Country of origin label monitoring of musky and common octopuses (Eledone spp. and Octopus vulgaris) by means of a portable near-infrared spectroscopic device

Country of origin label monitoring of musky and common octopuses (*Eledone spp.* and *Octopus vulgaris)* by means of a portable near-infrared spectroscopic device

Maria Olga Varrà, Sergio Ghidini, Maria Pia Fabrile, Adriana Ianieri, Emanuela Zanardi*

Department of Food and Drug, University of Parma, ~~Vi~~Strada del Taglio, 10, 43126 Parma, Italy

***CORRESPONDING AUTHOR:**

E-mail address: emanuela.zanardi@unipr.it (E. Zanardi)

Codice campo modificato

ABSTRACT

~~The recognized economic and nutritional value of cephalopods has recently led to a widespread capture fishery and distribution worldwide, thus increasing the possibility of fraudulent substitution of the products and posing into question the truthfulness of the geographic indications reported on the label.~~ Modern analytical techniques using miniaturized and portable near infrared (NIR) spectroscopy instruments are particularly suited for assessing the authenticity of fishery products since meeting the requirements of rapidity, eco-friendliness, cost-effectiveness, and easiness of application. The objective of the present study was to verify the suitability of use of a portable and ultra-compact NIR spectrometer combined with machine learning to characterize the geographic origin of two ~~widely consumed~~ octopus species. Replicate NIR spectra ~~in the~~ (908.1–1676.2 nm) ~~NIR region~~ of 118 musky and 29 common octopus specimens (*Eledone* spp. and *Octopus vulgaris*) from Portuguese Atlantic or Spanish Mediterranean fishing areas were recorded, pre-processed and elaborated via the following classification algorithms: orthogonal partial least square discriminant analysis (OPLS-DA), logistic regression (LR), random forest (RF), support vector machine (SVM), and multilayer perceptron-artificial neural network (MLP-ANN). When 7-fold cross validation was performed on 75% of data, the results showed that linear tools (OPLS-DA and LR) were the most powerful and stable techniques in recognizing the origin of both octopus species~~, with~~ (mean sensitivity, specificity, accuracy, and precision values above 98%) ~~and the lowest associated standard deviations.~~ During the external validation phase ~~(using 25% of the remaining spectral data)~~ OPLS-DA, SVM, and MLP-ANN performed better for common octopuses ~~(with no classification errors)~~, while LR and MLP-ANN for musky octopuses ~~(with only 2 and 3 Mediterranean samples misclassified, respectively).~~ The achieved outcomes suggest the combination of portable NIR spectroscopy and machine learning as a promising plan of action to be adopted for the creation of an integrated analytical platform with capabilities for automated data recording, processing, and reporting, which may be helpful for on-site and in-line monitoring of fishery products.

Abbreviations

first derivative, 1^{st}Der; second derivative, 2^{nd}Der; area under the receiver operating characteristic curves, AUROC; logistic regression, LR; multilayer perceptron artificial neural network, MLP-ANN; multiplicative scatter correction, MSC; near infrared, NIR; orthogonal partial least square-discriminant analysis, OPLS-DA; principal component analysis, PCA; principal component, PC; radial basis function, RBF; random forest, RF; root mean square error from cross-validation, RMSECV; standard normal variate, SNV; support vector machine, SVM;

Keywords: rapid methods; machine learning; chemometrics; food authenticity; geographical origin; cephalopods.

1. Introduction

Cephalopod mollusk species belonging to the Octopodidae and Eledonidae families (collectively known as octopuses) represent an important fishery resource, with high economic, social, cultural, and nutritional values, especially for Asiatic and Mediterranean countries. Common octopus (*Octopus vulgaris*) and horned and musky octopus (*Eledone spp.*) are the main octopus species produced and exploited for human consumption by Mediterranean and Central Eastern Atlantic countries. According to the most recent data, Italy, Spain, Portugal, Greece, and France together account for 7% and 95% of world and European production, respectively (European Market Observatory for Fisheries and Aquaculture Products (EUMOFA), 2020), being characterized by large- as well as small-scale artisanal octopus fisheries which are of extreme importance for local economy (Pita et al., 2021). Notwithstanding this, octopus catches by European Union countries have steadily declined over the past decade as a combined result of the effects of overfishing practices, new fisheries management policies focusing on sustainable practices, and climate change (Tinacci et al., 2020). On the other hand, to compensate the increased demand, large volumes of frozen octopus are now imported from third countries (Morocco and Mauritania mainly) (EUMOFA, 2020) and retailed by the local markets predominantly as thawed products. This market configuration increases the chances of species and fisheries suffering unfair competition, as well as illegal, unreported, and unregulated fishing activities being pursued, thus bringing about commercial frauds regarding the falsification of the geographic origin and traceability problems which, in turn, have important economic and sustainability repercussions (Fox et al., 2018). Indeed, fraudulent mislabeling of cephalopods and cephalopod-based products, occurring at any level of the supply chain, was reported more than two and three times frequently compared to that of crustaceans and fish (Guardone et al., 2017). Despite being economically motivated, mislabelling concerning the geographic origin of cephalopods may represent a safety risk for consumers due to the potential exposure to different contaminants and pollutants. Indeed, cephalopods originating from areas at risk of harmful algal bloom have been reported to accumulate several marine biotoxins, such as tetrodotoxin, saxitoxins, palitoxins, and domoic acid, thus possibly acting as toxin vectors to humans (Lopes et al., 2013; Whitelaw et al., 2019; Karlson et al., 2021). Public health consequences may also arise from the consumption of cephalopods from specific polluted fishing areas due to the presence of very high concentrations of heavy metals and persistent organic pollutants (Gomes et. al., 2013; Rjeibi et al., 2014; Roldán-Wong et al., 2018). Similarly, an actual food safety concern is related to the human exposure to microplastics through the consumption of seafood whose contamination can vary a lot among countries over the world (EFSA CONTAM Panel, 2016). Microplastics were in fact reported to be a vector for chemical contaminants (heavy metals, organochlorine pesticides, drug residues, polycyclic aromatic hydrocarbons, polychlorinated biphenyls, and polybrominated diphenyl ethers) (Brennecke et al., 2016; Camacho et al., 2019; De-la-Torre, 2020).

The prevention of food fraud should be based on a management system approach relying on two main supports: a proper vulnerability assessment system and the design and implementation of mitigation measures (Fox et al. 2018). That one is the framework within which the development of appropriate analytical methods, testing compliance of foods with their label descriptions, is set as a fundamental part of the management of food fraud incidents (Stadler et al., 2016). Recent attempts in literature and practice have suggested a few analytical laboratory methods as appropriate means to identify different types of frauds affecting cephalopod products. The issue of species mislabeling was addressed by using DNA barcoding (Guardone et al., 2017; Tatulli et al., 2020), while the replacement of fresh with frozen/thawed products was successfully identified through proteomics (Guglielmetti et al. 2018) or histological evaluation of tissues (Tinacci et al., 2020). Likewise, illicit water addition was uncovered by measuring electric conductivity and dielectric properties of samples (Mendes et al., 2018), as well as through the development of ad hoc fast 3D scanning methods (Han et al., 2020). Facing with the task of identifying the geographical origin of cephalopods is more challenging because of many pre-catch (e.g., seasonality, sizes) and post-catch (e.g., storage conditions) factors overlapping with the issue of interest and the lack of target measurable parameters providing the certainty of geographical authenticity (Esslinger et al., 2014; Varrà et al., 2021a). Nevertheless, considering the territorial nature, the small activity area, and the very small capability of metabolization (Oliveira et al. 2018; Arechavala-Lopez et al., 2019) cephalopod mollusks and, in particular, octopuses, appear to be good indicators of the seawater areas they inhabit. Based on this consideration, the use of comprehensive approaches relying on the combination of modern instrumental and advanced statistical methods, such as those based on near infrared (NIR) spectroscopy and machine learning, may represent a direct solution for the identification of the sources of origin of cephalopod stocks.

