Computers and Electronics in Agriculture 163 (2019) 104855

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

# Grapevine variety identification using "Big Data" collected with miniaturized spectrometer combined with support vector machines and convolutional neural networks

Armando M. Fernandes[a,*], Andrei B. Utkin[b], José Eiras-Dias[c], Jorge Cunha[c], José Silvestre[c], Pedro Melo-Pinto[d,e]

[a] INESC-ID Instituto de Engenharia de Sistemas e Computadores – Investigação e Desenvolvimento, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
[b] INOV INESC Inovação, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
[c] INIAV - Instituto Nacional de Investigação Agrária e Veterinária, Av. da República, Quinta do Marquês, 2780-157 Oeiras, Portugal
[d] CITAB-Centre for the Research and Technology of Agro-Environmental and Biological Sciences, Universidade de Trás-os-Montes e Alto Douro, 5000-801 Vila Real, Portugal
[e] Departamento de Engenharias, Escola de Ciências e Tecnologia, Universidade de Trás-os-Montes e Alto Douro, 5000-801 Vila Real, Portugal

ARTICLE INFO

*Keywords:*
Machine learning
Cultivar
Deep learning
Hyperspectral

ABSTRACT

Several experiments have been previously reported suggesting that the application of spectroscopy and machine learning allows the identification of grapevine varieties, however, up to now, the maximum number of varieties separated was twenty and the total number of sample spectra used does not go beyond the few hundreds. The present work aim is to answer the question: Is it possible to separate one variety from an enlarged group of other varieties when the number of samples is also significantly increased? With this in mind, a total of 35,833 spectra from leaves of 626 plants from 64 varieties were gathered for the study. This is a non-trivial evolution from previous works because it originates an increase in the variability of spectra which brings in a higher risk that a significant percentage of spectra of different varieties are equal and cannot be separated. Simultaneously, it was studied if a miniaturized and easy to use spectrometer could deliver data whose quality was enough to allow varieties separation even with data being collected in the field, non-destructively, and under uncontrolled solar lighting. This data was used to build support vector machines and convolutional neural networks for separating Touriga Nacional from 63 other varieties (including Touriga Franca) or Touriga Franca from 63 varieties (including Touriga Nacional), and the classification efficiencies are analysed.

## 1. Introduction

Variety identification is an important topic in viticulture because the wine quality potential is variety dependent and also because the consumers know and want wines of certain varieties which affects the price of grapes. It is therefore important to have methods that ensure trueness-to-type of plants that come out of nurseries. Conventionally, this variety identification is done using ampelography and ampelometry (Tomic et al., 2013) where an expert analyses and measures tens of grapevine features. However, the large number of features to analyse and the similarities between varieties make this a hard and laborious process that cannot be applied to hundreds of plants in a short time period. In addition, training a good ampelographer can take years. Even though ampelography and ampelometry are widely accepted and reliable methods there have been famous cases where producers thought that they were producing a certain variety but were in fact producing another (Tassie, 2010). This can have high costs for these producers due to the influence of the grapes variety on their commercial value. More recently, new DNA based methods (Tomic et al., 2013) have been developed that in spite of being highly reliable are still slow and expensive which prevents their extensive use. With the objective of creating simpler methods, in the last few years, spectroscopic methods have been combined with machine learning methods with promising results (Gutiérrez et al., 2015a; Gutiérrez et al., 2016; Gutiérrez et al., 2015b; Cao et al., 2010; Arana et al., 2005; Diago et al., 2013; Yang et al., 2012). Spectroscopy and machine learning has also been applied to grapevine clone identification but this topic is beyond the scope of the present article (Fernandes et al., 2014).

* Corresponding author.
*E-mail addresses:* arm.fernandes@gmail.com (A.M. Fernandes), andrei.utkin@inov.pt (A.B. Utkin), eiras.dias@iniav.pt (J. Eiras-Dias), jorge.cunha@iniav.pt (J. Cunha), jose.silvestre@iniav.pt (J. Silvestre), pmelo@utad.pt (P. Melo-Pinto).

https://doi.org/10.1016/j.compag.2019.104855
Received 12 July 2018; Received in revised form 12 April 2019; Accepted 10 June 2019
Available online 14 June 2019

Spectroscopy measures how electromagnetic radiation interacts with matter, i.e., how this radiation is absorbed or not depending on its wavelength. The ratio between the amount of light incident on a material and the amount of light coming from the material is called reflectance, and its plot versus the wavelength is called a reflectance spectrum. The necessity to use machine learning algorithms to process spectroscopic data comes from the large amount of information that this data contains. These algorithms learn from the spectral data to distinguish between different varieties. The current state-of-the-art in variety identification using spectroscopy and machine learning is described in Fernandes et al. (2018). Gutiérrez et al. (2015a) which separated 20 varieties and Cao et al. (2010) that used 197 samples for a single variety and 439 samples in total set the state-of-the-art in number of separated varieties and samples employed. The present work boosts these values by distinguishing samples of a certain variety from those of 63 other varieties; each variety to separate has more than 3000 samples and the total number of samples available for all varieties is 35833. This means a three-fold increase in the number of varieties, a 17-fold increase in number of samples in a variety and an 80-fold increase in the total number of samples employed. This leads to a more realistic and harder to solve problem than those in previously reported works. The reason is that there is an increased probability that the used dataset contains equal spectra from different varieties. This increase in number of varieties and samples also means an important step towards the creation of a robust grapevine variety identification system that can be commercialized. The present work reports the construction of machine learning classifiers capable of separating Touriga Franca ($TF_{var}$) or Touriga Nacional ($TN_{var}$) from the remaining varieties. When separating $TF_{var}$, $TN_{var}$ was added to the set of remaining varieties and vice-versa. The classifiers used were Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) and their results will be compared. To the best of the authors' knowledge this is the first time that CNN are being employed in grapevine variety identification even though they have been used once in rice variety identification (Qiu et al., 2018). The built classifiers were of the one-vs.-all type meaning that they are binary and indicate if a spectrum belongs to a certain variety or not. The classifiers were tested with data gathered in a different day of the training and validation data to minimize the influence, on the spectra separation, of environmental or biological parameters specific to a certain day. The choice of $TF_{var}$ and $TN_{var}$ as the main varieties to be separated has to do with their importance in Portugal in the production of the worldwide famous Port wine. In fact, $TF_{var}$ and $TN_{var}$ have each 7% (Ranking de castas, 2017), of the total grapevine area planted in Portugal making them the two most planted Portuguese autochthonous varieties in the country. Portugal which has one of the largest pools of autochthonous grapevine varieties in the world, 239 (Cunha et al., 2016), is actively working towards their preservation and dissemination.

