**Finding Nemo’s family: Enhancing NIR-Based Authentication of Mediterranean Anchovies- The Influence of Spectra Pre-processing and Machine Learning Techniques**

Nidhi Dalal (nidhi.dalal@unina.it)

Department of Agricultural Sciences, University of Naples Federico II, Portici (Naples), Italy

Tutor: Prof. Paola Adamo; Co-Tutor: Prof. Antonio G. Caporale

In this study, we explored the impact of spectra pre-processing techniques and machine learning algorithms on the authentication of anchovies using near-infrared (NIR) spectroscopy. Raw NIR spectra of anchovy samples underwent various pre-processing methods, either individually or in combination. Subsequently, the spectra were analyzed using unsupervised Principal Component Analysis (PCA) followed by Linear Discriminant Analysis (LDA). The results obtained from LDA were compared to those obtained using machine learning (ML) models, specifically Support Vector Machines (SVM). SVM outperformed PCA-LDA modelling in terms of classification accuracy and did not involve any data reduction due to its superior computing ability.

**Finding Nemo’s family: Miglioramento dell'autenticazione basata su spettroscopia NIR delle alici del Mar Mediterraneo: l'influenza delle tecniche di pre-elaborazione degli spettri e di Machine Learning**

In questo studio, abbiamo esplorato l'impatto delle tecniche di pre-elaborazione degli spettri e degli algoritmi di machine learning sull'autenticazione delle alici utilizzando la spettroscopia nel vicino infrarosso (NIR). Gli spettri NIR grezzi di campioni di alici sono stati sottoposti a vari metodi di pre-elaborazione, singolarmente o in combinazione. Successivamente, gli spettri sono stati analizzati utilizzando l'analisi delle componenti principali (PCA) non supervisionata seguita dall'analisi discriminante lineare (LDA). I risultati ottenuti dalla LDA sono stati confrontati con quelli ottenuti utilizzando modelli di machine learning (ML), nello specifico Support Vector Machines (SVM). SVM ha superato la modellazione PCA-LDA in termini di accuratezza della classificazione e non ha comportato alcuna riduzione dei dati grazie alla sua capacità di elaborazione superiore.

**Key words**: NIR spectroscopy, fish authentication, machine learning modelling, quality control, traceability

**1. Introduction**

# This oral communication highlights the primary outcomes of four activities with the following objectives:

# A1) Comparison of raw NIR spectra with processed spectra using various pre-processing methods, both individually and in combination, resulting in a total of nine different outputs.

# A2) Evaluation of the authentication models based on Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for each pre-processing technique employed.

# A3) Assessment of the authentication model based on Support Vector Machines (SVM) using linear and polynomial kernel for each pre-processing technique utilized.

# A4) Comparison of the authentication model performance between PCA-LDA and SVM approaches, determining their respective strengths and limitations.

# The presentation will provide a comprehensive overview of the findings from these activities, shedding light on the effectiveness of different pre-processing methods and the performance of PCA-LDA, linear kernel SVM and polynomial kernel SVM models in NIR-based authentication.

# **2. Application of NIR Spectroscopy for Authentication**

# Food authentication is of utmost importance in ensuring product quality, safety, and preventing fraudulent practices in the food industry. Near-infrared (NIR) spectroscopy has emerged as a powerful tool for food authentication due to its non-destructive nature and ability to provide rapid and reliable results. NIR spectroscopy utilizes the interaction of light with the molecular structure of food samples to generate spectral data. These spectra contain valuable information about the composition and characteristics of food products (Wang *et al*., 2017).

# Spectral pre-processing techniques play a crucial role in optimizing the quality of NIR modelling for food authentication. Pre-processing methods such as baseline correction, scatter correction, normalization, and outlier removal help to eliminate noise, enhance spectral features, and reduce unwanted variations. The appropriate selection and combination of pre-processing methods significantly impact the accuracy and robustness of the subsequent modelling process (Rinnan *et al*., 2009). This needs a trial-and-error approach to pick the best pre-processing method(s) suitable for the given food matrix so as not to have under or over pre-processing of the spectra.

