Talanta Open 12 (2025) 100494

Contents lists available at ScienceDirect

Talanta Open

journal homepage: www.sciencedirect.com/journal/talanta-open

Raman spectroscopy as a rapid tool for monitoring lactic acid concentration during wine malolactic fermentation directly in the winery

Anna Lisa Gilioli [a,b], Alessio Sacco [b,*], Andrea Mario Giovannozzi [b], Simone Giacosa [c], Antonella Bosso [d], Loretta Panero [d], Silvia Raffaela Barera [d], Stefano Messina [d], Marco Lagori [c,d], Silvia Motta [d], Massimo Guaita [d], Ettore Vittone [a], Andrea Mario Rossi [b]

[a] Physics Department, University of Turin, Via P. Giuria 1, 10125 Turin, Italy
[b] Istituto Nazionale di Ricerca Metrologica (INRiM), Strada delle Cacce 91, 10135 Turin, Italy
[c] Department of Agricultural, Forest and Food Sciences Department, University of Turin, Corso Enotria 2/C, 12051 Alba, Italy
[d] Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CREA), Via Pietro Micca 35, 14100 Asti, Italy

ARTICLE INFO

Keywords:
Raman spectroscopy
PLS regression
Lactic acid
Wine
Malolactic fermentation
Red winemaking

ABSTRACT

Lactic acid is mainly produced during the process of malolactic fermentation and evolution of its concentration is associated with the wine stabilization process and the quality of the final product. The quantitative analysis of lactic acid is carried out offline in the laboratory using various analytical techniques, the most used being high performance liquid chromatography (HPLC). Because of this, there is a clear demand in the winemaking community for analytical tools that allow real-time, fast and inexpensive quantification of lactic acid. An approach using Raman spectroscopy has positioned itself as a feasible alternative in this regard. The primary goal of this work is therefore to monitor the concentration of lactic acid (which changes rapidly during the malolactic fermentation process) in the analysed samples, specifically, Nebbiolo wine samples for making the Barolo wine. The collected Raman spectra using a portable Raman apparatus are processed using an algorithm that applies Partial Least Squares (PLS) regression to determine the lactic acid concentration for each sample. It proves to be a precise and reliable method that leads to the determination of a predictive model characterised by $R^2 = 0.76$ (on the validation set), $R^2 = 0.94$ (on the test set) and RMSE of the lactic acid concentration predicted by the model of 0.22 g/l (on the validation set) and 0.11 g/l (on the test set) respectively. This approach produces results comparable to those obtained via HPLC. Moreover, unlike the latter, it allows rapid and easy monitoring of the lactic acid concentration during fermentation directly in the winery.

Glossary

BT	back-thinned
CCD	charge-coupled device
CV	cross validation
DSS	decision supporting system
HCAs	hydroxycinnamic acids
HPLC	high performance liquid chromatography
LV	latent variable
MIR	mid-infrared
ML	machine learning
MSC	multiplicative scatter correction
MSE	mean square error
MST	mean total sum
NIR	near-infrared
ODR	orthogonal distance regression
PCR	principal component regression
PLS	partial least square
RMSE	root mean square error

Introduction

Lactic acid is a useful indicator for monitoring fermentation processes, and its concentration is closely related to the flavor and texture of the final product [1]. It is mainly produced during malolactic fermentation, an oenological process (lasting from a few weeks to months) in which lactic acid bacteria convert malic acid, with a sour taste, into lactic acid and carbon dioxide, with a softer flavor. This transformation

* Corresponding author.
E-mail address: a.sacco@inrim.it (A. Sacco).

https://doi.org/10.1016/j.talo.2025.100494
Received 8 January 2025; Received in revised form 23 May 2025; Accepted 28 May 2025
Available online 29 May 2025

contributes to the production of wines characterized by greater balance and persistence [2]. This transformation is often intentional (especially in red wines) and normally follows the alcoholic fermentation: a process by which ethyl alcohol and carbon dioxide are produced from the sugars contained in the grape must. Determining the concentration of malic and lactic acid allows malolactic fermentation to be monitored; this permits the evaluation of the evolution and the end of this transformation and the ability to act in favour of the protection of wine against spoilage, ensuring that the wine reaches biological stability and high qualitative standards.

Traditional methods for determining lactic acid concentration in wines include chromatography [3] and colorimetric techniques [4]. The former is the most commonly used approach, although it requires expensive and time-consuming equipment because the analysis is typically conducted in an outside laboratory. As a result, there is a growing need for on-site, real-time monitoring solutions to track lactic acid concentration during fermentation. To monitor lactic acid concentration and indicatively follow the fermentation process more efficiently, real-time and on-site monitoring methods are preferred. In this regard, vibrational spectroscopy techniques [5] and biosensor devices [6,7] have shown promise as viable alternatives. In particular, modern instrumentation using near-infrared (NIR), mid-infrared (MIR) and Raman spectroscopy has found wide applications in the food and beverage industry, including alcoholic beverage production [8,9]. They allow rapid and non-destructive measurements, have high specificity and sensitivity, and allow on-site measurements to be performed with portable instrumentation.

Prominent among these optical investigation techniques is Raman spectroscopy, which, due to the limited interference of water signals, is a valuable tool for analysing samples in aqueous environments. It is a material analysis technique that exploits the Raman effect in order to study the chemical composition of the sample by analysing the light beam scattered inelastically from it [10]. To the best of our knowledge, this technique has not yet been used to determine the concentration of lactic acid during malolactic fermentation in wine, probably due to the complexity of analysing the Raman spectra of lactic and malic acid, which in fact do not show clearly evident characteristics in the spectra of the wine under investigation. However, this difficulty can be curbed by integrating spectroscopic analyses with machine learning (ML) and multivariate analysis techniques. In fact, as technology has evolved, the latter have become increasingly useful tools for analytical science, to perform high-yield analyses and measurements of the sample [11,12]. By applying such techniques, it is possible to determine the pattern followed by a pre-labelled dataset to make predictions on new datasets and greatly accelerate experimental analysis and calculation. The use of these multivariate and ML methods has effectively contributed to a wide range of research, including spectroscopy data processing and its practical applications in the fields of analytical, physical and organic chemistry to obtain quantitative and qualitative information on the chemical composition of samples of interest. Consequently, there is a growing preference for the combined use of standard instrumental methods with multivariate and ML methods to determine patterns, optimise and automate the analysis of samples. In this study, multivariate analysis technique is used to analyse Raman spectra in order to identify data correlations and perform pattern recognition [13].

