International Journal of
Food Science
+Technology

Original article

Novel method for the detection of adulterants in coffee and the determination of a coffee's geographical origin using near infrared spectroscopy complemented by an autoencoder

Leah Munyendo,[1]* Daniel Njoroge,[2] Yanyan Zhang[3] & Bernd Hitzmann[1]

1 Department of Process Analytics and Cereal Science, University of Hohenheim, Garbenstr. 23, 70599, Stuttgart, Germany
2 Institute of Food Bio-resources Technology, Dedan Kimathi University of Technology, Private Bag, 10143, Dedan Kimathi, Nyeri, Kenya
3 Department of Flavor Chemistry, University of Hohenheim, Fruwirthstraße 12, 70599, Stuttgart, Germany

(Received 15 September 2022; Accepted in revised form 31 December 2022)

Summary Coffee authenticity is a foundational aspect of quality when considering coffee's market value. This has become important given frequent adulteration and mislabelling for economic gains. Therefore, this research aimed to investigate the ability of a deep autoencoder neural network to detect adulterants in roasted coffee and to determine a coffee's geographical origin (roasted) using near infrared (NIR) spectroscopy. Arabica coffee was adulterated with robusta coffee or chicory at adulteration levels ranging from 2.5% to 30% in increments of 2.5% at light, medium and dark roast levels. First, the autoencoder was trained using pure arabica coffee before being used to detect the presence of adulterants in the samples. Furthermore, it was used to determine the geographical origin of coffee. All samples adulterated with chicory were detectable by the autoencoder at all roast levels. In the case of robusta-adulterated samples, detection was possible at adulteration levels above 7.5% at medium and dark roasts. Additionally, it was possible to differentiate coffee samples from different geographical origins. PCA analysis of adulterated samples showed grouping based on the type and concentration of the adulterant. In conclusion, using an autoencoder neural network in conjunction with NIR spectroscopy could be a reliable technique to ensure coffee authenticity.

Keywords adulteration, autoencoder, chicory, coffee, geographical origin, NIR spectroscopy.

Introduction

Food authenticity has become a very challenging issue in the food industry, especially among high-value products such as coffee. Coffee is among the most consumed beverage globally with a consumption amounting to approximately 166.346 million bags during the 2020/2021 period (ICO, 2021). So far, over 100 species within the genus *Coffea* have been identified with *Coffea arabica* (arabica) and *Coffea canephora* (robusta) being the most widely cultivated from an economic and commercial point of view (Ferreira *et al*., 2019). There are quality differences between the two species resulting from differences in their genetic make-up, as well as the agronomic conditions for their cultivation. On the world market, arabica coffee, which accounts for more than 60% of the world's production, is generally regarded to have a superior cup quality and is exclusively sought after by most consumers (Barbosa *et al*., 2019; ICO, 2021). This makes it more expensive than robusta coffees.

Owing to the high commercial value of arabica coffees, their adulteration for economic gain with cheaper robusta coffee and other lower value materials, such as coffee husks, chicory, corn, barley, rice, wheat, spent coffee grounds, *etc*., has become a wide-spread practice. A food fraud report by the European Commission in 2018 highlighted that a high amount of coffee sold in the UK as 100% arabica coffee contained cheaper robusta coffee. Ten percent of the samples tested contained robusta coffee in proportions ranging from 1.6% to more than 21% (European Commission, 2018). In Brazil, the Brazilian Association of Coffee Industries carried out an inspection on 2400 brands present on the market, of which, 583 brands, representing 25% of Brazilian national brands, were found to be adulterated with coffee husks, rye, maize or brown sugar (Peixoto, 2009). The problem of

*Correspondent: E-mail: leah.munyendo@uni-hohenheim.de

doi:10.1111/ijfs.16283

adulteration is particularly concerning in roasted-ground coffee due to its colour and particle size. These illegal practices do not only result in regulation non-compliance regarding safety and quality, but may also imply a danger to consumers' health. On the other hand, unintentional mix-ups of the two coffee species (arabica and robusta) may happen along the processing line, necessitating quality control measures.

The sensory properties of coffee beverage are highly influenced by a coffee's geographical origin. On the global coffee trade market, it has become important to indicate the origin of coffee since there is an increasing consumer interest in high quality coffee from beans of known origin. Additionally, there are differences in coffee price depending on the origin (Barjolle *et al.*, 2017). Another important aspect in coffee trade is coffee blending done by mixing coffee from different geographical origins to meet consumer and market demand. Although this practice increases the commercial value and quality of the final product, it also enables fraud and adulteration practices involving both geographical and compositional practices (Šeremet *et al.*, 2022). In this respect, the possibility to verify the safety, quality and geographical origin of coffee would be highly useful.

Different analytical strategies employing biological, physical and chemical methods have been developed in the past years for this purpose. They include chromatographic methods (Martins *et al.*, 2018; Song *et al.*, 2019; Núñez *et al.*, 2020, 2021), DNA based approaches (Ferreira *et al.*, 2016; Uncu & Uncu, 2018; Couto *et al.*, 2019) and ultraviolet–visible spectrophotometry (UV–VIS) (Souto *et al.*, 2015; Rahman *et al.*, 2018). These techniques require skilled personnel, sophisticated and expensive instrumentation, are labor intensive, time consuming, environmentally unfriendly (chemicals used) and can detect only a few adulterants at a time (Burns & Walker, 2020). Therefore, robust analytical methods are necessary for effective assessment of coffee quality. Spectroscopic techniques such as infrared and nuclear magnetic resonance in combination with chemometrics have gained a substantial interest in this regard, especially for determining a coffee's geographical origin and detection of adulteration because of their high efficiency, ease of use, low-cost, rapidity and non-destructiveness (Toci *et al.*, 2016; Medina *et al.*, 2017; Mendes & Duarte, 2021). With an easing of chemometric computation, NIR spectroscopy has been used extensively to detect different adulterants such as rice, barley, coffee husks, corn, soil and chicory in coffee (Ebrahimi-Najafabadi *et al.*, 2012; Winkler-Moser *et al.*, 2015; Correia *et al.*, 2018; Forchetti & Poppi, 2020; Harohally & Thomas, 2021; Couto *et al.*, 2022). It has also been used for the determination of coffees' geographical origin (Medina *et al.*, 2017).

Considering coffee fraud is an important aspect of food safety, much simpler techniques are necessary for the detection of adulteration and determination of a coffee's geographical origin. An autoencoder is a type of artificial neural network that has been used in other fields for anomaly detection and could have a significant application in the coffee industry (Hasan *et al.*, 2016). Anomaly detection refers to a process of finding data that are significantly different from others in a data set. It can be classified into three classes based on the availability of the labels, that is, supervised, semi-supervised and unsupervised anomaly detection. Out of these, unsupervised anomaly detection is the most meaningful in practical applications where anomalies are detected in a data set without using any annotations (Cheng *et al.*, 2021). One powerful tool in modelling high-dimensional data in an unsupervised setting is a deep autoencoder. It consists of an encoder, which compresses the input data into a few latent space variables also known as a bottleneck, and a decoder that reconstructs the initial data from the latent space variables (Finke *et al.*, 2021). The two parts are trained together as one neural network to reconstruct the input data as well as possible, that is, with low reconstruction error by choosing a suitable loss function. From an anomaly detection perspective, the autoencoder is typically trained on a normal data set where it learns a representation that uses the structure of the training data and is thus specific for this set. Therefore, if it encounters new data that have different features from the training set, it should not be able to encode and decode these features resulting in a higher reconstruction error, which is an indicator of anomaly (Finke *et al.*, 2021).

Application of autoencoders in the food industry is still not common. However, there are some reports on its application in anomaly detection (*e.g.*, changes in the temperature of milk, its fat content, and the addition of cleaning solution or water) during milk processing using NIR spectroscopic data (Vasafi *et al.*, 2021). In the context of coffee, other artificial neural network approaches have been used to quantify adulterants in coffee and determine a coffee's geographical origin. Convolutional neural network techniques are reported as a feasible alternative to classical chemometrics for the quantification of coffee adulterants (Chakravartula *et al.*, 2022). Although many studies on the detection of adulterants in coffee and the determination of coffees' geographical origin using spectroscopic data exist in literature (Medina *et al.*, 2017; Flores-Valdez *et al.*, 2020; Wongsaipun *et al.*, 2021; Couto *et al.*, 2022), no study has focused on the use of autoencoders. The interpretation of results from the autoencoder neural network is much simpler compared to other techniques, thus, making it an appropriate approach for coffee fraud detection, as well as quality assessment. Therefore, the aim of this research was to evaluate the ability of an autoencoder to detect adulterants in roasted coffee, and to determine the geographical origin of roasted coffee. The hypothesis of this study was that NIR spectroscopy complemented by an autoencoder provides a feasible method to detect adulterants in roasted coffee and to determine a coffee's geographical origin (roasted).

