive Safety Assessment of Spinach

Bahareh Jamshidi[1], Najmeh Yazdanfar[2]

[1] Smart Agricultural Research Department, Agricultural Engineering Research Institute, Agricultural Research Education and Extension Organization (AREEO), Karaj, Iran.
[2] Iranian Institute of Research and Development in Chemical Industries, ACECR, Karaj, Iran.

Corresponding author: b.jamshidi@areeo.ac.ir

ARTICLE INFO

Article type:

Research Article

Article history:

Received 17 December 2024

Received in revised form 01 January 2025

Accepted 05 March 2025

Available Online 31 March 2025

Keywords:

Intelligent algorithms, Near-infrared spectroscopy, Orthogonal signal correction, Receiver operating characteristic, Subspace/discriminate ensemble.

ABSTRACT

In this study, an orthogonal signal correction (OSC)-based partial least squares (PLS) model and ensemble-based machine learning classifiers, combined with visible/near-infrared (Vis/NIR) spectroscopy, were proposed for non-destructive nitrate prediction in spinach leaves and sample safety evaluation. The OSC method was applied before developing the PLS model to enhance prediction accuracy. Spinach safety assessment was based on the maximum permissible nitrate accumulation level. Various ensemble classifiers, including subspace/discriminate, subspace/k-nearest neighbor, boosted trees, bagged trees, and random under-sampling boosted trees, were evaluated for distinguishing safe and unsafe samples. The best classification results were obtained using the subspace/discriminate ensemble classifier, achieving sensitivity, specificity, and accuracy of 95.24%, 98.73%, and 98.45% for the calibration dataset and 100%, 91.8%, and 92.31% for external validation. The receiver operating characteristic (ROC) curve indicated superior discrimination ability, with an area under the curve (AUC) of 0.95. Additionally, the best model demonstrated a high prediction speed of approximately 280 observations per second. These findings highlight that combining Vis/NIR spectroscopy with the subspace/discriminate ensemble classifier provides an effective, rapid, and non-invasive method for detecting nitrate contamination in spinach leaves, making it a promising approach for food safety monitoring.

Cite this article: Jamshidi, B., & Yazdanfar, N. (2025). Subspace/Discriminate Ensemble-based Machine Learning on Visible/Near-infrared Spectra as an Effective Procedure for Non-destructive Safety Assessment of Spinach. *Biomechanism and Bioenergy Research*, *4*(1), 59-73. https://doi.org/10.22103/bbr.2025.24539.1110

Publisher: Shahid Bahonar University of Kerman
DOI: https://doi.org/10.22103/bbr.2025.24539.1110

INTRODUCTION

Excessive consumption of nitrogen fertilizers in the soil increases nitrogen in the soil solution and consequently increases the uptake of nitrate in vegetables. However, nitrate reduction does not increase as much and accumulates in plant tissue. Many vegetables store nitrates that have absorbed more than their metabolic needs, and the concentration of nitrates in them increases. Vegetables with high nitrate accumulation are toxic to humans and cause methemoglobinemia disease and some types of cancer. Therefore, the maximum amount of nitrate, maximum allowable level of nitrate, is the most important safety index of vegetables. If the amount of nitrate is below the index, the vegetable is safe. Vegetables with the nitrate concentration above the index are unsafe. To ensure vegetable safety, there is a need to assess its nitrate concentration. Currently, nitrate concentration of the product is measured by laboratory methods which are difficult, destructive, costly, and very time-consuming. In recent decades, optical near-infrared (NIR) spectroscopy has been used for quality and safety assessment of various vegetables (Bachir et al., 2024; de Brito et al., 2022; Ito, 2014; Jamshidi, 2017; Jamshidi et al., 2016; Jamshidi et al., 2015; Li et al., 2024; Rahi et al., 2020). There is some research on some vegetables such as cucumber (Jamshidi & Yazdanfar, 2022), Japanese radish (Ito et al., 2003), komatsuna leaves (Itoh et al., 2011), and lettuce (Itoh et al., 2015) which confirms the ability of this non-destructive technology for rapid determination of nitrate concentration in the product.

For spinach leaves, Itoh et al. investigated the utilizing NIR spectroscopy for the measurement of nitrate concentration. They collected Vis/NIR spectra of small areas on the leaves in the range of 610–1050 nm and in transmittance mode (Itoh et al., 2011). Then, they cut away the measured areas from the leaves to measure their nitrate concentrations using a liquid chromatography analyzer. The spectra were pre-processed using mean-center and the standard normal variate (SNV) transformation. With an algorithm of wavelength selection, both partial least squares (PLS) and principal component regression (PCR) were used to predict the nitrate concentration of the samples. According to their results, the best accuracy was achieved from the PCR model after using mean-center method. Correlation coefficient of validation set was approximately 0.8. Moreover, Torres et al. collected reflectance NIR spectra of spinach samples using a portable miniature spectrophotometer at the spectral range of 908–1676 nm. Nitrate content of each sample was also measured using an RQFlex reflectometer (Torres et al., 2021). They used modified partial least squares (MPLS) regression after SNV and Detrend (DT), as well as derivatives of the spectra to estimate nitrate contents in the samples. Determination coefficient and standard error for the prediction set were respectively 0.62 and 688 mg /kg. In this research, new approaches for both nitrate content prediction of spinach leaves and discrimination of safe from unsafe spinach samples were investigated based on leaf spectra in Vis/NIR region (450–1000 nm). To this end, orthogonal signal correction (OSC) as an effective pre-processing technique was performed before developing the PLS models. Ensemble-based classifiers of subspace/discriminate, subspace/k-nearest-neighbor (KNN), boosted/trees, bagged/trees, and RUS (random under sampling)-Boosted/trees were also used and compared for detection of the nitrate contaminated samples. The basis of product safety was the maximum level of nitrate accumulation specified by World Health Organization (WHO) and Iranian National Standardization Organization (INSO) (2000 mg /kg (WHO, 1978; INSO, 2013)).