With the advances in instrumentatal technology, the latest years are witnessing a shift from the laboratory usage of stationary benchtop NIR equipments to miniature and portable devices for quality, safety, and authenticity testing of food of animal origin, including fish and seafood (Grassi et al., 2018; Cruz-Tirado et al., 2021; Dos Santos et al., 2020; Silva et al., 2020; Dos Santos Pereira et al., 2021; González-Mohino et al., 2020; Müller-Maatsch et al., 2021a; Pennisi et al., 2021; Yakes et al., 2021; Yu et al., 2020; Currò et al., 2022). Nevertheless, only one study proved the suitability of using portable NIR technology coupled with machine learning to monitor traceability of cuttlefish cephalopods (Currò et al., 2021). Portable instruments, besides being rugged, user-friendly, compact, ultra-light, and cheaper compared to traditional stationary instruments, allow direct analysis without sample processing and consumption, thus facilitating on-site or in-line analysis and globally minimizing times and costs associated with the analytical flow (Beć et al. 2021; McVey et al. 2021; Müller-Maatsch et al., 2021b). Moreover, methods based on the use of portable NIR devices meet the requirements and goals of 'White Analytical Chemistry', showing a sinergy between analytical, ecological, and practical attributes (Nowak et al., 2021). Indeed, the greenness of the approach (absence of waste generation and toxic solvents, low power consumption) is perfectly balanced both with analytical efficiency (accuracy, precision, sensitivity) and practical/economic efficiency (Nowak et al., 2021). Although these benefits ~~and~~ , the extensive research done, and the substantial progress in the development of more efficient computer algorithms, no reliable methods have~~s~~ been yet deployed in routine monitoring or accepted as an official standard since there is still the need to identify the proper machine learning algorithms able to speed up and simplify the analytical workflow and accurately ascertain the sought food authenticity features (Ríos-Reina et al., 2021).

Based on the above considerations, the goal of the present study is to propose a rapid, cheap, and eco-friendly analytical methodology to identify possible fraudulent mislabeling concerning the country of origin of different octopus species, which can be potentially useful for regulators, industry, and stakeholders, for the inspection and certification of cephalopods authenticity. To this end, a handheld, portable, and wireless NIR spectrometer was used to analyze two species of octopus originating from two different fishing areas (Spanish Mediterranean and Portuguese Atlantic), and the resulting spectral fingerprints were patterned by different traditional and modern machine learning tools in order to identify a fit-for-purpose methodology for their origin recognition.

2. Materials and Methods

2.1. Sample collection and handling

Three different batches of musky octopuses (*Eledone* spp., Cephalopoda: Octopodidae) of medium size (200–300 g total body weight, bw) and fished by means of otter trawls were collected during the autumn season from each of the two sampling sites chosen, corresponding to the FAO fishing areas 37.1.1 (i.e., Balearic waters of the Western Mediterranean Sea) and 27.9.a (i.e., Eastern Portuguese waters of the North-East Atlantic Ocean). Musky octopus from Mediterranean Sea and Atlantic Ocean accounted for 61 and 57 specimens, respectively. Similarly, three different batches of common octopuses (*Octopus vulgaris*, Cephalopoda: Octopodidae) of medium size (1300–1500 g bw) and caught by means of otter trawls in summer were retrieved from the FAO fishing area 37.1.1, accounting for a total of 29 specimens. Other 10 common octopus samples of the same size (1000–1500 g bw) and fishing season were collected from the FAO fishing areas 27.9.a. The sampling plan is graphically summarized in Figure 1.

All the samples were transported and delivered as fresh products and stored at freezing temperature (-21 ± 2°C) once arrived at the laboratory. Before analysis, the samples were defrosted in a refrigerator (4 ± 2°C per at least 18 hours) and the temperature gradually brought to room-conditions (20 ± 2°C) for 60–90 min. The octopuses were then manually skinned and dissected. Since the mantle was the flesh portion selected for the experiments, it was separated from the arms and left whole for subsequent spectral analysis.

2.2. MicroNIR setup and measurement

The octopus mantles were analyzed by using the ultracompact, portable and wireless NIR device MicroNIR OnSite-W (Viavi Solutions, Santa Rosa, CA) equipped with the spectral acquisition software MicroNIR Pro^{TM} (v.3.1, Viavi Solutions, Santa Rosa, CA, USA).

Diffuse reflectance NIR spectra were recorded in the 908.1–1676.2 nm region as 100 co-added scans and with an integration time of 10 ms. The spectral apparent resolution was 6.25 nm, hence each spectrum consisted of 125 reflectance points. NIR spectra were recorded by perpendicularly interfacing the acquisition window of NIR spectrometer with the surface of the sample. Spectral scanning was performed on four different points of the ventral and dorsal side of the mantle in order to collect heterogenicity of composition and thickness. The four replicate spectra were all individually used for statistical processing, thus resulting in two data matrices consisting of 472 spectra of musky octopus (118 samples × 4 spectral replicates) and 156 spectra of common octopus (39 specimens × 4 spectral replicates). Before NIR analysis and every 15 minutes during the analysis execution, the MicroNIR device was calibrated by recording a total absorbance (dark) reference spectrum (by leaving the lamps on and the acquisition window of the spectrometer empty) and a total reflectance reference spectrum (by using the external white diffuse reflectance standard disc Spectralon® 99%, LabSphere, North Sutton, NH, USA) to correct the background signal for the proper response over operation time and any little temperature changes.

2.3. Statistics and data modeling pipeline

The spectral data were exported, transformed to apparent absorbance values, mean-centered, and preprocessed by using standard normal variate (SNV), multiplicative scatter correction (MSC), first derivative (1st Der), and second derivative (2nd Der), alone or as combined spectral filters, to identify the best solution enhancing the signal-to-noise ratio and spectral resolution.

To achieve significant and robust results, each classification model was trained on 75% and validated on 25% of the common or musky octopus spectral pre-processed dataset. The low number of samples in the common octopus dataset hindered the possibility of averaging the four spectral replicates recorded for each sample. Therefore, it was decided to retain the four individual spectral replicates of each sample and to perform the 75:25 splits by randomly allocating them together into either the internal calibration sets (i.e., training sets) or the external validation sets (i.e., validation sets)~~.~~ as follows: i) 354 spectral data into the training set of musky octopuses (118 samples × 3 spectral replicates, i.e. 75% of total spectral replicates); ii) 118 spectral data into the validation set of musky octopuses (118 samples × 1 spectral replicate, i.e. 25% of total spectral replicates); iii) 156 spectral data into the training set of common octopuses (39 samples × 3 spectral replicates, i.e. 75% of total spectral replicates); iv) 39 spectral data into the training set of common octopuses (39 samples × 1 spectral replicate, i.e. 25% of total spectral replicates. Thereby, a balanced repartition of the two representative classes to be discriminated (i.e., the two geographical provenances of the samples), as well as the maximum independence among samples in the two sets were assured. In order to enhance statistical confidence with small sample sizes, training data were also 7-fold cross-validated (internal calibration phase), thus the models were trained on 6/7th and internally evaluated on 1/7th of the data. Considering that for small-sized datasets the whole performances of the machine learning models are influenced by the exact repartition of samples into the training and the ~~test~~ validation sets, the 75:25 splits were repeated 4 times for all the models. Hence, four different training and ~~testing~~ validation sets were generated for each octopus species. The final statistical outcomes resulting from training and validation stages were reported as means and standard deviations (Michelucci & Venturini, 2021).

