Food Control 145 (2023) 109502

Contents lists available at ScienceDirect

Food Control

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

Fusing one-class and two-class classification – A case study on the detection of pepper fraud

Martin Alewijn *, Vasiliki Akridopoulou, Tjerk Venderink, Judith Müller-Maatsch, Erika Silletti

Wageningen Food Safety Research (WFSR), Akkermaalsbos 2, 6708, WB, Wageningen, the Netherlands

ARTICLE INFO

Keywords:
High-level data fusion
Visible-near infrared spectrophotometry
Direct analysis in real time-mass spectrometry
Fraud
Classification

ABSTRACT

Black pepper is a commercially important commodity, which is susceptible for fraudulent additions. Analytical tools are capable of detection of specific additions, but in most published cases these tools and associated mathematical models are suitable for only one or a few predetermined adulterants. There is a need for methodology that can detect any addition without having to know the type of adulterant *a priori*. We analysed a dataset of 200 authentic black pepper samples and a total of 210 adulterated samples consisting of mixtures of black pepper and oil-dress pepper, spent pepper, coffee husk, coffee skin and papaya seeds, respectively. A small, non-destructive spectral tool, a visible-near infrared spectrophotometer, (VIS/NIR) and a slower and more expensive mass spectrometric tool, direct analysis in real time-mass spectrometer (DART-MS), were evaluated according to their performances in terms of adulteration detection for a number of machine learning modes. The often-used approach where an 'optimal' model is selected and employed yielded for VIS/NIR very reasonable results for most of the adulterants used, but no single model performed well for all adulterants in the dataset. However, high-level fusion modelling of both one- and two-class models developed for different adulterant types using a penalized excess scoring system led to a performance of typically >75% correct classification, regardless of the nature of the adulterant. DART-MS outperformed VIS/NIR but also led to no single model that was able to detect all adulterants present. From the small number of tested fusion strategies, again the penalized excess score outperformed the other fusion options and yielded perfect classification scores for all but one of the adulterants tested. This shows that this type of modelling for cases where the nature of a target is unknown is a promising approach. It is even speculated that this modelling approach is likely to be suitable for types of adulterant that were not used in the model development phase.

1. Introduction

The global and European demand for spices increased in recent years and is forecasted to grow even further in the next years in all sectors, i.e. industrial, food service and retail (FAO, 2022; Silvis, Van Ruth, Van Der Fels-Klerx, & Luning, 2017). In particular, pepper production has been growing from 511,000 tonnes in 2015 to 731,034 in 2019 worldwide. While 2020 the production was slightly less than the year before (714,296 tonnes), the trend on this spice production is expected to keep growing (FAO, 2022). With a growing demand, natural fluctuations in the products quality and lack of control measures, the spice food chain is vulnerable to food fraud (Silvis, Van Ruth, Van Der Fels-Klerx, & Luning, 2017; Ulberth, 2020). Reported cases of fraud in black pepper include addition/admixtures with chilli, papaya seeds, millet, Juniper berry, starch, ash, spent pepper, pepper husk pepper pinheads, buckwheat and mineral oil (Danezis, Tsagkaris, Brusic, & Georgiou, 2016; Dhanya, Syamkumar, & Sasikumar, 2009; Galvin-King, Haughey, & Elliott, 2018; Gul, Nasrullah, Nissar, Saifi, & Abdin, 2018; McGoverin, September, Geladi, & Manley, 2012; Negi, Pare, & Meenatchi, 2021; Orrillo et al., 2019; Osman et al., 2019; Paradkar, Singhal, & Kulkarni, 2001; Parvathy et al., 2014; Wilde Amelie, Simon, Haughey, & Elliott, 2019). In addition, berries from different *piper* species (e.g. P. Attenuatum and P. Galeatum, P. Longum), dried fruit of West Indian Lantana (i.e. *Lantana Camara*), stem and chaff of black pepper are often mixed with whole peppercorn (Sasikumar, Swetha, Parvathy, & Sheeja, 2016). Analytical tools to detect these additions are for instance DNA barcoding, polymerase chain reaction (PCR), gas chromatography–mass spectrometry (GC-MS), direct analysis in real time mass spectrometry (DART-MS), near infrared spectroscopy (NIR), and hyperspectral imaging (Danezis et al., 2016; Dhanya et al., 2009; Galvin-King et al., 2018; Gul et al.,

* Corresponding author.
E-mail address: martin.alewijn@wur.nl (M. Alewijn).

https://doi.org/10.1016/j.foodcont.2022.109502
Received 3 August 2022; Received in revised form 18 October 2022; Accepted 7 November 2022
Available online 10 November 2022

2018; McGoverin et al., 2012; Negi et al., 2021; Orrillo et al., 2019; Osman et al., 2019; Paradkar et al., 2001; Parvathy et al., 2014; Ulberth, 2020; Wilde Amelie et al., 2019). Currently, there does not seem to be any scientific way in determining the 'best' technique for a given matrix/fraud type combination, let alone for the case where multiple fraud types need to be detected. In this paper we explore two techniques: DART-MS and NIR. Both techniques are capable of generating broad chemical fingerprints of samples but are fundamentally very different and measure the samples from different angles. In spice authentication, DART-MS has been used to detect anisatin in star anise (Shen et al., 2012), to discriminate piper species (Chandraa, Bajpaia, Srivastvab, Kumarc, & Kumara, 2014), amongst others (Guo et al., 2017). Also, spectroscopic techniques have been used for spice authentication, applying straightforward single application models (Kaavya et al., 2020; Kucharska-Ambrożej & Karpinska, 2020; McGoverin et al., 2012) to multiple models for pepper (Hu, Yin, Ma, & Liu, 2018; Nobari-Moghaddam, Tamiji, Akbari-Lakeh, Khoshayand, & Haji-Mahmoodi, 2021) (Wilde Amelie et al., 2019).