Materials and methods

Materials

Raw arabica and robusta coffees were purchased from Buxtrade GmbH, An den Geestbergen 1, 21 614 Buxtehude, Germany, and Hochland Kaffee Hunzelmann GmbH und Co. KG, Germany, respectively. Raw chicory root was sourced from Detrade UG, Bruchstrasse 14 d, Stuhr, Germany. All samples were roasted at three different roast levels as indicated in Section 2.2. All samples were then ground and arabica coffee samples were adulterated with robusta coffee or chicory at different adulteration levels. The experimental conditions selected focused on different types and varying amounts of adulterants and roasting times with the aim of generating a fair number of distinct coffee samples that could be present on the market. The coffees were from the following regions: Kenya, Guatemala, Colombia and a blend from Central and South America (all arabica) and India (robusta coffee).

Sample roasting and preparation of adulterated arabica coffee

The coffee samples were roasted in a Gene Cafe CBR-101 coffee roaster at 240 °C for 10, 15 and 20 min, corresponding to light, medium and dark roasts, respectively. A blend of arabica coffee from Central and South America was used to prepare adulterated samples. Chicory root was roasted using the same roaster at the same temperatures but for 4, 5 and 6 min. Shorter times were necessary to achieve a similar colour to that of the coffee because of its small size and low moisture content. All the samples were ground with an electric grinder (Melitta Calibra EU 1027–01 Mill 160 W, Germany) on a fine grind setting. Afterwards, adulterated arabica coffee samples at different concentrations were prepared by separately weighing the ground samples, that is, chicory, arabica and robusta coffee and mixing mechanically using a 3D mixer (Turbula Willy A. Bachofen, Switzerland) for 5 min. The mixing of various proportions of adulterants yielded six classes of adulterated samples: three classes containing arabica coffee and chicory (light, medium and dark roast) and three classes containing arabica and robusta coffee (light, medium and dark roast). Each adulterant was added at adulteration levels ranging from 2.5% to 30% (w/w) in increments of 2.5% resulting in 72 adulterated samples. All adulteration levels were prepared in duplicate.

Acquisition of NIR spectra

The spectra of adulterated arabica coffee samples and arabica coffee from Kenya, Guatemala, Colombia, and a blend from Central and South America were obtained using a Fourier Transform NIR spectrometer (Bruker, Germany) equipped with OPUS software (Version 7, Bruker, Germany). The spectra were acquired in diffuse reflectance mode between the spectral range of 12 500 cm^{-1} and 3600 cm^{-1} with 4 cm^{-1} resolution resulting in 4615 channels. Each spectrum was recorded as an average of 64 scans. At each roast level, 212 spectra of the arabica coffee blend from Central and South America were obtained. For adulterated coffee samples and coffee from Kenya, Guatemala and Colombia, 50 spectra were recorded for each roast level (Table 1). NIR data were pre-processed using the standard normal variate (SNV) method to remove scattering effects prior to further analysis (Rinnan *et al.*, 2009). Preliminary results using different pre-processing methods showed SNV as the best technique for data transformation.

Exploratory data analysis

For the unsupervised method, Principal component analysis (PCA) models were built using Unscrambler software X version 10.3 (CAMO Software AS., Oslo, Norway). This enabled visualisation of patterns between pure arabica coffee and arabica adulterated with robusta coffee or chicory at different roast levels. To achieve this, all the wavenumbers of a spectrum were used and PCA score plots drawn with the first two principal components. Analysis was done based on the non-linear iterative partial least squares algorithm (NIPALS). Prior to PCA analysis, spectra of every sample were averaged to obtain five replications to reduce scattering among replications for easy visualisation.

Autoencoder neural network

Autoencoder is an unsupervised neural network that attempts to copy its input to its output through a back propagation learning procedure. Its architecture consists of an encoder and a decoder, where the information only moves forward, that is, from the input nodes, through the hidden nodes to the output nodes (Fig. 1). From the figure, there are fewer variables in the hidden layer (latent space) than in the input and output layers. Thus, the autoencoder is forced to extract in its latent space the variables that signify the

Table 1 Number of spectra/samples used for calibration set, selection of best neural network structure and prediction set

	Total number of recorded spectra	**Number of spectra for calibration set**	**Spectra for selection of the best neural network structure**	**Number of spectra for prediction set**
Adulterated arabica coffee samples and coffee from different regions				
Central and South_America blend	212	162	15	35
Kenya	50	0	15	35
Guatemala	50	0	15	35
Colombia	50	0	15	35
2.5%	50	0	15	35
5%	50	0	15	35
7.5%	50	0	15	35
10%	50	0	15	35
12.5%	50	0	15	35
15%	50	0	15	35
17.5%	50	0	15	35
20%	50	0	15	35
22.5%	50	0	15	35
25%	50	0	15	35
27.5%	50	0	15	35
30%	50	0	15	35

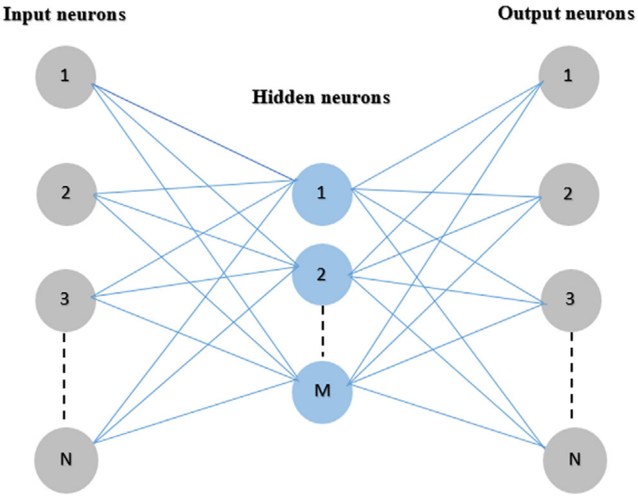

Figure 1 The autoencoder structure. M and N are the number of neurons in the hidden layer and in the input and output layer, respectively. Here N = 4,615 while M was tested with three to twenty neurons.

most prominent features of the input data, that is, it extracts correlations in the input data that allow for effective compression of the data (encoding). Finally, the decoder takes the output of the latent space and attempts to recreate the input (Finke *et al.*, 2021).

In this study, a three-layer (I*15*O) feedforward backpropagation neural network was applied. 'I', '15' and 'O' refer to the input, hidden and output neurons, respectively. At first, different autoencoder neural network structures were trained using spectra of pure arabica coffee (a blend from Central and South America). The best network structure was then selected and validated using adulterated coffee samples and arabica coffees from Kenya, Guatemala, Colombia, and a blend from Central and South America (not included in the training set). Specifically, 162 spectra of an arabica coffee blend from Central and South America at each roast level were used to train the autoencoder, that is, the autoencoder was trained with a spectra of light, medium and dark roasted coffee samples. Sixteen structures were tested, each consisting of one or two hidden layers. The number of neurons within these hidden layers ranged from three to 20. Before the training process, the data were randomly split into a training and a test set. Training was stopped when the mean square error (MSE) of the test set increased in order to avoid overfitting. MSE for each sample was calculated as follows during the training of the autoencoder.

$$\mathrm{MSE} = \frac{1}{n}\left(\vec{\mathrm{x}}_{\mathrm{in}}-\vec{\mathrm{x}}_{\mathrm{out}}\right)^{\mathrm{T}}\left(\vec{\mathrm{x}}_{\mathrm{in}}-\vec{\mathrm{x}}_{\mathrm{out}}\right)$$

where n is equal to the number of wavenumber channels used, and $\vec{\mathrm{x}}_{\mathrm{in}}$ and $\vec{\mathrm{x}}_{\mathrm{out}}$ are the spectra (as a vector) for the input and output of the autoencoder, respectively, while T refers to the transposed vector.

After autoencoder training, 15 spectra of each adulteration level, and those of coffees from Kenya, Guatemala, Colombia, and a blend from Central and South America (not included in the training set) were used to choose the best performing neural network structure in terms of prediction. The best neural network structure was then used for prediction with the remaining data

(Table 1). The highest MSE that occurred during autoencoder training was set as a limit to determine if a spectrum was different from the training set spectra. In the prediction set, MSE higher than the set limit gave important information on the ability of the neural network to detect adulteration and coffee from different regions, that is, samples with MSE higher than the highest recorded during training were considered detectable. The autoencoder neural network was developed using MATLAB software's proprietary script language (version 2019b) with Deep Learning Toolbox (version 13.0).

