Microchemical Journal 208 (2025) 112433

Contents lists available at ScienceDirect

Microchemical Journal

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

The impact of cluster resolution feature selection on pattern recognition and classification for detecting Sudan dye adulteration in palm oil

Joanna K. Kwao [a,c], Cheetham Mingle [b], John N. Addotey [d], Kwabena F.M. Opuni [a], Lawrence A. Adutwum [a,*]

[a] Department of Pharmaceutical Chemistry, School of Pharmacy, University of Ghana, P. O. Box LG 43, Legon, Accra, Ghana
[b] Food Physicochemical Laboratory, Food and Drugs Authority, P. O. Box CT 2783, Cantonments, Accra, Ghana
[c] Department of Chemistry and Biochemistry, University of South Carolina, Columbia, SC 29208, United States of America
[d] Department of Pharmaceutical Chemistry, Faculty of Pharmacy and Pharmaceutical Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

ARTICLE INFO

Keywords:
Chemometrics
Machine Learning
Adulteration
Palm Oil
Sudan Dyes
Feature Selection
FTIR

ABSTRACT

This study evaluates the performance of some commonly used chemometric and machine learning techniques such as principal component analysis (PCA), artificial neural network (ANN), k-nearest neighbors (KNN), logistic regression discriminant analysis (LRDA), partial least squares discriminant analysis (PLSDA), support vector machine (SVM), and gradient boosted decision tree (GBDT) on HATR – FTIR data for detecting Sudan dye adulteration in palm oil. We employed the Icoshift for data alignment and Savitzky-Golay smoothing to enhance the data quality. Cluster resolution feature selection (CRFS) selected 2.39 % of 3351 features. Using only the 80 selected features PCA models showed a clear separation between adulterated and pure palm oil samples and an improvement in explained variance which hitherto was not observed. LRDA, PLSDA and SVM showed improved training TPR, ACC and MCC after feature selection. KNN showed improvement all model quality parameters after feature selection.

1. Introduction

Adulteration of food with harmful chemical additives constitutes a major public health challenge [1–3]. This practice is often intentional and seeks to enhance the appearance and perceived quality of food products since the appearance of food is one of the easiest interpretable markers of quality for consumers [2,4] An example highlighted in literature is the adulteration of palm oil, ketchup, and chili products with Sudan dyes [5–7].

Sudan I, II, III, and IV are synthetic lipophilic dyes used in various industries and research laboratories as a coloring and staining agent. Structurally, they contain chromophoric azo groups (–N = N–) and extended π-conjugated systems, characterized by intense red color. This property makes these dyes attractive for use as food additives. However, evidence suggests that they are unsuitable for human consumption [1,2,4]. *In vitro* and *in vivo* studies have demonstrated the potential genotoxicity of Sudan dyes and their aromatic amine metabolites. Mechanisms of toxicity include formation of covalent DNA adducts, generation of reactive metabolites, and contact allergenicity [4,8]. Various international and independent regulatory agencies have recognized the potential toxicity of Sudan I, II, III and IV dyes and have banned or strictly limited their use as food additives [9–11].

Despite these findings, adulteration of palm oil with Sudan dyes remains persistent in Ghana. The visual appearance of palm oil is an easily interpretable quality marker for consumers, and this accounts for the frequent adulteration of market samples with Sudan dyes. Due to the limited information on Sudan dye carcinogenicity, no level of consumption can be considered safe. Therefore, routine evaluation of palm oil for the presence of Sudan dyes is quintessential for public safety.

Historically, the detection of Sudan dyes in palm oil has involved some form of chromatography [12–15]. These techniques are generally complex, time-consuming, and expensive. Qualitative analysis with spectroscopy is an effective and rapid alternative for detection and regulation of adulterated food products [16–18].

Spectroscopy, coupled with machine learning, data mining and chemometrics, is a rapid and non-destructive approach for detecting food adulteration [19–23]. These techniques have been applied for the authentication of honey [24–27], milk, meat [19,28,29] and other food

* Corresponding author.
E-mail address: adutwum@ug.edu.gh (L.A. Adutwum).

https://doi.org/10.1016/j.microc.2024.112433
Received 7 July 2024; Received in revised form 17 October 2024; Accepted 9 December 2024
Available online 12 December 2024

products. Chemometric methods (e.g. supervised and unsupervised) have been used to detect Sudan dyes in palm oil [22,23,30–32]. Data from spectroscopic analyses, including surface-enhanced Raman spectroscopy (SERS), visible-near infrared (VIS-NIR), have been analysed using unsupervised methods such as principal component analysis (PCA) and supervised methods such as linear discriminant analysis (LDA) and support vector machines (SVM) for the identification of samples adulterated with Sudan [22,23,30–32]. However, the models from these studies have limited capability for quality monitoring because they lack clear distinctions between pure and adulterated samples. Another study based on VIS-NIR data and PCA failed to provide external model validation, raising questions about their potential utility. To address this shortcoming, a subsequent study used a model classification rate which could be misleading when dealing with imbalanced data [33].