After data repartition, the spectra of the training sets were used to create two principal component analysis (PCA) models (one for each of the octopus species considered). PCA was used for the sake of screening data structure, to reveal any potential hidden correlations among samples and variables, and to detect potential outliers, otherwise detrimental for the accuracy and stability of the subsequent machine learning classification models. Nevertheless, since used as a preliminary investigation tool, PCA was computed only once, i.e., by using data included in only one of the four training sets created. The following supervised classification tools were then tested: orthogonal partial least square discriminant analysis (OPLS-DA), logistic regression (LR), random forest (RF), support vector machine (SVM), and multilayer perceptron artificial neural network (MLP-ANN). Calculations of PCA and OPLS-DA were performed by using the statistical software packages SIMCA (v. 16.0.2, Sartorius Stedim Data Analytics AB, Umea, Sweden), while calculations of LR, RF, SVM, and MLP-ANN were done by using IBM SPSS Modeler software (v. 18.2, SPSS Inc., Chicago, IL, USA).

2.3.1 Training parameters of the machine learning models

In this work, the number of predictive and orthogonal new latent variables of the calibration OPLS-DA models was estimated by cross-validation. The fitting and prediction abilities of the models were evaluated by analyzing the following parameters: R^2X (cumulative variability of the spectral data modelled by all the extracted latent variables), R^2Y (cumulative variation associated to class labels explained by all the extracted latent variables) , Q^2X (cumulative variability associated to class labels predicted by all the extracted latent variables), RMSECV (root-mean-squared error of cross validation) and RMSEP (root-mean-squared error of prediction).

The overall significance of the logistic regression equation to classify octopus samples was indeed estimated by the likelihood ratio Chi-square test (using -2 times the log of the likelihood as reference value) and taking into consideration the Cox-Snell pseudo-R^2 regression value.

As for RF models, their structures were created using the Gini Impurity Index as a tree branching and variable selection criterion. According to the default settings suggested by the software, the number of trees to be generated was beforehand set to a maximum of 100. To avoid over splitting, the maximum depth of the tree structure was set to 10 levels while the minimum number of samples to be included into each child node was set to 5. Finally, the number of split variables for each tree node was set at 11 (square root of the total number of variables).

For the training of the SVM models, the Radial Basis Function (RBF) was used as kernel function. The balance between the model complexity and training error was established by setting the regularization parameter C (box-constraint or penalty factor) to 90, the additional kernel function γ parameter to 0.1, and the regression precision parameter ε to 0.1, by following the default settings suggested by the software.

Finally, the architecture of the MLP-ANN was built automatically and included an input layer (containing the NIR spectral data), one single hidden layer with one hidden neuron (transforming the weighted sum of the inputs by a hyperbolic tangent activation function to generate the outputs), and an output layer (using SoftMax activation function to estimate the probability of samples belonging to each classification group).

The relative contribution of each NIR wavelength to the predictive models was measured through different functions, based on the machine learning algorithm employed: while model-dependent methods consisting on the evaluation of the of Variable influence on projection (VIP) index, t-test statistics, and mean squared error were applied for OPLS-DA, LR, and RF, respectively, model-independent method based on the calculation of the area under the receiver operating characteristic (ROC) curve values (AUROC) were applied both for SVM and MPL-ANN.

2.3.2. Evaluation and comparison of the classification models performances

The estimation of the goodness of each classification model was performed on multiple fronts. Considering that each classification model is characterized by its own statistic outputs, standardized metrics providing a direct comparison of the performances of the models were chosen.

Firstly, the mean accuracy, specificity, sensitivity, and precision parameters (Fawcett, 2006) were calculated from the confusion matrices reporting percentages of common and musky octopus samples of the training sets correctly classified in the proper class during the cross-validation process. Prediction capabilities of the models were then graphically inspected through the ~~area under the~~

~~receiver operating characteristic (ROC) curve values (AUROC)~~AUROC values for each of the two classes (Mediterranean, Atlantic) of the validation sets, which is an optimal compromise to summarize sensitivity and specificity. ROC curves for validation data were created by plotting the true positive rate (TPR or sensitivity) versus the false positive rate (FPR or 1-specificity) at all predicted probability cut-off values (Fawcett, 2006).

3. Results and Discussion

3.1. NIR spectral characteristics and correction

Pre-processing of NIR spectra is quite a mandatory step in common practice to minimize the systematic variation in the spectra deriving from light scattering. Non-linearity and multiplicative effects deriving from this variation appear in the form of baseline shifts and drifts which are not directly related to the chemical properties, but rather to the structural features and physical status of the sample (Rinnan et al., 2009).

As it can be observed form Figure 2, light scattering effects were found in the raw absorbance spectra of both musky and common octopus samples recorded by MicroNIR. Therefore, ~~,~~ different pre-processing techniques were applied to the raw spectra by finding a compromise among drifts/shifts minimization, an acceptable peak separation degree, and the addition of unwanted noise. As a result, MSC and 2nd Der were discarded, while transformation by SNV followed by 1st Der (Norris-Williams, quadratic polynomial order, 15 points gap) was selected the best suited combination of spectral filters since allowed to re-align NIR spectra and partially suppress broad bands, without over processing and potential information loss (Figure 2). Despite intrinsic differences related to species, the average SNV plus 1st Der spectra of musky and common octopuses from the two geographical provenances were not characterized by rough visual differences in the absorbance pattern. The predominant bands were found in the 950–1000 nm, 1100–1200 nm, and in the 1300–1450 nm NIR regions (Figure 2). Nevertheless, considering that the original maximum peaks in the raw spectra correspond to the zero-crossing segment of the 1st Der spectra, the predominant individual features observables within the above mentioned NIR regions for common octopus dataset were at 1194 and 1440 nm, where $–CH_2$ and $–CH_3$ bonds of aliphatic hydrocarbons were reported to absorb (Workman & Weyer, 2012). In the case of musky octopus spectral dataset, the first feature moved to a lower wavelength (1186 nm), while the second one was located at the same wavelength (1440 nm) (Figure 2). ~~Due to difficulties in sampling procedures and in the interpretation of the NIR spectra, only a few research works have focused on the correlation between NIR absorbance spectra and the geographical origin of fish and seafood.~~ Mostly, NIR wavelengths related to lipid absorption and, sometimes, proteins, were already identified as useful for the classification of fish by origin (Ghidini et al., 2019; Currò et al., 2021; Varrà et al., 2021b). ~~Nevertheless, the assumption behind the possibility of fingerprinting the origin of cephalopods by using lipid and protein spectral features rely on the well-known link existing between the characteristics of the specific marine environment (water saline composition and average temperature, sediments, currents, seasonal temperature changes, population, and availability of fish preys) and the parallel variation of the chemical constituents of the fish tissues (Saito et al., 1997~~ In this context, octopus species, since particularly sensitive to any variation of the aquatic ecosystem, can provide useful information about the seawaters of origin~~, including the environment pollution status~~ (Sillero-Ríos et al., 2018), ~~Indeed,~~ especially ~~-~~due to ~~previous studies demonstrated that both~~ variations of the fatty acid and elemental profiles ~~vary a lot among *O. vulgaris* populations sampled in different areas and, therefore, they can be successfully used as markers of geographical origin~~ (Arechavala-Lopez et al., 2019; Semedo et al., 2012). ~~On the other hand, also pollutants or toxic elements on the fishing area may be similarly reflected into octopus tissues, leading to questioning food safety. This is one of the reasons why the authentication of seafood according to the geographic origin represent an important prerequisite of food safety (Freitas et al., 2020).~~

The possible differences in composition pointed out in the NIR spectra might explain the results achieved upon applying PCA. Specifically, the PCA models for musky and common octopuses were characterized by 8 and 7 principal components (PCs), covering 99.3 and 99% of the total variance, respectively. From the score scatter plots of the first three PCs (Figure 3), it can be sated that none of the octopus samples was suspected of being an outlier, since not crossing the 95% confidence limits for Hotelling's T^2 defined by plot ellipse. At the same time, the PC1 collected the inter-origin variability only among common octopuses, but not among musky octopuses, which did not separate efficiently each other.