Results and discussion

Spectra overview of adulterated samples

Average raw spectra of pure and adulterated (30% robusta or 30% chicory) arabica coffee samples at light, medium and dark roast levels are presented in Fig. 2. The spectral profile as well as curve trend was similar for all the samples at different roast levels. Comparing pure arabica coffee and that adulterated with 30% robusta coffee at light (Fig. 2a), medium (Fig. 2b), and dark (Fig. 2c) levels, one can observe slight differences in intensities for light roasted samples along the spectrum. At light roast, chemical compounds in roasted coffee are less degraded and thus a sample adulterated with robusta may contain these compounds in different amounts compared to an unadulterated sample, which could explain the observed results. The spectra for medium and dark roasted samples overlapped, which may be associated with degradation of compounds that were important in differentiating the samples (Tfouni *et al.*, 2012). Differences in intensities between pure arabica coffee and that adulterated with 30% chicory were observed for light (Fig. 2d), medium (Fig. 2e) and dark (Fig. 2f) roast levels. However, for the dark roast level, one can observe distinct differences among the samples particularly in the region between 4000 and 6000 cm^{-1} mainly related to the C–H, N–H, C–C, O–H and C=O plus O–H vibrations, which may be correlated with organic acids, carbohydrates, proteins and chlorogenic acids (Santos *et al.*, 2016). It is important to note that chicory and coffee have different chemical compounds, and thus, adding it to coffee could change the composition of adulterated sample, which may explain these observations. A study by Gandra *et al.* (2017) demonstrates that adding adulterants to coffee reduces its phenolic compounds and caffeine content.

In general, characteristic peaks were attributed to the presence of water, carbohydrates, proteins, chlorogenic acids, caffeine and trigonelline and were observed in the following spectra regions; 3800–4400 cm^{-1} characterised mainly by C–H bond vibrations; 4500–4900 cm^{-1} by O–H and the 2nd overtone of C=O bond vibrations; 5000–5500 cm^{-1} by the 1st overtone and combination band of C=O and O–H bonds; 5600–6100 cm^{-1} by the 1st overtone of C–H bonds and 6500–7100 cm^{-1} by the 1st overtone of O-H stretching and deformation (Munyendo *et al.*, 2021).

PCA analysis

Principal component analysis analysis was done to visualise patterns between pure arabica coffee and that adulterated with robusta coffee or chicory at adulteration levels ranging from 2.5% to 30% in increments of 2.5%. (Fig. 3). Two-dimensional score plots for pure arabica and arabica coffee adulterated with robusta coffee or chicory at light, medium and dark roast levels are presented in Fig. 3a–c, respectively. In all the roasts, PC1 and PC2 accounted for more than 95% of the total explained variance. There is a clear separation of pure arabica coffee from the ones adulterated at all levels indicating possible differences in chemical composition among the samples (Gandra *et al.*, 2017). Additionally, PCA efficiently distinguished the samples based on the type of adulterant added, as well as the concentration. Samples with lower adulterant concentration (*i.e.*, 2.5%) were located close to pure arabica coffee while those with high concentration far away. Similar findings have been reported using different adulterants (Chakravartula *et al.*, 2022). Considering arabica coffee adulterated with chicory at dark roast level (Fig. 3c), there was a clear discrimination among all adulterant concentrations. The positioning of the samples was in the direction of the positive quadrant of PC1 as the adulteration level increases. In general, NIR was able to effectively identify differences among samples based on their chemical composition regardless of their roast level.

Detection of adulteration using autoencoder

The ability of autoencoder to detect adulteration of arabica coffee with chicory or robusta coffee at light, medium and dark roast levels was investigated.

Figure 4 presents autoencoder errors of training and prediction sets of light roasted samples. The limit (1.46 E-03 units^2) was set based on the highest autoencoders' MSE in the training set (Fig. 4a). In general, samples with MSE higher than the limit were considered abnormal, meaning they can be detected by autoencoder as samples different from pure arabica coffee. All samples adulterated with robusta coffee (Fig. 4b) and chicory (Fig. 4c) at different concentrations were above the limit. It can therefore be concluded that an autoencoder neural network was able

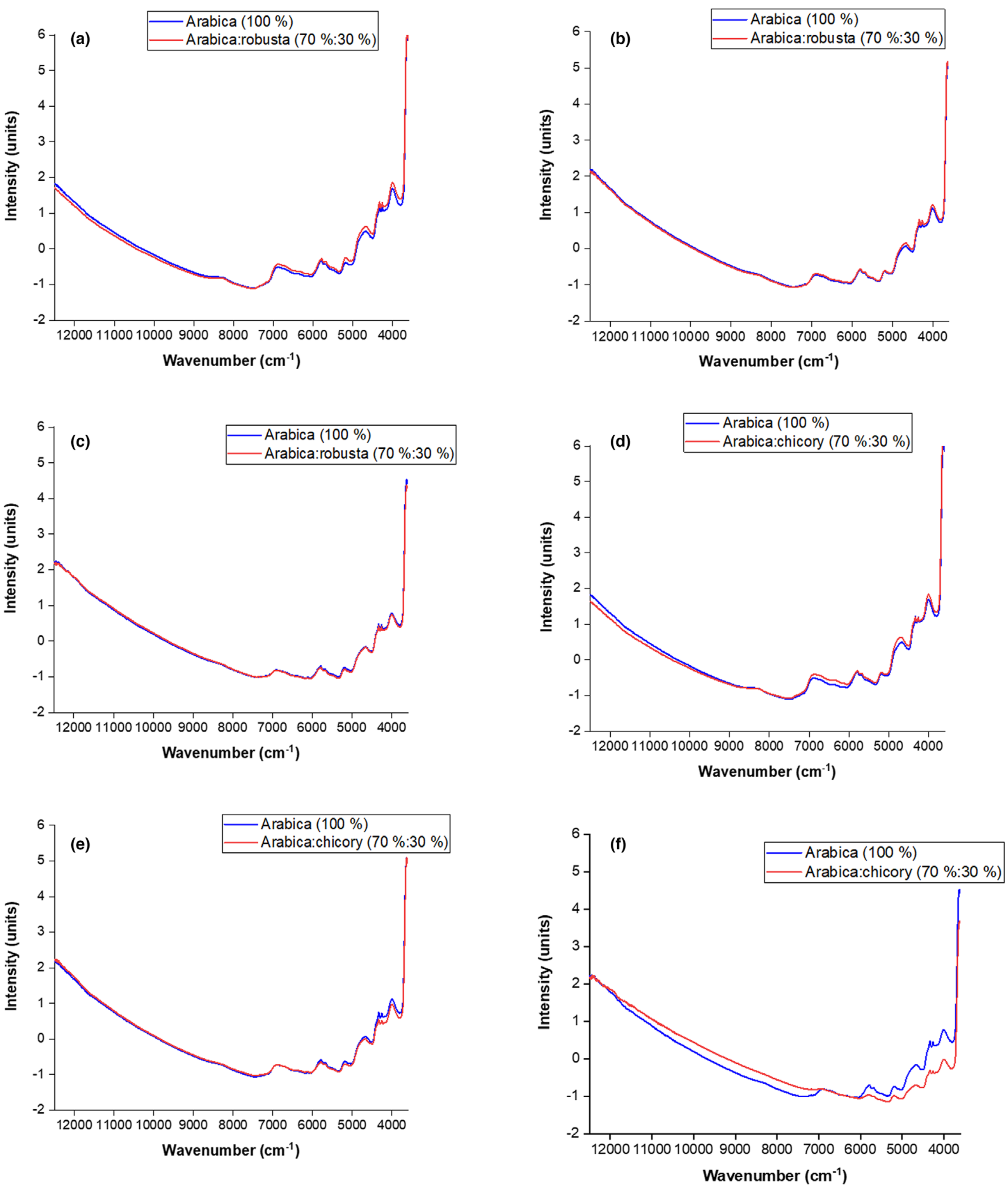

Figure 2 Average SNV transformed spectra of pure arabica coffee and that adulterated with either 30% robusta or 30% chicory. Spectra 2a–c are for arabica coffee adulterated with robusta coffee at light, medium and dark roasts, respectively, while spectra 2d–f are for arabica coffee adulterated with chicory at light, medium and dark roasts, respectively.