Despite the significant contributions of spectroscopy and chemometrics to the detection of Sudan dye adulterations in palm oil, only a few chemometric techniques have been studied. These are PCA, LDA and SVM. A more comprehensive evaluation of the most frequently used classification algorithms will facilitate the selection of an optimum tool for this application. This is important to eliminate false negatives in the detection of dangerous adulterants in food.

Additionally, it is also known that chemometric/machine learning models benefit from prior feature selection routines since it eliminates spurious sections of the spectra, which may impact the learning models adversely [34–36]. Unfortunately, the implementation of a feature selection routine before analysis in the case of Sudan dyes is conspicuously missing in the literature.

In this study, we compare some commonly used machine learning algorithms to determine the optimum algorithm for the detection of Sudan dyes in palm oil. We further explored the impact of a feature selection routine on pattern recognition and machine learning model performances. The following machine learning algorithms: artificial neural network discriminant analysis (ANN), gradient boosted decision tree (GBDT), k-nearest neighbor (KNN), logistic regression discriminant analysis (LRDA), partial least squares discriminant analysis (PLSDA) and support vector machine discrimination analysis (SVM) were used. The model performance was evaluated using the sensitivity, specificity, accuracy, and Matthew correlation coefficient (MCC)[37–40].

2. Experimental

2.1. Data collection

A total of ninety-eight (98) palm oil samples were obtained from the storage units of Ghana Food and Drugs Authority (FDA). These samples had been analyzed and the presence of Sudan dye or otherwise confirmed using high-performance liquid chromatography (HPLC) by the FDA before they were stored. The samples included 44 pure and 54 adulterated ones.

The IR spectra of the samples were obtained using the Perkin Elmer Spectrum 100 FT-IR spectrometer (Perkin Elmer, Serial Number: 67322) fitted with a horizontal attenuated total reflectance (HATR) accessory (Perkin Elmer, Serial Number: 67322). Spectra were collected within a wavenumber range of 650 – 4000 cm^{-1} at a resolution of 4 cm^{-1} with 4 replicate measurements using Perkin Elmer Spectrum ES (version 10.3). Each spectrum was saved individually as a file with a .sp extension.

Prior to sample measurement, the HATR plate was carefully and thoroughly cleaned in situ with acetone and soft lint-free tissue paper before filling in with the sample to be measured. The cleanliness of the plate was verified by checking the background after cleaning the plate.

The spectra for the pure samples were collected before those for the adulterated samples to prevent cross-contamination. A total of 392 spectra representing 176 and 216 for pure and confirmed adulterated samples, respectively were obtained.

2.2. Data processing and feature selection

Data importation, processing and analysis were performed using Matlab® 2022b (Mathworks, Natick, MA, USA) and PLS Toolbox® ver. 9.1 (Eigenvector Research, Manson, WA, USA).

The data was imported into the average of the replicate measurements for each sample was determined. The final data was organized into a matrix of samples in rows and wavenumbers in columns. The dataset matrix consisted of 98 × 3351 (samples/objects × wavenumbers/cm^{-1}).

Minor shifts in the spectra which occur from run to run were corrected using the interval correlation optimized shift algorithm (Icoshift) [41,42]. A Savitsky Golay smoothing algorithm was also applied with a smoothing window of 11 and a second-degree polynomial [43]. The aligned and smoothed data was organized and randomly assigned into two groups containing two-thirds (65 samples) and one-third (33 samples) for training and validation sets, respectively.

Feature selection was performed using the cluster resolution feature selection algorithm (CR-FS) [44–47]. The Fisher ratio ranking metric was used while the start and stop numbers for the CR-FS were automatically detected using the null probability analysis [47]. The feature selection was performed using a permutation number of 10. For each feature selection permutation, a random subset of the training set data was selected.

2.3. Machine learning model and model quality Estimation

Two sets of data were used for the downstream analysis. These were the original data incorporating all the variables and a second set which contains the variables retained during the feature selection process. Both data sets were auto scaled and normalized prior to model generation. The PCA and machine learning models were built using the training set data and evaluated with the external validation set for each of the datasets. Subsequently, the model quality parameters namely, sensitivity/true positive rate (TPR), specificity/true negative rate (TNR), accuracy (ACC), and MCC were computed for all the ML models [33,37].