As reviewed recently by Nobari-Moghaddam et al. (2021) and Reinholds, Bartkevics, Silvis, van Ruth, and Esslinger (2015), the detection of fraud in spices typically combine an analytical technique with chemometric techniques, which is a potentially powerful setup to detect deviation in normal natural variation in products. The published work on this matter usually shows good results, but this is generally only true for the combination of one matrix, one adulterant, one analytical technique and one chemometric model. For example, models are able to detect multiple adulterants in black pepper (Wilde Amelie et al., 2019). Nevertheless, in real life, one typically does not know up front which adulterants to test for, as all currently known and even perhaps new adulterants might be present. For routine situations, one would like to be able to detect any type of adulteration in a single workflow, with of course the best possible performance. Possible solutions to this problem are one-class classification (OCC), orthogonal techniques and/or data fusion. In one-class classification only the natural variation of the product in question is modelled, and any deviation could flag a sample as 'not-normal' (Müller-Maatsch, Alewijn, Wijtten, & Weesepoel, 2021), prompting for additional clarification of the reason. Although this kind of modelling is designed for detecting all abnormalities (which are visible given the analytical signal used), they generally have poorer performance in terms of detection rates than specifically designed binary models for specific adulterants (Bellinger, Sharma, Zaıane, & Japkowicz, 2017; Deng, Li, Liu, Guo, & Newsam, 2018). Orthogonal techniques apply multiple techniques to the same samples, such as Fourier-transform infrared spectroscopy (FTIR) and Liquid Chromatography Mass Spectrometry (LC-MS) for oregano (Black, Haughey, Chevallier, Galvin-King, & Elliott, 2016) and GC-MS, LC-MS, and DART-MS for origin classification of black pepper (Liang et al., 2021). Both papers on herbs and spices do not combine the results of both techniques, although orthogonal techniques have the potential to detect more types of adulterants, or could be used for verification of either of the results (Galvin-King et al., 2018). Combination of analytical techniques, or results from different chemometric models could be done through data fusion such as high, mid, and low-level data fusion (Oliveira, Cruz-Tirado, & Barbin, 2019). There are several examples of data fusion in spice authentication in literature, such as low-level data fusion (mid infrared spectroscopy (MIR) and NIR) in authentication of star anise (Shen et al., 2012) (Wang, Mei, Ni, & Kokot, 2014) and mid-level data fusion using DART-MS on oregano admixtures (Massaro et al., 2021). In most cases, as illustrated by (Oliveira, Cruz-Tirado, Roque, Teófilo, & Barbin, 2020), multiple models are produced but their results are not combined in any way.

In this study we investigated the possibility of using a combinatory approach by combining two different (orthogonal) analytical techniques, VIS/NIR spectroscopy and DART-MS, and combinations of binary and one-class chemometric models to detect a broad range of diverse adulterants in an efficient experimental procedure. We used papaya seeds and coffee husk as two extraneous replacement materials, and spent black pepper and black pepper dressed with mineral oil as two products that originate from black pepper with a lower quality.

2. Materials and methods

2.1. Pepper and adulterants sample set

A total of 200 black pepper samples (*Piper Nigrum* L.) were provided by four different Dutch and international companies operating in the spice business at different points in the spice supply chain. Samples were obtained directly from the farm (n = 100 sampled at every bag), sampled according to ISO 948:1980 (n = 90) at the trading company in the Netherlands, or otherwise sampled (n = 10, no information available). Samples were collected between September 2017 and October 2018 from four different countries of origin, i.e. Vietnam (n = 108), Brazil (n = 61), Indonesia (n = 27) and India (n = 4). All samples were provided as whole black pepper peppercorns with the exception of the samples from Indonesia which were provided as ground black pepper. To account for some black pepper-processing variability, 191 samples were sundried, 7 samples were additionally machine dried and 33 samples were further sterilized after being sundried. In addition, samples possessed different grades, with 144, 22 and 25 samples having grade I, II and II, respectively (Codex Alimentarius CXS 326–2017). Besides the country of origin, no information was available for 10 samples. Four types of adulterants were selected as relevant adulterants for black pepper based on personal communications with the industrial partners and procurement possibilities on the market. The adulterants used to prepare the adulterated samples included mineral oil-dressed black pepper (OD, n = 2), spent black pepper (SP, n = 4), coffee husk (CH, n = 4), coffee skin (CS, n = 1) and papaya seeds (PS, n = 3). One of the papaya seed samples was obtained from fresh papaya, purchased at a local store in the Netherlands. The other two papaya seed samples were produced in Vietnam. All other adulterants were provided by the previously mentioned spice companies.

2.2. Sample preparation

Whole black pepper peppercorn and adulterants were kept in the dark at 4 °C in aluminium seal bags for the duration of the study. The samples were allowed to equilibrate to room temperature just before further handling. Fifty grams of whole peppercorns or adulterants were weighted and grinded at 10.000 rpm for 90 s using a knife mill (Grindomix GM 200, Retsch, Haan, Germany). For the preparation of the adulterant papaya seeds from the fresh fruit, papaya seeds were stripped from the fruit, oven-dried (Universal Oven UF260, Memmert, Schwabach, Germany) at 70 °C for more than 24 h and placed in a desiccator for 30 min until a constant weight was achieved.

Ground black pepper was spiked with the respective adulterants at different concentrations. A total of 196 authentic black peppers mixed with adulterants, at concentrations in the range of 2.5%–40% (w/w), were prepared, as detailed in Table 1. Lower levels than 2.5% were considered of negligable relevance for economic adulteration, whereas higher levels were *a priori* considered as possibly detectable with the naked eye and therefore not relevant for an analytical approach.

2.3. VIS-NIR prototype measurements

The VIS/NIR measurements were performed using a modular prototype from OceanOptics (Duiven, the Netherlands). A sample-holder module was connected to a halogen lamp (HL-2000-FHSA, OceanOptics, The Netherlands) and two spectrophotometers: a FLAME-S-XR1-ES spectrophotometer (FLMS05361, OceanOptics, The Netherlands) for measurements in the visible wavelength range (VIS) and a FLAME-NIR spectrophotometer (FLMN01815, OceanOptics, The Netherlands) for the NIR region. VIS measurements were performed with 500 lines/mm

Table 1
Overview of datasets.

Technique	Unique samples[a]	Total scans[a]	Variables per scan	Population outliers
VIS/NIR	Tot: 410 BP: 200 Ad: 196 + 14 (OD: 36 + 2, SP: 56 + 4, CH: 52 + 4, CS: 12 + 1, PS: 40 + 3)	Tot: 2472 BP: 1352 Ad: 1120 (OD: 201, SP: 325, CH: 299, CS: 66, PS: 229)	400–1634 nm,~ 0.4 nm res ≤926 nm, 5.7 nm res >926 nm. 1418 variables	BP: 67 scans, 0 samples)
DART-MS	Tot: 187 BP: 42 Ad: 131 + 14 (OD: 24 + 2, SP: 36 + 4, CH: 36 + 4, CS: 12 + 1, PS: 23 + 3)	Tot: 892 BP: 256 Ad: 636 (OD: 112, SP: 180, CH: 176, CS: 56, PS: 112)	100–2000 m/z, 1 m/z resolution, 1901 variables	BP: 13 scans, 0 samples)

[a] Tot: Total number (authentic black pepper samples, adulterants and adulterated samples), BP: pure black pepper samples, Ad: Adulterated black pepper samples + the number of pure adulterants. Adulterants: OD: mineral oil-dressed black pepper, SP: spent black pepper, CH: coffee husk, CS: coffee skin, PS: papaya seeds (PS).

groove density grating (XR1), blazed at 250 nm and slit 100 μm. Integration time was set to 10 ms with 154 scans to average and a signal to noise reduction performed by boxcar smoothing value of 3. Spectra were recorded in the range between 400 and 926 nm. NIR measurements were performed with 150 lines/mm groove density grating (N33), 1.1 μm blaze and slit 200 μm, in the range between 926 and 1634 nm. Further, the integration time was set to 110 ms, with 14 scans to average and a boxcar smoothing value of 0. The spectra were obtained and processed using OceanView Software provided by OceanOptics. Before each measurement, the spectrophotometers were calibrated. Saturation level was set at 80% of the maximum saturation. The light background was adjusted with a grey reference (40%), provided by the manufacturer. The black lid cover of the sample-holder module was used for the calibration of dark background. VIS and NIR spectra were recorded of all samples in a randomized order. About 15 g of material were added to a 5 cm diameter glass Petri dish to obtain full coverage. Replicate measurements were achieved by recording the spectra of each sample on three different days.

