SegundoVillalta_articulo.pdf

Universidad Peruana Union

Detalles del documento

Identificador de la entrega

trn:oid:::29566:440186458

Fecha de entrega

17 mar 2025, 12:04 p.m. GMT-5

Fecha de descarga

17 mar 2025, 12:45 p.m. GMT-5

Nombre de archivo

SegundoVillalta_articulo.pdf

Tamaño de archivo

661.2 KB

11 Páginas

5265 Palabras

29.179 Caracteres

15% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Exclusions

- 560 Excluded Sources

Match Groups

75 Not Cited or Quoted 14%
Matches with neither in-text citation nor quotation marks

4 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

11% Internet sources

9% Publications

10% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **75** Not Cited or Quoted 14%
  Matches with neither in-text citation nor quotation marks
- **4** Missing Quotations 1%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

11% Internet sources
9% Publications
10% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.nal.usda.gov	<1%
2	Internet	enauka.gov.rs	<1%
3	Internet	mail.vrg.org	<1%
4	Internet	mts.intechopen.com	<1%
5	Internet	sciencetechindonesia.com	<1%
6	Internet	www.ukaazpublications.com	<1%
7	Publication	Hoeil Chung. "Applications of Near-Infrared Spectroscopy in Refineries and Impor...	<1%
8	Submitted works	IIT Delhi on 2022-04-06	<1%
9	Internet	ashs2024.sched.com	<1%
10	Internet	jurnal.ugm.ac.id	<1%

11 Internet

ri.diva-portal.org <1%

12 Internet

squ.elsevierpure.com <1%

13 Publication

Mehrad Mortazavi, Arjan Bains, Leili Afsah-Hejri, Reza Ehsani, Patricia J. LiWang. "... <1%

14 Publication

Zhan Li, Cheng Li, Yue Gao, Wenguang Ma, Yunye Zheng, Yongzhi Niu, Yajing Gua... <1%

15 Internet

eng.au.dk <1%

16 Internet

personal.article.dyndns.info <1%

17 Internet

search.crossref.org <1%

18 Internet

www.cbi.eu <1%

19 Publication

Bouzid, Emna. "Design and Implementation of a Machine Learning Predictive Mo... <1%

20 Submitted works

Colorado State University, Global Campus on 2023-08-10 <1%

21 Submitted works

Kingston University on 2023-03-31 <1%

22 Submitted works

University of Greenwich on 2024-04-12 <1%

23 Internet

ddescholar.acemap.info <1%

24 Submitted works

Griffth University on 2024-02-05 <1%

25 Publication

Liu Jiang, Kang Zheng, Xinyue Hou, Ibrahim A. Darwish. "Maillard-induced chia se... <1%

26 Submitted works

Massey University on 2024-09-15 <1%

27 Publication

P.V. Mohanan. "Artificial Intelligence and Biological Sciences", CRC Press, 2025 <1%

28 Publication

Robert Senga, Mahmoud Nasr, Manabu Fujii, Amal Abdelhaleem. "Sustainable val... <1%

29 Internet

research-repository.uwa.edu.au <1%

30 Internet

scirp.org <1%

31 Internet

www.sju.edu.in <1%

32 Publication

Mohamed Djamel Miara, Sami Youssef, Yassine Fendane, Hassan Boukcim et al. "... <1%

33 Internet

research-management.mq.edu.au <1%

34 Publication

"Dietary Phytochemicals", Springer Science and Business Media LLC, 2021 <1%

35 Internet

baadalsg.inflibnet.ac.in <1%

36 Internet

bmcbiotechnol.biomedcentral.com <1%

37 Internet

centaur.reading.ac.uk <1%

38 Internet

jap.ut.ac.ir <1%

39 Publication

Cristiana S. A. M. Maduro Dias, Helder Nunes, Carlos Vouzela, João Madruga, Alfre... <1%

40 Publication

Randall Bluffstone. "Environmental Economics and Ecosystem Services", Routledg... <1%

41 Internet

olis.com.ua <1%

42 Internet

repositorio.ucv.edu.pe <1%

43 Internet

www.erdkunde.uni-bonn.de <1%

44 Internet

www.esciencecentral.org <1%

45 Submitted works

CSU, Fullerton on 2024-01-29 <1%

46 Publication

Casado, C.. "Modelling the lactation curve of rabbit does: Towards a model includi... <1%

47 Internet

academic.oup.com <1%

48 Internet

cris.maastrichtuniversity.nl <1%

49 Internet

digitalcommons.andrews.edu <1%

50 Internet

malque.pub <1%

51 Internet

spatial-accuracy.org <1%

52 Publication

Brian G. Osborne. "Near-Infrared Spectroscopy in Food Analysis", Wiley, 2006 <1%

53 Publication

Dany-Yudet Millones-Liza, Elizabeth-Emperatriz Garcia-Salirrosas. "Development ... <1%

54 Submitted works

Higher Education Commission Pakistan on 2013-12-24 <1%

55 Publication

Poddar, Sukanya. "In Vitro Antihypertensive Potentials of Extensively Hydrolyzed ... <1%

56 Submitted works

Queen's University of Belfast on 2021-09-21 <1%

57 Submitted works

University College London on 2025-03-15 <1%

58 Submitted works

University of New South Wales on 2015-06-20 <1%

59 Internet

mrforum.com <1%

60 Internet

nationalsheep.org.uk <1%

61 Internet

oamjms.eu <1%

62 Internet

repositorio.autonoma.edu.pe <1%

63 Internet

repositorio.una.ac.cr <1%

64 Internet

researcherslinks.com <1%

65 Internet

revistas.unsm.edu.pe <1%

66 Internet

www.sciopen.com <1%

67 Submitted works

University of Durham on 2021-09-02 <1%

68 Submitted works

University of Stellenbosch, South Africa on 2017-09-13 <1%

69 Submitted works

University of Greenwich on 2024-09-06 <1%

Application of machine learning in near-infrared spectra for moisture prediction in ground cumin

Segundo Villalta[1,*] **and Javier Linkolk López-Gonzales**[1,*]

[1]Escuela de Posgrado, Universidad Peruana Unión, Lima, Peru

ABSTRACT

Cumin (Cuminum cyminum L.) is one of the most valued spices worldwide, and moisture content is a critical parameter for assessing its quality and safety. This study aimed to evaluate the performance of machine learning algorithms in the analysis of near-infrared (NIR) spectra for moisture prediction in ground cumin. Six cumin samples were collected and NIR spectra were obtained in the range of 1100-2100 nm. The spectra, expressed as absorbance, were pretreated using Savitzky-Golay (SG) and normalization independently, followed by optimization using principal component analysis (PCA). The complete and optimized data were modeled using multiple linear regression (MLR) and quadratic support vector machines (QSVM). The models were trained using 5-way cross-validation with 30 replicates, evaluating their performance using the coefficient of determination (R^2) and root mean square error (RMSE). On the full dataset, QSVM with SG pretreatment performed best (R^2 = 0.99; RMSE = 0.0582). For the optimized data, QSVM combined with normalization performed outstandingly (R^2 = 0.97; RMSE = 0.0919). These findings highlight that QSVM, thanks to its ability to capture nonlinear relationships, offers more accurate and robust predictions compared to MLR. This study suggests exploring advanced techniques based on deep learning and/or complementary algorithms that can improve the performance of predictive models.