Sensitivity/TPR is the probability of a positive prediction being a truly positive class. Specificity/TNR on the other hand is the measure of the probability that a negative prediction is truly negative. TPR, TNR, ACC and MCC are defined by the Eqs. (1)–(4) below where, ***TP***- number of positive samples predicted positive, ***TN*** – number of negative samples predicted negative, ***FP*** – number of negative samples predicted positive, and ***FN*** – number of positive samples predicted negative.

$$Sensivity/TPR = \frac{TP}{TP + FN} \tag{1}$$

$$Specificty/TNR = \frac{TN}{TN + FP} \tag{2}$$

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

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

3. Results and discussions

Adulteration of palm oil with Sudan dyes presents a major health risk to consumers and hence the need for a fast, reliable, and cheap method for quality monitoring cannot be overemphasized. Due to the non-destructive and cost-effective nature of spectroscopic methods, their use with machine learning techniques for food adulteration detection has become popular. Earlier studies achieved a lot of success, but few methods compared the performance of commonly used machine learning algorithms and the impact of feature selection on their performance.

Herein, we aimed to evaluate commonly used machine learning algorithms on spectroscopic data to compare their performance when they are applied to the detection of Sudan dye adulteration in palm oil. The workflow describing the data handling and analysis are shown in Fig. SI in the supporting information.

Shifts in spectroscopic data do occur though they are not as marked as in chromatographic data, baseline drifts and distortions in Fourier transform spectra also occur [48]. These distortions become more pronounced over the long-term use of the spectrometers. Even though equipment standardization can reduce the misalignment in spectra data it is still important to align the data before analysis. Hence, spectroscopic data must be aligned especially in a case like this where the entire spectra are employed. Icoshift is a useful algorithm for both spectroscopic and chromatographic data alignment [41,42]. Icoshift was used for the alignment of the data. Fig. 1a and 1b show the raw and aligned data, respectively. Employing second derivative to spectroscopic data have been demonstrated to improve the performance of machine learning algorithms [49,50]. Hence, a second derivative Savitsky Golay smoothing was applied to the aligned data, results of which are shown in Fig. 2. The application of derivatives to spectroscopic data improves the data by reducing noise, enhancing features, and reducing the impact of baseline drifts that may be present [16,51,52].

Cluster resolution feature selection (CRFS), a wrapper type of feature selection combines sequential backward elimination and forward selection using the cluster resolution metric to evaluate the contribution of features to the separation of samples in different classes in principal components space [44–47]. It has been widely used on both spectroscopic, chromatographic, and other kinds of data [47,53]. We implemented CRFS here without user-defined start and stop numbers. A permutation number of ten (10) was used and features surviving at least 70 % of the time were retained. Of the 3351 variables (wavenumbers), 80 representing 2.39 % were retained. Fig. 3a shows the feature survival rate after the 10 permutations. The location of the features that survived at least 70 % are shown in Fig. 3b. The frequency of features selected is shown in Fig. S2 and Table S1 in the SI. Thus, the dataset obtained after feature selection had only 80 wavenumbers.

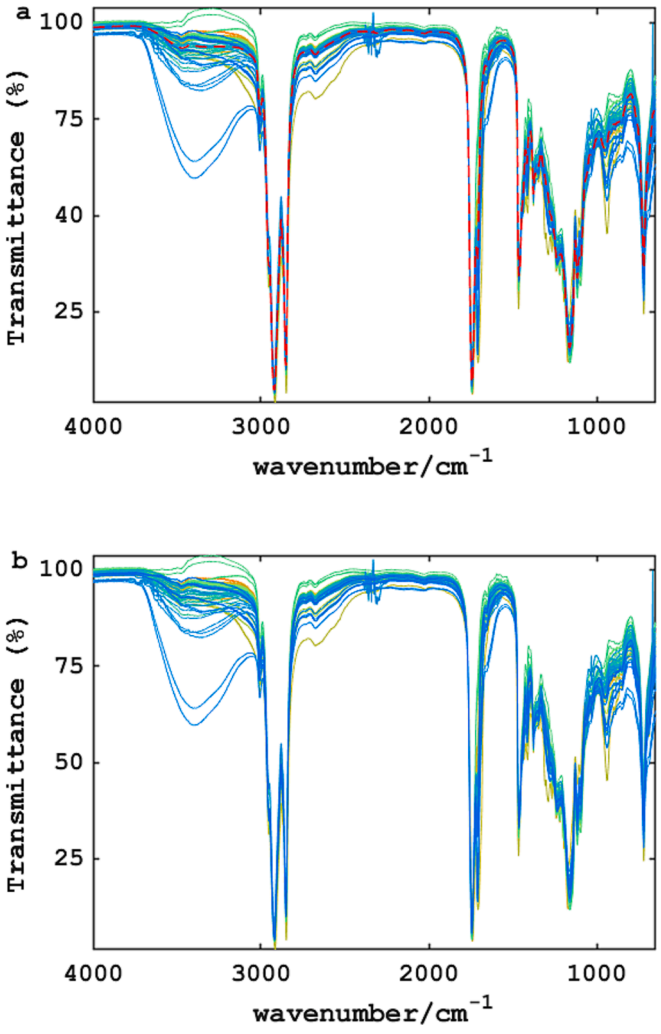

Fig. 1.