2.4. DARTS-MS measurements

Due to practical constraints only a subset of the available samples (Table 1) could be analysed using DART-MS. Per class (pure black peppers, adulterated samples at each level), a desired number of samples was randomly selected. 0.5 g of material, i.e., pure black pepper, pure adulterants and their mixture, were vigorously mixed with 5 ml ethyl acetate (LC-MS grade, Actu-All, Oss, the Netherlands) for 10 s in a vortex (Velp Scientifica, Usmate, Italy) and 5 min on a overhead shaker (Heidolph Instruments, Schwabach, Germany). The samples were centrifuged (Eppendorf centrifuge 5810, Hamburg, Germany) for 5 min at 1500 g and the supernatant transferred to an Eppendorf cup and stored at −20 °C until further analysis.

A DART ionisation source (Ionsense, MA, USA) was coupled with a high-resolution mass spectroscopy (Exactive, Thermo Scientific™), equipped with an x-y autosampler (Ionsense, MA, USA) that mounts a 96-place stainless steel mesh, used for sample application. The DART ion source operating conditions were: Helium flow is 3.5 L min^{-1}, capillary temperature is 275 °C, capillary voltage 60 V, tube lens voltage 100 V and skimmer voltage 20 V. The mass spectrometer operated in negative ion mode at 200 °C with a sample speed of 5 mm/s. The average of 10 scans/sample was taken as one measurement. Each sample was measured 4 times, individual replicates randomised over the mesh, and masses were acquired from 100 *m/z* to 2000 *m/z*. To eliminate variables (masses) with too much noise, variables with responses <1000 counts in all scans in the dataset were removed from modelling phase.

2.5. Data processing

All chemometrics have been done in R studio (Team, 2019). The data obtained from the two techniques above were initially processed separately, using all available scans (for dataset dimensions, see Table 1). After standard normal variate (SNV) (for VIS/NIR) and row-wise normalization (for DART-MS), principal component analysis (PCA) was calculated on all scans of authentic samples, and score vs orthogonal distance plots were used to manually and conservatively identify population outliers, which were excluded from all training sets (Table 1). Per technique, one-class, and two-class models for pure black pepper versus each of the available adulterants, in combination with various pre-processing treatments were screened in cross-validation mode - leave-10%-out: training sets consisted of 90%, test sets of 10% of the number of unique samples indicated in Table 1, for two-class models the sampling was stratified meaning that both authentic and non-authentic groups were split 90/10%. The models' performances were ranked on area under the receiver operating characteristic values (AUROC). Different combinations of preprocessing and models were tried as described previously (Müller-Maatsch et al., 2021). In calculation of the AUROCs, model-specific outliers were identified (and omitted) as authentic scans exceeding median class distance plus 3x median absolute deviation (MAD) for all authentic scans (Leys, Delacre, Mora, Lakens, & Ley, 2019). Models with >10% of scans being classified as outliers were considered overfitted and thus ignored. Based on those AUROC results, for each adulterant the two best-performing one-class and two-class models were selected. In a number of cases, the same model was selected for more than one adulterant, resulting in 12 and 9 selected models for VIS/NIR and DART, respectively. Sample scores were calculated as averages of the individual scans' classification scores, without removing model-specific outliers. Per-model classification was performed based on the classification scores, with upper- and lower classification limits determined by the range of the (cross validation (CV) left-out) scores for the pure pepper samples ignoring black pepper samples outside the median ± 3 x MAD range – the latter samples are considered misclassifications.

Finally, the models were combined, or fused, in a step that can be seen as the aggregation phase of ensemble learning (Sagi & Rokach, 2018). To the best of our knowledge, there is no established way to do so, and any conceivable procedure may be used. We investigated three possibilities: averaging, voting, and a penalized excess score (PES) system. For averaging, the sample class distance scores for each of the models were averaged across samples. Final classification boundaries were calculated as for the single models. In the voting system, the sample class distance scores per model were converted to a binary result using the authentic samples only. This was done in a leave-10%-out cross validation mode, where a median ±3 x MAD range interval for each of the 90% black pepper samples was determined. Left-out samples, both the 10% left-out black pepper samples and all adulterant and mixed samples, got a vote "in-class" when within this range, and "out-of-class" otherwise. The fused score (classification) was the majority vote per sample. In the penalized excess score approach, similarly, a leave-10%-out cross-validation on the authentic samples was performed, and the median ±3 x MAD range was determined. The scores for the left-out set within this range were set to 0. Excess scores (<> 3 x MAD) were squared, and these scores were summed per sample for all models. Final classification boundaries were set to a fixed limit of 3.

3. Results and discussion

The number of authentic samples and the diversity of conditions and geographical origins, together with the number of mixtures that were

made in-house gave a reasonable picture of the natural variation in black pepper samples. The diversity in possible adulterants, with 14 samples from 5 categories, was however relatively low. Although this set was quite comparable to many published works in the field of authenticity research, it is important to stress that this set gives only an indication of the performance of the described methods for black pepper authentication. A full validation of the method, for example according to (Alewijn, van der Voet, & van Ruth, 2016) or (Riedl, Esslinger, & Fauhl-Hassek, 2015), was found to be out of scope of this paper. Although the VIS/NIR data (Fig. 1) collected the data effectively simultaneously and provided results for the whole wavelength range displayed, noise levels in the middle of the range proved unacceptably high. This was probably caused by a low illumination of the light source in the device in this area, hence, the area was not used in modelling. The profiles obtained are visually similar to previously published studies (Orrillo et al., 2019; Wilde Amelie et al., 2019).

DART-MS data were binned to nominal mass range. After thresholding, all masses >970 *m/z* were discarded, and also from the raw data (Fig. 2) it is clear that there is little signal of compounds heavier than ~ 650 *m/z*. The largest peaks in the spectra match with the $[M\text{-}H]^-$ forms of several fatty acids (C16 and C18) (Duke, 2017), but at this stage no attempts were made for further identification of the peaks found. To our knowledge, there is no information available in literature on DART-MS on black pepper in negative mode, although some publications with (tentative) identification in positive mode are available (Chandraa et al., 2014; Liang et al., 2021).