## 2. Materials and methods

### 2.1. Samples

The spectroscopic measurements of leaves were done in the 25th, 26th, 27th and 28th of July of 2017, in Dois Portos, Portugal, 39°02′34.03″N 9°10′57.41″W, in the Portuguese ampelographic collection planted in 1988 at INIAV - *Instituto Nacional de Investigação Agrária e Veterinária* (www.iniav.pt). There was no precipitation during these days. The measurements were done in the field, non-destructively, i.e. no part of the grapevine was removed for measurement, and without touching the grapevines. The data collected was from Touriga Nacional ($TN_{var}$), Touriga Franca ($TF_{var}$) and a large blend of 62 varieties that were measured without being separated. The varieties in this blend are all those in Appendix A except for $TF_{var}$ and $TN_{var}$. When creating a classifier to identify $TN_{var}$ the data from variety $TF_{var}$ was added to the data of the blend and vice-versa. $TF_{var}$ and $TN_{var}$ are Portuguese autochthonous as well as 41 other varieties from the blend. The blend contains also eight varieties from Spain, three from France, and one from Italy, Greece, Hungary and Turkey and one from both Germany and Italy. The chosen varieties include the eight most planted in Portugal, and 21 of them, those that individually have more than 0.5% of the total area planted, correspond to 65% of the total vineyard area planted in Portugal when grouped. The dataset contains data from plants whose material was collected in different locations even though it was all planted in Dois Portos. The total number of plants measured was 626. $TN_{var}$ and $TF_{var}$ were each measured four times every day, while each of the other varieties in the blend was measured twice a day, equally divided between the morning and the afternoon. The plants of all varieties were positioned along 15 lines in the vineyard and each line was randomly visited throughout the day with the measurements being done from North to South or in the opposite direction. The morning and afternoon corresponded, respectively, to the time periods between 10:36 and 12:52 GMT + 1 and 13:41 and 17:05 GMT + 1. A total of 35,833 spectra were measured, with 3345 being $TN_{var}$, 3569 being $TF_{var}$ and 28,919 being from the variety blend that does not contain $TN_{var}$ or $TF_{var}$. The amount of data gathered per day and in the four days is shown in Table 1. Appendix A contains an estimate of the number of samples measured for the varieties in the blend. This estimate was obtained dividing the number of samples measured in each row of grapevines by the number of plants in the row. The reason is that only the number of samples per row, which contained several blend varieties, was noted because the spectra from each row of grapevines were collected without stops between each blend variety. The reason not to stop was that it would delay significantly the measurement process and reduce, also significantly, the total number of samples measured per day. In addition, knowing with greater accuracy the number of samples in the blend varieties does not interfere with the conclusions of the present study. $TN_{var}$ and $TF_{var}$ were measured individually and separately from the remaining varieties and, therefore, their number of samples is known exactly.

In the present work the spectral dataset was not augmented by artificially generating new samples from the measured ones. The reason was that it does not yet seem to exist a set of standard methods for this, contrarily to what happens with images. In fact, the recent work of Qiu et al. (2018), in rice variety identification, does not seem to use dataset augmentation. Consequently, artificial data generation is left for later work. Only copies of the spectra of the less populated class were added to convolutional neural training sets to balance the training data. With the Support Vector Machines the balancing was achieved by increasing the weights of spectra in the classes with fewer of them.

### 2.2. Acquisition method

Spectra were acquired with a portable Flame-S spectrometer from OceanOptics that weighs 265 g and has a power consumption of 1.25 W. It is computer controlled and powered via a USB port, operating between 360 and 1028 nm with a total of 2048 wavelength channels. The spectrometer detector was a Sony ILX511B linear silicon CCD array, the signal-to-noise ratio was 250:1, the dynamic range was 1300:1 for a single acquisition, the dark noise was 50 RMS counts, and the analog-to-digital conversion depth was 16 bit. The spectrometer temperature

**Table 1**
Number of spectra measured per day and variety.

| Date | $TN_{var}$ | $TF_{var}$ | Variety blend without $TF_{var}$ and $TN_{var}$ |
|---|---|---|---|
| July, 25th | 703 | 688 | 7009 |
| July, 26th | 789 | 843 | 7488 |
| July, 27th | 820 | 890 | 6997 |
| July, 28th | 1033 | 1148 | 7425 |
| Total | 3345 | 3569 | 28,919 |

was kept stable at 21 ± 0.2 °C by placing it inside an insulated box cooled by an 89 W Peltier thermoelectric cooler, that was switched on and off by a temperature controller. The heat on the hot side of the Peltier was removed by a cooling unit with 150 W of thermal dissipation power normally employed in computer CPU cooling. The spectrometer had an optical fiber attached that was used to collect the sunlight reflected from grapevine leaves. The spectra acquired corresponded to averaging 30 scans with an integration time of 9 ms, which meant a total measurement time per sample of 270 ms. The integration time was set so that the spectrometer measurements did not saturate. However, in the morning of July, 27th, the integration time had to be increased to 30 ms because the weather was cloudy and the lighting was smaller than in other days. In order to keep the total measurement time more or less the same so that the number of acquired spectra would not change much, the number of scans averaged per spectra was reduced to 10. Even though it was possible to exclude the data of the morning of the 27th of July from classifier creation it was kept because situations of poor sun lighting are highly probable to occur. The measurements were done by slowly walking along the grapevines and collecting with the optical fiber the light coming from these grapevines. The optical fiber was kept slightly above the grapevines and therefore the captured sunlight originated from the adaxial side of the leaves. The spectrometer was set to save one spectra/sample every 270 ms, therefore, measuring continuously for a few minutes, while walking, allows gathering hundreds of spectra. Namely, in one minute one can collect approximately 200 spectra.

### 2.3. Reflectance determination

The raw spectra of the grapevine leaves were converted into reflectance spectra using the formula:

$$R(\lambda) = \frac{LeafI(\lambda) - DI(\lambda)}{SpeI(\lambda) - DI(\lambda)}$$

where $R$ is the reflectance, λ is the wavelength, *LeafI* the signal intensity coming from the leaves, *SpeI* the signal intensity coming from a calibration plate and *DI* the dark intensity that is measured with the spectrometer fiber plugged. The calibration plate is made of white spectralon, which has the highest known diffuse reflectance over the spectrometer operating wavelengths. The mean of all spectra gathered is illustrated in Fig. 1 separated in the varieties $TF_{var}$, $TN_{var}$ and blend of varieties (other). It is also depicted the mean plus and minus one standard deviation. In the region below the 700 nm the variability of the data is practically the same for the groups $TF_{var}$, $TN_{var}$ and other. In the region between 700 and 950 nm one may see differences between the mean plus and minus one standard deviation for the three groups of varieties, however, there is a large overlap of the spectra of the tree

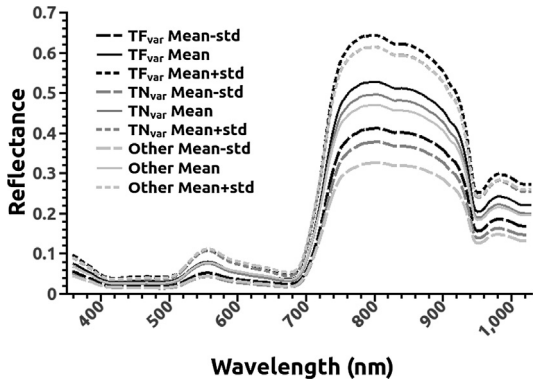


**Fig. 1.** Mean spectra and mean plus and minus one standard deviation (std) for the varieties Touriga Nacional ($TN_{var}$), Touriga Franca ($TF_{var}$) and the blend of multiple varieties (other). The data is from the 26th of July and corresponds to the training set.

groups shown. The curves of "Mean + std" for $TN_{var}$ and the blend of varieties overlap in the region between 700 and 950 nm. The spectra comparison can also be made using the second derivative of the spectra. In this situation, the mean curves are coincident across the whole spectrum for $TF_{var}$, $TN_{var}$ and blend of varieties (other). The same happens for the "Mean + std" and "Mean-std" curves, meaning that the overlap between spectra for the second derivative is total.