1 Introduction

Cumin (Cuminum cyminum L.) is one of the most widely used spices in gastronomy due to its distinctive spicy, strong and warm aroma and flavour (Chaudhry et al., 2020). Cumin has also been reported to contain a wide variety of phytochemicals that provide it with various biological activities, such as antimicrobial, antioxidant, anti-inflammatory, anticancer, antiproliferative, hepatoprotective and immunomodulatory properties (Mohammed et al., 2024). For this reason, cumin is used in various industrial sectors, including food and beverages, cosmetics and medicines (Chaudhry et al., 2020).

Cumin is usually marketed in powder form and one of the characteristics of this type of product is its low humidity, $\leq 10\%$ according to the Food and Agriculture Organization of the United Nations (FAO), 2022). This limits microbial development and guarantees a long shelf life. However, this condition is not always met due to inadequate processing techniques. In addition to cumin's susceptibility, it can be easily contaminated by pathogenic bacteria such as Bacillus cereus, Escherichia coli, Clostridium perfringens and Salmonella (Tannir et al., 2024).

To address this problem, continuous moisture analysis must be performed, where the standard technique is by heating and calculating weight loss; however, it is laborious, slow, consumes a lot of energy and destroys the samples (Tugnolo et al., 2021). In this context, near-infrared (NIR) spectroscopy emerges as a promising alternative as it is fast, simple, non-destructive, accurate, does not require reagents or prior sample preparation (Dias et al., 2021). NIR spectroscopy is based on the absorption of waves within the range 780 to 2500 nm concerning hydrogen-containing functional groups such as C–H, O–H, N–H and C–O, linked to the chemical composition and structure of the products (Ozturk et al., 2023; Zhang et al., 2022a).

A disadvantage of NIR spectroscopy is the large number of variables it yields, its multicollinearity, and its nonlinear relationship with the complex chemical composition of foods (Ozturk et al., 2023). Therefore, robust multivariate methods such as machine learning are required for spectral analysis, whose algorithms have been successfully combined with NIRS in food product evaluations (Zhang et al., 2022b). Despite the power of machine learning techniques, their capacity is subject to the type and amount of input data; therefore, other algorithms must be integrated for: a) preprocessing, to remove instrumental noise and outliers; and b) feature selection, to identify and extract a small subset of relevant variables (Zhang et al., 2022b).

There is no single approach for the treatment and analysis of spectral information (Fatemi et al., 2022); therefore, at least two algorithms should be tested and compared for each task. In this context, the aim of this work was to evaluate the performance of machine learning algorithms in the analysis of NIR spectra for the prediction of moisture in ground cumin.

2 Materials and methods

2.1 Raw material

Cumin powder (Cuminum cyminum) was purchased from six (6) different points of sale at the Bellavista market in Sullana, Piura, Peru in November 2023. It was visually ensured that the powder was free of other agents and/or external substances. The samples were stored in high-density polyethylene bags, under shade and in low humidity conditions.

2.2 NIR spectra acquisition

For spectra acquisition, a NIR spectrometer (Polytec PSS-A-T01, Polytec GmbH, Germany) equipped with a tungsten halogen lamp as light source and an Indium-Gallium-Arsenic detector was used. The equipment operated in the spectral range of 1100-2100 nm with a resolution of 2 nm. Prior to measurements, the device was turned on for a few minutes to stabilize the illumination and perform an internal calibration. Each cumin powder sample was placed on a continuously rotating platform, allowing the capture of 50 spectral profiles per sample. The fundamental principle of NIR spectroscopy is to irradiate the sample with light in the NIR range and record the reflected, absorbed or transmitted radiation (Li et al., 2020). Although the equipment generates spectra in terms of reflectance, the results are usually reported in absorbance in the literature. For this reason, absorbance spectra were obtained by transforming the inverse of the reflectance values into a logarithm.

2.3 Spectrum processing

An inherent disadvantage of NIR spectra is the presence of background interferences, including noise and overlapping bands, which can lead to redundancy, collinearity, and complex spectral structure unsuitable for direct analysis (Li et al., 2020). Therefore, preprocessing is critical to improve data quality.

The most commonly used methods in infrared spectroscopy include exclusion, filtering, derivatization, smoothing, baseline correction, and normalization (Mokari et al., 2023). Since selecting the most suitable pretreatment is challenging, two approaches were implemented in this study:

2.3.1 Savitzky-Golay (SG) smoothing

This method improves the spectrum quality by removing noise using numerical differentiation and a smoothing step based on a moving window. Key parameters include the window size (number of points) and the degree of the polynomial used for the fit (Mokari et al., 2023).

In this study, a second-order SG smoothing with a 21-point window was used.

2.3.2 Normalization

Spectral intensity can vary due to differences in measurement conditions, such as equipment alignment or light source power, even for the same material (Mokari et al., 2023). Normalization adjusts for these differences, ensuring comparability of spectra obtained under slightly different conditions. Furthermore, this method is useful for reducing fluctuations arising from physical characteristics of the sample, such as particle size or texture (Nagy et al., 2022).

In this study, scaling normalization was applied, in which each point of the centered spectrum was divided by its standard deviation, homogenizing the intensities and improving comparability.

2.4 Data development

Regression is one of the most popular, intuitive and simple forecasting methods, widely used in various applications (Etemadi & Khashei, 2021). The relationship between NIR spectra and response variables, such as moisture content, can be linear or non-linear. Therefore, the effectiveness of two different approaches addressing both possibilities was evaluated:

2.4.1 Regresión lineal múltiple

19 MLR is a supervised method frequently used to identify linear relationships between a dependent (predicted) variable and
19 multiple independent variables (predictors). This approach allows predictions to be made based on new values of the independent variables by fitting a linear equation (Etemadi & Khashei, 2021).

