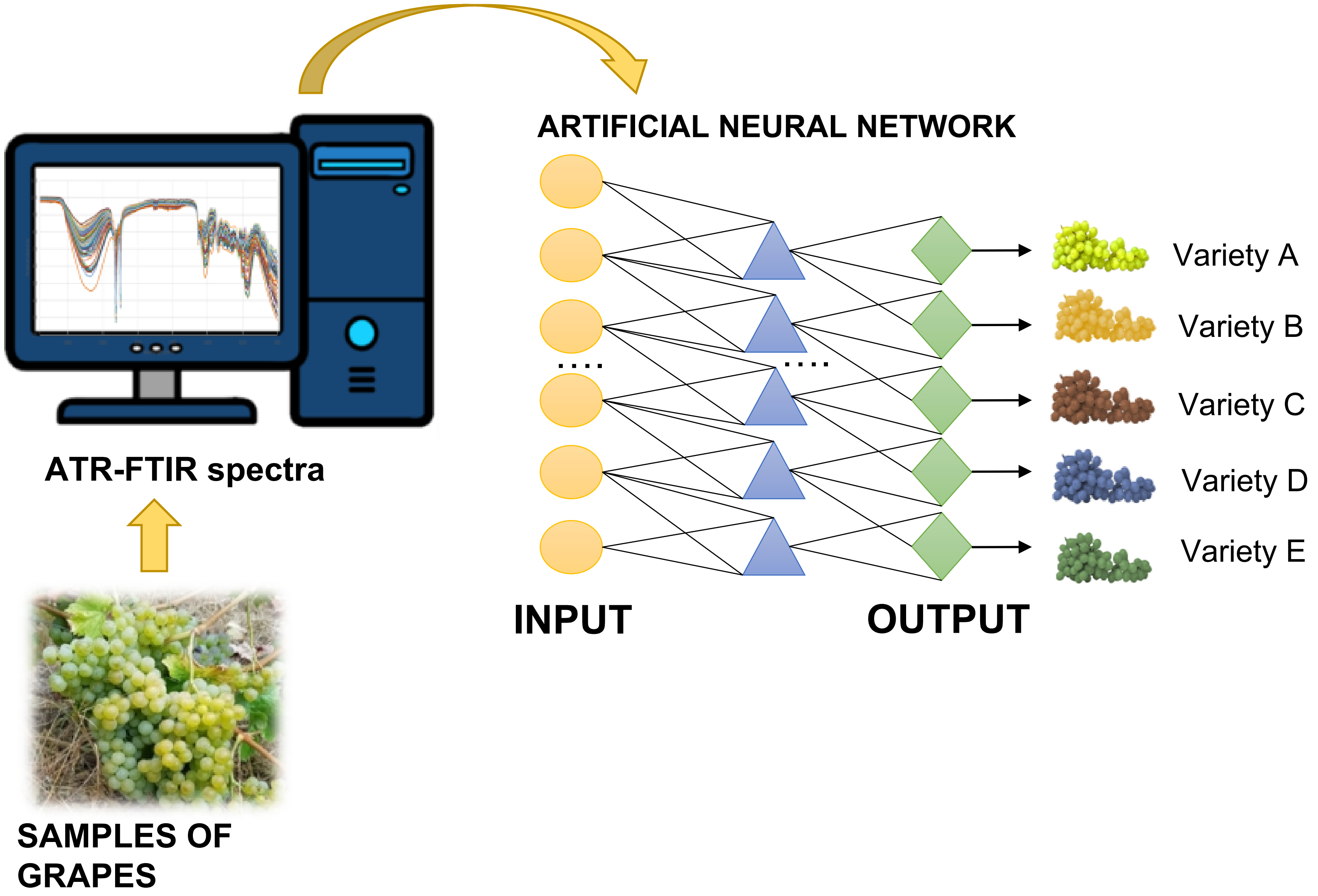
Highlights

- Fast Fourier Infrared (FTIR) analysis in grape skin provided information enough for the identification of the sample variety using Artificial Neural Networks (ANN);
- Attenuated total reflectance (ATR) allows recording spectra very fast without sample pre-treatment avoiding undesired structural changes of the samples;
- ANN together with Olden's Connection Weight Algorithm allowed identifying the principal compounds influencing the classification and ripeness;
- Pectin, polysaccharides and specially fructose, have the strongest influence in class and ripeness identification;


ARTIFICIAL NEURAL NETWORK
ATR-FTIR spectra
SAMPLES OF GRAPES
INPUT
OUTPUT
Variety A
Variety B
Variety C
Variety D
Variety E

Artificial Neural Network and Attenuated Total Reflectance-Fourier Transform Infrared Spectroscopy to identify the chemical variables related to ripeness and variety classification of grapes for Protected Designation of Origin wine production

Clarissa Murru, Christian Chimeno-Trinchet, Marta Elena Díaz-García, Rosana Badía-Laíño, Alfonso Fernández-González

Departamento de Química Física y Analítica, Universidad de Oviedo
C/ Julián Clavería s/n, 33006, Oviedo, Spain
Corresponding author: fernandezgalfonso@uniovi.es

Abstract

The vineyard grown in the territories included in the Protected Designations of Origin (PDO) classification of the European Union, present unique organoleptic properties of colour, aroma and flavour. Development of techniques for identifying grape varieties or ripeness among other characteristics, are key interesting for the PDO control and quality.

Attenuated total reflectance (ATR) allows fast recording spectra without sample pre-treatment, thus avoiding undesired physical and/or chemical changes of the sample. This method works in a rapid, non-destructive and easy-to-use way. The fast-fourier transform infrared spectroscopy (FTIR) analysis of five grape varieties (*Alabarín blanco, Mencía, Verdejo negro, Albarín negro* and *Carrasquín*) used for wine production of PDO *Vino de Cangas* provided information enough for the identification of grape class using artificial neural networks (ANN).

Despite the statistical similitude of the FTIR spectra among different grapes and maturity state, ANN resulted to be a helpful tool for classifying grape samples according to the variety or to their ripeness degree. Furthermore, compounds present in grapes that can most influence such classification can be outlined from the ANN. In this context, pectin and polysaccharides are especially significant in variety and ripeness identification,

whereas polyphenols and fructose provide useful information for ripeness degree classification of grapes.