### 2.4. Reflectance spectra preprocessing

The process of developing the best classifier involved applying several combinations of preprocessing methods to the reflectance spectra before the spectra were given to the classifiers. The methods employed were the Savitzky-Golay (SG) filter, logarithm, multiplicative scattering correction (MSC), standard normal variate (SNV), first derivative and second derivative (Rinnan et al., 2009). The SG filter consists of replacing each point of the spectra by the least squares fit with a polynomial, in the present case a first order one, of the surrounding points. It was applied to smooth the spectra. This is equivalent to replacing each point ($R_i$) by a linear combination ($SG_i$) of itself and its surrounding points through the equation:

$$SG_i = \sum_n C_n R_n$$

where the $C_n$ are coefficients that provide the polynomial least squares fit. Index $n$ includes point $i$ and some predefined number of surrounding points. The negative logarithm of the reflectance is a way to get values proportional to compound concentration. MSC aims at removing undesirable additive and multiplicative scattering effects from the data. This implies fitting each individual reflectance spectrum, $\boldsymbol{R_j}$, to the average training data spectrum, $\boldsymbol{R_{mean}}$, and then correcting the individual spectrum with the fit coefficients, $b_0$ and $b_1$, following the equations:

$$\boldsymbol{R_j} = b_0 + b_1 \boldsymbol{R_{mean}} + error$$

$$\boldsymbol{MSC_j} = \frac{\boldsymbol{R_j} - b_0}{b_1}$$

where $\boldsymbol{MSC_j}$ is the multiplicative scattering correction for each spectrum. The SNV of a spectrum, $\boldsymbol{SNV_j}$, is obtained by subtracting from each sample spectrum, $\boldsymbol{R_j}$, its mean value, $mean_j$, and dividing the result by its standard deviation, $std_j$, according to the equation:

$$\boldsymbol{SNV_j} = \frac{\boldsymbol{R_j} - mean_j}{std_j}$$

It is rather similar to MSC with the difference that SNV does not use information from an average training data spectrum. The first order derivative, $R^{(1)}$, has the ability to remove the additive effects, while the second order derivative, $R^{(2)}$, eliminates both additive and multiplicative effects. In the present article, the derivatives used are centered difference approximations with fourth order error (Butt, 2010), as shown below:

$$R_i^{(1)} = \frac{-R_{i+2} + 8R_{i+1} - 8R_{i-1} + R_{i-2}}{12}$$

$$R_i^{(2)} = \frac{-R_{i+2} + 16R_{i+1} - 30R_i + 16R_{i-1} - R_{i-2}}{12}$$

where $R_i$ is the reflectance at point $i$ and '$i + 1$' or '$i$-1' are points to the right or left of point $i$.

### 2.5. Support vector Machines (SVM)

Support vector machines for classification create a hyperplane that maximizes the distance to the points to be classified. This corresponds to solving an optimization problem that minimizes the norm of the weights given to the points that are used to create the hyperplane and

which are called the support vectors. However, a hyperplane that correctly separates all the points cannot be always found, being necessary to add a slack variable that introduces a penalization into the minimization problem for each point that is not properly classified. This optimization problem can be rewritten in a way that the objective function depends only of the training data, namely of dot products between training patterns. In the case that the problem is not linear and, therefore an hyperplane cannot be used to separate the patterns, one can use a kernel function that transforms the data into a space where the problem becomes linear. This is simply done by replacing the dot products between training patterns in the objective function by the kernel function transformation of these patterns. The SVM software used was LIBSVM (Chang and Lin, 2011).

### 2.6. Convolutional neural networks (CNN)

A CNN is composed of convolutional, pooling and fully connected layers. The objective of these layers in the CNN is to create high order features that improve the classification efficiency. The convolutional layers, one dimensional in the present work, consist of a set of feature maps with an associated receptive field of the size of only a small region of the input spectra. Each feature map output corresponds to the convolution (dot product) between the receptive field weights and all spectra points. This means that contiguous points in the feature map were determined in overlapping and contiguous regions of the input spectra. This way, several features are determined over the whole spectra. The big advantage of using convolutional layers is that they have much less weights to be learned than a fully connected neural network. The pooling layer is applied to the feature maps and performs a down-sampling. In this case, the feature maps were divided into non-overlapping regions and their maximum taken. Afterwards, dropout is applied meaning that at training time some elements contributing to the next layer are left out, or not, with a certain probability. After the convolution and pooling layers, one of each in the present work, there is a fully connected neural network that processes the features coming from the previous layers. In the present work the fully connected neural network had one hidden layer and the activation function in the convolutional layer and hidden neurons was a rectified linear unit while in the output neuron it was a sigmoid. The CNN weights were calculated using stochastic gradient descent to minimize the binary cross-entropy. The CNN were trained using KERAS (https://keras.io/) running in graphical processing units.

### 2.7. Validation methods

The acquired data was divided into training, validation and test sets. The training data was used to create the SVM and CNN classifiers, the validation data purpose was that of adjusting the classifier parameters and the test set was used to assess the real generalization ability of the created classifiers. The data from the 26th and 28th of July were used for training and testing, respectively. The 25th and 27th data were both employed in validation. The idea was to choose classifiers with equivalent results for both days. The reason is that some variability in data was expected between days due to various factors such as the biological material natural evolution, changes in solar lighting at the time of data acquisition or water availability in the soil. By using data from several days in the validation one hope to discard the influence of all these factors when choosing a classifier and keeping only the information relative to the grapevine varieties. The test set data was collected in a different day of the training and validation data to take into account the data variability between days when assessing the robustness of the created classifiers. Training, validating and testing with data from different days was already employed in a previous work of the article's authors that separates ten grapevine varieties using hyperspectral stem data instead of leaves (Fernandes et al., 2018). Most frequently, scientific works disregard these differences in the measured data and distribute data evenly in the training, validation and testing sets. However, there are studies where the difference in data, not between days but between years, is taken into account by training and testing classifiers with data from different years (Gomes et al., 2017).

The validation data was first used to select the SVM and CNN set of internal parameter values that provided classifiers with a variation in classification percentage between the two validation days of less than 10% in both classes of the binary classifiers. The 10% value was chosen to provide some flexibility to the classifier creation, since the data from the two days has some difference, but, simultaneously, it is not so large that allows the classifier to respond very differently to data that is reasonably similar. In SVM to calculate the area under relative operating characteristic (AUROC) curves, used to select the best classifiers, it was necessary to determine class probabilities that take longer to be obtained than the simplest SVM that can be used for classification. Consequently, the difference in percentages for the two validation days was determined for all possible classifier parameters before calculation of the class probabilities. The class probabilities were only calculated for the parameters sets whose SVM, without class probabilities, presented less than 10% classification percentage difference between the two validation days. For CNN, the class probabilities were always obtained since in CNN the class probability determination is not separate as with the SVM.