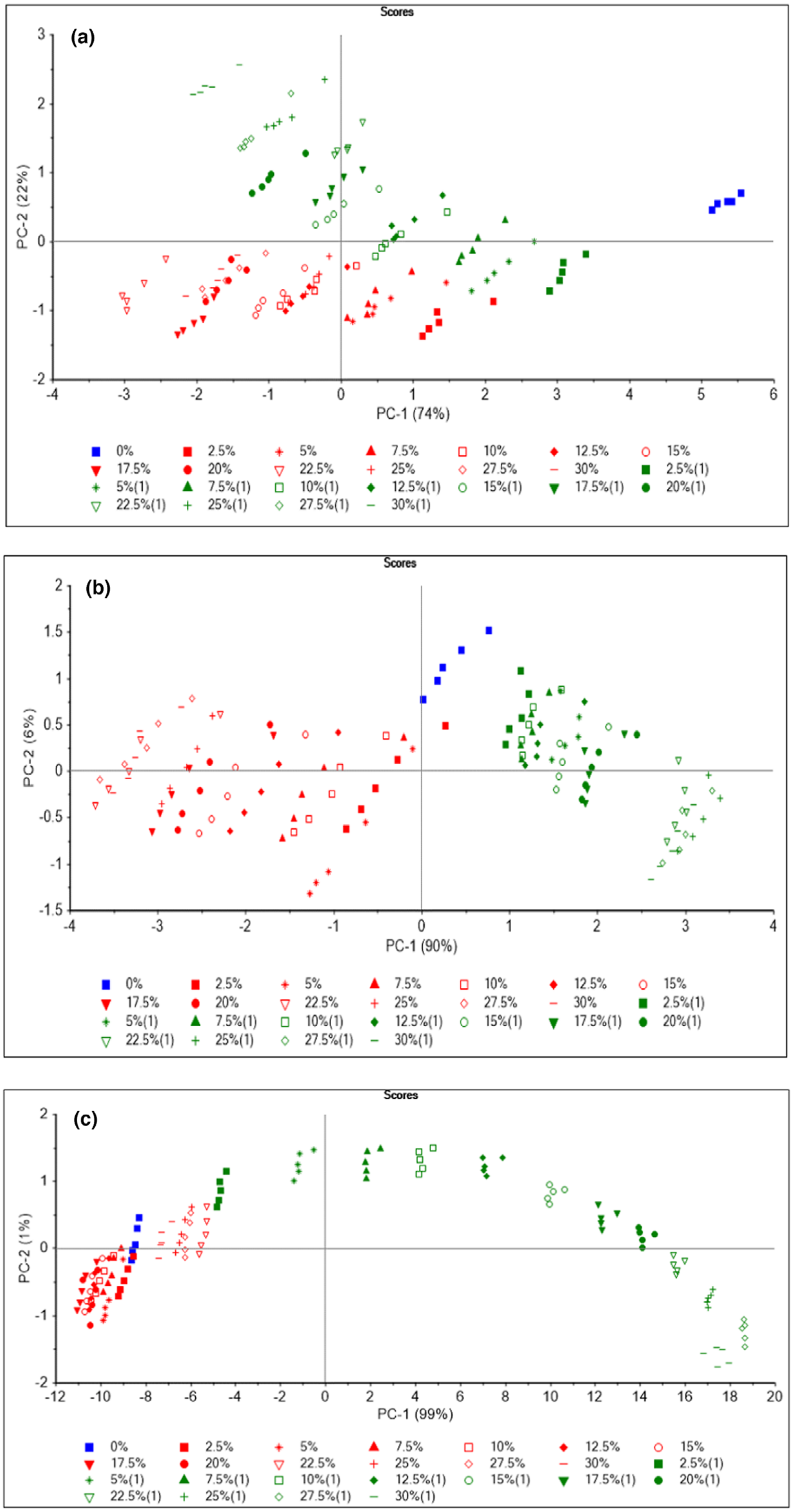

Figure 3 PCA score plots of pure and adulterated arabica coffee at levels ranging from 2.5% to 30%. Plots. 3a–c are for light, medium and dark roasts, respectively. Blue colour represents pure arabica; red and green are for arabica coffee adulterated with robusta and chicory, respectively.

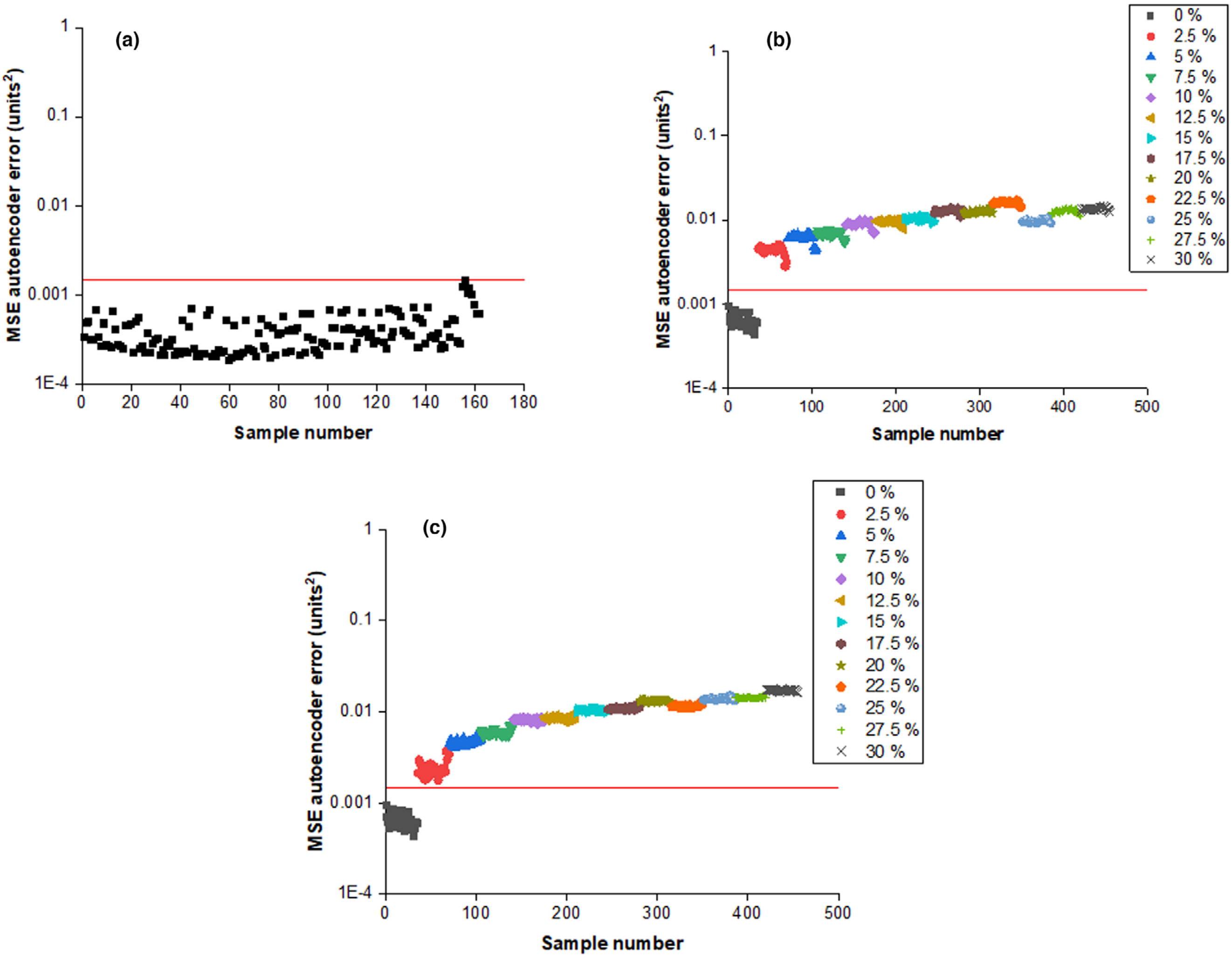

Figure 4 Autoencoders' MSE of light roasted samples. (a) Training set (n = 162) i.e. pure arabica coffee, (b) and (c) prediction sets (n = 455) i.e. arabica coffee adulterated with robusta coffee and chicory, respectively. Limit (red line) is equal to the highest MSE of the training set.

to detect arabica coffee samples at light roasting adulterated with robusta coffee or chicory even at very low concentrations.

Adulteration of all light roasted samples were detectable. On the contrary, not all samples adulterated with robusta coffee at medium roast were detected (Fig. 5b). MSE of some replications at 2.5%, 5% and 7.5% adulterations levels were below 1.69 E-03 units^2, which was set as the limit from the training set. In samples adulterated with 2.5% robusta coffee, 94% of the replications were lower than the limit (Table 2). Thus, autoencoder could not detect arabica coffee adulterated with 2.5% robusta coffee. Similar results were observed with samples adulterated with 5% and 7.5% robusta coffee, with 43% and 80% of their replications being lower than the limit, respectively. On the other hand, adulteration of arabica coffee with chicory was detectable in all the samples (Fig. 5c and Table 3).