# Given the large size of the NIR spectra generated (6000 data points for each NIR spectrum in this study), the general chemometric approach relies on data reduction using unsupervised methods such as PCA. The multicollinear nature of NIR spectra also benefits from PCA based data reduction. The principal components (PCs) thus obtained are used for supervised analysis. LDA, quadratic discriminant analysis (QDA), partial least squares regression-discriminant analysis (PLS-DA), soft independent modelling of class analogies (SIMCA), Orthogonal PLS-DA (OPLS-DA) are the most commonly used chemometric supervised techniques (Berrueta *et al*., 2007).

# While traditional chemometric techniques have been widely used in NIR spectroscopy, machine learning algorithms have demonstrated superiority in food authentication tasks. Machine learning models, particularly those based on Support Vector Machines (SVM), possess remarkable capabilities in handling complex and high-dimensional NIR data used in this study. They can effectively identify patterns and classify samples based on their spectral fingerprints, providing higher accuracy and more robust predictions compared to traditional methods. Machine learning algorithms can learn from large datasets and adapt to various spectral variations, making them more versatile in handling different food authentication scenarios. Additionally, machine learning models can allow for an easy and quick integration of multiple spectral and non-spectral parameters, enabling a comprehensive analysis of food samples and enhancing the overall authentication process (Song *et al*., 2020; Mishra *et al*., 2022). So far, only one study has been published on application of ML on fish authentication, with no comparison with traditional chemometric techniques.

# **3. Sampling Plan**

The selection of the sampling area and the number of samples is a crucial aspect of any study focusing on food traceability, particularly in the context of fish. It is essential to carefully consider these factors to ensure the accuracy and representativeness of the study results. When choosing samples, it is important to strive for an ideal representation of the natural variance of the analyte of interest within the target population. A representative sample can be defined as an aliquot of a material taken from a consignment, possessing all the essential characteristics of the bulk (Murray and Cowe, 2004).

In the case of fish, this natural variation can stem from factors such as fish age or size, the season of fishing, geographical location, fishing type (wild vs. farmed), freshness, food processing methods, water salinity, and water temperature, among others. Furthermore, the process of sampling can be influenced by various practical considerations. These include the available budget and timeframe for the project, the specific objectives of the study, and the logistical aspects associated with sample collection. Access to authentic samples, which accurately reflect the target population, also needs to be considered. By ensuring that the samples are selected in a manner that encompasses the key characteristics of the bulk, the study can provide reliable and meaningful insights into food traceability.

The samples from this study were obtained from local market in Portici (NA), Italy. The details of procurement of fresh anchovies from Tyrrhenian Sea and Adriatic Sea are provided in table 1.

|  |
| --- |
| **Table 1**: *Sampling plan of anchovies*  |
| Month (2022) | Tyrrhenian Sea | Adriatic Sea |
| January | Trial 12 (n= 6) | - |
| February | Trial 14 (n= 12) | Trial 13 (n= 12), Trial 15 (n= 8) |
| March | Trial 16 (n= 11) | Trial 18 (n= 10), Trial 19 (n= 9) |
| April | Trial 20 (n= 9) | Trial 21 (n= 11) |
| May | Trial 23 (n= 9) | - |
| September | Trial 27 (n= 8) | Trial 26 (n= 9) |
| October | Trial 28 (n= 6) | Trial 29 (n= 9) |
|  | N= 61 | N= 68 |

Care was taken to keep the number of samples equal in anchovies from both locations to maintain sample balance. The final sampling, however, depended on the seasonal availability of fresh anchovies from either locations in the local market.