MATERIALS AND METHODS

Figure 1 indicates the steps of research process.

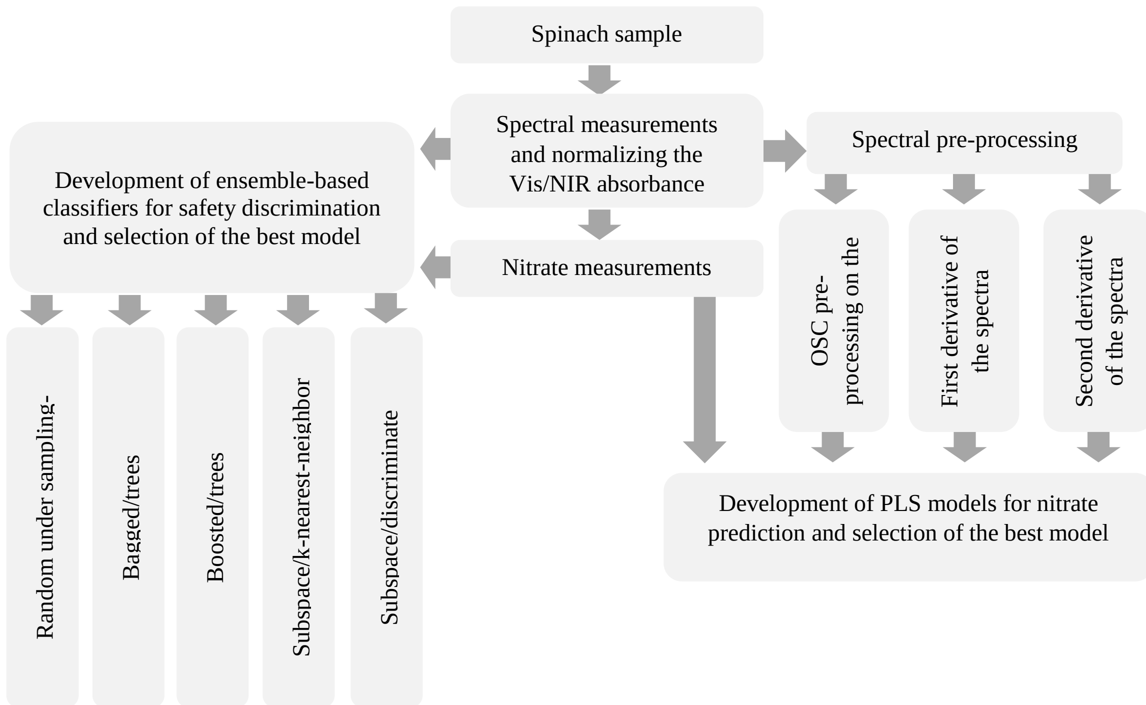


Figure 1. Flowchart of the steps in research process.

Spinach samples

Spinach samples were purchased on different days of the winter season (two seasons) from various local markets in Karaj city and west of Tehran city (the provinces of Alborz and Tehran, Iran). Each spinach sample was considered to contain one, two or three leaves depending on the leaf size. Consequently, a total of 360 spinach samples with different nitrate concentrations were used for the experiments. No preparation was done on the samples. Only damaged and yellow leaves were removed and thrown away. From the purchased samples on each day, one sample was chosen for measurement of the moisture content. Until the measurements which were done on the same day or one day after purchase, all the samples were kept at 3ºC in plastic bags. To reach ambient temperature, spinaches were placed in the laboratory before the experiments. Nitrate measurements using a reference method were carried out after spectral analysis. Before that, weight of each sample was determined using a digital balance.

Spectral measurements

Vis/NIR spectral measurements of the spinach leaves with different nitrate concentrations were done using a small portable spectrometer (GW–VIS, StellarNet Inc., USA) equipped with an optical fiber probe (R600-8-VisNIR, StellarNet Inc, USA), and a very small light source of tungsten halogen with a power of 5 W (LS1–Filter, StellarNet Inc, USA). Measurement mode and spectral range were interactance and 450–1000 nm, respectively. First, the spectra of white reference and dark were measured. For each spinach leaf, the spectra were then collected from various positions on the leaf blade, the petiole, and the stem (at least 20 positions for each leaf with 3 spectroscopic replicates per position). For spectra acquisition, the software of SpectraWiz (StellarNet Inc, USA) was used. If the sample contained two or three leaves, the spectra of each leaf was also collected. After that, the mean spectrum of the different positions on leaf or leaves was considered as the indicator spectrum for the sample and was transformed to absorbance value (log 1/R). Vis/NIR spectrum measurement method and the view of all absorption spectra of spinach samples are shown on Figure 2.