For both techniques, population outliers, i.e., scans (for the pure black peppers) that are unlike the other black pepper scans and would negatively influence the classification performance were discarded. For both techniques, about 5% of the individual scans were considered population outliers (Table 1). Removal of those scans led only to a reduction of the number of replicates for some samples and, importantly, no samples were discarded completely. It is therefore likely that outliers were due to analytical errors, for VIS/NIR perhaps uneven surfaces or insufficient sample thickness in the Petri dish, for DART-MS perhaps the selection of scans that were recorded on the edge of the sample spot and sample mesh.

For VIS/NIR, 6 one-class and 6 two-class models were selected based on the highest AUROCs for each of the five adulterants (Table 2). In the two-class models, only partial least squares discriminant analysis (PLS-DA) algorithms were selected, with different data pretreatments and different numbers of factors. It is interesting to note that the two best-performing models for CH proved to be models that were calculated for black pepper and another adulterant. This can have a number of possible reasons: i) it may prove that the model screening did not yield the 'optimal' model, ii) that the model had difficulties in capturing the real natural variation in the adulterant class due to the relatively low number of CH samples, iii) that the VIS/NIR spectra of this adulterant is very close to one of the other adulterants in the set, or iv) even more likely, a combination of these reasons. Likewise, CS and PS had consistent high AUROCs and consequently also high correct classification rates for all two-class models, regardless if samples with these particular adulterants were used in calculating the model or if the model was calculated for any of the other adulterants. This observation alone is already a good sign that authenticity models developed for a specific matrix/adulterant system potentially are suitable for other adulterants (in the same matrix) as well.

Judging from the results in Table 2, there do not seem to be single models that perform "excellent" for all 5 adulterants simultaneously in this set. However, in all adulterants except the OD adulterations, there are models with correct classification rates of >75% for those samples with adulterant concentrations of 10% and more. The OD adulterants are an exception here, as it concerns ground pepper, and the oil to make the berries appear more shiny is only applied on the surface. The actual concentration of oil in the ground sample is therefore much lower (exact quantities are unknown). For spectroscopic techniques, low concentrations are known to be difficult to detect. A good chance of detection of >10% adulterants using NIR is in line with previous observations (McGoverin et al., 2012; Wilde Amelie et al., 2019). Flawless results have been reported using ATR-FTIR (Orrillo et al., 2019), but this was a model aimed only at a single adulterant, i.e. papaya seed.

The individual models for DART-MS behave in many ways like the individual models for VIS/NIR. There is no single model that performs 'perfectly' for all adulterants, and 2-class models that were trained again on a specific adulterant are in some cases also very powerful against other (for that model unseen) adulterants. Although not all available samples could be measured using DART-MS and consequently fewer data were available to develop models, the DART-MS models seem to perform better than the VIS/NIR models. For each of the individual adulterants there are models that detect 100% of the adulterated pepper samples. This is likely the result of the more direct determination of constituents of the authentic peppers and adulterants compared to the

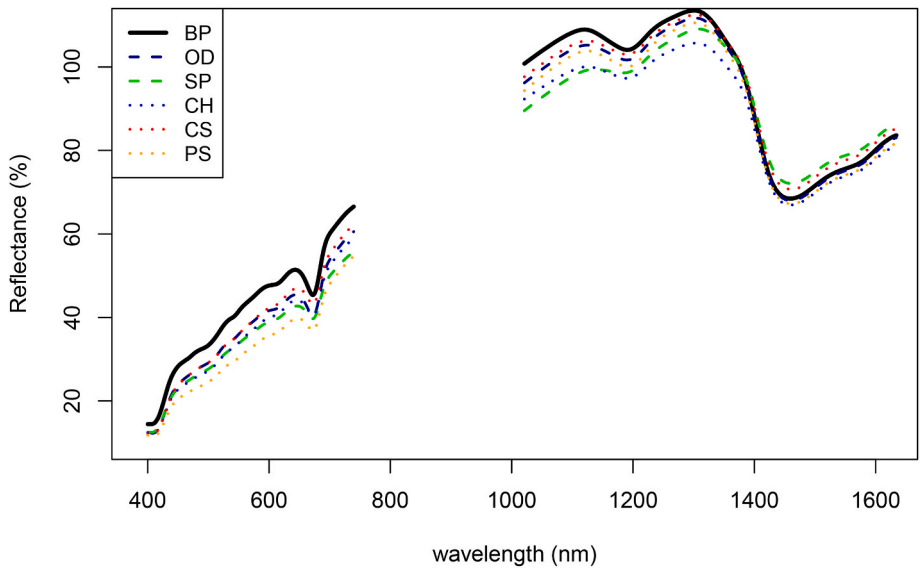

Fig. 1. VIS/NIR averaged raw spectra for pure black peppers (black line) and averages for the 10% adulterated samples (colored lines). BP= Black Pepper; OD = Oil Dressed Black Pepper, SP= Spent Pepper, CH = Coffee Husk, CS = Coffee Skin, PS = Papaya Seeds. The wavelength range between 740 and 1020 nm was considered too noisy for processing and was removed from this plot and further calculations.

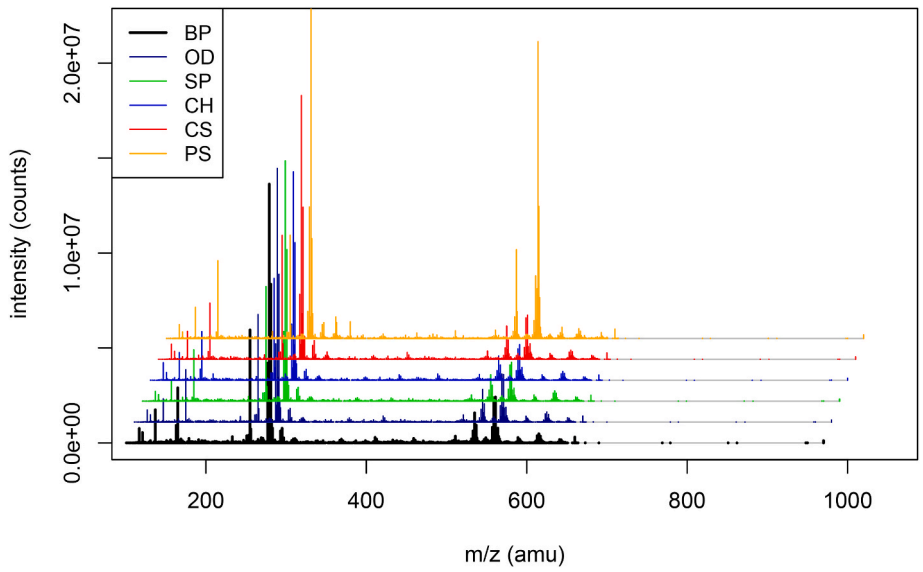

Fig. 2. DART-MS averaged raw data for pure black peppers (black line) and averages for the 10% adulterated samples (colored lines). BP= Black Pepper; OD = Oil Dressed Black Pepper, SP= Spent Pepper, CH = Coffee Husk, CS = Coffee Skin, PS = Papaya Seeds. Spectra are displayed in offset mode.