The best classifier and set of internal parameters and spectra preprocessing methods was that with the best AUROC determined for all patterns from the two validation days together. This allows to have classifiers robust to slight variations of the data. Only the best SVM or CNN classifiers were assessed using the test set, to ensure an unbiased measurement of generalization ability. The classification results that will be shown ahead were determined for the decision threshold that allowed the classifier to provide the largest absolute difference between true positive and false positive percentages in the ROC curves of the validation set. This optimal point in ROC corresponds to different classification percentages of those obtained with the SVM before class probability calculation and, consequently, the tables contain situations where the classification percentages between the two validation days are larger than 10%. However, this does not happen in the best SVM for Touriga Nacional or Touriga Franca identification.

## 3. Results

This section contains the results of the attempts to create two classifiers able to separate Touriga Nacional or Touriga Franca from the remaining varieties. In the case of the classifier for Touriga Nacional, Touriga Franca was included in the remaining varieties and vice-versa. Support vector machines (SVM) and convolutional neural networks (CNN) were both tested for each classifier.

### 3.1. Parameter values optimisation

The study of the best preprocessing method was done using the SVM and simultaneously with the SVM kernel optimization. Once the best set of preprocessing methods and their parameter values was determined it was employed in the training of the CNN. This was done this way to reduce the necessary amount of computation that is quite large for the CNN. The adjustment of the use of SG filter, logarithm, MSC, SNV and derivative as well as of the SVM kernel was done by selecting an initial set of these characteristics and afterwards trying to optimize it. Initially SG was set to 71, the logarithm was employed, and both MSC and SNV were not. These settings were employed with two kernels, the polynomial and the RBF, and with no derivative, with the first derivative or with the second derivative of the logarithm of reflectance. This allowed to choose the best kernel and derivative after six different tests. Next, the SG was varied between not used or used with windows of 51 or 91 pixels. Finally, it was tested if employing or not the logarithm, MSC, or SNV could improve the results.

**Table 2**
Calculation results for Touriga Nacional. The best SVM is in bold and the best overall result is in Gray. "Poly" stands for polynomial.

| Classifier | | Kernel | SG | log | MSC | SNV | Der | AUROC Val | Validation 1 | | Validation 2 | | AUROC Test | Test | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Not $TN_{var}$ (%) | $TN_{var}$ (%) | Not $TN_{var}$ (%) | $TN_{var}$ (%) | | Not $TN_{var}$ (%) | $TN_{var}$ (%) |
| Support vector machines | 1 | **Poly** | **71** | **Yes** | **No** | **No** | **No** | **0.73** | **66.74** | **72.55** | **61.08** | **69.76** | **0.79** | **63.02** | **81.90** |
| | 2 | Poly | 71 | Yes | No | No | 1 | 0.69 | 46.03 | 85.06 | 49.51 | 78.66 | | | |
| | 3 | Poly | 71 | Yes | No | No | 2 | 0.54 | 21.58 | 80.09 | 34.63 | 79.39 | | | |
| | 4 | RBF | 71 | Yes | No | No | No | No set of parameters allowed to have both validation results with difference smaller than 10% | | | | | | | |
| | 5 | RBF | 71 | Yes | No | No | 1 | | | | | | | | |
| | 6 | RBF | 71 | Yes | No | No | 2 | | | | | | | | |
| | 7 | Poly | 51 | Yes | No | No | No | 0.71 | 73.79 | 59.60 | 60.24 | 70.85 | | | |
| | 8 | Poly | 91 | Yes | No | No | No | 0.71 | 67.91 | 67.85 | 63.31 | 61.46 | | | |
| | 9 | Poly | No | Yes | No | No | No | 0.65 | 55.79 | 58.75 | 63.09 | 65.98 | | | |
| | 10 | Poly | 71 | No | Yes | No | No | 0.65 | 56.06 | 60.88 | 60.78 | 66.34 | | | |
| | 11 | Poly | 71 | Yes | No | Yes | No | 0.73 | 71.14 | 70.27 | 65.09 | 61.59 | | | |
| | 12 | Poly | 71 | No | No | No | No | 0.67 | 62.93 | 52.06 | 81.43 | 53.29 | | | |
| CNN | | - | 71 | Yes | No | No | No | 0.69 | 57.35 | 74.25 | 52.33 | 77.07 | 0.68 | 62.75 | 67.67 |

The SVM internal parameter optimization was done by grid search testing 25 or 625 different values combination when RBF or polynomial kernel were used, respectively. The difference in the number of combinations has to do with the fact that SVM with RBF kernel has the values of two parameters to adjust while polynomial kernel has four. With the CNN, for Touriga Nacional it was possible to test 28 randomly selected sets of parameter values during seven days of calculation. For Touriga Franca 14 different set of parameter values, also randomly chosen, were tested during approximately two days of calculation. In spite of the larger calculation time for Touriga Nacional the results are not better for this variety as it will be seen ahead.

### 3.2. Identification of Touriga Nacional

Table 2 contains the validation and test results for Touriga Nacional for multiple combinations of spectra preprocessing steps when using SVM or CNN. For this variety, the best spectra preprocessing consisted of applying a SG window with 71 pixels and the logarithm of reflectance and not using MSC, SNV, or derivatives. This corresponds to classifier number one that is shown in bold. An example of a preprocessed spectrum is shown in Fig. 2. Only the two top images are relevant for the present case. The use of RBF kernel is totally unadvised in this case since it did not allow to obtain equivalent results for both validation days as seen in classifiers 4, 5 and 6. This suggests that these classifiers focused on characteristics that varied from day to day or even on the noise but not on the characteristics relevant for variety classification. The use of the second derivative in classifier number three seems to originate a large drop in AUROC. The best SVM, number one, used polynomial kernel and obtained an AUROC of 0.73 in validation. The best internal parameters of this SVM were 0.1 for the regularisation cost, 0.1 and 1000 for the *gamma* and *coef0* parameters of the kernel in LIBSVM, respectively, and 5 for the polynomial degree. The results between two validation days for SVM number one had a difference of only 5.66 and 2.79 percentage points for non-$TN_{var}$ and $TN_{var}$ patterns, respectively, which is a good indication that the classifier might be using relevant variety characteristics and is consequently able to generalise. This is confirmed by the test AUROC of 0.79 that is slightly better than the validation AUROC. In test, the correct classification percentage was 63.02% and 81.90% for non-$TN_{var}$ and $TN_{var}$ patterns, respectively. The corresponding confusion matrix is shown in Table 3. The AUROC in test was better than in validation due to a clear improvement in $TN_{var}$ patterns classification.