Nevertheless, all the above aspects suggested a promising route for the application of supervised classification methods which could efficiently address the challenge of identifying the origin of octopuses by using NIR spectroscopy.

3.2. Analysis and comparison of machine learning models performances

~~In the field of food quality, safety and authenticity, there is a definite trend towards the automation and the use of smart fingerprint technologies, able to collect a huge amount of data to characterize foods and food systems in a comprehensive way, such as those based on miniaturized and portable spectroscopic sensors (Mcvey et al., 2021). This trend has been accompanied by a substantial progress in the development of more efficient computer algorithms and solutions, aimed to speed up and simplify the analytical workflow without sacrificing the reliability of the results. At the same time, coupling fingerprinting techniques with advanced computer-assisted data analysis offers the advantages of extending the domain of food applications thanks to the possibility of monitoring quality, safety, authenticity though one single analysis and, thus, preventing food fraud and food-borne illness.~~

In this work, five different powerful machine learning tools were tested against NIR spectroscopic data of musky and common octopus specimens of Mediterranean or Atlantic fishing origin included into the training sets, with the aim to develop a new tool for the prevention of potential frauds related to the falsification of the origin. The information embedded into the SNV + 1st Der pre-processed spectra was thus modelled by selecting OPLS-DA, LR, SVM, RF, and ANNs as supervised classification tools, which were initially tested on the training data by applying a cross-validation process (see *Section 2.3*).

The cross-validation results of the five tested machine learning tools are illustrated in Figure 4, while a summary of the modelling statistics is reported in *Supplementary Materials* (Tables S1, S2, S3, Figure S1). Each model was trained by considering all the spectral variables included into the dataset, corresponding to 125 NIR absorbance values (908.1–1676.2 nm spectra). The only exception was represented by RF models which automatically perform a feature selection to avoid overfitting. These models were built by using a total of 53 and 60 input variables for musky and common octopus classifications, respectively.

Contrary to what achieved when applying PCA (see *Section 3.1*) slightly better results were obtained when modelling musky octopuses compared to common octopuses, thus potentially confirming that the higher number of the analyzed samples still included in their NIR spectra a fraction of discriminant information related to the origin which was captured and described by the more powerful supervised algorithms. The performance metrics in cross validation were all above 96% except for RF models, which were characterized by the highest error rates (with approx. 6 and 16% of musky and common octopus wrongly recognized). The conventional linear algorithms were the most performant ones. Specifically, LR was found to be the best solution to recognize the origin labelling of common octopuses (average values of accuracy, specificity, sensitivity, and precision over 4 repetitions of 99.79 ± 0.43, 99.86 ± 0.29, 99.86 ± 0.29, and 99.60 ± 0.81%, respectively), while OPLS-DA showed the highest accuracy, specificity, sensitivity, and precision metrics for musky octopuses (average values over 4 repetitions of 99.58 ± 0.67, 99.59 ± 0.66, 99.59 ± 0.66, and 99.57 ± 0.68%, respectively) (Figure 4). In particular, the highest sensitivity (related to the true positive rates) and specificity (indicating the true negative rate) values shown by OPLS-DA and LR have both important positive consequences on the overall goodness of the discriminant methodology. In fact, sensitivity values indicate the degree of confidence in identifying the real authentic samples, thus having a direct impact on the economic side. On the other hand, specificity values indicate the degree of confidence in identifying the real non-authentic samples, with significant repercussions on the legal front.

As for computational performances of both MLP-ANN and SVM, also these tools showed very good predictions in cross-validation. SVM is one of the most frequently used machine leaning technique in food chemistry and authentication studies and, compared to linear classification methods, offers the advantages of adaptability towards the non-linear distribution of the data typical of NIR spectroscopy (Jiménez-Carvelo et al., 2019). The superiority of SVM methods over traditional linear classifiers combined with NIR spectral data in verifying different food authenticity claims has been demonstrated in several works (Cardoso & Poppi, 2021; Benes et al., 2020; Parastar et al., 2020; Sampaio et al., 2020; Bisutti et al., 2019). The suitability of using SVM to authenticate cephalopods (*Sepia officinalis*) has been also confirmed in a recent work, where its application yielded 83–100% balanced accuracy, 67–100% sensitivity, and 88–100% specificity for the classification of the samples according to 5 different geographical origins (Currò et al., 2021). It should be noted, however, that in the present work SVM, as well as RF, were characterized by very large standard deviations associated with all the metrics (Figure 4). This result might alert on the dependency of these techniques on the specific repartition of samples into training and validation sets, thus suggesting a potential instability and lack of robustness towards future prediction of unknown samples.

In conclusion, it can be stated that the data support the hypothesis that the simplest and the most interpretable classifiers (i.e., OPLS-DA and LR) also guarantee the best results in cross validation. The reason underlying this finding could be due to the existence of a direct linear rather than indirect correlation between NIR spectral patterns and the geographical origin of octopuses achieved by applying optimal spectral pre-processing operations. Therefore, although the complex nature of the samples, this correlation can be easily extrapolated by traditional linear techniques (Zareef et al., 2020). Evidence for this theory is however limited to the results obtained throughout the present research and it can be easily supposed that, with increasing sample size and non-linear variability (in terms of different fishing seasons, batches, sizes, storage times and temperatures), the more complex and flexible algorithms such as SVM, RF, and MLP-ANN might be the most performing ones.

Additionally, it is worth to say that the training of SVM and RF usually involves the identification of the best numerical values to be assigned to building parameters, so as to increase accuracy and performance of the final models. This operation is a very complex task requiring time, expertise, and efforts and, therefore, it does not sit well with the purpose of having a speedy, cost-effective, easy and fully exploitable procedure. In this work, the values for these building parameters (maximum number and depth of trees, minimum size and number of variables included into each child nodes for RF, as well as the *C, γ,* and *ε* parameters for SVM) were chosen by following the default settings recommended by the software and no complex operations such as manual search, use of genetic algorithms or grid search were performed (Phan et al., 2017).

3.2.1 Comparison of predictive NIR wavelengths

~~Given the different mathematical nature of the models being presented, also the relative strength of each NIR wavelength in guiding the classification of octopus samples based on the geographic provenance is expected be different. .~~For each predictive model computed, a different ranking of NIR wavelengths in terms of their importance (i.e., their relative contribution to the predictive models) was obtained since dependent on the mathematical function employed (see Section 2.3.1). Information about the first ten most influential NIR bands extracted as strong predictors of origin by individual cross-validated models obtained by training sets is provided in Table 1. Considering that the training phase was performed four consecutive times by changing sample datasets (*Section 2.3*), the kind and order of importance of the wavelengths were sometimes found to be different based on the dataset considered for modelling. For the sake of conciseness, those extracted by the most accurate of the four fitted models were reported.

NIR wavelength absorbance at 1453.2 nm was the most common important variable for musky octopuses, since it was extracted by four out five machine learning algorithms (OPLS-DA, RF, SVM, and MLP-ANN). Similarly, NIR band at 1632.8 had a strong influence on musky octopus samples discrimination for OPLS-DA, LR, and SVM. On the contrary, RF stood out the most from the other models because six out ten wavelengths (1081.5, 1075.3, 1465.6, 1137.3, 1341.7, 1093.9 nm) were exclusive predictors. As for common octopuses, absorbance peaks at 970 and 976.2 nm were shared as important predictors respectively by OPLS-DA, LR, RF, and SVM and by OPLS-DA, RF, SVM, and MLP-ANN, while that located at 963.8 by OPLS-DA, SVM, MLP-ANN (Table 1).