Dark roasted samples, showed similar findings as those of medium roasted samples (Fig. 6). The limit of detection was equal to 1.77 E-03 units^2. For arabica coffee adulterated with robusta coffee, some replications at 2.5%, 5% and 7.5% adulteration levels are below the limit (Fig. 6b). In the case of 2.5% and 5% levels, the percentage values lower than the limit are 80% and 89%, respectively, demonstrating the inability of autoencoder to detect them (Table 2). On the other hand, only 37% of replications at 7.5% were below the limit. The average of autoencoder errors at this level was 1.89 E-03 units^2 (Table 2), which was higher than the limit (1.77 E-03 units^2) since most of

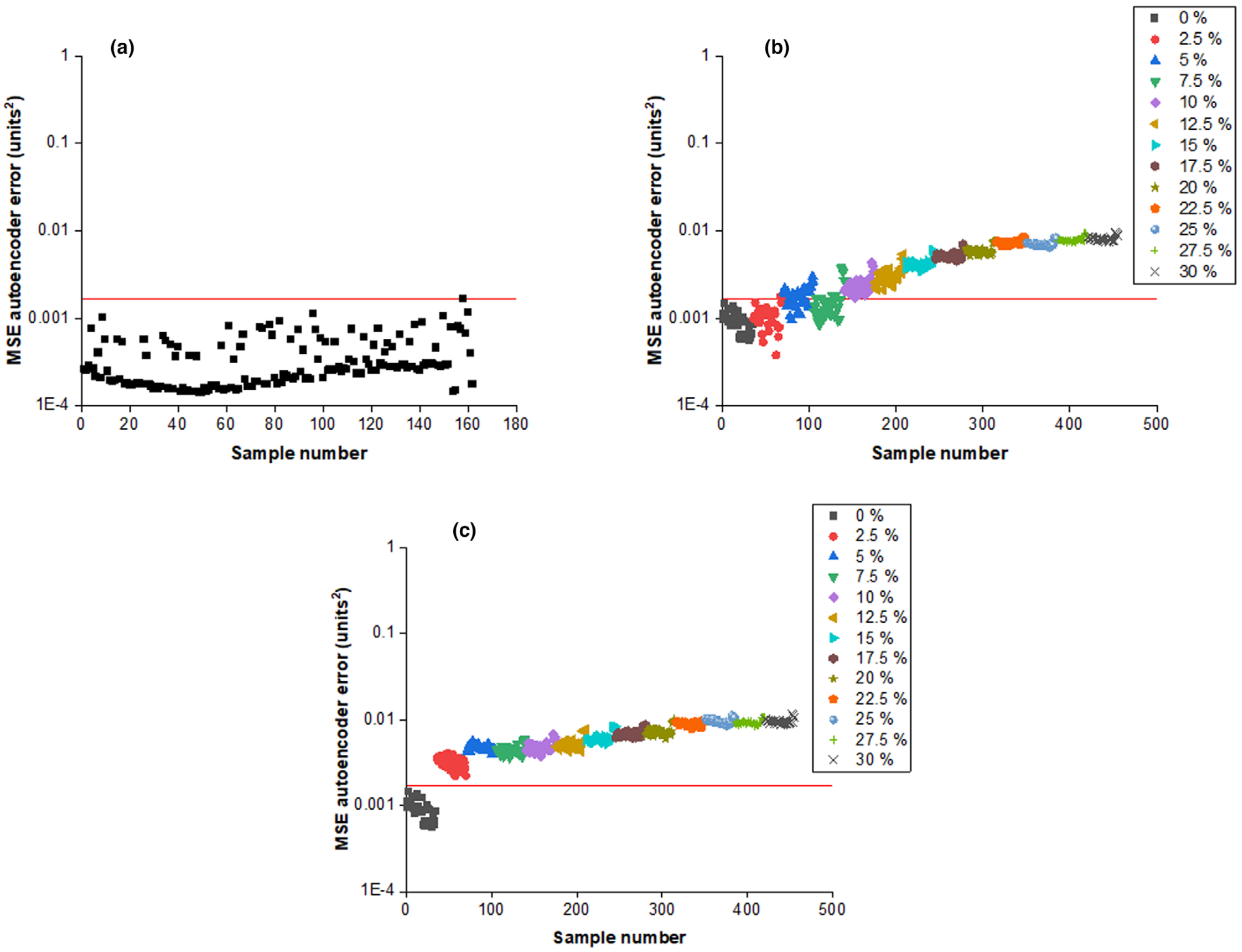

Figure 5 Autoencoders' MSE of medium roasted samples. (a) Training set (n = 162) i.e. pure arabica coffee, (b) and (c) prediction sets (n = 455) i.e. arabica coffee adulterated with robusta coffee and chicory, respectively. Limit (red line) is equal to the highest MSE of the training set.

the replications (63%) were above the limit (Fig. 6b). Just as for the medium roasted samples, adulteration of arabica coffee with chicory at all levels was detected (Fig. 6c).

It is important to note that at all roast levels, 0% adulterated arabica coffee was detected as not different from the one used in the training set. At all roast levels, arabica coffee adulterated with chicory was detectable. This could be explained by the change in the chemical composition of adulterated samples, particularly caffeine that is present in coffee and not in chicory (Nwafor *et al.*, 2017). As for the case of robusta coffee adulterated samples at medium and dark roasts, it was interesting to observe that low levels of adulteration, that is, 2.5%, 5% and 7.5% were not detectable. During coffee roasting, components in green beans are degraded through different chemical reactions to form other compounds. Vignoli *et al.* (2014) reported a decrease in the content of chlorogenic acids, trigonelline, furfural and hydroxymethylfurfural, and an increase in caffeine and melanoidins as roasting degree increased. Therefore, the inability of the autoencoder to detect low levels of adulteration may be attributed to the decrease of important compounds in robusta coffee due to the roasting effect. Couto *et al.* (2022) were able to detect arabica coffee adulterated with robusta from as low as 1% using NIR with PCA, though the information on the coffee roasting conditions are not provided so it is difficult to compare the sensitivity of the proposed method with the one in this study.

Table 2 Percentage of spectra whose autoencoders' MSE were below and above the limit plus MSE mean and standard deviation for samples in the prediction set (adulterated with robusta coffee)

	Light roast				Medium roast				Dark roast			
	Values below the limit (%)	Values above the limit (%)	Mean of MSE (units^2)	Standard deviation of MSE (units^2)	Values below the limit (%)	Values above the limit (%)	Mean of MSE (units^2)	Standard deviation of MSE (units^2)	Values below the limit (%)	Values above the limit (%)	Mean of MSE (units^2)	Standard deviation of MSE (units^2)
Adulteration level (%)												
0	100	0	6.29 E-04	1.17 E-04	100	0	8.97 E-04	2.51 E-04	100	0	8.61 E-04	2.12 E-04
2.5	0	100	4.37 E-03	5.39 E-04	94	6	1.07 E-03	3.15 E-04	80	20	1.46 E-03	4.33 E-04
5	0	100	6.12 E-03	6.83 E-04	43	57	1.77 E-03	4.63 E-04	89	11	1.49 E-03	3.62 E-04
7.5	0	100	6.91 E-03	5.93 E-04	80	20	1.51 E-03	7.37 E-04	37	63	1.89 E-03	3.76 E-04
10	0	100	8.97 E-03	7.71 E-04	0	100	2.41 E-03	5.85 E-04	3	97	2.63 E-03	4.52 E-04
12.5	0	100	9.53 E-03	5.56 E-04	0	100	3.11 E-03	7.15 E-04	0	100	3.33 E-03	3.77 E-04
15	0	100	1.03 E-02	4.80 E-04	0	100	4.20 E-03	4.91 E-04	0	100	3.93 E-03	3.63 E-04
17.5	0	100	1.29 E-02	5.57 E-04	0	100	5.30 E-03	4.87 E-04	0	100	4.62 E-03	3.01 E-04
20	0	100	1.25 E-02	6.04 E-04	0	100	5.95 E-03	6.15 E-04	0	100	4.75 E-03	6.21 E-04
22.5	0	100	1.60 E-02	7.73 E-04	0	100	7.31 E-03	3.52 E-04	0	100	7.41 E-03	5.88 E-04
25	0	100	9.72 E-03	4.08 E-04	0	100	7.11 E-03	4.57 E-04	0	100	6.17 E-03	8.25 E-04
27.5	0	100	1.26 E-02	6.46 E-04	0	100	7.85 E-03	4.69 E-04	0	100	6.43 E-03	9.16 E-04
30	0	100	1.34 E-02	5.67 E-04	0	100	8.04 E-03	4.82 E-04	0	100	5.77 E-03	8.18 E-04

Table 3 Percentage of spectra whose autoencoders' MSE were below and above the limit plus MSE mean and standard deviation for samples in the prediction set (adulterated with chicory)