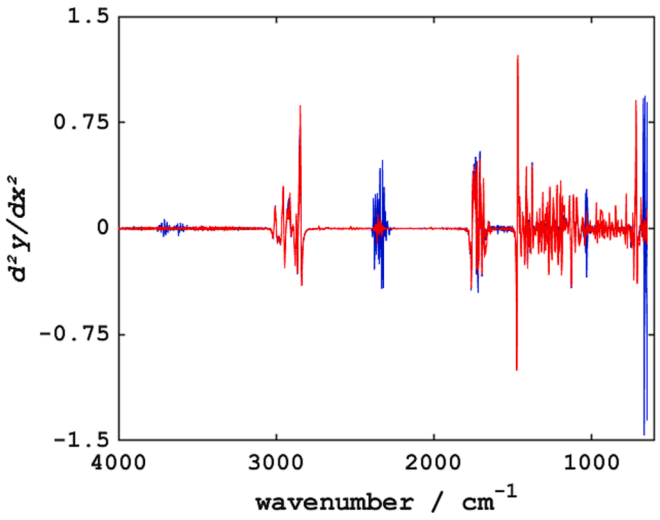

Fig. 2.

PCA is arguably one of the most used multivariate analysis techniques. We compared the separation performance of PCA on the data before and after feature selection. The result of this comparison is shown in Fig. 4a and 4b. From Fig. 4a, it can be observed that there is no clear separation between the adulterated and pure palm oil samples. Fig. 4b shows the PCA score plot after feature selection using only 80 variables shows a better separation. It can also be observed that the sum of explained variance in the first two principal components (PC 1 and PC 2) were 31.73 % and 22.01 %, for before and after feature selection, respectively. As this represents the underlying information in the data, it implies that the application of the feature selection does not lead to a loss of information. Even though PCA is not a classification algorithm, knowledge of the sample class can aid in its use for classification. The lack of separation between the pure and adulterated samples in Fig. 4a makes it nearly impossible to use the PCA model incorporating all the variable for classification. In Fig. 4b where only the selected features were used it can be observed that only four (4) pure samples fell outside the 95 % confidence ellipse of the pure class. It can also be observed one samples of the validation set of the adulterated class was also misclassified. This suggests that CRFS improves the PCA model and more significantly using only 80 of the features.

We further evaluated the performance of artificial neural network (ANN), gradient boosted decision tree (GBDT), *k*-nearest neighbour (KNN), logistic regression discriminant analysis (LRDA), partial least squares discriminant analysis (PLSDA) and support vector machine (SVM) on the identification of adulterated palm oil samples before and after feature selection. The model prediction sensitivity, specificity, accuracy, and MCC were used. Sensitivity and specificity measure the true positive rate and true negative rate, respectively while the overall model performance is evaluated by the accuracy. MCC is more informative than the accuracy for evaluating binary classification problems, as it considers the balance between all components of the confusion matrix (true positives, true negatives, false positives, and false negatives) [38,39].

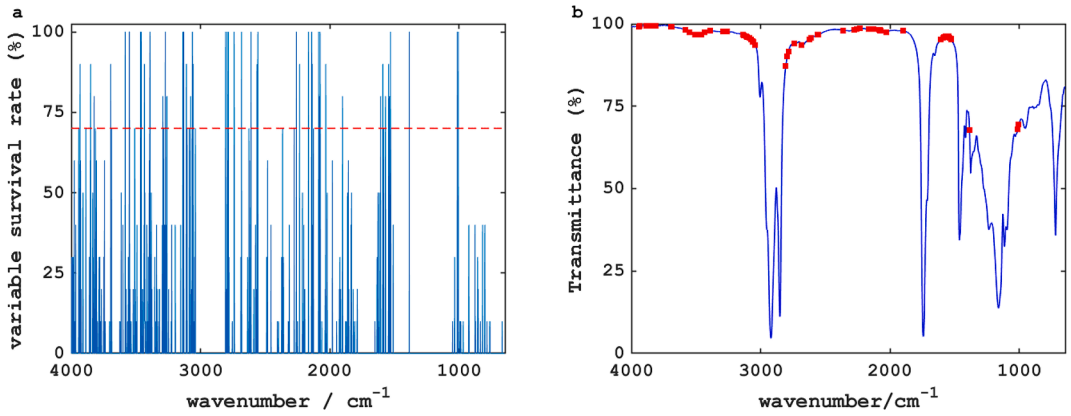

Fig. 3.

