Cultivar classification of Apulian olive oils: use of artificial neural networks for comparing NMR, NIR and merceological data

Giulio Binetti, Laura Del Coco, Rosa Ragone, Samanta Zelasco, Enzo Perri, Cinzia Montemurro, Raffaele Valentini, David Naso, Francesco Paolo Fanizzi, Francesco Paolo Schena

PII: S0308-8146(16)31424-8
DOI: http://dx.doi.org/10.1016/j.foodchem.2016.09.041
Reference: FOCH 19821

To appear in: *Food Chemistry*

Received Date: 30 May 2016
Revised Date: 2 September 2016
Accepted Date: 6 September 2016

Please cite this article as: Binetti, G., Coco, L.D., Ragone, R., Zelasco, S., Perri, E., Montemurro, C., Valentini, R., Naso, D., Fanizzi, F.P., Schena, F.P., Cultivar classification of Apulian olive oils: use of artificial neural networks for comparing NMR, NIR and merceological data, *Food Chemistry* (2016), doi: http://dx.doi.org/10.1016/j.foodchem.2016.09.041

Cultivar classification of Apulian olive oils: use of artificial neural networks for comparing NMR, NIR and merceological data

Giulio Binetti[a,‡*], Laura Del Coco[b,‡], Rosa Ragone[c,‡], Samanta Zelasco[d], Enzo Perri[d], Cinzia Montemurro[e], Raffaele Valentini[f], David Naso[a], Francesco Paolo Fanizzi[b*], Francesco Paolo Schena[c*]

[a] Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, Via E. Orabona, 4, 70125 Bari, Italia; e-mail: giulio.binetti@poliba.it, david.naso@poliba.it

[b] Dipartimento di Tecnologie Biologiche ed Ambientali, Università del Salento, Prov.le Lecce-Monteroni, 73100 Lecce, Italia; e-mail: laura.delcoco@unisalento.it, fp.fanizzi@unisalento.it

[c] Consorzio C.A.R.S.O., Università di Bari, Strada Provinciale Casamassima Km 3, 70010 Valenzano (Bari), Italia; e-mail: rosaragone@yahoo.it, paolo.schena@uniba.it

[d] Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria-Centro di ricerca per l'Olivicoltura e l'Industria Olearia, Contrada Li Rocchi, 87036 Rende (Cosenza), Italia; e-mail: samanta.zelasco@crea.gov.it, enzo.perri@crea.gov.it

[e] Dipartimento di Biologia e Chimica Agro-Forestale ed Ambientale, Sezione di Genetica e Miglioramento, Università di Bari, via Amendola 165/a, 70126 Bari, Italia; e-mail: cinzia.montemurro@uniba.it

[f] Oliveti Terra di Bari O.P. Olivicoli Soc. Coop. Agricola, 6/A, Via Brigata 6/A, 70124 Bari, Italia; e-mail: agrivalent@libero.it

[‡] These authors have contributed equally to this work.

* Corresponding authors: G. Binetti, F.P. Fanizzi, F.P. Schena

ABSTRACT

The development of an efficient and accurate method for extra-virgin olive oils cultivar and origin authentication is complicated by the broad range of variables (e.g., multiplicity of varieties, pedo-climatic aspects, production and storage conditions) influencing their properties. In this study, artificial neural networks (ANNs) were applied on several analytical datasets, namely standard merceological parameters, near-infra red data and ^{1}H nuclear magnetic resonance (NMR) fingerprints, obtained on mono-cultivar olive oils of four representative Apulian varieties (Coratina, Ogliarola, Cima di Mola, Peranzana). We analysed 888 samples produced at a laboratory-scale during two crop years from 444 plants, whose variety was genetically ascertained, and on 17 industrially produced samples. ANN models based on NMR data showed the highest capability to classify cultivars (in some cases, accuracy > 99%), independently on the olive oil production process and year; hence, the NMR data resulted to be the most informative variables about the cultivars.

Keywords:

Artificial neural networks; olive oil; cultivar classification; merceological analysis; near-infra red spectroscopy; nuclear magnetic resonance spectroscopy.

1. INTRODUCTION

Extra-virgin olive oil (EVOO) is considered as one of the best sources of "good" fatty acids and antioxidants, with positive effects on human health [1]. It is a complex food matrix, difficult to be analyzed. Monitoring quality (from harvesting through transformation to storage) and authentication (detection of adulterations, identification of geographical origin and variety) are nowadays the main challenges for olive oil industry and food control laboratories. In an attempt to protect customers and producers against false declarations, international organizations have established the guidelines for olive oil certification, indicating the methods to determine several chemical and physical parameters and reference limits. However, the official procedures often result to be inadequate to screen a large number of samples, time consuming, and insufficient for a quick and detailed examination [2]. As regards the authentication of the cultivar of EVOOs, the development of an efficient and accurate method is complicated by the broad range of variables influencing the olive oil properties: pedo-climatic aspects and process conditions interact with genetic characteristics.

The general strategy followed by researchers is to get the metabolic fingerprints of a large amount of oils obtained from many varieties, and then to build up models by means of multivariate statistical analyses (MVA) to predict unknown samples. Among the modern analytical methods, nuclear magnetic resonance (NMR) spectroscopy seems very attractive, because it requires an easy sample preparation and rapidly provides a complete metabolic profile of olive oils, giving information about either the lipid fraction and several minor compounds (sterols, tocopherols, polyphenols, oxidized products, etc.) [3-6]. Near infra-red (NIR) spectroscopy [7,8], gas and high-performance liquid chromatography with mass spectrometry (MS) [9,10], and electronic sensors have been exploited too [11,12].

MVA programs handle the large amounts of metabolic data produced by innovative techniques, and extract the main variables that discriminate between the categories under examination. Beside the exploratory methods, such as Principal Component, Hierarchical Cluster and Tree Clustering Analysis (unsupervised), ~~and~~ Linear and Quadratic Discriminant Analysis, Partial Least Square discriminant analysis (PLS-DA) ~~regression~~ (supervised) are some of the most extensively used classification approaches. However, sometimes their usage as a reliable approach to classify cultivars and geographic origins and to unravel adulterations has been negatively criticized [2].