In particular, the aim of this work is to use Raman spectroscopy combined with multivariate analysis techniques to determine the concentration of lactic acid during malolactic fermentation directly in the winery. This option is being studied as part of the QualShelL 'Wine Quality and Shelf-Life' project [14], which involves implementing new procedures for assessing grape quality and monitoring oenological processes in order to achieve high wine quality standards and to have rapid and innovative analytical tools to support decisions on oenological interventions in the winery (precision oenology). This study is carried out on samples of Nebbiolo wine for making the Barolo wine[1] [15] and the results obtained are compared with the currently most widely used method: high performance liquid chromatography (HPLC) [16].

The study conducted led to the determination of a valid approach for the determination of lactic acid concentration in wine samples. This method, unlike conventional methodologies, is fast, user-friendly, non-destructive and allows on-site measurements to be carried out; it also provides results that are consistent with those produced by the methodologies currently in use.

Materials and methods

Experimental design

The following experimental design was prepared in order to determine the concentration of lactic acid in wines during malolactic fermentation using Raman spectroscopy.

- Analysis of Nebbiolo wine samples obtained from 4 different wineries.
- Acquisition of Raman spectra of samples whose lactic acid content changed during malolactic fermentation using portable instrument. Measurements were carried out on 13 dates in one and a half months (November to December 2022) by taking 3 independent measurements for each date and cellar.
- Analysis of 150 Raman spectra[2] using multivariate techniques. Creation of a predictive model based on lactic acid concentration analysed through the method currently used (HPLC).
- The model was applied in order to predict the lactic acid concentration of 5 unknown Nebbiolo wine samples.

Materials

The samples analysed consist of Nebbiolo wines, collected after the alcoholic fermentation, destined to the production of Barolo designated wine. They were taken directly from the stainless steel tanks during the malolactic fermentation period. The wines under examination came from four different wineries, with codes A, B, C and D.

HPLC determination of lactic acid in Nebbiolo wine samples

Each wine sample underwent a preparation step. Specifically, 1 ml of sample was taken and then acidified with the same volume of 1 N o-phosphoric acid (in order to shift the acid balance towards the non-associated form). Subsequently, the sample was passed through a C18 cartridge to separate the hydrophilic phase (containing the acids) from the lipophilic phase (containing the polyphenolic fraction). The first phase is then collected in a 10 ml flask (together with the water used in the cartridge washing step following the passage of the wine) which is brought to volume using $5 \cdot 10^{-3}$M o-phosphoric acid. Next, the wine sample is subjected to filtration (pore diameter 0.2 μm) and collected in a 4 ml vial. Finally, to perform the quantitative analysis, the sample inside the vial, containing the lactic acid, is injected into the HPLC. The

[1] The samples analysed are wines from Nebbiolo grapes harvested in wineries in the Barolo production area. Following a specific period of ageing, this wine is defined as 'Barolo wine'. At the time of analysis, the above-mentioned ageing phase has not yet taken place, so we simply define the samples as Nebbiolo wines for making the Barolo wine.

[2] For one winery, the Raman spectra of the last two dates are missing (i.e.: 6 acquisitions in total). Therefore, the total of 156 spectra is reduced to 150.

quantitative determination of lactic acid is carried out by comparison with an external standard and construction of a calibration line, with four concentration points equal to 0.7–1.3–2.0–2.7 g/L of lactic acid. Each of the four points was prepared and injected three times [17]. The chromatographic conditions and acquisition parameters of the instrumental apparatus used are given below.

- Eluent: phosphoric acid ($5{\cdot}10^{-3}$M);
- flow rate: 0.6 ml/min;
- volume injected: 20 μl;
- HPLC column: RP C18, 4 mm × 250 mm;
- UV–Vis detector at 210 nm.

The output is a chromatogram for each injected sample (i.e. for each sampling point and winery) from which the mass concentration of lactic acid can be determined.

Raman spectrum acquisition of Nebbiolo wine samples

Measurement in three replicates (spectra) for each sampling point and winery were acquired in order to obtain repeated measurements for each population element. The three measurements were carried out considering a different portion of the sample each time. They can therefore be considered independent measurements.

To perform the Raman analysis, 3 ml of sample were taken and placed into a quartz cuvette. Then, using a customised portable Raman apparatus (using an optical fibre for illumination and collection), the Raman spectrum of the sample under investigation is acquired. The acquisition parameters used to obtain the spectra are given below.

- Spectrometer: Exemplar Pro (B&W Tek, Plainsboro, NJ, USA) featuring a highly sensitive deep cooled (−25 °C) back-thinned (BT) CCD detector, a spectral range of 190 −1100 nm and a spectral resolution of 0.6 nm;
- laser: BRM-785E (B&W Tek, Plainsboro, NJ, USA) characterised by a power output of 300 mW and a wavelength of 785 nm;
- optical fiber coupled Raman Trigger Probe BAC102 (B&W Tek, Plainsboro, NJ, USA): the probe is characterised by a working distance of 5.5 mm and the fibre has a numerical aperture of 0.22;
- acquired range of Raman shift: [50, 3400] cm^{-1};
- exposure time: 10 s;
- number of scans: 3;
- average spectral resolution: 11 cm^{-1};
- edge filter characterised by a Raman cut-off at 150 cm^{-1}

Data preprocessing

Before analysing the data, it is necessary to preprocess the acquired spectra in order to correct and transform them to be suitable for the statistical elaboration.

First, the tail of the Rayleigh peak was removed because it contained no information about the samples and was influenced by both elastically scattered photons and the absorption of the edge filter in the spectrometer; the noise was then reduced using the Savitzky-Golay smoothing algorithm implemented with the SciPy library [18] (using a number of points in the window equal to 9 and a polynomial of order 2). Multiplicative scatter correction (MSC), a data transformation technique used to compensate for additive and/or multiplicative effects, was then applied. Through this technique, the signal dispersion for each pixel can be reduced by correcting the data using a reference spectrum (the mean of all spectra) [19]. Next, the baseline, given by the fluorescence phenomenon, was removed by an iterative process using a polynomial of order seven and a tolerance factor of 10^{-4}. Finally, an L2 normalization was applied to the intensities of each spectrum; this step rescales these to have unit norm. In this way the spectra have the same scale and the data are comparable to each other. Hence, we obtain the pre-processed Raman spectra.

PLS regression application

After the data have been processed, it is possible to proceed to the dataset creation phase. The dataset consists of an X matrix and a Y column vector:

- The X matrix contains the intensity of each Raman spectrum. It is an $M \times N$ matrix where M is equal to the number of total spectra and N is equal to the number of RS acquired for each spectrum; in this case: 150 × 1545. So, for example, the first row of X matrix contains the intensities of the first Raman spectrum acquired.
- Y is a vector of length M (with M equal to the number of spectra acquired). The elements of this vector are the lactic acid concentrations (g/l) determined using HPLC. For example, the first element of Y is the lactic acid concentration associated with the first acquired spectrum.[3]

In order to study the data, PLS regression was applied to the dataset (this technique was implemented using Python's scikit-learn library [20]). It is a statistical method that, starting from the input data, identifies a linear regression model [21]. This chemometric regression technique was chosen because it is one of the best multivariate methods suitable for handling datasets in collinear situations [22].