A convolutional neural network trained with spectra preprocessed with the methods providing the best results for SVM was unable to surpass the best SVM AUROC. In fact, for the CNN the AUROC was 0.69 which compares to the 0.73 of the SVM number one in validation. In test, it is even worse with the CNN having AUROC of 0.68 when compared to the 0.79 of the SVM. In test, the CNN and SVM present rather similar results for non-$TN_{var}$ patterns; however, for $TN_{var}$ spectra the SVM is clearly better by more than 14 percentage points. The CNN parameters were 50 filters with kernel size 60 in the one dimensional convolution, the max pooling size was 10, the dropout 50% and the number of hidden units of the fully connect neural network was 30. The CNN structure and spectra preprocessing are shown in Fig. 3. It is interesting to point out that the difference between AUROC in validation and test in CNN is smaller than in SVM. Fig. 4 depicts the ROC curves for the best SVM in both validation and test sets, clearly showing the better results in test relative to validation.

### 3.3. Identification of Touriga Franca

Table 4 contains the validation and test results for Touriga Franca for various spectra preprocessing steps when using SVM or CNN. For Touriga Franca the best preprocessing meant using a SG window with 91 pixels, and using also a logarithm of reflectance, SNV and first derivative while not using MSC. An example of a preprocessed spectrum is shown in Fig. 2. The images depict the use of the four methods employed in the present case. The best kernel for SVM classification was RBF. The SVM parameters were 0.1 for the regularisation cost and 1000 for the kernel *gamma*. These parameters allowed obtaining an AUROC of 0.96 with the results between two validation days having a difference of only 5.55% and 3.63% for non-$TF_{var}$ and $TF_{var}$ patterns, respectively. The AUROC and correct classification percentages in test were similar to those in validation. The test AUROC was 0.97 for a correct classification percentage in non-$TF_{var}$ patterns of 92.94% and of 89.2% for $TF_{var}$ patterns. In the results one may see that not using a SG filter in classifier nine or using second derivative in classifier three and six led to small AUROC values of less than 0.7. Changing the SG window size from 71 to 91 between classifiers 5 and 8 leads to the same AUROC value, 0.95, and finally applying SNV in classifier 11 improves the AUROC to 0.96.

The CNN with the best preprocessing methods obtained an AUROC of 0.97 in validation which is better than the 0.96 of the SVM number 11. For the CNN the difference between validation days for non-$TF_{var}$ and $TF_{var}$ patterns was 7.4% and 0.16%, respectively. In the test set the CNN was again better than the SVM, with the former obtaining an AUROC of 0.98 versus the 0.97 of the latter. In test, the correct classification percentages were 91.63% for non-$TF_{var}$ patterns and 93.82% for $TF_{var}$ patterns. The corresponding confusion matrix is given in Table 5. The difference in AUROC between validation and test is small indicating good generalization. In fact, the ROC curves in Fig. 4 overlap significantly. The CNN parameters were 40 filters with kernel size 70

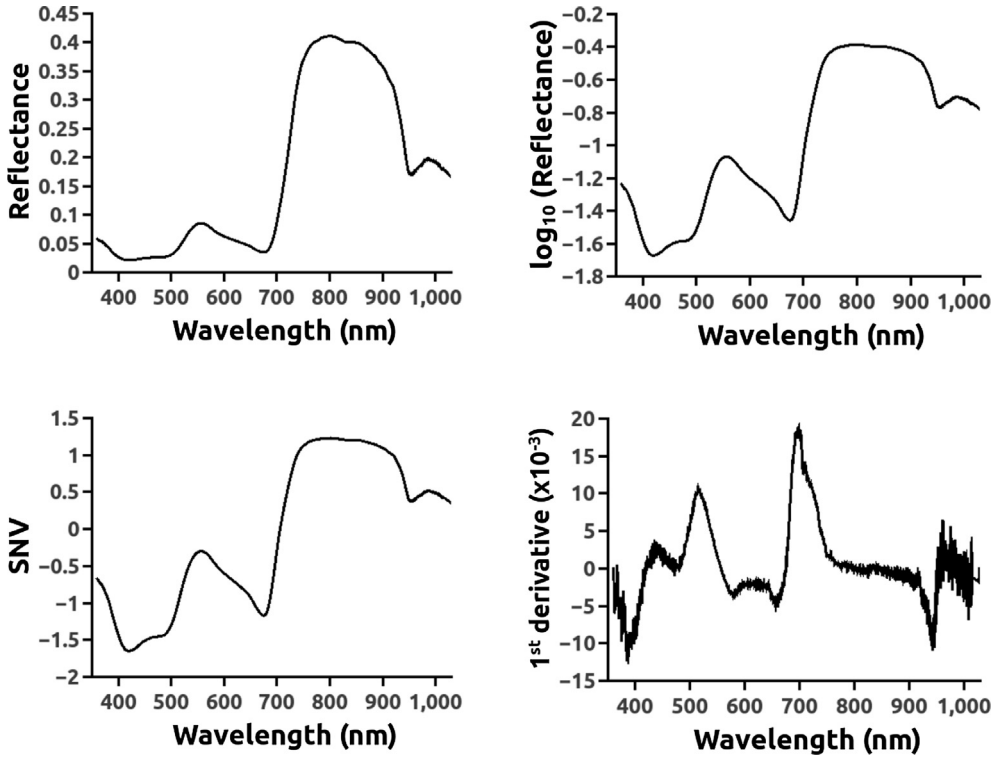


**Fig. 2.** Example of spectrum preprocessing. The top left image shows the reflectance spectrum after Savitzky-Golay filter, and the top right image the reflectance logarithm. In the bottom it is shown the application of standard normal variate (SNV) to the logarithm and finally a first derivative of SNV.

**Table 3**
Confusion matrix in the test set for the best classifier from Table 2, the SVM classifier number one. The values correspond to number of samples.

| | | Predicted variety | |
|---|---|---|---|
| | | Touriga Nacional | Not-Touriga Nacional |
| Actual variety | Touriga Nacional | 846 | 187 |
| | Not-Touriga Nacional | 3170 | 5403 |

for the one dimensional convolution, the max pooling size was 5, the dropout 20% and the number of hidden units of the fully connected neural network was 30. The CNN structure and spectra preprocessing are shown in Fig. 3. The difference between SVM and CNN in test was significantly smaller for Touriga Franca than for Touriga Nacional. The classification results for Touriga Franca were also significantly better than those for Touriga Nacional.

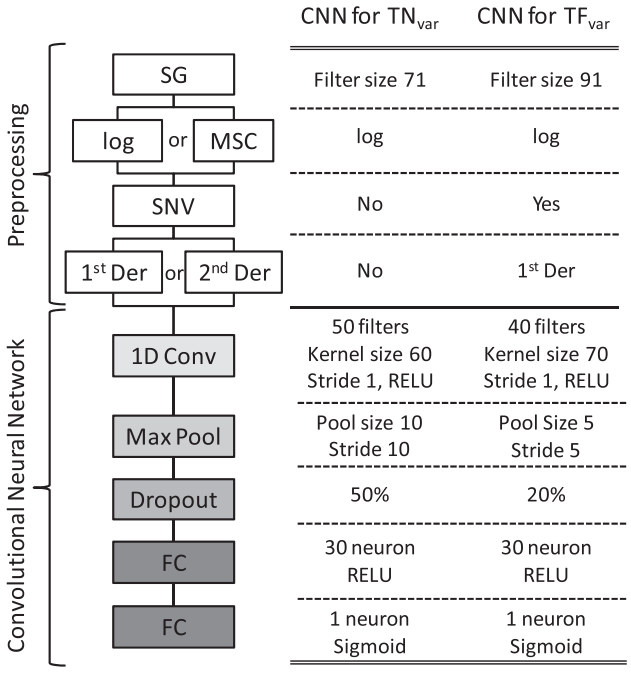


**Fig. 3.** Structure of the CNN and the spectra preprocessing used. The preprocessing is the same employed in the SVM with the best results.