Table 2
AUROCs and correct classification rates for the manually selected models for the VIS/NIR data. Results are based on left-out sets of cross validation. All available samples and all available replicates are used in the cross-validation, results for AUROC and correct classification are based on sample level (sample replicate model scores are averaged first), and pure and samples <10% are withheld from these calculations. The classification results for the fusion model for the lower adulteration levels are reported, as percentage and in absolute numbers, in the small table below the main table.

	VIS/NIR		AUROC					Correct classification (%)					
	type	model	OD	SP	CH	CS	PS	BP	OD	SP	CH	CS	PS
V1	2Class (PS)	PLSDA (3), SNV, 1D, RR	0.66	0.96	0.99	1.00	1.00	99	6	46	63	71	79
V2	2Class (PS)	PLSDA (5), SNV, 1D, TT	0.70	0.95	0.99	1.00	0.99	98	14	56	86	100	97
V3	2Class (OD)	PLSDA (9), SNV, 1D, RR	0.86	0.68	0.88	0.98	0.99	98	37	31	34	71	87
V4	2Class (OD)	PLSDA (5), SNV, 1D, RR	0.81	0.55	0.91	0.99	0.99	99	34	20	45	86	83
V5	2Class (SP)	PLSDA (9), SNV, 1D, RR	0.55	0.98	0.98	1.00	0.97	99	12	77	72	100	62
V6	2Class (CS)	PLSDA (7), SNV, 1D, RR	0.59	0.95	0.98	1.00	0.98	99	14	60	72	100	75
V7	OCC	kNN (7), SNV, 1D, RR	0.65	0.92	0.89	0.72	0.93	97	0	69	58	29	68
V8	OCC	kNN (2), SNV, 1D (25p), RR	0.64	0.89	0.95	0.81	0.97	99	1	65	53	29	63
V9	OCC	kNN (2), SNV, DWT	0.58	0.95	0.95	0.89	0.98	99	8	63	68	29	89
V10	OCC	PCAr (3), SNV, RR	0.52	0.93	0.94	0.86	0.98	93	11	78	77	57	94
V11	OCC	PCAr (3), SNV, 1D, RR	0.59	0.96	0.97	0.88	0.97	96	2	64	70	29	89
V12	OCC	OCSVM, SNV, 1D, RR	0.67	0.88	0.95	0.82	0.97	100	1	40	60	14	80
VF1	Fusion	PES	0.65	0.94	0.95	1.00	0.99	94	17	75	81	86	94
VF2	Fusion	Avg	0.64	0.93	0.94	0.78	0.97	98	2	65	65	29	86
VF3	Fusion	Voting	0.66	0.92	0.96	0.99	1.00	97	10	71	73	57	92

adulteration level	OD	SP	CH	CS	PS
2.5%	8% (2)	22% (9)	24% (8)	–	19% (5)
5%	10% (4)	32% (11)	37% (9)	50% (3)	43% (13)
10%	7% (2)	49% (19)	53% (15)	83% (4)	83% (26)
20%	0% (0)	84% (22)	88% (20)	–	100% (24)
40%	38% (7)	100% (25)	100% (30)	–	100% (30)
100%	50% (1)	100% (4)	100% (4)	100% (1)	100% (3)

Abbreviations: 1D: first derivative (11 point window); DWT: Discrete Wavelet Transform (la8 filter); kNN: k-Nearest Neighbors (number of neighbors considered between brackets); OCSVM: One-Class Support Vector Machines; PCAr: PCA residual distance (number of principal components considered between brackets); RR: Redundant variables removed (for sets of variables with correlation >97.5% the ones with lower variances are removed); TT: *t*-test: only variables with a *t*-test result on difference of means authentic vs target class samples of <0.001 are retained.

indirect nature of the VIS/NIR signal.

Apart from the fact that in practice one does not know *a priori* for which adulterant to test, multiple models also have the potential to enhance their performance. Although the models presented are all selected for performance (highest AUROCs), are based on the same spectral data, and even have had the exact same cross-validation split scheme (no random effects due to different splits), models have clearly different performances. That is also shown by the fact that their class distance scores are only weakly correlated in many cases. Fig. 3 shows the correlation plots for the VIS/NIR set, displaying the correlation of the class distance scores of pure samples and one type of adulterant/mix, for each pair of the selected models. Neighboring models do have generally high correlations, but typically not 1, and are thus not identical. There certainly is a clear pattern that one- and two-class models lead to structurally different score mechanisms, but also within those groups the different models bring different insights on the nature of the samples.

An illustration of the power of non-correlated, multiple models is given in Fig. 4. Here, class distances for an example set of two models for pure black pepper and one adulterant (SP) are plotted. The area is separated by the 95% quantile range for the pure black pepper samples for both models. Therefore, the green rectangle represents the area where samples would be considered 'authentic'. The red shaded areas is where the models agree that samples are non-authentic, and there is no added value in combining models. The orange and purple shaded area, however, represents the areas where one model would consider a sample normal, but the other model would not. This is the area where benefit can be made from combining models. But as the example also illustrates is the danger that authentic samples are in increased risk of being misclassified. A viable solution for model fusion avoids the latter, while still detecting more truly adulterated samples.

As mentioned above, there is no established way of method fusion, here we will apply just three variants on the data sets presented and study their potentials to detect more than one authenticity issue, limiting false negatives (Tables 2 and 3). It is clear that simple averaging the class distance model results for each sample yield poor results, the aggregate performs generally below average of the individual models for each adulterant, for both techniques, both in correct classification rates and in AUROC. Most probably this situation is similar to creating a two-class model (adulterated or not) when the adulterated class is clearly consisting of different constituents, adulterated samples may lie on different positions in the multivariate domain with regard to the authentic samples, and the average might overlap with the authentic sample space. The fusion approach by voting and by the PES both yield classification scores that are as good as, or close to, the performance of the best individual model. In the VIS/NIR case, the PES system is slightly better, although it also creates more false negatives. In both cases, however, the number of false negatives is still sufficiently low for most practical situations, and increased only very lightly at the benefit of overall performance gain. Note that the outliers (pure pepper misclassifications) reported in the individual models are largely different samples, so simply applying all individual models in series would also greatly affect the number of false negatives (data not shown). Data fusion using different models based on the same data therefore seems to have a clear benefit over using just one 'optimal' model in terms of performance. Moreover, when the nature of the adulterant is unknown, data from the different models in the aggregation phase also enables to plot the multivariate direction of possible deviations, which can be used