The input data are the observed variables (matrix X) and the predicted variables (vector Y). To identify the pattern that follows the data, PLS regression determines the multidimensional direction in X-space that explains the maximum multidimensional variance in Y-space. This technique projects the input data, using an appropriate transformation, into a new space with the objective of maximising the covariance between the X matrix and the Y vector defined in the new space. A new representation of the data is then obtained, defined with new variables called latent variables (LVs), in which dimensionality reduction can be applied and the intrinsic properties of the data can be highlighted [23].

This technique is particularly advantageous in this case as the sample matrix is complex and contains chemical compounds with similar atoms and molecular structures that have similar relevant RS windows. This similarity of output signal is also present in the analysis carried out using traditional methodology. However, applying PLS Regression further improves the discrimination of the compounds under investigation (as the resolution of the hand-held Raman instrument is not sufficient for the analysis of individual peaks for this particular application), leading to a better analysis of the compounds under investigation present in the sample.

Before applying PLS regression, the dataset (containing 150 Raman spectra corresponding to 50 samples) was divided into training, validation and test sets to implement the homologous phases. The training and validation phase was implemented using k-fold group cross-validation (CV) [24]. This is a statistical technique used to evaluate how the results of an analysis generalize to a dataset independent of the one used to build the model. It is a resampling method that uses different parts of the data to validate and train the generic model used over different iterations and works as shown below.

The dataset (which includes both the training and validation sets) is divided into k subgroups called Folds. Making use of a cycle we have that, at each iteration, one subgroup is used as the validation set and the remainder as the training set. The cycle ends when each subgroup has

[3] Lactic acid concentrations were determined by HPLC, used as a reference method, by taking a single measurement for each of the 50 wine samples, as this is an established technique with high precision and repeatability. In contrast, three Raman measurements were taken on different portions of each sample in order to account for intra-sample variability. Therefore, in order to construct the reference vector Y (for the calibration of the PLS model) the concentration obtained from the HPLC was repeated three times, assigning it to each of the corresponding Raman replicates, for a total of 150 measurements.

been used as a validation set; and finally the performance measure of the model is given as the average of the values calculated at each iteration.

This method provides information regarding how well the model generalizes and makes predictions on an independent and unknown dataset.

In this study, k-fold group CV was implemented using k=5, thus dividing the dataset into an 76 % training set and a 24 % validation set (excluding five data samples, corresponding to 15 Raman spectra, used for the testing phase).

Optimal PLS regression model and performance evaluation

The parameter R^2 and *RMSE* are used to determine the model that best describes the data (in the optimal case $R^2 = 1$ and $RMSE = 0$). The former is the coefficient of determination indicating the link between the variability of the data and the correctness of the statistical model used, the latter defines the quality of the predictions.

R^2 is defined by the equation given below [25]:

$$R^2 = 1 - \frac{MSE}{MST} \tag{1}$$

where *MSE* is the mean square error and *MST* is the mean total sum of squares, in particular:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$
$$MST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{2}$$

where: y_i are the measurements for each sample, $\hat{y}_j$ are the expected value for each, $\bar{y}$ is the average of all measurements and n is equal to the number of measurements.

Instead, the *RMSE* is the square root of the *MSE* and is the metric commonly used to assess the accuracy of a forecast model.

From the graphs of R^2 and *RMSE* as a function of the number of latent variables for the training and validation set, it is possible to determine the best model and thus the optimal number of latent variables. The latter corresponds to the abscissa of the minimum *RMSE* on the validation set (considering a value of R^2 that is sufficiently high).

After determining the optimal model, it can be applied to calculate the lactic acid concentration of new samples. In order to determine the accuracy of the model prediction on these new data, that is, the agreement between the measurements made and the predicted values, the *RMSE* (defined above) can be used, but also the *Precision* and *Bias* parameters (whose theoretical minimum value is 0 for both) can be considered. They are defined as follows:

$$Precision = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n-1}}$$
$$Bias = \frac{\sum_{j=1}^{m}\left|\bar{y}_i - \hat{y}_j\right|}{m} \tag{3}$$

where: y_i are the measurements for each sample in the test set, $\bar{y}_i$ is their mean and $\hat{y}_j$ are the expected value for each sample under examination. Additionally, n is equal to the number of replicates of each element of the test set and m is equal to the number of test samples. It is important to note that the second member of the first equation providing the precision is the average (–) calculated on the m samples of the test set.

Results and discussion

Analysis of Nebbiolo wine Raman spectrum and preprocessing

The Raman spectra acquired are shown in Fig. 1. From an initial analysis of the graphs, it can be seen that the spectra are characterised by a baseline given by the fluorescence phenomenon that will need to be corrected. These spectra are then preprocessed in order to transform and correct them appropriately before applying PLS regression, following the steps described in subsection "Data preprocessing". The preprocessed spectra are shown in Fig. 2, where the characteristic peaks of the sample matrix are highlighted.

The peaks present in the Raman shift window [2600 - 3100] cm^{-1} are due to CH_x stretching, while the intensity peak observed around 880 cm^{-1} probably originates from CC stretching of ethanol. The bands around 450 cm^{-1} and 1250 cm^{-1} can be associated with OCC and HCC stretching, respectively. Other peaks of weak intensity can be observed in the [1050–1450] cm^{-1} window and are presumably originated by hydroxycinnamic acids (HCAs). Finally, Raman scattering around 1000 cm^{-1} and 1600 cm^{-1} could be associated with the presence of phenolic compounds [26].