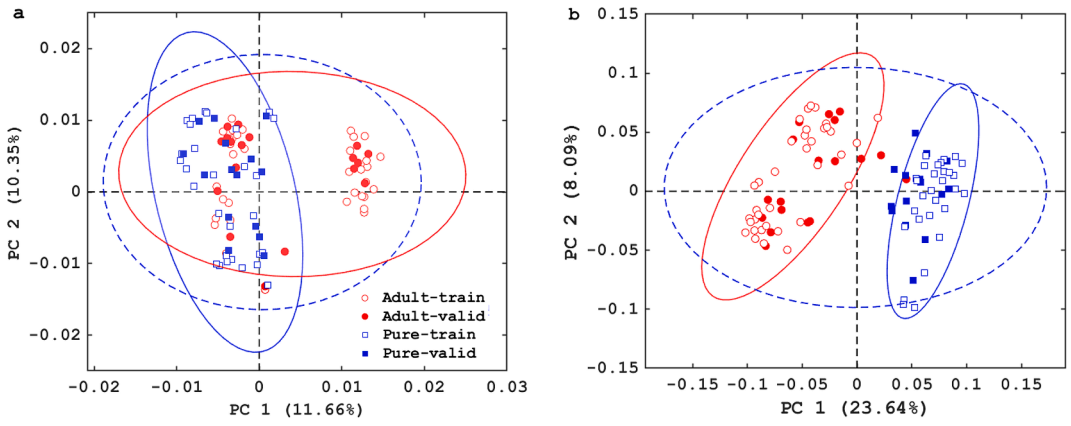

Fig. 4.

The model prediction plots for ML ANN, GBDT, KNN, LRDA, PLSDA and SVM for models generated using all the wavelengths (3551 variables) and CRFS selected wavelengths (80 variables) are shown in Figs. S3 – S8 in the supporting information. In general, the application of CRFS led to an improvement in the model quality parameter for the training set for KNN, PLSDA and SVM as shown in Table 1. From Table 1, TNR, ACC and MCC increased for training set increased after feature selection. In the case of PLSDA and SVM, excellent values were obtained for all quality parameters after feature selection. While applying the models to the external validation sets, all the algorithms except ANN and GBDT showed improved model results. It is known that high number of variables in machine learning applications risks over fitting and can also confuse learning algorithms [54–56]. The application of feature selection routines such as CRFS reduces redundancy, noise and computational requirements. An efficient variables selection contributes significantly to a parsimonious machine learning models which are desirable.

4. Conclusion

Our findings show that feature selection, specifically using the CRFS method, significantly enhances the performance of principal component

Table 1

	Before Feature Selection (3351 variables)				After Feature Selection (80)			
Training Model Quality								
Algorithm	TNR	TPR	ACC	MCC	TNR	TPR	ACC	MCC
ANN	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
GBDT	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
KNN	0.8621	0.9444	0.9077	0.8136	1.0000	0.9444	0.9692	0.9400
LRDA	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
PLSDA	1.0000	0.9722	0.9846	0.9694	1.0000	1.0000	1.0000	1.0000
SVM	0.9655	0.9444	0.9538	0.9389	1.0000	1.0000	1.0000	1.0000
Validation Model Quality								
ANN	1.0000	0.8889	0.9394	0.8856	0.8667	1.0000	0.9394	0.8832
GBDT	0.9333	1.0000	0.9697	0.9403	0.9333	0.9444	0.9394	0.8778
KNN	0.8667	0.8333	0.8485	0.6974	1.0000	0.9444	0.9697	0.9410
LRDA	1.0000	0.8889	0.9394	0.8856	1.0000	0.9444	0.9697	0.9410
PLSDA	0.9333	0.8889	0.9091	0.8192	1.0000	0.9444	0.9697	0.9410
SVM	0.9333	0.8889	0.9091	0.8192	1.0000	0.9444	0.9697	0.9410

analysis (PCA). KNN, LRDA, PLSDA and SVM showed improved model generalization as better results were obtained during model validation wit external validation sets. These findings underscore the importance of appropriate feature selection and the choice of machine learning algorithms to achieve reliable and accurate detection of Sudan dyes in palm oil.

CRediT authorship contribution statement

Joanna K. Kwao: . **Cheetham Mingle:** Writing – review & editing, Resources, Methodology, Data curation. **John N. Addotey:** . **Kwabena F.M. Opuni:** . **Lawrence A. Adutwum:** .

Funding

This research did not receive any specific grant from funding agencies.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.microc.2024.112433.

Data availability

Data will be made available on request.