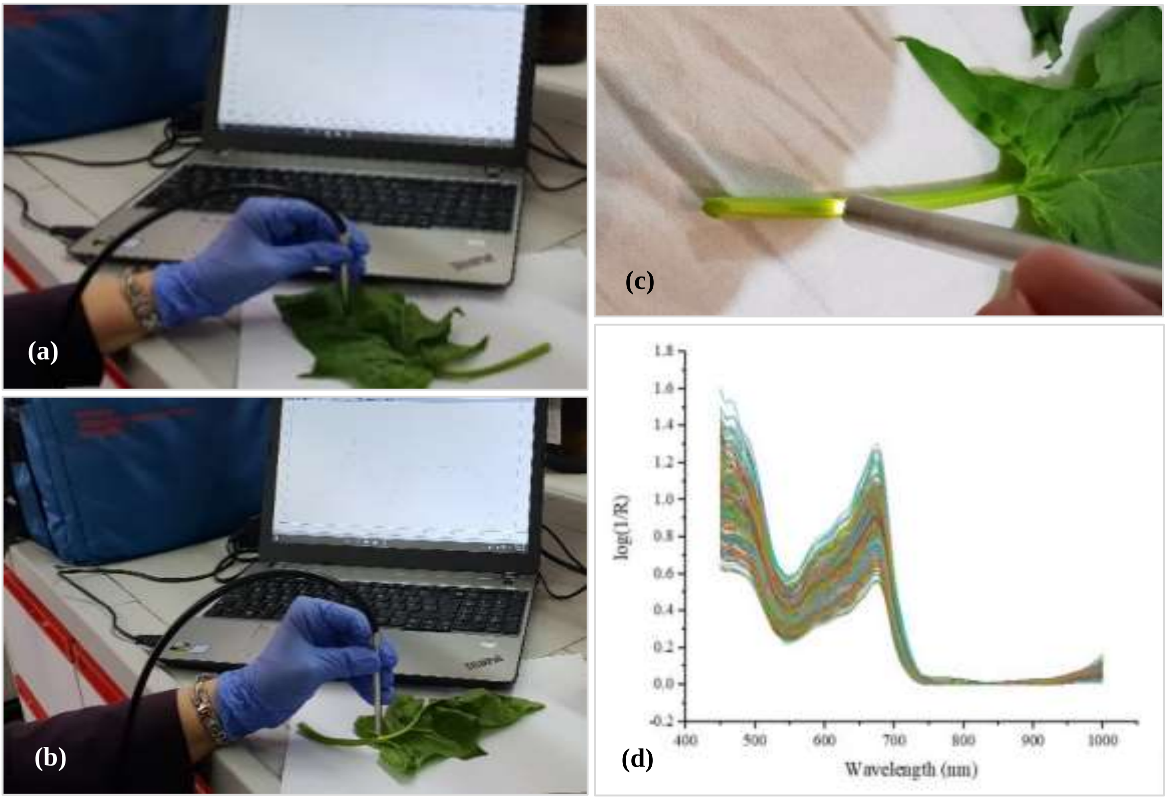


Figure 2. Measurement of Vis/NIR spectra from different positions on the spinach leaf blade (a), petiole (b), and stem (c). All absorbance spectra of the samples (d).

Nitrate measurements

Immediately after the spectroscopy, nitrate determination of the spinach samples was done based on the international standard No. 6635 (ISO, 1984) at Chemical Analysis Center (CAC) in Iranian Institute of Research and Development in Chemical Industries. The samples were then categorized into two groups of "safe″ and "unsafe″ based on their nitrate concentration measured in mg NO_3 /kg. In accord with WHO and INSO, the samples with no or with nitrate concentrations below 2000 mg /kg were assigned as "safe". The remained spinach samples with nitrate concentration above 2000 mg /kg were assigned as "unsafe" (WHO, 1978; INSO, 2013).

Data analysis and machine learning

First, the samples with abnormal spectra were excluded (37 samples). The spectra of the remained 323 samples were transformed to absorbance and the baseline was corrected. To neutralize the effect of hidden factors, the spectra were normalized by using mean normalization (MN) technique. The normalized spectra were then used for developing nitrate prediction and safety classification learner models. Data analysis and machine learning techniques were done by using Unscrambler X10.4 (CAMO Software AS, Norway) and MATLAB R2017a (Mathworks, Natick, MA, USA).

Development of nitrate prediction model

The spinach samples were divided to calibration and external validation sets, randomly. Therefore, approximately 80 percent of the samples (258 spinach samples including 237 “safe” and 21 “unsafe”) were used for constructing the calibration model. The remained 20 percent of the samples (65 spinach samples including 61 “safe” and 4 “unsafe”) were also selected for external validation (prediction) of the model.

PLS regression was used for development of nitrate prediction models in spinach samples. Before that, OSC as an effective pre-processing technique was performed to remove spectral features that are not related to the measured response, nitrate level (Biney et al., 2021). OSC derives and removes one or more background components that are orthogonal to the nitrate content, mathematically. Therefore, it removes less information about nitrate contents in the samples and makes a PLS model more accurate (Wold et al., 1998).

In this research, Thomas Fearn’s OSC method (Fearn, 2000) was used before building PLS calibration model. The results were then compared with those achieved by the model developed based on first derivative (D_1) and second derivative (D_2) of the samples’ spectra. To this end, the algorithm of Savitzky–Golay (SG) was used while the points of smoothing and the order of polynomial were 5 and 2, respectively.

Considering the 15 latent variables (LVs), PLS calibration models were built with 20 segments random cross validation method. To evaluate the calibration models, correlation coefficient of calibration (r_c) and standard error of calibration (SEC), as well as correlation coefficient of cross validation (r_{cv}) and standard error of cross validation (SECV) were considered. The performance of the calibration models for prediction of nitrate level in unknown samples were then assessed based on the correlation coefficient of prediction (r_p), standard error of prediction (SEP), and the ratio of performance to deviation (RPD).

Development of safety classification models

Development of hybrid architectures, known as ensemble methods, is an important research field in artificial intelligence (AI). These techniques use multiple learning algorithms for improvement of the performance. In ensemble classification, a collection of classifiers is used to classify unknown samples instead of just a single-classifier. All the classification models in the ensemble predict the class of each unknown sample. Then, the predictions of the classifiers are combined using a kind of voting system. Therefore, ensemble-based classifiers combine the predictive power of multiple individual classifiers and reduce the possibility of poor selection compared to the single-classifiers (Mienye & Sun, 2022; Mounce et al., 2017). Ensemble-based classifiers have several procedures such as random subspace, bagging (bootstrap aggregation), and boosting. Random subspace randomizes the algorithm of learning by randomly choosing a subset of attributes before the algorithm training. The predictions are then combined with the majority vote. In bagging approach, a set of trained models on random data is built. The outputs of the models are finally aggregated or combined by averaging. Boosting method works based on the voting or averaging of the multiple models so that the developed models are weighted based on the performance (Ashour et al., 2018; Mienye & Sun, 2022).

In this research, ensemble classifiers were utilized to develop the classification models for discrimination of safe and unsafe spinach samples. Different ensembles of subspace/discriminate, subspace/KNN, the boosted/trees, bagged/trees, and the RUS-Boosted/trees were performed and compared.