A comparison of the Raman spectra of pure lactic acid and pure malic acid (Figure S1 of the Supplementary Information) with the Nebbiolo wine spectrum (Fig. 2) shows that the characteristic peaks of the acid are

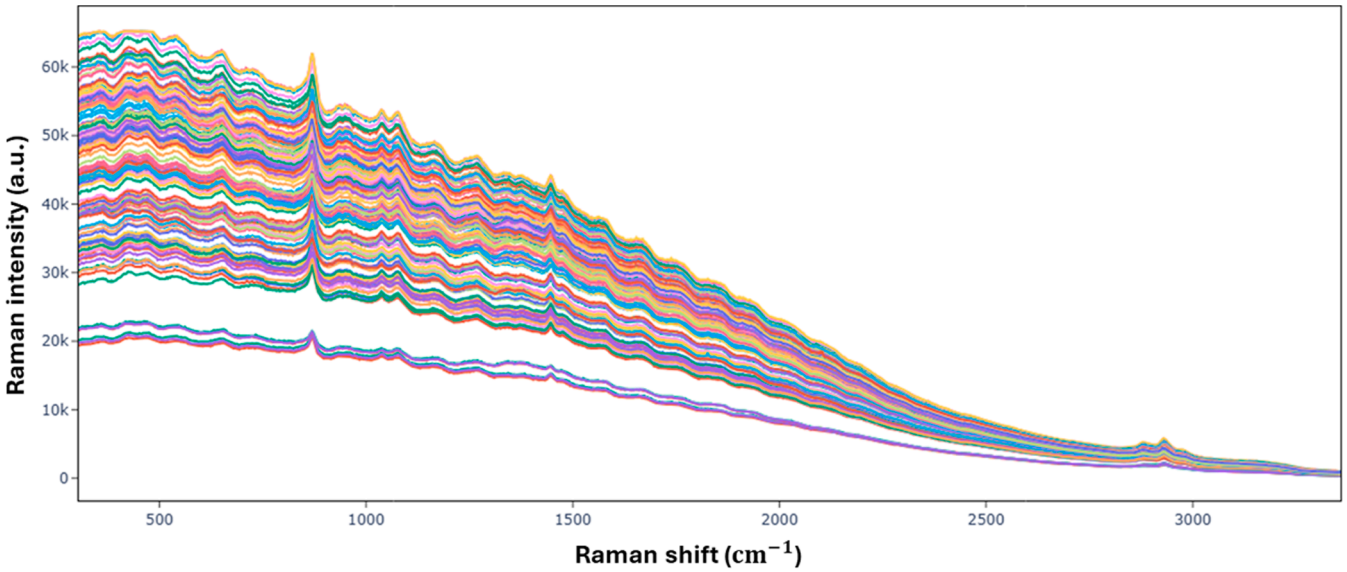

Fig. 1. Complete dataset of Raman spectra of Nebbiolo wine samples.

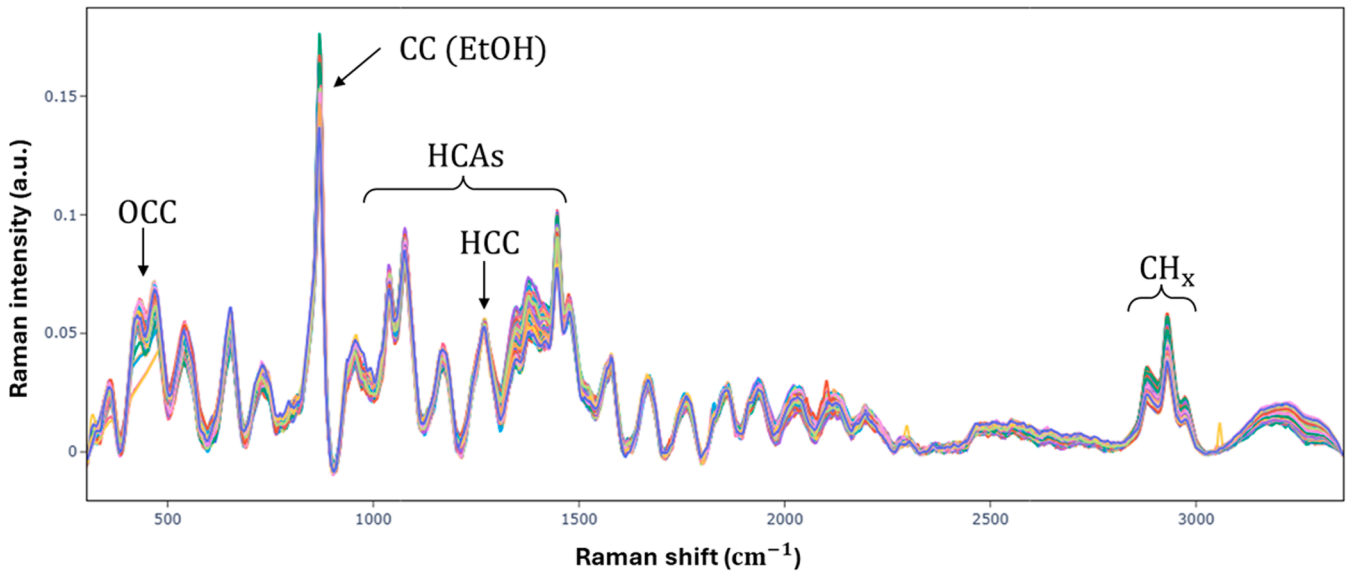

Fig. 2. Pre-processed Raman spectra highlighting the characteristic peaks of the sample under investigation.

obscured by various compounds present in the wine matrix under examination. For this reason, multivariate approach is necessary to find the correlation between the wine Raman spectra and the concentration of lactic acid analysed by HPLC.

PLS regression model

Following the preprocessing of the data, PLS regression is applied to the training set in order to determine the pattern followed by the data. First, a new representation of the data (which makes use of LVs) can be determined, which makes it possible to reduce the dimensionality of the data and highlight their intrinsic properties. For instance, Fig. 3 shows the scores plot of the first two LVs (in the model used in this study, a larger number of latent variables will be used, as explained below), from which it can be seen that:

- The dimensionality of the dataset was reduced; in fact, each spectrum is defined by one point in the scores plot, therefore by only two variables (LV1 and LV2). In contrast, before the application of this technique, each spectrum was defined by each row of the X matrix (1545 variables).
- The graph highlights the similarities and differences in the dataset. In fact, by colouring each point (Raman spectrum) with a colour scale proportional to the concentration of lactic acid in the associated sample, a specific trend followed by the data can be observed. Specifically, lactic acid concentration decreases in the LV1 direction and increases with LV2.

To determine the optimal number of latent variables, *RMSE* and R^2 are plotted as a function of the number of latent variables (for the training set and for the validation set) and the abscissa of the minimum *RMSE* on the validation set is determined; considering a sufficiently high R^2 (Fig. 4). This represents the optimal number of LVs and corresponds to the optimal PLS model describing the data, in this case:

- $RMSE = 0.22\ g/l$ and $R^2 = 0.76$ (on the validation set);
- number of LVs equal to 7.