References

[1] J. Spink, Safety of food and beverages: risks of food adulteration, Encyclopedia of Food Safety 3 (2014) 413–416, https://doi.org/10.1016/B978-0-12-378612-8.00300-0.
[2] Banerjee D, Chowdhary S, Chakraborty S, Bhattacharyya R. Recent advances in detection of food adulteration. Food Safety in the 21st Century: Public Health Perspective 2017:145–60. https://doi.org/10.1016/B978-0-12-801773-9.00011-X.
[3] E. Hong, S.Y. Lee, J.Y. Jeong, J.M. Park, B.H. Kim, K. Kwon, et al., Modern analytical methods for the detection of food fraud and adulteration by food category, J Sci Food Agric 97 (2017), https://doi.org/10.1002/jsfa.8364.
[4] T.M. Fonovich, Sudan dyes: Are they dangerous for human health, Drug Chem Toxicol 36 (2013) 343–352, https://doi.org/10.3109/01480545.2012.710626.
[5] C.G. Reile, M.S. Rodríguez, D.D. de Fernandes, S, Gomes A de A, Diniz PHGD, Di Anibal CV., Qualitative and quantitative analysis based on digital images to determine the adulteration of ketchup samples with Sudan I dye, Food Chem 328 (2020), https://doi.org/10.1016/j.foodchem.2020.127101.
[6] Hussain Khan M, Saleem Z, Ahmad M, Sohaib A, Ayaz H, Mazzara M. Hyperspectral imaging for color adulteration detection in red chili. Applied Sciences (Switzerland) 2020;10. https://doi.org/10.3390/app10175955.
[7] S.Y.S.S. Adade, H. Lin, S.A. Haruna, A.O. Barimah, H. Jiang, A.A. Agyekum, et al., SERS-based sensor coupled with multivariate models for rapid detection of palm oil adulteration with Sudan II and IV dyes, J. Food Compos. Anal. 114 (2022), https://doi.org/10.1016/j.jfca.2022.104834.
[8] Opinion of the Scientific Panel on Food Additives, Flavourings, Processing Aids and Materials in Contact with Food on a request from the commission related to Lutein for use in particular nutritional uses. EFSA Journal 2007;5. https://doi.org/10.2903/j.efsa.2007.315.
[9] C.H. Xing, Y. Wang, J.C. Liu, Z.N. Pan, H.L. Zhang, S.C. Sun, et al., Melatonin reverses mitochondria dysfunction and oxidative stress-induced apoptosis of Sudan I-exposed mouse oocytes, Ecotoxicol Environ Saf 225 (2021), https://doi.org/10.1016/j.ecoenv.2021.112783.
[10] T.B. Zanoni, E. Silva de Paula, G.J. Zocolo, M.V.B. Zanoni, D. Palma de Oliveira, Identification of Sudan III-(deoxy)-guanosine adducts formed in situ in a reaction with no catalyst, Toxicol Environ Chem 95 (2013), https://doi.org/10.1080/02772248.2014.892748.
[11] R.J. Bienstock, L. Perera, M.A. Pasquinelli, Molecular Modeling Study of the Genotoxicity of the Sudan I and Sudan II Azo Dyes and Their Metabolites, Front Chem 10 (2022), https://doi.org/10.3389/fchem.2022.880782.
[12] T.C. Pham, X.T. Dang, B.N. Nguyen, T.T. Vu, Determination of Sudan i and II in Food by High-Performance Liquid Chromatography after Simultaneous Adsorption on Nanosilica, J Anal Methods Chem 2021 (2021), https://doi.org/10.1155/2021/6664463.
[13] J.K. Adjei, V. Ahormegah, A.K. Boateng, H.K. Megbenu, O.S. Fast, easy, cheap, robust and safe method of analysis of Sudan dyes in chilli pepper powder, Heliyon 6 (2020), https://doi.org/10.1016/j.heliyon.2020.e05243.
[14] E. Ertaş, H. Özer, C. Alasalvar, A rapid HPLC method for determination of Sudan dyes and Para Red in red chilli pepper, Food Chem 105 (2007) 756–760, https://doi.org/10.1016/j.foodchem.2007.01.010.
[15] M.M. Dar, Detection of Sudan Dyes in Red Chilli Powder by Thin Layer Chromatography, J Allergy Ther (2012;S1.), https://doi.org/10.4172/scientificreports.586.
[16] Blanco M, Villarroya I. NIR spectroscopy: A rapid-response analytical tool. TrAC - Trends in Analytical Chemistry 2002;21. https://doi.org/10.1016/S0165-9936(02)00404-1.
[17] Y. Xu, P. Zhong, A. Jiang, X. Shen, X. Li, Z. Xu, et al., Raman spectroscopy coupled with chemometrics for food authentication: A review. TrAC - T, Trends Anal. Chem. 131 (2020), https://doi.org/10.1016/j.trac.2020.116017.
[18] L. Mandrile, L. Barbosa-Pereira, K.M. Sorensen, A.M. Giovannozzi, G. Zeppa, S. B. Engelsen, et al., Authentication of cocoa bean shells by near- and mid-infrared spectroscopy and inductively coupled plasma-optical emission spectroscopy, Food Chem 292 (2019) 47–57, https://doi.org/10.1016/j.foodchem.2019.04.008.