Discriminant analysis (DA) and KNN are two powerful classifier models in machine learning which their accuracies can be improved by using within a random subspace ensemble. Moreover, subspace ensembles have the merit of using less memory than ensembles with all predictors so that they can handle missing values. In this research, both subspace/discriminate ensemble and subspace/KNN ensemble were used for distinguishing safe from unsafe spinach samples.

On the other hand, decision trees are popular classifier models in machine learning that learn from an existing dataset and a tree-like structure can be achieved. It shows the relationships between the input predictors of the dataset and the desired class. Decision trees are fast and unstable. Therefore, they are suitable for ensembles because by using them within an ensemble the problem of instability is reduced (Mounce et al., 2017). Boosting, bagging (bootstrap aggregation) and random under sampling boosting (RUS-Boost) are three common methods for creating of decision trees ensembles. Therefore, the ensembles of boosted/trees, bagged/trees, and the RUS-Boosted/trees were also used for safety assessment of the spinach samples in this research. RUS-Boost sampling is particularly effective in classifying unbalanced data, meaning that some classes have significantly fewer members in the training dataset than others (Mounce et al., 2017).

Confusion matrix for all the ensemble classifiers was assessed to evaluate the performance of each classification model on both calibration and prediction sets. In the confusion matrix (Figure 3), positive data (unsafe samples) that were placed in the predicted class of positive are known as True Positive (TP), and negative data (safe samples) that were placed in the predicted class of negative are known as True Negative (TN). Moreover, negative data that were placed in the predicted class of positive are known as False Positive (FP), and positive data that were placed in the predicted class of negative are known as False Negative (FN).

True class	Negative	TN	FP
	Positive	FN	TP
		Negative	Positive
		Predicted class	

Figure 3. A confusion matrix with True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Positive data are 'unsafe' samples and negative data are 'safe' samples.

The classifier models were evaluated based on three parameters of sensitivity (Eq. 1), specificity (Eq. 2), and accuracy (Eq. 3) for calibration and prediction sets to distinguish the unsafe from the safe spinach samples. The optimal model should have the maximum values of sensitivity, specificity and accuracy. Compared to specificity and accuracy values, higher sensitivity in detection of unsafe spinach, nitrate-contaminated sample, is preferred to ensure the product safety.

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

To evaluate the performance of the ensemble classifiers, the results were also analyzed using receiver operating characteristic (ROC) curve, a graphical representation of the model performance. ROC indicates the trade-off between TP rates (TPR) or sensitivity and FP rate (FPR) or 1 – specificity. Classification models with the curves closer to the upper left corner of the ROC space have better performance. The closer the curves come to the 45º line, the less powerful are the classifier models. The area under the ROC curve (AUC) is a global measure of the model's ability for discrimination of the classes. An AUC of 1.0 demonstrates that the model has perfect ability for discriminating. While an AUC of 0.5 demonstrates a model with no discrimination ability (de Hond et al., 2022).

RESULTS AND DISCUSSION

Samples' statistics

The mean of weight for all spinach samples and the mean of moisture content of the selected samples were 6.08 (g) and 92.43%, respectively.

Figure 4 shows the histogram plot for nitrate concentrations in the spinach samples. As it can be seen, there was a great variation in nitrate concentration of the samples from 0 to 5215 mg /kg with the mean and standard deviation (SD) of 673.97 and 1039.22 mg /kg, respectively. The mean of nitrate accumulation in the spinach samples was lower than the maximum level of nitrate concentration in accord with WHO and INSO (2000 mg /kg). Table 1 shows the status of 'safe' and 'unsafe' samples in terms of nitrate accumulation. Most of the samples were safe without or with nitrate concentration below 2000 mg /kg ranged from 0 to 1995 mg /kg. Unsafe samples had nitrate concentration above 2000 mg /kg ranged from 2008 to 5215 mg /kg. The variation in nitrate concentrations of unsafe samples was more than that for safe samples.

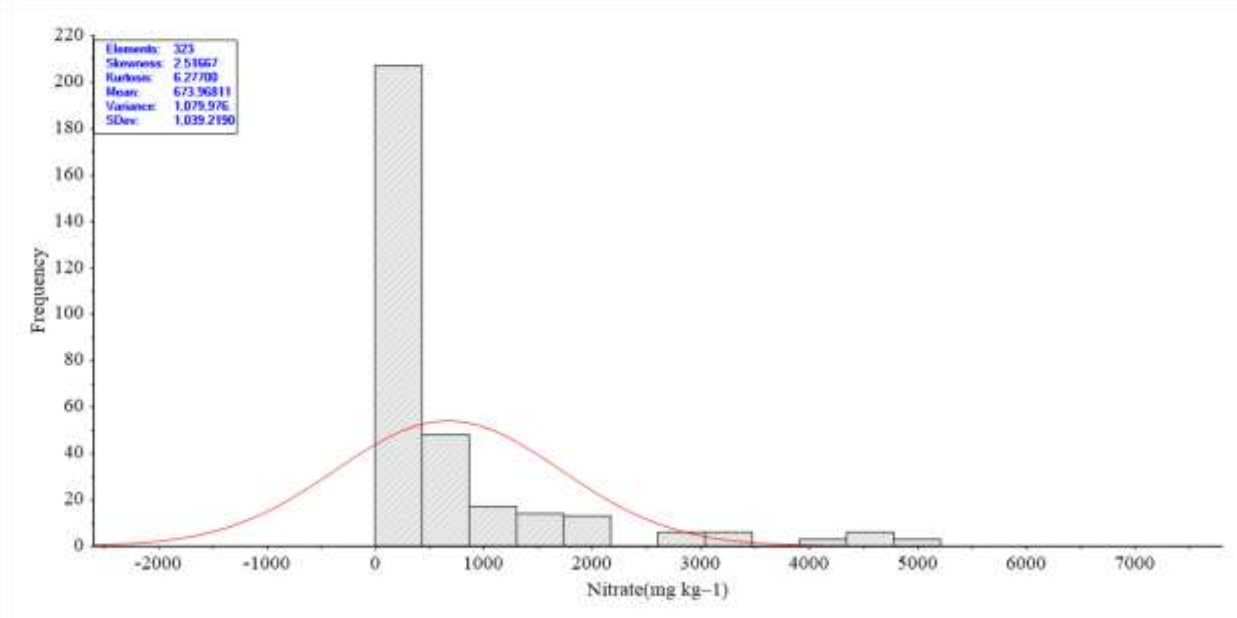


Figure 4. Histogram plot for nitrate contents in spinaches

Table 1. The status of nitrate concentration in both safe and unsafe spinach samples.