Recently, artificial neural networks (ANNs) have been introduced in food analysis [13]. ANNs are a set of mathematical methods, which attempt to mimic the functioning of the human brain [14]. They consist of sophisticated non-linear computational tools that are capable of modelling extremely complex functions learning by example: the data structure is automatically learnt from representative data by means of opportunely designed training algorithms. There are some examples of usage of ANNs for olive oil classification according to geographical origin, year of production, merceological category, adulteration, processing, and blending. Generally, in those works ANNs have been built using only one kind of analytical data, such as data obtained through mass spectrometry [15], NIR [16], electronic sensors [11,12], NMR [17,18], or traditional standardized methods [19,20]. The use of ANNs for cultivar classification of olive oils has not been completely explored. Bucci R. et al. have shown that, using chemometrics and ANNs and choosing the chemical indices routinely determined for the oil quality control as descriptors, it was possible to accurately attribute the cultivar of 153 Italian EVOOs obtained from five different varieties [21]. Peres A.M. et al. measured ten biometrical parameters of oil samples belonging to six Portuguese cultivars, collected in different groves during four crop years, and created an artificial neural network able of predicting the variety of unknown samples more accurately than using linear discriminant analysis [22].

In our study, ANN models were set up using multiple types of information, namely standard merceological parameters, NIR data, and NMR fingerprints, in order to find the most accurate ANN model for cultivar classification.

2. MATERIALS AND METHODS

2.1 Selection of plants and olive oil extraction

We focused on four representative cultivars of Apulia, namely Coratina, Ogliarola Barese, Cima di Mola, and Peranzana. Forty hundered-fifth plants were selected in 15 areas (30 plants per area) across the Foggia and Bari provinces, in order to cover different pedo-climatic regions. We chose 240 plants for Coratina, 60 for Ogliarola Barese, 90 for Cima di Mola, and 60 for Peranzana, based on phenotype characteristics. Every tree was marked with an identification code. Harvest was performed at the optimal olive ripening stage, in different periods depending on cultivar and growing conditions, in two subsequent crop years (2013-2014 and 2014-2015). Considering the very high number of samples, analyses were conducted in the same period for the same cultivar, following each other, compatibly with the time required for each test. The drupes harvested from each plant were milled within 24 hours, by using a mini olive press (Spremoliva C30 milling machine, Toscana Enologica Mori, Tavarnelle Val di Pesa, FI, Italy). About 25-30 kg of olives were processed in each working cycle, lasting approximately 2-3 hours: the machine was cleaned carefully after each cycle. The produced monocultivar olive oils were filtered and stored in sealed dark glass bottles at room temperature prior to analysis.

2.2 Genetic characterization

Molecular characterization was conducted on the 450 olive accessions in order to verify genetic correspondence with reference cultivars (Coratina, Ogliarola barese, Cima di Mola and Peranzana). Reference cultivars were collected from the CREA-OLI olive collection located in Mirto Crosia (CS, Italy). A set of highly discriminant microsatellite molecular markers was used (DCA3, DC5,

DCA8, DCA15, DCA18, GAPU71b, GAPU101) for this purpose [23]. Genomic DNA was extracted from young leaves, dried in silica gel for 2-3 days. An amount of 50 mg of dried leaf tissue was ground using Tissuelyser II (Qiagen, Hilden, Germany); DNA was extracted using the GenElute™ Plant Genomic DNA Miniprep Kit (Sigma-Aldrich, St. Louis MO, USA), and quantified at NanoDrop 2000 UV-Vis Spectrophotometer (Thermo Scientific, Waltham, MA, USA). PCR amplification was carried out using KAPA3G Plant PCR Kit (Kapa Biosystems, Wilmington, MA, USA). Amplicons were analyzed by using a 16-capillary DNA sequencer (3130 Applied Biosystems, Foster City, CA, USA), equipped with the GeneMapper 3.7v software. Two laboratories conducted the molecular analysis (CREA-OLI and DISSPA, University of Bari). To verify the reproducibility of the results, a ring test was conducted on 15 samples randomly chosen.

2.3 Merceological analysis

Twenty-three merceological parameters were measured and subsequently used as inputs for the ANNs: total tocopherols, total phenols, free acidity, peroxide value, UV spectrophotometric indices (K_{232} = UV absorbance at 232 nm, K_{270} = UV absorbance at 270 nm, $\Delta K = K_{232} - K_{270}$), NI R.T.6.9 and NI R.T. 7.4 (unidentified peaks), C14:0 (myristic acid), C16:0 (palmitic acid), C16:1is and C16:1c (isomers of palmitoleic acid), C17:0 (eptadecanoic acid), C17:1 (eptadecenoic acid), C18:0 (stearic acid), C18:1 (oleic acid), C18:2 (linoleic acid), C18:3 (linolenic acid), C20:0 (arachidic acid), C20:1 (eicosenoic acid), C22:0 (behenic acid), C24:0 (lignoceric acid). All the measurements were carried out according to the official methods of the European Regulation/Commission Regulation EEC no. 2568/91 and its subsequent modifications (EC Reg. 2568/1991) [24]. Moreover, the quantification of phenols and tocopherols was performed according to the method described by COI [25].

2.4 Near Infra-Red (NIR) Spectrometry

136 The Near Infra-Red (NIR) analysis was conducted using the XDS analyser instrument (Foss,
137 Analytical A/S, Denmark) equipped with an infrared reading system. The XDS NIR instrument is
138 supported by the RINA (Remote Internet Analysis) software suite. A dispersive grating
139 monochromator permits a highest signal/noise ratio, thus it guarantees an efficient analysis of
140 complex matrices and dispersions. For each olive oil sample, an aliquot was weighted in a 1.5 mL
141 quartz cuvettes for NIR analysis, without any preliminary treatment. Spectra were acquired at
142 constant temperature of 40°C, according to standard instrument procedures for olive oil analysis, in
143 the wavelength range from 700 nm to 2500 nm. Three measurements were performed for each
144 sample, in order to minimize errors due to instrumental fluctuations. Twenty parameters were
145 evaluated: acidity, peroxides, K_{232}, K_{270}, ΔK, palmitic acid, palmitoleic acid, eptadecanoic acid,
146 eptadecenoic acid, stearic acid, oleic acid, linoleic acid, linolenic acid, arachidic acid, eicosenoic
147 acid, polyphenols, tocopherols, methyl esters, ethyl esters, methyl ethyl esters.