2.4.2 Quadratic support vector machine

Originally designed for classification problems, the SVM model has also been adapted for regression applications (Zeng
44 et al., 2021). SVM uses the structural risk minimization principle and Vapnik-Chervonenkis dimension theory to handle high-dimensional data with remarkable generalization and accuracy (Zeng et al., 2021). For nonlinear problems, a kernel
27 function is used to project the data into a higher-dimensional space, allowing to find and construct an optimal hyperplane that fits the data. In this study, a variant of SVM adapted for regression, known as QSVM, was used.

2.5 Model optimization

In multivariate analysis, datasets may contain variables with high collinearity, which makes interpretation and processing difficult. Reducing dimensionality by selecting the most relevant spectral features not only improves computational efficiency but also facilitates model training and validation (Mokari et al., 2023).

To address this issue, principal component analysis (PCA), an unsupervised method widely used in chemometrics, was implemented. PCA allows compressing spectral information by identifying underlying correlation patterns and projecting the data into a lower-dimensional space. This process generates new orthogonal axes called principal components (PC), ordered by the amount of variance explained. The first principal component (PC1) captures the largest variance in the dataset, followed by the second (PC2), and so on (Mokari et al., 2023).

Although PCA is not a modeling method by itself, principal components that explained 95% of the total variance were selected as inputs to build optimized models based on MLR or QSVM.

2.6 Calculating model performance

A 5-step cross-validation strategy (*cv*) with 30 repetitions will be followed for training and validation of the full and optimized
23 models. The prediction ability of the models was evaluated using the coefficient of determination (R^2) and the root mean square error (*RMSE*) defined by Equations 1 and 2, respectively,

45 $$R^2_{cv} = \frac{\sum_{i=1}^{n} (\bar{y}_i - y_i)^2}{\sqrt{\sum_{i=1}^{n} (\bar{y}_i - y_t)^2}} \tag{1}$$

$$RMSE_{CV} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\bar{y}_1 - y_i)^2} \tag{2}$$

where $\bar{y}_i$ and y_i are the response variables of the i-th (i^{th}) sample for prediction and reference and n is the number of samples.

68 Moisture content was determined by oven drying following the indications of the official methods of the Association of Official Analytical Chemists (AOAC International, 2016). The analyses were performed in triplicate and expressed in %.

2.7 Data analysis

28 Moisture analyses were presented as mean ± standard deviation using box-and-whisker plots. One-way analysis of variance
58 (ANOVA) was used for hypothesis testing at a significance level of 5%, and Tukey's post hoc test at 5% was used to determine significant differences between samples. Tests were run in Minitab 18.0 (Minitab Inc., USA). Spectra preprocessing and feature selection will be performed in Unscrambler X 10.4 (CAMO Software, Norway) and Weka (University of Waikato, New Zealand), respectively. Development, training, and validation of regression models were conducted in Matlab R2023a (The MathWorks, Inc., USA).

Spectra preprocessing and feature selection were performed in Unscrambler X 10.4 (CAMO Software, Norway) and Weka (University of Waikato, New Zealand), respectively. Model development, training and validation were conducted in Matlab R2023a (The MathWorks, Inc., USA).

30 3 Results and discussion

3.1 Moisture content.

The moisture content of the samples showed a slight variation (Figure 1), with values ranging between 9.11 ± 0.00% and 10.55
62 ± 0.01%. Significant differences ($p < 0.05$) were found between the samples, except between those with means of 10.55% and 10.51%. According to the Peruvian Table of Food Composition, the average moisture content in cumin is 8.1% (Reyes García et al., 2017), which shows some proximity to the results obtained in this study.

On the other hand, FAO (2022) indicates that cumin powder must contain less than 10% moisture. In this work, four samples exceeded this limit, reaching up to 10.03%, which could be attributed to an inefficient drying process. Although the difference is minimal, a higher moisture content can favor the proliferation of microorganisms, compromising the quality of the product. In general, spices are susceptible to mold attack, which, under high humidity conditions, can produce dangerous compounds such as mycotoxins, with serious implications for human health (Cicero et al., 2020).

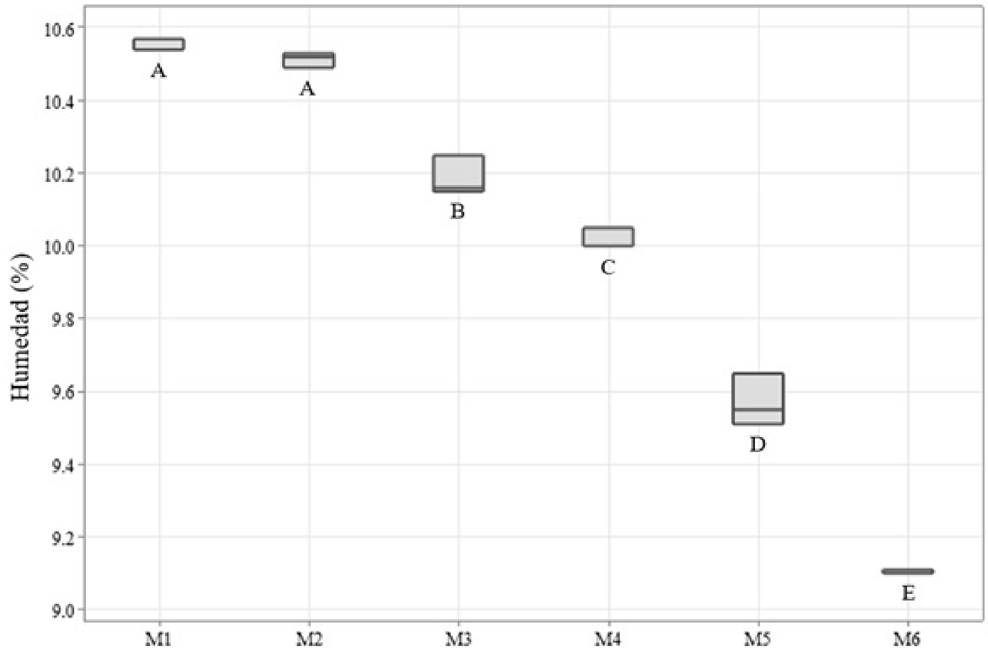


Figure 1. Moisture content of cumin samples.

3.2 Original and pre-treated NIR spectra

Figure 2 presents the NIR spectra of the cumin samples in different formats: in reflectance (**Figura 2(a)**) and in absorbance (**Figura 2(b)**). The behavior observed in these spectra is consistent with previous studies (de Lima et al., 2020; Florian-Huamán et al., 2022). The differences in intensities are due to the superposition of overtone bands related to molecular rotational and vibrational transitions, characteristic of each cumin sample (Li et al., 2020).