From the above results, it seems that NIR bands around the 1137.3 and 1341.7 nm, which are potentially related to aliphatic and aromatic hydrocarbons (Workman & Weyer, 2012), were strongly involved in differentiating Atlantic from Mediterranean musky octopuses, thus corroborating what emerged from the visual inspection of the pre-processed spectra discussed in *Section 3.1*. However, whereas OH-group absorption bands (1453 nm, 1075.3–1093.9 nm, and 963.8–976.2 nm) were clearly influent, the additional contribution of water to musky and octopus discrimination by origin should not be overlooked (Workman & Weyer, 2012). Considering that sample processing before NIR recording was standardized and freezing/defrosting as well as acquisition procedures performed at the same time/temperature conditions, it could be inferred that the water content varied across specimens to such an extent that inter-origin difference was higher than inter-individual one in both octopus species. Nevertheless, another hypothesis can be derived from the principles of aquaphotomics, according to which, since one single biological compound is solvated by many water molecules, the NIR response to individual biological compounds was amplified by water absorption which, indirectly, contributed to the achievement of high accuracy in prediction (Muncan & Tsenkova, 2019).

3.3. Independent evaluation of the predictivity of machine learning models

The ability of the ~~fitted~~ trained models to generalize beyond the training data and confidently assign the correct labels of geographical origin to future candidate octopus samples was estimated by using the ~~samples~~ 25% of data (i.e., one out four spectral replicates for each sample) belonging the validation sets~~)~~ and previously excluded from the calibration phase (see *Section 2.3*). The resulting label assignments for 118 ~~musky~~ and 39 spectral data of musky and common, respectively, included into ~~of~~ the validation sets are reported in the confusion matrices plotted in Figure 5. In agreement to what observed for PCA, but in contrast to results of cross-validation, the absolute best outcomes in external validation were achieved when recognizing the origin of common octopuses: three out five models (OPLS-DA, SVM, and MLP-ANN) predicted both the Mediterranean and Atlantic origin of common octopus validation ~~these~~ samples with 100% accuracy (Figure 5). More in detail, OPLS-DA, SVM, and MLP-ANN classifiers were all characterized by mean AUROC values of 1, with OPLS-DA and MLP-ANN also showing the lowest associated standard deviation values for the prediction of Mediterranean and Atlantic common octopus samples, respectively. On the contrary, despite the optimistic results in cross-validation, the analysis of the LR confusion matrix reported in Figure 5 revealed the poorest performances, since one Atlantic (10%) and nine Mediterranean common octopus validation samples (31%) were misclassified. The associated AUROC values were in fact 0.638 ± 0.002 and 0.631 ± 0.019 and, thus, quite close to the randomly guess rate of class membership recognition of 0.5. In this instance, it is important to reiterate that the ~~better~~ best classification of common octopuses compared to musky octopuses is likely to be the consequence of the smaller-sized group which determined the inclusion into the models of a smaller amount of variation available for self-learning which, in turn, hindered finding the best predictive correlation between NIR wavelengths and octopus origins. Hence, definite conclusions on this aspect could not be drawn, but it can be assumed that, although the noisy and collinear nature of the NIR spectra, none of the trained models underwent overfitting, as can be seen from the similarity between training and validation results.

As for musky octopuses, no models were able to recognize the provenance with 100% accuracy (Figure 5). The maximum correct classification rates in validation were shown by SVM and LR predicting the Mediterranean samples (98 and 97%, respectively) and by LR and MLP-ANN predicting the Atlantic ones (100%). Although SVM and MLP-ANN classifiers had high mean AUROC values, the lowest associated standard deviations were found for LR classifiers (AUROC= 0.989 ± 0.009 for Mediterranean samples; AUROC= 0.992 ± 0.009 for Atlantic samples). Shifting the focus from machine learning classifiers to single classes (i.e., geographical origins), it can be also noticed that, whatever the octopus species considered, samples from Atlantic Ocean were better recognized than samples from Mediterranean sea. From this finding it could be inferred that the small extension of the water surface and the close proximity to the coast of the Eastern Portuguese waters of the North-East Atlantic Ocean (FAO fishing area 27.9.a) are reflected in a more uniform environment which, in turn, might be responsible for the composition of octopus originating from this area to be more stable and preserved compared to that of octopuses from Western Mediterranean waters (FAO fishing areas 37.1.1). In fact, if on the one hand Western Mediterranean Sea is a semi-enclosed area characterized by stable temperature and salinity, the North-East Atlantic Ocean is characterized by a continuous input of organic matter from the Portuguese coast. This contributes to the permanent availability of prey and constant accessibility to food, which is the main factor influencing the fatty acid composition of octopuses (Massutí et al., 2004). Indeed, in the present work, the same lipid composition was hypothesized to be determinant in differentiating samples according to their country of origin (*Section 3.1*). As an example, concentrations of monounsaturated and n-6 polyunsaturated fatty acids were found to be constantly lower in Eastern Atlantic populations of *O. vulgaris* than Western Mediterranean ones, while concentrations of total fatty acids and n-3-6 polyunsaturated fatty acids higher in Atlantic populations (Arechavala-Lopez et al., 2019; Torrinha et al., 2014).

4. Conclusions

The results achieved from this study indicate that portable NIR sensors ~~a potential integrated analytical platform combining portable and miniature NIR spectroscopy and machine learning might be a suitable solution to~~ can identify with great accuracy the geographical origin of two ~~widely consumed~~ octopus species widely consumed in Europe,~~coming from Mediterranean or Atlantic fishing areas,~~ thus helping fraud prevention and having a direct impact on the quality and safety of the products. Regardless of the classification method employed, equally good results were achieved when fitting the models. Nevertheless, musky octopuses (*Eledone* spp.) were better modelled compared to common octopuses (*O. vulgaris*) when using traditional linear algorithms, (OPLS-DA and LR), thus suggesting the presence of a direct linear relationship between NIR spectra and the provenances of octopuses which can be easily extracted with increasing sample sizes. ~~Following the validation of the fitted models, OPLS-DA, SVM, and MLP-ANN allowed to achieve the maximum label recognition rates of common octopuses, while LR and MLP-ANN performed better for musky octopuses. Additionally, r~~Regardless of the octopus species considered, the origin label estimates were better for the Atlantic sample population compared to the Mediterranean one, probably because of the specific characteristics of the fishing waters, which contributed to make Atlantic population more homogenous from a compositional point of view.

In conclusion, the obtained data might be transferred to the fish chain environment and, provided their constant validation, find concrete application for the protection of the reputation of national and regional traditional fisheries. This application may materialize in the direct interface of a portable NIR system to the production flow and its customization based on the products to be handled and the specific industrial facility processes, so as to online monitor quality, authenticity, and safety of cephalopods and contribute to the development of quality certification schemes.

~~Given the promising outcomes, future research will be focused on the creation of multi-class classification models for the detection of commercial fraud, including additional fishing areas and, possibly, also different species of octopus. In this context, the most important impact under a future perspective would be the setting up of a tool to promote and protect the reputation of national and regional traditional fisheries and which would also help in the development of quality certification schemes.~~

~~If this is shown to be possible and satisfying results are achieved, then the next step might be the exploitation of the methodology for complementary applications addressing food safety and surveillance, such as the detection of contaminants, residues, and food additives. This way, a single integrated methodology might be used to characterize in a comprehensive way the fishery products found in the marketplace and ensure high quality and safety standards.~~

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.

Acknowledgements

The authors gratefully acknowledge Dr. Livio Artale (Proqualys S.r.l, Italy) for providing octopus samples.

Funding

This work was supported by the University of Parma (Italy).

~~**CRediT author statement**~~

~~**Maria Olga Varrà:** Conceptualization; Investigation, Formal analysis, Methodology, Writing – Original Draft; **Sergio Ghidini:** Conceptualization, Supervision, Methodology, Writing – Review & Editing; **Maria Pia Fabrile:** Formal analysis, Methodology; **Adriana Ianieri:** Funding acquisition, Project administration, Writing – Review & Editing; **Emanuela Zanardi:** Conceptualization, Supervision, Writing – Review & Editing; Project administration.~~

Appendix A. Supplementary materials

The following are the Supplementary data to this article:

Table 1

Comparison among the ten most important predictive NIR wavelengths used by each machine learning tool to categorize the octopus samples of the training sets by their origin.