148 **2.5 Nuclear Magnetic Resonance (NMR) Spectrometry**

149 The mono-cultivar EVOO samples, belonging to the four most representative varieties of Apulia,
150 collected during two crop years, were analyzed by ^{1}H NMR spectroscopy. Briefly, each NMR
151 sample was prepared dissolving ~140 mg of olive oil in $CDCl_3$ and adjusting the mass ratio of olive
152 oil:$CDCl_3$ to 13.5%:86.5%. Next, 600 µL of the prepared mixture was transferred into a 5-mm
153 NMR tube. ^{1}H NMR spectra were recorded on a Bruker Avance spectrometer (Bruker, Karlsruhe,
154 Germany), operating at 400.13 MHz, T = 300 K, equipped with a PABBI 5-mm inverse detection
155 probe incorporating a *z* axis gradient coil. NMR experiments were performed under full automation
156 for the entire process after loading individual samples on a Bruker Automatic Sample Changer
157 (BACS-60), interfaced with the software IconNMR (Bruker). Automated tuning and matching,
158 locking and shimming, and calibration of the 90° hard pulse P(90°) were done for each sample
159 using standard Bruker routines, ATMA, LOCK, TOPSHIM and PULSECAL, to optimize NMR
160 conditions. For each sample, after a 5-min waiting period for temperature equilibration, standard

one-dimensional (^{1}H ZG) NMR experiments were performed. The relaxation delay (RD) and acquisition time (AQ) were set to 4 s and ~3.98 s, respectively, resulting in a total recycle time of ~7.98 s. FIDs were collected into time domain (TD) = 65536 (64 k) complex data points by setting: spectral width (SW) = 20.5524 ppm (8223.685 Hz), receiver gain (RG) = 4, number of scans (NS) = 16. The accumulation of 16 scans was preferred because of some metabolites present in high concentrations [26].

The NMR raw data set was pre-processed using Topspin 2.1 and AMIX 3.9.15 (Bruker BioSpin GmbH, Germany). The FIDs were multiplied by an exponential line broadening function (0.3 Hz) before Fourier transformation and automatically phased. Spectra were referenced to the TMS signal at 0.00 ppm, used as an internal standard and obtaining good peak alignment.

We focused on the spectral region within 10.00-0.5 ppm, excluding the signal of the residual non-deuterated chloroform and its carbon satellites (7.6-6.9 ppm). This region was reduced in small intervals (buckets) of equal size (0.04 ppm) by applying the rectangular bucketing procedure, obtaining 221 buckets. Total sum normalization was applied to minimize small differences due to total olive oil concentration and/or acquisition conditions among samples. The Pareto scaling method, which is performed by dividing the mean-centered data by the square root of the standard deviation, was then applied to the variables. Each bucket (NMR variable) was ~~labeled~~ labelled with its average value of chemical shift (ppm), and subsequently used as ANN input. Multivariate analysis (Principal Component Analysis, PCA) of NMR bucket-reduced data was carried out using SIMCA 14 software (Umetrics, Umea, Sweden).

All chemical reagents for analyses were of analytical grade. Deuterated chloroform ($CDCl_3$ 99.8%-d) containing tetramethylsilane TMS (0.03% v/v) was purchased from Armar Chemicals (Döttingen, Switzerland).

2.6 Artificial Neural Networks

185 The most common artificial neural network is the Multi-Layer Perceptron (MLP), successfully
186 used also for classification and pattern recognition [27-29]. It is a universal function approximator
187 which can solve non-linearly separable problems and learn any arbitrarily complex linear function
188 with an arbitrary accuracy level [30,31]. In general, a MLP is composed by one input layer with p
189 inputs, one or more hidden layers with n hidden neurons, and one output layer with q outputs
190 (Figure 1). The output of the j-th hidden neuron is computed as

191 $$y_j^h = f^h\left(\sum_{k=1}^{p} w_{k,j}^h x_k + b_j^h\right),\ j = 1,\ldots,n,$$

192 where f^h is the activation function of the hidden neuron, $w_{k,j}^h$ is the weight between the input x_k
193 and the hidden neuron j, and b_j^h is a bias term for the hidden neurons. Then, considering a MLP
194 with one hidden layer, the i-th output is computed as

195 $$y_i = f^o\left(\sum_{j=1}^{n} w_{j,i}^o y_j^h + b_i^o\right),\ i = 1,\ldots,q,$$

196 where f^o is the activation function of the output neuron, $w_{j,i}^o$ is the weight between the hidden
197 neuron j and the output neuron i, and b_i^o is a bias term for the output neuron.

198 In this study, the Multi-Layer Perceptron model was used for olive oil cultivar classification, and
199 was configured as a pattern recognition network with sigmoidal and softmax activation functions
200 for hidden and output neurons, respectively [28]. The training process was performed using the
201 Scaled Conjugate Gradient algorithm to minimize the cross-entropy cost function. The dataset was
202 normalized with a zero mean and unit standard deviation transformation. Then, the dataset was
203 divided into training, validation, and test sets: the training set was used for the learning process, the
204 validation set was used during the learning process to avoid overfitting by adopting the early stop
205 strategy [29], and the test set was used to properly evaluate the classifier performance on an

independent data set. In addition, a 5-fold cross validation was performed to limit the bias of the model performance associated with a random sampling of the training data. This means that, 5 folds were randomly created with roughly the same cultivar proportion as in the initial dataset; then, 5 different MLP models were trained by using different folds for training, validation and test. We did many trials, varying the number of hidden layers from 1 to 2, the number of hidden neurons from 3 to 20, and re-training each network 1000 times with different random initial weights. Finally, we chose the best model based on the accuracy, i.e. a global criterion defined as the percentage of correct prediction in the dataset. Considering the confusion matrix (i.e., an $m \times m$ table that compares the classifier outputs with the actual values in the dataset), the accuracy was defined as:

$$a = \frac{TP + TN}{(TP + TN + FP + FN)},$$

where TP, TN, FP and FN are the true positive, true negative, false positive and false negative values, respectively. Then, since 5 MLP models were trained on different folds using the k-folds cross validation, the classifier performance was defined as the average value of the accuracies for the 5 trained MLP models.