Keywords: Artificial Neural Networks; Grapes; Connection Weight Algorithm; ATR-FTIR

1. Introduction

Commonly, viniculture is restricted to territories where the exposition to the sun lasts for long periods of the year. This makes the countries around the Mediterranean Sea outstanding places for grape culture and wine industry, thus becoming the most famous producers and exporters of wine. Spain is known for being a sunny country which dedicates huge extensions of terrain to viniculture, having seventy-five different Protected Denominations of Origin (PDO) for wine.

Polyphenolic compounds constitute an important aspect in the quality of grapes and wines and can be found in high concentrations in the skin of fruits, having important and different roles as secondary metabolites [1]. Polyphenolic compounds can be divided into two groups: non-flavonoid (hydroxybenzoic and hydroxycinnamic acids and stilbenes) and flavonoid compounds (anthocyanins, flavan-3-ols and flavonols). Among the flavonoid compounds, anthocyanins are the family of polyphenols responsible for colour in grapes and young wines, while flavan-3-ols (monomeric cathechins and proanthocyanidins) are mainly responsible for the astringency, bitterness and structure of wines [2]. For its part, flavonols (quercetin, myricetin, kaempferol, isorhamnetin and their glycosides), contribute to bitterness. In grape berries, flavonols are the most abundant phenolic compounds in grape skins, while grape seeds are rich in flavan-3-ol [3]. The concentration of phenolic compounds in grapes depends on the variety of grapevine and it is influenced by viticultural and environmental factors [4].

Another important group of chemicals that provide useful information for characterization of different varieties of fruit are those located in the skin cell wall. The primary cell wall of plants mainly consists of various polysaccharides (pectins, hemicelulloses and cellulose) and comparably, smaller amounts of structural glycoproteins, phenolic esters, minerals and enzymes [5]. Plant cell walls and their constitutive polysaccharide networks are vital with regard to the mechanical properties of the plant organ, such as stiffness or strength.

Chemometric techniques coupled to Near Infrared (NIR), FTIR or ATR-FTIR have been successfully applied for identifying plant leaves [6], for studying adulteration of cumin seed oil [7] or grape nectars [8], for determining the geographic origin of chardonnay grapes [9], for classifying different brands of fruit wines [10] or for identifying apples used in the production of cider [11].

The aim of the present work is to use the absorption bands in the mid-IR region, which reveals information about the type of molecules present in the grape skins in a fast, powerful and non-destructive way. The basis of the measurements relies on the wavelength-dependent interaction of light with the skin grape components. The FTIR technique, coupled with the use of chemometric procedures to extract the information from the IR spectrum [10,11], provides an accurate, reliable method suitable for discriminating grape varieties despite the quite similar composition of their skins. Also, FTIR-chemometrics, may provide important information for assessing ripeness degree classification of grapes.

Results obtained demonstrate that FTIR coupled to chemometrics allows the consistent identification of several grape varieties used for the production of *PDO Vino de calidad de Cangas (Wine Cangas Quality)*, which must be exclusively made with the admitted and/or authorized grape varieties, as the listed in the legislation (Table 1) [12]. The sample grapes used in this work come from a small vineyard in northern Spain (Cangas de Narcea, Asturias), endowed with an especial microclimate and soil suitable for viniculture. The wines derived from this vineyard own PDO according to the classification of the European Union [12, 13] and present unique organoleptic characteristics in terms of colour, aroma and flavour, looking clean, bright and a right alcohol / acidity balance.

Table 1 *Varieties of grapes allowed in PDO ''Vino de Calidad de Cangas''. The varieties used in this study appear in bold.*

Accepted varieties	**Albarín blanco**
	Albillo
	Garnacha tintorera
	Mencía
	Picapoll blanco
	Extra
	Verdejo negro
Authorized varieties	**Albarín negro**
	Carrasquín
	Godello
	Gewurztraminer
	Merlot
	Pinot noir
	Syrah

2 Materials and methods

2.1 *Grape samples and leaves collection*

Grapes of five varieties *Albarín blanco (AB), Mencía (MN), Verdejo negro (VN), Albarín negro (AN)* and *Carrasquín (CQ)* were kindly provided by "Bodegas Vidas" cellar. Every week (along 3 weeks) three different clusters of three different plants (nine clusters) were collected for every variety. Three different grapes were collected from every cluster, yielding 27 grapes per variety and week (a total of 135 grapes per week). During the third week, grapes from varieties AB and VN could not be collected due to the industrial needs of the vineyard.

Leaves for every variety were collected every week along three weeks. A single leaf was taken every time from three different plant to avoid its further damage.

2.2 Instrumentation

A Varian 670-IR spectrometer equipped with a DLaTGS detector and a diamond-based Golden Gate ATR device was used for all the measurements. Mathematical data processing and calculations were performed with MatLab R2018a from Mathworks.

2.3 *Measurement protocol*

Grapes were thoroughly washed with distilled water prior to analysis. A thin skin layer was cut using a scalpel and the external part brought into close contact with the ATR diamond. Every grape skin was sampled three times and its spectrum was recorded from 600 cm^{-1} to 4000 cm^{-1} with resolution 4 cm^{-1} (average of 16 scans). A final number of 1053 spectra were recorded. Leaves were analysed without any previous treatment taking the FTIR spectra in different points of their surface. A total number of 135 spectra were recorded. Unused grapes and leaves were frozen for future needs.