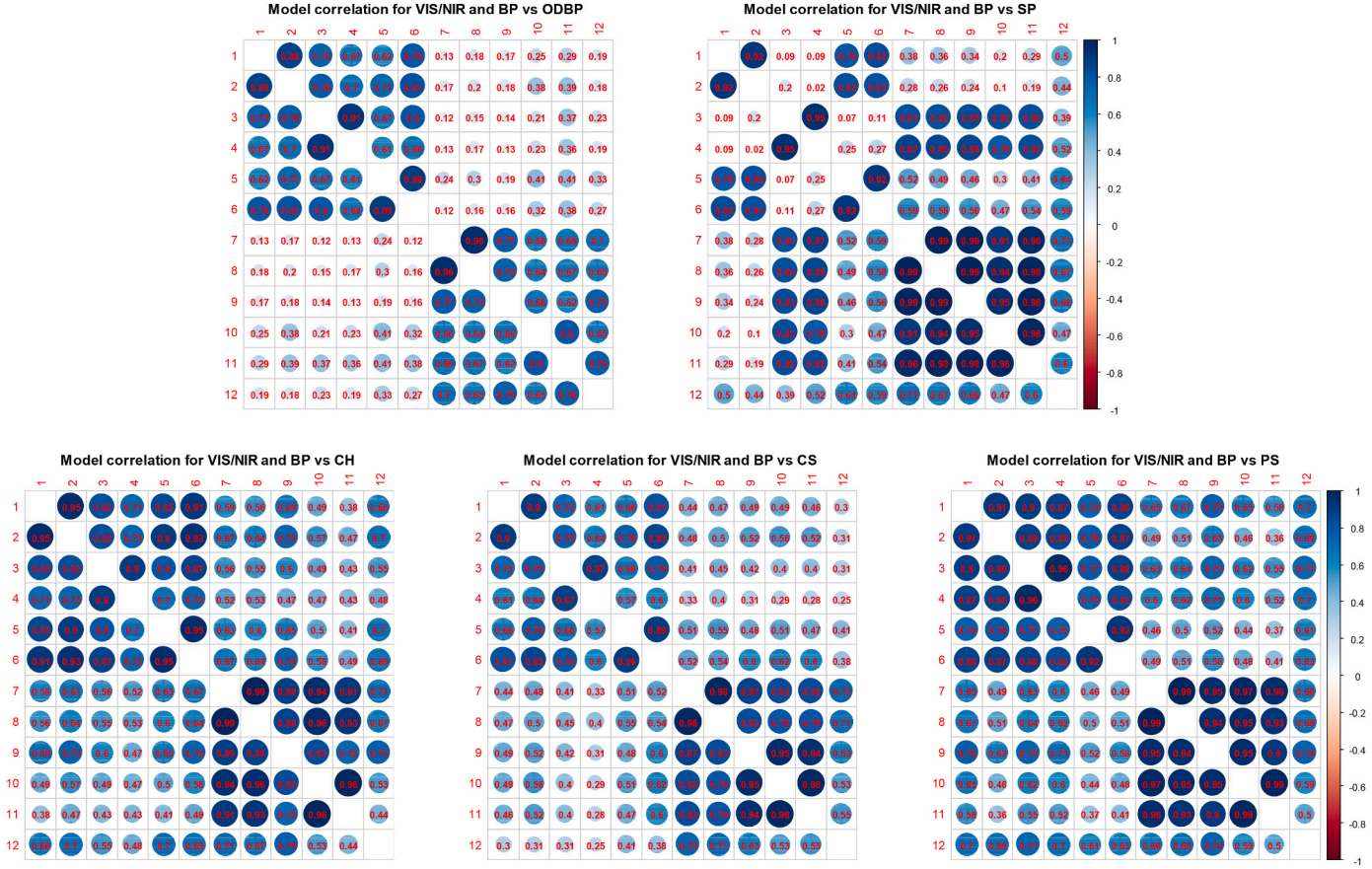

Fig. 3. Correlation of classification model scores from the VIS/NIR data, per sample, black peppers and one adulterant/mix, for the models selected manually. Although generated on the same dataset and resulting from the left-out sets from the same cross-validation instances, there is little correlation in many cases, presumably adding to the strength of model combinations.

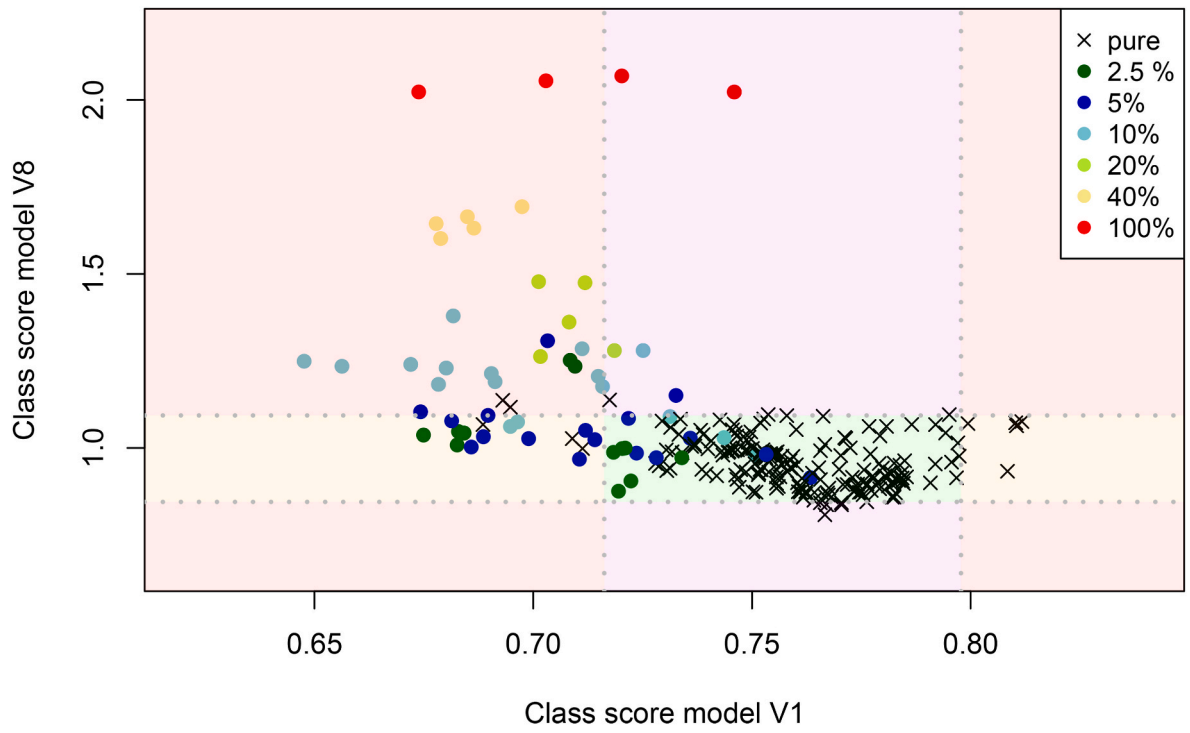

Fig. 4. Illustration of the power of multiple models: sample class scores for pure black peppers and samples adulterated with the indicated concentrations of spent pepper (SP), using models V1 and V8 (manually selected, see Table 2) based on the VIS/NIR data. Grey dotted lines represent 95% confidence intervals for both models. Area shading; *green:* pure samples according to both models; *yellow:* pure according to model V8 but adulterated according to model V1; *purple:* pure according to model V1 but adulterated according to model V8; *red:* adulterated according to both individual models.