The LASSO (*Least Absolute Shrinkage and Selection Operator*) algorithm [32,33], was used to identify the most informative predictors for the olive oil cultivars among all the predictors. Indeed, there were cases of data correlation/redundancy; for example, the NMR signals at 1.30 ppm, 2.82 ppm, and 0.9 ppm refer, respectively, to the methylenic, bis-allylic and methyl protons of the same molecule (linolenic acid). Thus, the LASSO algorithm was used to reduce the number of predictors in the regression models by selecting the most informative predictors, and to produce shrinkage estimates with potentially lower predictive errors than ordinary least squares.

In addition to the LASSO algorithm, two heuristic approaches based on the standard deviation of each predictor were also considered, to further investigate the importance of the predictors for the

229 olive oil cultivar. The rationale was that, an input with a smaller range of variability may be
230 expected to provide a lower influence on the ANN output compared to an input with a greater range
231 of variability. Thus, in the first heuristic approach we chose the predictors with greater standard
232 deviations (I), while in the second one we chose the predictors with greater standard deviations
233 normalized by their respective average values (II).

234 Different ANNs were trained using different combinations of variables: the full dataset, the
235 subsets of variables provided by LASSO, and the two subsets of variables selected according to
236 standard deviation and normalized standard deviation criteria (Table 1), in order to find the best
237 model for cultivar classification performance and to compare the prediction capability of the
238 different predictors among merceological, NIR, and NMR data. The MATLAB software was used
239 for LASSO analysis and neural networks training and validation.

240 **3. RESULTS**

241 **3.1 Genetic data**

242 The set of microsatellites used for the molecular characterization (DCA3, DC5, DCA8, DCA15,
243 DCA18, GAPU71b, GAPU101) discriminated efficiently all the analyzed accessions, showing a
244 unique molecular profile corresponding to the four reference cultivars (Table S1). Molecular data
245 were highly comparable between CREA-OLI and DISSPA laboratories, except at the DCA9
246 (182/194 vs 172/186) and DCA18 (175/177 vs 177/179) loci for 'Peranzana' and 'Ogliarola'
247 accessions, respectively, where allele assignations were different. Ring test conducted on 15
248 accessions randomly chosen confirmed the reproducibility of the analysis. Genetic analysis showed
249 a discrepancy for only 6 out of 450 accessions, which were excluded from the study. Among them,
250 four accessions were riconducible to Apulian varieties (Simona, Pasola and Cima di Melfi), one was
251 different from the reference profile at two loci (DCA9, DCA18), and the last one did not correspond

252 to any known cultivar (Table S1). In conclusion, 444 olive oil samples of each crop year were used
253 for further analyses.

254 **3.2 Merceological data**

255 Average and standard deviation of all merceological parameters are reported in Table S2.
256 Coratina samples presented the highest content of phenols, higher than the minimum level
257 (>250mg*Kg^{-1}) necessary to boast healthy claims, as established by Reg. CEE 2568/91 and its
258 subsequent modifications [24]. Another interesting aspect of Coratina oils concerned the content of
259 eicosenoic acid, that was higher than the maximum value fixed at 40% by regulations in 22% of
260 samples.

261 **3.3 NIR data**

262 Results obtained through NIR spectroscopy are summarized in Table S3. NIR profiles, recorded
263 in both crop years, showed that Coratina samples were enough different from samples belonging to
264 the other three cultivars. Coratina oils presented the highest content of oleic and eicosenoic acids
265 and the lowest content of all the other fatty acids and total tocopherols. The average values of
266 acidity, peroxides, K_{232} and K_{270}, and ΔK of all samples, were compatible with the "extra-virgin"
267 definition, based on the limits fixed by the Commission Regulation EC. No 1989/2003 [34].

268 **3.4 NMR data**

269 ^{1}H NMR profiling of olive oils dissolved into deuterated chloroform is a well-established
270 technique in metabolomics and for classification of olive oil cultivars [6,35]. NMR spectroscopy
271 combines targeted and non-targeted analysis within one single measurement and provides a
272 remarkable level of reproducibility of the data. Moreover, due to a highly reproducible and very
273 detailed fingerprinting, it is possible to differentiate samples even if only small changes occur.
274 Determination of both fatty acid profile and unsaponifiable fraction is usually obtained from proton

^{1}H NMR spectrum according to literature data [26]. The olefinic protons –C*H*=C*H*– of all unsaturated fatty acids were assigned at 5.4-5.3 ppm, the proton signal at 5.14 ppm was assigned to >C*H*OCOR of sn 1,2 DGs; bis-allylic protons (=CHCH_2CH=) of linolenic and linoleic acids were assigned at 2.85-2.70 ppm, the methylene (CH_2) protons of at 1.2 ppm, and the terminal methyl group protons of all saturated and unsaturated chains at 1.0-0.8 ppm. Signals in the range between 4.75 and 4.55 ppm referred to different terpenes, while protons of aldehydes and phenolic compounds resonate in the range 9.7-9.1 and 7.0-5.6 ppm, respectively.

Interestingly, among the 221 buckets constituting the reduced NMR spectrum of each sample, only 24 buckets were selected on the basis of their ability to discriminate between cultivars, and then used for training ANNs (see Section 3.5). A preliminary work on a MVA analysis related to the complete NMR data set (221 buckets for 900 samples) has been already reported [36]. On the first attempt, in order to reveal a general data grouping of all the samples, an unsupervised PCA analysis was applied to the whole data (^{1}H NMR-bucket-reduced spectra). Visual inspection of three dimensional PCA scoreplot, reported in Figure 2, showed a certain degree of separation in particular for the Coratina samples, while a certain degree of overlap was observed among the three remaining classes, Cima Di Mola, Ogliarola and Peranzana.

3.5 ANN data

We exploited the ANN methodology for cultivar classification of 888 mono-cultivar olive oil samples obtained from the two crop years. Different sets of data, chosen according to the four criteria described in Section 2.6, were used to train ANNs, in order to find the most accurate ANN model, and consequently the most informative analytical technique (Table 1).