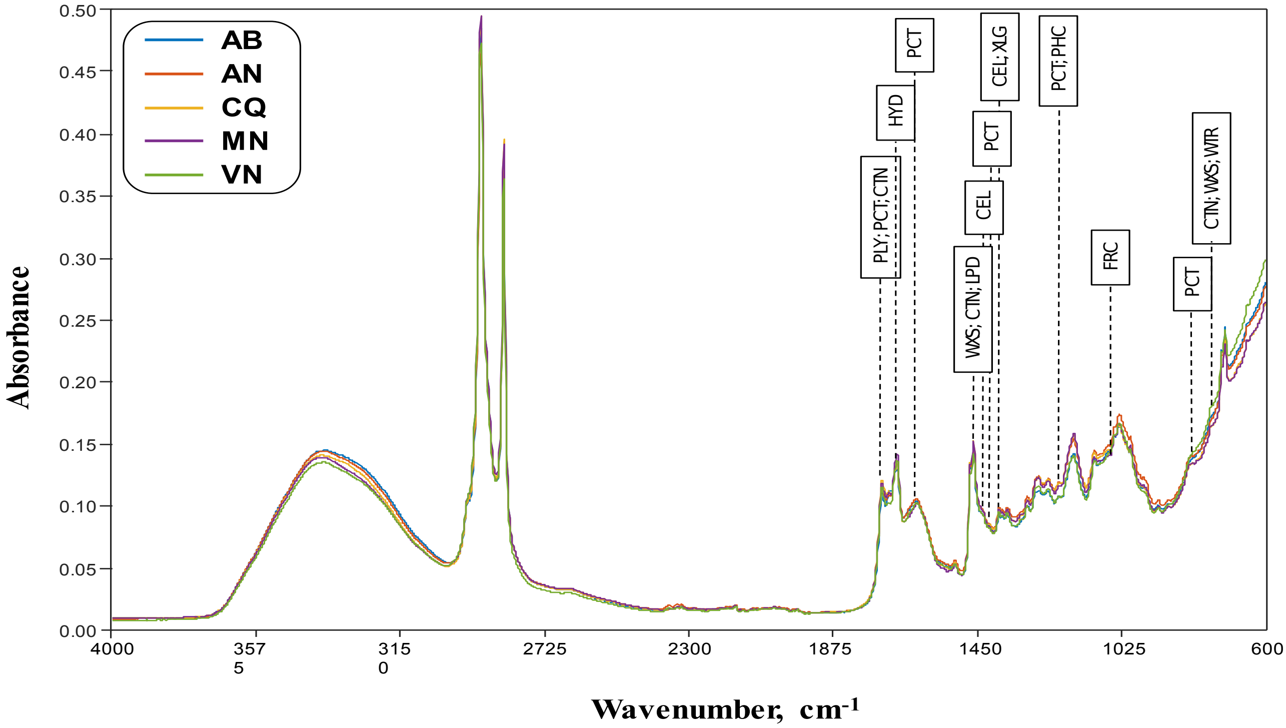
2.4 Artificial Neural Network (ANN) training

Data were randomly selected between training (85%) and test (15%) datasets. Performance of ANN was checked with cross-validation (15%) of training dataset. The selected ANN for this work is a two-layer feed-forward network with a simple perceptron with sigmoidal activation, and the network was trained with a scaled conjugate gradient backpropagation. Four different ANN were trained: for classifying grapes (Gr-ANN), for identifying ripeness (Ri-ANN) and for identifying the grape variety and ripeness from the leaf spectra (LeGr-ANN and LeRi-ANN), each of which consisted on an input layer with forty variables, a hidden layer with 10 neurons and an output layer with 5 components (Gr-ANN and LeGr-ANN) or 3 components (Ri-ANN and LeRi-ANN). In the context of this work, ripeness means the number of weeks passed since the first sampling (week 1). Forty different mid-IR peaks were selected, whose areas were used as input variables. These peaks were selected taking into account the absorption maxima at which the main chemical components of the grape skin absorbed IR radiation. Every peak was normalized with the MatLab *mapminmax* function, so the input data were in the range [-1,1]. Every spectrum was taken as the average of 16 scans, providing a good signal-to-noise ratio. Furthermore, the use of peak areas instead of heights contributed to minimize the effect of the noise in the signal. Consequently, no further noise-reduction protocol was followed so as not to overload the system with calculations.

3 Results and discussion

The mid-IR spectra provide precise information about the chemical groups present in the skin of the grapes. In this case the peaks of the spectra obtained (Figure 1) correspond to the following functional groups: 2916 and 2849, stretching (CH_2); 1733, stretching (C=O) ester; 1687 stretching (C=O) acid; 1629 stretching (COO^-); 1470, 1386 and 761 bending (CH_2); 1210 and 825 ring vibration; 1060 glycosidic bond (C-O-C); 960 bending(C-O). The spectra of the different varieties of grapes were very similar to the naked eye. In order to check whether this similitude was statistically significant, a study of the correlation coefficient of every variety pair was carried out with the aim to evaluate if the spectral difference for every variety pair was significantly different from zero.

Figure 1 *Mean spectra of the grape skin of the five tested varieties in this work: albarín blanco (AB), albarín negro (AB), carrasquín (CQ), mencía (MN) and verdejo negro (VN). Cellulose (CEL), cutin (CTN), fructose (FRC), hydroxycinnamic acids (HYD), lipids (LPD), pectins (PCT), phenolic compounds (PHC), polyesters (PLY), water (WTR), waxes (WXS) and xyloglucan (XLG).*



In a first step, the correlation coefficient as suggested by Varmuza et al. [14] was determined (Equation 1) and the probability p associated to the Student's t value was then calculated from that correlation coefficient.

$$\mathrm{COR}_{a-b} = \frac{z_a' z_b}{\|z_a\|\|z_b\|} \quad \text{(Equation 1)}$$

where z_a and z_b are the mean-centred absorbance spectra calculated according to Equation 2:

$$\vec{z_a} = \vec{x_a} - \vec{1}\cdot\overline{x_a} \quad \text{(Equation 2)}$$

In which x_a the vector containing the absorbances of compound a, $\vec{1}$ the vector (1,1,1,…) and $\overline{x_a}$ the mean value of the absorbances of compound a. The degrees of freedom are the number of wavenumbers scanned, 1765.

Using Li's approach [15], the null-hypothesis (H_0='the spectra are not correlated') is discarded if p value falls below 0.05. Results are collected in Table 2.

Table 2 *Similitude of IR absorbance spectra of the different varieties of grapes according to COR and Student's t.*

COR *t*	**AN**	**CQ**	**MN**	**VN**
AB	0.9973 *13.5841*	0.9984 *17.4448*	0.9983 *17.2682*	0.9959 *11.0407*
AN		0.9985 *18.2114*	0.9985 *18.3300*	0.9964 *11.7384*
CQ			0.9996 *35.3545*	0.9933 *8.6135*
MN				0.9941 *9.1972*

The probability associated for the t-values shown in Table 2 are below 10^{-5} in every case and, consequently it is possible to discard the null-hypothesis of not being correlated. These results could be understood taking into account that the chemical composition of the grapes skin is correlated in the different grape varieties, even in AB which is the only one white grape in the study.