In order to visually assess the performance of the model, the model's

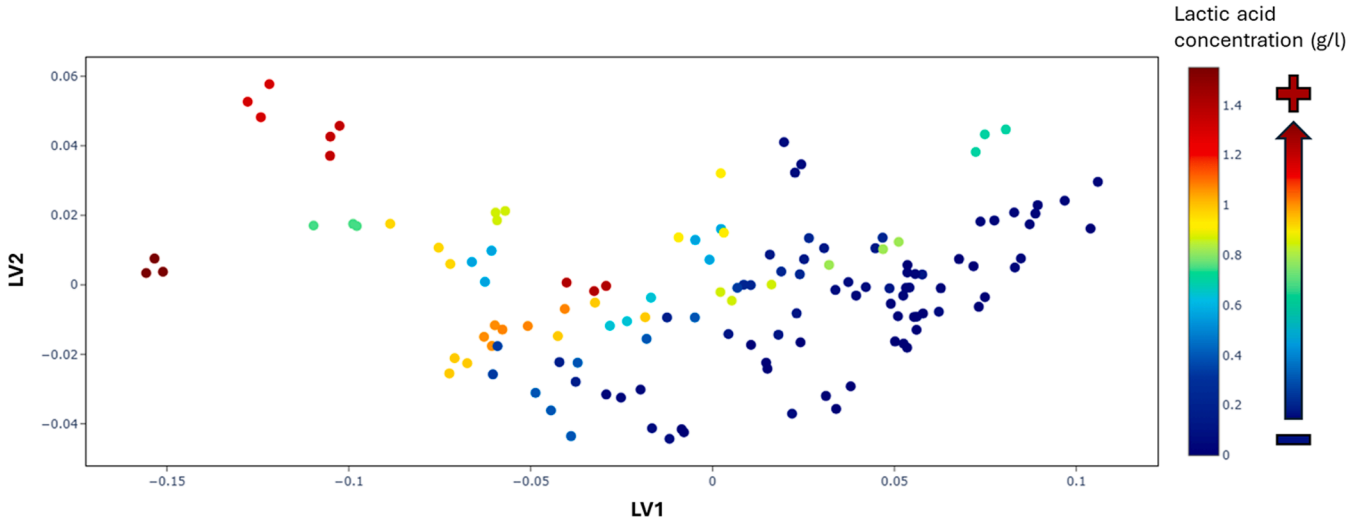

Fig. 3. New data representation. Each point defines a Raman spectrum of Nebbiolo wine sample; it is also coloured following a colour code according to the concentration of lactic acid in the sample.

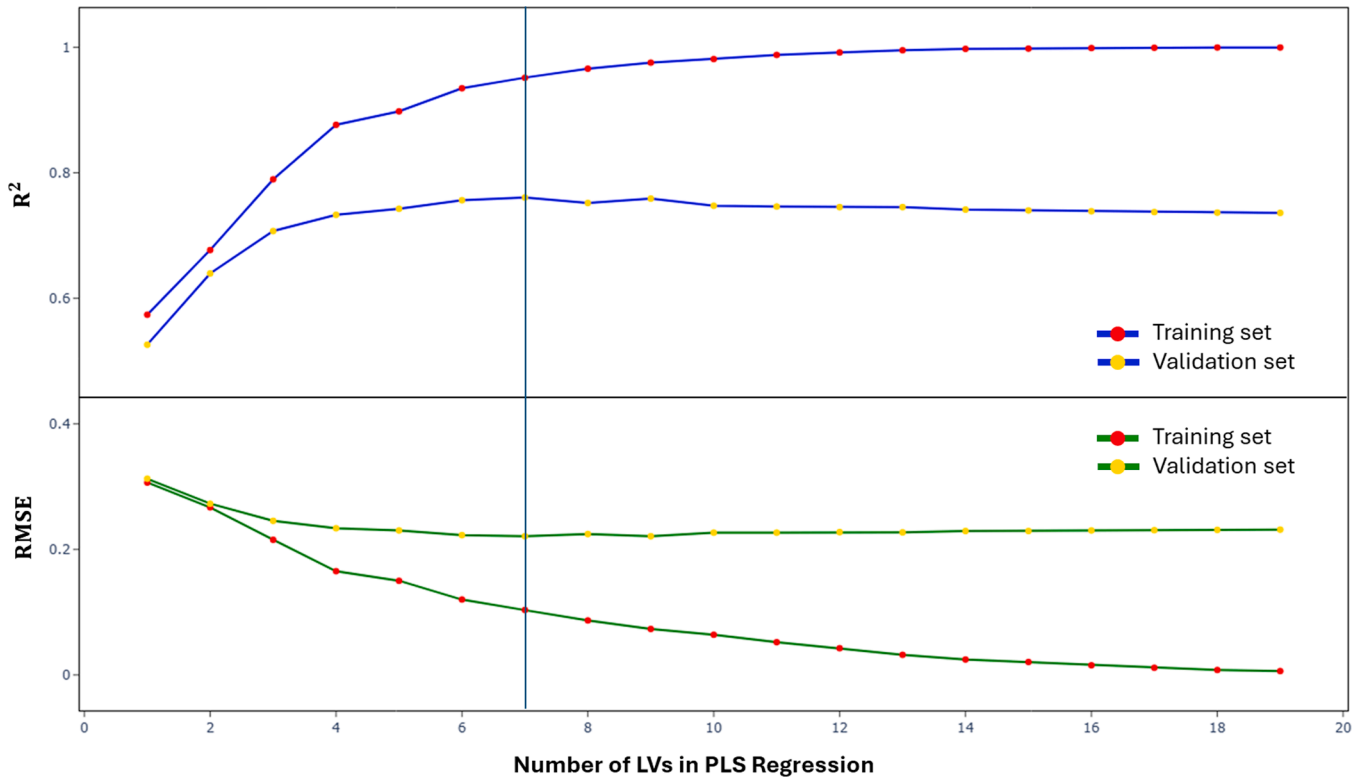

Fig. 4. R^2 and *RMSE* as a function of the number of latent variables, for the training and validation set.

predicted lactic acid concentrations on the calibration set (considering the entire dataset excluding the test set) as a function of those provided by the reference method (HPLC) can be represented in a cartesian graph (Fig. 5). The blue points represent the spectra from the calibration set (training and validation set), whose coordinates are the model-predicted and reference lactic acid concentrations; while the purple curve is the linear fit of these points. The green curve representing the bisector of the first quadrant; indicates the expected trend where the concentrations determined by the two techniques are equal. The linear fit of the calibration set data points was calculated by applying orthogonal distance regression (ODR), a regression technique that considers the uncertainty of both the independent and dependent variable [27]. In this specific case, the uncertainty of the independent variable is the mean uncertainty of the HPLC measurements (equal to 0.09 g/l), while the uncertainty of the dependent variable is the root mean square error (*RMSE*) of the lactic acid concentration predicted by the model on the calibration set (equal to 0.10 g/l).