We recorded globally 43 “traditional” variables for merceological and NIR analyses and 221 “innovative” variables for NMR analysis. ANN models were trained using data gathered from the

two crop years, and were validated on an independet test set composed by 176 samples not used during the training process. They showed similar optimal performances independently on the data type. In fact, the values of classification accuracy (i.e., the prediction rate on the independent test set) of the ANNs trained with all 43 merceological and NIR variables and the ANNs trained with all 221 NMR variables were 98.9% and 99.0%, respectively (Table 2). Moreover, ANN models showed similar complexity: the ANNs trained with "traditional" variables and ANNs trained with "innovative" variables were composed by 2 layers with 20 and 17 hidden neurons, respectively (Table 2).

The trained ANN models were also used to independently classify the complete dataset of the two crop years. The classification accuracy remained high independently on the crop year. In fact, the values of classification accuracy of the ANNs trained with "traditional" variables and ANNs trained with "innovative" variables were 99.5% and 99.4% for the first crop (2013-2014) and 99.6% and 99.7% for the second crop (2014-2015), respectively (Table 3).

The application of the LASSO algorithm allowed us to select a subset of 29 variables for merceological and NIR analyses (Table 1). The ANNs trained with this subset performed as well as the ANNs trained with all 43 traditional variables, having values of accuracy about 99% (Table 2, and 3). This means that, the findings removed by the LASSO algorithm were not informative for the purpose of cultivar classification of olive oils. Analogously, ANNs created using the subset of 24 NMR variables selected by LASSO presented a classification capability similar to that of ANNs created using all 221 NMR variables (Table 2, and 3). Hence, only few regions of the ^{1}H-NMR spectrum obtained from an olive oil sample were very informative about its cultivar. Moreover, reducing the number of predictors, the complexity of the ANN models was also reduced, from 2 layers with 20 hidden neurons to 2 layers with 13 hidden neurons for traditional variables and from 2 layers with 17 hidden neurons to 1 layer with 16 hidden neurons (Table 2).

Since inputs with different variability may have different influence on the ANN outputs, we chose other two subsets of variables to be used for ANN training based on two criteria: (I) greater standard deviations or (II) greater standard deviations normalized by their respective average values. According to (I), we selected 3 traditional and 5 innovative variables, whereas according to (II) we selected 4 traditional and 10 innovative variables (Table 1). In both cases, ANN models trained with the subsets of merceological and NIR data presented significantly lower accuracy in cultivar classification than models trained with NMR data, i.e. 80.5% and 91.6% using (I), and 60.8% and 97.5% using (II), respectively (Table 2). Similar trends were also observed when using the full dataset of the two independent crop years: performances ranged from 57.5% to 81.6% for "traditional" data and from 92.4% to 99.5% for "innovative" variables (Table 3).

3.6 Validation on industrially produced olive oils

The ANNs described above, that were obtained using data from monocultivar olive oils produced at a laboratory-scale, were tested on a group of 17 monocultivar olive oils industrially produced, belonging to the four cultivar considered. For the 17 testing samples, we recorded all the same merceological, NIR, and NMR variables, as done with the previuos collection of 888 olive oils. The aim of this step was to verify if the trained ANNs could work with commercialized (large-scale produced) olive oils as well as with oils obtained by mini olive press, despite of differences in operating conditions (volume, instrumentation, storage, etc…). Table 4 shows the values of accuracy in assigning the cultivar of the testing samples relative to the trained ANN models.

The ANN model built using the subset of 24 NMR variables, found by application of LASSO algorithm, correctly classified 15 out of 17 testing samples, with the highest accuracy (88.2%) among the compared ANN models.

4. DISCUSSION

In our study, for the first time, ANN models were set up using multiple types of information, namely standard merceological parameters, NIR data ~~profiles~~, and NMR fingerprints, in order to find the most accurate ANN model for cultivar classification. For training and validation of the ANNs, we used data from 888 mono-cultivar olive oil samples produced at a laboratory scale, belonging to four varieties (ascertained by genetic analysis), and collected during two crop years. The ANNs were also tested on a smaller dataset composed by 17 samples industrially produced, to verify if the trained ANNs could work with these as well, despite of differences in operating conditions (volume, instrumentation, storage, etc…).

Overall, the ANNs seems to be an excellent approach to classify olive oils according to cultivar; in particular, the ANN models based on NMR data shows the best performance, when considering the samples produced either at a laboratory scale and at a large scale.

In fact, the high values of accuracy reported for ANN models built with NMR data (in some cases > 99%) suggest that, NMR spectroscopy supplied a quantity of information, useful for cultivar classification, similar or higher than merceological analysis and NIR spectroscopy together. Moreover, the method of olive oil milling has a weak influence on performances of ANNs trained with NMR data (Table 4). Most information are contained in narrow regions of the entire NMR spectrum, namely the 24 variables extracted by LASSO. Consequently, they are the most suitable attibutes for classification of samples according to the cultivars examined here. In details, these spectral regions comprise the peaks of phenolic compounds (NMR resonances at 6.86 6.74, 6.58, 6.30 ppm), aldehydes (9.06, 7.98, 7.94 ppm), acyl groups of all TGs (triglycerides, 5.34 ppm), acyl groups of *sn* 1,2 DGs (diglycerides, 5.14, 4.94, 3.70 ppm), fatty acids such as linoleic and linolenic acids (2.82, 2.78, 1.30 ppm), and cycloartenol (0.58 ppm). Our results confirm previuos literature findings. Indeed, the content of different polyphenols and aldehydes has been largely associated with the organoleptic properties typical of a cultivar [2]. Differences in the level of saturated and unsaturated fatty acids have been observed among cultivars [37]. Cycloartenol, that is a triterpenoid of the sterol class, has already been found to be affected by cultivar, as well as triterpene alcohol

composition and total triterpene alcohol content [38]. On the other hand, the level of *sn* 1,2 DGs or, more correctly, the *sn* 1,2 DGs / *sn* 1,3 DGs ratio, has been correlated with the degree of lipid degradation due to the activity of hydrolytic enzymes, that prevalently depends on oil production and storage processes [39].

In conclusion, we assess that the combination of NMR fingerprinting with ANN modelling could provide an effective, robust, and rapid method for classifying olive oil cultivars.