## 4. Discussion

In the present work, the analysed samples were leaves with the spectra being collected non-destructively in the field. Gutiérrez et al. (2015a), Gutiérrez et al. (2016), Gutiérrez et al. (2015b), in three different works, has also collected leaf spectra non-destructively, however, in the present work, contact to the sample was unnecessary, contrarily to Gutiérrez et al. (2015a), Gutiérrez et al. (2016), Gutiérrez et al. (2015b) works, allowing therefore a faster sample collection. This was rather important to enable collecting an extremely large number of spectra as it was done. Diago et al. (2013) have also used leaves for variety discrimination but Cao et al. (2010) and Arana et al. (2005) employed grape berries, instead. The equipment employed by Gutiérrez et al. (2015a), Gutiérrez et al. (2016), Gutiérrez et al. (2015b) operated in the wavelengths between 1600 nm and 2400 nm while our miniaturized spectrometer worked in the 400–1000 nm range which usually allows for less expensive equipment. Gutiérrez et al. (2015a) has separated the largest number of varieties up to now, 20, using in the study 20 samples per variety. There are other works where the number of samples per variety was significantly larger, namely Cao et al. (2010) that used between 115 and 197 samples, Arana et al. (2005) with 144 samples and Diago et al. (2013) with 100. However, these works classified only 3, 2 and 3 varieties, respectively. These values have been

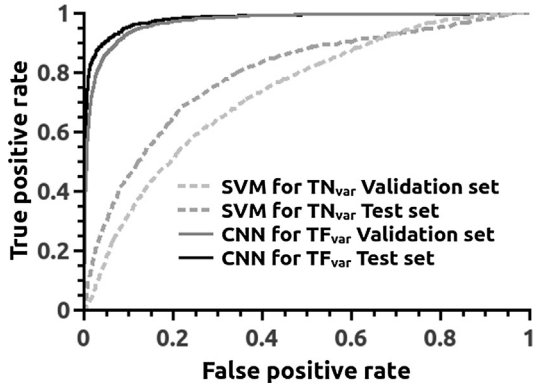


**Fig. 4.** Relative operating characteristic curves for SVM of Touriga Nacional ($TN_{var}$) and CNN of Touriga Franca ($TF_{var}$) in both validation and test sets.

clearly surpassed in the present work by separating Touriga Nacional ($TN_{var}$) or Touriga Franca ($TF_{var}$) from 63 other varieties, using a total of 35,833 spectra from which 3345 and 3569 were from $TN_{var}$ and $TF_{var}$, respectively.

In the previously published works mentioned above reaching more than 90% of correct classification for a certain variety was quite common, even though in Gutiérrez et al. (2015a) there were seven varieties out of twenty whose correct classification percentages were smaller than 90%. Gutiérrez et al. (2015a) separated Touriga Nacional from other varieties, as we now do, and reached a 100% classification percentage for both $TN_{var}$ and non-$TN_{var}$ spectra. The expectable impact of scaling up the problem was a larger difficulty to reach these large classification efficiencies. This is exactly what happened for $TN_{var}$ and non-$TN_{var}$ spectra in the present work with correct classification percentages now dropping to 81.90% and 63.02%, respectively. This means that the amount of false positives, i.e., spectra that are not from $TN_{var}$ but are classified as $TN_{var}$ is still too large, namely 36.08%, which is equal to one minus the correct classification percentage for non-$TN_{var}$ spectra. This drop shows that previous results with less samples and varieties must be viewed with caution as they may not hold with more demanding datasets. Nevertheless, the scaling up might not be the only major influence on the results because $TF_{var}$, that was analysed under the same conditions as $TN_{var}$, reached correct classification percentages of 93.82% and 91.63% for $TF_{var}$ and non-$TF_{var}$ patterns, respectively. Unfortunately, $TF_{var}$ was not used by Gutiérrez et al. (2015a) so that a comparison could be made. The results suggest that some varieties are easier to separate than others independently of the number of samples or varieties under analysis. With this in mind it is possible that the drop

**Table 5**
Confusion matrix in the test set for the best classifier from Table 4, the CNN. The values correspond to number of samples.

| | | Predicted variety | |
|---|---|---|---|
| | | Touriga Franca | Not-Touriga Franca |
| Actual variety | Touriga Franca | 1077 | 71 |
| | Not-Touriga Franca | 708 | 7750 |

in $TN_{var}$ results relative to Gutiérrez et al. (2015a) is originated by a harder to separate group of varieties in the present work. Another possible justification for the decrease in $TN_{var}$ classification percentages could be the fact that the miniaturized spectrometer did not provide data with sufficient quality for variety separation, however, the better results for $TF_{var}$ seem to contradict this hypothesis.

The difference in the results of $TN_{var}$ and $TF_{var}$ in the present work is most probably due to the existence of one or various leaf features in $TF_{var}$ that lead to spectra which are more easily distinguishable from other varieties spectra than the $TN_{var}$ spectra, however, conducting the (complex) study to understand how reflectance spectra are influenced by a large number of leaf features is out of the scope of the present article; consequently, this hypotheses cannot be proven or refuted at this time. As seen in the $TF_{var}$ case, machine learning methods can be employed with good results without the explicit knowledge of how the leaf features affect the reflectance spectra. This is fundamental in cases such as the present one where the reflectance spectra have extremely complex relations with the sample features. However, in the $TN_{var}$ case it might be necessary to gain some understanding of these relations in order to achieve better classification results or, at least, to reach the conclusion that the results cannot be improved.

In the future, a possibility to improve the obtained results is the use of multiple measurements from the same grapevine. If the correct classification percentages obtained here are observed per plant then it will be possible to correctly classify 100% of the grapevines, even for $TN_{var}$, because there will always be a larger number of measurements in the plant with the correct classification rather than with the wrong one.