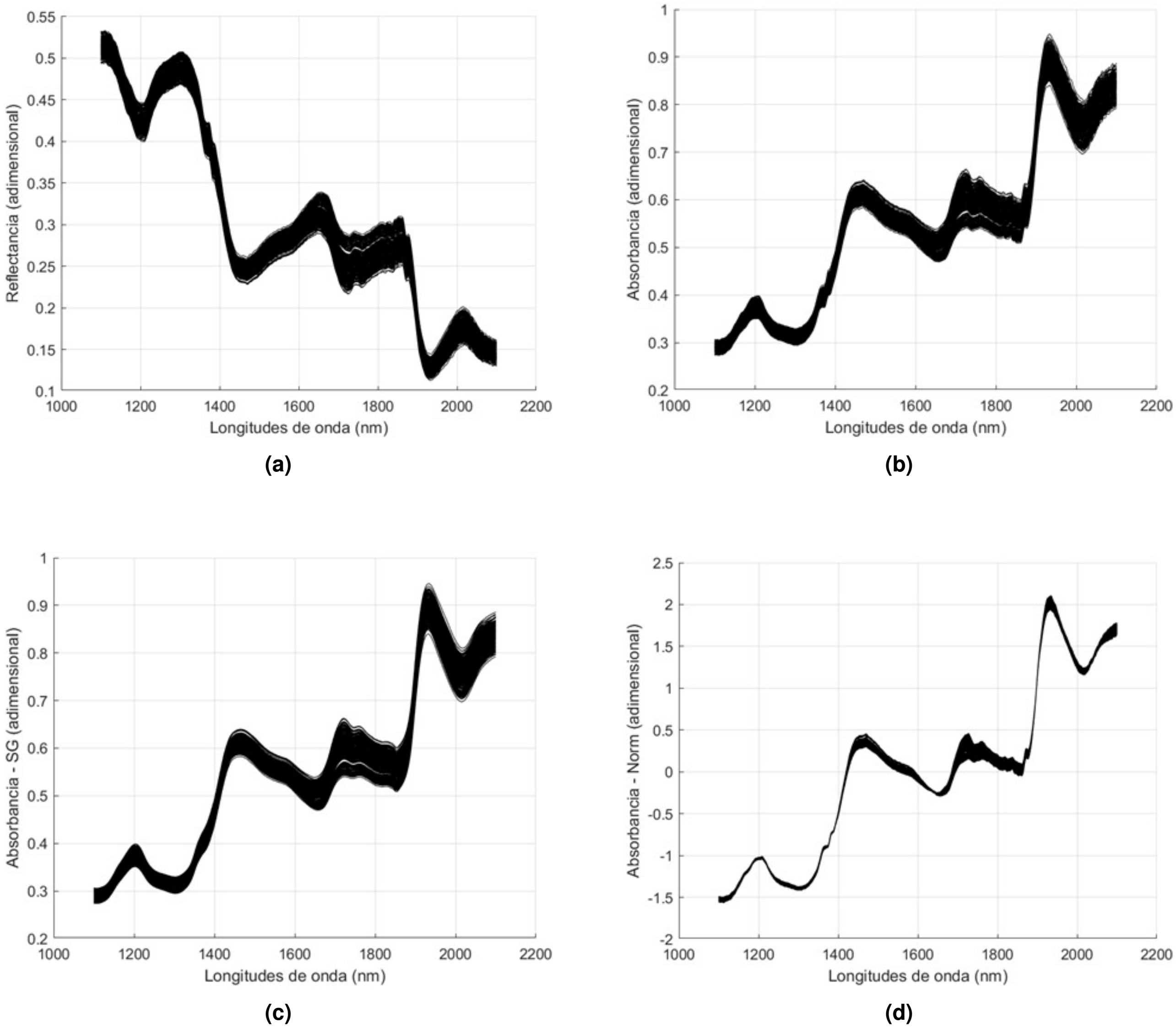


Figure 2. NIR spectra of cumin samples in (a) reflectance, (b) absorbance, (c) SG-pretreated absorbance, and (d) normalized absorbance mode.

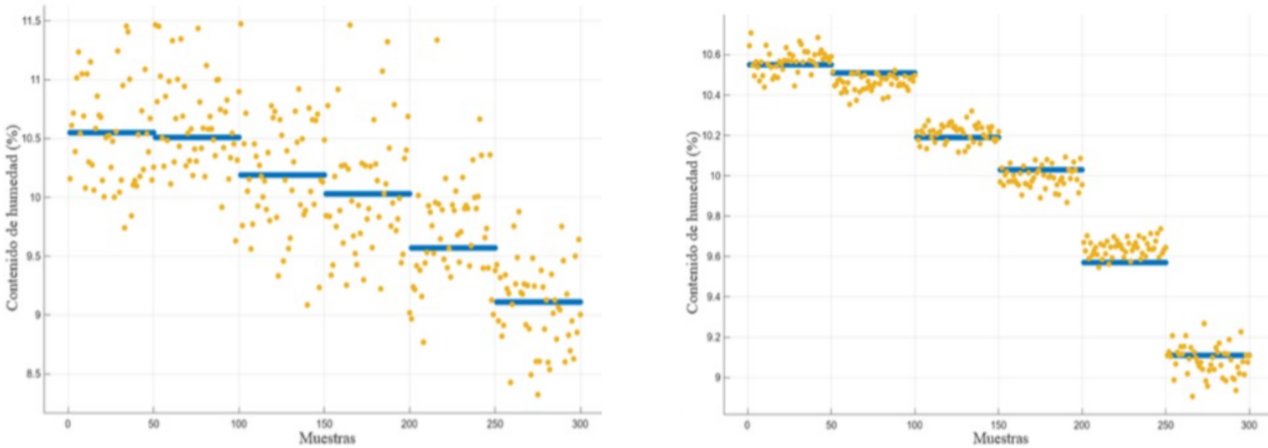
Figura 2(c) shows the absorbance spectrum smoothed with the Savitzky-Golay (SG) algorithm, evidencing a notable decrease in noise, especially in the 1700 to 2100 nm region. In contrast, the normalized spectrum is presented in **Figura 2(d)**. Although the normalization process corrects certain discrepancies, it can also result in a loss of spectral information. The effectiveness of each pretreatment will be evaluated in the metrics of the predictive models, since their impact varies depending on the dataset (Nagy et al., 2022). In some cases, pretreatments can even worsen the modeling performance.

The peaks at around 1200 nm, 1460 nm, 1720 nm and 1760 nm are associated with C –H bonds of methylene and essential oils, first O –H overtone associated with water or cellulosic compounds, C – H2 bands, N – H bonds of polyamides and flavone, as well as bands linked to the content of lipids, fatty acids, amino acids, lignin, aliphatic hydrocarbon and polyphenols (de Lima et al., 2020; Florian-Huamán et al., 2022; Yang et al., 2021).