The following table shows the fitting parameters, Z-score and p-value (Table 1):

Plotting the graph loadings as a function of Raman shift (Figure S2 of

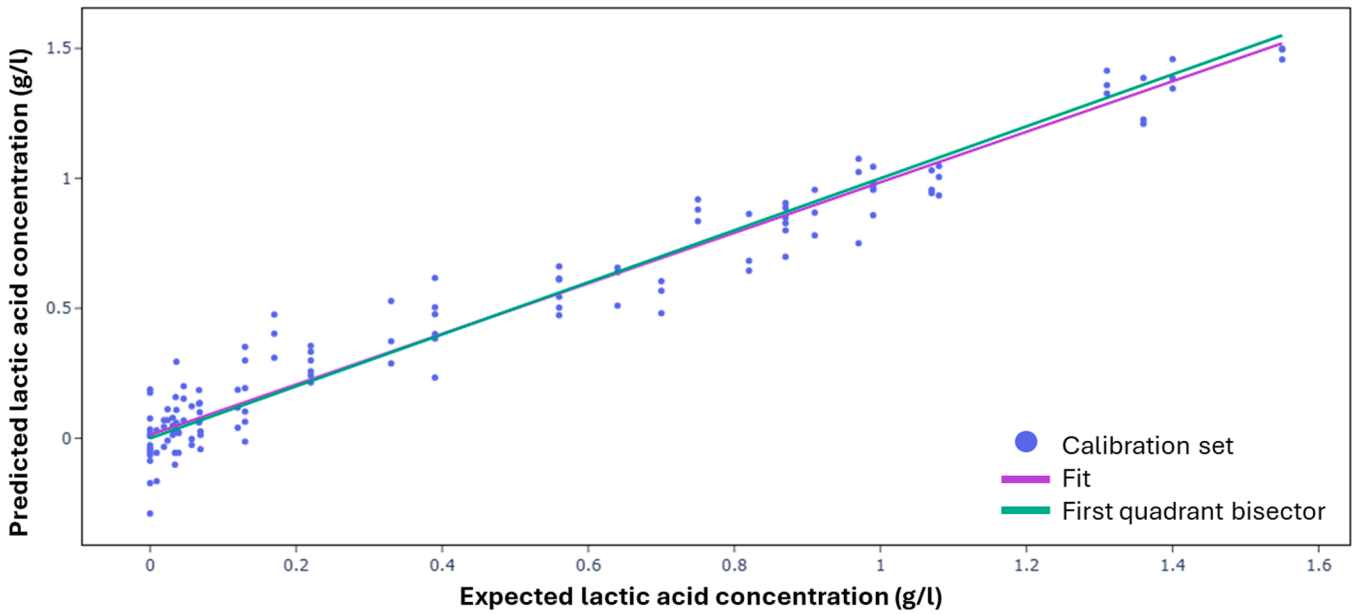

Fig. 5. Lactic acid concentration predicted by the model from Raman spectra as a function of that measured by HPLC for the training set.

Table 1
Fit parameters and statistical test.

Fit parameters	Measurements	Expected values	Z-score	p-value
m	0.97 ± 0.02	1	−1.47	0.14
q	0.01 ± 0.01 g/l	0 g/l	1.01	0.31

the Supplementary Information) it shows the loadings plot of the first latent variable. This graph shows the Raman shift most relevant for the determination of the model; these are those referring to the greater ordinates (in modulus). In this specific case, the most relevant Raman shift is the peak at about 800 cm^{-1} of the ethanol CC stretching and the peak of the CH_x stretching around 3000 cm^{-1} These peaks are among the most intense in the Raman spectrum of lactic acid (Figure S1 of the supplementary information), we therefore have confirmation that the most relevant RS windows in the determination of the pattern are precisely those referring to the most intense peaks in the spectrum of the acid under investigation.

Finally, the histograms of the latent variables (Figure S3 of the Supplementary Information) show the explained variance of *X* (Raman spectra) and of *Y* (lactic acid concentration). By summing up the explained variance of all LVs used in the model (in this case 7) it is possible to know the percent information maintained despite the data transformation. In the new data representation, the retained information contained in *X* is equal to 70 %, while for *Y* is equal to 95 %.

Testing new data

The model was applied to predict the lactic acid concentration of five new Nebbiolo wine samples randomly extracted from the dataset (corresponding to 15 Raman spectra and thus 10 % of the total dataset). Taking into consideration the graph in Fig. 5, but considering the spectra of test set (red dots) instead of those of the calibration set; the plot shown in Fig. 6 is obtained. Furthermore, this graph shows the error bars considering the uncertainty associated with the values predicted by the lactic acid concentration model, equal to the sample standard deviation (σ), and the confidence interval of $(1.96{\cdot}\sigma)$; determined considering a 5 % significance level.

Using *RMSE*, it is possible to determine the accuracy of the model's prediction on the test set, i.e., the agreement between the measurements made and the expected values. It is defined as explained in the subsection 'Optimal PLS regression model and performance evaluation'; the *RMSE* value for the test set is given in the equation below:

$$RMSE = 0.11\ g/l \tag{4}$$

In addition, the *Precision* and *Bias* parameters (also defined in the subsection 'Optimal PLS regression model and performance evaluation') were used to evaluate the performance of the model by considering a number of replicates for each test set element (*n*) equal to 3 and a number of test samples (*m*) equal to 5. The calculated parameters for the five unknown samples are shown below:

$$\begin{aligned} Precision &= 0.07\ g/l \\ Bias &= 0.07\ g/l \end{aligned} \tag{5}$$

The theoretical minimum value of these parameters is zero, so the model is able to predict the lactic acid concentration in unknown samples with high accuracy.

Comparison of HPLC and Raman spectroscopy

Finally, a comparison between the lactic acid concentration predicted by the model and that provided by the HPLC for the five unknown samples is reported. For each of these, three spectra were acquired, therefore the measurement of lactic acid concentration associated with each sample is the average of the concentrations referred to the three

Table 2
Comparison of the lactic acid concentration predicted by the model and provided by HPLC for the five unknown samples.

Sample	Lactic acid concentration by HPLC (g/l)	Lactic acid concentration by Raman Spectroscopy (g/l)	Z-score on test set	P-value
1	0.06 ± 0.05	0.05 ± 0.04	−0.18	0.86
2	0.19 ± 0.06	0.04 ± 0.10	−1.55	0.12[1]
3	0.36 ± 0.07	0.38 ± 0.08	0.29	0.77
4	0.89 ± 0.10	0.90 ± 0.2	−0.14	0.89
5	1.2 ± 0.1	1.13 ± 0.04	−0.95	0.34

[1] The discrepancy in sample 2 is because the samples analysed are real samples of a complex nature (we are in fact going to determine the concentration of a particular acid in a matrix rich in other chemical compounds). Sample 2 satisfies the Z-test and has a p-value greater than 0.05. It is therefore a statistically significant sample.