	Light roast				Medium roast				Dark roast			
	Values below the limit (%)	Values above the limit (%)	Mean of MSE (units^2)	Standard deviation of MSE (units^2)	Values below the limit (%)	Values above the limit (%)	Mean of MSE (units^2)	Standard deviation of MSE (units^2)	Values below the limit (%)	Values above the limit (%)	Mean of MSE (units^2)	Standard deviation of MSE (units^2)
Adulteration level (%)												
0	100	0	6.29 E-04	1.17 E-04	100	0	8.97 E-04	2.51 E-04	100	0	8.61 E-04	2.12 E-04
2.5	0	100	2.40 E-03	4.89 E-04	0	100	3.08 E-03	4.72 E-04	0	100	1.03 E-02	1.21 E-03
5	0	100	4.62 E-03	3.93 E-04	0	100	4.67 E-03	3.79 E-04	0	100	2.28 E-02	2.76 E-03
7.5	0	100	5.93 E-03	4.14 E-04	0	100	4.45 E-03	5.67 E-04	0	100	3.77 E-02	2.72 E-03
10	0	100	8.12 E-03	2.57 E-04	0	100	4.78 E-03	6.62 E-04	0	100	5.19 E-02	3.28 E-03
12.5	0	100	1.04 E-02	3.34 E-04	0	100	5.31 E-03	8.79 E-04	0	100	7.19 E-02	4.03 E-03
15	0	100	1.09 E-02	2.09 E-04	0	100	6.12 E-03	8.29 E-04	0	100	9.48 E-02	3.38 E-03
17.5	0	100	1.31 E-02	2.86 E-04	0	100	6.86 E-03	6.79 E-04	0	100	1.15 E-01	4.24 E-03
20	0	100	1.16 E-02	2.94 E-04	0	100	7.31 E-03	8.89 E-04	0	100	1.32 E-01	4.20 E-03
22.5	0	100	1.16 E-02	4.14 E-04	0	100	9.01 E-03	5.05 E-04	0	100	1.47 E-01	3.25 E-03
25	0	100	1.39 E-02	4.08 E-04	0	100	9.57 E-03	6.20 E-04	0	100	1.63 E-01	2.12 E-03
27.5	0	100	1.42 E-02	2.65 E-04	0	100	9.18 E-03	5.74 E-04	0	100	1.81 E-01	2.30 E-03
30	0	100	1.70 E-02	3.46 E-04	0	100	9.50 E-03	5.95 E-04	0	100	1.67 E-01	5.71 E-03

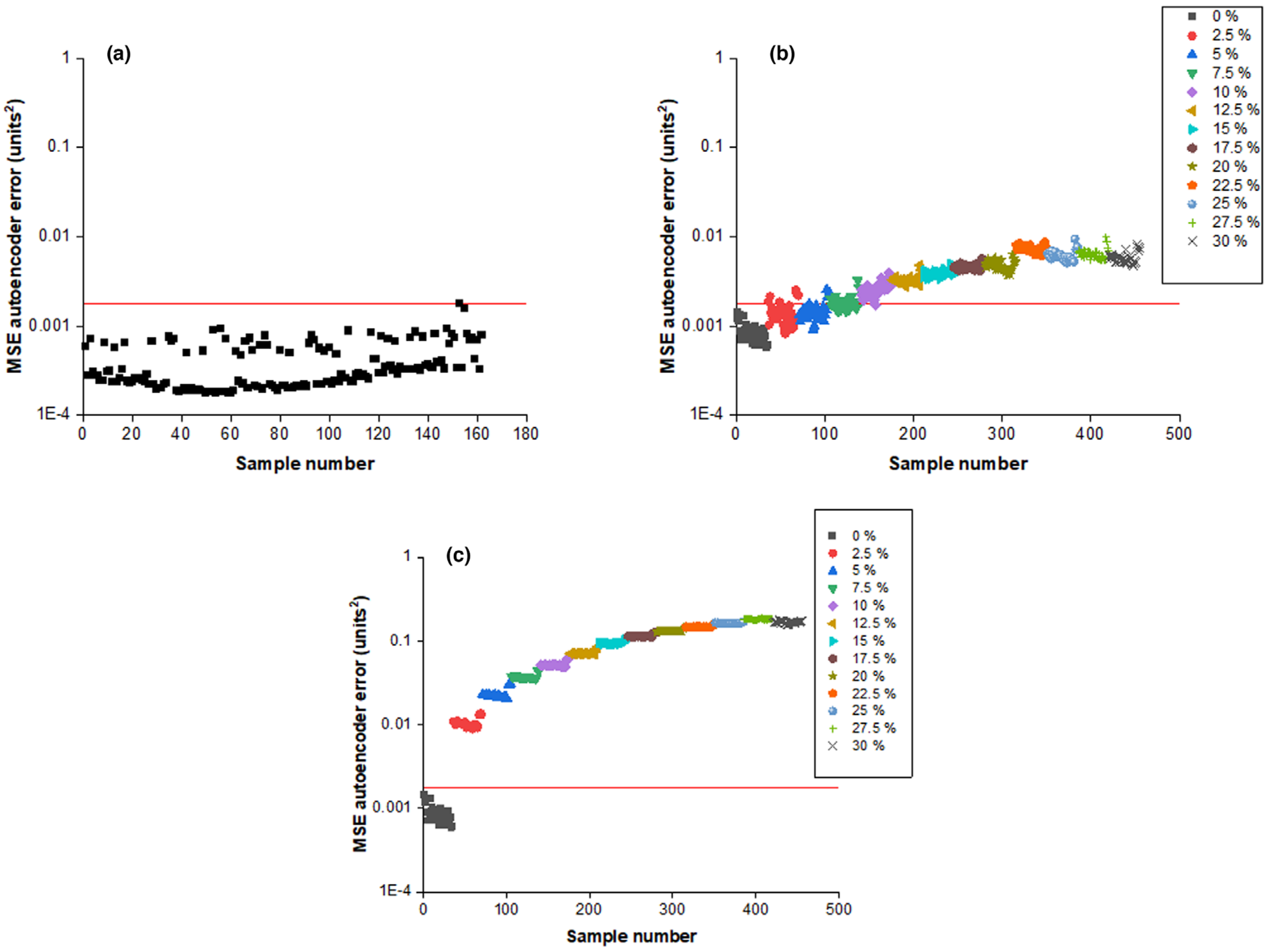

Figure 6 Autoencoders' MSE of dark roasted samples. (a) Training set (n = 162) i.e. pure arabica coffee, (b) and (c) prediction sets (n = 455) i.e. arabica coffee adulterated with robusta coffee and chicory, respectively. Limit (red line) is equal to the highest MSE of the training set.

Geographical origin

With the aim to investigate the ability of autoencoder to detect differences in geographical origin of roasted coffee, a coffee blend from Central and South America was used to train the autoencoder (training set). The prediction set consisted of coffee from Kenya, Guatemala and Colombia. The highest autoencoders' MSE of the samples in the training set was established as the limit and thus, samples in the prediction set having errors higher than the limit were from a different geographic origin (Table 4). For light and dark roasts, the limit was 1.46 E-03 units^2 and 1.77 E-03 units^2, respectively. Coffee samples from Kenya, Guatemala and Colombia were above the limit (Fig. 7d, f) for both light and dark roasting. The average MSE of these samples were all higher than the limit (Table 4). Therefore, the autoencoder was able to differentiate the samples based on their country of origin given the samples used in the training set were a coffee blend from Central and South America.

Different results were observed for medium roasted samples. The limit from the training set was 1.69 E-03 units^2. Only samples from Guatemala and Colombia were clearly detectable by the autoencoder since all the replications were above the limit (Table 4). For samples from Kenya, 14% of the replications were below the limit (Table 4). Nevertheless, the average of the errors was 1.87 E-03 units^2, higher than the limit, with 86% of the replications being so. Since during coffee roasting some compounds are degraded and others formed (Vignoli *et al.*, 2014), it is possible that at a medium roast, chemical components that were important in differentiating samples from Kenya and a blend of Central and South America were similar. Thus, the autoencoder detected the samples as not

Table 4 Percentage of spectra whose autoencoders' MSE were below and above the limit plus MSE mean and standard deviation for samples in the prediction set (coffee from different geographical origin)

	Light roast				Medium roast				Dark roast			
	Values below the limit (%)	Values above the limit (%)	Mean of MSE (units^2)	Standard deviation of MSE (units^2)	Values below the limit (%)	Values above the limit (%)	Mean of MSE (units^2)	Standard deviation of MSE (units^2)	Values below the limit (%)	Values above the limit (%)	Mean of MSE (units^2)	Standard deviation of MSE (units^2)
Geographic origin												
Central and South_America blend	100	0	6.29 E-04	1.17 E-04	100	0	8.97 E-04	2.51 E-04	100	0	8.61 E-04	2.12 E-04
Kenya	0	100	3.09 E-03	1.88 E-04	14	86	1.87 E-03	2.05 E-04	0	100	3.94 E-03	4.36 E-04
Guatemala	0	100	2.77 E-02	1.15 E-03	0	100	1.10 E-02	4.81 E-04	0	100	1.08 E-02	4.87 E-04
Colombia	0	100	5.97 E-03	4.88 E-04	0	100	5.07 E-03	3.22 E-04	0	100	4.66 E-03	6.23 E-04

different from each other even though they were from different geographical origin.