# **4. Materials and Methods**

Fresh anchovies from Tyrrhenian Sea and Adriatic Sea were cleaned (removal of head, guts, bone, and internal viscera), washed with distilled water and sorted into different groups based on weight classes with a weight difference of 2g each.10-12 fishes from each group were homogenized together. The homogenate was freeze-dried and pulverized with ball-milling machine until 60% of it can pass through 100-mesh sieve and 80% can pass through 60-mesh size sieve. Perkin Elmer FT-IR 9700 with NIRA (Near-Infrared Accessory) was used for spectra acquisition. Three spectra were acquired from each sample resulting in a total of 387 spectra (interleaved) from 129 samples. They were acquired over a wavelength of 1000-2500 nm with the resolution of 4 cm-1, and 64 scans per spectrum.

The spectral pre-processing was performed on Unscrambler X (Version 10.4, 64 bit). Further data analysis was performed using Python 3.10.4. The Python libraries used for data analysis were Sci-Kit learn 1.2.2, Panda 2.0.2, Numpy 1.21.6, OS 2.1.4, and MatPlotLib 3.7.1. NIR spectra were pre-processed using techniques outlined in Table 2. Pre-processing techniques were chosen after a thorough review of literature for NIR based authentication of fish, meat, and other food products.

|  |
| --- |
| **Table 2***: List of pre-processing techniques used for NIR spectra* |
| Pre-processing | Code |
| Raw spectra | Raw |
| Standard Normal Variate  | SNV |
| Savitzky-Golay 1st derivate with 15 points of smoothening | SG-1d |
| Savitzky-Golay 2nd derivate with 15 points of smoothening  | SG-2d |
| Multiplicative Scatter Correction | MSC |
| Standard Normal Variate + Savitzky-Golay 1st derivate with 15 points of smoothening | SNV+SG-1d |
| Standard Normal Variate + Savitzky-Golay 2nd derivate with 15 points of smoothening | SNV+SG-2d |
| Standard Normal Variate + Multiplicative Scatter Correction | SNV+MSC |
| Multiplicative Scatter Correction + Savitzky-Golay 1st derivate with 15 points of smoothening | MSC+SG-1d |
| Multiplicative Scatter Correction + Savitzky-Golay 2nd derivate with 15 points of smoothening | MSC+SG-2d |

For chemometric modelling, the bulk of the NIR data with 6000 data points was reduced to 7 PCs by PCA. Initial PCA was checked for outliers and the whole process repeated to get final PCs for LDA. LDA was used to build a discrimination model using with training set and testing set of 0.8 and 0.2 respectively. For machine learning, SVM algorithm with liner and polynomial kernel using training and testing set of 0.8 and 0.2 respectively. Complete NIR spectra were used for SVM classification.

# **5. Results and Discussion**

## **5.1 Spectral changes due to pre-processing**

## Raw NIR spectra, comprising numerous variables representing chemical constituents, suffer from multi-collinearity due to overlapping bands. Individual analytes may absorb at multiple wavelengths, leading to potential misreading. Weak absorption bands and non-linear scatter effects further complicate data interpretation. Baseline shifts and scatter effects, caused by light scattering and path length differences, introduce undesired variations (Rinnan *et al.*, 2009). Therefore, pre-processing is essential to mitigate these effects and enhance subsequent exploratory analysis. Image 1 shows the changes in the NIR spectra after spectral pre-processing.



**Figure 1**: *Changes in NIR spectra after spectral pre-processing*

## **5.2 PCA-LDA analysis**

Training score, testing score and cross-validation scores of PCA-LDA to discriminate between anchovies from Tyrrhenian Sea and Adriatic Sea are shown in Table 3.