3.3 Full MLR and QSVM models

Figure 3 presents the response graphs of the MLR and QSVM models for the prediction of moisture content in the cumin samples.

In **Figure 3**, corresponding to the MLR model, a notable dispersion around the predicted moisture values is observed. This suggests that the MLR model does not adequately capture the variations in the data, probably due to its linear nature. This limits its ability to fit possible non-linear relationships between NIR spectra and moisture content, which significantly affects its predictive accuracy.



(a) (b)

Figure 3. Response graph of (a) MLR and (b) QSVM.

On the other hand, **Figura 3(b)** showing the results of the QSVM model shows a lower dispersion at each moisture level. This indicates a superior ability of the QSVM to capture the non-linear features present in the data. The quadratic structure of the QSVM model, combined with the use of kernel functions, allows for a better adaptation to the inherent complexities in the spectral relationship between NIR and moisture content. This finding is in agreement with previous studies, such as that of Goyal et al. (2024), who pointed out that NIR spectra are highly non-linear due to the combined bands, reinforcing the advantage of non-linear models in this context.

Comparative studies also support these observations. For example, Sanaiefar et al. (2018) evaluated different models (MLR, SVM with kernel, artificial neural networks, and Bayesian networks) in predicting olive oil quality from dielectric spectra, finding that SVM consistently outperformed the other approaches. Likewise, Zhang et al. (2016) highlighted the superiority of the LS-SVM model in predicting aspartic acid in rapeseed leaves using NIR spectra, compared to MLR and PLSR.

3.4 PCA

Figure 4 presents the PCA applied to the NIR spectra under different pretreatment conditions: raw, SG-smoothed and normalized. In **Figure 4(a)** the PCA applied to the raw spectra shows that the first two principal components (PC1 and PC2) explain 97.99% of the total variance. However, a significant overlap is observed between samples with different moisture levels. This suggests that the original data contain noise, scattering or other effects that may make it difficult to identify clear patterns related to moisture content.

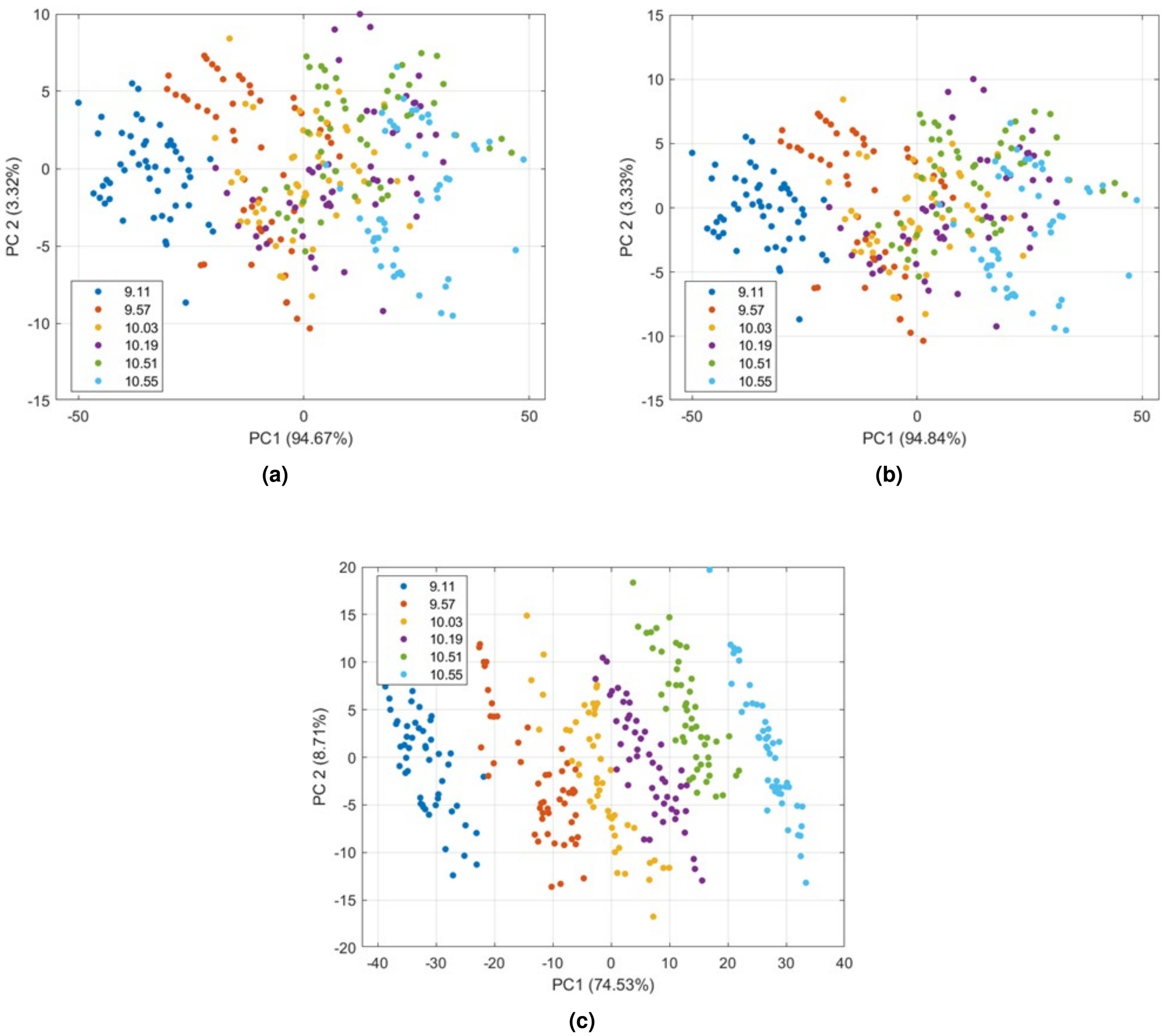


Figure 4. PCA plot of (a) raw, (b) SG pretreated and (c) Normalized spectra.

Figure 4(b) corresponding to the PCA applied to the SG-smoothed data shows similar results, explaining 98.17% of the total variance. Despite the noise reduction in the spectra due to the SG pretreatment, no significant improvement is achieved in the separation between humidity groups, indicating that this pretreatment did not introduce significant changes in the structure of the data.