The geographical discrimination of coffee was mainly possible because of the differences in the chemical composition among the samples. Important factors that affect chemical composition of coffee includes soil, quantity of rainfall, species, temperature, agricultural practices, and altitude, which may differ from one country to another (Bilge, 2020; Zhu *et al.*, 2021).

At all roast levels, a coffee blend from Central and South America (not included in the training set) was part of the prediction set to check the efficiency of the autoencoder. All replications of this coffee blend were below the limit at all roast levels (Fig. 7d–f). This demonstrates that the autoencoder was able to detect that the coffee was not different from the one used for training. Other chemometric techniques in conjunction with NIR spectroscopy have shown their ability to differentiate coffee from different geographical origin (Minh *et al.*, 2022). Authentication of a coffee's geographical origin is becoming gradually of interest considering its influence on the sensory properties of the beverage and ultimately its price. Usually, skilled tasters whose responses are subjective evaluate the origin of coffee. Additionally, it is a challenge for one taster to reliably identify a large number of coffee origins (Baqueta *et al.*, 2019). Therefore, the autoencoder presents a fast, non-subjective and simple technique for authenticating the origin of roasted coffee.

Conclusion

The authenticity of coffee regarding its safety and geographical origin is important for consumers, processors and traders. As a new method of machine learning with potential applications in the food industry, deep autoencoder represents a powerful tool to detect adulterants in coffee, as well as differentiate coffees from different geographical origins. Arabica coffee adulterated with as low as 2.5% chicory at all roasts were detectable. For robusta adulterated samples, detection was possible at adulteration levels above 7.5% at medium and dark roasts. Additionally, it was possible to differentiate coffee samples from different geographical origins. PCA proved to be a suitable chemometric model for the visualisation of data, where samples were grouped based on the type and concentration of the adulterant. In general, autoencoder proved to be feasible in detecting adulterants in coffee as well as for the geographical origin of roasted arabica coffee. Therefore, this method could be adopted in the coffee industry as a quality control tool to verify the authenticity of their products. In this research, only two adulterants were investigated and thus it would be of interest to examine the capability of autoencoder to

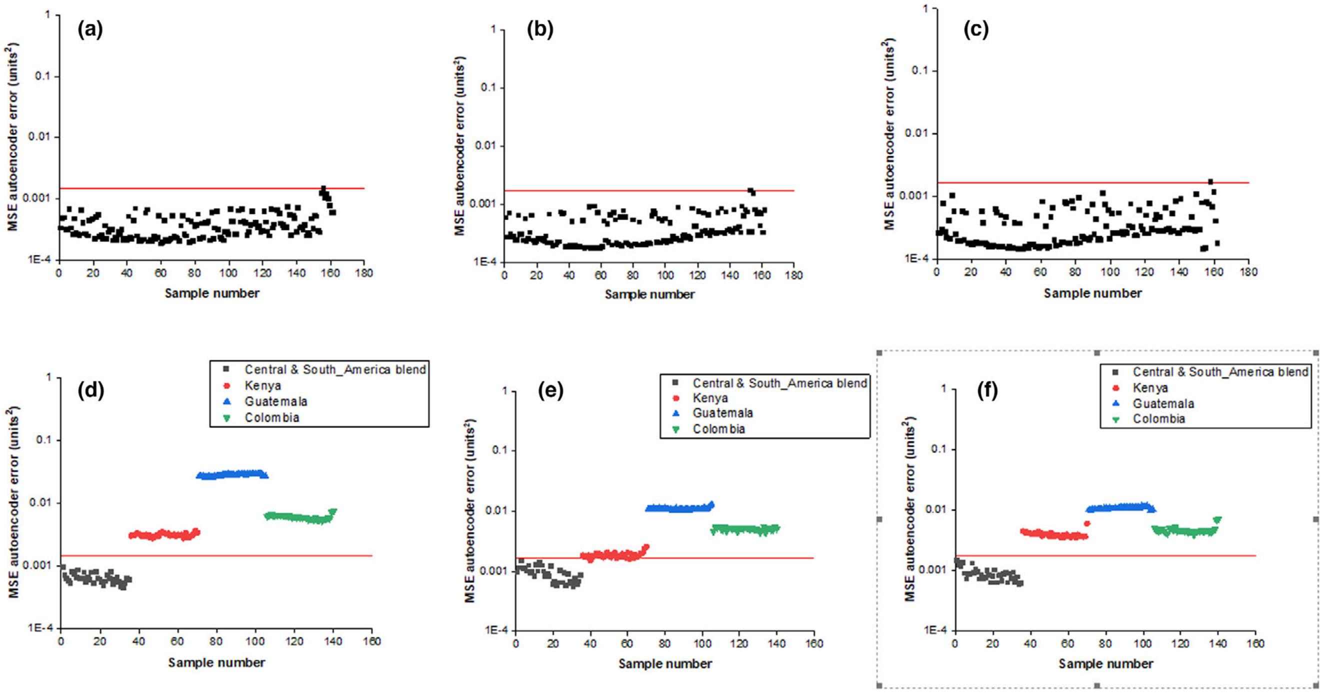

Figure 7 Autoencoders' MSE of coffee samples from different geographical origins and roast levels. (a–c) Training sets (n = 162) at light, medium and dark roast levels, respectively. (d–f) Prediction sets (n = 140) at light, medium and dark roast levels, respectively. Limit (red line) is equal to the highest MSE of the training set.

detect other possible adulterants and mixtures of different adulterants in coffee samples.

Acknowledgment

We thank Olivier Paquet-Durand and Majharul Islam Babor for their assistance in developing autoencoder neural network and data analysis. Open Access funding enabled and organized by Projekt DEAL.

Author contributions

Leah Masakhwe Munyendo: Formal analysis (lead); funding acquisition (lead); investigation (lead); methodology (equal); writing – original draft (lead); writing – review and editing (equal). **Daniel Njoroge:** Supervision (equal); validation (equal); writing – review and editing (equal). **Yanyan Zhang:** Supervision (equal); validation (equal); writing – review and editing (equal). **Bernd Hitzmann:** Conceptualization (lead); methodology (equal); supervision (lead); validation (lead); writing – original draft (equal); writing – review and editing (equal).

Conflict of interest

The authors declare no conflict of interest.

Funding information

National Research Fund of Kenya and the German Academic Exchange Service (DAAD), funding ID 57524989, funded this research.

Data availability statement

Research data are not shared.

References

Baqueta, M.R., Coqueiro, A. & Valderrama, P. (2019). Brazilian coffee blends: A simple and fast method by near-infrared spectroscopy for the determination of the sensory attributes elicited in professional coffee cupping. *Journal of Food Science*, **84**, 1247–1255.

Barbosa, M.S.G., Scholz, M.B.S., Kitzberger, C.S.G. & Benassi, M.T. (2019). Correlation between the composition of green arabica coffee beans and the sensory quality of coffee brews. *Food Chemistry*, **292**, 275–280.

Barjolle, D., Quiñones-Ruiz, X.F., Bagal, M. & Comoé, H. (2017). The role of the state for geographical indications of coffee: case studies from Colombia and Kenya. *World Development*, **98**, 105–119.

Bilge, G. (2020). Investigating the effects of geographical origin, roasting degree, particle size and brewing method on the physicochemical and spectral properties of arabica coffee by PCA analysis. *Food Science and Technology*, **57**, 3345–3354.

Burns, D.T. & Walker, M. (2020). Critical review of analytical and bioanalytical verification of the authenticity of coffee. *Journal of AOAC International*, **103**, 283–294.

Chakravartula, S.S.N., Moscetti, R., Bedini, G., Nardella, M. & Massantini, R. (2022). Use of convolutional neural network (CNN) combined with FT-NIR spectroscopy to predict food adulteration: A case study on coffee. *Food Control*, **135**, 108816.

Cheng, Z., Wang, S., Zhang, P., Wang, S., Liu, X. & Zhu, E. (2021). Improved autoencoder for unsupervised anomaly detection. *International Journal of Intelligent Systems*, **36**, 7103–7125.

Correia, R.M., Tosato, F., Domingos, E. *et al.* (2018). Portable near infrared spectroscopy applied to quality control of Brazilian coffee. *Talanta*, **176**, 59–68.

Couto, C.C., Freitas-Silva, O., Oliveira, E.M.M., Sousa, C. & Casal, S. (2022). Near-infrared spectroscopy applied to the detection of multiple adulterants in roasted and ground arabica coffee. *Foods*, **11**, 61.

Couto, C.C., Santos, T.F., Mamede, A.M.G.N. *et al.* (2019). Coffea arabica and C. canephora discrimination in roasted and ground coffee from reference material candidates by real-time PCR. *Food Research International*, **115**, 227–233.