Models	Important predictors (wavelengths, nm)*	
	Musky octopus	Common octopus
OPLS-DA	1453.2; 1459.4; 1447; 1632.8; 1440.8; 1564.7; 1558.5; 1626.6; 1570.9; 1434.6	982.4; 976.2; 988.6; 970; 963.8; 957.7; 951.5; 994.8; 1193; 1199.2
LR	1552.3; 1614.3; 1577.1; 1632.8; 1564.7; 1589.5; 1316.9; 1397.5; 1409.8; 1583.3	957.7; 951.5; 970; 1304.5; 1298.3; 982.4; 1316.9; 1248.8; 1236.4; 1224
RF	1453.2; 1081.5; 1459.4; 1323.1; 1075.3; 1465.6; 1137.3; 1341.7; 1447; 1093.9	970; 1180.7; 1304.5; 976.2; 1174.5; 1298.3; 1205.4; 1434.6; 1440.8; 1193.0
SVM	1632.8; 1626.6; 1620.5; 1614.3; 1608.1; 1453.2; 1601.9; 1595.7; 1589.5; 1583.3	1583.3; 1248.8; 963.8; 1316.9; 994.8; 970; 976.2; 1007.2; 982.4; 1180.7
MLP-ANN	1453.2; 982.4; 1100.1; 1069.2; 1620.5; 1106.3; 1205.4; 1007.2; 1001; 1186.8	1025.8; 1459.4; 976.2; 1019.6; 988.6; 963.8; 1242.6; 1211.6; 1236.4; 1069.2;

* For all the models, wavelengths are sorted in descending order of predictive importance. The reported wavelengths refer to the most accurate among the four trained models.

References

Arechavala-Lopez, P., Capó, X., Oliver-Codorniú, M., Sillero-Rios, J., & Busquets-Cortés, C. (2019). Fatty acids and elemental composition as biomarkers of Octopus vulgaris populations : Does origin matter ? *Marine Pollution Bulletin*, *139*, 299–310. https://doi.org/10.1016/j.marpolbul.2018.12.048

Beć, K. B., Grabska, J., & Huck, C. W. (2021). Principles and Applications of Miniaturized Near-Infrared (NIR) Spectrometers. *Chemistry - A European Journal*, *27*(5), 1514–1532. https://doi.org/10.1002/chem.202002838

Benes, E., Bajusz, D., Gere, A., Fodor, M., & Rácz, A. (2020). Comprehensive chemometric classification of snack products based on their near infrared spectra. *LWT - Food Science and Technology*, *133*, 110130. https://doi.org/10.1016/j.lwt.2020.110130

Bisutti, V., Merlanti, R., Serva, L., Lucatello, L., Mirisola, M., Balzan, S., Tenti, S., Fontana, F., Trevisan, G., Montanucci, L., Contiero, B., Segato, S., & Capolongo, F. (2019). Multivariate and machine learning approaches for honey botanical origin authentication using near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, *27*(1), 65–74. https://doi.org/10.1177/0967033518824765

Brennecke, D., Duarte, B., Paiva, F., Caçador, I., & Canning-Clode, J. (2016). Microplastics as vector for heavy metal contamination from the marine environment. *Estuarine, Coastal and Shelf Science*, 178, 189-195. https://doi.org/10.1016/j.ecss.2015.12.003

Camacho, M., Herrera, A., Gómez, M., Acosta-Dacal, A., Martínez, I., Henríquez-Hernández, L. A., & Luzardo, O. P. (2019). Organic pollutants in marine plastic debris from Canary Islands beaches. *Science of the total environment*, *662*, 22-31. https://doi.org/10.1016/j.scitotenv.2018.12.422

Cardoso, V. G. K., & Poppi, R. J. (2021). Non-invasive identification of commercial green tea blends using NIR spectroscopy and support vector machine. *Microchemical Journal*, *164*, 106052. https://doi.org/10.1016/j.microc.2021.106052

Cruz-Tirado, J. P., Lucimar da Silva Medeiros, M., & Barbin, D. F. (2021). On-line monitoring of egg freshness using a portable NIR spectrometer in tandem with machine learning. *Journal of Food Engineering*, *306*, 110643. https://doi.org/10.1016/j.jfoodeng.2021.110643

Currò, S., Balzan, S., Serva, L., Boffo, L., Ferlito, J. C., Novelli, E., & Fasolato, L. (2021). Fast and Green Method to Control Frauds of Geographical Origin in Traded Cuttlefish Using a Portable Infrared Reflective Instrument. *Foods*, *10*, 1678. https://doi.org/10.3390/ foods10081678

Currò, S., Fasolato, L., Serva, L., Boffo, L., Ferlito, J. C., Novelli, E., & Balzan, S. (2022). Use of a portable near-infrared tool for rapid on-site inspection of freezing and hydrogen peroxide treatment of cuttlefish (Sepia officinalis). *Food Control*, *132*, 108524. https://doi.org/10.1016/j.foodcont.2021.108524

De-la-Torre, G. E. (2020). Microplastics: an emerging threat to food security and human health. *Journal of food science and technology*, *57*(5), 1601-1608. https://doi.org/10.1007/s13197-019-04138-1

Dos Santos, D. A., Coqueiro, A., Gonçalves, T. R., Carvalho, J. C., Bezerra, J. S., Matsushita, M., de Oliveira, C. A. L., Março, P. H., Valderrama, P., & Ribeiro, R. P. (2020). Omega-3 and Omega-6 Determination in Nile Tilapia's Fillet Based on MicroNIR Spectroscopy and Multivariate Calibration. *Journal of the Brazilian Chemical Society*, *31*(9), 1883–1890. https://doi.org/10.21577/0103-5053.20200082

Dos Santos Pereira, E. V., de Sousa Fernandes, D. D., de Araújo, M. C. U., Diniz, P. H. G. D., & Maciel, M. I. S. (2021). In-situ authentication of goat milk in terms of its adulteration with cow milk using a low-cost portable NIR spectrophotometer. *Microchemical Journal*, *163*. https://doi.org/10.1016/j.microc.2020.105885

EFSA CONTAM Panel (EFSA Panel on Contaminants in the Food Chain) (2016). Presence of microplastics and nanoplastics in food, with particular focus on seafood. *Efsa Journal*, *14*(6), e04501. https://doi.org/10.2903/j.efsa.2016.4501

Esslinger, S., Riedl, J., & Fauhl-Hassek, C. (2014). Potential and limitations of non-targeted fi ngerprinting for authentication of food in offi cial control. *Food Research International*, *60*, 189–204. https://doi.org/10.1016/j.foodres.2013.10.015

European Market Observatory for Fisheries and Aquaculture Products (EUMOFA). (2020). *Octopus in the EU. Price structure in the supply chain. Focus on Italy, Spain and Greece.* Publications Office of the European Union. https://doi.org/10.2771/87203

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fox, M., Mitchell, M., Dean, M., Elliott, C., & Campbell, K. (2018). The seafood supply chain from a fraudulent perspective. *Food Security*, *10*, 939–963. https://doi.org/10.1007/s12571-018-0826-z

~~Freitas, J., Vaz-pires, P., Câmara, J.S., 2020. From aquaculture production to consumption : Freshness , safety , traceability and authentication , the four pillars of quality. Aquaculture 518, 734857. https://doi.org/10.1016/j.aquaculture.2019.734857~~

Ghidini, S., Varrà, M. O., Dall’Asta, C., Badiani, A., Ianieri, A., & Zanardi, E. (2019). Rapid authentication of European sea bass (Dicentrarchus labrax L.) according to production method, farming system, and geographical origin by near infrared spectroscopy coupled with chemometrics. *Food Chemistry*, *280*, 321–327. https://doi.org/10.1016/j.foodchem.2018.12.075