Once confirmed that the different varieties are all correlated, the next step was to guess whether the spectra could be considered statistically undistinguishable or not. For doing that, a statistical study of the spectral differences between each possible variety pair was carried out. If two spectra are similar, the mean value of the absorbance differences, as defined in Equation 3, should be 0.

$$Dif_{a-b} = \overline{(\vec{a} - \vec{b})} \quad \text{(Equation 3)}$$

However, spectral differences didn't follow a normal distribution according to the Kolmogorov-Smirnov test and, therefore, Student's t was not applicable. Alternatively, we applied a Wilcoxon Signed Rank test to check whether the median was 0 for all possible variety pairs (avoiding self-comparisons). For each possible combination it resulted to be $p<10^{-5}$ except for pair MN-VN, with $p= 0.476$. This means that the spectral differences of every variety pair had a median significantly different from zero, with the exception of MN-VN.

With these results in mind, a good chance for recognizing the grape variety using an ANN (Gr-ANN) was expected. To tackle it, four different ANN, one at every ripeness status (1 week, 2 weeks or 3 weeks) and a fourth (pooling together all the data) were assayed. The results collected in Tables S1 (Supplementary Information) represent the matches (percentage of grapes correctly classified) and reliability (percentage of the grapes classified into a variety, which really belongs to it, or 100 - percentage of false positives) for the test dataset classification at three different stages of ripening, as well as the area under ROC curve AUC (Table 3). Unfortunately, no AB or VN grapes could be collected the last week due to the industrial needs of the vineyard.

Table 3 *Results of the classification of the training dataset for every variety at different ripeness states.*

Gr-ANN		**AB**	**AN**	**CQ**	**MN**	**VN**	**Average**
1st week	Matches	76.9%	100%	100%	91.7%	84.6%	90.6%
	Reliability	83.3%	90.9%	86.7%	91.7%	100%	90.5%
AUC (1st week)		0.9979	0.9980	0.9917	0.9993	0.9994	
2nd week	Matches	100%	77.8%	75.0%	85.7%	100%	87.7%
	Reliability	94.4%	63.6%	90.0%	92.3%	100%	88.1%
AUC (2nd week)		0.9997	0.9625	0.9739	0.9690	0.9997	
3rd week	Matches	---	92.9%	100%	92.9%	---	95.2%
	Reliability	---	100%	80.0%	100%	---	93.3%
AUC (3rd week)		---	0.9816	0.9798	0.9840	---	
All pooled together	Matches	95.5%	70.3%	81.0%	77.4%	92.3%	83.3%
	Reliability	87.5%	81.3%	81.0%	66.7%	100%	83.3%

AUC (all)	0.9981	0.9585	0.9456	0.9591	0.9971

In all the cases, the area under ROC curve (AUC) value is over 0.96, indicating a good performance of the classification. It is worth noting that a diminution of the global performance of the network was observed when the ripeness of the grape was not taken into consideration (83.3% average matches) when compared to the performance for every week separately (90.6%, 87.7% and 95.2% average matches). In order to identify the origin of this effect, we trained a second ANN (Ri-ANN) to evaluate the ripeness degree regardless of the grape variety, whose results are summarised in Table 4 and S2.

Table 4 *Results of the identification of the ripening week regardless of grape variety.*

Ri-ANN		**1st week**	**2nd week**	**3rd week**
Mixed varieties	Matches	88.7%	83.1%	80.0%
	Reliability	85.5%	83.1%	84.2%
Area under ROC curve		0.9844	0.9562	0.9758

Our overall success rates in the grape classification of 91.2% (average of weekly classification) or 83.3% (pooling all weeks together) as well as the success rate for identifying the ripeness degree of 83.9% is better than the success rates obtained by Gambetta et al. for identifying the geographical origin of Chardonnay grapes (81%-83%) [9]. Although better results can be found in the literature too (success rate 97.2%) [16], they are not directly comparable to ours as the classification was carried out just for only two varieties (Viura and Chardonnay) instead of five as in the present work. Cozzolino et al. [17] in a two-case classification (Chardonnay and Riesling) also present poorer results (86%) when using the grape juice instead of the grapes themselves.

The success rate of the Ri-ANN was lower than that of Gr-ANN, thus suggesting a stronger dependence of the IR spectra on the grape variety rather than on the ripeness degree. This fact was expectable considering the chemical changes that the grape skin may suffer over the short period of three weeks. On the other hand, since the whole pool of 40 variables were used in Gr-ANN and Ri-ANN, it was possible that those variables influencing more the ripeness degree were contributing to mask the variety identification and vice-versa. For this reason, new approaches were carried out to evaluate which experimental variables were influencing the most every trained ANN.