[19] A.I. Ropodi, E.Z. Panagou, G.J.E. Nychas, Data mining derived from food analyses using non-invasive/non-destructive analytical techniques; determination of food authenticity, quality & safety in tandem with computer science disciplines, Trends Food Sci Technol 50 (2016) 11–25, https://doi.org/10.1016/J.TIFS.2016.01.011.
[20] H.J. He, D.W. Sun, Microbial evaluation of raw and processed food products by Visible/Infrared, Raman and Fluorescence spectroscopy, Trends Food Sci Technol 46 (2015) 199–210, https://doi.org/10.1016/J.TIFS.2015.10.004.
[21] Monago-Maraña O, Durán-Merás I, Muñoz de la Peña A, Galeano-Díaz T. Analytical techniques and chemometrics approaches in authenticating and identifying adulteration of paprika powder using fingerprints: A review. Microchemical Journal 2022;178. https://doi.org/10.1016/j.microc.2022.107382.
[22] S. Yao-Say Solomon Adade, H. Lin, H. Jiang, S.A. Haruna, A. Osei Barimah, M. Zareef, et al., Fraud detection in crude palm oil using SERS combined with chemometrics, Food Chem 388 (2022), https://doi.org/10.1016/j.foodchem.2022.132973.
[23] C.V. Di Anibal, L.F. Marsal, M.P. Callao, I. Ruisánchez, Surface Enhanced Raman Spectroscopy (SERS) and multivariate analysis as a screening tool for detecting Sudan I dye in culinary spices, Spectrochim Acta A Mol Biomol Spectrosc 87 (2012) 135–141, https://doi.org/10.1016/J.SAA.2011.11.027.
[24] M. Shiddiq, Z. Zulkarnain, V. Asyana, H. Aliyah, Identification of Pure and Adulterated Honey Using Two Spectroscopic Methods, J Phys Conf Ser 1351 (2019), https://doi.org/10.1088/1742-6596/1351/1/012022.
[25] D. Valinger, L. Longin, F. Grbeš, M. Benković, T. Jurina, J. Gajdoš Kljusurić, et al., Detection of honey adulteration – The potential of UV-VIS and NIR spectroscopy coupled with multivariate analysis, Lwt 145 (2021), https://doi.org/10.1016/j.lwt.2021.111316.
[26] D.A. Magdas, F. Guyon, C. Berghian-Grosan, M.C. Muller, Challenges and a step forward in honey classification based on Raman spectroscopy, Food Control 123 (2021) 107769, https://doi.org/10.1016/j.foodcont.2020.107769.
[27] A.A. Boateng, S. Sumaila, M. Lartey, M.B. Oppong, K.F.M. Opuni, L.A. Adutwum, Evaluation of chemometric classification and regression models for the detection of syrup adulteration in honey, LWT 163 (2022), https://doi.org/10.1016/j.lwt.2022.113498.
[28] Fengou LC, Spyrelli E, Lianou A, Tsakanikas P, Panagou EZ, Nychas GJE. Estimation of Minced Pork Microbiological Spoilage through Fourier Transform Infrared and Visible Spectroscopy and Multispectral Vision Technology. Foods 2019, Vol 8, Page 238 2019;8:238. https://doi.org/10.3390/FOODS8070238.
[29] A.I. Ropodi, E.Z. Panagou, G.J.E. Nychas, Multispectral imaging (MSI): A promising method for the detection of minced beef adulteration with horsemeat, Food Control 73 (2017) 57–63, https://doi.org/10.1016/J.FOODCONT.2016.05.048.
[30] S.S. Andoh, K. Nyave, B. Asamoah, B. Kanyathare, T. Nuutinen, C. Mingle, et al., Optical screening for presence of banned Sudan III and Sudan IV dyes in edible palm oils, Food Addit Contam Part A Chem Anal Control Expo Risk Assess 37 (2020) 1049–1060, https://doi.org/10.1080/19440049.2020.1726500.
[31] S.S. Andoh, T. Nuutinen, C. Mingle, M. Roussey, Qualitative analysis of Sudan IV in edible palm oil, Journal of the European Optical Society 15 (2019), https://doi.org/10.1186/s41476-019-0117-0.
[32] E. Teye, C. Elliott, L.K. Sam-Amoah, C. Mingle, Rapid and nondestructive fraud detection of palm oil adulteration with Sudan dyes using portable NIR spectroscopic techniques, Food Addit Contam Part A Chem Anal Control Expo Risk Assess 36 (2019) 1589–1596, https://doi.org/10.1080/19440049.2019.1658905.
[33] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using matthews correlation coefficient metric, PLoS One 12 (2017) e0177678, https://doi.org/10.1371/journal.pone.0177678.
[34] I. Guyon, An Introduction to Variable and Feature Selection 1 Introduction, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[35] I. Guyon, A. Elisseeff, Feature Extraction, Foundations and Applications: An introduction to feature extraction, Stud. Fuzziness Soft Comput. 207 (2006) 1–25, https://doi.org/10.1007/978-3-540-35488-8_1.
[36] B.E. Boser, I.M. Guyon, V.N. Vapnik, Training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, 1992, pp. 144–152.
[37] T. Loong, Clinical review Understanding sensitivity and specificity with the right, BMJ 327 (2003) 716–719, https://doi.org/10.1136/bmj.327.7417.716.