	Safe samples (N = 298)			Unsafe samples (N = 25)		
	Range	Mean	SD	Range	Mean	SD
Nitrate (mg /kg)	0–1995	415.32	489.86	2008–5215	3757.00	887.93

N = number of the samples

Vis/NIR spectra

As it can be seen in Figure 2d, the absorbance was related to the color compounds of spinach in the visible region of the samples' spectra. In all spectra, a strong absorption peak at 675 nm due to the content of chlorophyll *a* in the spinach samples was found. Itoh et al. reported a similar peak around the wavelength of 650 nm, the absorption region of chlorophyll *b* and total chlorophyll (Itoh et al., 2011). Furthermore, Kara and Dasgan, reported that there is a strong correlation between nitrate concentration and total chlorophyll in vegetables. In the NIR region (Kara & Dasgan, 2018), there was an increasing trend similar to the achievements reported by Itoh et al. which could be caused by the second overtones of functional groups of N–H and O–H (Itoh et al., 2011).

Nitrate prediction results

Table 2 presents the nitrate concentration estimation results in the spinaches by using PLS models constructed after performing D_1, D_2, and OSC on the spectra. Although the results of PLS model developed with D_2 for calibration set as well as cross validation and external validation (prediction) were better than those achieved by using the model constructed based on D_1, the best results were obtained when PLS performed after OSC. It was noted that using OSC before building PLS calibration model is an effective approach for estimation of nitrate content in spinach samples based on the spectra of the leaves in Vis/NIR region. The RPD of the OSC-based PLS model was close to 2 where the model could be adequate for screening (Wiedemair et al., 2019).

Table 2. The results of the PLS models developed for prediction of nitrate concentration in the spinach samples.

Pre-processing	LVs	Calibration		Cross Validation		External Validation		RPD
		r_c	SEC (mg /kg)	r_{cv}	SECV (mg/kg)	r_p	SEP (mg/kg)	
D_1	15	0.78	655.81	0.62	831.03	0.71	708.95	1.46
D_2	15	0.93	386.19	0.80	628.41	0.75	685.45	1.52
OSC	15	0.94	347.66	0.85	557.01	0.81	590.02	1.76

Bold values indicate the best prediction model.

The estimation results of nitrate accumulation in spinach leaves for the best calibration model in this research were better than the results reported by Itoh et al. in terms of SEC and r_c (347.66 mg /kg and 0.94 compared to 542 mg /kg and 0.86, respectively) (Itoh et al., 2011). The prediction results of the model for the external validation set were slightly weaker than the results achieved by Itoh et al. in terms of SEP and r_p (590.02 mg /kg and 0.81 compared to 403.6 mg /kg and 0.84, respectively) (Itoh et al., 2011). However, they measured nitrate concentrations only in the small areas of the leaves where the spectra were collected and the number of samples examined in the mentioned research was totally 48 samples and not included large amounts of nitrate concentration (e.i. more than 3800 mg /kg).

The best model in this research had much better results in comparison with the results reported by Torres et al. (2020) (SECV = 557.01 mg /kg against SECV = 766 mg /kg) for prediction of nitrate content in spinach leaves. They used NIR spectra in the range of 908–1676 nm (without visible region) and analyzed the spectra based on linear variable filter (LVF) technology.

On the other hand, the best results of this research were better compared to the results obtained by Torres et al. (2021) for estimation of nitrate content in spinach leaves in terms of SEC, coefficient of determination for calibration set (R_c^2), SECV, coefficient of determination for cross validation set (R_{cv}^2), SEP, coefficient of determination for prediction set (R_p^2), and RPD (347.66 mg /kg, 0.90, 557.01 mg /kg, 0.72, 590.02 mg /kg, 0.66, and 1.76 compared to 675 mg /kg, 0.57, 708 mg /kg, 0.53, 688 mg /kg, 0.62, and 1.53, respectively). Meanwhile, they did not consider very low and high nitrates (less than 70 and more than 3900 mg /kg).

Correlation between the estimated and the measured nitrate accumulation in the spinach samples for the prediction set by using the developed OSC-based PLS model is illustrated in Figure 5.

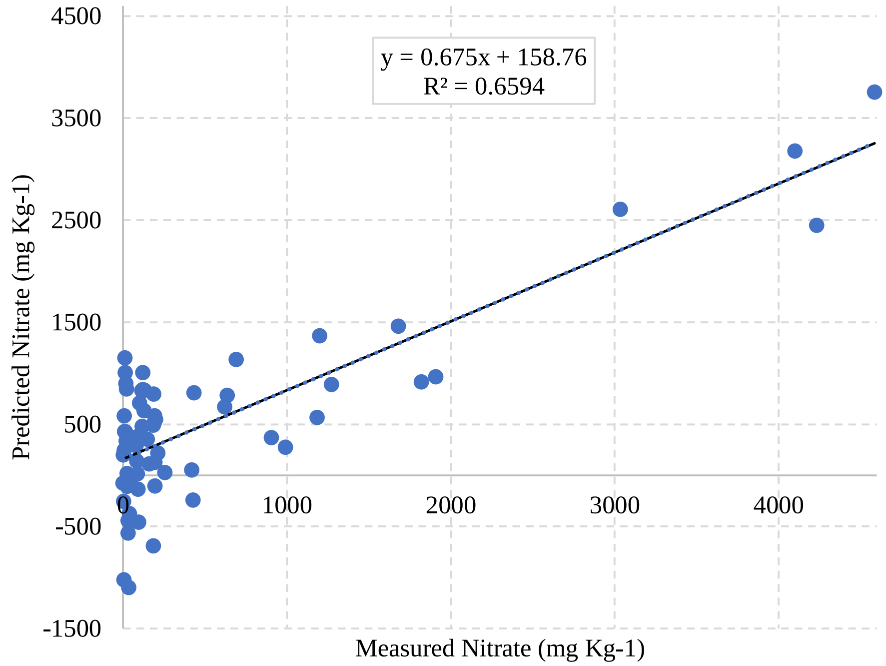


Figure 5. Correlation between the measured and the predicted nitrate contents in the spinach leaves for prediction set using the OSC-based PLS model.