Several algorithms have been described with this purpose, being the Connection Weight Algorithm as proposed by Olden et al. [18] one of the most accurate. The Connection Weight Algorithm was carried out independently for each output neuron (that is, each target variety). Details are collected in Tables S3 and S4. When analysing the three Gr-ANN trained with a controlled ripeness degree, the critic variables resulted to be 35, 33, 27, 24, 16, 5 and 3; when checking the Gr-ANN trained with all the grape samples regardless of the ripeness degree, the variables selected were 35, 3, 36, 33, 5 and 2. Variables 35, 33, 5 and 3 were common to both lists, suggesting that they had the most weight in the variety identification. Finally, we classified the variables according to the number of times they appear considering all the classifications together (1st week, 2nd week, 3rd week and all Gr-ANN) finding as main variables 35, 33, 3, 5, 36, 27 and 16 (sorted in decreasing importance). Similarly, the application of the Olden's Connection Weight Algorithm to the Ri-ANN showed that the ripeness-related variables were 6, 30, 19 and 9. Table 5 summarises these variables and their assignation to chemical compounds in grape skin [10, 19 - 24]. The peak at 1210 cm^{-1} (variable #16) was not easy to assign, but considering the FTIR spectra accessible from the *Spectral Database for Organic Compounds SDBS* [25], apple pectin exhibes intense absorption at 1250 cm^{-1} and citrus pectin at 1210 cm^{-1}, it was plausible that the grape absorption at 1210 cm^{-1} arose from pectin too, although other authors assign this band to phenolic compounds [17]. The IR band corresponding to variable #27 is that at 761 cm^{-1}. Although this band is difficult to assign too, it is quite close to the δ (CH_2) rocking from cutin and waxes as reported by Heredia-Guerrero et al [23] and can also be assigned to water, according to Cozzolino et al.[17]. Fructose, with an absorption peak at 1070 cm^{-1} has also been described as the most contributing variable to the identification of the geographical origin of Chardonnay grapes by Gambetta et al.[9]. Fructose, at 1070 cm^{-1}, together with water, at 780 cm^{-1}, and phenolic compounds, at 1256 cm^{-1}, seem to play also an important role in the identification between Chardonnay and Riesling varieties in grape juices according to Cozzolino's work [17]. These wavenumbers are consistent with the variables shown in Table 5. It is clear that pectin has a strong influence both in the variety identification and in the ripeness degree. Polyphenols and sugar (fructose) are closely related to the ripeness degree as already described [24], and appear as important variables in our results too.

Once identified the main variables, the network was trained again using only the most influencing variables (35, 33, 3, 5, 36, 27 and 16 for Gr-ANN and 6, 30, 19 and 9 for Ri-ANN), but less satisfactory results were obtained (best match rate for Gr-ANN 74.4%, best reliability for Gr-ANN 79.2%; best match rate for Ri-ANN 82.5%, best reliability for Ri-ANN 64.4%). Despite using the most representative variables, a drastic reduction in the number of them impaired the success rate.

Table 5 *Chemical assignation of the main influencing variables [10 ,19 - 24]*

Variable #	Associated to	Wavenumber	Compound
3	Variety	1733 cm^{-1}	Polyesters, pectins, cutin
5	Variety	1687 cm^{-1}	Hydroxycinnamic acids
6	Ripeness	1629 cm^{-1}	Pectin
9	Ripeness	1386 cm^{-1}	Cellulose, Xyloglucan
16	Variety	1210 cm^{-1}	Possibly pectin or phenolic compounds
19	Ripeness	1063 cm^{-1}	Fructose
27	Variety	761 cm^{-1}	Probably cutin and waxes; possible water

30	Ripeness	1470 cm^{-1}	Waxes, cutin, lipids
33	Variety	1417 cm^{-1}	Carboxylate (pectin ester group)
35	Variety	825 cm^{-1}	Pectin
36	Variety	1433 cm^{-1}	Cellulose

Once found which variables were mainly involved in the classification, we tried to understand the confusion matrixes of Gr-ANN (regardless of ripeness state) and Ri-ANN. These matrixes were prepared with the whole dataset (training, validation and test data) and are shown in Figure 2. The main confusions occur with varieties AN, CQ and MN which are more frequently misclassified than VN and AB. It is important to state that AN is the worst identified variety (poorest number of matches) and MN the most wrongly chosen (poorest reliability). These facts could be explained considering that AN is the variety which shares more variables with other varieties (5 variables with three different varieties, see Table 6) and, therefore, it is more likely to be misclassified. On the other hand, MN is the only variety which has at least one variable in common with the others (Table 6). Sharing a variable with every variety makes easier for them to be included in a given category (poor reliability).

Figure 2 *Confusion matrixes considering training, validation and test datasets for Gr-ANN (left) and Ri-ANN (right). Red colour remarks worst results.*

		Input class					Reliability
		AB	AN	CQ	MN	VN	
Output class	AB	156	2	1	4	2	94.5%
	AN	0	187	23	12	2	83.5%
	CQ	3	21	190	20	1	80.9%
	MN	2	32	26	207	0	77.5%
	VN	1	0	3	0	157	97.5%
Matches		96.3%	77.3%	78.2%	85.2%	96.9%	**85.3%**

		Input week			Reliability
		1st	2nd	3rd	
Output week	1st	376	26	6	92.2%
	2nd	23	354	41	84.7%
	3rd	6	25	195	86.3%
Matches		92.8%	87.4%	80.6%	**87.9%**

Table 6 *Variables in common in the different varieties.*

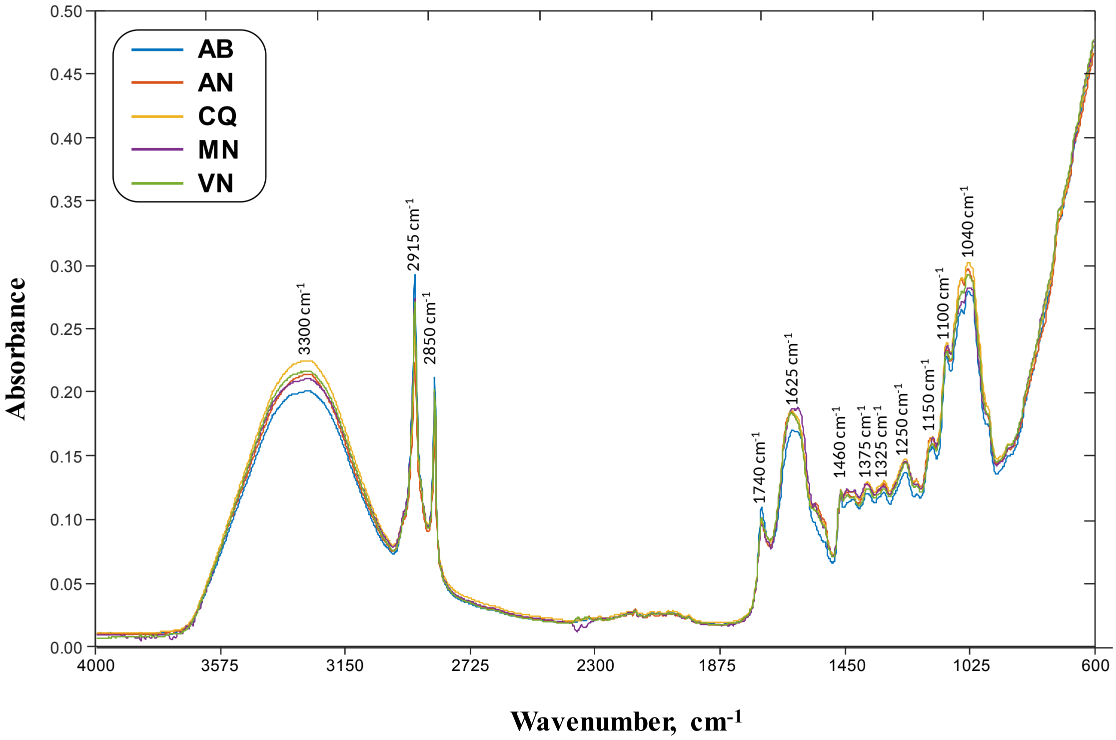
Variety	Shares	Details
AB	3 vars with 3 varieties	AN (#33 and #36), CQ (#3), MN (#3)