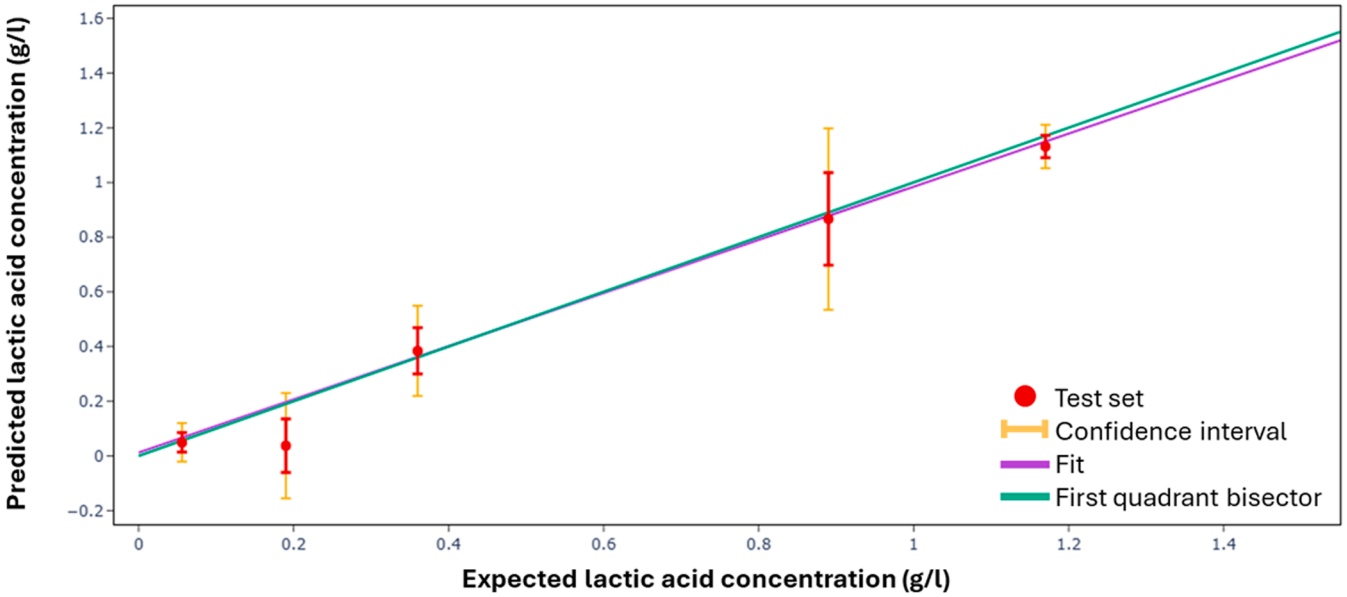

Fig. 6. Lactic acid concentration predicted by the model from Raman spectra as a function of that measured by HPLC for the test set.

spectra. The uncertainty, however, is equal to the sample standard deviation. Table 2 shows the values of lactic acid concentration measured by the two techniques, the Z-score and the p-value. The z-score was calculated as the distance in standard deviations between the observed measure and the population mean. From this, the p-value was then calculated using a normal distribution table for a two-tailed test. With a significance level set at 5 %, if the resulting p-value is greater than 0.05, this implies that the difference between the observed measure and the expected value is not statistically significant, indicating that the measures are compatible with the expected value.

From the values in Table 2, it can be observed that the results obtained by Raman spectroscopy are consistent with those obtained by HPLC. Performing a Z-test on the results of the unknown samples yields a p-value greater than 0.05 for all samples; we can therefore conclude that the concentrations provided by the model are not statistically significantly different from the values measured by HPLC.

Therefore, the predictions on the test samples are comparable with the expected values, and are accurate and precise; in fact, bias and precision are both equal to 0.07 g/l and $R^2 = 0.94$ (on the test set).

Conclusions

This work is part of a line of research aimed at developing precision oenology through the study and development of analytical methods capable of collecting real-time information on wine composition. These methods make it possible to monitor the evolution of wines during the vinification and ageing process and provide information to support decisions on the oenological practices to be carried out in the cellar (DSS decision supporting system). In this specific case, the study used Raman spectroscopy, machine learning techniques and multivariate modelling methods to quantify the concentration of lactic acid in Nebbiolo wine samples in order to indicatively monitor the phenomenon of malolactic fermentation. This new methodology is easy and user-friendly; in fact, the procedure followed to analyse the chemical compound under investigation consists of just two basic steps: acquisition of the sample's Raman spectrum and sending it to the multivariate regression algorithm. Finally, the latter will return the absolute concentration of lactic acid in the sample (in g/l). This methodology, unlike the conventionally used methods, e.g. HPLC, can allow on-site measurements (using portable Raman equipment). Furthermore, it is a non-destructive and fast technique (no sample preparation step required). Finally, it can be noted that the study is aimed at Nebbiolo wine samples for the production of Barolo wine. However, by modifying the pre-processing step of the spectra and re-training the model, it is possible to generalise the model to other wine types of similar or higher malic and lactic acid concentration.

The constructed PLS regression model showed good repeatability and robustness (bias and precision both 0.07 g/l) and is characterised by $R^2 = 0.76$ (on the validation set) and $R^2 = 0.94$ (on the test set) for concentrations in the range $[0,\ 1.6]$ g/l. Moreover, the RMSE of lactic acid concentration predicted by the model is 0.22 g/l (on the validation set) and 0.11 g/l (on the test set). In conclusion, the results obtained show that the Raman-PLS approach can be successfully used to determine the concentration of lactic acid during malolactic fermentation (directly in the winery) with a high accuracy and repeatability comparable to currently used methodologies such as HPLC.

CRediT authorship contribution statement

Anna Lisa Gilioli: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Alessio Sacco:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Andrea Mario Giovannozzi:** Writing – review & editing, Supervision. **Simone Giacosa:** Writing – review & editing. **Antonella Bosso:** Writing – review & editing, Resources, Project administration, Funding acquisition, Conceptualization. **Loretta Panero:** Writing – review & editing, Investigation, Data curation. **Silvia Raffaela Barera:** Writing – review & editing. **Stefano Messina:** Writing – review & editing. **Marco Lagori:** Writing – review & editing. **Silvia Motta:** Writing – review & editing. **Massimo Guaita:** Writing – review & editing. **Ettore Vittone:** Writing – review & editing, Supervision. **Andrea Mario Rossi:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

Declaration of competing interest

All authors declare that they have no conflicts of interest.