Classification results

Figure 6 shows the classification results by different ensemble classifiers for discrimination of safe and unsafe spinach samples. The calculated parameters of sensitivity, specificity, and accuracy for the ensemble classifiers are also reported in Table 3.

All the calibration models except the RUS-Boosted/trees ensemble, had accuracy above 91% for classifying the unsafe and safe samples. The best accuracy was obtained by subspace/discriminate ensemble (98.45%). Although the classifier models of boosted/trees ensemble and bagged/trees ensemble had slightly better specificities than the subspace/discriminate ensemble classifier (100% and 99.58% against 98.73%), their sensitivities were extremely low and unacceptable (0% and 14.28%, respectively). In contrast, the subspace/discriminate ensemble classifier had excellent sensitivity (95.24%). Due to the fact that the sensitivity of the models is more important than their specificities for detection of the unsafe samples, the subspace/discriminate ensemble classifier was chosen as the best classification model to distinguish unsafe from safe spinach samples. The sensitivity of the best model for external validation was 100%. Moreover, the specificity and accuracy of this model were respectively 91.8% and 92.31% for external validation set. These results indicate that the developed model is very appropriate for sample screening and final confirmation of spinach leaf contamination. It was concluded that ensemble classifiers with random subspace are better than those with bagging or boosting approaches for distinguishing the unsafe from the safe samples based on Vis/NIR spectra of the spinach leaves. It was also noted that RUS-Boosted/trees ensemble is more sensitive than boosted/trees and bagged/trees ensemble models for detecting the unsafe spinach samples because one class (unsafe) had fewer members than the other class (safe).

Classification set

True class	Predicted class: Safe	Predicted class: Unsafe
Safe	234	3
Unsafe	1	20

Prediction set

True class	Predicted class: Safe	Predicted class: Unsafe
Safe	56	5
Unsafe		4

(a)

Classification set

True class	Predicted class: Safe	Predicted class: Unsafe
Safe	229	8
Unsafe	12	9

Prediction set

True class	Predicted class: Safe	Predicted class: Unsafe
Safe	61	
Unsafe	1	3

(b)

Classification set

True class	Predicted class: Safe	Predicted class: Unsafe
Safe	237	
Unsafe	21	

Prediction set

True class	Predicted class: Safe	Predicted class: Unsafe
Safe	61	
Unsafe	4	

(c)

Classification set

True class	Predicted class: Safe	Predicted class: Unsafe
Safe	236	1
Unsafe	18	3

Prediction set

True class	Predicted class: Safe	Predicted class: Unsafe
Safe	61	
Unsafe	3	1

(d)

Classification set

True class	Predicted class: Safe	Predicted class: Unsafe
Safe	195	42
Unsafe	7	14

Prediction set

True class	Predicted class: Safe	Predicted class: Unsafe
Safe	57	4
Unsafe		4

(e)

Figure 6. Confusion matrix for subspace/discriminate ensemble (a), subspace/KNN ensemble (b), boosted/trees ensemble (c), bagged/trees ensemble (d), and RUS-Boosted/trees ensemble (e) classifiers for discrimination of safe and unsafe spinach samples

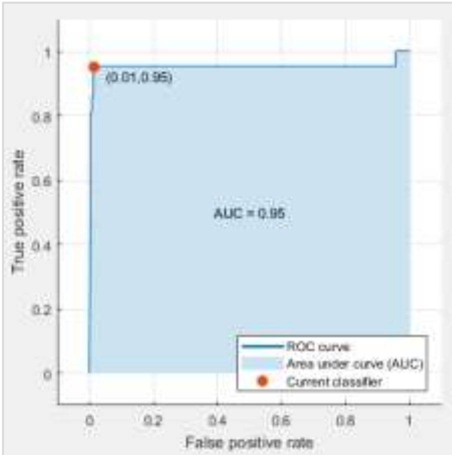
Table 3. Results of the ensemble classifiers for discrimination of "unsafe" from "safe" spinach samples.

Classifier model	Calibration			External Validation		
	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
subspace/discriminate ensemble	95.24	98.73	98.45	100	91.8	92.31
subspace/KNN ensemble	42.86	96.62	92.25	75	100	98.46
boosted/trees ensemble	0	100	91.86	0	100	93.85
bagged/trees ensemble	14.28	99.58	92.63	25	100	95.38
RUS-Boosted/trees ensemble	66.67	82.28	81.01	100	93.44	93.85

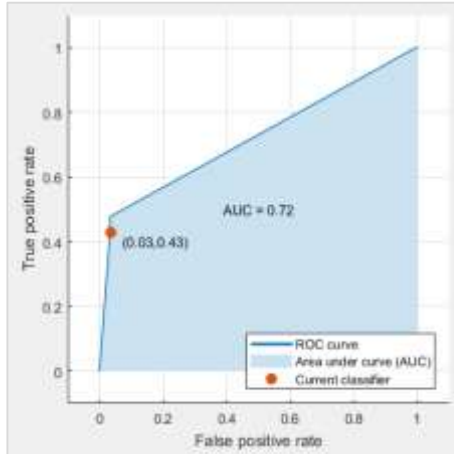
Bold values indicate the best ensemble classifier.