All the pre-processed spectra performed better than raw spectra in general. The best performing pre-processing was that of SNV+SG-1d followed closely by MSC+SG-1d. As can be seen by Figure 1, SG-1d and SG-2d led to a large-scale spectral transformation. However, 2nd derivate spectra appeared over-pre-processed and thus did not perform as well as 1st derivate. The multicollinearity of NIR spectra in these cases of PCA was only reduced, not completely eliminated and thus the discrimination was not entirely accurate. Reduction in accuracy can also be attributed to seasonal and size differences between the anchovies from same location. Anchovies have been documented to have seasonal changes based on changes in food availability and size. Since, the sampling was completely random and did not pick anchovies of one size, the variance in NIR data from the same batch of anchovies was high.

|  |
| --- |
| **Table 3***: Results of PCA-LDA discrimination on raw and pre-processed NIR spectra* |
| Pre-processing | Training Score | Testing Score | Cross Validation |
| Raw | 0.74 | 0.67 | 0.65 |
| SNV | 0.87 | 0.90 | 0.91 |
| SG-1d | 0.89 | 0.90 | 0.92 |
| SG-2d | 0.81 | 0.85 | 0.83 |
| MSC | 0.84 | 0.89 | 0.87 |
| SNV + SG-1d | 0.93 | 0.96 | 0.95 |
| SNV + SG-2d | 0.82 | 0.86 | 0.84 |
| SNV + MSC | 0.89 | 0.92 | 0.92 |
| MSC + SG-1d | 0.94 | 0.95 | 0.92 |
| MSC + SG-2d | 0.80 | 0.85 | 0.81 |

## **5.3 Machine learning analysis**

## SVM was chosen over all other ML approaches in its efficiency in handling large sets of multicollinear nature, like the NIR spectra. Linear and polynomial kernels were used for classification between Tyrrhenian and Adriatic anchovies. The classification results are presented in Table 4.

|  |
| --- |
| **Table 4***: Results of SVM classification on raw and pre-processed NIR spectra* |
| Pre-processing | SVM linear kernel | SVM polynomial kernel |
| Train | Test | Train | Test |
| Raw | 0.64 | 0.54 | 0.69 | 0.55 |
| SNV | 0.82 | 0.91 | 0.89 | 0.90 |
| SG-1d | 0.89 | 0.96 | 0.98 | 1.00 |
| SG-2d | 0.80 | 0.76 | 0.81 | 0.85 |
| MSC | 0.81 | 0.80 | 0.89 | 0.78 |
| SNV + SG-1d | 1.00 | 0.97 | 1.00 | 1.00 |
| SNV + SG-2d | 0.86 | 0.93 | 0.88 | 0.89 |
| SNV + MSC | 0.85 | 0.80 | 0.89 | 0.81 |
| MSC + SG-1d | 0.98 | 1.00 | 1.00 | 0.99 |
| MSC + SG-2d | 0.93 | 0.91 | 0.92 | 0.82 |

In general, polynomial kernel performed better than linear kernel in this case since the data was highly multicollinear and the whole spectra was used instead of PCs. The pre-processing techniques remarkably well, both individually and in combination. Most pre-processing techniques had accuracy of more than 0.90 in training and testing set. SNV+SG-1d performed the best followed by MSC+SG-1d and SG-1d.

## **5.4 Comparison between PCA-LDA, SVM linear and SVM polynomial**

SVM polynomial outperformed LDA-PCA and SVM linear in terms of accuracy of classification of anchovies between the two locations. SVM was better able to capture and process the non-linear seasonal and size differences between anchovies from the same location thereby picking up intricate relations in the dataset. Use of SVM also removed the step of data reduction and made the whole step of data analysis shorter and less time consuming. This was possible to high computing ability of ML techniques. LDA relied on certain assumptions about the data distribution and relationship between variables given by PCA. SVM models, on the other hand, are more flexible and can make predictions without strong assumptions, reducing the risk of model bias (Song *et al*., 2020; Mishra *et al*., 2022). Some pre-processing techniques performed better with PCA-LDA than SVM, which can be attributed to the reduction in multicollinearity by PCA.