[38] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom 21 (2020) 1–13, https://doi.org/10.1186/s12864-019-6413-7.
[39] D. Chicco, N. Tötsch, G. Jurman, The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation, BioData Min 14 (2021) 1–22, https://doi.org/10.1186/S13040-021-00244-Z/TABLES/5.
[40] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta (BBA) Protein Struct. 405 (1975) 442–451, https://doi.org/10.1016/0005-2795(75)90109-9.
[41] F. Savorani, G. Tomasi, S.B. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, J. Magn. Reson. 202 (2010) 190–202, https://doi.org/10.1016/j.jmr.2009.11.012.
[42] G. Tomasi, F. Savorani, E.SB. Icoshift, An effective tool for the alignment of chromatographic data, J Chromatogr A 1218 (2011), https://doi.org/10.1016/j.chroma.2011.08.086.
[43] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal Chem 36 (1964) 1627–1639.
[44] N.A. Sinkov, J.J. Harynuk, Cluster resolution: a metric for automated, objective and optimized feature selection in chemometric modeling, Talanta 83 (2011) 1079–1087, https://doi.org/10.1016/j.talanta.2010.10.025.
[45] N.A. Sinkov, J.J. Harynuk, Three-dimensional cluster resolution for guiding automatic chemometric model optimization, Talanta 103 (2013) 252–259.
[46] M.S. Armstrong, A.P. de la Mata, J.J. Harynuk, An efficient and accurate numerical determination of the cluster resolution metric in two dimensions, J Chemom 35 (2021), https://doi.org/10.1002/cem.3346.
[47] L.A. Adutwum, A.P. de la Mata, H.D. Bean, J.E. Hill, J.J. Harynuk, Estimation of start and stop numbers for cluster resolution feature selection algorithm: an empirical approach using null distribution analysis of Fisher ratios, Anal Bioanal Chem 409 (2017) 6699–6708, https://doi.org/10.1007/s00216-017-0628-8.
[48] F. Zhang, X. Tang, L. Li, Origins of Baseline Drift and Distortion in Fourier Transform Spectra, Molecules 27 (2022), https://doi.org/10.3390/molecules27134287.
[49] T. Shi, H. Liu, Y. Chen, T. Fei, J. Wang, G. Wu, Spectroscopic diagnosis of arsenic contamination in agricultural soils, Sensors (switzerland) 17 (2017), https://doi.org/10.3390/s17051036.
[50] X. Zhang, Z. Gao, Y. Yang, S. Pan, J. Yin, X. Yu, Rapid identification of the storage age of dried tangerine peel using a hand-held near infrared spectrometer and machine learning, J near Infrared Spectrosc 30 (2022), https://doi.org/10.1177/09670335211057232.
[51] R. Gill, T.S. Bal, A.C. Moffat, The Application of Derivative UV-Visible Spectroscopy in Forensic Toxicology, J. Forensic Sci. Soc. 22 (1982), https://doi.org/10.1016/S0015-7368(82)71466-5.
[52] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra. TrAC - T, Trends Anal. Chem. 28 (2009), https://doi.org/10.1016/j.trac.2009.07.007.
[53] A.O.O. Oliynyk, L.A.A. Adutwum, J.J.J. Harynuk, A. Mar, Classifying crystal structures of binary compounds AB through cluster resolution feature selection and support vector machine analysis, Chem. Mater. 28 (2016) 6672–6681, https://doi.org/10.1021/acs.chemmater.6b02905.
[54] S.M. Vieira, J.M.C. Sousa, U. Kaymak, Fuzzy criteria for feature selection, Fuzzy Sets Syst 189 (2012) 1–18, https://doi.org/10.1016/j.fss.2011.09.009.
[55] J.M. Cadenas, M.C. Garrido, R. Martínez, Feature subset selection Filter - Wrapper based on low quality data, Expert Syst Appl 40 (2013) 6241–6252, https://doi.org/10.1016/j.eswa.2013.05.051.
[56] B. Duval, J.-K. Hao, Advances in metaheuristics for gene selection and classification of microarray data, Brief Bioinform 11 (2010) 127–141, https://doi.org/10.1093/bib/bbp035.