Table 3
AUROCs and correct classification rates for the manually selected models for the DART-MS data. Results are based on left-out sets of cross validation. All available samples and all available replicates are used in the cross-validation, results for AUROC and correct classification are based on sample level (sample replicate model scores are averaged first), and pure and samples <10% are withheld from these calculations. The classification results for the fusion model for the lower adulteration levels are reported, as percentage and in absolute numbers, in the small table below the main table.

DART-MS			AUROC					Correct classification (%)					
	type	model	OD	SP	CH	CS	PS	BP	OD	SP	CH	CS	PS
D1	2Class (CS)	PLSDA (3), RN	0.70	0.91	0.94	0.92	0.74	100	7	19	25	71	73
D2	2Class (CS)	SIMCA (5), RN	0.74	0.83	0.84	0.85	0.63	100	21	56	69	14	40
D3	2Class (CS)	RF, ^{10}log	0.75	0.56	0.67	0.60	1.00	100	100	88	75	14	60
D4	2Class (OD)	PLSDA (5), RN	0.62	0.71	0.86	0.70	0.60	100	0	25	56	14	87
D5	2Class (OD)	RF, ^{10}log	0.74	0.56	0.86	0.60	0.57	100	0	100	75	86	47
D6	2Class (PS)	PLSDA (9), RN	0.53	0.63	0.57	1.00	0.55	100	7	25	25	0	100
D7	2Class (PS)	OCSVM, ^{10}log	0.69	0.92	0.79	1.00	0.68	98	100	100	100	100	40
D8	OCC	PCAr, ^{10}log, RR	0.66	0.62	0.80	0.82	0.95	95	50	31	56	29	80
D9	OCC	Maha, RR	0.58	0.79	0.54	0.97	0.82	100	0	25	56	14	87
DF1	Fusion	Average	0.67	0.84	0.65	0.94	0.79	100	7	25	56	14	80
DF2	Fusion	Voting	1.00	1.00	0.98	1.00	1.00	100	100	100	100	86	100
DF3	Fusion	PES	1.00	1.00	1.00	1.00	1.00	100	100	100	100	71	100

adulteration level	OD	SP	CH	CS	PS
2.5%	100% (6)	92% (12)	89% (18)	–	100% (5)
5%	83% (6)	92% (12)	83% (6)	100% (6)	100% (6)
10%	100% (6)	100% (12)	100% (6)	67% (6)	100% (6)
40%	100% (6)	–	100% (6)	–	100% (6)
100%	100% (2)	100% (4)	100% (4)	100% (1)	100% (3)

Abbreviations: ^{10}log: logarithmic transformation (base 10); Maha: Mahalanobis distance; OCSVM: One-Class Support Vector Machines; PCAr: PCA residual distance (number of principal components considered between brackets); RF: Random Forest; RN: Row-wise Normalization; RR: Redundant variables removed (for sets of variables with correlation >97.5% the ones with lower variances are removed); SIMCA: Soft independent modelling of class analogies.

to discriminate between the known possible adulterants.

Note that we did not perform data fusion based on the results from two different analytical techniques. In the case of two imperfect predictions it is hypothesized that additional analytical data gives the model even more diverse insight, and thus a better chance on synergy across models. As the DART-MS models in this study perform (near-)

perfect for the given set, there is little room to show significant improvement of the results, and we aim to investigate this in an upcoming paper.

4. Conclusion

Both VIS/NIR and DART-MS are viable techniques to detect a variety of possible adulterants in black pepper samples. For each of the five analysed adulterants it is possible to make 'traditional' chemometric models that are capable to detect fraud-relevant levels of adulterants. The OD adulteration for VIS/NIR was the only exception. In practice, one does not know if and which adulterant is present in any given sample, hence, it is desired to check for all possible adulterants simultaneously. Applying different individual models in sequence multiplies the risk in obtaining false positives (authentic samples being misclassified), and it seems better to apply model fusion into a final score. The added benefit was found to be that the combined test performs well for a broad set of adulterants. The number of real-life possible adulterants in black pepper is practically countless, and is unfeasible to develop and test all options. While this study only involved five adulterants, it is believed that a wide variety of other adulterants will be detected using the fusion technique, as long as there is measurable effect on the analytical technique applied. A survey on broader, unseen, adulterants would be a valuable next step. Extended validation of the current models with new black pepper samples, from different origins and different harvest periods, would be another prerequisite before the current method would be ready for routine use.

CRediT authorship contribution statement

Martin Alewijn: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Vasiliki Akridopoulou:** Investigation. **Tjerk Venderink:** Methodology, Investigation. **Judith Müller-Maatsch:** Writing – original draft, Writing – review & editing. **Erika Silletti:** Conceptualization, Supervision.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgements

This work was funded by the Topsector Agri&Food project AF17038, Fast on-site methods for food safety and authenticity (https://topsectoragrifood.nl/project/af-17038-snelle-on-site-methoden-voor-voedselveiligheid-en-authenticiteit/). The authors gratefully acknowledge the industrial partners who provided the majority of the samples used in this study.

References

Alewijn, M., van der Voet, H., & van Ruth, S. (2016). Validation of multivariate classification methods using analytical fingerprints–concept and case study on organic feed for laying hens. *Journal of Food Composition and Analysis, 51*, 15–23.

Bellinger, C., Sharma, S., Zaiane, O. R., & Japkowicz, N. (2017). Sampling a longer life: Binary versus one-class classification revisited. In *First international workshop on learning with imbalanced domains: Theory and applications* (pp. 64–78). PMLR.

Black, C., Haughey, S. A., Chevallier, O. P., Galvin-King, P., & Elliott, C. T. (2016). A comprehensive strategy to detect the fraudulent adulteration of herbs: The oregano approach. *Food Chemistry, 210*, 551–557.

Chandraa, P., Bajpaia, V., Srivastvab, M., Kumarc, K. B. R., & Kumara, B. (2014). Metabolic profiling of piper species by direct analysis in real time mass spectrometry combined with principal component analysis. *Analytical Methods, 6*, 4234–4239.

Danezis, G. P., Tsagkaris, A. S., Brusic, V., & Georgiou, C. A. (2016). Food authentication: State of the art and prospects. *Current Opinion in Food Science, 10*, 22–31.

Deng, X., Li, W., Liu, X., Guo, Q., & Newsam, S. (2018). One-class remote sensing classification: One-class vs. binary classifiers. *International Journal of Remote Sensing, 39*(6), 1890–1910.