## 5. Relevance of the developed method

Up to now the available methods for variety identification, ampelography and DNA based analysis are not effective in terms cost or measurement time. Ampelography requires a long training time of the experts in order to be safely applied and the analysis of each plant cannot be done in a few seconds because it requires analysing various plant traits. DNA analysis cannot be made in the field neither in a few

**Table 4**
Calculation results for Touriga Franca. The best SVM is in bold and the best overall result is in Gray.

| Classifier | | Kernel | SG | log | MSC | SNV | Der | AUROC Val | Validation 1 | | Validation 2 | | AUROC Test | Test | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Not $TF_{var}$ (%) | $TF_{var}$ (%) | Not $TF_{var}$ (%) | $TF_{var}$ (%) | | Not $TF_{var}$ (%) | $TF_{var}$ (%) |
| Support vector machines | 1 | Poly | 71 | Yes | No | No | No | 0.83 | 74.39 | 83.87 | 64.5 | 80.11 | | | |
| | 2 | Poly | 71 | Yes | No | No | 1 | 0.95 | 88.62 | 89.10 | 94.54 | 84.27 | | | |
| | 3 | Poly | 71 | Yes | No | No | 2 | 0.68 | 64.39 | 65.41 | 58.72 | 69.89 | | | |
| | 4 | RBF | 71 | Yes | No | No | No | 0.91 | 92.41 | 77.47 | 84.51 | 83.93 | | | |
| | 5 | RBF | 71 | Yes | No | No | 1 | 0.95 | 87.90 | 88.08 | 93.55 | 87.98 | | | |
| | 6 | RBF | 71 | Yes | No | No | 2 | 0.61 | 55.81 | 53.78 | 68.65 | 55.17 | | | |
| | 7 | RBF | 51 | Yes | No | No | 1 | 0.95 | 86.50 | 87.06 | 91.36 | 86.85 | | | |
| | 8 | RBF | 91 | Yes | No | No | 1 | 0.95 | 88.55 | 89.39 | 92.21 | 88.54 | | | |
| | 9 | RBF | No | Yes | No | No | 1 | 0.63 | 68.48 | 53.49 | 68.67 | 50.56 | | | |
| | 10 | RBF | 91 | No | Yes | No | 1 | 0.80 | 82.77 | 65.41 | 72.04 | 67.75 | | | |
| | 11 | **RBF** | **91** | **Yes** | **No** | **Yes** | **1** | **0.96** | **89.24** | **90.26** | **94.79** | **86.63** | **0.97** | **92.94** | **89.2** |
| | 12 | RBF | 91 | No | No | Yes | 1 | No set of parameters allowed to have both validation results with difference smaller than 10% | | | | | | | |
| CNN | | - | 91 | Yes | No | Yes | 1 | 0.97 | 86.49 | 93.31 | 93.89 | 93.15 | 0.98 | 91.63 | 93.82 |

minutes; it can only be applied by highly trained personnel in well equipped laboratories, and involves consumables use. All this prevents the application of these methods to the classification of large number of plants in a reduce time period. The method now proposed can classify one plant in just a few seconds and this can be done by an ordinary person after minimal training, which is impossible with the other two methods. This way, growers can control themselves the variety of the plants in their vineyards. In addition, the method involves only investment cost since it does not require any consumables for each analysis. Making variety classification available to everyone has advantages for both producers and consumers by preventing plant mislabeling which improves wine trueness-to-type as well as wine quality by avoiding improper variety blends.

## 6. Conclusions

The present work has shown that it is possible to separate spectra of leaves from the grapevine varieties Touriga Nacional ($TN_{var}$) or Touriga Franca ($TF_{var}$) from spectra of 62 other varieties plus $TF_{var}$ or $TN_{var}$, respectively, when more than 35,000 spectra are used, even though the efficiency of this separation can be rather different depending on the varieties used. The work has also shown that it is possible to collect these large amounts of data in a relatively small amount of time, namely four days. The amount of varieties and samples employed represents an unprecedented quality improvement in the representativeness of the data used in this type of research.

In the case of $TN_{var}$, the support vector machine (SVM) provided better test results than the convolutional neural network (CNN) and allowed to correctly classify 63.02% of the non-$TN_{var}$ spectra and 81.90% of the $TN_{var}$ spectra. For $TF_{var}$ it was the CNN that gave the best results with the non-$TF_{var}$ and the $TF_{var}$ spectra having correct classification percentages of 91.63% and 93.82%, respectively. The classification percentages were obtained for a test set whose data was gathered in a different day of the training and validation data which increases the results credibility since the impact of variations between the days such as biological material evolution, changes in solar lighting or water availability in the soil was also taken into account. The significantly better results for $TF_{var}$ than for $TN_{var}$ indicate that it is easier to separate some varieties than others and also that the scaling up of the number of spectra and varieties considered in the present work does not necessarily lead to smaller classification efficiencies. In addition, the large classification percentages for $TF_{var}$ suggest that the data gathered with the miniature spectrometer has enough quality to allow a highly efficient identification of spectra at least for some varieties. Even though the correct classification percentages for $TN_{var}$ were not as good as one would like there is no doubt that the classification was better than random. This opens up the possibility of having good plant classification percentages by making a majority vote with multiple measurements of the same plant, but only if the classification percentages obtained for multiple plants still hold for a single plant.

In spite of the large classification efficiencies obtained for Touriga Franca the technology has still a long way to go because it is necessary to prove that classifiers trained with data from a certain year and location can correctly classify data from other years and locations, which must be done using a large number of samples and varieties. It is also important to understand how many times per year should data be collected in order to have an all year long valid classifier. Obtaining data to study all these topics will be quite demanding but necessary to make the technology evolve. Finally, developing methods to understand why Touriga Nacional presents worse results than Touriga Franca would probably improve our knowledge on how the variety separation is being made and might allow finding ways to improve this separation.

## Acknowledgements

The authors thank Mr. António Manuel Fernandes for his help in all the logistics related to the experiments. Armando Fernandes acknowledges a post doctoral grant with number SFRH/BPD/108060/2015 from Fundação para a Ciência e a Tecnologia. The authors acknowledge financial support through projects: National Funds by FCT - Portuguese Foundation for Science and Technology, under the project UID/AGR/04033/2019; project INTERACT – “Integrative Research in Environment, Agro-Chains and Technology”, no. NORTE-01-0145-FEDER-000017, in its line of research entitled VitalityWine, co-financed by the European Regional Development Fund (ERDF) through NORTE 2020 (North Regional Operational Program 2014/2020); and project –¨Symbiotic technology for societal efficiency gains: Deus ex Machina (DEM)¨,no. NORTE-01-0145-FEDER-000026, co-financed by the European Regional Development Fund (ERDF) through NORTE 2020 (North Regional Operational Program 2014/2020).