CONCLUSIONS

The application of ANNs evidence that NMR data showed the highest capability to classify cultivars ~~are the most informative variables~~. ANN modelling of NMR data provide an accurate and efficient approach for cultivar classification of olive oils and prediction of unknown samples, independently from methods of milling and year of production. In addition, it is known that NMR technique is advantageous in terms of required time and costs (especially compared to traditional merceolgical methods). As well, the ANN approach presents its own advantages compared to other chemometric techniques, e.g. it is a non-linear method, fitting better to the data; no particular manipulation of raw data is needed; the MLP design is relatively simple, presenting connections in parallel and sequence between neurons; it learns and does not need reprogramming, thus implementation is not difficult; it can handles data of different origins more easily than using other approaches.

However, the limit of the ANN models is that they are applicable exclusively to monovarietal samples belonging to the four varieties (Coratina, Ogliarola, Cima di Mola, Peranzana) used for their training.

Acknowledgments

This study was supported by a grant from the Italian Ministry of University and Research-MIUR (PON01_01958 PIVOLIO). We gratefully acknowledge Oliveti Terra di Bari (Bari, Italy) and Olearia Basile (Andria, BT, Italy) farms for sample supply.

References

[1] G. Sindona, A Marker of Quality of Olive Oils: The Expression of Oleuropein, Olives and Olive Oil in Health and Disease Prevention (2010) 95–100.

[2] P. Dais, E. Hatzakis, Quality assessment and authentication of virgin olive oil by NMR spectroscopy: A critical review, Anal. Chim. Acta. 765 (2013) 1–27.

[3] L. Del Coco, S.A. De Pascali, V. Iacovelli, G. Cesari, F.P. Schena, F.P. Fanizzi, Following the olive oil production chain: 1D and 2D NMR study of olive paste, pomace, and oil, Eur. J. Lipid Sci. Technol. 116 (2014) 1513–1521.

[4] C. Lucas-Torres, A. Pérez, B. Cabañas, A. Moreno, Study by ^{31}P NMR spectroscopy of the triacylglycerol degradation processes in olive oil with different heat-transfer mechanisms, Food Chem. 165 (2014) 21–8.

[5] R.I.M. Almoselhy, M.H. Allam, M.H. El-Kalyoubi, A.A. El-Sharkawy, ^{1}H NMR spectral analysis as a new aspect to evaluate the stability of some edible oils, Ann. Agric. Sci. 59 (2014) 201–206.

[6] F. Longobardi, A. Ventrella, C. Napoli, E. Humpfer, B. Schütz, H. Schäfer, et al., Classification of olive oils according to geographical origin by using ^{1}H NMR fingerprinting combined with multivariate analysis, Food Chem. 130 (2012) 177–183.

[7] M. Casale, N. Sinelli, P. Oliveri, V. Di Egidio, S. Lanteri, Chemometrical strategies for feature selection and data compression applied to NIR and MIR spectra of extra virgin olive oils for

416 cultivar identification., Talanta. 80 (2010) 1832–7.

417 [8] M. Casale, R. Simonetti, Review: Near infrared spectroscopy for analysing olive oils, J.
418 Near Infrared Spectrosc. 22 (2014) 59–80.

419 [9] A. Bajoub, A. Carrasco-Pancorbo, E.A. Ajal, N. Ouazzani, A. Fernández-Gutiérrez,
420 Potential of LC-MS phenolic profiling combined with multivariate analysis as an approach for the
421 determination of the geographical origin of north Moroccan virgin olive oils, Food Chem. 166
422 (2015) 292–300.

423 [10] P. Agozzino, G. Avellone, D. Bongiorno, L. Ceraulo, S. Indelicato, S. Indelicato, et al.,
424 Determination of the cultivar and aging of Sicilian olive oils using HPLC-MS and linear
425 discriminant analysis, J. Mass Spectrom. 45 (2010) 989–995.

426 [11] M.S. Cosio, D. Ballabio, S. Benedetti, C. Gigliotti, Geographical origin and authentication
427 of extra virgin olive oils by an electronic nose in combination with artificial neural networks, Anal.
428 Chim. Acta. 567 (2006) 202–210.

429 [12] L.G. Dias, A. Fernandes, A.C. a Veloso, A. a S.C. Machado, J. a. Pereira, A.M. Peres,
430 Single-cultivar extra virgin olive oil classification using a potentiometric electronic tongue, Food
431 Chem. 160 (2014) 321–329.

432 [13] E. Funes, Y. Allouche, G. Beltrán, A. Jiménez, A Review: Artificial Neural Networks as
433 Tool for Control Food Industry Process, J. Sens. Technol. 5 (2015) 28–43.

434 [14] J. Zupan, J. Gasteiger, Neural networks: A new method for solving chemical problems or
435 just a passing phase?, Anal. Chim. Acta. 248 (1991) 1–30.

436 [15] R. Goodacre, D.B. Kell, Pyrolysis mass spectrometry and its applications in biotechnology,
437 Curr. Opin. Biotechnol. 7 (1996) 20–8.

438 [16] F. Marini, A.L. Magrì, R. Bucci, A.D. Magrì, Use of different artificial neural networks to
439 resolve binary blends of monocultivar Italian olive oils, Anal. Chim. Acta. 599 (2007) 232–40.

440 [17] D.L. García-González, L. Mannina, M. D'Imperio, A.L. Segre, R. Aparicio, Using ^{1}H and
441 ^{13}C NMR techniques and artificial neural networks to detect the adulteration of olive oil with
442 hazelnut oil, Eur. Food Res. Technol. 219 (2004) 545–548.

443 [18] S. Rezzi, D.E. Axelson, K. Héberger, F. Reniero, C. Mariani, C. Guillou, Classification of
444 olive oils using high throughput flow ^{1}H NMR fingerprinting with principal component analysis,
445 linear discriminant analysis and probabilistic neural networks, Anal. Chim. Acta. 552 (2005) 13–24.

446 [19] F. Marini, Artificial neural networks in foodstuff analyses: Trends and perspectives A
447 review, Anal. Chim. Acta. 635 (2009) 121–131.

448 [20] S.F. Silva, C.A.R. Anjos, R.N. Cavalcanti, R.M.D.S. Celeghini, Evaluation of extra virgin
449 olive oil stability by artificial neural network, Food Chem. 179 (2015) 35–43.

450 [21] R. Bucci, A.D. Magrí, A.L. Magrí, D. Marini, F. Marini, Chemical authentication of extra
451 virgin olive oil varieties by supervised chemometric procedures, J. Agric. Food Chem. 50 (2002)
452 413–418.