The ROC curves of the calibration models of different five ensemble classifiers for distinguishing unsafe from safe spinach samples are illustrated in figure 7a to figure 7e. As it can be seen, the subspace/discriminate ensemble classifier gave curve closer to the upper left corner of the ROC space with an AUC of 0.95. Therefore, this classifier model had the best performance and discriminating ability than other ensemble classifiers. The weakest classifier was boosted/trees ensemble model with AUC of 0.51. Compared to boosted/trees and bagged/trees ensemble classifiers, RUS-Boosted/trees ensemble had better ROC curve for detection of unsafe samples with AUC of 0.86. These achievements confirmed the obtained results from the confusion matrix.

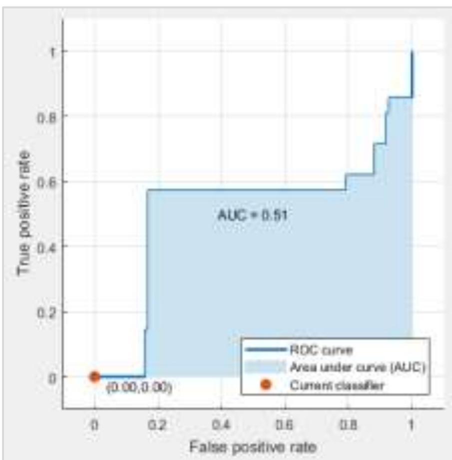
On the other hand, the boosted/trees ensemble model took the longest computation time with a prediction speed of approximately 770 observations per second (~770 obs/s). The shortest computation time was related to the subspace/KNN ensemble classifier with prediction speed of ~120 obs/s. The prediction speeds in the ensembles of bagged/trees and RUS-Boosted/trees were ~370 and ~530 obs/s, respectively. However, a reasonable time of computation in the subspace/discriminate ensemble model with prediction speed of ~280 obs/s indicated that this random subspace-based model can be suitable for rapid classification of the samples. Overall, the subspace/discriminate ensemble classifier was suggested as an effective model in distinguishing unsafe from safe spinach samples based on Vis/NIR spectra of the leaves with a training time of 11.175 seconds.



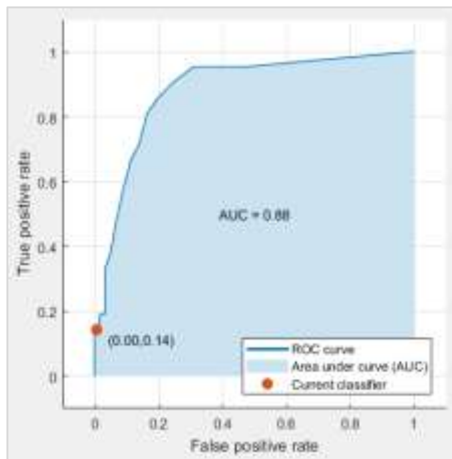
(a)



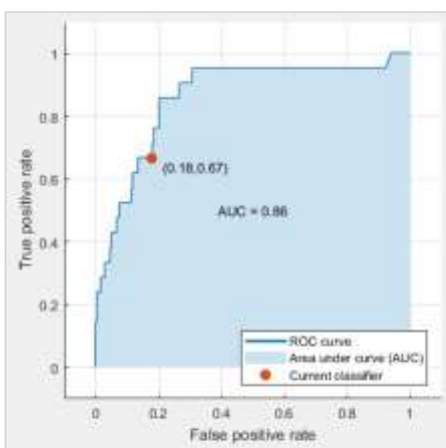
(b)



(c)



(d)



(e)

Figure 7. The ROC curves for ensembles of subspace/discriminate (a), subspace/KNN (b), boosted/trees (c), bagged/trees (d), and RUS-Boosted/trees (e) for discrimination of safe and unsafe spinach samples

This research offers significant contribution for rapid and non-destructive safety evaluation of spinach samples in terms of nitrate concentrations based on optical spectroscopy in Vis/NIR region of the electromagnetic spectrum (450–1000 nm) by purposing new approaches of signal pre-processing and machine learning techniques. Based on the results, utilizing the OSC before developing the PLS model is proposed as an effective procedure for prediction of nitrate concentration in spinach leaves. The ability of OSC-based PLS model was acceptable so that it can be used for initial screening of the spinach samples in terms of the nitrate concentrations. Employing the ensemble-based classifiers with three approaches of random subspace, bagging, and boosting proved that the combination of Vis/NIR spectroscopy and the subspace/discriminate ensemble classifier realizes the best accuracy and the performance compared to the other classifiers. Therefore, this novel approach is very suitable for non-invasive and fast discrimination of the unsafe from the safe samples based on the maximum level of nitrate contents in accord with WHO and INSO. The proposed ensemble-based machine learning algorithm can be useful for developing an expert screening system for the spinach leaves. For further researches, it is recommended to investigate other ensemble-based classifiers and to strengthen the classification models by increasing the dataset size.

FUNDING SOURCES

This work was funded by the Agricultural Research Education and Extension Organization (AREEO) [project number 24-14-14-032-970618].

DECLARATION OF COMPETING INTEREST

The authors declare no competing interests.

REFERENCES

Ashour, A. S., Guo, Y., Hawas, A. R., & Xu, G. (2018). Ensemble of subspace discriminant classifiers for schistosomal liver fibrosis staging in mice microscopic images. *Health information science and systems*, *6*, 1-10. https://doi.org/10.1007/s13755-018-0059-8

Bachir, N., Haddarah, A., SEPULCRE, F., & PUJOLA, M. (2024). Derivation of a no significant risk level (NSRL) of acrylamide in potato-based synthetic models and validation by NIR spectroscopy. *Food Chemistry Advances, 4*, 100652. https://doi.org/10.1016/j.focha.2024.100652