Dhanya, K., Syamkumar, S., & Sasikumar, B. (2009). Development and application of SCAR marker for the detection of papaya seed adulteration in traded black pepper powder. *Food Biotechnology, 23*(2), 97–106.

Duke, J. A. (2017). *Handbook of phytochemical constituents of GRAS herbs and other economic plants*. Routledge.

FAO. (2022). Faostat. License: CC BY-NC-SA 3.0 IGO. Extracted from: https://www.fao.org/faostat/en/#data/QCL. Date of Access. (Accessed 11 July 2022).

Galvin-King, P., Haughey, S. A., & Elliott, C. T. (2018). Herb and spice fraud; the drivers, challenges and detection. *Food Control, 88*, 85–97.

Gul, I., Nasrullah, N., Nissar, U., Saifi, M., & Abdin, M. Z. (2018). Development of DNA and GC-MS fingerprints for authentication and quality control of *Piper nigrum* L. And its adulterant *Carica papaya* L. *Food Analytical Methods, 11*(4), 1209–1222.

Guo, T., Yong, W., Jin, Y., Zhang, L., Liu, J., Wang, S., et al. (2017). Applications of DART-MS for food quality and safety assurance in food supply chain. *Mass Spectrometry Reviews, 36*(2), 161–187.

Hu, L., Yin, C., Ma, S., & Liu, Z. (2018). Assessing the authenticity of black pepper using diffuse reflectance mid-infrared Fourier transform spectroscopy coupled with chemometrics. *Computers and Electronics in Agriculture, 154*, 491–500.

Kaavya, R., Pandiselvam, R., Mohammed, M., Dakshayani, R., Kothakota, A., Ramesh, S., et al. (2020). Application of infrared spectroscopy techniques for the assessment of quality and safety in spices: A review. *Applied Spectroscopy Reviews, 55*(7), 593–611.

Kucharska-Ambrożej, K., & Karpinska, J. (2020). The application of spectroscopic techniques in combination with chemometrics for detection adulteration of some herbs and spices. *Microchemical Journal, 153*, Article 104278.

Leys, C., Delacre, M., Mora, Y., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology, 32*.

Liang, J., Sun, J., Chen, P., Frazier, J., Benefield, V., & Zhang, M. (2021). Chemical analysis and classification of black pepper (Piper nigrum L.) based on their country of origin using mass spectrometric methods and chemometrics. *Food Research International, 140*, Article 109877.

Massaro, A., Negro, A., Bragolusi, M., Miano, B., Tata, A., Suman, M., et al. (2021). Oregano authentication by mid-level data fusion of chemical fingerprint signatures acquired by ambient mass spectrometry. *Food Control, 126*, Article 108058.

McGoverin, C. M., September, D. J., Geladi, P., & Manley, M. (2012). Near infrared and mid-infrared spectroscopy for the quantification of adulterants in ground black pepper. *Journal of Near Infrared Spectroscopy, 20*(5), 521–528.

Müller-Maatsch, J., Alewijn, M., Wijtten, M., & Weesepoel, Y. (2021). Detecting fraudulent additions in skimmed milk powder using a portable, hyphenated, optical multi-sensor approach in combination with one-class classification. *Food Control, 121*, Article 107744.

Negi, A., Pare, A., & Meenatchi, R. (2021). Emerging techniques for adulterant authentication in spices and spice products. *Food Control*, Article 108113.

Nobari-Moghaddam, H., Tamiji, Z., Akbari-Lakeh, M., Khoshayand, M. R., & Haji-Mahmoodi, M. (2021). Multivariate analysis of food fraud: A review of NIR based instruments in tandem with chemometrics. *Journal of Food Composition and Analysis*, Article 104343.

Oliveira, M., Cruz-Tirado, J., Roque, J., Teófilo, R., & Barbin, D. (2020). Portable near-infrared spectroscopy for rapid authentication of adulterated paprika powder. *Journal of Food Composition and Analysis, 87*, Article 103403.

Oliveira, M. M., Cruz-Tirado, J., & Barbin, D. F. (2019). Nontargeted analytical methods as a powerful tool for the authentication of spices and herbs: A review. *Comprehensive Reviews in Food Science and Food Safety, 18*(3), 670–689.

Orrillo, I., Cruz-Tirado, J., Cardenas, A., Oruna, M., Carnero, A., Barbin, D. F., et al. (2019). Hyperspectral imaging as a powerful tool for identification of papaya seeds in black pepper. *Food Control, 101*, 45–52.

Osman, A. G., Raman, V., Haider, S., Ali, Z., Chittiboyina, A. G., & Khan, I. A. (2019). Overview of analytical tools for the identification of adulterants in commonly traded herbs and spices. *Journal of AOAC International, 102*(2), 376–385.

Paradkar, M. M., Singhal, R. S., & Kulkarni, P. R. (2001). A new TLC method to detect the presence of ground papaya seed in ground black pepper. *Journal of the Science of Food and Agriculture, 81*(14), 1322–1325.

Parvathy, V. A., Swetha, V. P., Sheeja, T. E., Leela, N. K., Chempakam, B., & Sasikumar, B. (2014). DNA barcoding to detect chilli adulteration in traded black pepper powder. *Food Biotechnology, 28*(1), 25–40.

Reinholds, I., Bartkevics, V., Silvis, I. C., van Ruth, S. M., & Esslinger, S. (2015). Analytical techniques combined with chemometrics for authentication and determination of contaminants in condiments: A review. *Journal of Food Composition and Analysis, 44*, 56–72.

Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica Acta, 885*, 17–32.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1249.

Sasikumar, B., Swetha, V., Parvathy, V., & Sheeja, T. (2016). Advances in adulteration and authenticity testing of herbs and spices. In *Advances in food authenticity testing* (pp. 585–624). Elsevier.

Shen, Y., van Beek, T. A., Claassen, F. W., Zuilhof, H., Chen, B., & Nielen, M. W. F. (2012). Rapid control of Chinese star anise fruits and teas for neurotoxic anisatin by Direct Analysis in Real Time high resolution mass spectrometry. *Journal of Chromatography A, 1259*, 179–186.

Silvis, I., Van Ruth, S., Van Der Fels-Klerx, H., & Luning, P. (2017). Assessment of food fraud vulnerability in the spices chain: An explorative study. *Food Control, 81*, 80–87.

Team, R. C. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ulberth, F. (2020). Tools to combat food fraud–a gap analysis. *Food Chemistry, 330*, Article 127044.

Wang, Y., Mei, M., Ni, Y., & Kokot, S. (2014). Combined NIR/MIR analysis: A novel method for the classification of complex substances such as illicium verum hook. F. And its adulterants. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 130*, 539–545.

Wilde Amelie, S., Simon, A., Haughey, P. G.-K., & Elliott, C. T. (2019). The feasibility of applying NIR and FT-IR fingerprinting to detect adulteration in black pepper. *Food Control, 100*, 1–7.