AN	4 vars with 3 varieties	AB (#33 and #36), MN (#35), VN (#2 and #35)
CQ	2 vars with 3 varieties	AB (#3), MN (#3) and VN (#5)
MN	2 vars with 4 varieties	AB (#3), AN (#35), CQ (#3) and VN (#35)
VN	3 vars with 3 varieties	AN (#2 and #35), CQ (#5) and MN (#35)

Concerning the confusion matrix for the Ri-ANN, it was clear that every week was mistaken with each other in a similar extent with the exception of weeks 1st and 3rd. This was easily explained if we notice that every variety share with each other just two variables: 1st and 2nd share #6 and #19, 2nd and 3rd share #6 and #9 and 1st and 3rd share #6 and #30. However, as shown in Table S4, variables #6 and #30 are the two with the most weight in their respective weeks.

So as to get more information, the possibility of identification of the grape variety through the FTIR spectrum of the leaves was evaluated. Since the chemical composition of the leaves was not expected to change with the fruit ripeness, results regarding this identification were expected to be poor. Mean spectra of the leaves of the five different vines are shown in Figure 3. Similarly, to the statistical analysis of the grapes, leaves showed a high correlation degree (details in Table S5) with probabilities below 10^{-5} which allow discarding the null-hypothesis of not being correlated. As in the case of grapes, the spectral difference between two varieties yielded non-normal distributions according to the Kolmogorov-Smirnov test, so we performed again a Wilcoxon signed rank test. Every possible combination showed a p value below 10^{-3} with the exception of AN-VN (p=0.0569), CQ-VN (p=0.7033) and MN-VN (p=0.1400). This implies that these pairs are very similar, without a statistically significant difference. With this information, poorer results than those obtained with the grapes were expected.

__Figure 3__ Mean spectra of the leaves of the five tested varieties in this work: Albarín Blanco (AB), Albarín Negro (AB), Carrasquín (CQ), Mencía (MN) and Verdejo Negro (VN).



The results in Tables 7 and S6 collect the matches (percentage of leaves correctly classified) and reliability (percentage of the leaves classified into a variety, really belongs to that variety, 100-percentage of false positives) for the test dataset classification without considering the degree of matureness, as well as the area under ROC curve.

__Table__ 7 Results of the classification of the training dataset for the leaves.

LeGr-ANN		AB	AN	CQ	MN	VN
Leaves	Matches	50.0%	75.0%	41.7%	50.0%	40.0%
	Reliability	25.0%	54.5%	445.5%	66.7%	66.7%
Area under ROC curve		0.8677	0.8588	0.8198	0.9021	0.7875

Finally, we wanted to check whether the leaves change enough during the maturation of the grape to obtain the ripeness degree of the fruit from the IR-spectrum of the leaf. We trained then a new ANN with the IR data obtained from the leaves (LeRi-ANN) whose results are shown in Tables 8 and S7.

__Table 8__ Results of the identification of the ripening week regardless of grape variety obtained from the leaves.

LeRi-ANN		1st week	2nd week	3rd week
Mixed varieties	Matches	80.0%	54.4%	22.2%
	Reliability	63.2%	42.9%	100.0%
Area under ROC curve		0.8891	0.7955	0.8510

No further studies on the LeGr-ANN and LeRi-ANN were performed due to the poor results obtained in the above experiments.

4 Conclusions

Despite being statistically similar, the FTIR spectra of the grape skin retain information enough to enable the identification of the grape variety using Artificial Neural Networks. ANN resulted to be a good choice for identifying the grape varieties involved in the *PDO Vino de Calidad* de Cangas production as well as the ripeness degree. The use of Olden's Connection Weight Algorithm allowed identifying the most influencing wavenumbers and chemical compounds, indicating that pectin was important for both identification of variety and the ripeness degree. As expected, fructose played an important role in the ripeness degree while polyphenols do not seem to affect the identification of the samples. Similar studies on the grape leaves did not yield relevant results because the chemical composition evolution of the studied plants was not as different as the one of the grapes in the collection time range.

Acknowledgements

Authors would like to acknowledge Beatriz Pérez-García and *Bodegas Vidas* wine cellar for providing the leaves and grapes for this study. We would like to acknowledge the

Ministerio de Economía y Competitividad and Fondo Europeo de Desarrollo Regional (MINECO/FEDER) by the financial support under the project MAT2015-66747-R.