Gomes, F., Oliveira, M., Ramalhosa, M. J., Delerue-Matos, C., & Morais, S. (2013). Polycyclic aromatic hydrocarbons in commercial squids from different geographical origins: levels and risks for human consumption. *Food and chemical toxicology*, *59*, 46-54. https://doi.org/10.1016/j.fct.2013.05.034

González-Mohino, A., Pérez-Palacios, T., Antequera, T., Ruiz-Carrascal, J., Olegario, L. S., & Grassi, S. (2020). Monitoring the processing of dry fermented sausages with a portable NIRS device. *Foods*, *9*(9), 1–12. https://doi.org/10.3390/foods9091294

Grassi, S., Casiraghi, E., & Alamprese, C. (2018). Handheld NIR device: A non-targeted approach to assess authenticity of fish fillets and patties. *Food Chemistry*, *243*, 382–388. https://doi.org/10.1016/j.foodchem.2017.09.145

Guardone, L., Tinacci, L., Costanzo, F., Azzarelli, D., Amico, P. D., Tasselli, G., Magni, A., Guidi, A., Nucera, D., & Armani, A. (2017). DNA barcoding as a tool for detecting mislabeling of fishery products imported from third countries : An official survey conducted at the Border Inspection Post of Livorno-Pisa (Italy). *Food Control*, *80*, 204–216. https://doi.org/10.1016/j.foodcont.2017.03.056

Guglielmetti, C., Manfredi, M., Brusadore, S., Sciuto, S., Esposito, G., Giuseppe, P., Magnani, L., Gili, S., Marengo, E., Luigi, P., & Mazza, M. (2018). Two-dimensional gel and shotgun proteomics approaches to distinguish fresh and frozen-thawed curled octopus (*Eledone cirrhosa*). *Journal of Proteomics*, *186*, 1–7. https://doi.org/10.1016/j.jprot.2018.07.017

Han, C., Choi, H., Jo, S., Na, H., Kim, M. K., Kim, M., & Lee, J. (2020). Development of a 3D scanning method to discriminate blocks of Octopus minor with surplus water gain. *Food Chemistry*, *303*, 125414. https://doi.org/10.1016/j.foodchem.2019.125414

Jiménez-Carvelo, A. M., González-Casado, A., Bagur-González, M. G., & Cuadros-Rodríguez, L. (2019). Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review. *Food Research International*, *122*, 25–39. https://doi.org/10.1016/j.foodres.2019.03.063

Karlson, B., Andersen, P., Arneborg, L., Cembella, A., Eikrem, W., John, U., West, J.J., Kerstin, K., Kobos, J., Lehtinen, S., Lundholm, N., Mazur-Marzec, H., Naustvoll, L., Poelman, M., Provoost, P., De rijcke, M. & Suikkanen, S. (2021). Harmful algal blooms and their effects in coastal seas of Northern Europe. *Harmful Algae*, 101989. https://doi.org/10.1016/j.hal.2021.101989

Lopes, V. M., Lopes, A. R., Costa, P., & Rosa, R. (2013). Cephalopods as vectors of harmful algal bloom toxins in marine food webs. *Marine drugs*, *11*(9), 3381-3409. https://doi.org/10.3390/md11093381

Massutí, E., Gordon, J. D. M., Moranta, J., Swan, S. C., Stefanescu, C., & Merrett, N. R. (2004). Mediterranean and Atlantic deep-sea fish assemblages: Differences in biomass composition and size-related structure. *Scientia Marina*, *68*(S3), 101–115. https://doi.org/https://doi.org/10.3989/scimar.2004.68s3101

Mcvey, C., Elliott, C. T., Cannavan, A., Kelly, S. D., Petchkongkaew, A., & Haughey, S. A. (2021). Portable spectroscopy for high throughput food authenticity screening: Advancements in technology and integration into digital traceability systems. *Trends in Food Science & Technology*, *118*, 777–790. https://doi.org/10.1016/j.tifs.2021.11.003

Mendes, R., Schimmer, O., Vieira, H., & Teixeira, B. (2018). Control of abusive water addition to *Octopus vulgaris* with non-destructive methods. *Journal of the Science of Food and Agriculture*, *98*, 369–376. https://doi.org/10.1002/jsfa.8480

Michelucci, U., & Venturini, F. (2021). Estimating Neural Network’ s Performance with Bootstrap: A Tutorial. *Machine Learning and Knowledge Extraction*, *3*, 357–373. https://doi.org/10.3390/make3020018

Müller-Maatsch, J., Alewijn, M., Wijtten, M., & Weesepoel, Y. (2021a). Detecting fraudulent additions in skimmed milk powder using a portable, hyphenated, optical multi-sensor approach in combination with one-class classification. *Food Control*, *121*, 107744. https://doi.org/10.1016/j.foodcont.2020.107744

Müller-Maatsch, J., Bertani, F. R., Mencattini, A., Gerardino, A., Martinelli, E., Weesepoel, Y., & Van Ruth, S. (2021b). The spectral treasure house of miniaturized instruments for food safety, quality and authenticity applications: A perspective. *Trends in Food Science and Technology*, *110*, 841–848. https://doi.org/10.1016/j.tifs.2021.01.091

Muncan, J., & Tsenkova, R. (2019). Aquaphotomics-From Innovative Knowledge to Integrative Platform in Science and Technology. *Molecules*, *24*(15), 2742. https://doi.org/10.3390/molecules24152742

Nowak, M., Wietecha-pos, R., & Pawliszyn, J. (2021). White Analytical Chemistry: an approach to reconcile the principles of Green Analytical Chemistry and functionality. *Trends in Analytical Chemistry*, *138*, 116223. https://doi.org/10.1016/j.trac.2021.116223

Oliveira, M., Gomes, F., Torrinha, Á., João, M., Delerue-matos, C., & Morais, S. (2018). Commercial octopus species from different geographical origins : Levels of polycyclic aromatic hydrocarbons and potential health risks for consumers. *Food and Chemical Toxicology*, *121*, 272–282. https://doi.org/10.1016/j.fct.2018.09.012

Parastar, H., van Kollenburg, G., Weesepoel, Y., van den Doel, A., Buydens, L., & Jansen, J. (2020). Integration of handheld NIR and machine learning to “Measure & Monitor” chicken meat authenticity. *Food Control*, *112*, 107149. https://doi.org/10.1016/j.foodcont.2020.107149

Pennisi, F., Giraudo, A., Cavallini, N., Esposito, G., Merlo, G., Geobaldo, F., Acutis, P. L., Pezzolato, M., Savorani, F., & Bozzetta, E. (2021). Differentiation between fresh and thawed cephalopods using NIR spectroscopy and multivariate data analysis. *Foods*, *10*(3), 1–14. https://doi.org/10.3390/foods10030528

Phan, A. V., Nguyen, M. Le, & Bui, L. T. (2017). Feature weighting and SVM parameters optimization based on genetic algorithms for classification problems. *Applied Intelligence*, *46*(2), 455–469. https://doi.org/10.1007/s10489-016-0843-6

Pita, C., Roumbedakis, K., Fonseca, T., Matos, F. L., Pereira, J., Villasante, S., Pita, P., Bellido, J. M., Gonzalez, A. F., García-Tasende, M., Lefkaditou, E., Adamidou, A., Cuccu, D., Belcari, P., Moreno, A., & Pierce, G. J. (2021). Fisheries for common octopus in Europe: socioeconomic importance and management. *Fisheries Research*, *235*, 105820. https://doi.org/10.1016/j.fishres.2020.105820