## Appendix A. List of used varieties

The list uses the Portuguese names in general but provides also a well-known international name in parenthesis for non-Portuguese varieties. In front of each variety is given the grape color, “R”, “W” and “Ro” for red, white and rosé, respectively.

| Variety | Country of origin | Number of plants | Estimated number of samples |
|---|---|---|---|
| Touriga Nacional R | Portugal (PRT) | 6 | See exact values in Table 1 |
| Touriga Franca R | PRT | 7 | See exact values in Table 1 |
| Alfrocheiro R | PRT | 7 | 322 |
| Alicante Bouschet R | France (FR) | 5 | 279 |
| Alvarinho W | PRT | 5 | 230 |
| Aragonez R (Tempranillo Tinto) | Spain (SP) | 17 | 936 |
| Azal W | PRT | 6 | 290 |
| Baga R | PRT | 11 | 613 |
| Batoca W | PRT | 14 | 614 |
| Beba W | SP | 7 | 339 |
| Bical W | PRT | 18 | 787 |
| Boal Espinho W | PRT | 8 | 354 |
| Cabinda R | PRT | 4 | 205 |
| Caínho W | PRT | 6 | 253 |
| Carrasquenho R | PRT | 7 | 390 |
| Castelão Branco W | PRT | 6 | 305 |
| Castelão R | PRT | 29 | 1484 |
| Cornifesto R | PRT | 7 | 322 |
| Dedo de Dama Ro | PRT | 5 | 261 |
| Diagalves W (Mantuo) | SP | 7 | 346 |
| Dona Maria W | PRT | 7 | 346 |

| | | | |
|---|---|---|---|
| Douradinha W | PRT | 4 | 204 |
| Fernão Pires W | PRT | 32 | 1549 |
| Fernão Pires Rosado Ro | PRT | 3 | 145 |
| Folha de Figueira W | PRT | 5 | 211 |
| Gewürztraminer Ro | Germany/Italy (IT) | 6 | 313 |
| Gouveio W | PRT | 6 | 305 |
| Jaen R (Mencía) | SP | 13 | 570 |
| Jampal W | PRT | 7 | 346 |
| Loureiro W | SP | 7 | 295 |
| Malvasia Fina W | PRT | 17 | 839 |
| Malvasia Fina Roxa Ro | PRT | 6 | 305 |
| Marufo R | PRT | 13 | 598 |
| Moscatel Graúdo W (Muscat of Alexandria) | GREECE | 14 | 684 |
| Mourisco R | PRT | 7 | 365 |
| Nevoeira R | PRT | 6 | 276 |
| Patorra R | PRT | 7 | 322 |
| Perle de Csaba W (Csaba Gyoengye) | HUNGARY | 7 | 342 |
| Pinot Blanc W | FR | 7 | 295 |
| Praça W | PRT | 7 | 295 |
| Primavera R | PRT | 5 | 256 |
| Rabigato W | PRT | 6 | 266 |
| Rabo de Ovelha W | PRT | 26 | 1152 |
| Rodo R (Mondeuse Noir) | FR | 7 | 365 |
| Sarigo W (Cayetana Blanca) | SP | 6 | 266 |
| Sezão R | PRT | 7 | 339 |
| Síria W | PRT | 50 | 2022 |
| Sultanina W | TURKEY | 7 | 342 |
| Tamarez W | PRT | 21 | 843 |
| Tinta Caiada R (Parraleta) | SP | 21 | 1043 |
| Tinta Francisca R | PRT | 11 | 482 |
| Tinta Miúda R (Graciano) | SP | 11 | 559 |
| Tinta Negra R | PRT | 12 | 556 |
| Tintem R | PRT | 7 | 307 |
| Tinto Pegões R | PRT | 2 | 88 |
| Tália W (Trebbiano Toscano) | IT | 11 | 549 |
| Trincadeira R | PRT | 7 | 339 |
| Five different varieties with unknown names | | 26 | 1210 |
| Vinhão R | PRT | 7 | 339 |
| Vital W | PRT | 13 | 661 |

## References


Arana, I., Jaren, C., Arazuri, S., 2005. Maturity, variety and origin determination in white grapes (Vitis vinifera L.) using near infrared reflectance technology. J. Infrared Spectrosc. 13, 349–357.

Butt, R., 2010. Introduction to Numerical Analysis using MATLAB. Jones and Bartlett Publishers, Sudbury, Mass.

Cao, F., Wu, D., He, Y., 2010. Soluble solids content and pH prediction and varieties discrimination of grapes based on visible–near infrared spectroscopy. Comput. Electron. Agric. 71, S15–S18. https://doi.org/10.1016/j.compag.2009.05.011.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 1–27. https://doi.org/10.1145/1961189.1961199.

Cunha, J., Ibáñez, J., Teixeira-Santos, M., et al., 2016. Characterisation of the Portuguese grapevine germplasm with 48 single-nucleotide polymorphisms. Aust. J. Grape Wine Res. 22, 504–516. https://doi.org/10.1111/ajgw.12225.

Diago, M.P., Fernandes, A.M., Millan, B., et al., 2013. Identification of grapevine varieties using leaf spectroscopy and partial least squares. Comput. Electron. Agric. 99, 7–13. https://doi.org/10.1016/j.compag.2013.08.021.

Fernandes, A., Gomes, V., Melo-Pinto, P., 2018. A review of the application to emergent subfields in viticulture of local reflectance and interactance spectroscopy combined with soft computing and multivariate analysis. In: Cruz Corona, C. (Ed.), Soft Computing for Sustainability Science. Springer International Publishing, Cham, pp. 87–115.

Fernandes, A.M., Melo-Pinto, P., Millan, B., et al., 2014. Automatic discrimination of grapevine (Vitis vinifera L.) clones using leaf hyperspectral imaging and partial least squares. J. Agric. Sci. 1–11. https://doi.org/10.1017/S0021859614000252.

Fernandes, A., Utkin, A., Eiras-Dias, J., et al., 2018. Assessment of grapevine variety discrimination using stem hyperspectral data and AdaBoost of random weight neural networks. Appl. Soft Comput. 72, 140–155. https://doi.org/10.1016/j.asoc.2018.07.059.

Gomes, V., Fernandes, A., Martins-Lopes, P., et al., 2017. Characterization of neural network generalization in the determination of pH and anthocyanin content of wine grape in new vintages and varieties. Food Chem. 218, 40–46. https://doi.org/10.1016/j.foodchem.2016.09.024.

Gutiérrez, S., Tardaguila, J., Fernández-Novales, J., Diago, M.P., 2015b. Data mining and non-invasive proximal sensing for precision viticulture. In: Proceedings of the 2nd Int. Electron. Conf. Sens. Appl.

Gutiérrez, S., Tardaguila, J., Fernández-Novales, J., Diago, M.P., 2015a. Support vector machine and artificial neural network models for the classification of grapevine varieties using a portable NIR spectrophotometer. PloS One 10, e0143197.

Gutiérrez, S., Tardaguila, J., Fernández-Novales, J., Diago, M.P., 2016. Data mining and NIR spectroscopy in viticulture: applications for plant phenotyping under field conditions. Sensors 16, 236.

Qiu, Z., Chen, J., Zhao, Y., et al., 2018. Variety identification of single rice seed using hyperspectral imaging combined with convolutional neural network. Appl. Sci. 8, 212. https://doi.org/10.3390/app8020212.

Ranking de castas, 2017. Inst Vinha E Vinho. http://www.ivv.gov.pt/np4/35/?newsId=7501&fileName=Ranking_Castas_2017.xls.

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

Tassie, L., 2010. Vine identification–knowing what you have. Grape Wine Res Dev Corp Fact Sheet Aust Gov.

Tomic, L., Stajner, N., Javornik, B., 2013. Characterization of grapevines by the use of genetic markers. In: Sladonja, B. (Ed.), The Mediterranean Genetic Code - Grapevine and Olive. InTech.

Yang, H.Q., Luo, W.Q., Wang, W.J., 2012. Nondestructive discrimination of grape seed varieties using UV-VIS-NIR spectroscopy and chemometrics. Appl. Mech. Mater. Trans. Tech. Publ. 89–94.