On the other hand, **Figure 4 (c)** based on normalized data shows that the two principal components explain a smaller percentage of the total variance (83.24%). However, normalization allowed for a better separation between cumin samples with different moisture levels. This can be attributed to the fact that normalization removed some of the redundant variability in the spectra, allowing the PCA to focus on more relevant features to distinguish the groups. Despite this, this process also reduced the variance captured by PC1 and PC2, evidencing the trade-off between noise removal and preservation of total spectral variability.

These findings highlight the utility of PCA as a tool to assess patterns and the influence of pretreatments. According to Goyal et al. (2024), PCA is effective in identifying general patterns in naturally structured data, although these can be affected by specific pretreatments. Gomes et al. (2023) also highlighted that exploratory analysis using PCA can provide valuable preliminary information to define the relevance of data in subsequent modeling. Thus, combining PCA with pretreatment assessments allows for quick decisions on preparing data for more accurate predictive models.

3.5 Optimized MLR and QSVM models

Table 1 shows the summary of the results of the MLR and QSVM models for the prediction of the moisture content of the cumin samples.

Dataset		MLR		QSVM	
		R^2	RMSE	R^2	RMSE
Completo	Crudo	0.41	0.397	0.98	0.0678
	SG	0.46	0.3674	**0.99**	**0.0582**
	Norm	0.24	0.4482	0.98	0.0649
Optimizado	Crudo	0.77	0.2457	0.79	0.2365
	SG	0.77	0.2452	0.79	0.2354
	Norm	0.95	0.1144	**0.97**	**0.0919**

Table 1. Metrics of the developed models.

SG pretreatment showed minimal improvement in the MLR model performance, probably due to its ability to increase the signal-to-noise ratio without removing relevant information from the spectra (Goyal et al., 2024). Similar results were observed in the modelling of fresh fruits (Mishra et al., 2021). On the other hand, normalization deteriorated the model performance, possibly because by removing intensity variations between spectra (Goyal et al., 2024), crucial information for prediction was lost, as evidenced in the authentication of paprika varieties (Kolasinac et al., 2022). In the case of QSVM models, normalization did not significantly affect the accuracy when using complete data, while SG pretreatment slightly increased the accuracy.

With PCA-optimized data, MLR models showed substantial improvement, suggesting that dimensionality reduction enabled capturing clearer patterns in the data. PCA likely removed uninformative variables and reduced noise, facilitating a better fit of the MLR model (Mokari et al., 2023). In contrast, the QSVM model showed a reduction in accuracy when using raw or SG-pretreated data, indicating a possible omission of relevant variables. However, the performance of the QSVM with optimized and normalized data was comparable to that obtained with full data, suggesting that this model can handle the high dimensionality of the spectra. From a practical perspective, the use of PCA-optimized data represents a viable solution without compromising model performance.

Overall, QSVM models significantly outperformed MLR models, with R^2 values ranging from 0.98 to 0.99 and RMSEs ranging from 0.0678 to 0.0582 when using full data. With PCA-optimized data, R^2s ranged from 0.79 to 0.97 and RMSEs ranged from 0.2365 to 0.0919. These differences underscore the superior ability of QSVM to model nonlinear relationships in the data.

Regarding the comparison between linear and non-linear models, a common limitation in the literature is the use of less robust calibration and validation methods, such as the division of small data sets assuming independence (Rocha et al., 2020). Therefore, the implementation of more rigorous approaches, such as cross-validation, applied in this study, is recommended.

The superiority of the QSVM model is due not only to its ability to handle non-linear data, but also to the intrinsic properties of the SVM. This approach optimizes generalization and minimizes errors arising from unanticipated data during training, in addition to mitigating overfitting (Siripatrawan & Makino, 2024). For example, Yang et al. (2022) showed that, depending on preprocessing, linear SVM models outperformed KNN models in identifying illegal drugs in food, and that quadratic SVMs were superior to their linear versions.

Although QSVM showed promising results, other approaches could also be explored. Abdel-Sattar et al. (2021) reported that quadratic SVM outperformed MLR in predicting ber fruit mass, but ANNs performed even better. Deep learning has also been shown to be superior in many cases (Zhou et al., 2019). For example, Bai et al. (2024) employed NIR spectra to estimate the quality of Chinese medicinal herbs, where a Bayesian optimized long-term memory model outperformed other approaches, including ANNs.

It should be noted that not in all cases the non-linear model is superior. For example, Jin et al. (2023) found that PLSR was more effective than SVM in detecting lamb freshness. Furthermore, linear models such as MLR can be improved by modifications in hyperparameters or by incorporating additional algorithms, such as LASSO or ridge regression, which also address overfitting issues (Goyal et al., 2024).

46 In the present study, the QSVM with SG pretreatment provided the best fit to the full data (R^2 = 0.99 and RMSE = 0.0582), as shown in **Figure 5(a)**.

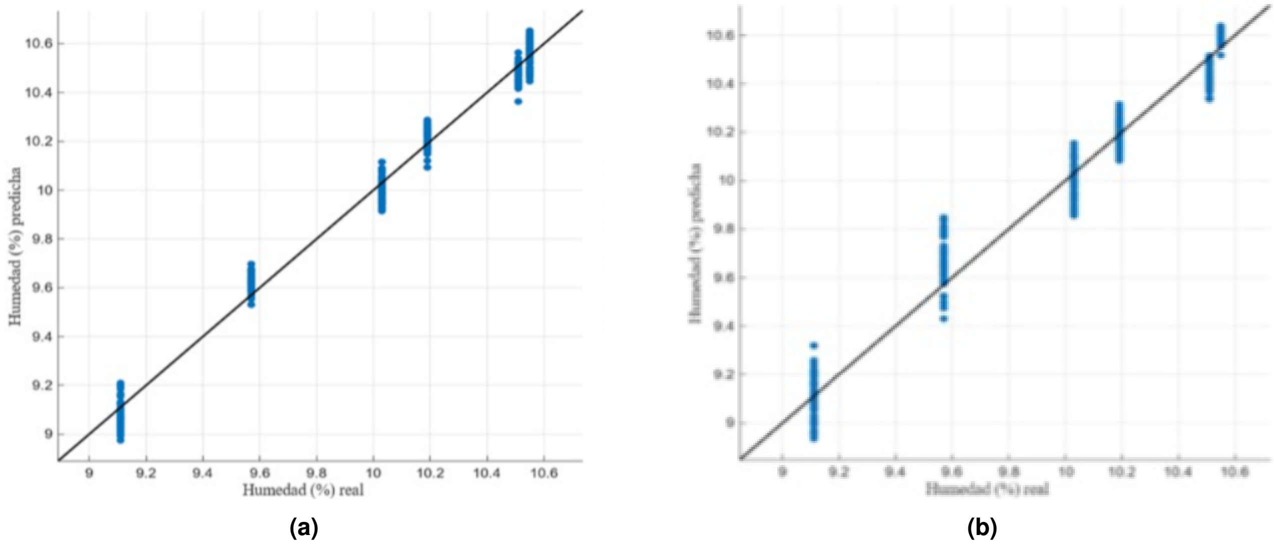


Figure 5. Improved QSVM models of (a) SG-pretreated full spectra and (b) normalized and optimized spectra.