Rinnan, Å., Berg, F. van den, & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. In *TrAC - Trends in Analytical Chemistry*, *28*, 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Ríos-Reina, R., Camiña, J. M., Callejòn, R. M., & Azcarate, S. M. (2021). Trends in Analytical Chemistry Spectralprint techniques for wine and vinegar characterization, authentication and quality control: Advances and projections. *Trends in Analytical Chemistry*, *134*, 116121. https://doi.org/10.1016/j.trac.2020.116121

Rjeibi, M., Metian, M., Hajji, T., Guyot, T., Chaouacha-Chékir, R. B., & Bustamante, P. (2014). Interspecific and geographical variations of trace metal concentrations in cephalopods from Tunisian waters. *Environmental monitoring and assessment*, *186*(6), 3767-3783. https://doi.org/10.1007/s10661-014-3656-2

Roldán-Wong, N. T., Kidd, K. A., Ceballos-Vázquez, B. P., & Arellano-Martínez, M. (2018). Is there a risk to humans from consuming octopus species from sites with high environmental levels of metals?. *Bulletin of environmental contamination and toxicology*, *101*(6), 796-802. https://doi.org/10.1007/s00128-018-2447-9

~~Saito, H., Ishihara, K., & Murase, T. (1997). The fatty acid composition in tuna (bonito, *Euthynnus pelamis*) caught at three different localities from tropics to temperate. *Journal of the Science of Food and Agriculture*, *73*(1), 53–59. https://doi.org/10.1002/(SICI)1097-0010(199701)73:1<53::AID-JSFA707>3.0.CO;2-5~~

Sampaio, P. S., Castanho, A., Almeida, A. S., Oliveira, J., & Brites, C. (2020). Identification of rice flour types with near-infrared spectroscopy associated with PLS-DA and SVM methods. *European Food Research and Technology*, *246*(3), 527–537. https://doi.org/10.1007/s00217-019-03419-5

Semedo, M., Reis-Henriques, M. A., Rey-Salgueiro, L., Oliveira, M., Delerue-Matos, C., Morais, S., & Ferreira, M. (2012). Science of the Total Environment Metal accumulation and oxidative stress biomarkers in octopus (*Octopus vulgaris*) from Northwest Atlantic. *Science of the Total Environment*, *433*, 230–237. https://doi.org/10.1016/j.scitotenv.2012.06.058

Sillero-Ríos, J., Sureda, A., Capó, X., Oliver-Codorniú, M., & Arechavala-Lopez, P. (2018). Biomarkers of physiological responses of Octopus vulgaris to different coastal environments in the western Mediterranean Sea. *Marine Pollution Bulletin*, *128*, 240–247. https://doi.org/10.1016/j.marpolbul.2018.01.032

Silva, L. C. R., Folli, G. S., Santos, L. P., Barros, I. H. A. S., Oliveira, B. G., Borghi, F. T., Santos, F. D. do., Filgueiras, P. R., & Romão, W. (2020). Quantification of beef, pork, and chicken in ground meat using a portable NIR spectrometer. *Vibrational Spectroscopy*, *111*. https://doi.org/10.1016/j.vibspec.2020.103158

Stadler, R. H., Tran, L. A., Cavin, C., Zbinden, P., & Konings, E. J. M. (2016). Analytical approaches to verify food integrity: Needs and challenges. *Journal of AOAC International*, *99*(5), 1135–1144. https://doi.org/10.5740/jaoacint.16-0231

Tatulli, G., Cecere, P., Maggioni, D., Galimberti, A., & Pompa, P. P. (2020). A Rapid Colorimetric Assay for On-Site Authentication of Cephalopod Species. *Biosensors*, *10*(190). https://doi.org/10.3390/bios10120190

Tinacci, L., Armani, A., Scardino, G., Guidi, A., Nucera, D., Miragliotta, V., & Abramo, F. (2020). Selection of Histological Parameters for the Development of an Analytical Method for Discriminating Fresh and Frozen / Thawed Common Octopus (*Octopus vulgaris*) and Preventing Frauds along the Seafood Chain. *Food Analytical Methods*, *13*, 2111–2127. https://doi.org/10.1007/s12161-020-01825-0

Torrinha, A., Cruz, R., Gomes, F., Casal, S., & Morais, S. (2014). Octopus Lipid and Vitamin E Composition: Interspecies, Interorigin, and Nutritional Variability. *Journal of Agricultural and Food Chemistry*, *62*, 8508–8517. https://doi.org/10.1021/jf502502b

Varrà, M. O., Ghidini, S., Husáková, L., Ianieri, A., & Zanardi, E. (2021a). Advances in Troubleshooting Fish and Seafood Authentication by Inorganic Elemental Composition. *Foods*, *10*, 270. https://doi.org/10.3390/foods10020270

Varrà, M. O., Ghidini, S., Ianieri, A., & Zanardi, E. (2021b). Near infrared spectral fingerprinting: A tool against origin-related fraud in the sector of processed anchovies. *Food Control*, *123*. https://doi.org/10.1016/j.foodcont.2020.107778

Whitelaw, B. L., Cooke, I. R., Finn, J., Zenger, K., & Strugnell, J. M. (2019). The evolution and origin of tetrodotoxin acquisition in the blue-ringed octopus (genus Hapalochlaena). *Aquatic*

Toxicology, 206, 114-122. https://doi.org/10.1016/j.aquatox.2018.10.012

Workman, J. ., & Weyer, L. (2012). *Practical Guide and Spectral Atlas for Interpretive Near-Infrared Spectroscopy* (2nd ed.). CRC Press (Taylor & Francis group).

Yakes, B. J., Ellsworth, Z., Karunathilaka, S. R., & Crump, E. (2021). Evaluation of Portable Sensor and Spectroscopic Devices for Seafood Decomposition Determination. *Food Analytical Methods*, *14*, 2346–2356. https://doi.org/10.1007/s12161-021-02064-7

Yu, H. D., Zuo, S. M., Xia, G., Liu, X., Yun, Y. H., & Zhang, C. (2020). Rapid and Nondestructive Freshness Determination of Tilapia Fillets by a Portable Near-Infrared Spectrometer Combined with Chemometrics Methods. *Food Analytical Methods*, *13*(10), 1918–1928. https://doi.org/10.1007/s12161-020-01816-1

Zareef, M., Chen, Q., Hassan, M. M., Arslan, M., Hashim, M. M., Ahmad, W., Kutsanedzie, F. Y. H., & Agyekum, A. A. (2020). An Overview on the Applications of Typical Non-linear Algorithms Coupled With NIR Spectroscopy in Food Analysis. *Food Engineering Reviews*, *12*(2), 173–190. https://doi.org/10.1007/s12393-020-09210-7

Figure captions

Figure 1. Sampling area extensions of musky and common octopuses.

Figure 2. Effect of the application of different pre-processing filters (Standard Normal Variate, SNV; first derivative, 1st Der) on the quality and usefulness of the 908.1–1676.2 nm spectra recorded by MicroNIR (Atlantic samples: red lines; Mediterranean samples: red lines). The main spectral features in SNV + 1st Der spectral patterns are highlighted by dotted marker boxes.

Figure 3. 3-D score scatter plot from PCA applied to musky and common octopuses.

Figure 4 Main figures of merit (mean ± standard deviation) of the five different machine learning tools (OPLS-DA, LR, RF, SVM, and RF) obtained by cross-validation of the training samples for the characterization of the geographic origin of musky and common octopuses.

Figure 5. Confusion matrices (mean classification rates) and corresponding areas under the curve (AUC) values of receiver operating characteristic (AUROC) ~~AUROC~~ values (mean ± standard deviation) resulting from the origin prediction of 118 musky and 39 common octopus ~~test~~ spectral samples (included into the external validation sets) by the five different classifiers (OPLS-DA, LR, RF, SVM and MLP-ANN). Correct classification rates in confusions matrices are included into green boxes (Med: Mediterranean samples, Atl: Atlantic samples).