References

1. Plant secondary metabolites: Occurrence, structure and role in the human diet, A. Crozier, MN. Clifford, H. Ashihara, Blackwell Publishing, Oxford, England **2006**

2. Phenolic substances in grapes and wine and their significance, VL. Singleton, P. Essau, Academic Press, New York **1969**

3. “Phenolic compounds in skins and seeds of ten grape *Vitis vinifera* varieties grown in a warm climate”, R. Rodríguez-Montealegre, R. Romero-Peces, JL. Chacón-Vozmediano, J. Martínez-Gascueña, E. García-Romero, *J. Food Compos. Anal.* 19 687-693 **2006** doi: doi.org/10.1016/j.jfca.2005.05.003

4. “Flavonoid compositional differences of grapes among site test plantings of Cabernet franc”, F. Broussaud, V. Cheynier, C. Asselin, M. Moutounet, *Am. J. Enology Vitic.* 50 277-284 **1999** accesible through http://www.ajevonline.org/content/50/3/277.article-info

5. “Use of FT-IR Spectra and PCA to the Bulk Characterization of Cell Wall Residues of Fruits and Vegetables Along a Fraction Process” , M. Szymanska-Chargot, A. Zdunek, *Food Biophys*. (2013) 8:29–42 DOI 10.1007/s11483-012-9279-7

6. “Attenuated total reflectance spectroscopy of plant leaves: a tool for ecological and botanical studies”, BR. da Luz, New Phytologist, 172 305-318 **2006** doi: 10.1111/j.1469-8137.2006.01823.x

7. “Rapid detection of authenticity and adulteration of cold pressed black cumin seed oil: A comparative study of ATR-FTIR spectroscopy and synchronous fluorescence with multivariate data analysis”, FN. Arslan, G. Akin, SNK. Elmas, I. Yilmaz, HG. Jannsen, A. Kenar, *Food Control*, 98 323-332 **2019** doi: 10.1016/j.foodcont.2018.11.055

8. “Detection of adulterants in grape nectars by attenuated total reflectance Fourier-transform mid-infrared spectroscopy and multivariate classification strategies”, CSW. Miaw, MM. Sena, SVC. de Souza, MP. Callao, I. Ruisánchez, *Food Chem.* 266 254-261 **2018** doi: 10.1016/j.foodchem.2018.06.006

9. “Classification of Chardonnay Grapes According to Geographical Indication and Quality Grade Using Attenuated Total Reflectance Mid-infrared Spectroscopy”, JM. Gambetta, D. Cozzolino, SEP. Bastian, DW. Jeffery, *Food Anal. Methods* 12 (1) 239-245 **2019** doi: 10.1007/s12161-018-1355-2

10. “Chemometric assisted Fourier Transform Infrared (FTIR) Spectroscopic analysis of fruit wine samples: optimizing the initialization and convergence criteria in the non-negative factor analysis algorithm for developing a robust classification model”, K. Kumar, A. Giehl, CD. Patz, *Spectrochim. Acta*, *Part A* 209 22-31 **2019** doi: doi.org/10.1016/j.saa.2018.10.024

11. “Easy-to-use analytical approach based on ATR–FTIR and chemometrics to identify apple varieties under Protected Designation of Origin (PDO)”, A. Fernández-González, J.M. Montejo-Bernardo, H. Rodríguez-Prieto, Z. Castaño-Monllor, R. Badía-Laíño, ME. Díaz-García, *Comput. Electron. Agric.* 108 166-172 **2014**

12. Boletín Oficial del Principado de Asturias 287, 27173-27181, 2008

13. Commission delegated regulation (EU) No 664/2014, Official Journal of the European Union, L 179/17, 2014

14. “Spectral similarity versus structural similarity: infrared spectroscopy”, Varmuza, K., Karlovits, M., Demuth, W., Anal. Chim. Acta 490 313–324 **2003**

15. “A comparative study of point-to-point algorithms for matching spectra”, J. Li, D.B. Hibbert, S. Fuller, G. Vaughn, *Chemometrics and intelligent laboratory systems* 82 (1-2) 50-58 **2006** doi: 10.1016/j.chemolab.2005.05.015

16. “Maturity, variety and origin determination in white grapes (*Vitis vinifera L.*) using near infrared reflectance technology”, Arana I., Jarén C., Arazuri S., *J. Near Infrared Spectrosc.* 13 349-357 **2005** doi: 10.1255/jnirs.566

17. “Varietal Differentiation of Grape Juice Based on the Analysis of Near- and Mid-infrared Spectral Data”, Cozzolino D., Cynkar W., Shah N., Smith P., *Food Anal. Methods* 5 381-387 **2012** doi: 10.1007/s12161-011-9249-6

18. “An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data”, J.D. Olden, M.K. Joy, R.G. Death, Ecological modelling 178 389-397 **2004** doi: 10.1016/j.ecolmodel.2004.03.013

19. “Linking ATR-FTIR and Raman features to phenolic extractability and other attributes in grape skin”, J. Nogales-Bueno, B. Baca-Bocanegra, A. Rooney, J.M. Hernández-Hierro, F.J. Heredia, H.J. Byrne, *Talanta* 167 44-50 **2017** doi: 10.1016/j.talanta.2017.02.008

20. “Use of FT-IR Spectra and PCA to the Bulk Characterization of Cell Wall Residues of Fruits and Vegetables Along a Fraction Process”, M. Szymanska-Chargot, A. Zdunek, *Food Biophys.* 8 29-42 **2013** doi: 10.1007/s11483-012-9279-7

21. Socrates, G., **2001**. Infrared and Raman Characteristic Group Frequencies: Tables and Charts, third ed. Wiley

22. *Spectral database for organic compounds SDBS*, https://sdbs.db.aist.go.jp/sdbs/cgi-bin/cre_index.cgi, compounds ID 2656 and 2695. Last accessed 22/03/2019

23. “Infrared and raman spectroscopic features of plant cuticles: a review”, J.A. Heredia-Guerrero, J.J. Benítez, E. Domínguez, I.S. Bayer, R. Cingolani, A. Athanassiou, A. Heredia, *Front. Plant Sci.*5 305 1-14 **2014**, doi: 10.3389/fpls.2014.00305

24. “Array of biosensors for discrimination of grapes according to grape variety, vintage and ripeness”, C. Medina-Plaza, J.A. de Saja, J.A. Fernández-Escudero, E. Barajas, G. Medrano, M.L. Rodríguez-Méndez, *Anal. Chim. Acta* 947 16-22 **2016** doi: 10.1016/j.aca.2016.10.032

25. Spectral Database for Organic Compounds SDBS, https://sdbs.db.aist.go.jp/sdbs/cgi-bin/cre_index.cgi, last accessed 5 april 2019