Between the two SVM kernels, polynomial kernel outperformed the linear kernel, since most relationships in NIR spectra are non-linear which were better captured by the polynomial kernel. The polynomial kernel implicitly maps the original feature space into a higher-dimensional feature space. By doing so, it enables SVM models to find non-linear decision boundaries that can effectively separate data points in the transformed space. This ability to operate in a higher-dimensional feature space can help SVM models better handle multicollinear NIR datasets, where the variables may not be linearly separable in the original feature space (Patle and Chouhan, 2013). In addition, the polynomial kernel allows for the detection and representation of interactions between variables. In the case of multicollinearity, where variables may be correlated with each other, the polynomial kernel can capture the complex interactions and dependencies between these variables. This flexibility enables the SVM model to better understand the underlying structure of the data and make more accurate predictions (Ben-Hur *et al*., 2008).

The anchovies from two locations in Mediterranean Sea were thus classified due to changes in their chemical composition which can be further attributed to differences in food availability between the two locations. This was shown by analysis of nutritional and fatty acids composition of fish. Large scale differences were observed in fatty acid composition, especially omega-3 fatty acids which were then measured by the NIR spectra. (The nutritional data was not complete on the day of submission and thus not shown here).

# **6. Conclusions and Future Perspectives**

# Spectral pre-processing in as integral part of NIR based food authentication and depends on the food matrix and instrument used for NIR spectra. SNV+SG-1d, followed closely by MSC-SG-1d and SG-1d turned out to be the best spectral pre-treatment based on discrimination and classification results by PCA-LDA and SVM respectively. The SVM polynomial kernel outperformed LDA-PCA and SVM linear in accurately classifying anchovies between different locations. SVM exhibited superior capability in capturing and processing the non-linear seasonal and size differences within anchovies from the same location, uncovering intricate relations in the dataset and eliminated the need for data reduction. Overall, the polynomial kernel outperformed the linear kernel by effectively handling the non-linear relationships present in NIR spectra by mapping the feature space into a higher-dimensional space enabling the SVM models to discover non-linear decision boundaries and better handle multicollinear NIR datasets.

Although ML handled the NIR data considerably well, there is still scope of bettering the process by spectral selection to further reduce the bulk of data and ease the process of data fusion. Data fusion, which involves using data from two more analytical techniques, can create more robust and chemically comprehensive model.

# **8. References**

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., & Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS computational biology*, *4*(10), e1000173.

Berrueta, L. A., Alonso-Salces, R. M., & Héberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of chromatography A*, *1158*(1-2), 196-214.

Cozzolino, D. (2016). Near Infrared Spectroscopy and Food Authenticity. *Advances in Food Traceability Techniques and Technologies: Improving Quality Throughout the Food Chain*, 119–136. https://doi.org/10.1016/B978-0-08-100310-7.00007-7

 Mishra, P., Passos, D., Marini, F., Xu, J., Amigo, J. M., Gowen, A. A., ... & Nordon, A. (2022). Deep learning for near-infrared spectral data modelling: Hypes and benefits. *TrAC Trends in Analytical Chemistry*, 116804.

Murray, I., & Cowe, I. (2004). Near-infrared spectroscopy in agriculture. American Society of Agronomy. *Crop Science Society of America, Soil Science Society of America*.

Patle, A., & Chouhan, D. S. (2013, January). SVM kernel functions for classification. In *2013 International Conference on Advances in Technology and Engineering (ICATE)* (pp. 1-9). IEEE.

Rinnan, Å., Van Den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, *28*(10), 1201-1222.

Song, W., Jiang, N., Wang, H., & Guo, G. (2020). Evaluation of machine learning methods for organic apple authentication based on diffraction grating and image processing. *Journal of Food Composition and Analysis*, *88*, 103437.

Wang, L., Sun, D. W., Pu, H., & Cheng, J. H. (2017). Quality analysis, classification, and authentication of liquid foods by near-infrared spectroscopy: A review of recent research developments. *Critical reviews in food science and nutrition*, *57*(7), 1524-1538.