Acknowledgements

This work was funded through the PSR QualShelL by Regione Piemonte – Programma di Sviluppo Rurale (FEASR) operazione 16.1.1

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.talo.2025.100494.

Data availability

Data will be available upon reasonable request.

References

[1] E. Caplice, G.F. Fitzgerald, Food fermentations: role of microorganisms in food production and preservation, Int. J. Food Microbiol. 50 (1999) 131–149, https://doi.org/10.1016/S0168-1605(99)00082-3.
[2] I. Gil-Sánchez, B.B. Suáldea, M.V. Moreno-Arribas, Chapter 6 - malolactic fermentation, in: A. Morata (Ed.), Red Wine Technol, Academic Press, 2019, pp. 85–98, https://doi.org/10.1016/B978-0-12-814399-5.00006-2.
[3] E.F. López, E.F. Gómez, Simultaneous determination of the major organic acids, sugars, glycerol, and ethanol by HPLC in grape musts and white wines, J. Chromatogr. Sci. 34 (1996) 254–257, https://doi.org/10.1093/chromsci/34.5.254.
[4] S.B. Barker, W.H. Summerson, The colorimetric determination of lactic acid in biological material 138 (1941) 535–554.
[5] S. Chandra, J. Chapman, A. Power, J. Roberts, D. Cozzolino, Origin and regionality of wines—The role of molecular spectroscopy, Food Anal. Methods 10 (2017) 3947–3955, https://doi.org/10.1007/s12161-017-0968-1.
[6] P. Giménez-Gómez, M. Gutiérrez-Capitán, F. Capdevila, A. Puig-Pujol, C. Fernández-Sánchez, C. Jiménez-Jorquera, Monitoring of malolactic fermentation in wine using an electrochemical bienzymatic biosensor for L-lactate with long term stability, Anal. Chim. Acta 905 (2016) 126–133, https://doi.org/10.1016/j.aca.2015.11.032.
[7] P. Giménez-Gómez, M. Gutiérrez-Capitán, F. Capdevila, A. Puig-Pujol, C. Fernández-Sánchez, C. Jiménez-Jorquera, Robust L-malate bienzymatic biosensor to enable the on-site monitoring of malolactic fermentation of red wines, Anal. Chim. Acta 954 (2017) 105–113, https://doi.org/10.1016/j.aca.2016.11.061.
[8] L. Mandrile, G. Zeppa, A.M. Giovannozzi, A.M. Rossi, Controlling protected designation of origin of wine by Raman spectroscopy, Food Chem. 211 (2016) 260–267, https://doi.org/10.1016/j.foodchem.2016.05.011.
[9] S.B. Rodriguez, M.A. Thornton, R.J. Thornton, Discrimination of wine lactic acid bacteria by Raman spectroscopy, J. Ind. Microbiol. Biotechnol. 44 (2017) 1167–1175, https://doi.org/10.1007/s10295-017-1943-y.
[10] R.L. McCreery, Raman Spectroscopy for Chemical Analysis, John Wiley & Sons, 2005.
[11] Y. Qi, D. Hu, Y. Jiang, Z. Wu, M. Zheng, E.X. Chen, Y. Liang, M.A. Sadi, K. Zhang, Y. P. Chen, Recent progresses in machine learning assisted raman spectroscopy, Adv. Opt. Mater. 11 (2023) 2203104, https://doi.org/10.1002/adom.202203104.
[12] I.S. Arvanitoyannis, M.N. Katsota, E.P. Psarra, E.H. Soufleros, S. Kallithraka, Application of quality control methods for assessing wine authenticity: use of multivariate analysis (chemometrics), Trends. Food Sci. Technol. 10 (1999) 321–336, https://doi.org/10.1016/S0924-2244(99)00053-9.
[13] D. Cozzolino, W. Cynkar, N. Shah, P. Smith, Technical solutions for analysis of grape juice, must, and wine: the role of infrared spectroscopy and chemometrics, Anal. Bioanal. Chem. 401 (2011) 1475–1484, https://doi.org/10.1007/s00216-011-4946-y.
[14] Qualshell project, (n.d.). https://qualshell.com/.
[15] K. O'Keefe, Barolo and Barbaresco - the King and Queen of Italian Wine, University of California Press, 2014.
[16] A. Robles, M. Fabjanowicz, T. Chmiel, J. Płotka-Wasylka, Determination and identification of organic acids in wine samples. Problems and challenges, TrAC

Trends Anal. Chem. 120 (2019) 115630, https://doi.org/10.1016/j.trac.2019.115630.
[17] P. Cane, Il controllo di qualità dei vini mediante HPLC: determinazione degli acidi organici, Enotecnico 26 (1990) 69–72.
[18] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 contributors, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods 17 (2020) 261–272, https://doi.org/10.1038/s41592-019-0686-2.
[19] M.R. Maleki, A.M. Mouazen, H. Ramon, J.D. Baerdemaeker, Multiplicative scatter correction during on-line measurement with near infrared spectroscopy, Biosyst. Eng. 96 (2007) 427–433, https://doi.org/10.1016/j.biosystemseng.2006.11.014.
[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[21] K. Yoshida, Y. Shimizu, J. Yoshimoto, M. Takamura, G. Okada, Y. Okamoto, S. Yamawaki, K. Doya, Prediction of clinical depression scores and detection of changes in whole-brain using resting-state functional MRI data with partial least squares regression, PLoS One 12 (2017) 1–21, https://doi.org/10.1371/journal.pone.0179638.
[22] G.G. Dumancas, S. Ramasahayam, G. Bello, J. Hughes, R. Kramer, Chemometric regression techniques as emerging, powerful tools in genetic association studies, TrAC Trends Anal. Chem. 74 (2015) 79–88, https://doi.org/10.1016/j.trac.2015.05.007.
[23] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130, https://doi.org/10.1016/S0169-7439(01)00155-1.
[24] I.K. Nti, O. Nyarko-Boateng, J. Aning, others, performance of machine learning algorithms with different K values in K-fold cross-validation, Int. J. Inf. Technol. Comput. Sci. 13 (2021) 61–71.
[25] D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ. Comput. Sci. 7 (2021) e623, https://doi.org/10.7717/peerj-cs.623.
[26] V. Deneva, I. Bakardzhiyski, K. Bambalov, D. Antonova, D. Tsobanova, V. Bambalov, D. Cozzolino, L. Antonov, Using raman spectroscopy as a fast tool to classify and analyze Bulgarian wines—A feasibility study, Molecules 25 (2020), https://doi.org/10.3390/molecules25010170.
[27] Paul T Boggs, Janet R Donaldson, Orthogonal distance regression:, (1989). https://doi.org/10.6028/NIST.IR.89-4197.