Gr-ANN		AB	AN	CQ	MN	VN
Training (1st week)	Matches	100%	100%	98.1%	100%	98.3%
	Reliability	98.0%	98.3%	100%	100%	100%
Validation (1st week)	Matches	94.4%	92.3%	85.7%	100%	100%
	Reliability	100%	92.3%	92.3%	100%	81.8%
Test (1st week)	Matches	76.9%	100%	100%	91.7%	84.6%
	Reliability	83.3%	90.9%	86.7%	91.7%	100%
Training (2nd week)	Matches	100%	85.0%	80.7%	86.0%	100%
	Reliability	98.2%	83.6%	83.6%	86.0%	100%
Validation (2nd week)	Matches	100%	75%	83.3%	70.6%	100%
	Reliability	100%	64.3%	76.9%	92.3%	92.3%
Test 2nd week	Matches	100%	77.8%	75.0%	85.7%	100%
	Reliability	94.4%	63.6%	90.0%	92.3%	100%
Training 3rd week	Matches	---	88.1%	94.8%	94.3%	---
	Reliability	---	94.5%	91.7%	90.9%	---
Validation 3rd week	Matches	---	85.7%	93.3%	92.9%	---
	Reliability	---	100%	87.5%	92.9%	---
Test 3rd week	Matches	---	92.9%	100%	92.9%	---
	Reliability	---	100%	80.0%	100%	---
Training (all)	Matches	95.8%	79.5%	78.4%	85.7%	98.2%
	Reliability	96.6%	84.0%	80.4%	80.0%	96.5%
Validation (all)	Matches	100%	73.5%	73.5%	88.6%	95.8%
	Reliability	91.7%	83.3%	83.3%	76.5%	100%
Test (all)	Matches	95.5%	70.3%	81.0%	77.4%	92.3%
	Reliability	87.5%	81.3%	81.0%	66.7%	100%

Table S1 - *Matches* represent the matches (percentage of grapes correctly classified) and *reliability* (percentage of the grapes classified into a class, which really belongs to that class) of the classification at three different stages of ripening.

Ri-ANN		**1st week**	**2nd week**	**3rd week**
Training	Matches	93.8%	88.4%	80.7%
	Reliability	94.1%	84.7%	86.3%
Validation	Matches	91.9%	87.7%	80.6%
	Reliability	89.1%	86.4%	89.3%
Test	Matches	88.7%	83.1%	80.0%
	Reliability	85.5%	83.1%	84.2%

Table S2 - *Matches* and *reliability* of the identification of the degree of ripening regardless of the grape class.

Network		**Most influencing variables**
Gr-ANN 1st week	AB	33 4 35 14 7
	AN	6 33 16 24 30
	CQ	5 23 12 3 27
	MN	6 27 18 16 35
	VN	27 35 5 23 3
	Total	35 ~ 27 (12% each) > 33 ~ 23 ~ 16 ~ 6 ~ 5 ~ 3 (8% each)
Gr-ANN 2nd week	AB	3 36 4 14 33
	AN	29 24 36 16 40
	CQ	35 24 3 5 38
	MN	35 26 17 4 12
	VN	35 28 2 32 9
	Total	35 (12%) > 36 ~ 24 ~ 4 ~ 3 (8% each)
Gr-ANN 3rd week	AN	33 27 25 24 36
	CQ	33 16 5 27 25
	MN	7 40 11 32 28
	Total	33 ~ 27 ~ 25 (13.3% each)
SELECTED		35 (9.2%), 33 and 27 (7.7% each), 24, 16, 5 and 3 (6.2% each)
Gr-ANN All	AB	3 36 4 33 17
	AN	33 2 35 16 36
	CQ	3 11 5 26 18
	MN	6 3 12 35 13
	VN	35 28 2 5 29
	Total	35 ~ 3 (12%) > 36 ~ 33 ~ 5 ~ 2 (8%)
SELECTED		35 and 3 (12%), 36, 33, 5 and 2 (8%)
GLOBAL SELECTION		35 (9%), 33 and 3 (7.8% each), 5 (6.7%), 36, 27 and 16 (5.6% each)

Table S3 – Most influencing variables on Gr- ANN. In the *total* section, variables appear sorted according to their number of apparition in that ANN. In the selected variables we show the percentage of apparition of that variable.

Network		Most influencing variables
Ri-ANN	1^{st}	6 30 3 19 7
	2^{nd}	6 11 1 9 19
	3^{rd}	6 30 9 27 31
	Total	6 (20%) > 30 (13.3%) ~ 19 (13.3%) ~ 9 (13.3%)
SELECTED		6 (20%), 30, 19 and 9 (13.3%)

Table S4 – Most influencing variables on Ri- ANN. In the *total* section, variables appear sorted according to their number of apparition in that ANN.

COR *t*	**AN leaf**	**CQ leaf**	**MN leaf**	**VN leaf**
AB leaf	0.9970 *12.8809*	0.9974 *13.8404*	0.9990 *22.3439*	0.9989 *21.3025*
AN leaf		0.9995 *31.6109*	0.9984 *17.6564*	0.9990 *22.3439*
CQ leaf			0.9987 *19.5925*	0.9994 *28.8545*
MN leaf				0.9994 *28.8545*

Table S5 - Similitude of IR absorbance spectra of the different leaves according to COR and Student's t.

LeGr-ANN		**AB**	**AN**	**CQ**	**MN**	**VN**
Training	Matches	70.8%	54.1%	59.2%	54.7%	33.3%
	Reliability	57.1%	51.1%	73.6%	45.4%	48.7%
Validation	Matches	64.0%	58.3%	48.4%	71.8%	44.4%
	Reliability	64.0%	52.5%	71.4%	59.6%	48.0%
Test	Matches	70.8%	52.9%	57.6%	74.4%	54.2%
	Reliability	60.7%	51.4%	79.2%	60.4%	72.2%

Table S6 - *Matches* represent the matches (percentage of grapes correctly classified) and *reliability* (percentage of the grapes classified into a class, which really belongs to that class) of the classification at three different stages of ripening.

LeRi-ANN		1st week	2nd week	3rd week
Training	Matches	75.8%	70.3%	18.6%
	Reliability	68.8%	55.1%	48.5%
Validation	Matches	71.4%	69.8%	21.9%
	Reliability	66.2%	56.4%	58.3%
Test	Matches	82.5%	71.4%	15.8%
	Reliability	64.4%	60.8%	54.5%

Table S7 - *Matches* and *reliability* of the identification of the degree of ripening regardless of the grape class from the IR spectra of the leaves.