Figure 5(b) reveals the best fit for the PCA-optimized dataset, using QSVM and a prior normalization (R^2 = 0.97 and RMSE = 0.0919). Mokari et al. (2023) reported that, PCA is sensitive to the scale of the data; thus, data normalization is crucial, particularly if spectroscopic data are explored.

Since the QSVM model using all wavelengths had similar performance to the optimized QSVM model, the latter can be integrated into low-cost portable devices for real-time analysis (Nagy et al., 2022), contributing to the development of innovative and accessible tools to address food insecurity.

4 Conclusions

The moisture prediction performance of ground cumin was evaluated using preprocessing such as SG and normalization, combined with MLR and QSVM models applied on NIR spectra. The analyzed samples showed moisture levels between 9.11% and 10.55%, some above the maximum recommended limit (10%), which could compromise product quality and consumer safety. SG pretreatment improved the performance of both models by increasing the signal-to-noise ratio, while normalization negatively affected the performance of MLR and did not impact QSVM. On the other hand, PCA significantly improved the results by removing non-informative variables.

54 Overall, for the full data, QSVM with SG presented the best performance (R^2 = 0.99 and RMSE = 0.0582). For the optimized data, QSVM with prenormalization was the most accurate (R^2 = 0.97 and RMSE = 0.0919). These results highlight QSVM as a robust and accurate tool for moisture prediction in ground cumin, with potential applications in portable devices and real-time routine analysis.

References

1. Abdel-Sattar, M., Aboukarima, A. M., & Alnahdi, B. M. (2021). Application of artificial neural network and support vector regression in predicting mass of ber fruits (Ziziphus mauritiana Lamk.) based on fruit axial dimensions. *PLoS ONE*, 16(1), e0245228. https://doi.org/10.1371/journal.pone.0245228 [43]
2. Agencia Agraria de Noticias. (2021). Perú importó comino por US$ 6 millones durante 2020. https://agraria.pe/index.php/noticias/peru-importo-comino-por-us-6-millones-durante-2020-23423
3. Bai, L., Zhang, Z.-T., Guan, H., Liu, W., Chen, L., Yuan, D., Chen, P., Xue, M., & Yan, G. (2024). Rapid and accurate quality evaluation of Angelicae Sinensis Radix based on near-infrared spectroscopy and Bayesian optimized LSTM network. *Talanta*, 275, 126098. https://doi.org/10.1016/j.talanta.2024.126098 [14] [11]
4. CBI. (2023). The European market potential for cumin seeds. https://www.cbi.eu/market-information/spices-herbs/cumin/market-potential [18]
5. Chaudhry, Z., Khera, R. A., Hanif, M. A., Ayub, M. A., & Sumrra, S. H. (2020). Cumin. In Hanif, M. A., Nawaz, H., Khan, M. M., & Byrne, H. J. (Eds.), *Medicinal Plants of South Asia* (pp. 165-178). Elsevier. https://doi.org/10.1016/B978-0-08-102659-5.00013-6 [33] [12]

6. Chen, C., Yang, B., si, R., Chen, C., Chen, F., Gao, R., Li, Y., Tang, J., & Lv, X. (2021). Fast detection of cumin and fennel
59 using NIR spectroscopy combined with deep learning algorithms. *Optik*, 242, 167080. https://doi.org/10.1016/j.ijleo.2021.167080

32 7. Cicero, N., Gervasi, T., Durazzo, A., Lucarini, M., Macri, A., Nava, V., Giarratana, F., Tardugno, R., Vadala, R., &
22 Santini, A. (2020). Mineral and Microbiological Analysis of Spices and Aromatic Herbs. *Foods*, 11(4), 548. https://doi.org/10.3390/foods11040548

8. de Lima, A. B. S., Batista, A. S., de Jesus, J. C., Silva, J. J., de Araújo, A. C. M., & Santos, L. S. (2020). Fast quantitative detection of black pepper and cumin adulterations by nearinfrared spectroscopy and multivariate modeling. *Food Control*,
61 107, 106802. https://doi.org/10.1016/j.foodcont.2019.106802

50 9. Dias, C. S. A. M. M., Nunes, H. P., Melo, T. M. M. V., Rosa, H. J. D., Silva, C. C. J., & Borba, A. E. S. (2021). Application of
7 Near Infrared Reflectance (NIR) spectroscopy to predict the moisture, protein, and fat content of beef for gourmet hamburger
60 preparation. *Livestock Science*, 254, 104772. https://doi.org/10.1016/j.livsci.2021.104772

10. FAO. (2022). Codex Alimentarius: Normas para el comino CXS 327-2017, adoptada en 2017, enmendada en 2021,
36 2022. https://www.fao.org/fao-who-codexalimentarius/committees/committee-detail/related-standards/es/?committee=CCSCH

16 11. Fatemi, A., Singh, V., & Kamruzzaman, M. (2022). Identification of informative spectral ranges for predicting major
66 chemical constituents in corn using NIR spectroscopy. *Food Chemistry*, 383, 132442. https://doi.org/10.1016/j.foodchem.2022.132442

47 12. Galvin-King, P., Haughey, S. A., & Elliot, C. T. (2021). Garlic adulteration detection using NIR and FTIR spectroscopy
37 and chemometrics. *Journal of Food Composition and Analysis*, 96, 103757. https://doi.org/10.1016/j.jfca.2020.103757

13. Gomes, A. A., Azcarate, S. M., Spanik, I., Khvalbota, L., & Goicoechea, H. C. (2023). Pattern recognition techniques in
24 food quality and authenticity: A guide on how to process multivariate data in food analysis. *TrAC Trends in Analytical Chemistry*, 164, 117105. https://doi.org/10.1016/j.trac.2023.117105

14. Goyal, R., Singha, P., Singh, S. K. (2024). Spectroscopic food adulteration detection using machine learning: Current
31 challenges and future prospects. *Trends in Food Science & Technology*, 146, 104377. https://doi.org/10.1016/j.tifs.2024.104377

15. Jin, P., Fu, Y., Niu, R., Zhang, Q., Zhang, M., Li, Z., & Zhang, X. (2023). Non-Destructive Detection of the Freshness
57 of Air-Modified Mutton Based on Near-Infrared Spectroscopy. *Foods*, 12(14), 2756. https://doi.org/10.3390/foods12142756