453 [22] A.M. Peres, P. Baptista, R. Malheiro, L.G. Dias, A. Bento, J.A. Pereira, Chemometric
454 classification of several olive cultivars from Trás-os-Montes region (northeast of Portugal) using
455 artificial neural networks, Chemom. Intell. Lab. Syst. 105 (2011) 65–73.

456 [23] V. Alba, C. Montemurro, W. Sabetta, A. Pasqualone, A. Blanco, SSR-based identification
457 key of cultivars of Olea europaea L. diffused in Southern-Italy, Sci. Hortic. (Amsterdam) 123
458 (2009) 11–16.

459 [24] European Commission, Regulation No 2568/91 on the characteristics of olive oil and olive-

residue oil and on the relevant methods of analysis (1991).

[25] International Olive Council, COI/T.20/Doc No 29 Determination of biophenols in olive oils by HPLC (2009) 1–8.

[26] A. Barison, C.W.P. da Silva, F.R. Campos, F. Simonelli, C.A. Lenz, A.G. Ferreira, A simple methodology for the determination of fatty acid composition in edible oils through ^{1}H NMR spectroscopy, Magn. Reson. Chem. 48 (2010) 642–650.

[27] G.P. Zhang, Neural networks for classification: a survey, IEEE Trans. Syst. Man Cybern. Part C Applications, 30 (2000) 451–462.

[28] C.M. Bishop, Neural networks for pattern recognition, 2005.

[29] C.M. Bishop, Pattern Recognition and Machine Learning, 2006.

[30] K. Hornik, M. Stinchcombe, H. White, Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks, Neural Networks. 3 (1990) 551–560.

[31] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Networks 4 (1991) 251–257.

[32] R. Tibshirani, R.S. Society, Regression and shrinkage via the LASSO, J R Stat Soc, Ser B. 58 (1996) 267–288.

[33] R. Tibshirani, Regression shrinkage and selection via the LASSO: A retrospective, J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (2011) 273–282.

[34] European Commission, Regulation No 1989/2003 amending Regulation No 2568/91 on the characteristics of olive oil and olive-pomace oil and on the relevant methods of analysis (2003).

[35] S. Piccinonna, R. Ragone, M. Stocchero, L. Del Coco, S.A. De Pascali, F.P. Schena, et al.,

481 Robustness of NMR-based metabolomics to generate comparable data sets for olive oil cultivar
482 classification. An inter-laboratory study on Apulian olive oils, Food Chem. 199 (2016) 675–683.

483 [36] Fanizzi, F.P., Del Coco, L., Girelli, C.R., De Pascali, S.A. Harvest year effects on Apulian
484 EVOOs evaluated by 1H NMR based metabolomics. IV International Conference on Foodomics,
485 Cesena (FC), Book of abstract (2015) 31–32.

486 [37] R.M. Alonso-Salces, K. Héberger, M. V. Holland, J.M. Moreno-Rojas, C. Mariani, G.
487 Bellan, et al., Multivariate analysis of NMR fingerprint of the unsaponifiable fraction of virgin olive
488 oils for authentication purposes, Food Chem. 118 (2010) 956–965.

489 [38] R. Aparicio, G. Luna, Characterisation of monovarietal virgin olive oils, Eur. J. Lipid Sci.
490 Technol. 104 (2002) 614–627.

491 [39] B. Nieva-echevarría, E. Goicoechea, M.J. Manzanos, M.D. Guillén, A method based on 1 H
492 NMR spectral data useful to evaluate the hydrolysis level in complex lipid mixtures, FRIN. 66
493 (2014) 379–387.

494

495 **Table 1**

496 Feature selection results for merceological, NIR and NMR data (L acid: linoleic acid; Ln acid:
497 linolenic acid; DGs: diglycerides; TGs: triglycerides).

Data	Merceological + NIR		NMR
Full dataset	43 variables (23 + 20 variables)		221 variables
LASSO selection	29 variables		24 variables
	merc_tocopherols_tot	NIR_acidity	NMR_906, 798, 794 (aldehydes)
	merc_phenols_tot	NIR_peroxide	NMR_686, 674, 658, 630 (phenols)
	merc_peroxide	NIR_K232	NMR_534 (TGs C*H*=C*H)*
	merc_K232	NIR_K270	NMR_514, 494 (*sn* 1,2 DGs >C*H*OCOR)
	merc_K270	NIR_palmitoleic_acid	NMR_466 (terpenes)
	merc_C14	NIR_eptadecanoic_acid	NMR_390
	merc_C16_1IS	NIR_eptadecenoic_acid	NMR_370 (*sn* 1,2 DGs CH_2OH)

	merc_C16_1C merc_C17_0 merc_C17_1 merc_C18_1 merc_C20_0 merc_C24_0	NIR_stearic_acid NIR_oleic_acid NIR_linoleic_acid NIR_linolenic_acid NIR_eicosenoic_acid NIR_tocopherols NIR_methyl_esters NIR_ethyl_esters NIR_methylethyl_esters	NMR_326 NMR_282 (Ln acid =CHCH_2CH=) NMR_278 (L acid =CHCH_2CH=) NMR_254 NMR_218 NMR_198 (Acyl groups CH_2CH=CH) NMR_190 (Acyl groups OCOCH_2CH_2) NMR_158 (Acyl groups OCOCH_2CH_2) NMR_154 (Acyl groups OCOCH_2CH_2) NMR_130 (CH_2 Ln and L acids) NMR_058 (CH_2 cycloartenol)
Selection with standard deviation criterion	3 variables: merc_tocopherols_tot merc_phenols_tot NIR_tocopherols		5 variables: NMR_130 (CH_2 Ln and L acids) NMR_198 (Acyl groups CH_2CH=CH) NMR_534 (TGs C*H*=C*H*) NMR_278 (L acid =CHCH_2CH=) NMR_158 (Acyl groups OCOCH_2CH_2)
Selection with normalized standard deviation criterion	4 variables: merc_C24_0 merc_C17_0 merc_C17_1 NIR_ethyl_esters		10 variables: NMR_058 (CH_2 cycloartenol) NMR_278 (L acid =CHCH_2CH= NMR_514 (*sn* DGs 1,2 >C*H*OCOR) NMR_190 (Acyl groups OCOCH_2CH_2) NMR_370 (*sn* 1,2 DGs CH_2OH) NMR_154 (Acyl groups OCOCH_2CH_2) NMR_674 (phenolic compounds –Ph-*H*) NMR_658 (phenolic compounds –Ph-*H*) NMR_282 (Ln acid =CHCH_2CH=) NMR_198 (Acyl groups CH_2CH=CH)

498
499
500

501 **Table 2**

502 Architecture of the best ANNs and accuracy computed on the independent test set composed by
503 samples of both the two crop years.