Biney, J. K. M., Blöcher, J. R., Borůvka, L., & Vašát, R. (2021). Does the limited use of orthogonal signal correction pre-treatment approach to improve the prediction accuracy of soil organic carbon need attention? *Geoderma, 388*, 114945. https://doi.org/10.1016/j.geoderma.2021.114945

de Brito, A. A., Campos, F., dos Reis Nascimento, A., Damiani, C., da Silva, F. A., de Almeida Teixeira, G. H., & Júnior, L. C. C. (2022). Non-destructive determination of color, titratable acidity, and dry matter in intact tomatoes using a portable Vis-NIR spectrometer. *Journal of Food Composition and Analysis, 107*, 104288. https://doi.org/10.1016/j.jfca.2021.104288

de Hond, A. A., Steyerberg, E. W., & Van Calster, B. (2022). Interpreting area under the receiver operating characteristic curve. *The Lancet Digital Health*, *4*(12), e853-e855. https://doi.org/10.1016/S2589-7500(22)00188-1

Fearn, T. (2000). On orthogonal signal correction. *Chemometrics and intelligent laboratory systems, 50*(1), 47-52. https://doi.org/10.1016/S0169-7439(99)00045-3

ISO (International Organization for Standardization). (1984). *Fruit, vegetable and derived products - Determination of nitrite and nitrate content - Molecular absorption spectrometric method* (Tech. Rep., Standard No. 6635).

INSO (Iranian National Standardization Organization). (2013). *Maximum levels for nitrates in agricultural products* (Standard No. 16596). (in Persian)

Ito, H. (2014). New ways to evaluate the quality of vegetables using instruments. *Japan Agricultural*

Research Quarterly: JARQ, *48*(2), 111-120. https://doi.org/10.6090/jarq.48.111

Ito, H., Horie, H., Ippoushi, K., & Azuma, K. (2003). Potential of Visible-Near Infrared (VIS-NIR) Spectroscopy for Non-destructive Estimation of Nitrate Content Oin Japanese Radishes. International Conference on Quality in Chains. An Integrated View on Fruit and Vegetable Quality 604,

Itoh, H., Nomura, K., Shiraishi, N., Uno, Y., Kuroki, S., & Ayata, K. (2015). Continuous measurement of nitrate concentration in whole lettuce plant by visible-near-infrared spectroscopy. *Environmental Control in Biology*, *53*(4), 205-215. https://doi.org/10.2525/ecb.53.205

Itoh, H., Tomita, H., Uno, Y., & Shiraishi, N. (2011). Development of method for non-destructive measurement of nitrate concentration in vegetable leaves by near-infrared spectroscopy. *IFAC Proceedings Volumes*, *44*(1), 1773-1778. https://doi.org/10.3182/20110828-6-IT-1002.00738

Jamshidi, B. (2017). Non-destructive safety assessment of agricultural products using Vis/NIR spectroscopy. *Nir News*, *28*(1), 4-8. https://doi.org/10.1177/0960336016687519

Jamshidi, B., Mohajerani, E., & Jamshidi, J. (2016). Developing a Vis/NIR spectroscopic system for fast and non-destructive pesticide residue monitoring in agricultural product. *Measurement*, *89*, 1-6. https://doi.org/10.1016/j.measurement.2016.03.069

Jamshidi, B., Mohajerani, E., Jamshidi, J., Minaei, S., & Sharifi, A. (2015). Non-destructive detection of pesticide residues in cucumber using visible/near-infrared spectroscopy. *Food Additives & Contaminants: Part A*, *32*(6), 857-863. https://doi.org/10.1080/19440049.2015.1031192

Jamshidi, B., & Yazdanfar, N. (2022). Development of a spectroscopic approach for non-destructive and rapid screening of cucumbers based on maximum limit of nitrate accumulation. *Journal of Food Composition and Analysis*, *110*, 104513. https://doi.org/10.1016/j.jfca.2022.104513

Kara, M., & Dasgan, H. (2018). Nitrate, nitrite and chlorophyll contents in parsley and their relationships with each other. XXX International Horticultural Congress IHC2018: International Symposium on Fruit and Vegetables for Processing, International 1292,

Li, S., Li, J., Wang, Q., Shi, R., Yang, X., & Zhang, Q. (2024). Determination of soluble solids content of multiple varieties of tomatoes by full transmission visible-near infrared spectroscopy. *Frontiers in Plant Science*, *15*, 1324753. https://doi.org/10.3389/fpls.2024.1324753

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *Ieee Access*, *10*, 99129-99149. https://doi.org/10.1109/ACCESS.2022.3207287

Mounce, S., Ellis, K., Edwards, J., Speight, V., Jakomis, N., & Boxall, J. (2017). Ensemble decision tree models using RUSBoost for estimating risk of iron failure in drinking water distribution systems. *Water Resources Management*, *31*, 1575-1589. https://doi.org/10.1007/s11269-017-1595-8

Rahi, S., Mobli, H., Jamshidi, B., Azizi, A., & Sharifi, M. (2020). Different supervised and unsupervised classification approaches based on visible/near infrared spectral analysis for discrimination of microbial contaminated lettuce samples: Case study on E. coli ATCC. *Infrared Physics & Technology*, *108*, 103355. https://doi.org/10.1016/j.infrared.2020.103355

Torres, I., Sánchez, M.-T., Vega-Castellote, M., Luqui-Muñoz, N., & Pérez-Marín, D. (2021). Routine NIRS analysis methodology to predict quality and safety indexes in spinach plants during their growing season in the field. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *246*, 118972. https://doi.org/10.1016/j.saa.2020.118972

Wiedemair, V., Langore, D., Garsleitner, R., Dillinger, K., & Huck, C. (2019). Investigations into the performance of a novel pocket-sized near-infrared spectrometer for cheese analysis. *Molecules*, *24*(3), 428. https://doi.org/10.3390/molecules24030428

Wold, S., Antti, H., Lindgren, F., & Öhman, J. (1998). Orthogonal signal correction of near-infrared spectra. *Chemometrics and intelligent laboratory systems*, *44*(1-2), 175-185. https://doi.org/10.1016/S0169-7439(98)00109-9

WHO (World Health Organization). (1978). *Nitrates, nitrites and N–nitroso compounds* (Environmental Health Criteria No. 5).