Ebrahimi-Najafabadi, H., Leardi, R., Oliveri, P., Casolino, M.C., Jalali-Heravi, M. & Lanteri, S. (2012). Detection of addition of barley to coffee using near infrared spectroscopy and chemometric techniques. *Talanta*, **99**, 175–179.

European Commission. (2018). Food fraud summary, May 2018. Online refernce included in article URL https://knowledge4policy.ec.europa.eu/publication/food-fraud-summary-may-2018_en. Accessed 26/5/2022

It shows the problem of coffee adulteration on the market. It is important for my research as it illustrates the need of developing simple methods for quality assessment.

Ferreira, T., Farah, A., Oliveira, T.C., Lima, I.S., Vitório, F. & Oliveira, E.M.M. (2016). Using real-time PCR as a tool for monitoring the authenticityof commercial coffees. *Food Chemistry*, **199**, 433–438.

Ferreira, T., Shuler, J., Guimarães, R. & Farah, A. (2019). Introduction to coffee plant and genetics. In: *Coffee: Production, Quality and Chemistry*. Pp. 1–25. London: Royal Society of Chemistry.

Finke, T., Krämer, M., Morandini, A., Mück, A. & Oleksiyuk, I. (2021). Autoencoders for unsupervised anomaly detection in high energy physics. *Journal of High Energy Physics*, **2021**, 161.

Flores-Valdez, M., Meza-Márquez, O.G., Osorio-Revilla, G. & Gallardo-Velázquez, T. (2020). Identification and quantification of adulterants in coffee (*Coffea arabica* L.) using FT-MIR spectroscopy coupled with chemometrics. *Foods*, **9**, 851.

Forchetti, D.A.P. & Poppi, R.J. (2020). Detection and quantification of adulterants in roasted and ground coffee by NIR hyperspectral imaging and multivariate curve resolution. *Food Analytical Methods*, **13**, 44–49.

Gandra, F.P.P., Lima, A.R., Ferreira, E.B., Pereira, M.C.A. & Pereira, R.G.F.A. (2017). Adding adulterants to coffee reduces bioactive compound levels and antioxidant activity. *Journal of Food and Nutrition Research*, **5**, 313–319.

Harohally, N.V. & Thomas, C. (2021). Quick NIR based method for ascertaining coffee and chicory percentage in a mixture. *ACS Food Science & Technology*, **1**, 524–528.

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K. & Davis, L.S. (2016). Learning temporal regularity in video sequences. In: *IEEE Conference on Computer Vision and Pattern Recognition*, Pp. 733–742. Piscataway, USA: Institute of Electrical and Electronics Engineers (IEEE).

ICO. (2021). Online reference inlcuded in the artcle. http://www.ico.org/trade_statistics.asp?section=Statistics. Accessed 2/06/2022

Martins, V.D.C., Godoy, R.L.D.O., Gouvêa, A.C.M.S. *et al.* (2018). Fraud investigation in commercial coffee by chromatography. *Food Quality and Safety*, **22**, 121–133.

Medina, J., Caro Rodríguez, D., Arana, V.A., Bernal, A., Esseiva, P. & Wist, J. (2017). Comparison of attenuated Total reflectance Mid-Infrared, near infrared, and 1H-nuclear magnetic resonance spectroscopies for the determination of Coffee's geographical origin. *International Journal of Analytical Chemistry*, **2017**, 1–8.

Mendes, E. & Duarte, N. (2021). Mid-infrared spectroscopy as a valuable tool to tackle food analysis: A literature review on coffee, dairies, honey, olive oil and wine. *Foods*, **10**, 477.

Minh, Q.N., Lai, Q.D., Minh, H.N. *et al.* (2022). Authenticity green coffee bean species and geographical origin using near-infrared spectroscopy combined with chemometrics. *International Journal of Food Science and Technology*, **57**, 4507–4517.

Munyendo, L., Njoroge, D. & Hitzmann, B. (2021). The potential of spectroscopic techniques in coffee analysis—A review. *Processes*, **10**, 71.

Núñez, N., Collado, X., Martínez, C., Saurina, J. & Núñez, O. (2020). Authentication of the origin, variety and roasting degree of coffee samples by non-targeted HPLC-UV fingerprinting and chemometrics. Application to the detection and quantitation of adulterated coffee samples. *Food*, **9**, 378.

Núñez, N., Saurina, J. & Núñez, O. (2021). Authenticity assessment and fraud quantitation of coffee adulterated with chicory, barley, and flours by untargeted HPLC-UV-FLD fingerprinting and chemometrics. *Food*, **10**, 840.

Nwafor, I.C., Shale, K. & Achilonu, M.C. (2017). Chemical composition and nutritive benefits of chicory (*Cichorium intybus*) as an ideal complementary and/or alternative livestock feed supplement. *The Scientific World Journal*, **3**, 1–11.

Peixoto, P. (2009). Fraudes atingem 25% das marcas de cafe, diz associacao. Folha de Sao Paulo, April 28, 2009.

Rahman, M.Z.F.B.A., Chong, H.W. & Lim, V. (2018). UV-visible chemometrics approach for the determination of selected adulterants in claimed premixed coffee. *Malaysian Journal of Medicine and Health Sciences*, **14**, 147–152.

Rinnan, A., Berg, F. & Engelsen, S.B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry*, **28**, 1201–1222.

Santos, J.R., Viegas, O., Páscoa, R.N.M.J., Ferreira, I.M.P.L.V.O., Rangel, A.O.S.S. & Lopes, J.A. (2016). In-line monitoring of the coffee roasting process with near infrared spectroscopy: measurement of sucrose and colour. *Food Chemistry*, **208**, 103–110.

Šeremet, D., Fabečić, P., Cebin, A.V., Jarić, A.M., Pudić, R. & Komes, D. (2022). Antioxidant and sensory assessment of innovative coffee blends of reduced caffeine content. *Molecules*, **27**, 448.

Song, H.Y., Jang, H.W., Debnath, T. & Lee, K.G. (2019). Analytical method to detect adulteration of ground roasted coffee. *International Journal of Food Science and Technology*, **54**, 256–262.

Shows the ability of chromatographic method to detect coffee adulteration. It is significant for comparison with an autoencoder in terms of time needed to carry out, cost, labour etc.

Souto, U.T.D.C.P., Barbosa, M.F., Dantas, H.V. *et al.* (2015). Identification of adulteration in ground roasted coffees using UV–vis spectroscopy and SPA-LDA. *LWT - Food Science and Technology*, **63**, 1037–1041.

Tfouni, S.A.V., Serrate, C.S., Carreiro, L.B. *et al.* (2012). Effect of roasting on chlorogenic acids, caffeine and polycyclic aromatic hydrocarbons levels in two Coffea cultivars: Coffea arabica cv. Catuaı ’ Amarelo IAC-62 and Coffea canephora cv. Apoata˜ IAC-2258. *International Journal of Food Science and Technology*, **47**, 406–415.

Toci, A.T., Farah, A., Pezza, H.R. & Pezza, L. (2016). Coffee adulteration: more than two decades of research. *Critical Reviews in Analytical Chemistry*, **46**, 83–92.

Provides a summary of methods that have been used to detect adulteration. It is important as it highlights what can be improved.

Uncu, A.T. & Uncu, A. (2018). Plastid trnH-psbA intergenic spacer serves as a PCR-based marker to detect common grain adulterants of coffee (*Coffea arabica* L.). *Food Control*, **91**, 32–39.

Vasafi, P.S., Paquet-Durand, O., Brettschneider, K., Hinrichs, J. & Hitzmann, B. (2021). Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy. *Journal of Food Engineering*, **299**, 110510.

It is important as it shows the ability of an autoencoder to be used in food industry in detecting anomalie.

Vignoli, A., Viegas, C., Bassoli, D. & Benassi, M.T. (2014). Roasting process affects differently the bioactive compounds and the antioxidant activity of arabica and robusta coffees. *Food Research International*, **61**, 279–285.

Winkler-Moser, J.K., Singh, M., Rennick, K.A. *et al.* (2015). Detection of corn adulteration in Brazilian coffee (Coffea arabica) by tocopherol profiling and near-infrared (NIR) spectroscopy. *Journal of Agricultural and Food Chemistry*, **63**, 10662–10668.

Wongsaipun, S., Theanjumpol, P., Muenmanee, N., Boonyakiat, D., Funsueb, S. & Kittiwachana, S. (2021). Application of artificial neural network for tracing the geographical origins of coffee bean in northern areas of Thailand using near infrared spectroscopy. *Chiang Mai Journal of Science*, **48**, 163–175.

Zhu, M., Long, Y., Ma, Y. *et al.* (2021). Comparison of chemical and fatty acid composition of green coffee bean (*Coffea arabica* L.) from different geographical origins. *LWT - Food Science and Technology*, **40**, 110802.