Data	Merceological + NIR	NMR
Full dataset	98.9% 2 layers with 20 hidden neurons	99.0% 2 layers with 17 hidden neurons
LASSO selection	98.9% 2 layers with 13 hidden neurons	98.3% 1 layer with 16 hidden neurons
Selection with standard deviation criterion	80.5% 1 layer with 16 hidden neurons	91.6% 1 layer with 12 hidden neurons

Selection with normalized standard deviation criterion	60.8% 2 layers with 12 hidden neurons	97.5% 1 layer with 15 hidden neurons

504

505

506

507 **Table 3**

508 Accuracy for the best ANNs evaluated on the full dataset of the two crop years.

Data	**Merceological + NIR**		**NMR**	
Crop year	2013/2014	2014/2015	2013/2014	2014/2015
Full dataset	99.5%	99.6%	99.4%	99.7%
LASSO selection	99.7%	99.2%	98.8%	99.5%
Selection with standard deviation criterion	81.6%	79.6%	92.4%	92.6%
Selection with normalized standard deviation criterion	57.5%	64.0%	97.8%	99.5%

509

510

511

512 **Table 4**

513 Accuracy for the best ANNs evaluated on the independent test of industrially produced samples.

Data	Merceological + NIR	NMR
Full dataset	47.0%	82.3%
LASSO selection	52.9%	88.2%
Selection with standard deviation criterion	47.0%	58.8%
Selection with normalized standard deviation criterion	23.5%	76.5%

514

515 **Figure Captions**

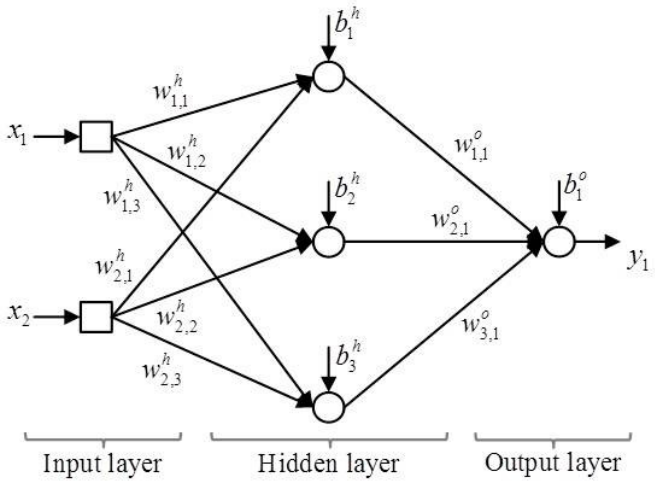
516 **Figure 1**. An example of MLP with 2 input neurons, 3 hidden neurons in the hidden layer and 1
517 output neuron

518 **Figure 2.** 3D PCA scoreplot for ^{1}H NMR-bucket-reduced spectra of monovarietal EVOO
519 samples, showing the general data grouping of all the samples.

520

521

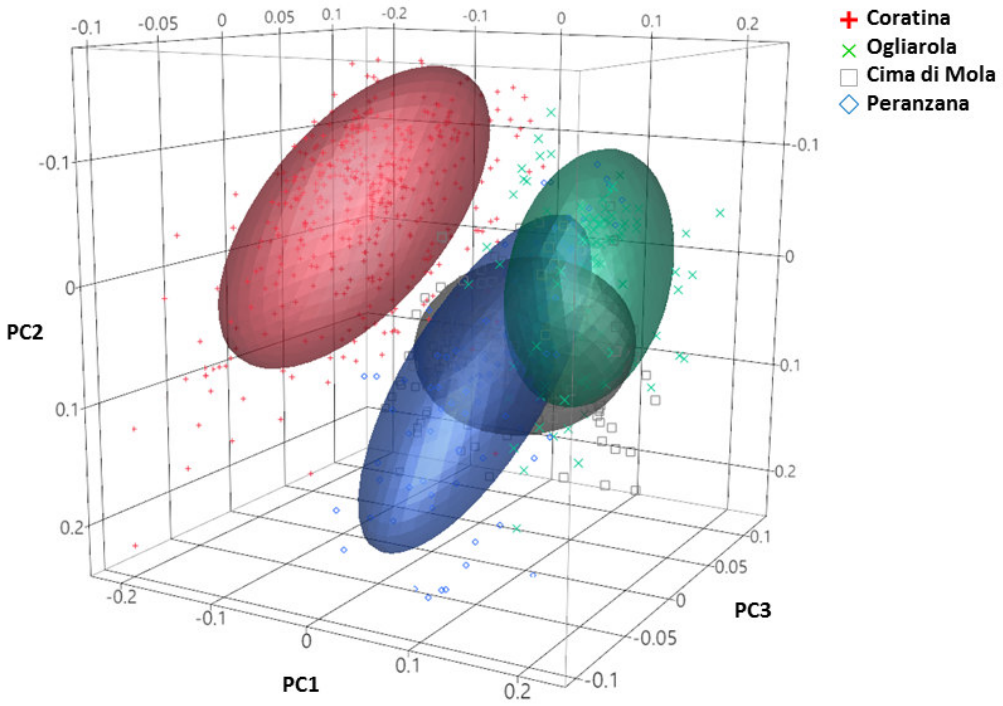
522



523

524

525



526

527

528

1. Cultivar discriminating ability of ANNs on Apulian monocultivar EVOOs was studied.
2. Merceological, NIR and ^{1}H NMR data were used as ANNs training sets.
3. ANN models based on NMR data showed the highest accuracy in classifying cultivars.
4. The most information about cultivars was contained in very few NMR peaks.
5. Performance was not influence by the milling method nor the crop year.