2 16. Kolasinac, S., Pécinar, I., Danojevic, D., & Stevanovic, Z. D. (2022). Raman spectroscopy coupled with chemometric
25 modeling approaches for authentication of different paprika varieties at physiological maturity. *LWT – Food Science and Technology*, 162, 113402. https://doi.org/10.1016/j.lwt.2022.113402

17. Liu, Z., Zhang, R., Yang, C., Hu, B., Luo, X., Li, Y., & Dong, C. (2022). Research on moisture content detection method
8 during green tea processing based on machine vision and near-infrared spectroscopy technology. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 271, 120921. https://doi.org/10.1016/j.saa.2022.120921

18. Malvandi, A., Feng, H., & Kamruzzaman, M. (2022). Application of NIR spectroscopy and multivariate analysis for
4 Nondestructive evaluation of apple moisture content during ultrasonic drying. *Spectrochimica Acta Part A: Molecular and
65 Biomolecular Spectroscopy*, 269, 120733. https://doi.org/10.1016/j.saa.2021.120733

19. Mishra, P., Rutledge, D. N., Roger, J.-M., Wali, K., & Khan, H. A. (2021). Chemometric pre-processing can negatively
1 affect the performance of near-infrared spectroscopy models for fruit quality prediction. *Talanta*, 229, 122303. https://doi.org/10.1016/j.talanta.2021.122303

5 20. Mohammed, F. S., Sevindik, M., Uysal, I., Çesko, C., & Koraqi, H. (2024). Chemical composition, biological activities,
63 uses, nutritional and mineral contents of cumin (Cuminum cyminum). *Measurement: Food*, 14, 100157. https://doi.org/10.1016/j.meafoo.2024.100157

21. Ozturk, S., Bowler, A., Rady, A., & Watson N. J. (2023). Near-infrared spectroscopy and machine learning for classification
34 of food powders during a continuous process. *Journal of Food Engineering*, 341, 111339. https://doi.org/10.1016/j.jfoodeng.2022.111339

22. Pu, Y., Pérez-Marín, D., O'Shea, N., & Garrido-Varo, A. (2021). Recent Advances in Portable and Handheld NIR Spectrometers and Applications in Milk, Cheese and Dairy Powders. *Foods*, 10, 2377. https://doi.org/10.3390/foods10102377

3 **23.** Reyes García, M., Gómez-Sánchez Prieto, I., & Espinoza Barrientos, C. (2017). Tablas peruanas de composición de
42 alimentos. Ministerio de Salud, Instituto Nacional de Salud. https://cdn.www.gob.pe/uploads/document/
file/4565836/Tablas-peruanas.pdf?v=1684253633

24. Rocha, W. F. C., do Prado, C. B., & Blonder, N. (2020). Comparison of Chemometric Problems in Food Analysis using Non-Linear Methods. *Molecules*, 25(13), 3025. https://doi.org/10.3390/molecules25133025

17 **25.** Sanaeifar, A., Jafari, A., & Golmakani, M.-T. (2018). Fusion of dielectric spectroscopy and computer vision for quality
15 characterization of olive oil during storage. *Computers and Electronics in Agriculture*, 145, 142-152. https://doi.
org/10.1016/j.compag.2017.12.035

26. Singh, N., Yadav, S. S., Kumar, S., & Narashiman, B. (2021). A review on traditional uses, phytochemistry, pharmacology, and clinical research of dietary spice Cuminum cyminum L. *Phytoterapy Research*, 35, 5007-5030. https://doi.org/10.1002/ptr.7133

27. Siripatrawan, U., & Makino, Y. (2024). Assessment of food safety risk using machine learning-assisted hyperspectral
26 imaging: Classification of fungal contamination levels in rice grain. *Microbial Risk Analysis*, 27-28, 100295. https:
//doi.org/10.1016/j.mran.2024.100295

28. Tannir, H., Debs, E., Mansour, F., & Chamas, S. (2024). Machine learning models for predicting quality in walnut oil:
11 Application of infrared spectroscopy and chemical composition. *Food Chemistry*, 399, 133974. https://doi.org/
10.1016/j.foodchem.2024.133974

29. Tugnolo, A., Giovenzana, V., Malegori, C., Oliveri, P., Casson, A., Curatitoli, M., Guidetti, R., & Beghi, R. (2021). A
7 reliable tool based on near-infrared spectroscopy for the monitoring of moisture content in roasted and ground coffee: A
56 comparative study with thermogravimetric analysis. *Food Control*, 130, 108312. https://doi.org/10.1016/j.
foodcont.2021.108312

30. Wang, Z., Wu, Q., & Kamruzzaman, M. (2022). Portable NIR spectroscopy and PLS based variable selection for
48 adulteration detection in quinoa flour. *Food Control*, 138, 108970. https://doi.org/10.1016/j.foodcont.
2022.108970

31. Yang, B., Chen, C., Chen, F., Chen, C., Tang, J., Gao, R., & Lv, X. (2021). Identification of cumin and fennel from different
4 regions based on generative adversarial networks and near infrared spectroscopy. *Spectrochimica Acta Part A: Molecular
1 and Biomolecular Spectroscopy*, 260, 119956. https://doi.org/10.1016/j.saa.2021.119956

1 **32.** Yang, F., Wang, C., Yu, H., Guo, Y., Cheng, Y., Yao, W., & Xie, Y. (2022). Establishment of the thin-layer chromatography-
surface-enhanced Raman spectroscopy and chemometrics method for simultaneous identification of eleven illegal drugs in
49 anti-rheumatic health food. *Food Bioscience*, 49, 101842. https://doi.org/10.1016/j.fbio.2022.101842

14 **33.** Zhang, S., Liu, S., Shen, L., Chen, S., He, L., & Liu, A. (2022a). Application of near-infrared spectroscopy for the
20 nondestructive analysis of wheat flour: A review. *Current Research in Food Science*, 5, 1305-1312. https://doi.org/
10.1016/j.crfs.2022.08.006

21 **34.** Zhang, W., Kasun, L. C., Wang, Q. J., Zheng, Y., & Lin, Z. (2022b). A Review of Machine Learning for Near-Infrared
69 Spectroscopy. *Sensors*, 22(24), 9764. https://doi.org/10.3390/s22249764

13 **35.** Zhou, L., Zhang, C., Liu, F., Qiu, Z., & He, Y. (2019). Application of Deep Learning in Food: A Review. *Comprehensive
Reviews in Food Science and Food Safety*, 18, 1793-1811. https://doi.org/10.1111/1541-4337.12492